

# *Proactive Transfer Learning*

Seungwhan Moon

CMU-LTI-18-003

Language Technologies Institute  
School of Computer Science  
Carnegie Mellon University  
5000 Forbes Ave., Pittsburgh, PA 15213  
[www.lti.cs.cmu.edu](http://www.lti.cs.cmu.edu)

**Thesis Committee:**

Jaime Carbonell, Chair

Yiming Yang

Louis-Philippe Morency

Jude Shavlik (University of Wisconsin-Madison)

*Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy  
In Language and Information Technologies*

© 2018 Seungwhan Moon

**Keywords:** Proactive Transfer Learning, Transfer learning, Proactive learning, Multi-modal learning, Zeroshot learning, Knowledge graph embeddings, Negative transfer problems,

*To my family I dedicate this thesis.*



## Abstract

Humans learn from heterogeneous knowledge sources and modalities - through various books, lectures, communications, textual or visual - and yet given a novel task we are able to leverage the combined knowledge base to make comprehensive inferences for solutions. When necessary, we also actively mine and query diverse sources of information to acquire knowledge we want. Hence, learning is *combinatorial* across heterogeneous sources, and *proactive* in novel knowledge acquisition.

In this thesis, we exploit the proactive transfer learning framework which follows this metaphor and provides a unique solution for learning a novel low-resource task by (1) leveraging multiple existing heterogeneous knowledge sets, as well as by (2) querying an optimized subset of target samples to produce additional maximally-useful annotations. Specifically, we study a completely heterogeneous transfer learning (CHTL) task where source and target datasets differ in both feature and label spaces (*e.g.* text corpora in different languages, describing different topics). For the unique challenges of CHTL, we formulate a common latent subspace learning problem in which an optimal subset of source and target data are projected through shared network pathways, subsequently reducing the joint supervised and unsupervised loss. To ground heterogeneous labels into common space, we utilize embeddings obtained from an external knowledge graph or a language model. In addition, we describe a novel dynamic proactive learning (DPAL) task where we optimize queries with regards to multiple selection strategies (“learn to active-learn”), given multiple annotators with varying expertise. Lastly, by combining these two lines of work, we present the proactive transfer learning (PTL) framework which exploits unlabeled target samples while progressively improving *transferability* between source and target domains, thereby maximizing information gain from both transfer learning and conventional active learning approaches. We demonstrate the efficacy of the proposed framework via various low-resource multimodal transfer learning tasks, such as hetero-lingual text classifications, image-aided named entity disambiguation for social media posts, etc.



## Acknowledgments

Several years ago I embarked on this academic journey, fueled with curiosity about how I could make machines learn and think better, and filled partly with uncertainty, not knowing where to begin. Had it not been for the generous and patient support and guidance from numerous people, this work would not have been possible. While this journey is far from being over, at the completion of this thesis I would like to take a moment to acknowledge all the people who have inspired me to keep pedaling through.

I first and foremost would like to thank my advisor Jaime Carbonell for his enormous help, support and inspiration throughout this journey. With his profound knowledge and vision in various fields of machine learning, Jaime has helped me shape my research in proactive learning and transfer learning throughout numerous discussions. There certainly were many moments where my algorithms would not produce positive results - or things simply wouldn't work, hitting a dead-end right before paper deadlines. I cannot thank Jaime enough for always being patient and enthusiastic in those moments, encouraging me to persist and pursue innovative new directions. I am also thankful for the very institute that Jaime has founded and led ever since, which I am now proud to be an alum of.

Special gratitude goes to Professor Yiming Yang, Louis-Phillippe Morency, and Jude Shavlik, who I had the honor to have on my thesis committee. The feedback and guidance I received from my thesis committee was invaluable in expanding my ideas as well as in improving the details from my thesis proposal.

I was also fortunate to have opportunities to work at various places as an intern, learning invaluable skills in industry settings under the support of many mentors and advisors. I would like to first thank Samsung Advanced Institute of Technology which was my first academic foray into the deep learning research community. My mentors Hosub Lee, Yeha Lee and Young Sang Choi were patient and kind enough to help jumpstart me on this exciting path.

I became an Imagineer under the guidance of my wonderful mentors Gunhee Kim and Leonid Sigal at Disney Research. Their incredible work ethic and professionalism have always inspired me throughout my PhD studies. The research we conducted at Disney also led me to continue down the path of multimodal and transfer learning of textual and visual representations.

To the sounds of splashing waves on the sandy Venice beach at Snapchat Research, Leonardo Neves, Vitor Carvalho and I managed to write two papers and two patents over the summer - which says something about our work ethic. I learned what the expression "*work hard and play hard*" really meant.

During my last PhD internship at Facebook, I was challenged to work on some of the most complex and the largest problems in the world. I am grateful for my mentors Xiao Wu and Hongyan Zhou for allowing me to explore and drive this challenging and exciting research project, where I was able to combine and apply all

of the knowledge and skills I learned from my PhD training. I am extremely excited to continue my journey at Facebook after graduation.

Throughout various projects from class, internships, and externally funded programs, I had the privilege to have collaborated with a number of friends, colleagues, and mentors: Akash Bharadwaj, Vitor Carvalho, William Casey, Volkan Cirik, Chris Dyer, David French, Hyeju Jang, Peter Jansen, Yohan Jo, Eliezer Kanal, Gunhee Kim, Suyoun Kim, Adhiguna Kuncoro, Yu-Hsin Kuo, Brian Lindauer, Lara Martin, Calvin McCarter, Leonardo Neves, Saloni Potdar, Carolyn Rose, Qinlan Shen, Leonid Sigal, Khe Chai Sim, Nathan VanHoudnos, Haohan Wang, Bronwyn Woods and Evan Wright. All the research discussions we had over regular meetings, coffee, lunches, dinners, all-nighters, etc. not only helped me get through my PhD, but will also remain as the fun and colorful memories of my PhD studies.

My collegiate and graduate education was made possible through the generous financial grants from the various institutions. I wish to express my sincere gratitude and appreciation to Samsung Scholarship, Language Technologies Institute, and Franklin W. Olin College of Engineering.

I would also like to thank our wonderful university staff for their unfailing support and assistance with many of my last-minute requests: Stacey Young, Mary Jo Bensasi, Alison Day, Alison Chiocchi, and many more.

Special thanks goes to the Evive Station, a water fountain / automatic bottle cleaning machine installed on the 6th floor of the Gates building that has kept me hydrated throughout many difficult headache-pounding days. The machine also tells me that I have saved 2,522 plastic bottles throughout my PhD life, making me feel good about saving the planet.

The CMU Korean students community has always provided me with various support along the way. I am indebted to Gihyuk, Jihoon, Sungho, Suyoun, Youngsuk, Yesul, Jungeun, Joohyung, Junsung, Jisu, Jiyeon, Jihee, Taegyun, Youngwook, Kiwan, Serim, Junwoo, Sunhee, Seungmoon, Kijung, Sungwook, Daegun, Min Kyung, Eunyong, Eunjin, Euiwoong, Jay-Yoon, Jaejun, Jeongyeob, Hyeju, Jinhee, Se-Joon, Philgoo, and many more, for memories and laughs we shared and the supportive group we built.

Words cannot describe how thankful I am for my family, both in Korea and in the US. Their unconditional love and selfless nurturing has been instrumental in shaping me into the person I have become. While miles lie between us, my heart is always with them. I love you, and I cannot wait to see my baby nephew soon.

Lastly, to Mariah - thank you for always encouraging me, inspiring me, and supporting me as I strive to achieve my goals. Thank you for enduring this journey with me, through the pain of being apart on the other ends of the continent for 5 years, through the frustration of failures and the joy of accomplishments along the road. Without you, I would have been lost. A lifetime of adventures awaits us, and I am thrilled to be on it with the love of my life.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Overview of the Thesis . . . . .	2
1.2.1	Transfer Learning from Heterogeneous Data Sources . . . . .	2
1.2.2	Dynamic proactive learning from heterogeneous experts and selection strategies . . . . .	4
1.2.3	Proactive transfer learning . . . . .	5
1.2.4	Applications in multi-modal transfer learning . . . . .	6
1.3	Thesis Statement . . . . .	6
1.4	Organization of the Document . . . . .	7
<b>2</b>	<b>Completely Heterogeneous Transfer Learning (CHTL)</b>	<b>9</b>
2.1	Introduction . . . . .	9
2.2	Problem Formulation . . . . .	10
2.3	Proposed Approach . . . . .	11
2.3.1	Label Embeddings . . . . .	12
2.3.2	Unsupervised Representation Learning for Features . . . . .	13
2.3.3	CHTL Network . . . . .	13
2.4	Empirical Evaluation . . . . .	17
2.4.1	Baselines . . . . .	17
2.4.2	Simulation on Synthetic Datasets . . . . .	19
2.4.3	Application: Hetero-lingual Text Classification . . . . .	21
2.5	Related Work . . . . .	28
2.6	Summary . . . . .	30
<b>3</b>	<b>CHTL Applications: Multimodal Transfer Learning for Named Entity Disambiguation</b>	<b>31</b>
3.1	Introduction . . . . .	32
3.2	Proposed Methods . . . . .	34
3.2.1	Notations . . . . .	35
3.2.2	Textual and Visual Contexts Features . . . . .	36
3.2.3	Lexical Embeddings: Deep Levenshtein . . . . .	37
3.2.4	Label Embeddings from Knowledge Graph . . . . .	38
3.2.5	Deep Zeroshot MNED Network (DZMNED) . . . . .	39

3.3	Empirical Evaluation	40
3.3.1	Datasets	41
3.3.2	Baselines	41
3.3.3	Results	43
3.4	Related Work	45
3.5	Summary	46
<b>4</b>	<b>Multi-source CHTL with Unsupervised Transferrable Adversarial Network Training</b>	<b>49</b>
4.1	Introduction	49
4.2	Method	50
4.2.1	Transferable Adversarial Network (TAN)	50
4.3	Empirical Evaluation	52
4.3.1	Simulation on Synthetic Datasets: Multiple Sources	52
4.3.2	Application: Multimodal Text-aided Image Scene Recognition	54
4.4	Related Work	55
4.5	Summary	56
<b>5</b>	<b>Proactive Learning with Multiple Heterogeneous Labelers</b>	<b>57</b>
5.1	Introduction	57
5.2	Method	59
5.2.1	Proactive Learning with Multiple Domain Experts	59
5.2.2	Expertise Estimation	61
5.2.3	Density-based Sampling for Multi-classification Tasks	63
5.3	Experimental Evaluation	64
5.3.1	Multi-class Information Density	65
5.3.2	Multiple Experts	67
5.4	Summary	71
<b>6</b>	<b>Learn to Active Learn: Dynamic Active Learning with Multiple Selection Strategies</b>	<b>73</b>
6.1	Introduction	73
6.2	Problem Formulation	75
6.3	Dynamic Proactive Learning Framework	76
6.3.1	Feature Space	76
6.3.2	Learning	78
6.4	Empirical Evaluation	80
6.4.1	Task	80
6.4.2	Baselines	81
6.4.3	Results	82
6.5	Related Work	83
6.6	Summary	84

<b>7</b>	<b>Learn to Active Transfer: Dynamic Proactive Transfer Learning</b>	<b>87</b>
7.1	Introduction . . . . .	87
7.2	Methods . . . . .	88
7.2.1	Proactive Transfer Learning Strategies . . . . .	88
7.2.2	Integration with DPAL . . . . .	90
7.3	Empirical Evaluation . . . . .	91
7.4	Related Work . . . . .	94
7.5	Summary . . . . .	94
<b>8</b>	<b>Conclusions</b>	<b>97</b>
	<b>Bibliography</b>	<b>101</b>



# List of Figures

1.1	<b>Proactive Transfer Learning Overview.</b> Given a novel task with scarce labels, we address the following classes of problems: (1) Completely Heterogeneous Transfer Learning (CHTL), which leverage existing heterogeneous knowledge sets to extract relevant information (Section 1.2.1), (2) Dynamic Proactive Learning (DPAL) settings in which we query multiple oracles an optimized subset of target samples to produce additional maximally-useful annotations by learning an optimal strategy from data and taking into account cost and expertise of annotations (Section 1.2.2), and (3) Proactive Transfer Learning (PTL) framework which simultaneously addresses the above problems, finding the optimal balance (Section 1.2.3). . . . .	3
2.1	An illustration of the proposed approach. Source and target datasets lie in different feature spaces ( $\mathbf{x}_S \in R^{M_S}$ , $\mathbf{x}_T \in R^{M_T}$ ), and describe different categories ( $\mathcal{Z}_S \neq \mathcal{Z}_T$ ). First, categorical labels are embedded into the dense continuous vector space ( <i>e.g.</i> via text embeddings learned from unsupervised documents.) The objective is then to learn projections $\mathbf{f}$ , $\mathbf{g}$ , and $\mathbf{h}$ jointly such that $\mathbf{g}$ and $\mathbf{h}$ map the source and target data to the latent common feature space, from which $\mathbf{f}$ can project to the same space as the embedded label space. Note that the shared projection $\mathbf{f}$ is learned from both the source and the target datasets, thus we can more robustly predict a label for a projected instance by finding its nearest label term projection. The attention mechanism filters and suppresses irrelevant source samples, and the denoising autoencoder $\mathbf{g}^\theta$ and $\mathbf{h}^\theta$ improve robustness with unsupervised training. . . . .	12
2.2	Attentional Heterogeneous Transfer. The choice of $\alpha$ determines which samples to attend or to suppress in propagating the loss values to the transfer network. . . . .	15
2.3	The proposed method (a) and the baseline networks (b-e). At test time, the nearest neighbor-based models (a,c) return the nearest label in the embedding space ( $\mathcal{Y}$ ) to the projection of a test sample, whereas the $n$ -way softmax layer (SM) classifiers (b,d,e) are trained to produce categorical labels from their respective final projection. We use the notation $\mathbf{W}_-$ to refer to $\mathbf{W}_t$ and $\mathbf{W}_s$ , as they share the same architecture. . . . .	19

2.4	Dataset generation process. (a) Draw a pair of source and target label embeddings ( $\mathbf{y}_{S,m}, \mathbf{y}_{T,m}$ ) from each of $M$ Gaussian distributions, all with $\sigma = \sigma_{\text{label}}$ (source-target label heterogeneity). For a random projection $\mathbf{P}_S, \mathbf{P}_T$ , (b) Draw synthetic source samples from new Gaussian distributions with $\mathcal{N}(\mathbf{P}_S \mathbf{y}_{S,m}, \sigma_{\text{diff}})$ , $\forall m \in \{1, \dots, M\}$ . (c) Draw synthetic target samples from $\mathcal{N}(\mathbf{P}_T \mathbf{y}_{T,m}, \sigma_{\text{diff}})$ , $\forall m$ . The resulting source and target datasets have heterogeneous label spaces (each class randomly drawn from a Gaussian with $\sigma_{\text{label}}$ ), as well as heterogeneous feature spaces ( $\mathbf{P}_S \neq \mathbf{P}_T$ ). . . . .	20
2.5	Simulation results with varying source-target heterogeneity (X-axis: $\sigma_{\text{label}}$ , Y-axis: accuracy) at different $\%_{L_T}$ . Baselines: CHTL: ATT+AE (black solid; proposed approach), CHTL: ATT (red dashes), CHTL (green dash-dots), ZSL (blue dots). . . . .	21
2.6	<b>Hetero-lingual text classification learning curves</b> of error rates (%) on the target task at (a) $\%_{L_S} = 100\%$ and varying percentages of target samples available ( $\%_{L_T}$ ), and at (b) $\%_{L_T} = 0.1\%$ and varying percentages of source samples available ( $\%_{L_S}$ ), each experiment averaged over 10-fold runs. (a,b) Source: RCV-1, Target: FR, (c,d) S: RCV-1, T: CTK. . . . .	24
2.7	t-SNE visualization of the projected source (R8) and target (GR) instances, where (a), (b) are learned without the transferred knowledge (ZSL), and (c), (d) use the transferred knowledge (CHTL). . . . .	26
2.8	<b>Visualization of attention</b> (source: R8, target: GR). Shown in the figure is the 2-D PCA representation of source instances (blue circles), source instances with attention (black circles), and target instances (red triangles) projected in the embedded label space ( $\mathbb{R}^{M_E}$ ). Mostly the source instances that overlap with the target instances in the embedded label space are given attention during training. . . . .	28
3.1	Examples of (a) a traditional NED task, focused on disambiguating polysemous entities based on surrounding textual contexts, and (b) the proposed Multimodal NED task for short media posts, which leverages both visual and textual contexts to disambiguate an entity. Note that mentions are often lexically inconsistent or incomplete, and thus a fixed candidates generation method (based on exact mention-entity statistics) is not viable. . . . .	32
3.2	The main architecture of our Multimodal NED network. We extract contextual information from an image, surrounding words, and lexical embeddings of a mention. The modality attention module determines weights for modalities, the weighted projections of which produce label embeddings in the same space as knowledge-base (KB) entity embeddings. We predict a final candidate by ranking based on similarities with KB entity knowledge graph embeddings as well as with lexical embeddings. . . . .	35
3.3	Deep Levenshtein, which predicts approximate Levenshtein scores between two strings. As a byproduct of this model, the shared Bi-LSTM can produce lexical embeddings purely based on lexical property of character sequences. . . . .	37

4.1	(a) Standard GAN architecture and (b) Adverarial feature learning through Transferable Adversarial Network Architecture (TAN). For TAN, we define a unified generator $G$ for both source and target domains which generates samples in the chosen modality $m$ , represented with a learnable high-dimensional token parameter, from its label representation $F(x)$ . Then, a unified discriminator distinguishes between a real input (encoded) and a generated one for both source and target domains, each identified with its corresponding modality token. . . . .	50
4.2	Image scene classification task with varying amount of target image dataset labels (S: Wiki) with fully labeled text sources (a) Czech CTK and (b) RCV1 datasets. Knowledge graph embeddings from FreeBase are used as label embeddings. Chance accuracy (the most frequent class): 15.7%. . . . .	54
5.1	Comparison of error rates on the Diabetes 130 U.S. Hospitals Dataset with different cost ratios (when expertise was estimated via ground truth samples). The X-axis denotes the normalized total cost, and the Y-axis denotes the classification error. Our proposed methods are marked as * in the legends. . . . .	67
5.2	Comparison of error rates on the 20 Newsgroup Dataset with different cost ratios (when expertise was estimated via ground truth samples). The X-axis denotes the normalized total cost, and the Y-axis denotes the classification error. Our proposed methods are marked as * in the legends. . . . .	68
5.3	Final error rate at convergence as a function of the initial budgets set aside for expertise estimation on the UCI datasets, in proportion to the total budget spent to acquire labels until it reaches convergence. . . . .	72
6.1	Motivation for dynamic proactive learning, illustrated with UCI Landsat Satellite Dataset [59] projected on 2-dimensional space via PCA. Each row represents different stages (iteration=200 and 800) of active learning. Drawn in the background are the 5 class-dividing hyperplanes learned thus far, and each dot represents an unlabeled sample. (a) shows actual future improvement of the learner in cross-validation performance when a ground-truth label for a corresponding sample is obtained. (c)-(e) show utility values measured by various active learning strategies. While none of the single strategies serves as a consistently reliable indicator, (b) the dynamic proactive learning framework (DPAL) predicts utility values that match closely to actual future improvement. . . . .	74
6.2	Error rates at normalized cost of queried instances annotated by noiseless labelers, on (a) 20 newsgroups, (b) MNIST, (c) Covertypes datasets. . . . .	78
6.3	Error rates at normalized cost of queried instances annotated with a pool of noised labelers (labeler noise ratio = 0.3 for non-expertise classes, and 0 for expertise classes.), on (a) 20 newsgroups, (b) MNIST, (c) Covertypes datasets. . . . .	80
6.4	Normalized DPAL weight transitions for (a), (c), (e): a single noiseless labeler scenario and (b), (d), (f): multiple noised labelers scenario. . . . .	85

7.1	Ablation studies of DPAL network architecture. (a) Utility is measured for each active learning strategy, and an optimal linear weight is learned for element-wise composition of weighted utility. (b) A deep neural network replaces the linear weights, allowing for more complex composition of strategies (with less interpretability). (c) Encoded data features are appended as input to the DPAL network. (d) We employ an attention module to determine importance weight of each active learning strategy. . . . .	91
7.2	Proactive transfer learning results for various source and target dataset pairs. $X$ -axis: % of queried samples, $Y$ -axis: classification error rates. DPAL approaches combine multiple strategies to compose an optimal strategy. . . . .	92
7.3	<b>Visualization of DPAL selection strategy attention</b> on an example case of RCV1 $\rightarrow$ FR datasets pair. For each DPAL update step (column), the strategy attention module amplifies the most informative selection strategy (darker) while attenuating less important or noisy selection strategies (lighter). The model makes final predictions of the weighted utility based on weighted signals from all selection strategies. Strategies used - MD: Maximal marginal distribution overlap, PE: Projected embeddings entropy, D: target sample density, E: target sample class-conditional entropy. . . . .	93

# List of Tables

2.1	Overview of datasets. $ \mathcal{Z} $ : the number of categories. . . . .	22
2.2	<b>Hetero-lingual text classification</b> test accuracy (%) on the target task, given a fully labeled source dataset and a partially labeled target dataset ( $\%_{L_T}$ ), averaged over 10-fold runs. Label embeddings with word2vec. * and ** denote $p < 0.05$ and $p < 0.01$ paired t-test improvement over the non-CHTL baseline. . . . .	23
2.3	<b>CHTL with attention</b> test accuracy (%) on the target task, at varying $K$ (number of clusters for attention), averaged over 10-fold runs. $\%_{L_T} = 0.1$ . . . . .	25
2.4	<b>CHTL with varying label embedding methods</b> (W2V: word embeddings, G2V: knowledge graph embeddings, Wi ki : Wikipedia document representation, Rand: random vector embeddings): test accuracy (%) on the target task averaged over 10-fold runs. $\%_{L_T} = 0.1$ . Method: CHTL: 2fC+ATT+AE. . . . .	26
2.5	Comparison of performance (CHTL) with varying intermediate embedding dimensions, averaged over 10-fold runs. . . . .	27
3.1	NED performance on the <i>SnapCaptionsKB</i> dataset at Top-1, 3, 5, 10, 50 accuracies. The classification is over 1M entities. Candidates generation methods: N/A, or over a fixed number of candidates generated with methods: $m \rightarrow e$ hash list and kNN (lexical neighbors). * and ** denote $p < 0.05$ and $p < 0.01$ paired t-test improvement over its comparing baselines and ablation studies ( <b>non-zeroshot</b> vs. <b>zeroshot</b> (ours), <b>W+C</b> (ours) vs <b>W+C+V</b> (ours), <b>non-transfer</b> (ours) vs. <b>transfer</b> (ours final). . . . .	43
3.2	MNED performance (Top-1, 5, 10 accuracies) on SnapCaptionsKB with varying qualities of KB embeddings. Model: DZMNED (W+C+V) . . . . .	44
3.3	Error analysis: <b>when do images help NED?</b> Ground-truth (GT) and predictions of our model with vision input (W+C+V) and the one without (W+C) for the <u>underlined</u> mention are shown. For interpretability, visual tags (label output of InceptionNet) are presented instead of actual feature vectors. . . . .	45
4.1	Target task accuracy with varying source-target heterogeneity ( $\sigma_{\text{label}}$ ). 1 source is used unless otherwise noted. . . . .	53
5.1	Overview of Datasets. . . . .	65
5.2	Comparison of error rates of MCID vs DWUS vs US . . . . .	66
5.3	Comparison of error rates on the UCI Datasets (when expertise was estimated via ground truth samples) . . . . .	69

5.4	Comparison of error rates (when expertise was estimated via majority vote) . . .	70
6.1	Overview of datasets. . . . .	81
6.2	Normalized proactive learning costs at error rate convergence for each dataset with varying DPAL $ L^0 / L $ ratios. Bold denotes the best performance for each test, and * denotes the statistically significant improvement ( $p < 0.05$ ). . . . .	81

# Chapter 1

## Introduction

### 1.1 Motivation

We as humans learn from heterogeneous knowledge sources and modalities - through various books, lectures, communications, textual or visual, etc. - and yet given a novel task, we are able to leverage the combined knowledge base and “connect the dots” to make comprehensive inferences for solutions. When necessary, we also actively *mine* diverse sources of information to acquire specific knowledge we want. Hence, learning is *combinatorial* across heterogeneous sources, and *proactive* in novel knowledge acquisition.

In machine learning, information also lies in heterogeneous forms, often as in different tabular datasets with unique feature and label sets across diverse domains. However, most of the standard machine learning algorithms build and train a model around a single tabular dataset, and thus the model learning process is practically isolated and independent for each task. While models with complex computational capability such as deep neural networks allow for scalability with respect to a large single-source dataset, these approaches are not scalable or applicable to a large number of novel and often low-resourced tasks. Consider, for example, a state-of-the-art deep visual network trained on billions of labeled images, which achieves near or beyond human performance in most object categorization tasks. The application of this trained visual model is limited, however, in that it can only take as input images in the same feature format, and that it can only be used for the same single discriminative task - while its conceptual understanding of the visual world in general may be able to aid other cognitive tasks such as natural language understanding, etc. that share common contextual grounds.

In order for a machine to be intelligent and independent enough to cope with diverse novel

tasks with varying availability of resource, it needs an ability to acquire knowledge from already-abundant related source of information, albeit heterogeneous in nature to target tasks, without requiring a heavily curated, task-specific dataset. In addition, when there does not exist enough relevant information readily available, it needs to be able to actively query novel knowledge from oracles (*e.g.* human annotators, experiments, etc.), preferably in an optimized manner to avoid inefficient or redundant information acquisition. Lastly, it is important that these two approaches for acquiring novel knowledge are simultaneously optimized to create a synergy with regards to each other in order to best improve target task performance - *e.g.* by querying oracles for unlabelled knowledge *in order to* most effectively transfer additional knowledge from existing sources, or in reverse direction, by transferring knowledge from sources to better identify crucial knowledge area to query.

To this end, we propose a general framework called *Proactive Transfer Learning* which addresses and implements such crucial abilities, formulated as a joint optimization task with multiple sub-problems.

## 1.2 Overview of the Thesis

In this thesis, we primarily focus on a low-resource domain learning scenario (scarce label data) where conventional data-driven supervised machine learning approaches are not successful. There are two main branches of literature in machine learning that address a low resource scenario: The first is **transfer learning**, which aims at utilizing knowledge obtained from a related source into a target task. While transfer learning has been successful in many cases, most of the approaches require that related source domains to be in homogeneous features or labels space, confining the scope of its applicability. The second line of work is called **active learning**, which assumes availability of oracles ('teachers'), and actively queries optimized subset of unlabeled examples for labels to progressively build a better model.

Advancing the previous literature in these two areas, we exploit a new **proactive transfer learning** framework (Figure 1.1) which addresses the following classes of problems:

### 1.2.1 Transfer Learning from Heterogeneous Data Sources

We study a transfer learning framework where source and target datasets are heterogeneous in both feature and label spaces, hence called *Completely Heterogeneous Transfer Learning*

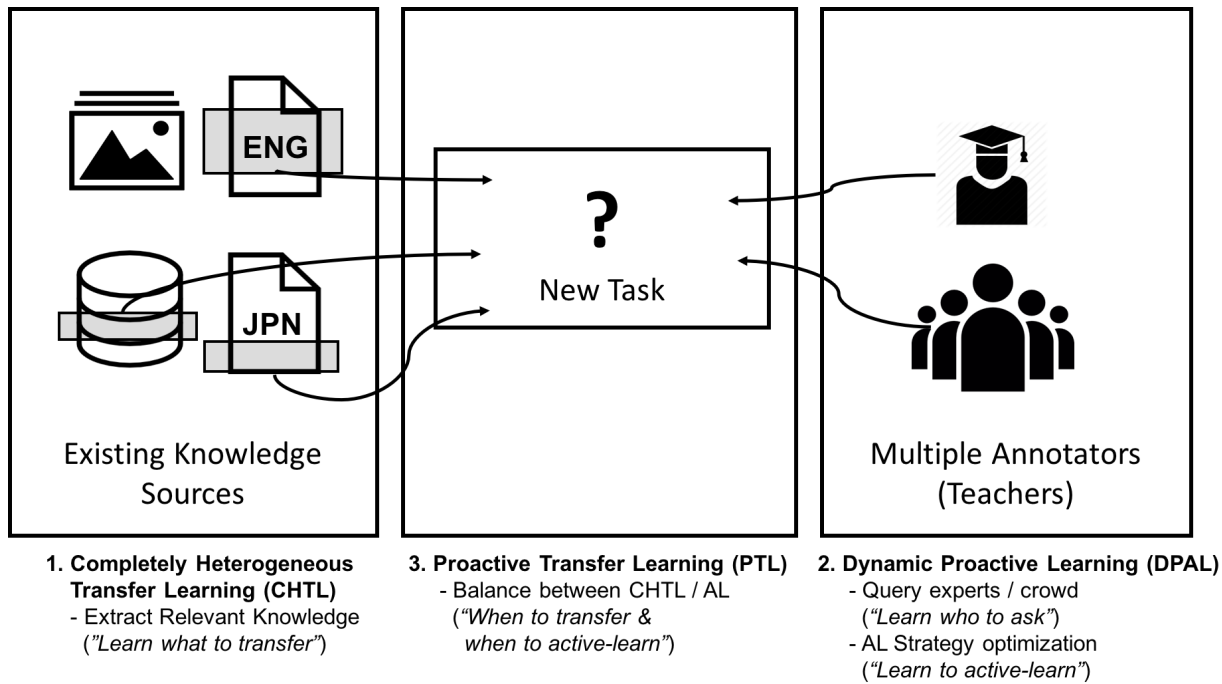


Figure 1.1: **Proactive Transfer Learning Overview**. Given a novel task with scarce labels, we address the following classes of problems: (1) Completely Heterogeneous Transfer Learning (CHTL), which leverage existing heterogeneous knowledge sets to extract relevant information (Section 1.2.1), (2) Dynamic Proactive Learning (DPAL) settings in which we query multiple oracles an optimized subset of target samples to produce additional maximally-useful annotations by learning an optimal strategy from data and taking into account cost and expertise of annotations (Section 1.2.2), and (3) Proactive Transfer Learning (PTL) framework which simultaneously addresses the above problems, finding the optimal balance (Section 1.2.3).

(CHTL). Unlike traditional transfer learning approaches, we do not require explicit relations between source and target tasks, such as source-target correspondent parallel dataset or label space homogeneity. CHTL has the following key components:

- **Common subspace learning** (Chapter 2): In order to ground heterogeneous source and target tasks into a common subspace, we formulate a joint learning problem in which each heterogeneous input source is mapped to a unified label embeddings space. To represent categorical labels as label embeddings, we borrow word embeddings obtained from a separate language model and knowledge graph embeddings transferred from an external knowledge base. We formulate the training objective based on combined hinge rank loss for both source and target classification tasks in the label embedding space [72].

- **Attentional transfer - “what and what not to transfer”** (Chapter 2): As we do not assume explicit relations between source and target tasks *a priori*, it is crucial to determine *what and what not to transfer* from source knowledge which has varying degree of relatedness. Hence, we define the attentional transfer module that selects an optimized subset of source data to maximize information gain from transferred knowledge. Specifically, we pre-cluster source samples into discrete sets based on their semantic proximity, and build the attentional transfer module (cluster-wise weighted average with softmax layer over sources) to attenuate or amplify the supervised loss back-propagation per each cluster [74].
- **Unsupervised knowledge transfer** (Chapter 4): In order to leverage vastly abundant unlabeled source and target data, we develop several unsupervised learning techniques for CHTL networks. Specifically, we propose the generative adversarial learning architecture for CHTL which take as input encoded source instances and a high-dimensional token that represents each modality (learnable parameter), from which the unified generator network generates an adversarial output in label embeddings space. The main CHTL network is trained via min-max training between the generator and the discriminator in addition to supervised training, producing a more robust classifier for label embeddings space. The training objective shares most of the components and pathways of the main network with the supervised training objectives, hence making a more robust classifier especially for low-resourced source tasks [75].

## 1.2.2 Dynamic proactive learning from heterogeneous experts and selection strategies

We also study an active learning framework which aims at acquiring optimized subset of labels given a fixed budget (*e.g.* asking minimal number of questions to oracles while maximizing information gain). Specifically, we study optimality of sample complexity with regards to (1) multiple annotators with varying expertise, and (2) multiple active learning selection strategies:

- **Query optimization among multiple annotators - “Learn who to ask”** (Chapter 5): Traditional active learning approaches often assume the existence of a single omniscient labeling “oracle”, whereas in real world scenarios it is more common to have multiple sources of annotations with different reliabilities or areas of expertise. To enhance practical reach of active learning algorithms, we formulate an iteration-wise greedy utility opti-

mization problem where utility is defined for each annotator and sample pair as expected information gain attenuated by its unique cost and estimated expertise [71, 76].

- **Selection strategies optimization - “Learn to active-learn”** (Chapter 6): Most of the recent work on active learning do not address the time-varying progress of learner’s knowledge state, and instead adhere to a static policy invariant with respect to new observations, resulting in non-optimal instance selection. We therefore design an adaptive proactive learning framework that can address beyond determining when to explore or to exploit, and aim at finding exactly what and how much to explore or exploit, with a careful balance that can approximate the ground-truth optimal strategy. This balance pattern is highly dependent on underlying distribution of a dataset as well as stream of labels obtained from annotators, and thus optimal selection strategies cannot be known *a priori*. As such, we develop a separate *strategy network* which *learns to active-learn* by composing a new strategy as a weighted ensemble of multiple existing strategies, where optimal weights are learned from active learning history data [73].

### 1.2.3 Proactive transfer learning

Given two options of addressing a low resource domain learning - active learning and transfer learning - we advocate for a system which can optimize with regards to each other. Specifically, we identify two synergistic active learning objectives - querying to improve transfer of existing knowledge, and querying to expand knowledge from unlabeled target set. Towards this goal, we define a dynamic proactive transfer learning framework which (1) progressively builds *bridges* between target and source domains in order to improve transfer accuracy (hence improving target task accuracy), and (2) exploit or explore target domains to improve target task accuracy, where the strategy is balanced and optimized adaptive to dataset pairs and active learning phases.

- **Estimation of transferred knowledge information gain from label expansion - “Transfer bridges”** (Chapter 7): We study a novel task of evaluating an unlabeled target sample for its expected information gain (when labeled) by its contribution to more robust knowledge transfer from source data. We propose several metrics that characterize these aspects, such as source-target marginal density overlap in the projected label embeddings space, etc. [74].
- **Dynamic proactive learning** (Chapter 7): We study a unique framework where active learning and transfer learning approaches are evaluated concurrently. We first characterize

expected information gain from a new label by (1) traditional active learning strategies such as uncertainty and density-based utilities, and (2) its expected gain from knowledge transfer, as discussed above. Lastly, we employ our dynamic active learning approaches to learn an optimized and balanced strategy that ultimately improves the target task accuracy the most.

## 1.2.4 Applications in multi-modal transfer learning

We evaluate the efficacy of the proposed approaches in various multi-modal learning domains. We also run the proposed algorithms on exhaustive simulation settings with varying control parameters to measure empirical characteristics of the method.

- **Hetero-lingual text classification (text; various languages)** (Chapter 2): We study a low-resource text classification task for various languages, where classification is aided by transferred knowledge from other text classification datasets of different languages and classes (topics). We demonstrate the efficacy of the proposed algorithm with various configurations of text datasets, such as 20Newsgroups, RCV1/2, Wikipedia articles dataset, etc. [72, 74].
- **Multimodal scene classification (image + text)** (Chapter 4): We also study an image-based scene classification task (target), leveraging transferred knowledge from heterogeneous (non-parallel) text datasets as source [75]. We use Wikipedia multimodal dataset to demonstrate the efficacy.
- **Multimodal named entity disambiguation (text + image)** (Chapter 3): We show that the CHTL training can be applied for a new NLP + Computer Vision multimodal learning task we introduce called Multimodal named entity disambiguation, which leverages both text and an accompanying image in order to disambiguate entities in text. We show that the proposed multimodal approaches outperform other conventional text-only approaches, and the performance further improves when knowledge is transferred from heterogeneous unsupervised images. We use Snapchat Captions dataset for this task [77, 78].

## 1.3 Thesis Statement

Given a novel low-resource task, a key ability of an intelligent learning framework includes inferring knowledge from existing heterogeneous sources (*transfer learning*), and actively querying

oracle(s) for knowledge acquisition (*active learning*). I show that the proposed proactive transfer learning framework can produce a unified representation where adequate subset of heterogeneous source knowledge can be drawn in inferring of a novel task, and optimally decide to acquire more information from oracles to improve transfer accuracy or target performance, or both. I discuss applications of the proposed framework in various multimodal transfer learning tasks, such as hetero-lingual text classifications or multimodal named entity disambiguation, etc.

## 1.4 Organization of the Document

The organization of this thesis is as follows. In Chapter 2, we describe the proposed Completely Heterogeneous Transfer Learning (CHTL) framework which attempts to transfer knowledge from a data source with heterogeneous feature and label spaces. We then extend the framework that allows for unsupervised knowledge transfer via a model we propose called Transferable Adversarial Network (TAN) in Chapter 4. We provide a special application case study for the CHTL framework in the Multimodal Named Entity Disambiguation task that leverages transferred knowledge from both textual and visual contexts in Chapter 3. As for the active learning side of the contributions, we provide a formulation for our Dynamic Proactive Learning (DPAL) with multiple labelers (Chapter 5) as well as with multiple selection strategies (Chapter 6). Chapter 7 presents the proactive transfer learning (PTL) framework which combines the CHTL approaches in the DPAL framework, addressing the problem of *when and what to transfer*. We give our concluding remarks and proposed directions for future work in Chapter 8



# Chapter 2

## Completely Heterogeneous Transfer Learning (CHTL)

### 2.1 Introduction

The notion of enabling a machine to learn a new task by leveraging an auxiliary source of knowledge has long been the focus of transfer learning. While many different flavors of transfer learning approaches have been developed, most of these methods assume explicit relatedness between source and target tasks, such as the availability of source-target correspondent instances (*e.g.* multi-view / multimodal learning), or the class relations information for multiple datasets sharing the same feature space (*e.g.* zero-shot learning, domain adaptation), etc. These approaches have been effective in their respective scenarios, but very few limited studies have investigated learning from heterogeneous knowledge sources that lie in both different feature and label spaces. See Section 2.5 for the detailed literature review.

Given an unforeseen target task with limited label information, we seek to mine useful knowledge from a plethora of heterogeneous knowledge sources that have already been curated, albeit in different feature and label spaces. To address this challenging scenario we first need an algorithm to estimate *how* the source and the target datasets may be related. One common aspect of any dataset for a classification task is that each instance is eventually assigned to some abstract concept(s) represented by its category membership, which often has its own *name*. Inspired by the Deep Visual-Semantic Embedding (DeViSE) model [35] which assigns the unsupervised word embeddings to label terms, we propose to map heterogeneous source and target labels into the same vector embedding space, from which we can obtain their semantic class relations.

Specifically, we use word embeddings learned from a language model and entity embeddings induced from a knowledge graph to represent labels. Using information from the class relations as an anchor, we first attempt to uncover a shared latent subspace where both source and target features can be mapped. Simultaneously, we learn a shared projection from this intermediate layer into the final embedded labels space, from which we can predict labels using the shared knowledge. We learn these projections by optimizing for the joint supervised loss for both source and target, as well as the unsupervised auto-encoder loss for reconstructing source and target. Lastly, as we do not assume explicit relations between source and target tasks *a priori*, it is crucial to determine *what and what not to transfer* from multiple source knowledge. Towards this goal, we propose an attentional transfer module which selects and attends to an optimized subset of source data to transfer knowledge from, ignoring unnecessary or confounding source instances that exhibit a negative impact in learning the target task.

We evaluate the proposed combined approach on a unique learning problem of a *hetero-lingual* text classification task, where the objective is to classify a novel target text dataset given only a few labels along with a source dataset in a different language, describing different classes from the target categories. While this is a challenging task, the empirical results show that the proposed approach improves over the baselines.

The rest of the chapter is organized as follows: we position our approach in relation to the previous work in Section 2.5, and formulate the completely heterogeneous transfer learning (CHTL) problem in Section 2.2. Section 2.3 describes in detail the proposed proactive transfer learning framework and presents the optimization problem. The empirical results are reported and analyzed in Section 2.4, and we give our concluding remarks in Section 2.6.

**Our contributions** in this chapter are three-fold: we propose (1) a novel transfer learning method with both heterogeneous feature and label spaces, (2) a novel attentional transfer module that mitigates negative transfer by selectively amplifying or attenuating subsets of source knowledge, and (3) we evaluate the proposed approach with extensive simulation studies as well as a novel transfer learning problem, the *hetero-lingual* text classification task.

## 2.2 Problem Formulation

We formulate the proposed framework for learning a target multiclass classification task given a source dataset with heterogeneous feature and label spaces as follows: We first define a dataset for the target task  $\mathbf{T} = \{\mathbf{X}_T, \mathbf{Y}_T, \mathbf{Z}_T\}$ , with the target task features  $\mathbf{X}_T = \{\mathbf{x}_T^{(i)}\}_{i=1}^{N_T}$  for  $\mathbf{x}_T \in$

$\mathbb{R}^{M_T}$ , where  $N_T$  is the target sample size and  $M_T$  is the target feature dimension, the ground-truth labels  $\mathbf{Z}_T = \{\mathbf{z}_T^{(i)}\}_{i=1}^{N_T}$ , where  $\mathbf{z}_T \in \mathcal{Z}_T$  for a categorical target label space  $\mathcal{Z}_T$ , and the corresponding high-dimensional label descriptors  $\mathbf{Y}_T = \{\mathbf{y}_T^{(i)}\}_{i=1}^{N_T}$  for  $\mathbf{y}_T \in \mathbb{R}^{M_E}$ , where  $M_E$  is the dimension of the embedded labels, which can be obtained from *e.g.* unsupervised word embeddings, etc. We also denote  $L_T$  and  $UL_T$  as a set of indices of labeled and unlabeled target instances, respectively, where  $|L_T| + |UL_T| = N_T$ . For a novel target task, we assume that we are given zero or a very few labeled instances, thus  $|L_T| = 0$  or  $|L_T| \ll N_T$ . Similarly, we define a heterogeneous source dataset  $\mathbf{S} = \{\mathbf{X}_S, \mathbf{Y}_S, \mathbf{Z}_S\}$ , with  $\mathbf{X}_S = \{\mathbf{x}_S^{(i)}\}_{i=1}^{N_S}$  for  $\mathbf{x}_S \in \mathbb{R}^{M_S}$ ,  $\mathbf{Z}_S = \{\mathbf{z}_S^{(i)}\}_{i=1}^{N_S}$  for  $\mathbf{z}_S \in \mathcal{Z}_S$ ,  $\mathbf{Y}_S = \{\mathbf{y}_S^{(i)}\}_{i=1}^{N_S}$  for  $\mathbf{y}_S \in \mathbb{R}^{M_E}$ , and  $L_S$ , accordingly. For the source dataset we assume  $|L_S| = N_S$ . Note that in general, we assume  $M_T \neq M_S$  (heterogeneous feature space) and  $\mathcal{Z}_T \neq \mathcal{Z}_S$  (heterogeneous label space).

Our goal is then to build a robust classifier ( $\mathcal{X}_T \rightarrow \mathcal{Z}_T$ ) for the target task, trained with  $\{\mathbf{x}_T^{(i)}, \mathbf{y}_T^{(i)}, \mathbf{z}_T^{(i)}\}_{i \in L_T}$  as well as transferred knowledge from  $\{\mathbf{x}_S^{(i)}, \mathbf{y}_S^{(i)}, \mathbf{z}_S^{(i)}\}_{i \in L_S}$ .

## 2.3 Proposed Approach

Our approach aims to leverage a source data that lies in different feature and label spaces from a target task. Transferring knowledge directly from heterogeneous spaces is intractable, and thus we begin by obtaining a unified vector representation for different source and target categories. Specifically, we utilize (1) a skip-gram based language model that learns semantically meaningful vector representations of words, as well as (2) a knowledge graph model that learns embeddings for entities in accordance with their known relations. We then map our categorical source and target labels into either of the embedding spaces (Section 2.3.1). In parallel, we learn compact representations for the source and the target features that encode abstract information of the raw features (Section 2.3.2), which allows for more tractable transfer through affine projections. Once the label terms for the source and the target datasets are anchored in the word embedding space, we first learn projections into a new latent common feature space from the source and the target feature spaces ( $\mathbf{g}$  and  $\mathbf{h}$ ), respectively, from which  $\mathbf{f}$  maps the joint features into the embedded label space (Section 2.3.3(a)). Lastly, we add an attentional transfer module that filters subsets of source knowledge that hurts the transfer performance (Section 2.3.3(a)). Figure 2.1 shows the illustration of the proposed approach.

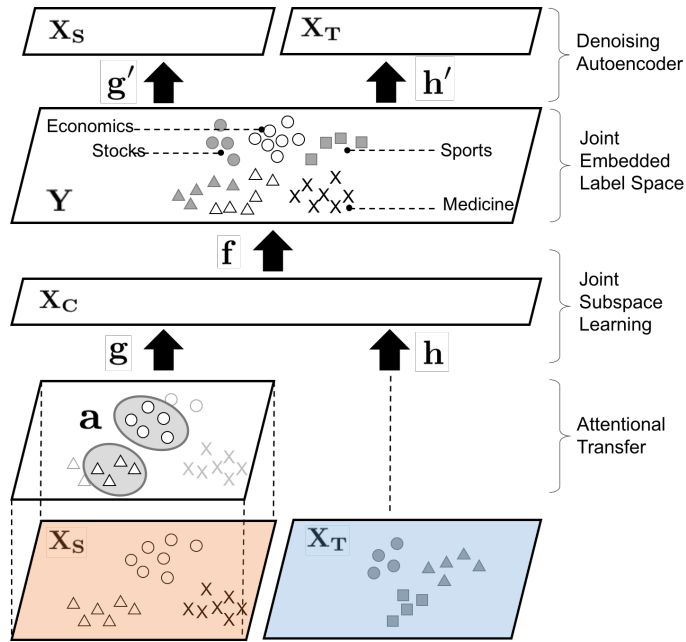


Figure 2.1: An illustration of the proposed approach. Source and target datasets lie in different feature spaces ( $\mathbf{x}_S \in \mathbb{R}^{M_S}$ ,  $\mathbf{x}_T \in \mathbb{R}^{M_T}$ ), and describe different categories ( $\mathcal{Z}_S \neq \mathcal{Z}_T$ ). First, categorical labels are embedded into the dense continuous vector space (e.g. via text embeddings learned from unsupervised documents.) The objective is then to learn projections  $f$ ,  $g$ , and  $h$  jointly such that  $g$  and  $h$  map the source and target data to the latent common feature space, from which  $f$  can project to the same space as the embedded label space. Note that the shared projection  $f$  is learned from both the source and the target datasets, thus we can more robustly predict a label for a projected instance by finding its nearest label term projection. The attention mechanism filters and suppresses irrelevant source samples, and the denoising autoencoder  $g^\theta$  and  $h^\theta$  improve robustness with unsupervised training.

### 2.3.1 Label Embeddings

**Language Model Based Label Embeddings:** The skip-gram based language model [69] has proven effective in encoding semantic information of words, which can be trained from unsupervised text. We use the obtained label term embeddings as *anchors* for source and target datasets, and drive the target model to learn indirectly from source instances that belong to semantically similar categories. In this work, we use 300-D word embeddings trained from the Google News dataset<sup>1</sup> (about 100 billion words).

<sup>1</sup>word2vec: <https://code.google.com/archive/p/word2vec/>

**Knowledge Graph Induced Label Embeddings:** We also use embeddings induced from knowledge graph entities and their relations [12, 80, 108]. On a nutshell, most of the methods attempt to build a model which takes as an input a triplet of two entities and their known relation, and outputs a score indicating correctness of the relation:

$$\mathcal{L} = \sum_{\mathcal{S}(\mathbf{y}_i, \mathbf{y}_j, \mathbf{r})} \sum_{(\mathbf{y}_i, \tilde{\mathbf{y}}, \mathbf{r})} l(f_{\mathbf{r}}(\mathbf{y}_i, \mathbf{y}_j), f_{\mathbf{r}}(\mathbf{y}_i, \tilde{\mathbf{y}})) \quad (2.1)$$

where  $\mathcal{S}(\mathbf{y}_i, \mathbf{y}_j, \mathbf{r})$  indicates all the known relations (positive triplets) where  $\mathbf{r}$  describes the relation between  $\mathbf{y}_i$  and  $\mathbf{y}_j$ ,  $(\mathbf{y}_i, \tilde{\mathbf{y}}, \mathbf{r})$  are negative triplets,  $f_{\mathbf{r}}$  determines the correctness score for a given triplet, and  $l$  determines the distance between a positive triplet and a negative triplet.

Specifically, we use Holographic Embeddings of Knowledge Graph (HOLE) to capture 300-dimensional vector embeddings of entities from WordNet [70].

**Wikipedia Document Representation:** Lastly, we take a tf-idf representation of each Wikipedia document, and use the dimension-reduced vectors via the latent semantic analysis (LSA) method [42] (dimension: 300). The assumption is that the resulting domain-free vectors would produce a high similarity score for related articles.

### 2.3.2 Unsupervised Representation Learning for Features

In order to project source and target feature spaces into the joint latent space effectively, as a pre-processing step we first obtain abstract and compact representations of raw features to allow for more tractable transformation. Unlike the similar zero-shot learning approaches [35], we do not use the embeddings obtained from a fully supervised network (e.g. the activation embeddings at the top of the trained visual model), because we assume the target task is scarce in labels. For our experiments with text features, we use LSA to transform the raw tf-idf features into a 200-D low-rank approximation.

### 2.3.3 CHTL Network

We describe the architecture for CHTL network as follows. We first define  $\mathbf{g} : \mathbb{R}^{M_s} \rightarrow \mathbb{R}^{M_c}$  and  $\mathbf{h} : \mathbb{R}^{M_T} \rightarrow \mathbb{R}^{M_c}$  to denote the projections with the respective sets of learnable parameters  $\mathbf{W}_g$  and  $\mathbf{W}_h$  that project source and target features into a latent common joint space, where the mappings can be learned with deep neural networks, kernel machines, etc. We consider feed-forward deep neural networks as well as a linear transformation layer for the choice of  $\mathbf{f}$ ,  $\mathbf{g}$ , and

h. Similarly, we define  $\mathbf{f} : \mathbb{R}^{M_C} \rightarrow \mathbb{R}^{M_E}$  with parameters  $\mathbf{W}_f$  which maps from the common feature space into the embedded label space.

We now describe several architectures for the CHTL algorithms to learn all of these parameters  $\mathbf{W} = \{\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h\}$  simultaneously.

### 2.3.3 (a) Feed-forward CHTL

We propose to solve the following joint optimization problem with hinge rank losses (similar to [35]) for both source and target.

$$\min_{\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h} \mathcal{L}_{\text{HR}}(\mathbf{S}, \mathbf{W}_g, \mathbf{W}_f) + \mathcal{L}_{\text{HR}}(\mathbf{T}, \mathbf{W}_h, \mathbf{W}_f) + \mathcal{R}(\mathbf{W}) \quad (2.2)$$

where

$$\mathcal{L}_{\text{HR}}(\mathbf{S}, \mathbf{W}_g, \mathbf{W}_f) = \frac{1}{|L_S|} \sum_{i=1}^{jL_S} \sum_{\tilde{\mathbf{y}} \in \mathbf{y}_S^{(i)}} \max[0, \epsilon - \mathbf{f}(\mathbf{g}(\mathbf{x}_S^{(i)})) \cdot \mathbf{y}_S^{(i)} + \mathbf{f}(\mathbf{g}(\mathbf{x}_S^{(i)})) \cdot \tilde{\mathbf{y}}]$$

$$\mathcal{L}_{\text{HR}}(\mathbf{T}, \mathbf{W}_h, \mathbf{W}_f) = \frac{1}{|L_T|} \sum_{j=1}^{jL_T} \sum_{\tilde{\mathbf{y}} \in \mathbf{y}_T^{(j)}} \max[0, \epsilon - \mathbf{f}(\mathbf{h}(\mathbf{x}_T^{(j)})) \cdot \mathbf{y}_T^{(j)} + \mathbf{f}(\mathbf{h}(\mathbf{x}_T^{(j)})) \cdot \tilde{\mathbf{y}}]$$

$$\mathcal{R}(\mathbf{W}) = \lambda_f \|\mathbf{W}_f\|^2 + \lambda_g \|\mathbf{W}_g\|^2 + \lambda_h \|\mathbf{W}_h\|^2$$

where  $\mathcal{L}_{\text{HR}}(\cdot)$  is the hinge rank loss for source and target,  $\tilde{\mathbf{y}}$  refers to the embeddings of other label terms in the source and the target label space except the ground truth label of the instance,  $\epsilon$  is a fixed margin,  $\mathcal{R}(\mathbf{W})$  is a weight decay regularization term, and  $\lambda_f, \lambda_g, \lambda_h \geq 0$  are regularization constants. We use  $\epsilon = 0.1$  for all of our experiments.

In essence, we train the weight parameters to produce a higher dot product similarity between the projected source or target instance and the word embedding representation of its correct label than between the projected instance and other incorrect label term embeddings. The intuition of the model is that the learned  $\mathbf{W}_f$  is a shared and more generalized linear transformation capable of mapping the joint intermediate subspace into the embedded label space.

We solve Eq.2.2 efficiently with stochastic gradient descent (SGD), where the gradient is estimated from a small minibatch of samples.

Once  $\mathbf{f}$ ,  $\mathbf{g}$ , and  $\mathbf{h}$  are learned, at test time we build a label-producing nearest neighbor (1-NN) classifier for the target task as follows:

$$1\text{-NN}(\mathbf{x}_T) = \underset{\mathbf{z} \in Z_T}{\operatorname{argmax}} \mathbf{f}(\mathbf{h}(\mathbf{x}_T)) \cdot \mathbf{y}_z \quad (2.3)$$

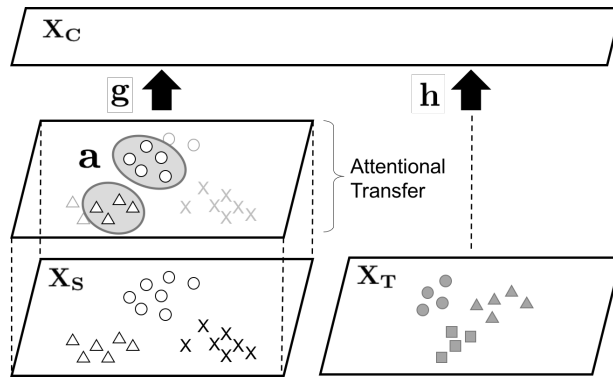


Figure 2.2: Attentional Heterogeneous Transfer. The choice of  $a$  determines which samples to attend or to suppress in propagating the loss values to the transfer network.

where  $y_z$  maps a categorical label term  $z$  into its word embeddings space. Similarly, we can build a NN classifier for the source task as well, using the projection  $f(g(\cdot))$ .

### 2.3.3 (b) Attentional Transfer at Instance Level - What And What Not To Transfer

Note that most of the previous approaches on transfer learning do not take into account different instance-level heterogeneity within a source dataset, often leading to undesirable *negative transfer*. Specifically, CHTL can suffer from brute-force merge of heterogeneous sources because explicit relations between source and target knowledge are not known in both instance and dataset-level.

We speculate that there are certain instances within the source task that are more likely to be *transferable* than other samples. Inspired by successes of attention mechanism from recent literature [14, 111], we propose an approach that selectively transfers useful knowledge by focusing only to a subset of source knowledge while avoiding others that may have a harmful impact on target learning. Specifically, the attention mechanism learns a set of parameters that specify a weight vector over a discrete subset of data, determining its relative importance or relevance in transfer. To enhance computational tractability we first pre-cluster the source dataset into  $K$  number of clusters  $S_1, \dots, S_K$ , and formulates the following joint optimization problem that learns the parameters for the transfer network as well as a weight vector  $\{\alpha_k\}_{k=1:K}$ :

$$\begin{aligned}
& \min_{\mathbf{a}, \mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h} \mu \sum_{k=1}^K \alpha_k \cdot \mathcal{L}_{\text{HR:K}}(\mathbf{S}_k, \mathbf{W}_g, \mathbf{W}_f) + \mathcal{L}_{\text{HR}}(\mathbf{T}, \mathbf{W}_h, \mathbf{W}_f) + \mathcal{R}(\mathbf{W}) \\
& \text{where} \\
& \alpha_k = \frac{\exp(\mathbf{a}_k)}{\sum_{k=1}^K \exp(\mathbf{a}_k)}, \quad 0 < \alpha_k < 1 \\
& \mathcal{L}_{\text{HR:K}}(\mathbf{S}_k, \mathbf{W}_g, \mathbf{W}_f) = \frac{1}{|L_{S_k}|} \sum_{i \in L_{S_k}} \sum_{\tilde{\mathbf{y}} \in \mathbf{y}_S^{(i)}} \max[0, \epsilon - \mathbf{f}(\mathbf{g}(\mathbf{x}_S^{(i)})) \cdot \mathbf{y}_S^{(i)} + \mathbf{f}(\mathbf{g}(\mathbf{x}_S^{(i)})) \cdot \tilde{\mathbf{y}}]
\end{aligned} \tag{2.4}$$

where  $\mathbf{a}$  is a learnable parameter that determines the weight for each cluster,  $\mathcal{L}_{\text{HR:K}}(\cdot)$  is a cluster-level hinge loss for the source dataset,  $L_{S_k}$  is a set of source indices that belong to a cluster  $\mathbf{S}_k$ , and  $\mu$  is a hyperparameter that penalizes  $\mathbf{a}$  and  $\mathbf{f}$  for simply optimizing for the source task only. Note that  $\mathbf{W}_f$  is shared by both source and target networks, and thus the choice of  $\mathbf{a}$  affects both  $\mathbf{g}$  and  $\mathbf{h}$ . Essentially, the attention mechanism works as a regularization over source dataset, suppressing the loss values for non-attended samples in knowledge transfer. Figure 2.2 illustrates the proposed approach.

We solve Eq.2.4 with a two-step alternating descent optimization. The first step involves optimizing for the source network parameters  $\mathbf{W}_g, \mathbf{a}, \mathbf{W}_f$  while the rest are fixed, and the second step optimizes for the target network parameters  $\mathbf{W}_h, \mathbf{W}_f$  while others are fixed.

In our experiments we use  $K$ -means clustering algorithm [4], but any other distance-based clustering algorithms can work as well.

### 2.3.3 (c) Denoising CHTL

We formulate unsupervised transfer learning with the CHTL architecture for added robustness, which is especially beneficial when labeled target data is scarce. Specifically, we add denoising auto-encoders where the pathway for predictions,  $\mathbf{f}$ , is shared and trained by both source and target through the joint subspace, thus benefiting from unlabelled source and target data. Finally, we formulate the CHTL learning problem with both supervised and unsupervised losses as follows:

$$\min_{\mathbf{a}, \mathbf{W}} \mu \sum_{k=1}^K \frac{\alpha_k}{|L_{S_k}|} \cdot \mathcal{L}_{\text{HR:K}}(\mathbf{S}_k) + \mathcal{L}_{\text{HR}}(\mathbf{T}) + \mathcal{L}_{\text{AE}}(\mathbf{S}, \mathbf{T}; \mathbf{W})$$

where

$$\begin{aligned} \mathcal{L}_{\text{AE}}(\mathbf{S}, \mathbf{T}; \mathbf{W}) = & \frac{1}{|UL_S|} \sum_{i=1}^{jUL_S} \|\mathbf{g}^\theta(\mathbf{f}(\mathbf{g}(\mathbf{x}_S^{(i)}))) - \mathbf{x}_S^{(i)}\|^2 \\ & + \frac{1}{|UL_T|} \sum_{j=1}^{jUL_T} \|\mathbf{h}^\theta(\mathbf{f}(\mathbf{h}(\mathbf{x}_T^{(j)}))) - \mathbf{x}_T^{(j)}\|^2 \end{aligned} \quad (2.5)$$

where  $\mathcal{L}_{\text{AE}}$  is the denoising auto-encoder loss for both source and target data (unlabelled),  $\mathbf{g}^\theta$  and  $\mathbf{h}^\theta$  reconstruct input source and target respectively, and the learnable weight parameters are defined as  $\mathbf{W} = \{\mathbf{W}_f, \mathbf{W}_g, \mathbf{W}_h, \mathbf{W}_g^\theta, \mathbf{W}_h^\theta\}$ . The final architecture is illustrated in Figure 2.1.

Algorithm 1 provides the detailed implementation steps of the CHTL approach. The algorithmic complexity for target label inference on CHTL is largely bounded by feed-forward prediction and search of the nearest label embeddings, and thus determined concisely by  $O((n_f + n_h) \cdot M_C^3 + |\mathcal{Z}_T| \cdot M_E)$ , where  $n_f$  and  $n_h$  denote the number of neural layers within function  $\mathbf{f}$  and  $\mathbf{h}$  respectively,  $M_C$  is the number of neurons in each layer,  $|\mathcal{Z}_T|$  refers to the number of known target categories, and  $M_E$  is the dimension of the label embeddings space. We assume the use of the naive matrix multiplication algorithm, which has a asymptotic run-time of  $O(n^3)$ .

## 2.4 Empirical Evaluation

We validate the effectiveness of the proposed approaches via extensive simulations (Section 2.4.2) as well as a real-world application in hetero-lingual text classification (Section 2.4.3) with the baselines described in Section 2.4.1.

### 2.4.1 Baselines

In our experiments we use a source dataset within heterogeneous feature and label spaces from a target dataset. Most of the previous transfer learning approaches that allow only one of input or output spaces to be heterogeneous thus cannot be used as baselines (see Section 2.5 for the detailed comparison). We therefore compare the proposed heterogeneous transfer approach with the following baseline networks (illustrated in Figure 2.3):

---

**Algorithm 1** Completely Heterogeneous Transfer Learning (CHTL)

---

**Input:** source data  $\mathbf{S}$ , target data  $\mathbf{T}$ , target sample  $\mathbf{x}_T$ , total number of batches  $B$

**A. CHTL Training (alternating back-prop)**

- Randomly initialize  $\mathbf{a}$ ,  $\mathbf{W}_f$ ,  $\mathbf{W}_T$ ,  $\mathbf{W}_S$  (truncated normal)

**while**  $batch_{++} \leq B$  **or** not converged **do**

1. Optimize  $\mathbf{W}_S$ ,  $\mathbf{a}$ ,  $\mathbf{W}_f$  by solving

$$\min_{\mathbf{W}_S, \mathbf{a}, \mathbf{W}_f} \mu \sum_{k=1}^K \frac{\alpha_k}{|L_{S_k}|} \cdot \mathcal{L}_{HR:K}(\mathbf{S}_k) + \mathcal{L}_{AE}(\mathbf{S}; \mathbf{W})$$

2. Optimize  $\mathbf{W}_T$ ,  $\mathbf{W}_f$  by solving

$$\min_{\mathbf{W}_T, \mathbf{W}_f} \mathcal{L}_{HR}(\mathbf{T}) + \mathcal{L}_{AE}(\mathbf{T}; \mathbf{W})$$

**end while**

**B. Inference with CHTL**

- Obtain projected label embeddings  $\mathbf{f}(\mathbf{h}(\mathbf{x}_T))$

- Initialize  $\min\_sim(\mathbf{x}_T) = \inf$

**for all**  $\mathbf{z} \in \mathcal{Z}_T$  **do**

$\text{sim}(\mathbf{x}_T, \mathbf{z}) = \mathbf{f}(\mathbf{h}(\mathbf{x}_T)) \cdot \mathbf{y}_z$  (similarity score in the projected label embeddings space)

$\min\_sim(\mathbf{x}_T)$ ,  $1\text{-NN}(\mathbf{x}_T) = \text{sim}(\mathbf{x}_T, \mathbf{z})$ ,  $\mathbf{z}$  if  $\min\_sim(\mathbf{x}_T) < \text{sim}(\mathbf{x}_T, \mathbf{z})$

**end for**

**return**  $1\text{-NN}(\mathbf{x}_T)$

---

- CHTL: ATT+AE (**proposed approach**; completely heterogeneous transfer learning (CHTL) network with attention and auto-encoder loss): the model is trained with the joint optimization problem in Eq.2.5.
- CHTL: ATT (CHTL with attention only): the model is trained with the optimization problem in Eq.2.4. We evaluate this baseline to isolate the effectiveness of the attention mechanism.
- CHTL (CHTL without attention or auto-encoder loss; [72]): the model is trained with the model without attention as in Eq.2.2.
- ZSL (Zero-shot learning networks with distributed word semantic embeddings; [35]): the model is trained for target dataset only with label embeddings  $\mathbf{Y}_T$  obtained from a dis-

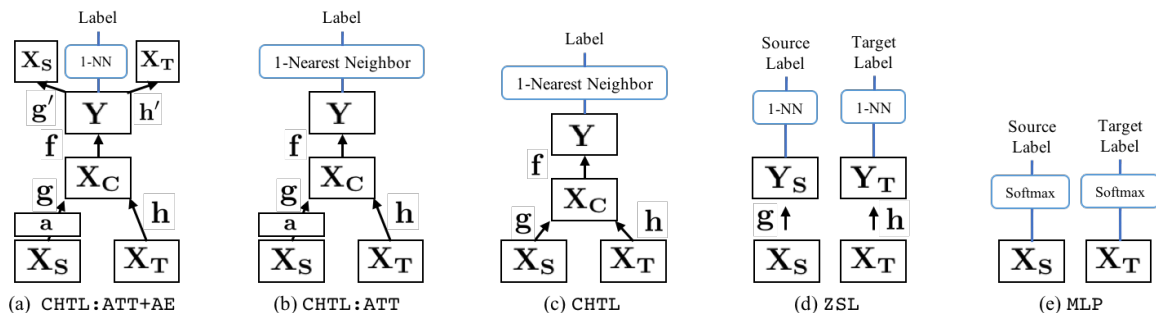


Figure 2.3: The proposed method (a) and the baseline networks (b-e). At test time, the nearest neighbor-based models (a,c) return the nearest label in the embedding space ( $\mathcal{Y}$ ) to the projection of a test sample, whereas the  $n$ -way softmax layer (SM) classifiers (b,d,e) are trained to produce categorical labels from their respective final projection. We use the notation  $\mathbf{W}_-$  to refer to  $\mathbf{W}_t$  and  $\mathbf{W}_s$ , as they share the same architecture.

tributed word semantics language model. We solve the following optimization problem:

$$\min_{\mathbf{W}_T} \frac{1}{|L_T|} \sum_{j=1}^{|L_T|} l(\mathbf{T}^{(j)}) \quad (2.6)$$

where the loss function is defined as follows:

$$l(\mathbf{T}^{(j)}) = \sum_{\tilde{y} \in \mathcal{Y}_T^{(j)}} \max[0, \epsilon - \mathbf{h}(\mathbf{x}_T^{(i)}) \cdot \mathbf{y}_T^{(j)} + \mathbf{h}(\mathbf{x}_T^{(j)}) \cdot \tilde{y}] + \mathcal{R}(\mathbf{W}_h)$$

- ZSL: AE (ZSL with autoencoder loss): we add the autoencoder loss to the objective to 2.6.
- MLP (A feedforward multi-layer neural network): the model is trained for a target dataset only with categorical labels.

For each of the CHTL variations, we vary the number of fully connected (FC) layers (*e.g.* 2fc, 3fc,  $\dots$ ) as well as the label embedding methods as described in Section 2.3.1 (semantic word embeddings (W2V); knowledge graph induced embeddings (G2V); wikipedia document representations (Wiki)); random vector embeddings (Rand)).

## 2.4.2 Simulation on Synthetic Datasets

We generate multiple pairs of source and target synthetic datasets and evaluate the performance with average classification accuracies on target tasks. Specifically, we aim to analyze the per-

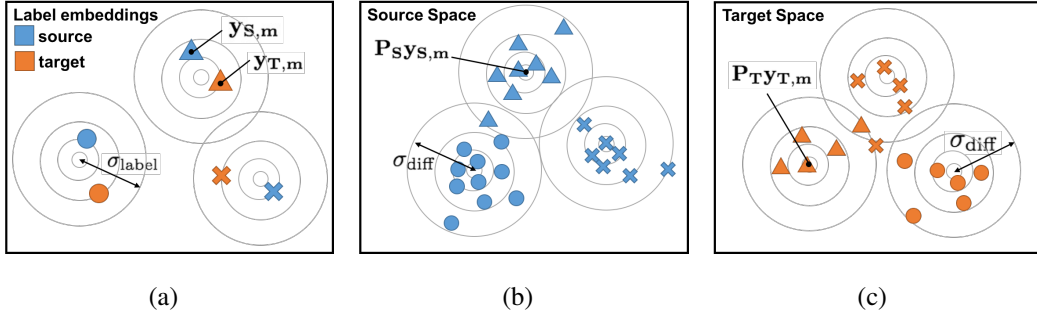


Figure 2.4: Dataset generation process. (a) Draw a pair of source and target label embeddings  $(\mathbf{y}_{S,m}, \mathbf{y}_{T,m})$  from each of  $M$  Gaussian distributions, all with  $\sigma = \sigma_{\text{label}}$  (source-target label heterogeneity). For a random projection  $\mathbf{P}_S, \mathbf{P}_T$ , (b) Draw synthetic source samples from new Gaussian distributions with  $\mathcal{N}(\mathbf{P}_S \mathbf{y}_{S,m}, \sigma_{\text{diff}}), \forall m \in \{1, \dots, M\}$ . (c) Draw synthetic target samples from  $\mathcal{N}(\mathbf{P}_T \mathbf{y}_{T,m}, \sigma_{\text{diff}}), \forall m$ . The resulting source and target datasets have heterogeneous label spaces (each class randomly drawn from a Gaussian with  $\sigma_{\text{label}}$ ), as well as heterogeneous feature spaces ( $\mathbf{P}_S \neq \mathbf{P}_T$ ).

formance of the proposed approaches with varying source-target heterogeneity at varying task difficulty levels.

**Datasets Generation** (Figure 2.4): we generate synthetic source and target datasets each with  $M$  different classes,  $\mathbf{S} = \{\mathbf{X}_S, \mathbf{Y}_S\}$ , and  $\mathbf{T} = \{\mathbf{X}_T, \mathbf{Y}_T\}$ , such that their embedded label space are heterogeneous with a controllable hyperparameter  $\sigma_{\text{label}}$ . We first generate  $M$  isotropic Gaussian distributions  $\mathcal{N}(\mu_m, \sigma_{\text{label}})$  for  $m \in \{1, \dots, M\}$ . From each distribution we draw a pair of source and target label embeddings  $\mathbf{y}_{S,m}, \mathbf{y}_{T,m} \in \mathbb{R}^{M_E}$ . Intuitively, source and target datasets are more heterogeneous with a higher  $\sigma_{\text{label}}$ , as the drawn pair of source and target embeddings is farther apart from each other. We then generate source and target samples each with a random projection  $\mathbf{P}_S \in \mathbb{R}^{M_S \times M_E}, \mathbf{P}_T \in \mathbb{R}^{M_T \times M_E}$  as follows:

$$\begin{aligned} \mathbf{X}_{S,m} &\sim \mathcal{N}(\mathbf{P}_S \mathbf{y}_{S,m}, \sigma_{\text{diff}}), \quad \mathbf{X}_S = \{\mathbf{X}_{S,m}\}_{1 \leq m \leq M} \\ \mathbf{X}_{T,m} &\sim \mathcal{N}(\mathbf{P}_T \mathbf{y}_{T,m}, \sigma_{\text{diff}}), \quad \mathbf{X}_T = \{\mathbf{X}_{T,m}\}_{1 \leq m \leq M} \end{aligned}$$

where  $\sigma_{\text{diff}}$  affects the label distribution classification difficulty. We denote  $\%_{L_T}$  as the percentage of target samples labeled, and assume that only a small fraction of target samples is labeled ( $\%_{L_T} \ll 1$ ).

For the following experiments, we set  $N_S = N_T = 4000$  (number of samples),  $M = 4$  (number of source and target dataset classes),  $M_S = M_T = 20$  (original feature dimension),  $M_E = 15$  (embedded label space dimension),  $K = 12$  (number of attention clusters),  $\sigma_{\text{diff}} =$

0.5,  $\sigma_{\text{label}} \in \{0.05, 0.1, 0.2, 0.3\}$ , and  $\%_{L_T} \in \{0.005, 0.01, 0.02, 0.05\}$ . We repeat the dataset generation process 10 times for each parameter set. We obtain 5-fold results for each dataset generation, and report the overall average accuracy in Figure 2.5.

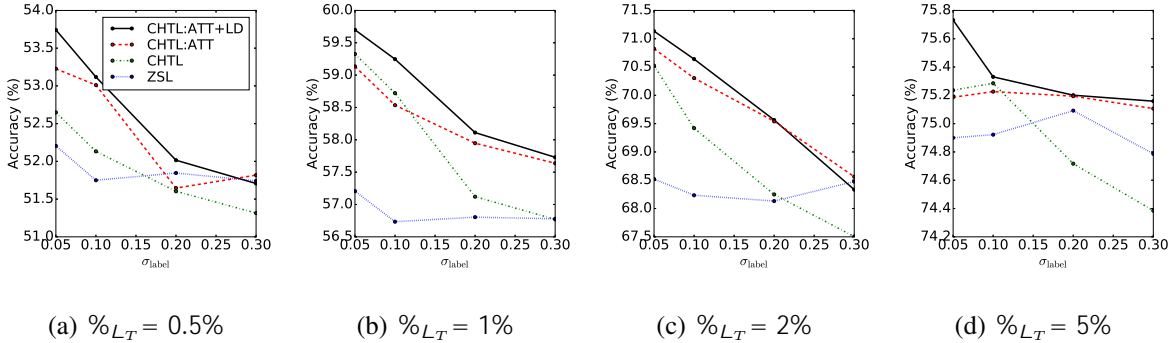


Figure 2.5: Simulation results with varying source-target heterogeneity (X-axis:  $\sigma_{\text{label}}$ , Y-axis: accuracy) at different  $\%_{L_T}$ . Baselines: CHTL: ATT+AE (black solid; proposed approach), CHTL: ATT (red dashes), CHTL (green dash-dots), ZSL (blue dots).

**Sensitivity to Source-Target Heterogeneity:** each subfigure in Figure 2.5 shows the performance of the baselines with varying  $\sigma_{\text{label}}$  (source-target heterogeneity). In general, the CHTL baselines outperforms ZSL, but the performance degrades as heterogeneity increases. However, the attention mechanism (CHTL: ATT+AE and CHTL: ATT) is generally effective with higher source-target heterogeneity, suppressing the performance drop. Note that the performance improves in most cases when the attention mechanism is combined with the auto-encoder loss (AE).

**Sensitivity to Target Label Scarcity:** we evaluate the tolerance of the algorithm at varying target task difficulty, measured with varying percentage of target labels given. Evidently, providing a smaller number of target labels makes the task more difficult, and thus Figure 2.5(a) shows the lowest target task accuracy across every baseline. Note that when a small number of labels are given (Figure 2.5(a)), the improvement due to CHTL algorithms is weak, indicating that the transfer learning algorithm requires a sufficient number of target labels to build proper anchors with source knowledge. Note also that while the performance gain of CHTL algorithms begins to degrade as the target task approaches the saturation error rate (Figure 2.5(d)), the attention mechanism (CHTL: ATT) is more robust to this degradation and avoids negative transfer.

### 2.4.3 Application: Hetero-lingual Text Classification

We apply the proposed approach to learn a target text classification task given a source text dataset with both a heterogeneous feature space (*e.g.* a different language) and a label space (*e.g.*

Table 2.1: Overview of datasets.  $|\mathcal{Z}|$ : the number of categories.

Dataset	$ \mathcal{Z} $	Label Terms ( <i>e.g.</i> )
RCV-1 (RCV1)	116	‘forecasts’, ‘accounts’, ‘money’ ‘equity’, ‘sports’, ‘acquisitions’, ...
Czech CTK Corpus (CTK)	60	‘european unions’, ‘soccer’, ‘education’, ‘sports’, ‘criminality’, ‘prague’, ...
20 Newsgroups (20NEWS)	20	‘politics’, ‘religion’, ‘electronics’, ‘motorcycles’, ‘baseball’, ‘sale’, ...
Reuters Multilingual (FR,SP,GR,I T)	6	‘corporate’, ‘finance’, ‘economics’ ‘performance’, ‘government’, ‘equity’
Reuters R8 (R8)	8	‘acquisition’, ‘interest’, ‘money’ ‘crude’, ‘trade’, ‘grain’, ...

describing different categories).

**The datasets** we use are summarized in Table 2.1. Note that the RCV-1 [57] (English: 804,414 documents), the 20 Newsgroups<sup>2</sup> (English: 18,846 documents), the Reuters Multilingual [2] (French: 26,648, Spanish: 12,342, German: 24,039, Italian:12,342 documents), the Czech CTK corpus (Czech: 14,695 documents), the R8<sup>3</sup> (English: 7,674 documents) datasets describe different categories with varying degrees of relatedness. The original categories of some of the datasets were not in the format compatible to our word embeddings dictionary or knowledge graph entities. We manually replaced those label terms to the semantically close words that exist in the dictionary or in the knowledge graph (*e.g.* `sci.med`  $\rightarrow$  ‘medicine’, etc.). While our current framework does not support multi-label classification, each article in the RCV-1 and CTK dataset is associated with multiple categories. To relax the problem into a single-label classification task, we assign each article the least frequent category across the dataset among categories associated with each article.

**Setup:** We assume a scenario where only a small fraction of the target samples are labeled (1% for RCV-1 and CTK, 0.1% for the rest) whereas the source dataset is fully labeled, and create

<sup>2</sup><http://qwone.com/~jason/20Newsgroups/>

<sup>3</sup><http://csmining.org/index.php/>

r52-and-r8-of-reuters-21578.html

Table 2.2: **Hetero-lingual text classification** test accuracy (%) on the target task, given a fully labeled source dataset and a partially labeled target dataset ( $\%_{L_T}$ ), averaged over 10-fold runs. Label embeddings with word2vec. \* and \*\* denote  $p < 0.05$  and  $p < 0.01$  paired t-test improvement over the non-CHTL baseline.

Datasets		Target Task Accuracy (%)							
Source	Target	MLP	ZSR	(: AE)	CHTL	(: ATT)	(: ATT+AE)	(: 2fc)	<b>(: 2fc+ATT+AE)</b>
RCV1	FR	39.4	55.7	56.5	57.5	58.9	58.9	58.7	<b>59.0*</b>
	SP	43.8	46.6	50.7	52.3	53.4	53.5	52.8	<b>54.2**</b>
	GR	37.7	51.1	52.0	56.4	57.3	58.0	57.3	<b>58.4**</b>
	IT	31.8	46.2	46.9	49.1	50.6	<b>51.2**</b>	49.5	51.0
20NEWS	FR	39.4	55.7	56.5	57.7	58.2	58.4	57.0	<b>58.6*</b>
	SP	43.8	46.6	50.7	52.1	52.8	52.3	52.3	<b>53.1*</b>
	GR	37.7	51.1	52.0	56.2	56.9	<b>57.5**</b>	55.9	57.0
	IT	31.8	46.2	46.9	47.3	48.0	<b>48.1</b>	47.3	47.7
R8	FR	39.4	55.7	56.5	56.5	56.4	57.2	55.9	<b>57.7</b>
	SP	43.8	46.6	50.7	50.6	51.3	<b>51.8*</b>	50.8	51.2
	GR	37.7	51.1	52.0	57.8	56.5	56.4	57.0	<b>58.0**</b>
	IT	31.8	46.2	46.9	49.7	50.4	<b>50.5*</b>	49.4	<b>50.5*</b>
FR					61.8	62.6	62.8	61.5	62.3
SP	R8	48.1	62.8	<b>63.5</b>	67.3	66.7	67.1	67.4	<b>67.7*</b>
GR					64.1	65.1	65.5	64.4	65.3
IT					62.0	63.4	<b>64.1</b>	61.6	63.0
RCV1						53.8	54.0	56.2	55.7
20NEWS	CTK	35.6	51.5	52.1	53.0	53.1	53.8	54.6	54.5
FR					52.5	53.4	53.4	53.6	53.8

various heterogeneous source-target pairs from the datasets summarized in Table 2.1. Table 2.2 reports the text classification results for the target task in this experimental setting. The results are averaged over 10-fold runs, and for each fold we randomly select  $\%_{L_T}$  of the target train instances to be labeled. Bold denotes the best performing model for each test, and \* denotes the statistically significant improvement ( $p < 0.05$ ) over other methods. Figure 2.6 reports the learning curve of text classification results at varying percentages of source and target samples available in the same experimental settings, each experiment averaged over 10-fold runs.

**Parameters:** We tune the parameters of each model with the following search space (bold indicate the choice for our final model): intermediate embeddings dimension ( $M_C$ ): {20, 40,

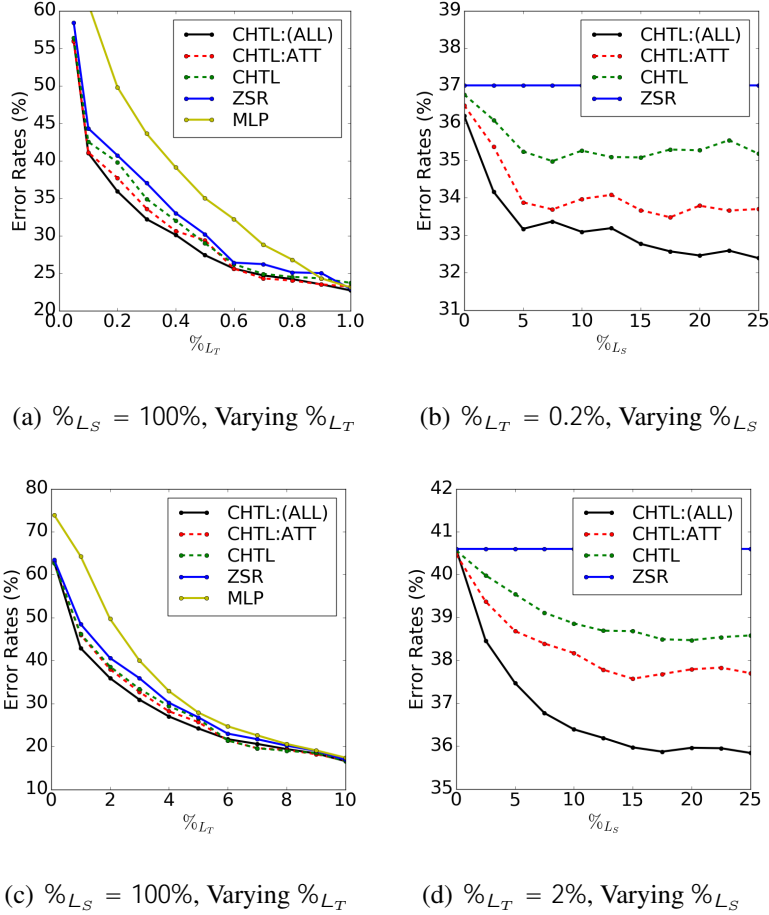


Figure 2.6: **Hetero-lingual text classification learning curves** of error rates (%) on the target task at (a)  $\%L_S = 100\%$  and varying percentages of target samples available ( $\%L_T$ ), and at (b)  $\%L_T = 0.1\%$  and varying percentages of source samples available ( $\%L_S$ ), each experiment averaged over 10-fold runs. (a,b) Source: RCV-1, Target: FR, (c,d) S: RCV-1, T: CTK.

80, 160, **320**, 640}, label embeddings dimension ( $M_E$ ): {100, 200, **300**}, knowledge graph embeddings size: {100, 200, **300**}, and attention cluster sizes ( $K$ ): {10, 20, **40**, 80}. We optimize the parameters with Adagrad [28] optimizer with batch size 100, learning rate 0.01, epsilon  $10^{-8}$ , and decay 0.01.

**Main Results:** Table 2.2 shows that all of the CHTL algorithms outperform the ZSR and MLP baselines, which indicates that knowledge from heterogeneous source domain does benefit target task. In addition, the proposed approach (CHTL:  $2f_C + \text{ATT} + \text{AE}$ ) outperforms other baselines in most of the cases, showing that the attention mechanism ( $K = 20$ ) as well as the denoising autoencoder loss improve the transfer performance ( $M_C = 320$ ,  $M_E = 300$ , label: word

Table 2.3: **CHTL with attention** test accuracy (%) on the target task, at varying  $K$  (number of clusters for attention), averaged over 10-fold runs.  $\%_{L_T} = 0.1$ .

Datasets		Accuracy (%)			
S	T	$K = 10$	$K = 20$	$K = 40$	$K = 80$
RCV1	FR	57.9	58.1	58.9	58.5
20NEWS	FR	57.7	58.0	58.2	58.3

embeddings). While having two fully connected layers (2FC) does not necessarily help CHTL performance by itself due to a small number of labels available for target data, it ultimately performs better when combined with the auto-encoder loss. Note that while both ZSL and MLP do not utilize source knowledge, ZSR with a nearest-neighbor classifier shows a huge improvement over MLP, due to the small number of categorical training labels. ZSL benefits from autoencoder loss as well, but the improvement is not as significant as in CHTL variations. Most of the results parallel the simulation results with the synthetic datasets, auguring well for the generality of our proposed approach. It can also be observed that given the same target (*e.g.* FR, SP, GR, IT), CHTL generally performs better with a densely labeled source dataset that has a larger category overlap (*e.g.* RCV-1), than a sparsely labeled source such as R8.

Figure 2.6 shows the baseline performances on the target tasks (a,b: RCV1→FR and c,d: RCV1→CTK) at (a,c) varying percentages of target amount available ( $\%_{L_T}$ ) with full source ( $\%_{L_S} = 100\%$ ), and at (b,d) varying percentages of source amount ( $\%_{L_S}$ ) for a partially available target. It can be seen that CHTL: (ALL) outperforms other baselines at most varying amounts of target samples, although the performance gap starts to degrade as more target samples are available. This result unfortunately indicates that while CHTL approach greatly benefits the target task when target labels are scarce, the auxiliary transferred knowledge can not complement already well-trained model with sufficient labels. Note also that the baselines with attention modules tend to suppress the performance degradation at higher  $\%_{L_T}$ , which coincides with the simulation results as well, showing its efficacy of avoiding negative transfer.

We observe that CHTL performance improves as more source is available (larger  $\%_{L_S}$ ), although the performance saturates fairly quickly as sufficient source is present. ZSR and MLP baselines do not leverage external source knowledge, hence the performance is constant across varying  $\%_{L_S}$ .

**Sensitivity to Attention Size ( $K$ )** for the CHTL: ATT baseline is reported in Table 2.3. Intuitively,  $K \approx N_S$  leads to a potentially intractable training while  $K \approx 1$  limits the ability to attend

Table 2.4: **CHTL with varying label embedding methods** (W2V: word embeddings, G2V: knowledge graph embeddings, Wi ki : Wikipedia document representation, Rand: random vector embeddings): test accuracy (%) on the target task averaged over 10-fold runs.  $\%_{L_T} = 0.1$ . Method: CHTL: 2fC+ATT+AE.

Datasets		Accuracy (%)			
S	T	W2V	G2V	Wi ki	Rand
RCV1	FR	59.0	59.4	53.4	48.7
20NEWS	FR	58.6	58.9	53.5	51.8
R8	FR	57.7	57.0	54.2	52.1

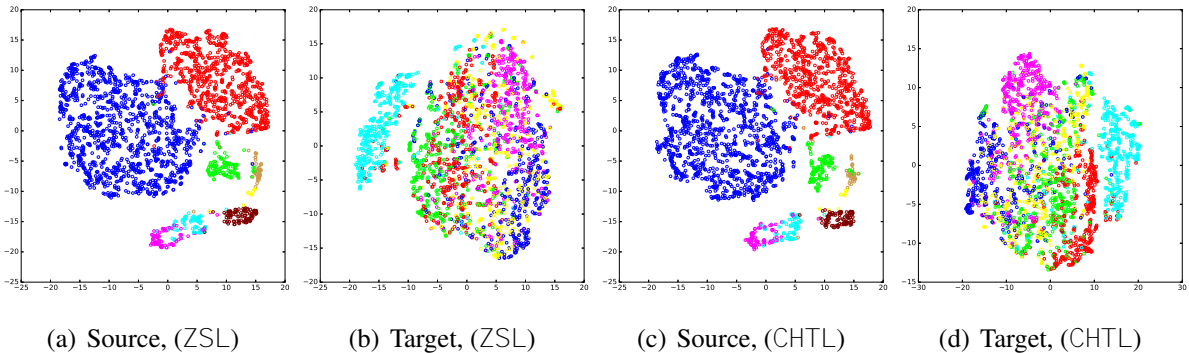


Figure 2.7: t-SNE visualization of the projected source (R8) and target (GR) instances, where (a), (b) are learned without the transferred knowledge (ZSL), and (c), (d) use the transferred knowledge (CHTL).

to subsets of source dataset, and thus one may expect an optimal value of  $K$  to exist. However, we do not observe statistically significant difference among the chosen set of parameters. We therefore set  $K = 40$  for all of the experiments, which yields the highest average accuracy on the two datasets.

**Choice of Label Embedding Methods** (Table 2.4): While W2V and G2V embeddings result in comparable performance with no significant difference, Wi ki (naive document representation of Wiki articles) and Rand embeddings perform much poorly. This shows that the quality of label embeddings is crucial in transfer of knowledge through CHTL. While Wikipedia data generally provide rich information about each entry, a more careful representation of articles must be devised.

**Feature visualization:** To visualize the projection quality of the proposed approach, we plot the t-SNE embeddings [104] of the source and the target instances (R8→GR;  $\%_{L_T} = 0.1$ ),

Table 2.5: Comparison of performance (CHTL) with varying intermediate embedding dimensions, averaged over 10-fold runs.

Datasets		Test Accuracy (%) vs. $M_C$					
S	T	20	40	80	160	320	640
20NEWS	FR	54.6	56.8	55.3	56.4	<b>57.7</b>	57.1
R8	FR	55.9	54.3	55.1	<b>57.0</b>	56.5	56.7

projected with CHTL and ZSL, respectively (Figure 2.7). We make the following observations: (1) The target instances are generally better discriminated with the projection learned from CHTL which transfers knowledge from the source dataset, than the one learned from ZSL. (2) The projection quality of the source samples remains mostly the same. Both of these observations accord with the results in Table 2.2.

**Sensitivity to the embedding dimension:** Table 2.5 compares the performance of CHTL with varying embedding dimensions ( $M_C$ ) at the intermediate layer. We do not observe statistically significant improvement for any particular dimension, and thus we simply choose the embedding dimension that yields the highest average value on the two dataset pairs ( $M_C = 320$ ) for all of the experiments.

**Visualization of Attention:** Figure 2.8 illustrates the effectiveness of the attention mechanism with an exemplary transfer learning task (source: R8, target: GR, method: CHTL: ATT,  $K = 20$ ,  $\%_{L_T} = 0.1$ ). Shown in the figure is the 2-D PCA representation of source instances (blue circles), top 5 source clusters with the highest weights ( $\alpha_k$ ) (black circles), and target instances (red triangles) projected in the embedded label space ( $R^{M_E}$ ) via  $f(g(\cdot))$  (source) or  $f(h(\cdot))$  (target). The source instances that overlap with some of the target instances in the label space (near source label terms ‘interest’ and ‘trade’ and target label term ‘finance’) are given the most attention, which thus serve as an *anchor* for knowledge transfer. Some of the source instances that are far from other target instances (near source label term ‘crude’) are also given high attention, which may be chosen to reduce the source task loss which is averaged over the attended instances. It can be seen that other heterogeneous source instances that may yield negative impact to knowledge transfer are effectively suppressed.

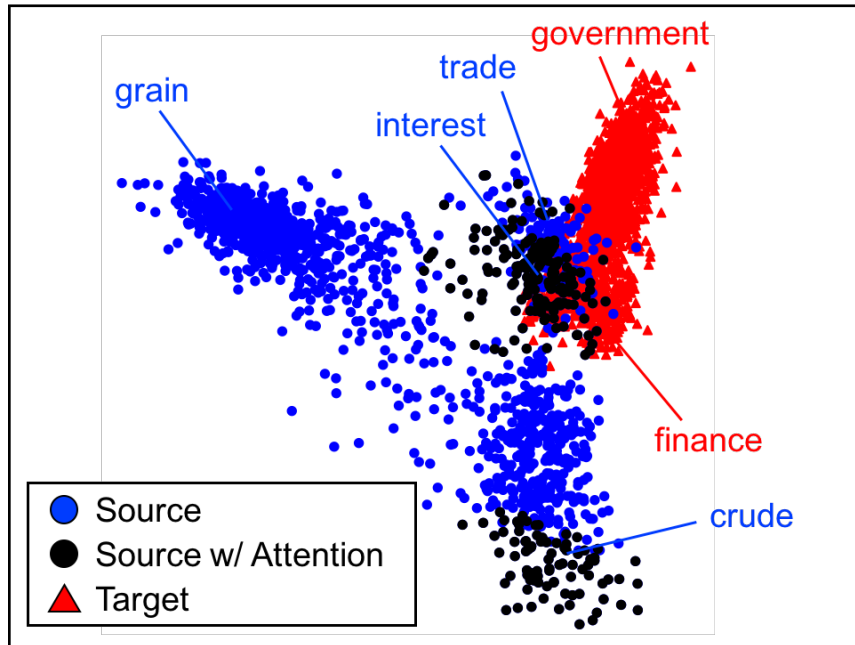


Figure 2.8: **Visualization of attention** (source: R8, target: GR). Shown in the figure is the 2-D PCA representation of source instances (blue circles), source instances with attention (black circles), and target instances (red triangles) projected in the embedded label space ( $\mathbb{R}^{M_E}$ ). Mostly the source instances that overlap with the target instances in the embedded label space are given attention during training.

## 2.5 Related Work

Note that CHTL tackles a broader range of problems than prior transfer learning approaches in that they often require parallel datasets with source-target correspondent instances (*e.g.* Hybrid Heterogeneous Transfer Learning (HHTL) [122] or CCA-based methods for a multi-view learning problem [107]), and that they require either homogeneous feature spaces [50, 62] or label spaces [20, 27, 101]. We provide a comprehensive list of related work below.

**Transfer learning with heterogeneous feature spaces:** Multi-view representation learning approaches aim at learning from heterogeneous “views” (feature sets) of multi-modal parallel datasets. The previous literature in this line of work include Canonical Correlation Analysis (CCA) based methods [22] with an autoencoder regularization in deep nets [107], translated learning [20], Hybrid Heterogeneous Transfer Learning (HHTL) [122] with marginalized stacked denoising autoencoders (mSDA) [16], [41], etc., all of which require source-target correspondent parallel instances. When parallel datasets are not given initially, [121] propose an active learning

scheme for iteratively finding optimal correspondences, or for text domain [101] propose to generate correspondent samples through a machine translation system despite noise from imperfect translation. The Heterogeneous Feature Augmentation (HFA) method [27] relaxes this limitation for a shared homogeneous binary classification task. Our approach generalizes all the previous work by allowing for heterogeneous label spaces between source and target, thus not requiring explicit source-target correspondent instances or classes.

**Transfer learning with a heterogeneous label space** : Zero-shot learning aims at building a robust classifier for unseen novel classes in the target task, often by relaxing categorical label space into a distributed vector space via transferred knowledge. For instance, [68] uses image co-occurrence statistics to describe a novel image class category, while [35, 36, 58, 88, 97, 109, 120] embed labels into semantic word vector space according to their label terms, where textual embeddings are learned from auxiliary text documents in an unsupervised manner. More recently, [50] proposes to learn domain-adapted projections to the embedded label space. While these approaches are reported to improve robustness and generalization on novel target classes, they assume that source datasets are in the same feature space as the target dataset (*e.g.* image). We extend the previous research by adding the joint objective of uncovering relatedness among datasets with heterogeneous feature spaces, via anchoring the semantic relations between the source and the target label embeddings.

**Domain adaptation with homogeneous feature and label spaces** often assumes a homogeneous class conditional distribution between source and target, and aims to minimize the difference in their marginal distribution. The previous approaches include distribution analysis and instance re-weighting or re-scaling [15, 46, 96, 124], subspace mapping [110], basis vector identification via regularization or sparse coding [50, 51], or via layerwise deep adaptation [62]. [83] provide an exhaustive survey on other traditional transfer learning approaches.

**Attention-based learning**: The proposed approach is largely inspired by the attention mechanism widely adapted in the recent deep neural network literature for various applications [14, 100, 111, 117]. The typical approaches train a parameter set for recurrent neural networks (*e.g.* an LSTM) which during the decoding step determines a weight over annotation vectors, or in other words a relative importance vector over discrete subsets of input. The attention mechanism can be seen as a regularization technique preventing overfitting during training, and in our case avoiding negative transfer.

Limited studies have investigated **negative transfer**, most of which propose to prevent negative effects of transfer by measuring dataset- or task-level relatedness via parameter comparison

in Bayesian models [5, 89]. Our approach practically avoids instance-level negative transfer, by determining *which* knowledge within a source dataset to suppress or attend in learning of a transfer network.

## 2.6 Summary

We summarize our contributions in this chapter as follows: We address a unique challenge of leveraging transferable knowledge in the heterogenous case, where labeled source data differs from target data in both feature and label spaces. To this end, (1) we propose a novel framework for heterogeneous transfer learning to discover the latent subspace to map the source into the target space, from which it simultaneously learns a shared final projection to the embedded label space. (2) An extensive empirical evaluation on both the simulations and the hetero-lingual text classification task demonstrate the efficacy of each part of the proposed approach. CHTL approaches are specifically effective for a sparsely labeled target, especially when aided with a well labeled source dataset with a large category coverage.

## Chapter 3

# CHTL Applications: Multimodal Transfer Learning for Named Entity Disambiguation

The CHTL network architecture is flexible and thus applicable for a number of downstream tasks. In previous chapter, we showed that the CHTL approach can be used for hetero-lingual text classification tasks (domains: text in multiple different languages) as an immediate application. In this chapter, we demonstrate that the CHTL approach of leveraging heterogeneous knowledge sets can be applied to the multimodal transfer learning task for named entity disambiguation (which requires joint learning of images and text), while keeping the CHTL network architecture mostly the same.

Specifically, we introduce the new Multimodal Named Entity Disambiguation (MNED) task, which leverage short captions and accompanying images to disambiguate an entity in multimodal social media posts such as Snapchat or Instagram captions. Social media posts bring significant challenges for disambiguation tasks because 1) ambiguity not only comes from polysemous entities, but also from inconsistent or incomplete notations, 2) very limited context is provided with surrounding words, and 3) there are many emerging entities often unseen during training. To this end, we build a new dataset called *SnapCaptionsKB*, a collection of Snapchat image captions submitted to public and crowd-sourced stories, with named entity mentions fully annotated and linked to entities in an external knowledge base. We then build a deep zeroshot multimodal network for MNED that 1) extracts contexts from both text and image, and 2) predicts correct entity in the knowledge graph embeddings space, allowing for zeroshot disambiguation of entities un-

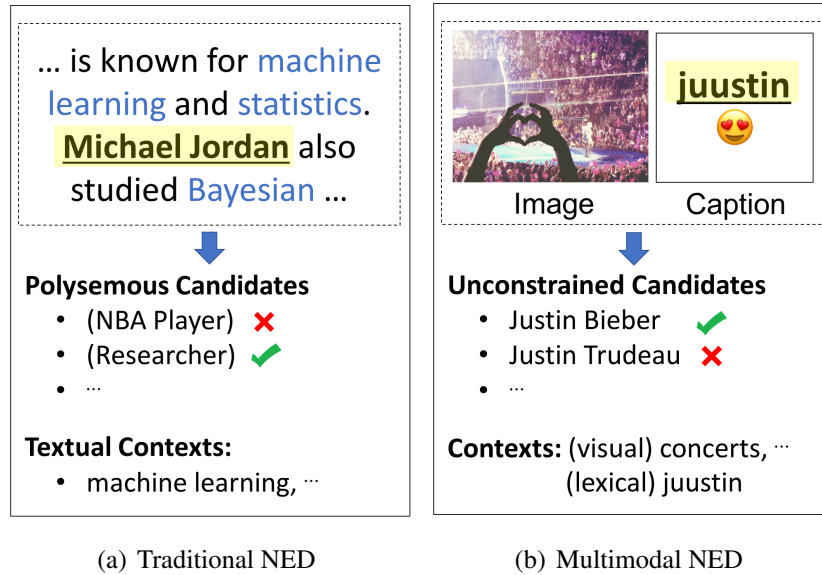


Figure 3.1: Examples of (a) a traditional NED task, focused on disambiguating polysemous entities based on surrounding textual contexts, and (b) the proposed Multimodal NED task for short media posts, which leverages both visual and textual contexts to disambiguate an entity. Note that mentions are often lexically inconsistent or incomplete, and thus a fixed candidates generation method (based on exact mention-entity statistics) is not viable.

seen in training set as well. Finally, we leverage the vastly available external heterogeneous data (separate sources of images as well as traditional text-only NED datasets) with the CHTL training, showing that knowledge transfer can further improve the multimodal training. The proposed model, along with the multimodal CHTL training, significantly outperforms the state-of-the-art text-only NED models.

### 3.1 Introduction

Online communications are increasingly becoming fast-paced and frequent, and hidden in these abundant user-generated social media posts are insights for understanding users and their preferences. However, these social media posts often come in unstructured text or images, making massive-scale opinion mining extremely challenging. Named entity disambiguation (NED), the task of linking ambiguous entities from free-form text *mention* to specific *entities* in a pre-defined knowledge base (KB), is thus a critical step for extracting structured information which leads to its application for recommendations, advertisement, personalized assistance, etc.

While many previous approaches on NED been successful for well-formed text in disam-

biguating polysemous entities via context resolution, several additional challenges remain for disambiguating entities from extremely short and coarse text found in social media posts (*e.g.* “juuustin 🥰” as opposed to “I love Justin Bieber / Justin Trudeau / *etc.*”). In many of these cases it is simply impossible to disambiguate entities from text alone, due to enormous number of surface forms arising from incomplete and inconsistent notations. In addition, social media posts often include mentions of newly emerging entities unseen in training sets, making traditional context-based entity linking often not viable.

However, as popular social media platforms are increasingly incorporating a mix of text and images (*e.g.* Snapchat, Instagram, Pinterest, *etc.*), we can advance the disambiguation task to incorporate additional visual context for understanding posts. For example, the mention of ‘juuustin’ is completely ambiguous in its textual form, but an accompanying snap image of a concert scene may help disambiguate or re-rank among several lexical candidates (*e.g.* Justin Bieber (a pop singer) versus Justin Trudeau (a politician) in Figure 3.1).

To this end, we introduce a new task called Multimodal Named Entity Disambiguation (MNED) that handles unique challenges for social media posts composed of extremely short text and images, aimed at disambiguating entities by leveraging both textual and visual contexts. We then propose a novel zeroshot MNED model, which obtains visual context vectors from images with a CNN [54], and combines with textual context extracted from a bidirectional LSTM [29] (Section 3.2.2). In addition, we obtain embeddings representation of 1M entities from a knowledge graph, and train the MNED network to predict label embeddings of entities in the same space as corresponding knowledge graph embeddings (Section 3.2.4). This approach effectively allows for zeroshot prediction of unseen entities, which is critical for scarce-label scenario due to extensive human annotation efforts required. We also develop a lexical embeddings model that determines lexical similarity between a mention and potential entities, to aid in prediction of a correct entity (Section 3.2.3). Lastly, we present both multi-view (using the image-caption parallel pairs only) and multimodal transfer learning approaches (using extra heterogeneous sources for knowledge transfer) to train the MNED network. Section 3.2.5 details the model combining the components above.

Note that our method takes different perspectives from the previous work on NED [31, 44, 113] in the following important ways. First, while most of the previous methods generate fixed “candidates” for disambiguation given a mention from mention-entity pair statistics (thus disambiguation is limited for entities with exact surface form matches), we do not fixate candidate generation, due to intractable variety of surface forms for each named entity and unforeseen

mentions of emerging entities. Instead, we have a lexical model incorporated into the discriminative score function that serves as soft normalization of various surface forms. Second, we extract auxiliary visual contexts for detected entities from user-generated images accompanied with textual posts, which is crucial because captions in our dataset are substantially shorter than text documents in most other NED datasets. To the best of our knowledge, our work is the first in using visual contexts for the named entity disambiguation task. See Section 3.4 for the detailed literature review.

**Our contributions** are as follows: for the new MNED task we introduce, we propose a deep zeroshot multimodal network with (1) a CNN-LSTM hybrid module that extracts contexts from both image and text, (2) a zeroshot learning layer which via embeddings projection allows for entity linking with 1M knowledge graph entities even for entities unseen from captions in training set, and (3) a lexical language model called *Deep Levenshtein* to compute lexical similarities between mentions and entities, relaxing the need for fixed candidates generation. We show that the proposed approaches successfully disambiguate incomplete mentions as well as polysemous entities, outperforming the state-of-the-art models on our newly crawled *SnapCaptionsKB* dataset, composed of 12K image-caption pairs with named entities annotated and linked with an external KB. Lastly, (4) we show that extra knowledge can be transferred from external knowledge sources, proving the efficacy of the CHTL training in multimodal transfer learning applications.

## 3.2 Proposed Methods

Figure 3.2 illustrates the proposed model modified from the original CHTL network, which maps each multimodal social media post data to one of the corresponding entities in the KB. Given a multimodal input that contains a mention of an ambiguous entity, we first extract textual and visual features contexts with RCNNs and Bi-LSTMs, respectively (Section 3.2.2). We also obtain lexical character-level representation of a mention to compare with lexical representation of KB entities, using a proposed model called *Deep Levenshtein* (Section 3.2.3). We then get high-dimensional label embeddings of KB entities constructed from a knowledge graph, where similar entities are mapped as neighbors in the same space (Section 3.2.4). Finally, we aggregate all the contextual information extracted from surrounding text, image, and lexical notation of a mention, and predict the best matching KB entity based on knowledge graph label representation and lexical notation of KB entity candidates (Section 3.2.5).

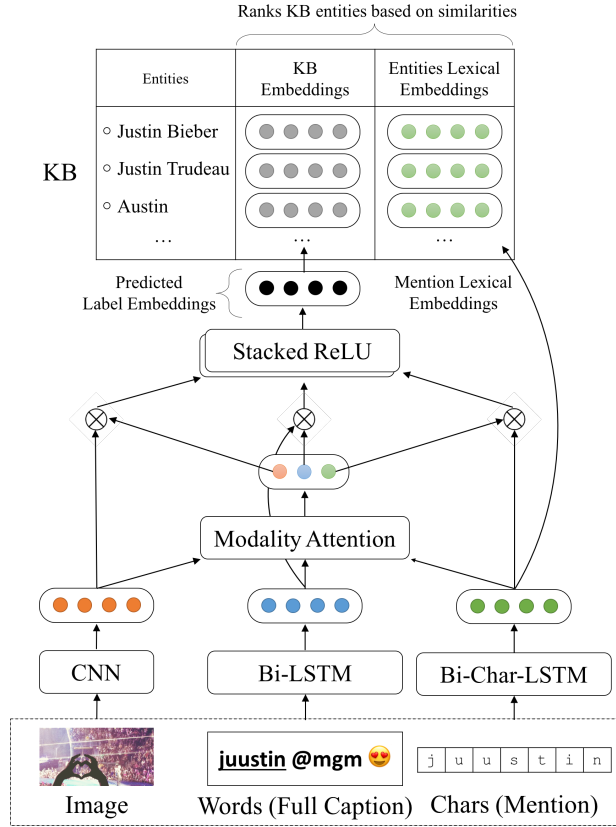


Figure 3.2: The main architecture of our Multimodal NED network. We extract contextual information from an image, surrounding words, and lexical embeddings of a mention. The modality attention module determines weights for modalities, the weighted projections of which produce label embeddings in the same space as knowledge-base (KB) entity embeddings. We predict a final candidate by ranking based on similarities with KB entity knowledge graph embeddings as well as with lexical embeddings.

### 3.2.1 Notations

Let  $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$  a set of  $N$  input social media posts samples for disambiguation, with corresponding ground truth named entities  $\mathbf{Y} = \{\mathbf{y}^{(i)}\}_{i=1}^N$  for  $\mathbf{y} \in \mathbf{Y}_{\text{KB}}$ , where  $\mathbf{Y}_{\text{KB}}$  is a set of entities in KB. Each input sample is composed of three modalities:  $\mathbf{x} = \{\mathbf{x}_w; \mathbf{x}_v; \mathbf{x}_c\}$ , where  $\mathbf{x}_w = \{\mathbf{x}_{w:t}\}_{t=1}^{L_w}$  is a sequence of words with length  $L_w$  surrounding a mention in a post,  $\mathbf{x}_v$  is an image associated with a post (Section 3.2.2), and  $\mathbf{x}_c = \{\mathbf{x}_{c:t}\}_{t=1}^{L_c}$  is a sequence of characters comprising a mention (Section 3.2.3), respectively. We denote high-dimensional feature extractor functions for each modality as:  $\mathbf{w}(\mathbf{x}_w)$ ,  $\mathbf{c}(\mathbf{x}_c)$ ,  $\mathbf{v}(\mathbf{x}_v)$ . We represent each output label in two modalities:  $\mathbf{y} = \{\mathbf{y}_{\text{KB}}; \mathbf{y}_c\}$ , where  $\mathbf{y}_{\text{KB}}$  is a knowledge base label embeddings representation

(Section 3.2.4), and  $y_c$  is a character embeddings representation of KB entities (Section 3.2.3: Deep Levenshtein).

We formulate our zeroshot multimodal NED task as follows:

$$y = \operatorname{argmax}_{y' \in \mathcal{Y}_{KB}} \operatorname{sim}(\mathbf{f}_{x \rightarrow y}(\mathbf{x}), y')$$

where  $\mathbf{f}_{x \rightarrow y}$  is a function with learnable parameters that project multimodal input samples ( $\mathbf{x}$ ) into the same space as label representations ( $y$ ), and  $\operatorname{sim}(\cdot)$  produces a similarity score between prediction and ground truth KB entities.

### 3.2.2 Textual and Visual Contexts Features

**Textual features:** we represent textual context of surrounding words of a mention with a Bi-LSTM language model [29] with distributed word semantics embeddings. We use the following implementation for the LSTM.

$$\begin{aligned} \mathbf{i}_t &= \sigma(\mathbf{W}_{xi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1}) \\ \mathbf{c}_t &= (1 - \mathbf{i}_t) \odot \mathbf{c}_{t-1} \\ &\quad + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\mathbf{x}_{w;t} + \mathbf{W}_{hc}\mathbf{h}_{t-1}) \\ \mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\mathbf{x}_{w;t} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t) \\ \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \\ \mathbf{w}(\mathbf{x}_w) &= [\overrightarrow{\mathbf{h}}_{L_w}; \overleftarrow{\mathbf{h}}_{L_w}] \end{aligned} \tag{3.1}$$

where  $\mathbf{h}_t$  is an LSTM hidden layer output at decoding step  $t$ , and  $\mathbf{w}(\mathbf{x}_w)$  is an output textual representation of bi-directional LSTM concatenating left and right context at the last decoding step  $t = L_w$ . Bias terms for gates are omitted for simplicity of formulation.

For the Bi-LSTM sentence encoder, we use pre-trained word embeddings obtained from an unsupervised language model aimed at learning co-occurrence statistics of words from a large external corpus. Word embeddings are thus represented as distributional semantics of words. In our experiments, we use pre-trained embeddings from Stanford GloVe model [84].

**Visual features:** we take the final activation of a modified version of the recurrent convolutional network model called Inception (GoogLeNet) [102] trained on the ImageNet dataset [91] to classify multiple objects in the scene. The final layer representation ( $\mathbf{v}(\mathbf{x}_v)$ ) thus encodes discriminative information describing what objects are shown in an image, providing cues for disambiguation.

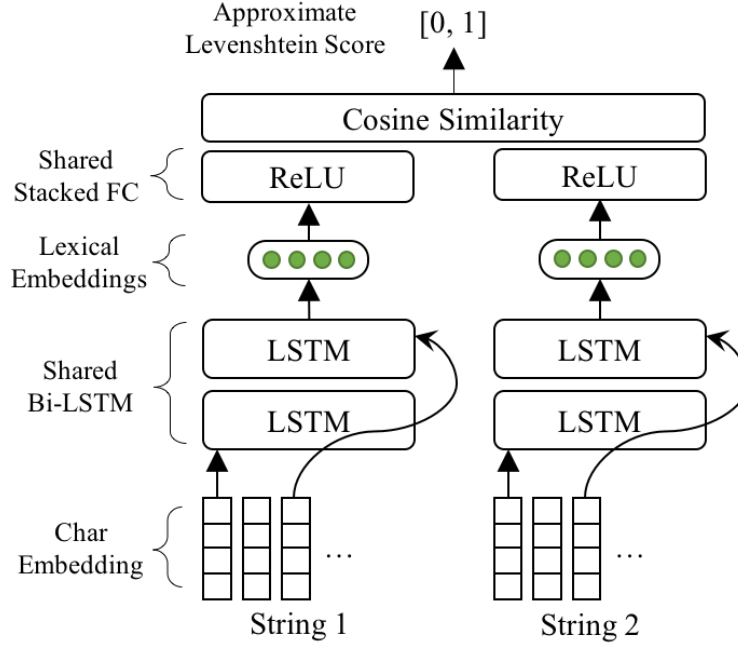


Figure 3.3: Deep Levenshtein, which predicts approximate Levenshtein scores between two strings. As a byproduct of this model, the shared Bi-LSTM can produce lexical embeddings purely based on lexical property of character sequences.

### 3.2.3 Lexical Embeddings: Deep Levenshtein

While traditional NED tasks assume perfect lexical match between mentions and their corresponding entities, in our task it is important to account for various surface forms of mentions (nicknames, mis-spellings, inconsistent notations, etc.) corresponding to each entity. Towards this goal, we train a separate deep neural network to compute approximate Levenshtein distance which we call Deep Levenshtein (Figure 3.3), composed of a shared bi-directional character LSTM, shared character embedding matrix, fully connected layers, and a dot product merge operation layer. The optimization is as follows:

$$\min_c \left\| \frac{1}{2} \left( \frac{\mathbf{c}(\mathbf{x}_c) \cdot \mathbf{c}(\mathbf{x}_c^\ell)}{\|\mathbf{c}(\mathbf{x}_c)\| \|\mathbf{c}(\mathbf{x}_c^\ell)\|} + 1 \right) - \text{sim}(\mathbf{x}_c, \mathbf{x}_c^\ell) \right\|^2 \quad (3.2)$$

where  $\mathbf{c}(\mathbf{x}_c) = [\overrightarrow{\mathbf{h}}_{c:L_c}; \overleftarrow{\mathbf{h}}_{c:L_c}]$

where  $\mathbf{c}(\cdot)$  is a bi-directional LSTM output vector for a character sequence defined similar as in Eq.3.1,  $\text{sim}(\cdot)$  is an output of the Deep Levenshtein network, producing a normalized similarity score with a range [0,1] based on Levenshtein edit distance, and  $(\mathbf{x}_c, \mathbf{x}_c^\ell)$  is any pair of two

strings. We generate millions of these pairs as training data by artificially corrupting seed strings by varying degrees (addition, deletion, replacement).

Once trained, it can produce a purely lexical embedding of a string without semantic allusion (via  $c(\cdot)$ ), and predict lexical similarity between two strings based on their distance in the embedding space. On an intuitive level, this component effectively bypasses normalization steps, and instead incorporates lexical similarities between input mentions and output KB entities into the overall optimization of the disambiguation network.

We use by-product  $c(\cdot)$  network to extract lexical embeddings of mentions and KB entities, and freeze  $c$  in training of the disambiguation network. We observe that this approach significantly outperforms alternative ways to obtain character embeddings (*e.g.* having a character Bi-LSTM as a part of the disambiguation network training, which unnecessarily learns semantic allusions that are prone to errors when notations are inconsistent.)

### 3.2.4 Label Embeddings from Knowledge Graph

Due to the overwhelming variety of (newly trending) entities mentioned over social media posts, at test phases we frequently encounter new named entities that are unseen in the training data. In order to address this issue, we propose a zeroshot learning approach [35] by inducing embeddings obtained from knowledge graphs on KB entities. Knowledge graph label embeddings are learned from known relations among entities within a graph (*e.g.* ‘IS-A’, ‘LOCATED-AT’, etc.), the resulting embeddings of which can group similar entities closer in the same space (*e.g.* ‘pop stars’ are in a small cluster, ‘people’ and ‘organizations’ clusters are far apart, etc.) [12, 80, 108]. Once high-level mapping from contextual information to label embeddings is learned, the knowledge-graph based zeroshot approach can improve the entity linking performance given ambiguous entities unseen in training data. In brief formulation, the model for obtaining embeddings from a knowledge graph (composed of subject-relation-object  $(s, r, o)$  triplets) is as follows:

$$P(\mathbb{1}_r(s, o) = 1 | \mathbf{e}, \mathbf{e}_r, \theta) = \text{score}(\mathbf{e}(s), \mathbf{e}_r(r), \mathbf{e}(o)) \quad (3.3)$$

where  $\mathbb{1}_r$  is an indicator function of a known relation  $r$  for two entities  $(s, o)$  (1: valid relation, 0: unknown relation),  $\mathbf{e}$  is a function that extracts embeddings for entities,  $\mathbf{e}_r$  extracts embeddings for relations, and  $\text{score}(\cdot)$  is a deep neural network that produces a likelihood of a valid triplet.

In our experiments, we use the 1M subset of the Freebase knowledge graph [8] to obtain label embeddings with the Holographic KB implementation by [80].

### 3.2.5 Deep Zeroshot MNED Network (DZMNED)

Using the contextual information extracted from surrounding text and an accompanying image (Section 3.2.2) and lexical embeddings of a mention (Section 3.2.3), we build a Deep Zeroshot MNED network (DZMNED) which predicts a corresponding KB entity based on its knowledge graph embeddings (Section 3.2.4) and lexical similarity (Section 3.2.3) with the following objective:

$$\min_{\mathbf{W}} \mathcal{L}_{\text{KB}}(\mathbf{x}, \mathbf{y}_{\text{KB}}; \mathbf{W}_{\mathbf{w}}, \mathbf{W}_{\mathbf{v}}, \mathbf{W}_{\mathbf{f}}) + \mathcal{L}_c(\mathbf{x}_c, \mathbf{y}_c; \mathbf{W}_c)$$

where

$$\mathcal{L}_{\text{KB}}(\cdot) = \frac{1}{N} \sum_{i=1}^N \sum_{\mathbf{y} \neq \mathbf{y}_{\text{KB}}^{(i)}} \max[0, \mathbf{y} \cdot \mathbf{y}_{\text{KB}}^{(i)} - \mathbf{f}(\bar{\mathbf{x}}^{(i)}) \cdot (\mathbf{y}_{\text{KB}}^{(i)} - \mathbf{y})^>]$$

$$\mathcal{L}_c(\cdot) = \frac{1}{N} \sum_{i=1}^N \sum_{\mathbf{y} \neq \mathbf{y}_c^{(i)}} \max[0, \mathbf{y} \cdot \mathbf{y}_c^{(i)} - \mathbf{c}(\mathbf{x}_c^{(i)}) \cdot (\mathbf{y}_c^{(i)} - \mathbf{y})^>]$$

$\mathcal{R}(\mathbf{W})$ : regularization

where  $\mathcal{L}_{\text{KB}}(\cdot)$  is the supervised hinge rank loss for knowledge graph embeddings prediction,  $\mathcal{L}_c(\cdot)$  is the loss for lexical mapping between mentions and KB entities,  $\bar{\mathbf{x}}$  is a weighted average of three modalities  $\mathbf{x} = \{\mathbf{x}_w, \mathbf{x}_v, \mathbf{x}_c\}$  via the modality attention module.  $\mathbf{f}(\cdot)$  is a transformation function with stacked layers that projects weighted input to the KB embeddings space,  $\tilde{\mathbf{y}}$  refers to the embeddings of negative samples randomly sampled from KB entities except the ground truth label of the instance,  $\mathbf{W} = \{\mathbf{W}_{\mathbf{f}}, \mathbf{W}_c, \mathbf{W}_{\mathbf{w}}, \mathbf{W}_{\mathbf{v}}\}$  are the learnable parameters for  $\mathbf{f}$ ,  $\mathbf{c}$ ,  $\mathbf{w}$ , and  $\mathbf{v}$  respectively, and  $\mathcal{R}(\mathbf{W})$  is a weight decay regularization term. Note that the proposed MNED model architecture and objective are mostly the same with the CHTL network, hence CHTL training can be directly applied.

Similarly to [77], we formulate the **modality attention** module for our MNED network as follows, which selectively attenuates or amplifies modalities:

$$[\mathbf{a}_w; \mathbf{a}_c; \mathbf{a}_v] = \sigma(\mathbf{W}_m \cdot [\mathbf{x}_w; \mathbf{x}_c; \mathbf{x}_v] + \mathbf{b}_m) \quad (3.4)$$

$$\alpha_m = \frac{\exp(\mathbf{a}_m)}{\sum_{m \in \{w, c, v\}} \exp(\mathbf{a}_m)} \quad \forall m \in \{w, c, v\}$$

$$\bar{\mathbf{x}} = \sum_{m \in \{w, c, v\}} \alpha_m \mathbf{x}_m \quad (3.5)$$

where  $\alpha = [\alpha_w; \alpha_c; \alpha_v] \in \mathbb{R}^3$  is an attention vector, and  $\bar{\mathbf{x}}$  is a final context vector that maximizes information gain.

Intuitively, the model is trained to produce a higher dot product similarity between the projected embeddings with its correct label than with an incorrect negative label in both the knowledge graph label embeddings and the lexical embeddings spaces, where the margin is defined as the similarity between a ground truth sample and a negative sample.

When visual embeddings are not provided, the modality attention can be defined in a similar way as follows:

$$\begin{aligned}
 [\mathbf{a}_w; \mathbf{a}_c] &= \sigma(\mathbf{W}_m \cdot [\mathbf{x}_w; \mathbf{x}_c] + \mathbf{b}_m) \\
 \alpha_m &= \frac{\exp(\mathbf{a}_m)}{\sum_{m' \in \{w, c\}} \exp(\mathbf{a}_{m'})} \quad \forall m \in \{w, c\} \\
 \bar{\mathbf{x}} &= \sum_{m \in \{w, c\}} \alpha_m \mathbf{x}_m
 \end{aligned} \tag{3.6}$$

At test time, the following label-producing nearest neighbor (1-NN) classifier is used for the target task (we cache all the label embeddings to avoid repetitive projections):

$$1\text{-NN}(\mathbf{x}) = \underset{(y_{\text{KB}}, y_c) \in \mathcal{Y}_{\text{KB}}}{\operatorname{argmax}} \mathbf{f}(\bar{\mathbf{x}}) \cdot \mathbf{y}_{\text{KB}} + \mathbf{g}(\mathbf{x}_c) \cdot \mathbf{y}_c \tag{3.7}$$

In summary, the model produces (1) projection of input modalities (mention, surrounding text, image) into the knowledge graph embeddings space, and (2) lexical embeddings representation of mention, which then calculates a combined score of contextual (knowledge graph) and string similarities with each entity in  $\mathcal{Y}_{\text{KB}}$ .

When auxiliary unimodal data (*e.g.* traditional NED datasets or entity associated images) are available, we leverage those data to transfer knowledge, complementary to multimodal training. Since the input is unimodal, we keep the weights and the overall architecture the same, except the modality attention module (plain averaging operation is used instead).

### 3.3 Empirical Evaluation

**Task:** Given a caption and an accompanying image (if available), the goal is to disambiguate and link a target mention in a caption to a corresponding entity from the knowledge base (1M subset of the Freebase knowledge graph [8]).

### 3.3.1 Datasets

Our **SnapCaptionsKB** dataset is composed of 12K user-generated image and textual caption pairs where named entities in captions and their links to KB entities are manually labeled by expert human annotators. These captions are collected exclusively from snaps submitted to public and crowd-sourced stories (aka *Live Stories* or *Our Stories*). Examples of such stories are “New York Story” or “Thanksgiving Story”, which are aggregated collections of snaps for various public venues, events, etc. Our data do not contain raw images, and we only provide textual captions and obfuscated visual descriptor features extracted from the pre-trained InceptionNet. We split the dataset randomly into train (70%), validation (15%), and test sets (15%). The captions data have average length of 29.5 characters (5.57 words) with vocabulary size 16,553, where 6,803 are considered unknown tokens from Stanford GloVE embeddings [84]. Named entities annotated in the dataset include many of new and emerging entities found in various surface forms. To the best of our knowledge, our *SnapCaptionsKB* is the only dataset that contains image-caption pairs with human-annotated named entities and their links to KB entities. For **auxiliary training**, we use the WNUT NED datasets [21], which is a challenging dataset with short and noisy texts mostly crawled from social media sites such as Twitter, along with 250K images obtained from Twitter dumps that contain image URLs, each crawled and associated with entities from the SnapCaptionsKB dataset.

### 3.3.2 Baselines

We report performance of the following state-of-the-art NED models as baselines, with several candidate generation methods and variations of our proposed approach to examine contributions of each component (W: word, C: char, V: visual).

**Candidates generation:** Note that our zeroshot approach allows for entity disambiguation without a fixed candidates generation process. In fact, we observe that the conventional method for fixed candidates generation harms the performance for noisy social media posts with many emerging entities. This is because the difficulty of entity linking at test time rises not only from multiple entities ( $e$ ) linking to a single mention ( $m$ ), but also from each entity found in multiple surface forms of mentions (often unseen at train time). To show the efficacy of our approach that does not require candidates generation, we compare with the following candidates generation methods:

- $m \rightarrow e$  hash list: This method retrieves KB entity ( $e$ ) candidates per mention ( $m$ ) based

on exact  $(m, e)$  pair occurrence statistics from a training corpora. This is the most predominantly used candidates generation method [31, 44, 113]. Note that this approach is especially vulnerable at test time to noisy mentions or emerging entities with no or a few matching candidate entities from training set.

- k-NN: We also consider using lexical neighbors of mentions from KB entities as candidates. This approach can be seen as soft normalization to relax the issue of having to match a variety of surface forms of a mention to KB entities. We use our Deep Levenshtein (Section 3.2.3) to compute lexical embeddings of KB entities and mentions, and retrieves Euclidean neighbors (and their polysemous entities) as candidates.

**NED models:** We choose as baselines the following state-of-the-art NED models for noisy text, as well as several configurations of our proposed approach to examine contributions of each component (W: word, C: char, V: visual).

- sDA-NED (W only) [44]: uses a deep neural network with stacked denoising autoencoders (sDA) to encode bag-of-words representation of textual contexts and to directly compare mentions and entities.
- ARNN (W only) [31]: uses an Attention RNN model that computes similarity between word and entity embeddings to disambiguate among fixed candidates.
- Deep Zeroshot (W only): uses the deep zeroshot architecture similar to Figure 3.2, but uses word contexts (caption) only.
- **(proposed)** DZMNED + Deep Levenshtein + InceptionNet with modality attention (MA) + knowledge transfer (W+C+V): is the proposed approach as described in Figure 3.2, with extra knowledge transfer from external sources.
- **(proposed)** DZMNED + Deep Levenshtein + InceptionNet with MA (W+C+V): is the same proposed approach, trained without the additional knowledge transfer.
- **(proposed)** DZMNED + Deep Levenshtein + InceptionNet w/o MA (W+C+V): concatenates all the modality vectors instead.
- **(proposed)** DZMNED + Deep Levenshtein (W+C): only uses textual context.
- **(proposed)** DZMNED + Deep Levenshtein w/o modality attention (W+C): does not use the modality attention module, and instead concatenates word and lexical embeddings.

Modalities	Model	Candidates Generation	Accuracy (%)				
			Top-1	Top-3	Top-5	Top-10	Top-50
W	ARNN [31]	$m \rightarrow e$ list	51.2	60.4	66.5	66.9	66.9
W	ARNN [31]	5-NN (lexical)	35.2	43.3	45.0	-	-
W	ARNN [31]	10-NN (lexical)	31.9	40.1	44.5	50.7	-
W	sDA-NED [44]	$m \rightarrow e$ list	48.7	57.3	66.3	66.9	66.9
W	Zeroshot	N/A	43.6	63.8	67.1	70.5	77.2
W + C	DZMNED	N/A	67.0	72.7	74.8	76.8	85.0
W + C	DZMNED + MA	N/A	67.8	73.5	74.8	76.2	84.6
W + C + V	DZMNED	N/A	67.2	74.6	77.7	80.5	88.1
W + C + V	DZMNED + MA	N/A	<b>68.1</b>	<b>75:5*</b>	78.2	80.9	87.9
W + C + V	DZMNED + MA + Transfer	N/A	68.0	75.2	<b>79:4*</b>	<b>82:3*</b>	<b>88.5</b>

Table 3.1: NED performance on the *SnapCaptionsKB* dataset at Top-1, 3, 5, 10, 50 accuracies. The classification is over 1M entities. Candidates generation methods: N/A, or over a fixed number of candidates generated with methods:  $m \rightarrow e$  hash list and kNN (lexical neighbors). \* and \*\* denote  $p < 0.05$  and  $p < 0.01$  paired t-test improvement over its comparing baselines and ablation studies (**non-zeroshot** vs. **zeroshot** (ours), **W+C** (ours) vs **W+C+V** (ours), **non-transfer** (ours) vs. **transfer** (ours final)).

### 3.3.3 Results

**Parameters:** We tune the parameters of each model with the following search space (bold indicate the choice for our final model): character embeddings dimension: {25, 50, **100**, 150, 200, 300}, word embeddings size: {25, 50, **100**, 150, 200, 300}, knowledge graph embeddings size: {**100**, 200, 300}, LSTM hidden states: {50, **100**, 150, 200, 300}, and  $\bar{x}$  dimension: {25, 50, **100**, 150, 200, 300}. We optimize the parameters with Adagrad [28] with batch size 10, learning rate 0.01, epsilon  $10^{-8}$ , and decay 0.1.

**Main Results:** Table 3.1 shows the Top-1, 3, 5, 10, and 50 candidates retrieval accuracy results on the *Snap Captions* dataset. We see that the MNED proposed approach significantly outperforms the baselines which all use fixed candidates generation method. Note that  $m \rightarrow e$  hash list-based methods, which retrieve as candidates the KB entities that appear in the training set of captions only, has upper performance limit at 66.9%, showing the limitance of fixed candidates generation method for unseen entities.  $k$ -NN methods which retrieve lexical neighbors of

KB Embeddings	Top-1	Top-5	Top-10
Trained with 1M entities	<b>68.1</b>	<b>78.2</b>	<b>80.9</b>
Trained with 10K entities	60.3	72.5	75.9
Random embeddings	41.4	45.8	48.0

Table 3.2: MNED performance (Top-1, 5, 10 accuracies) on SnapCaptionsKB with varying qualities of KB embeddings. Model: DZMNED (W+C+V)

mention (in an attempt to perform soft normalization on mentions) also do not perform well. Our proposed zeroshot approaches, however, do not fixate candidate generation, and instead compares combined contextual and lexical similarities among all 1M KB entities, achieving much higher upper performance limit (Top-50 retrieval accuracy reaches 88.1%). This result indicates that the proposed zeroshot model is capable of predicting for unseen entities as well. The lexical sub-model can also be interpreted as functioning as soft neural mapping of mention to potential candidates, rather than heuristic matching to fixed candidates.

In addition, when visual context is available (W+C+V), the performance generally improves over the textual models (W+C), showing that visual information can provide additional contexts for disambiguation. The modality attention module also adds performance gain by re-weighting the modalities based on their informativeness.

Lastly, we see that with additional training with knowledge trasfered from heterogeneous sources, the performance improves especially at top-5, 10, 50 results. This result is significant in that the proposed MNED architecture does not need to be bound to a multimodal NED data (the annotated data of which is rare and hard to obtain), and rather it can leverage vastly available existing sources to improve the performance.

**Error Analysis:** Table 3.3 shows example cases where incorporation of visual contexts affects disambiguation of mentions in textual captions. For example, polysemous entities such as ‘Jordan’ in the caption “*Taking the new Jordan for a walk*” or ‘CID’ as in “*LETS GO CID*” are hard to disambiguate due to the limited textual contexts provided, while visual information (*e.g.* visual tags ‘footwear’ for Jordan, ‘DJ’ for CID) provides similarities to each mention’s distributional semantics from other training examples. Mentions unseen at train time (‘STEPHHHH’, ‘murica’) often resort to lexical neighbors by (W+C), whereas visual contexts can help disambiguate better. A few cases where visual contexts are not helpful include visual tags that are not related to mentions, or do not complement already ambiguous contexts.

Caption ( <u>target</u> )	Visual Tags	GT	Top-1 Prediction	
			(W+C+V)	(W+C)
“YA BOI <u>STEPHHHH</u> ”	sports equip, ball, parade, ...	Stephen Curry	(=GT)	Stephenville
+ “Taking the new <u>Jordan</u> for a walk”	footwear, shoe, sock, ...	Air Jordan	(=GT)	Michael Jordan
“out for <u>murica</u> ’s bday 🤩”	parade, flag, people, ...	U.S.A.	(=GT)	Murcia (Spain)
“Come on now, <u>Dre</u> ”	club, DJ, night, ...	Dr. Dre	(=GT)	Dre Kirkpatrick
“LETS GO <u>CID</u> ”	drum, DJ, drummer, ...	CID (DJ)	(=GT)	CID (ORG)
- “kick back hmu for <u>addy</u> .”	weather, fog, tile, ...	Adderall	GoDaddy	(=GT)
“@ <u>Sox</u> to see 3 4 get retired! 🏆🍷”	sunglasses, stadium, ...	Red Sox	White Sox	White Sox

Table 3.3: Error analysis: **when do images help NED?** Ground-truth (GT) and predictions of our model with vision input (W+C+V) and the one without (W+C) for the underlined mention are shown. For interpretability, visual tags (label output of InceptionNet) are presented instead of actual feature vectors.

**Sensitivity to Knowledge Graph Embeddings Quality:** The proposed approach relies its prediction on entity matching in the KB embeddings space, and hence the quality of KB embeddings is crucial for successful disambiguation. To characterize this aspect, we provide Table 3.2 which shows MNED performance with varying quality of embeddings as follows: KB embeddings learned from 1M knowledge graph entities (same as in the main experiments), from 10K subset of entities (less triplets to train with in Eq.3.3, hence lower quality), and random embeddings (poorest) - while all the other parameters or the architecture are kept the same. It can be seen that the performance notably drops with lower quality of KB embeddings. When KB embeddings are replaced by random embeddings, the network effectively prevents the contextual zeroshot matching to KB entities and relies only on lexical similarities, achieving the poorest performance.

### 3.4 Related Work

**NED task:** Most of the previous NED models leverage local textual information [31, 44] and/or document-wise global contexts [17, 40, 45, 86], in addition to other auxiliary contexts or priors for disambiguating a mention. Note that most of the NED datasets (*e.g.* TAC KBP [47], ACE [9], CoNLL-YAGO [45], etc.) are extracted from standardized documents with web links such as Wikipedia (with relatively ample textual contexts), and that named entity disambiguation

specifically for short and noisy social media posts are rarely discussed. Note also that most of the previous literature assume the availability of “candidates” or web links for disambiguation via mention-entity pair counts from training set, which is vulnerable to inconsistent surface forms of entities predominant in social media posts.

Our model improves upon the state-of-the-art NED models in three very critical ways: (1) incorporation of visual contexts, (2) addition of the zeroshot learning layer, which allows for disambiguation of unseen entities during training, and (3) addition of the lexical model that computes lexical similarity entities to correctly recognize inconsistent surface forms of entities.

**Multimodal learning** studies learning of a joint model that leverages contextual information from multiple modalities in parallel. Some of the relevant multimodal learning task to our MNED system include the multimodal named entity recognition task [77], which leverages both text and image to classify each token in a sentence to named entity or not. In their work, they employ an entity LSTM that takes as input each modality, and a softmax layer that outputs an entity label at each decoding step. Contrast to their work, our MNED addresses unique challenges characterized by zeroshot ranking of 1M knowledge-base entities (vs. categorical entity types prediction), incorporation of an external knowledge graph, lexical embeddings, etc. Another is the multimodal machine translation task [30, 98], which takes as input text in source language as well as an accompanying image to output a translated text in target language. The intuition behind multimodal translation is that images can provide contextual information that may disambiguate confounding words or phrases in source language. These models usually employ a sequence-to-sequence architecture (*e.g.* target language decoder takes as input both encoded source language and images) often with traditional attention modules widely used in other image captioning systems [100, 111]. To the best of our knowledge, our approach is the first multimodal learning work at incorporating visual contexts for the NED task.

### 3.5 Summary

As a multimodal transfer learning application, we introduce a new task called Multimodal Named Entity Disambiguation (MNED), applied on short user-generated social media posts that are composed of text and accompanying images. Our proposed MNED model improves upon the state-of-the-art unimodal NED models by 1) extracting visual contexts complementary to textual contexts, 2) by leveraging lexical embeddings into entity matching which accounts for various surface forms of entities, removing the need for fixed candidates generation process, and 3) by

performing entity matching in the distributed knowledge graph embeddings space, allowing for matching of unseen mentions and entities by context resolutions. We show that the proposed MNED model achieves the best performance with extra knowledge transfer from vastly available external knowledge sources, showing the efficacy of the heterogeneous transfer learning approaches in NLP applications.



# Chapter 4

## Multi-source CHTL with Unsupervised Transferrable Adversarial Network Training

### 4.1 Introduction

In previous chapters, we described CHTL networks that leverage knowledge from a single heterogeneous source in learning of a low-resourced target task. While we have shown successful applications of the proposed network, several challenges still remain. First, we observe that the transfer accuracy is largely limited by the number of target labels available, which in the proposed architecture is crucial in learning a robust mapping function from source to target. Ideally, the network needs to be able to leverage the vastly available unlabelled samples to remedy the scarce labels problem. In addition, given the vast number of knowledge sources available, we advocate for a system which can utilize all of the multiple sources available, ideally attenuating or amplifying each knowledge source adaptively to the target task, instead of relying on transferred knowledge from a single source.

To this end, we propose a new joint unsupervised optimization for heterogeneous transfer network which effectively leverages both unlabeled source and target data, leading to enhanced discriminative power in both tasks. Specifically, we propose the new Transferrable Adversarial Network (TAN) architecture, which modifies and applies the popular unsupervised generative adversarial network (GAN) on our CHTL network to allow for unsupervised adversarial feature learning. Instead of learning a separate discriminator and a generator for each source and

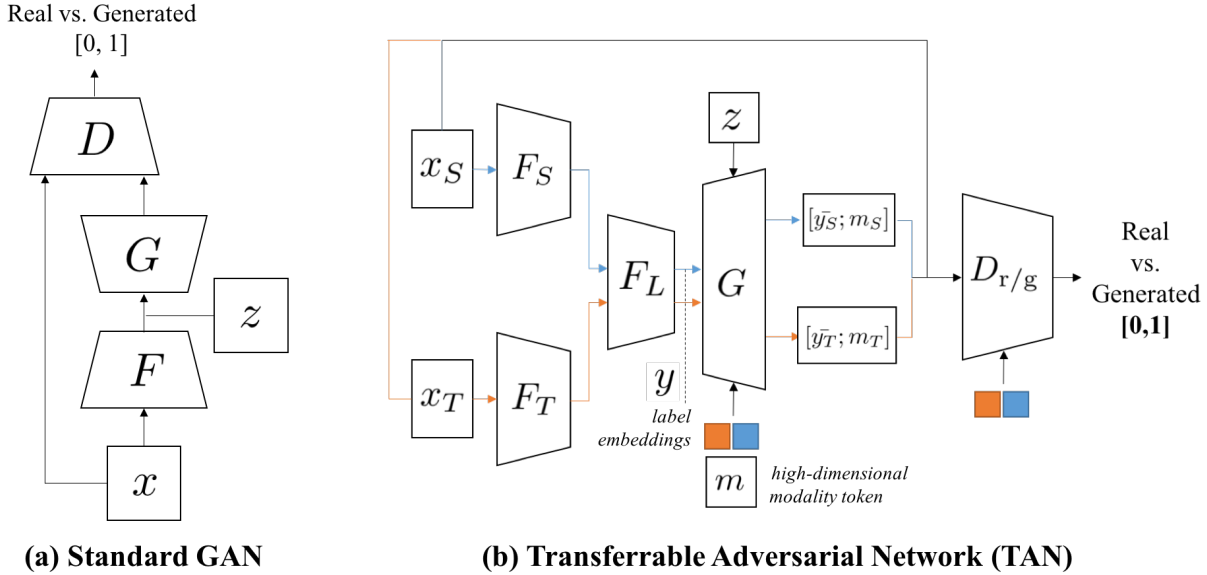


Figure 4.1: (a) Standard GAN architecture and (b) Adversarial feature learning through Transferrable Adversarial Network Architecture (TAN). For TAN, we define a unified generator  $G$  for both source and target domains which generates samples in the chosen modality  $m$ , represented with a learnable high-dimensional token parameter, from its label representation  $F(x)$ . Then, a unified discriminator distinguishes between a real input (encoded) and a generated one for both source and target domains, each identified with its corresponding modality token.

target, we improve the generality of the GAN feature learning and cross-modality learning by defined a unified discriminator and a generator with a modality identifier token. We observe that unsupervised training allows for more tractable learning of deep transfer networks, whereas the previous literature was confined to shallow transfer models due to a small number of labeled target data. In addition, we find that unsupervised transfer learning help the application of the CHTL training algorithms for multiple heterogeneous sources.

## 4.2 Method

### 4.2.1 Transferable Adversarial Network (TAN)

In previous chapters, we have presented several feature learning techniques for CHTL (*e.g.* stacked denoising autoencoders, etc.), which are crucial for learning a novel task with scarce labels. While these methods have shown success in many domains, transfer of knowledge across

more heterogeneous domains (*e.g.* images and text) require a more robust feature learning method to identify and build mapping among domains. To this end, we propose an adversarial feature learning method which leverages transferred knowledge from multiple heterogeneous domains. Figure 4.1 shows the illustration of the TAN architecture, compared against a standard GAN architecture.

Typically, a generative adversarial network **GAN** can be framed as a minimax two-player game between a generator and a discriminator, where a generator does its best to create an adversarial sample that is hard for a discriminator to discern from an original sample:

$$\min_G \max_D \mathbb{E}_x \mathbb{E}_X \log [D(x)] + \mathbb{E}_{x:z} \mathbb{E}_X \mathbb{E}_Z \log [1 - D \circ G([F(x); z])] \quad (4.1)$$

where  $F$  is a network that encodes original input  $x \sim \mathcal{X}$ ,  $G$  is a generator that takes as input encoded sample  $F(x)$  and noise input  $z$  to produce an adversarial sample, and  $D$  is a discriminator that produces a scalar probability  $[0,1]$  of a sample coming from the original distribution  $\mathcal{X}$ .

A variety of GAN architectures have been explored [13, 23, 103], most of which only requires unlabelled samples to train and often lead to robust feature representations. Advancing the previous literature on GAN training, we propose a unique transferrable adversarial network (TAN) which builds upon the CHTL network for supervised transfer learning, and adds adversarial feature learning components from unsupervised transferred knowledge. Specifically, we design the learning process such that a unified generator and a discriminator can be learned for each CHTL network that composes of multiple source and target modalities. This approach maximizes the reusability of the shared pathway within the CHTL network, unsupervised training of which can better aggregate heterogeneous knowledge into unified representation. To do this, we first define a *modality token*, a high-dimensional parameter vector that is learned for each modality as an identifier and taken as input for the generator and the discriminator. We then perform adversarial training within each modality and across multiple modalities, ensuring the reusability of the pathway across multiple domains. The following equation describes the entire TAN-CHTL objective.

$$\mathcal{L}_{\text{sup:S}} = \max_{F_S; F_L} \mathbb{E}_{x:y} \mathbb{E}_{X_S} \mathbb{E}_{Y_S} \log [\text{sim}(F(x), y)] \quad (4.2)$$

$$\mathcal{L}_{\text{sup:T}} = \max_{F_T; F_L} \mathbb{E}_{x:y} \mathbb{E}_{X_T} \mathbb{E}_{Y_T} \log [\text{sim}(F(x), y)] \quad (4.3)$$

$$\begin{aligned} \mathcal{L}_{\text{adv:r/g}} = \min_{G;m} \max_{D_{\text{r/g}}} & \mathbb{E}_{x:z} \mathbb{E}_{X_S} \mathbb{E}_Z \log [D_{\text{r/g}}([x; m_S; z])] \\ & + \mathbb{E}_{x:z} \mathbb{E}_{X_T} \mathbb{E}_Z \log [D_{\text{r/g}}([x; m_T; z])] \\ & + \mathbb{E}_{x:z} \mathbb{E}_{X_S} \mathbb{E}_Z \log [1 - D_{\text{r/g}}([G(F(x)); m_S; z])] \\ & + \mathbb{E}_{x:z} \mathbb{E}_{X_T} \mathbb{E}_Z \log [1 - D_{\text{r/g}}([G(F(x)); m_T; z])] \end{aligned} \quad (4.4)$$

where  $\mathcal{L}_{\text{sup:S}}$  and  $\mathcal{L}_{\text{sup:T}}$  refer to the CHTL supervised loss for source and target respectively,  $(x, y)$  refers to a paired sample and its label embeddings representation,  $F$  is the CHTL network that outputs projected label embeddings as defined in Section 2.3.3,  $m_S, m_T \in \mathbb{R}^{C_m}$  are high-dimensional modality embeddings for source and target. Note that the learning pathway for  $G$  and  $D$  include the feature encoder  $F_L$  shared by both source and target, allowing for unsupervised knowledge transfer. Note also that learning of TAN with multiple sources can easily be expanded by adding each source objective into the final objective. When multiple modalities are available as source, distances among the learned modality embeddings define inherent heterogeneity among different modalities, where more homogeneous modalities share more parameters from the generator and the discriminator. Following the standard GAN training procedure, we utilize alternating minmax optimization for unsupervised training of  $G$  and  $D$ .

## 4.3 Empirical Evaluation

We validate the effectiveness of the proposed unsupervised feature learning with TAN via simulations (Section 4.3.1) as well as a real-world application (multimodal image scene recognition: Section 4.3.2).

### 4.3.1 Simulation on Synthetic Datasets: Multiple Sources

We generate multiple configurations of source and target synthetic datasets and evaluate the performance with average classification accuracies on target tasks. Specifically, we aim to analyze the performance of the proposed approaches with varying source-target heterogeneity at

Configurations	$\sigma_{\text{label}}$			
	0.0	0.05	0.1	0.2
ZSR	40.2			
CHTL	41.0	40.8	40.9	40.6
CHTL:AE	42.4	41.7	41.2	41.5
CHTL:TAN	<b>43.7</b>	42.6	42.3	42.0
(2 Sources)	<b>43.8</b>	<b>43.1</b>	42.4	41.7
(10 Sources)	<b>43.8</b>	<b>43.4</b>	42.6	41.3

Table 4.1: Target task accuracy with varying source-target heterogeneity ( $\sigma_{\text{label}}$ ). 1 source is used unless otherwise noted.

varying task difficulty. We follow the similar dataset generation process as in Section 2.4.2, and generate multiple synthetic source datasets and a single target dataset with  $M$  isotropic Gaussian distributions. Each dataset generation is repeated 10 times. For the following experiment, we set  $N_S = N_T = 50000$ ,  $M_S = M_T = 100$  (original feature dimension),  $K = 100$  (number of classes), and  $\%_{L_T} = 1\%$ . Learnable parameters are optimized over the following search space for each configuration:  $C_m \in \{80,160,320,640\}$  (modality embeddings size),  $M_E \in \{80,160,320,640\}$ . We optimize the parameters with Adagrad [28] with batch size 1, learning rate 0.01, epsilon  $10^{-9}$ , and decay 0.1. We obtain 5-fold results for each dataset generation, and report the overall average accuracy in Table 4.1. Bold denotes statistically significant improvement over the CHTL baseline.

**Main results:** we observe that the proposed TAN approaches outperform the CHTL baselines, with significant improvement at lower source-target heterogeneity, showing the effectiveness of the adversarial transfer learning method. Specifically, we observe that the TAN model is able to leverage knowledge from a large number of source datasets that are similar to target dataset ( $K = 10$  at  $\sigma_{\text{label}} = 0.05, 0.1$ ), building more robust label embeddings from unsupervised source data with a similar class. As source-target heterogeneity increases, multiple heterogeneous datasets only induce more negative transfer, hence we do not see statistically significant improvement over the baseline.

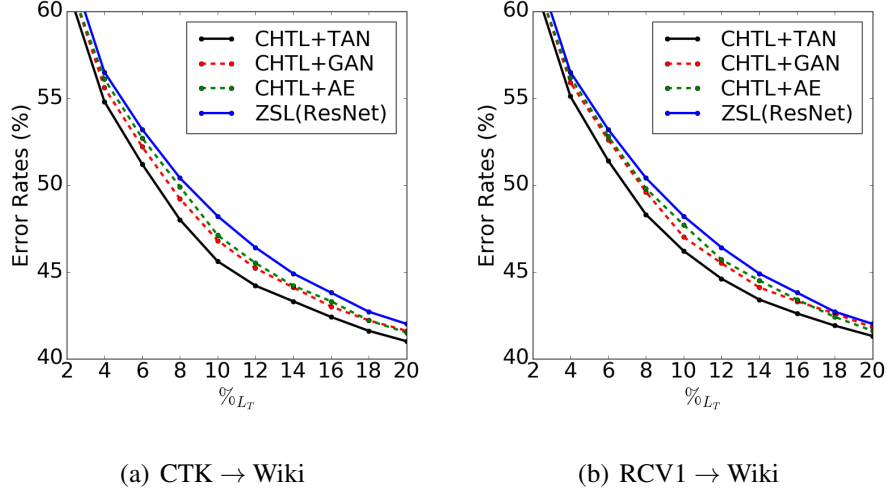


Figure 4.2: Image scene classification task with varying amount of target image dataset labels (S: Wiki) with fully labeled text sources (a) Czech CTK and (b) RCV1 datasets. Knowledge graph embeddings from FreeBase are used as label embeddings. Chance accuracy (the most frequent class): 15.7%.

### 4.3.2 Application: Multimodal Text-aided Image Scene Recognition

We study multimodal text-aided image scene recognition task for the Wikipedia Crossmodal dataset [85], which consists of 2,669 images across 30 categories (*e.g.* art, history, sport, etc.). We formulate an image scene recognition task of classifying an input image as one of the categories, trained along with other text datasets such as RCV1 (116 classes) and Czech CTK (60 classes) datasets. Note that while most of the object categorization image datasets (*e.g.* ImageNet [91]) often composes of the label space that is distinct from text datasets (hence making CHTL training in those datasets more challenging), the Wikipedia image dataset and the text classification datasets we use share similar class coverage of abstract concepts or topics, allowing for more tractable transfer of knowledge. As baselines of unsupervised training, we also consider CHTL networks with (1) autoencoder loss as well as (2) standard GAN representation learning loss. For GANs, we learn separate generators and discriminators for each source and target, without the use of modality embeddings that constitute the unified generator and the discriminator:

$$\begin{aligned}
\mathcal{L}_{\text{gan}} = \min_G \max_{D_S} & \mathbb{E}_{x:z} \mathbb{E}_{X_S} \mathbb{E}_Z \log [D_S([x; z])] + \mathbb{E}_{x:z} \mathbb{E}_{X_T} \mathbb{E}_Z \log [D_T([x; z])] \\
& + \mathbb{E}_{x:z} \mathbb{E}_{X_S} \mathbb{E}_Z \log [1 - D_S([G_S(F(x)); z])] + \mathbb{E}_{x:z} \mathbb{E}_{X_T} \mathbb{E}_Z \log [1 - D_T([G_T(F(x)); z])]
\end{aligned}
\tag{4.5}$$

**Parameters:** We tune the parameters of each model with the following search space (bold indicate the choice for our final model): modality embeddings size ( $C_m$ ): {80, **160**, 320, 640} intermediate embeddings dimension ( $M_C$ ): {20, 40, 80, 160, **320**, 640}, label embeddings dimension ( $M_E$ ): {100, 200, **300**}, knowledge graph embeddings size: {100, 200, **300**}. For image classifier, we use the pre-trained ResNet [43] as a visual descriptor extractor and add another layer with to project to the label embeddings space. We optimize the parameters with Adagrad [28] optimizer with batch size 100, learning rate 0.01, epsilon  $10^{-8}$ , and decay 0.01.

**Main results:** Figure 4.2 shows the results on the image scene classification task with varying amount of target Wikipedia image dataset labels available ( $\%_{L_T}$ ) with fully labeled text sources ( $\%_{L_S} = 100\%$ ): (a) Czech CTK and (b) RCV1 datasets. It can be seen that the CHTL network trained with TAN outperform other baselines at most varying amounts of target samples, showing the efficacy of the TAN training. While other CHTL baselines outperform the ZSL baseline built on top of ResNet, the difference is not as significant. Note also that CHTL:AE and CHTL:GAN achieve similar performance, which both optimizes for reconstruction loss for source and target separately. This separate training results in the non-optimal sharing of the intermediate  $F$  label predictor network, showing less significant improvement. The performance gap between TAN and baselines is more distinct with using the Czech CTK dataset as a source than the RCV1 dataset, showing that CHTL performance is bound to the degree of homogeneity between source and target.

## 4.4 Related Work

**Adversarial feature learning** approaches [23, 87] have recently been proposed, where the adversarial objective leads to robust capturing of the latent features. These work shift the conventional focuses of GAN approaches in their application for generation of images, and study the impact of the generated features (with varying arbitrary complexity) in downstream tasks. We extend the previous literature to heterogeneous transfer learning scenario by (1) applying GAN training to the CHTL network, and by (2) inducing modality embeddings to allow for synced knowledge transfer across multiple sources.

**GAN training for domain adaptation:** there have been recent work which apply the unsupervised GAN training for domain adaptation tasks. For example, [13, 103] propose a unique GAN architecture where the adversarial objective is to discern the target domain with the generated samples from source domain, adversarially resembling target distribution. [60] proposes

a Coupled GAN (CoGAN) which learns a shared network to model joint distribution of multi-domain images, without requiring tuples of corresponding images.

While the previous literature have shown successful applications in domain or transfer adaptation tasks in computer vision, limited work have addressed the adversarial transfer learning as unsupervised feature learning techniques. In addition, the key difference to our proposed work is that TAN addresses the applicability of adversarial training in the heterogeneous transfer learning settings (for both source and target), improving the practical reach of the algorithm.

**Multi-source transfer learning:** a number of work have addressed the problem of optimally transferring knowledge from multiple available sources [33, 37, 51, 112]. [33], for example, attempts to discover a latent subspace that adjusts and transforms multiple sources, from which the joint training can be conducted. [51] explores kernel mean matching approaches to transfer knowledge and re-weight instances from multiple sources with an application in protein-protein interactions prediction. Our proposed multi-source approach is unique in that we generalize the problem for multiple heterogeneous sources and tasks (classes), and that thus a unique unsupervised feature learning method is employed.

## 4.5 Summary

We propose a new method for training the completely heterogeneous transfer learning network with unsupervised transfer loss, with an aim to make more robust projections with deeper transfer networks. Specifically, we propose the new Transferrable Adversarial Network (TAN) architecture, which modifies and applies the popular unsupervised generative adversarial network (GAN) on our CHTL network to allow for unsupervised adversarial feature learning. The adversarial feature learning essentially leverages knowledge transferred from multiple heterogeneous sources, consequently improving downstream task performance. Results on synthetic datasets with varying heterogeneity and task difficulty provide new insights on the conditions and parameters in which TAN can succeed. The proposed approach is general and thus can be applied in other domains, as indicated by the domain-free simulation results.

# Chapter 5

## Proactive Learning with Multiple Heterogeneous Labelers

### 5.1 Introduction

The challenge in many machine learning or data mining tasks is that while unlabeled instances are abundant, acquiring their class labels often requires extensive human effort. The active learning paradigm addresses the challenge of insufficient labels by interactively optimizing the selection of queries [56, 90]. Several studies have shown that active learning reduces the sample complexity in a variety of applications, including network / graph analysis [10], text mining [1, 65], etc. However, active learning relies on tacit assumptions which prove limiting for real problems and applications. Primarily, active learning assumes the existence of a single omniscient labeling “oracle”, whereas in real life it is more common to have multiple sources of annotations with different reliabilities or areas of expertise. In addition, active learning assumes that labeling different instances incurs uniform cost, regardless of the difficulty or the expected accuracy inherent in each annotation task. Some research has addressed cost-sensitive active learning, but only with respect to instances and features [67, 76].

Proactive learning has been proposed as a means to relax the unrealistic assumption of a single omniscient labeling oracle, permitting multiple labelers with different accuracies, different availabilities and different costs [24, 25, 106, 116]. This line of work has shown that proactive learning extends its reach to practical applications by combining estimation (learning) of the labeler accuracy with maximum utility of the labeler and instance selection. However, the prior work assumes that labeler accuracy is independent of labels in multi-class problems, calculating

only an average accuracy across labels. In this paper, we address this limitation by explicitly estimating the dependency between label and labeler in order to optimize the assignment of new instances to labelers based on class priors or best-estimate of class membership.

We illustrate the novel contribution of the proposed method in the following example. Consider, for instance, the multi-class problem of a medical diagnosis of a patient with a disease that we know very little about (uncertainty in data): given multiple physicians who specialize in completely different areas (e.g. an oncologist, a cardiologist, or an internal medicine doctor), we need to assign one of the experts to diagnose the patient correctly (proactive learning selection). Querying an expert who has the best overall diagnosis performance across multiple domains (assuming uniform reliability across classes) may not give the most accurate results (analogous to the previous proactive learning methods [25]), because the chosen expert might lack knowledge in the specific disease of the patient. If we know that the patient has seemingly cancer symptoms (posterior class probability) and that an oncologist usually has the deepest understanding of cancer issues (estimated labeler accuracy given a specific class), we can leverage this information to better delegate a task to its respective expert. Note that this is indeed a real world challenge, as shown in our experiment with the Diabetes dataset in Section 5.3.

Similarly, our new method estimates labeler accuracy on a per-class basis, or per subset-of-classes basis, providing a considerably new level of flexibility in proactive learning. A probabilistic approach to model annotator accuracy in a binary classification task was proposed by [114], but modeling annotator performance over the entire data in multi-classes is a complex task that requires a large number of training examples. Our method efficiently reduces the cost and complexity involved in estimating the labeler accuracy over multiple classes by employing the reduced per-class estimation method.

Another approach proposed by [19, 82, 105] that handles unreliable annotators in crowdsourcing scenarios is to query multiple annotators repetitively to estimate the ground truth label for each instance. The integration of judgements from crowd is typically done via majority vote or selective sampling, but these work do not comprise estimating individual per-task (per-class) expertise for selective recruitment of crowd members. In addition, these methods are not desirable in active learning scenarios because querying multiple annotators repeatedly for a single instance incurs multiple costs, whereas our method tries to find the one most cost-efficient expert who can answer the query reliably.

The framework that we propose is flexible and can work with any instance selection criterion or any supervised learning method. In order to further improve the effect of the proposed algo-

rithm, we also propose a new density-based sampling strategy for multi-classes that considers the concept of *conflictivity* of the label distribution. We integrate our metrics as an ensemble method [6, 26, 66], and show that our selection criterion outperforms the traditional density-weighted sampling methods [79, 123], especially when there are multiple unreliable annotators.

The rest of the paper is organized as follows: Section 5.2 describes in detail the proposed proactive learning framework and presents the new density-based sampling strategy. The empirical results are reported and analyzed in Section 5.3, and we give our concluding remarks and proposed future work in Section 5.4.

## 5.2 Method

In this section, we present a proactive learning method for multi-classification tasks when multiple domain experts are present. In our scenario, we assume that there exist multiple narrow experts and one meta oracle. A narrow expert has expertise in the subset of classes from the data, and each expert’s expertise may or may not overlap. The probability of getting a correct answer given a query depends on the difficulty of the classification task for the expert. In other words, a narrow expert is more reliable in annotating the data for which the ground truth labels are within the expert’s expertise. A meta oracle, on the other hand, has expertise in every category. The cost of each expert or a meta oracle varies depending on the difficulty of the task, the skewness of the data, and its range of expertise areas. We experiment with various combinations of cost ratios to simulate different real-world situations.

### 5.2.1 Proactive Learning with Multiple Domain Experts

In proactive learning, we jointly select the optimal oracle and the instance at which the current system’s performance would best improve. As such, the solution to the problem is casted as a utility maximization subject to a budget constraint [24]. The objective of the problem can thus be formulated as:

$$\begin{aligned} & \max_{S, UL} E[V(S)] - \lambda \left( \sum_k t_k \cdot C_k \right) \\ \text{s.t. } & \sum_k t_k \cdot C_k \leq B, \quad \sum_k t_k = |S| \end{aligned} \tag{5.1}$$

where  $S$  is the set of instances to be sampled,  $UL$  is the set of unlabeled samples, and  $E[V(S)]$  is the expected value of information of the sampled data to the learning algorithm.  $V(S)$  may be

---

**Algorithm 2** Proctive Learning with Multiple Experts

---

**Input:** a multiclass classifier  $f$ , the pre-defined set of classes  $\mathcal{C}$ , labeled data  $L$ , unlabeled data  $UL$ , budget  $B$ , oracles  $k \in K$  with cost  $C_k$ , each with expertise in some classes

**Output:**  $f$

Obtain  $P(ans|y = c, k), \forall c \in \mathcal{C}, k \in K$  from Algorithm 3

Let  $C_T$  be the cost spent so far,  $C_T = 0$

**while**  $C_T < B$  **do**

    Train  $f$  on  $L$

    Choose  $(x, k) = \operatorname{argmax}_{k \in K; x \in UL} U(x, k)$  (Eq. 5.3)

    Query the label  $y = \operatorname{query}(x, k)$

$L = L \cup \{(x, y)\}$ ,  $UL = UL - \{(x, y)\}$

$C_T = C_T + C_k$

**end while**

---

replaced by any active learning selection criterion, such as the uncertainty-based sampling [55].  $k \in K$  denotes the chosen oracle from the set of experts, and  $\lambda$  is a weighting parameter that determines how much the value of information is penalized by the oracle cost.  $C_k$  and  $t_k$  refer to the cost of the chosen expert  $k$  and the number of times it is queried, respectively.  $B$  is the total amount of budget for querying oracles. However, Equation 5.1 is a complex optimization problem because the learning function is updated at every iteration while the samples to be queried and their labels are unknown to the learner. Therefore, we employ a greedy approximation of the problem which chooses a small batch of samples to be queried at every iteration that maximizes the utility under the budget constraint:

$$(x, k) = \operatorname{argmax}_{x \in UL; k \in K} U(x, k) \quad (5.2)$$

where  $U(x, k)$  refers to a utility score when a sample  $x$  is annotated by an oracle  $k$ . We define the utility score such that it incorporates the reliability and the cost of an oracle as well as the base value of information of an instance. This ensures that the learner does not always choose the most reliable, and the most costly oracle, but encourages the learner to select the most cost-effective pair of an instance and an expert that is likely to give a correct answer. Thus, we can formulate the utility score as follows:

$$U(x, k) = \frac{V(x) \cdot P(ans|x, k)}{C_k} \text{ for } k \in K \quad (5.3)$$

where  $V(x)$  is the value of the information of the sampled data to the learner, and  $P(ans|x, k)$  is the probability of receiving the correct answer from an expert  $k$  given the sample  $x$ . We therefore assign a higher utility for the instances that have a higher value of information and a higher probability of being labeled correctly, while having a cheaper cost of annotation. Algorithm 2 describes the cost-optimized proactive learning process using the utility function formulated above. In most of the real world datasets, however, the accuracy information of the labeling sources  $P(ans|x, k)$  is not given to the learner, and thus it needs to be estimated prior to the active selection process. While there may be various ways to estimate the accuracy of the labeling sources, the challenge is to minimize the number of queries that need to be made to each expert when there does not exist any query history a priori. The next section describes an efficient implementation of estimating expertise of labeling sources through selective sampling and the reduced per-class estimation.

## 5.2.2 Expertise Estimation

We assume that the oracle’s expertise is distinctly aligned over a subset of classes rather than over the entire distribution of the data. We can then reduce the estimated labeling accuracy  $P(ans|x, k)$  as follows:

$$E[P(ans|x, k)] = \sum_{c \in \mathcal{C}} P(y = c|x) \cdot P(ans|k, y = c) \quad (5.4)$$

where  $\mathcal{C}$  is the set of categories in a multi-classification task,  $P(y = c|x)$  is the class posterior probability of the label for the sample  $x$  being  $c$  (predicted by the learner), which is an estimate of the true underlying label density.  $P(ans|k, y = c)$  is the estimated probability of the expert  $k$  answering correctly for the label  $c$ . In other words, we integrate the learner’s prediction of an instance with the expert’s class-wide labeling accuracy. With the given formulation of  $P(ans|x, k)$ , the utility function favors the samples that have a higher probability of belonging to a certain label  $c$ , which an expert  $k$  is has expertise in. The meta-oracle will almost always have a higher value for  $P(ans|x, k)$ , but the overall utility will be dampened by a higher  $C_k$  as in Equation 5.3.

We consider two different scenarios for estimating  $P(ans|k, y = c)$  (detailed in Algorithm 3): (1) when there are labeled samples already available (assuming ground truth), and (2) when there is no labeled sample at all. If we are given the ground-truth labels for  $n$  instances, we inquire for the labels of those instances to each expert  $k$  and compute the labeler accuracy per class

---

**Algorithm 3** Expertise Estimation for Multiple Experts

---

**Input:** Labeled data  $L$ , unlabeled data  $UL$ , oracles  $k \in K$  each with expertise in some of the classes

**Output:**  $P(ans|y = c, k) \forall c \in \mathcal{C}, k \in K$

**if**  $|L| > 0$  for each  $c \in \mathcal{C}$  **then**

**for** each  $x$  and ground truth label  $(x, z) \in L$  **do**

**for** each  $k \in K$  **do**

$y^{(k)} = query(x, k)$

      Update  $P(ans|k, y = c)$  with  $h(y^{(k)}, z)$

**end for**

**end for**

**else**

  Choose  $n$  samples randomly from  $UL$

**for** each sample  $x$  **do**

    Initialize  $v(c) = 0 \forall c \in \mathcal{C}$

**for** each  $k \in K$  **do**

$y^{(k)} = query(x, k)$

$v(y^{(k)}) = v(y^{(k)}) + 1$

**end for**

$y^{maj} = \max_{c \in \mathcal{C}} v(c)$

    Set  $P(ans|k, y = y^{maj})$  with  $h(y^{(k)}, y^{maj}) \forall k$

**end for**

**end if**

---

with the available ground-truth labels. Therefore, we define the empirical labeler accuracy per class as follows:

$$P(ans|k, y = c) = \frac{1}{n} \sum_{i=1}^n h(y_i^{(k)}, z_i) \forall k \in K \quad (5.5)$$

where  $y_i^{(k)}$  is the prediction of  $x_i$  by an expert  $k$ ,  $z_i$  is the ground-truth label of  $x_i$ , and  $h(y_i^{(k)}, z_i) \in \{1, 0\}$  is an indicator function which is equal to 1 if  $y_i^{(k)} = z_i$  and 0 otherwise. When there is no labeled sample available, we choose  $n$  samples from the unlabeled set, and inquire for the label of each sample to every expert. We estimate the ground-truth label of each instance by majority vote on experts ( $= y^{maj}$ ), and compute  $P(ans|k, y = c)$  with  $h(y_i^{(k)}, y_i^{maj})$ . Note that  $P(ans|k, y = c)$  is independent of  $x$ , which thus gives only a brief class-sensitive belief about the

expert’s labeling accuracy. This simplified estimation allows for practical benefits in estimating labeler accuracy given the limited budget for the expertise discovery phase. In our experiments (Section 5.3.2), we show that this brief knowledge of class-wide expertise greatly improves the performance when incorporated into the active learning selection formula. We also present the empirical analysis of the performance for varying degrees of errors for the estimated expertise.

### 5.2.3 Density-based Sampling for Multi-classification Tasks

While our framework is flexible and can work with any selection strategies, we propose a new density-based sampling method for multiple classes with multiple imperfect oracles to further improve the effect of the proposed algorithm.

Some of the most notable work done on the density-weighted uncertainty sampling (DWUS) strategies for active learning include the pre-clustering method [79, 123], which incorporates the prior density  $p(x)$  of the data in the selection criterion. This method encourages the selection of more representative samples (e.g. centroids of denser clusters) at each query iteration, and avoids repetitively querying the samples that are in the same cluster.

We extend the previous work to accommodate for a multi-classification problem where labels are acquired from unreliable experts. First of all, if the expert that labeled a sample in a cluster is not reliable, we should in fact encourage querying samples from that cluster until we obtain a more credible label. Second, if a cluster encompasses conflicting opinions, or a cluster is placed over the decision boundaries, we should encourage querying from that cluster to better tune the decision boundaries between neighboring classes.

As such, we propose a new multi-class information density (MCID) as follows, which comprises of three components: (1) density, (2) unknownness, and (3) conflictivity. The density component measures how densely samples are positioned around a given point, and the unknownness component measures how many samples are labeled thus far. The conflictivity component measures how heterogeneous the label distribution is around a given sample. The conflictivity term encourages the learner to favor a cluster that still has conflicting and unresolved class distribution over a slightly denser cluster with unanimous class distribution.

A simple yet efficient implementation of MCID is to pre-cluster the dataset and calculate the three components in each cluster locally. For a given cluster  $q \in Q$ , where  $Q$  refers to a set of clusters of the dataset and  $q$  is a set of labeled and unlabeled samples within the cluster, the

MCID of a sample is defined as follows:

$$\begin{aligned} \rho(q, x) = & p(x) \cdot \frac{|q_{UL}|}{|q|} \\ & \cdot \left( - \sum_{c \in \mathcal{C}} P(y = c|q) \cdot \log P(y = c|q) \right) \end{aligned} \quad (5.6)$$

where  $\rho(q, x)$  is the MCID of a sample  $x$  in a cluster  $q$ ,  $p(x)$  is the density at a point  $x$ ,  $q_{UL}$  is a set of unlabeled samples within the cluster, and  $\mathcal{C}$  is the set of label classes. We induce  $p(x)$  using a  $|Q|$  Gaussian mixture model with weights  $P(q)$ , hence  $p(x) = \sum_{q \in Q} p(x|q)P(q)$ , where  $p(x|q)$  is a multivariate Gaussian sharing the same variance  $\sigma^2$  [79]:

$$p(x|q) = (2\pi)^{-\frac{d}{2}} \sigma^{-d} \exp\left\{-\frac{\|x - c_q\|^2}{2\sigma^2}\right\} \quad (5.7)$$

where  $c_q$  is the centroid of the cluster  $q$ . We estimate the cluster prior  $P(q)$  via an EM procedure:

$$\begin{aligned} P(q|x_i) &= \frac{P(q) \exp\left\{-\frac{\|x_i - c_q\|^2}{2\sigma^2}\right\}}{\sum_{q \in Q} P(q) \exp\left\{-\frac{\|x_i - c_q\|^2}{2\sigma^2}\right\}} \\ P(q) &= \frac{1}{N} \sum_{i=1 \dots N} P(q|x_i) \end{aligned} \quad (5.8)$$

where  $N$  is the size of the sample set. The second term in Equation 5.6 measures the proportion of samples known at each iteration. The last term is the entropy of class distribution within the cluster, which approximates the conflictivity of the cluster.

Note that the MCID measure does not contain any knowledge about how informative each individual point is. Therefore, the ultimate value function of an instance is given as a combination of the basis selection criteria  $\phi(x)$  (e.g. the uncertainty-based selection [55], etc.) and the MCID. Therefore:

$$V(x) = \beta \phi(x) + \rho(q, x) \quad (5.9)$$

where  $\beta \in (-\infty, \infty)$  is a weight parameter. For simplicity, in the following experiments, we use  $\beta = 1$  and  $\phi(x) = H(x) = -\sum_{y \in \mathcal{Y}} P(y|x) \cdot \log P(y|x)$ , or the entropy of the probability distribution [93].

### 5.3 Experimental Evaluation

Table 5.1 shows the summary of the datasets we used in our experiments. The Diabetes 130 U.S. Hospitals dataset [99] contains the attributes that identify the medical specialty of each annotator, as well as the specific diagnosis type (label) and medical records (attributes) of each patient

Table 5.1: Overview of Datasets.

Dataset	# Experts	# Classes	Size
Diabetes 130 U.S. Hospitals	3	3	13300
20 Newsgroups	5	20	7000
Landsat Satellite	3	6	3000
Image Segmentation	3	7	2310
Vehicle	4	4	946

instance. For our experiment, we make a subset of the dataset by choosing the three frequent diagnosis types, and consider three major medical specialties (Internal Medicine, Family/General Practice, Surgery-General). Each instance is annotated by only one annotator with a single medical specialty, and therefore we assume that labels we query come from an expert classifier model which is trained over the instances that each respective expert has annotated.

The rest of the datasets in Table 5.1 do not have any annotator information, and thus we simulate multiple narrow experts as follows. We assume that the narrow experts’ expertise does not overlap but together they cover every category. For example, we train 5 narrow experts for the 20 Newsgroups dataset, each specializing in 4 ( $=20/5$ ) unique classes (See Table 5.1). In order to simulate the reliability of the oracles with different expertise, we assume that a narrow expert resembles a classifier trained on the dataset of which the labels of the samples in its non-expertise categories are partially noised. The noise ratio was adjusted so that the overall labeling accuracy is around 50% for each non-expertise category. The meta oracle is trained on the entire dataset without any artificial noise. This simulates a realistic situation where every annotator has a varying degree of non-zero error rates on different classes. The results are averaged over 10 runs for every experiment.

### 5.3.1 Multi-class Information Density

We compared the proposed multi-class information density (MCID) method (detailed in Section 5.2.3) on several datasets with two other baseline selection criteria: (1) *US*, which uses the traditional uncertainty sampling (US) method, (2) *DWUS*, which employs the widely used density-only weighted uncertainty sampling (DWUS) method [79, 123].

Table 5.2 shows the classification error rates at four different stages of active learning (cost = 0.25, 0.50, 0.75, 1.0), where each label is obtained from a randomly chosen narrow expert

Table 5.2: Comparison of error rates of MCID vs DWUS vs US

Dataset	Cost	Classification Error Rates		
		MCID	DWUS	US
Diabetes	0.25	0.402	0.411	0.423
	0.50	0.374*	0.407	0.409
	0.75	0.362*	0.399	0.400
	1.00	0.354*	0.393	0.398
20 Newsgroups	0.25	0.508	0.521	0.516
	0.50	0.431*	0.470	0.488
	0.75	0.388*	0.428	0.453
	1.00	0.350*	0.381	0.388
Vehicle	0.25	0.333	0.335	0.350
	0.50	0.281	0.294	0.301
	0.75	0.260*	0.279	0.286
	1.00	0.242*	0.266	0.271

to allow for a realistic and heterogeneous label distribution. Both *DWUS* and *MCID* methods outperform the baseline (*US*), which greatly saves the annotation cost to converge. There is a time-variant performance difference on these two methods: the *DWUS* method performs almost the same as the *MCID* method at the beginning, which indicates that the conflictivity component of the measurement does not improve the performance when not enough labels are given. Once enough labels are given (cost  $\geq 0.5$ ), the *MCID* method outperforms the previous density-only weighted baseline (*DWUS*). Statistically significant improvements ( $p < .05$ ) over the baselines at each cost are marked as \*.

### 5.3.2 Multiple Experts

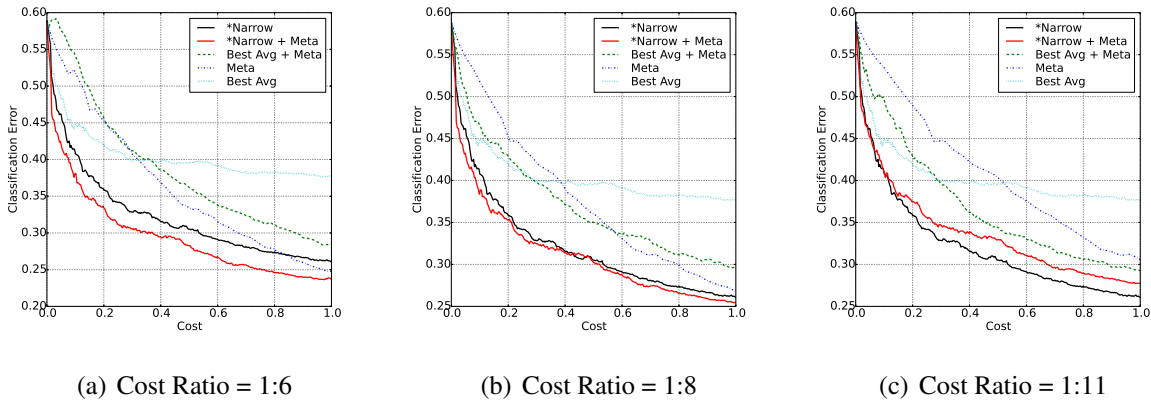


Figure 5.1: Comparison of error rates on the Diabetes 130 U.S. Hospitals Dataset with different cost ratios (when expertise was estimated via ground truth samples). The X-axis denotes the normalized total cost, and the Y-axis denotes the classification error. Our proposed methods are marked as \* in the legends.

The following figures show the results for the proposed proactive learning algorithm on five different datasets: Diabetes 130 U.S. Hospitals, 20 Newsgroups, UCI Landsat Satellite, UCI Statlog Image Segmentation, and UCI Vehicle. For each dataset, we vary the cost ratio of a narrow expert to the meta oracle, the initial number of labeled / unlabeled samples to estimate  $P(ans|k, y = c)$ , and the proactive learning methods.

Figures 5.1 and 5.2 show the performance of the proposed algorithm with varying cost ratios on the Diabetes dataset and the 20 Newsgroups dataset, respectively. Due to space constraints, we present the rest of the results in Tables 5.3 and 5.4.

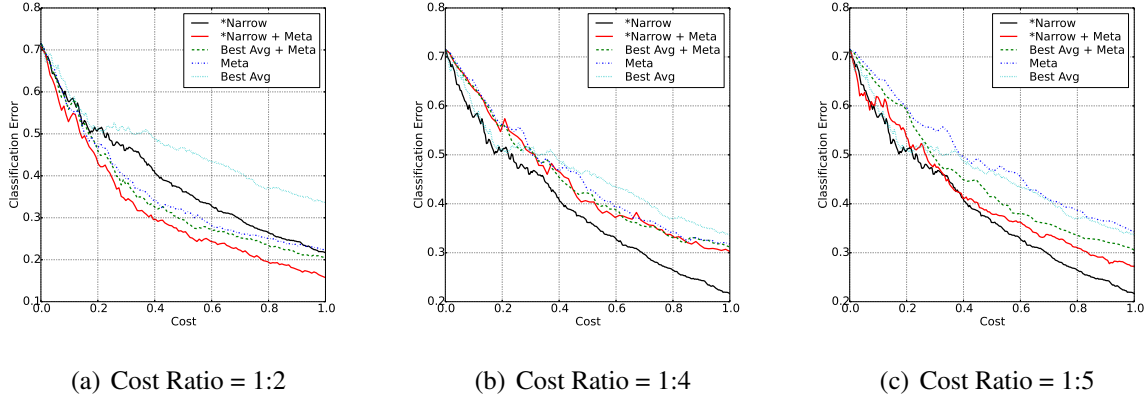


Figure 5.2: Comparison of error rates on the 20 Newsgroup Dataset with different cost ratios (when expertise was estimated via ground truth samples). The X-axis denotes the normalized total cost, and the Y-axis denotes the classification error. Our proposed methods are marked as \* in the legends.

There are three baselines that were considered: (1) *Best Avg* (Dotted Cyan), where the learner always asks one of the narrow experts that has the highest average  $P(ans|x, k)$  under the assumption that labelers are uniformly accurate across multiple classes, (2) *Meta* (Dashdot Blue), where for a higher price, the learner always asks the meta oracle that has expertise in every category, and (3) *Best Avg+Meta* (Dashed Green), which jointly chooses between the more reliable meta oracle and the fallible narrow expert under the uniform reliability assumption. Note that the baseline (3) refers to the proactive learning method proposed earlier by [25]. For all of the baseline methods, we use our proposed MCID method as a criterion for the instance selection.

The two proactive learning methods that we propose (marked as \*) are: (1) *Narrow* (Solid Black), where the learner selectively chooses the best pair of a sample and a narrow expert that yields the highest utility at each iteration, and (2) *Narrow+Meta* (Solid Red), which also includes the meta oracle in the pool of experts. We show that each proposed method has an advantage over each other depending on the availability and the affordability of the meta oracle.

In all of our experiments, *Narrow* and *Narrow+Meta* significantly outperform the *Best Avg* baseline. When the meta oracle is expensive (as in Figures 5.1(c), 5.2(b), 5.2(c)), *Narrow* significantly outperform the *Meta* and the *Best Avg+Meta* baseline ( $p < .01$ ). In reality, the meta oracle would be significantly more expensive than the narrow experts or it may not exist at all. The results are thus promising because the proposed method can perform very well even in the absence of the meta oracle. When the meta oracle is cheaper (Figures 5.1(a), 5.2(a)), on the other hand,

Table 5.3: Comparison of error rates on the UCI Datasets (when expertise was estimated via ground truth samples)

Dataset	Cost Ratio	Cost	Classification Error				
			*Narrow	*Narrow +Meta	Best Avg +Meta	Meta	Best Avg
Landsat Satellite	1:2	0.25	0.311	0.329	0.335	0.331	0.326
		0.50	0.217	0.223	0.246	0.283	0.249
		0.75	0.133	0.155	0.166	0.237	0.195
		1.00	<b>0.069</b>	0.098	0.119	0.185	0.128
Image Segmentation	1:2	0.25	0.111	0.126	0.142	0.203	0.169
		0.50	0.064	0.080	0.081	0.060	0.130
		0.75	0.050	0.047	0.050	0.043	0.113
		1.00	0.045	<b>0.032</b>	0.041	<b>0.029</b>	0.118
Vehicle	1:1.5	0.25	0.302	0.260	0.262	0.252	0.325
		0.50	0.231	0.201	0.211	0.210	0.261
		0.75	0.166	0.141	0.179	0.168	0.229
		1.00	0.132	<b>0.103</b>	0.148	0.139	0.215

Table 5.4: Comparison of error rates (when expertise was estimated via majority vote)

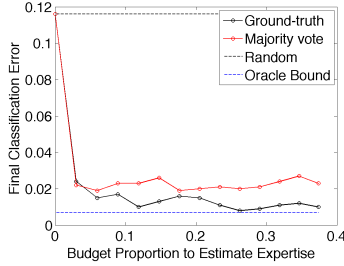
Dataset	Cost Ratio	Cost	Classification Error				
			*Narrow	*Narrow +Meta	Best Avg +Meta	Meta	Best Avg
Diabetes	1:11	0.25	0.355	0.375	0.430	0.471	0.421
		0.50	0.317	0.346	0.348	0.401	0.399
		0.75	0.276	0.301	0.306	0.343	0.372
		1.00	<b>0.269</b>	0.283	0.297	0.308	0.367
20 Newsgroups	1:5	0.25	0.461	0.490	0.501	0.559	0.521
		0.50	0.366	0.394	0.431	0.472	0.466
		0.75	0.289	0.335	0.343	0.397	0.395
		1.00	<b>0.221</b>	0.298	0.306	0.355	0.357
Landsat Satellite	1:2	0.25	0.321	0.331	0.351	0.329	0.334
		0.50	0.217	0.236	0.283	0.279	0.248
		0.75	0.138	0.184	0.223	0.234	0.199
		1.00	<b>0.078</b>	0.116	0.129	0.185	0.126
Image Segmentation	1:2	0.25	0.154	0.169	0.171	0.206	0.177
		0.50	0.087	0.060	0.062	0.062	0.135
		0.75	0.056	0.042	0.039	0.045	0.112
		1.00	0.042	<b>0.026</b>	<b>0.028</b>	<b>0.031</b>	0.118
Vehicle	1:1.5	0.25	0.301	0.251	0.246	0.250	0.326
		0.50	0.248	0.210	0.209	0.208	0.277
		0.75	0.182	0.167	0.170	0.169	0.263
		1.00	0.141	<b>0.112</b>	0.142	0.140	0.247

the joint *Narrow+Meta* method outperforms both the *Meta* baseline and the *Narrow* method ( $p < .01$ ), which indicates that the proposed algorithm jointly optimizes between the meta oracle and the narrow experts in the most cost-efficient way. While the joint *Best Avg+Meta* baseline [24] outperforms the other two baselines when the cost ratio is high, the improvement is not as significant as in our proposed methods for this experiment. This is because the previous work fails to capture the noisy labeler accuracy which varies by class. Tables 5.3 and 5.4 show similar results on other UCI datasets. Note that the proposed algorithm works successfully even when there is no ground truth sample available to estimate expertise. While the ground truth case generally performs better than the majority vote estimation method, they eventually converge at almost the same accuracy level (Tables 5.3 and 5.4).

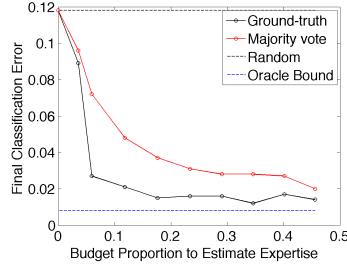
Figure 5.3 shows the difference in the final error rate at convergence as a function of the initial budget that was set aside to estimate expertise of each narrow expert. We assume that acquiring a label to estimate expertise incurs the same cost as querying an expert during the active learning process. If the learner spends a large enough budget to estimate expertise, it can more accurately delegate an instance to a narrow expert that has expertise for the chosen instance. *Ground truth* (Solid Black) represents the convergence accuracy when we employ the *Narrow* method, where the expertise was estimated using the ground truth samples with the marked proportion of the budget. *Majority vote* (Solid Red) refers to the *Narrow* method where the expertise was estimated using the majority vote method. As a baseline, we present the final accuracy when there is no prior knowledge of expertise, thus randomly choosing an expert at each iteration (Dotted Black). We also present an oracle bound (Dotted Blue), where we assume that we have perfect estimation of expertise of each expert, thus delegating an instance to the correct expert every time. For all of the UCI datasets that were tested, the results show that the proposed method works significantly better than the baseline even with a limited budget to estimate expertise. This result shows that even with the imperfect estimation of expertise we can still improve the performance greatly. The *ground truth* method utilizes improved estimation of expertise, thus outperforming the *majority vote* method.

## 5.4 Summary

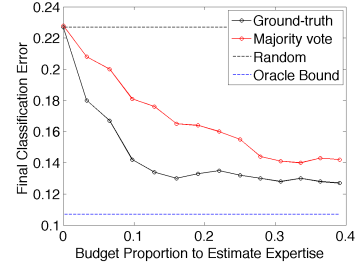
The novel contributions of this chapter are as follows: we proposed an efficient proactive learning algorithm for which there are multiple class-sensitive experts with varying costs whose expertise are distinctly aligned over multiple classes. The proposed method formulates a cost-



(a) Landsat Satellite



(b) Image Segmentation



(c) Vehicle

Figure 5.3: Final error rate at convergence as a function of the initial budgets set aside for expertise estimation on the UCI datasets, in proportion to the total budget spent to acquire labels until it reaches convergence.

driven decision framework which maximizes the utility across oracle-example pairs. We showed that our algorithm efficiently delegates each narrow expert to an unlabeled instance that the chosen expert is most likely to have expertise in. The empirical results on the datasets with both real and simulated experts demonstrate the effectiveness of this approach under different cost conditions. Specifically, when there exists an affordable meta oracle, the proposed algorithm jointly optimizes between the meta oracle and the narrow experts. Our approach works sufficiently well even with the imperfect estimation of expertise due to a limited budget. We also implemented a new density metric for multi-class classification which considers the *conflictivity* of the label distribution. The result shows an improvement over the traditional density-only-weighted method, especially when the annotators are not reliable.

# Chapter 6

## Learn to Active Learn: Dynamic Active Learning with Multiple Selection Strategies

### 6.1 Introduction

Most of the recent work on active learning does not address the time-varying progress of learner knowledge state adaptively, and instead adheres to a static policy invariant with respect to new observations, resulting in non-optimal instance selection.

Figure 6.1 illustrates the motivation of this work and the efficacy of the proposed approach. It is clearly observed in this illustration that (1) no single strategy can directly predict the best actual future improvement all the time, that (2) ground-truth *trends* of the preferred mode of strategies change over time as more samples are annotated, and that (3) sometimes there is no single selection strategy that can single-handily predict future improvement, whereas there exists an ensemble of strategies that predicts better. In reality there are numerous other factors that affect a true underlying optimal strategy than what it is illustrated in this figure, such as the time-varying improvement of labeler accuracy, many of which are subject to change as new observations are made. We therefore aim to design an adaptive proactive learning framework that can address beyond determining when to explore or to exploit, and aim at finding exactly what and how much to explore or exploit, with a careful balance that can approximate the ground-truth optimal strategy. Note that this balance pattern is highly dependent on underlying distribution of a dataset as well as stream of labels obtained from annotators, and thus optimal selection strategies cannot be known *a priori*. As such, we propose to learn a new strategy as a composition of existing strategies, where optimal combination is assumed to be conditional on the labels

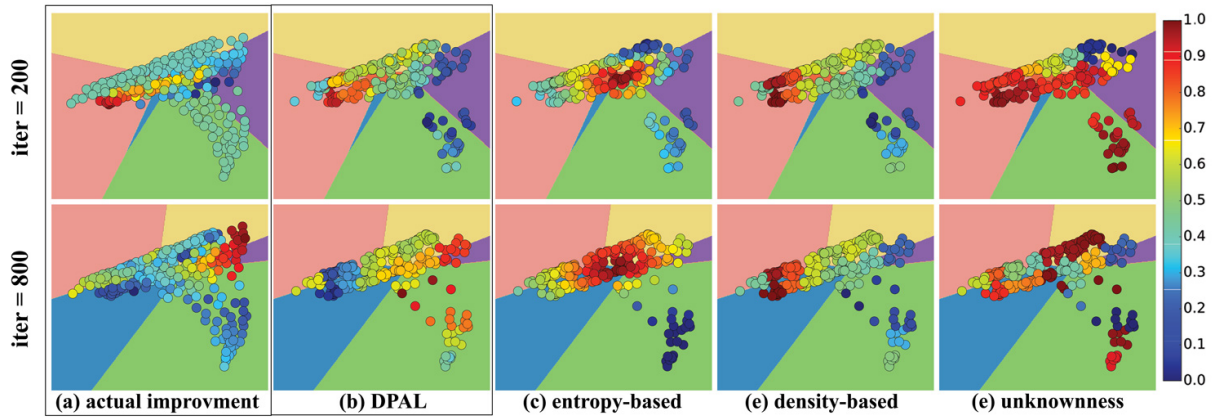


Figure 6.1: Motivation for dynamic proactive learning, illustrated with UCI Landsat Satellite Dataset [59] projected on 2-dimensional space via PCA. Each row represents different stages (iteration=200 and 800) of active learning. Drawn in the background are the 5 class-dividing hyperplanes learned thus far, and each dot represents an unlabeled sample. (a) shows actual future improvement of the learner in cross-validation performance when a ground-truth label for a corresponding sample is obtained. (c)-(e) show utility values measured by various active learning strategies. While none of the single strategies serves as a consistently reliable indicator, (b) the dynamic proactive learning framework (DPAL) predicts utility values that match closely to actual future improvement.

obtained so far. Additionally, we propose to update this strategy adaptively to a new trend as the true optimal strategy varies with progress of learner and new observations.

In this paper, we formulate the problem of finding optimal proactive learning strategy as dynamic adjustment of weights on ensemble of multiple selection strategies. In essence, we assume that utility of an unlabeled sample can be composed of weighted combination of values measured by multiple strategies. The active learning process then reduces to a utility maximization task where utility values of samples are computed based on the learned optimal strategy.

**Our contributions** are three-fold: (1) We propose a dynamic proactive learning (DPAL) framework that optimizes between multiple strategies based on the previous annotation history, extending the scope of adaptive active learning beyond the simple exploration-exploitation trade-off. In addition, we consider diverse dimensions of data utility in relation to the proactive learning scenario with multiple imperfect annotators, thus accounting for time-varying estimate of annotators expertise; (2) We formulate our approach using a structural SVM framework, which thus can accommodate any active learning criterion or loss function for measuring importance of different strategies; (3) We dynamically update our strategy given growing observations, thus being robust to time-varying trend of ground-truth optimal strategy.

## 6.2 Problem Formulation

We define a pool-based multi-class proactive learning scenario as follows. We have  $X = \{x_1, \dots, x_N\}$  and corresponding ground-truth multi-class labels  $Z = \{z_1, \dots, z_N\}$ , each of which is identified with a category  $c \in \mathcal{C}$ . We are given a pool of annotators  $K = \{k^{[1]}, \dots, k^{[M]}\}$ , who have expertise in different areas of the input space with varying degrees. We denote  $y_n^{[m]}$  as a label of  $x_n$  annotated by  $k^{[m]} \in K$ . For simplicity we do not allow duplicate label assignment of a sample by multiple annotators, thus  $y_n = y_n^{[m]}$  and  $k_n = k^{[m]}$  if  $x_n$  has been annotated by some annotator  $k^{[m]}$ , and null otherwise. In a semi-supervised setting, we assume a small subset of labeled set  $L = \{(x_n, y_n, k_n) \mid n \in I_L\}$  are known to the learner, where  $I_L \subset \{1, \dots, N\}$ , and  $|I_L| \ll N$ . The unlabeled learning pool then can be defined as  $UL = \{(x_n, y_n, k_n) \mid n \in I_{UL}\}$  for  $I_{UL} = \{1, \dots, N\} \setminus I_L$ . The proactive learning task is then to choose a subset of  $UL$  and a sequence of  $k \in \mathcal{K}$  for each instance that will best improve the learner performance, under a fixed budget constraint  $B > 0$ .

We employ a conventional greedy utility maximization approach [3, 71], where expected utility of a query is measured from a real-valued function  $U(x, k, L) : \mathcal{X} \times \mathcal{K} \times \mathcal{L} \rightarrow \mathbb{R}^+$ ,

for a sample  $x \in X$  annotated by an annotator  $k \in K$ , given a set of data labeled so far  $L \in \mathcal{L}$ . The objective then reduces to finding a pair  $(x, k)$  with the highest utility at each iteration, which when annotated, gives the best expected improvement to the learner, *e.g.*  $(x, k) = \operatorname{argmax}_{(x,k) \in X \times K} U(x, k, L)$ .

Note that the performance of an active learner thus naturally depends on how well we estimate the true utility function of a problem. Next, we describe how we define the utility function  $U$  and learn the most optimal strategy.

### 6.3 Dynamic Proactive Learning Framework

We design our DPAL system based upon the structural SVM (*e.g.* [48, 119]). We set the target utility function as our discriminant function, and assume that the discriminant function is linear in the feature vector  $\phi(x, k, L)$ , which describes the utility features measured by multiple strategies (as defined in Section Feature Space) given a sample  $x$ , annotator  $k$ , and labeled set  $L$ . Therefore, the objective of the problem at every proactive learning iteration can be formulated as:

$$(x, k) = \operatorname{argmax}_{(x,k) \in X \times K} U(x, k, L) = \operatorname{argmax}_{(x,k) \in X \times K} \mathbf{w} \cdot \phi(x, k, L)$$

Note that  $\mathbf{w}$  determines the weighted importance of each feature (strategy) to the learner, and that  $\phi$  is a function of  $L$ , which is subject to change at every iteration. We therefore learn the best weight distribution  $\mathbf{w}$  through training and *repeat* this process to dynamically adjust weights as the labeled set  $L$  expands. Algorithm 4 explains in detail how we learn and update the optimal  $\mathbf{w}$  by leveraging the past annotation history. At each active learning iteration we learn a new classifier  $f$  using the labeled set  $L$  with a support vector machine with RBF kernels [18].

#### 6.3.1 Feature Space

The feature vector  $\phi$  defines various factors or strategies that account for utility of a query, and thus the exact design is subject to user's choice. In this work, we decompose the feature vector into two main components:

$$\mathbf{w} \cdot \phi(x, k, L) = \mathbf{w}_1 \cdot \phi_1(x, k, L) + \mathbf{w}_2 \cdot \phi_2(x, k, L) \tag{6.1}$$

where the first component  $\phi_1(x, k, L)$  includes a set of features that describe the inherent value of sample  $x$  in relation to the currently available labeled set  $L$  (*e.g.* uncertainty [55], density [79,

---

**Algorithm 4** Weight Update for DPAL

---

**Input:** current  $\mathbf{w}$ , labeled data  $L$ , parameters  $Q$  (number of clusters),  $G$  (maximum number of comparing pairs), and  $|L^\theta|/|L|$  (learning history rate).

**Output:** Updated  $\mathbf{w}$

**def** updateDPAL( $\mathbf{w}$ ,  $L$ )

- Initialize  $\mathbf{w}$  randomly if  $\mathbf{w} == \text{null}$
- Choose  $L^\theta \subset L$  by sequential order of annotation, given  $|L^\theta|/|L|$
- Let  $\{L_q\}_{q=1; \dots; Q}$  be  $Q$  clusters within  $L \setminus L^\theta$
- Generate  $\min(G, \binom{Q}{2})$  pairs from  $\{(L_i, L_j)\}_{i \neq j}$
- $\mathbf{w} := \operatorname{argmin}_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{n=1}^N \xi_n$  s.t. Eq. 6.8

**return**  $\mathbf{w}$

**end def**

---

[123], etc.), whereas  $(x, k, L)$  consists of features that relate to an annotator  $k$  (e.g. probability of obtaining a correct answer from a labeler [115], etc.).

In this work, we choose to decompose the first component  $(x, L) = (x, k, L) \in \mathbb{R}^4$  into four elements, each of which measures uncertainty, density, unknownness, and conflictivity of a sample  $x$ , respectively:

$$_1(x, L) = - \sum_{c \in \mathcal{C}} P(y = c|x; L) \cdot \log P(y = c|x; L) \quad (6.2)$$

$$_2(x, L) = \rho(x|X) \quad (6.3)$$

$$_3(x, L) = -\rho(x|\{x_n|\forall n \in I_L\}) \quad (6.4)$$

$$_4(x, L) = - \sum_{c \in \mathcal{C}} P(y = c|q; L) \cdot \log P(y = c|q; L) \quad (6.5)$$

where  $_1(x, L)$  is the entropy of class posterior distribution of  $x$  given a labeled set  $L$  which measures uncertainty [93],  $_2(x, L)$  is the inherent density of samples around  $x$  in its distribution,  $_3(x, L)$  is the unknownness of a sample  $x$  given the labeled data  $L$  that penalizes local regions that have already been explored, which we estimate as inverse of observed density of labeled samples (independent of their labels), and  $_4(x, L)$  is the conflictivity of labels within its local cluster  $q \in Q$  [71]. We estimate density for Eq.6.3 and Eq.6.4 with the non-parametric Gaussian kernel method.

The intuition is that at the beginning of active learning phase  $\mathbf{w}$  will put a higher importance to  $_2(x, L)$  and  $_3(x, L)$  to encourage exploration, and gradually put more emphasis on  $_1(x, L)$

to reduce the global entropy.  $\mathcal{H}_4(x, L)$  reduces the local entropy at conflicting regions (clusters), thus fine-tuning the decision boundaries towards the end of the active learning phase.

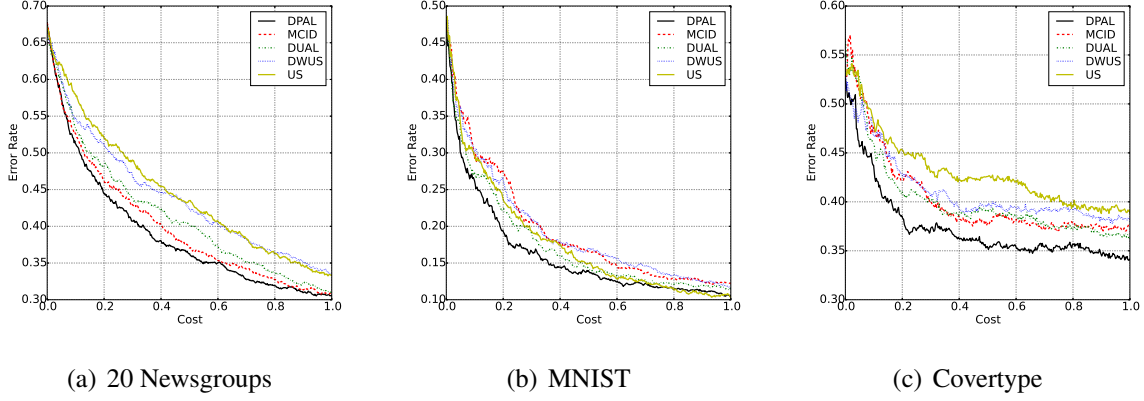


Figure 6.2: Error rates at normalized cost of queried instances annotated by noiseless labels, on (a) 20 newsgroups, (b) MNIST, (c) Covertypes datasets.

### 6.3.2 Learning

We choose to employ a single-valued metric for the second component  $\mathcal{H}_2(x, k, L)$ , which measures the expected probability of getting a correct answer given an annotator, thus  $\mathcal{H}_2(x, k, L) \in \mathbb{R}^1$ . Specifically, we define:

$$\begin{aligned} \mathcal{H}_2(x, k, L) &= P(ans|x, k) \\ &= \frac{1}{Z} \sum_{c \in \mathcal{C}} P(y = c|x; L^{[k]}) \cdot \log P(y = c|x; L^{[k]}) \end{aligned} \quad (6.6)$$

where  $L^{[k]} \subset L$  is a set of labels annotated by an annotator  $k$ ,  $Z$  is an entropy normalization constant, and  $P(y|x; L^{[k]})$  is the class-posterior probability of an annotator model trained on the annotation observations by  $k$ . Note that if an annotator model is confident of its label, its class-posterior entropy will be low, and vice versa. The probability  $P(ans|x, k)$  of getting a correct answer for  $x$  given  $k$  is thus estimated as negative entropy of class-posterior probability of its annotator model, assuming annotator confidence correlates with its accuracy. Note that this approach allows for estimation of  $P(ans|x, k)$  without ground-truth samples to compare against, and that the accuracy of this estimation generally improves over time as we observe more labels from each annotator.

In order to learn the optimal vector  $w$  in Eq.(6.1) that represents the true current weight preference, we leverage the most recently annotated data from  $L$ . Specifically, we choose a

subset  $L^\theta \subset L$  that contains the oldest samples by their sequential order of annotation, and re-evaluate how informative the most recent annotations  $(x, y, k) \in L \setminus L^\theta$  were, given a model trained on  $L^\theta$ . We typically choose  $L^\theta$  such that  $|L^\theta|/|L| = 0.8$ . The learning objective of the structural SVM [119] is then:

$$\min_{\mathbf{w}; \xi_n} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{N} \sum_{n=1}^N \xi_n \quad (6.7)$$

$$\begin{aligned} \text{s.t. } \mathbf{w} \cdot ((x_i, k_i, L^\theta) - (x_j, k_j, L^\theta)) & \quad (6.8) \\ & \geq ((x_i, k_i, L^\theta), (x_j, k_j, L^\theta)) - \xi_i, \\ & \quad \forall_{i \neq j} (x_i, k_i), (x_j, k_j) \in L \setminus L^\theta \end{aligned}$$

where  $\xi_n$  is a slack variable and  $C$  is a regularization parameter,  $(x_i, k_i)$  and  $(x_j, k_j)$  are two samples being compared drawn from the set  $L \setminus L^\theta$ . The loss function  $((x_i, k_i, L^\theta), (x_j, k_j, L^\theta))$  is defined such that it penalizes a less optimal pair of the two given  $L^\theta$  that leads to a less significant performance improvement (once it is added to  $L^\theta$ ). We propose to use the difference in total entropy decrease as a loss function:

$$\begin{aligned} & ((x_i, k_i, L^\theta), (x_j, k_j, L^\theta)) \quad (6.9) \\ & = -\frac{1}{Z} \sum_{n \in L^\theta \cup L} \left( \eta(x_n | L^\theta \cup (x_i, y_i)) - \eta(x_n | L^\theta \cup (x_j, y_j)) \right) \end{aligned}$$

where  $\eta$  is the entropy of class posterior probability of a sample given a labeled set, and  $Z$  is a normalization constant. In essence, Eq.(6.9) measures the observed relative increase in total entropy (uncertainty), thus penalizing a less optimal choice of a sample and an annotator.

In Eq.(6.8),  $(x_i, k_i), (x_j, k_j) \in L \setminus L^\theta$  can be any possible combination of two instance and annotator pairs, and thus the size of comparing pairs of  $(x, k)$  is exponential. To cope with this issue, we limit the generation of negative  $(x, k) \in L \setminus L^\theta$  to a fixed number  $G$  according to the allowed computational resource. Another challenge to the proposed loss function is that performance improvement signal is often not strong enough when only a single data point is added to a training set. Therefore, instead of comparing a pair of single samples per loss computation, we take a batch approach by first clustering samples in  $L \setminus L^\theta$  by their proximity in utility space, and then treating an average of each cluster as an input to the optimization problem in Eq.(6.7). This batch update approach not only boosts improvement signal for more robust loss function computation, but also further reduces possible combination pairs for faster optimization. We use alternating optimization that has been widely used for solving structural SVM problems (*e.g.*

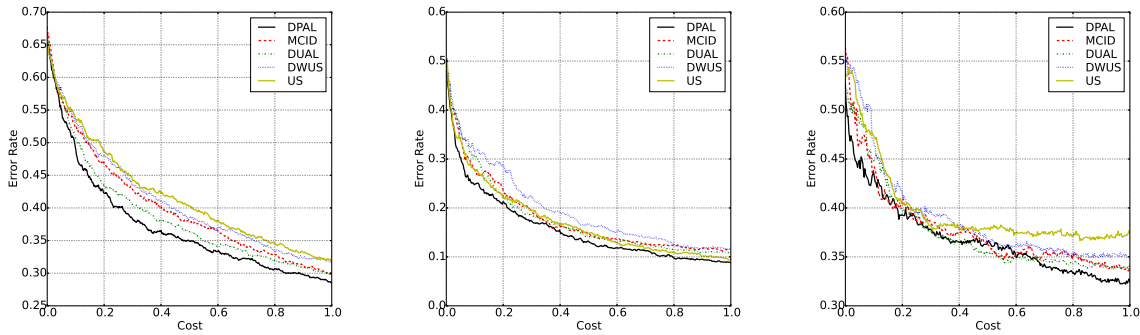
[52, 119]) at first with random initialization of  $\mathbf{w}$ , and update the optimal  $\mathbf{w}$  periodically (every 10% budget consumption). For each periodic update for  $\mathbf{w}_t$  at step  $t$  we use the previous weight vector  $\mathbf{w}_{t-1}$  as an initialization point, and add a smoothing regularization term  $\lambda \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2$  to Eq.6.7.

## 6.4 Empirical Evaluation

Below we demonstrate the efficacy of the proposed DPAL framework on several datasets ([53, 59]: Table 6.1) on a proactive learning task against several baseline methods. The results are averaged over 10-fold runs for every experiment.

### 6.4.1 Task

We evaluate the performance of each baseline in a proactive learning scenario [61, 71, 115], given a dataset and a pool of simulated annotators each with different expertise, as well as in a traditional active learning scenario given a single noiseless annotator. We start with a small percentage of labeled samples (0.5%), and at each iteration each learning algorithm chooses a pair of a sample and an annotator to expand its labeled samples pool. The goal is to reach a desired accuracy exhausting as little budget as possible. We generate multiple class-sensitive simulated experts for each dataset (as in Table 6.1) with varying noise levels, by building classifier models each of which is trained on a set of samples that has ground-truth labels for its expertise area and random labels (at a given noise ratio) for non-expertise areas.



(a) 20 Newsgroups

(b) MNIST

(c) Covertypes

Figure 6.3: Error rates at normalized cost of queried instances annotated with a pool of noised labelers (labeler noise ratio = 0.3 for non-expertise classes, and 0 for expertise classes.), on (a) 20 newsgroups, (b) MNIST, (c) Covertypes datasets.

Table 6.1: Overview of datasets.

Dataset	# Experts	# Classes	Size
20 Newsgroups	5	20	18846
MNIST	5	10	70000
Covertypes	5	10	58000

Table 6.2: Normalized proactive learning costs at error rate convergence for each dataset with varying DPAL  $|L^0|/|L|$  ratios. Bold denotes the best performance for each test, and \* denotes the statistically significant improvement ( $p < 0.05$ ).

Dataset	$ L^0 / L $			
	0.6	0.7	0.8	0.9
20 Newsgroups	1.13	1.07	<b>1.00*</b>	1.16
MNIST	1.08	<b>0.97</b>	1.00	0.99
Covertypes	1.04	1.10	<b>1.00</b>	1.21

## 6.4.2 Baselines

We compare our DPAL method against the following methods from literature: US (uncertainty sampling; [93] for the single annotator scenario, and [115] for the multiple noised annotators scenario), DWUS (density weighted uncertainty sampling; [123]), DUAL (exploration-exploitation switch; [26]), and MCI D (multi-class information density; [71]). The difference between DWUS and DUAL is that while DWUS measures utility of a sample as multiplicative composition of its entropy and density, DUAL alternates between uncertainty-based and density-based sampling strategies around an optimal switching point. MCI D is one of the state-of-the-art sampling strategies that combines DWUS with unknownness and conflictivity as a multiplicative ensemble to better handle multi-classification active learning problems. Note that most of the referenced papers for the baselines above do not address optimal selection of an annotator and a sample given estimated annotator accuracies. Therefore, we apply and use as a baseline the approach proposed by [115], which is to first select a sample based on its distribution-dependent strategy and then to delegate an annotator with the highest estimated probability of giving the correct answer for the chosen sample. Unlike the proactive annotator selection scheme proposed by [115], our DPAL framework directly integrates estimated annotator accuracy with base utility of a sample in a jointly optimal way. To separate out the effect of DPAL’s joint selection of annotator-sample

pairs, we run our experiments with varying noise levels of annotator expertise, including a noise-free case (a perfect oracle annotator).

### 6.4.3 Results

**Main results:** Figure 6.2 shows the active learning curve at varying amount of annotation cost on different datasets, when labels were annotated by a single noiseless oracle. This is a conventional active learning scenario, and thus we do not consider optimal selection of annotators in this experiment. For most of the datasets, it can be seen that our DPAL method reaches the accuracy at near convergence faster than the baselines, significantly saving the budget it requires to reach the same accuracy. The baselines which heuristically aims to leverage exploration-exploitation balance (MCI D, DUAL, DWUS) tend to outperform the US baseline which solely aims to exploit the uncertain samples, however the performance boost is not as strong as with DPAL. Among the baselines aside from DPAL, there is no baseline that consistently wins across the datasets evaluated in this experiment. Note for example that MCI D does not perform well on some of the datasets in this case, because the noiseless labels tend to neutralize the efficacy of the conflictivity term in MCI D, which favors to dissolve region around locally heterogeneous labels. DPAL learns to suppress non-optimal strategies and balance preferable strategies, leading to better performance overall.

Figure 6.3, on the other hand, assumes a proactive learning scenario, where annotators are simulated with a class-sensitive noise ratio according to their expertise. A proactive learner thus tries to assign the most knowledgeable annotator for a chosen sample, although the accuracy of expertise estimation is not perfect at the beginning, and tends to improve over time. Because annotated samples include noised labels due to non-optimal expert assignment, it can be seen that baseline performance is different from Figure 6.2. This result indicates that an optimal strategy for a proactive learning problem is highly dependent on the labelling accuracy of annotators as well as the inherent distribution of each dataset. DPAL learns its optimal strategy from the past annotation history, and thus in general outperforms other baselines on most of the datasets more consistently. This result is consistent with the result in Figure 6.2, which indicates that DPAL is flexible to incorporate any number of sampling strategies and optimize for the best ensemble weights given the pool of multiple selection strategies. Note that in practice one can choose to halt proactive active learning process at any desired accuracy, as DPAL tends to be more effective towards the beginning of annotation compared to other baselines.

**Dynamic weight transitions:** Figure 6.4 shows the dynamic weight transitions for multiple

selection strategies (a,c,e: a single noiseless labeler scenario (same setting as Figure 6.2), b,d,f: multiple domain experts scenario (same setting as Figure 6.3)), normalized to sum to 1 (U: uncertainty, D: density, K: unknownness, C: conflictivity, E: estimated annotator expertise). Note that the different components of the weight vector outweigh others at different active learning stages, confirming the observation that there is no single optimal active learning strategy consistently dominant (Figure 6.1). Note also that the optimal weights are different across the datasets, which shows the need for dynamic weight adjustment learned with DPAL rather than with a heuristic approach. It can be seen that the strategies that encourage exploration (density, unknownness) tend to be given higher weights at earlier stages, while the strategies that encourage exploitation (uncertainty, conflictivity) tend to be more dominant towards the end of the active learning process, which intuitively is a desirable strategy. The weight for estimated expertise of annotators is suppressed at the beginning when expertise estimation is unreliable due to the small number of labeled samples by each annotator to examine with. Later in the active learning process the expertise estimation weight is assigned higher weight, which leads to a more optimal selection of annotators and samples. These weight adjustment behaviors can explain the efficacy of the DPAL approach as demonstrated in Figures 6.2 and 6.3.

**Sensitivity to DPAL hyperparameters:** Table 6.2 shows the DPAL performance at varying  $|L^0|/|L|$  ratios ( $= 0.6, 0.7, 0.8, 0.9$ ), averaged over 10-fold cross validation runs. Each row represents the normalized proactive learning cost to reach convergence in error rate for each dataset, showing how each configuration saves learning budget compared to others. Intuitively, when  $|L^0|/|L|$  is lower, DPAL evaluates the contribution of each sample conditioned on  $L^0$  further back in the annotation history, and thus the learned weights  $w$  might not be optimal for current evaluation. When  $|L^0|/|L|$  is higher, there is less annotation history to leverage ( $L \setminus L^0$ ) for learning the optimal strategy, thus being more prone to over-fitting. We do not observe statistically significant improvement for any particular ratio consistent across all of the datasets, and thus for all of our experiments we simply choose  $|L^0|/|L| = 0.8$  which yields the best average value on the three datasets.

## 6.5 Related Work

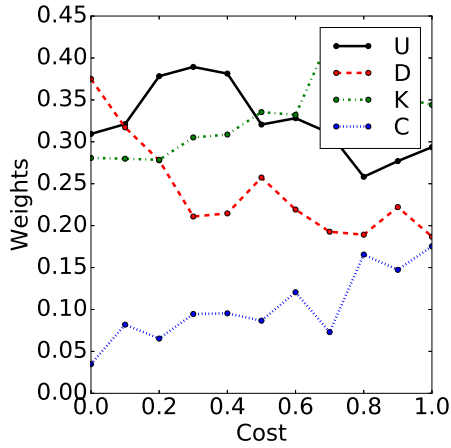
**Exploration-exploitation balance in active learning** is well studied and closely related to our work [7, 11, 63, 66, 81, 118]. The main idea behind these work is that a learning model can quickly improve its performance by first *exploring* diverse areas in its data distribution, and then

by “fine-tuning” its hypothesis hyper-plane via *exploitation*. For example, [26] proposes the dual strategy for active learning (DUAL) which switches between the two alternating policies (uncertainty and density) pivoting around a pre-defined threshold value. Our approach extends their work in terms of scalability and flexibility because we avoid use of heuristics, and instead optimize strategy based on stream of observed labels. [63] obtains a similar exploration-exploitation balance by extending the conventional query-by-committee (QBC) methods [34, 64, 94] under a stream-based active learning setting. More recently, several studies have investigated active learning in a multi-armed bandit framework [38, 92], which allow for exploration in hypothesis space by calculating lower confidence bounds on the risk of pulling each hypothesis. However, most of these work do not address other diverse dimensions that can lead to better active improvement, such as time-varying expertise level of annotators or learner’s estimation of annotator expertise given a growing number of observations. Our approach takes into account multiple strategies of user’s choice, extending the previous work on exploration-exploitation balance by delving more deeply and precisely into the question of exactly where in data to explore, how much to exploit, and *by asking whom*.

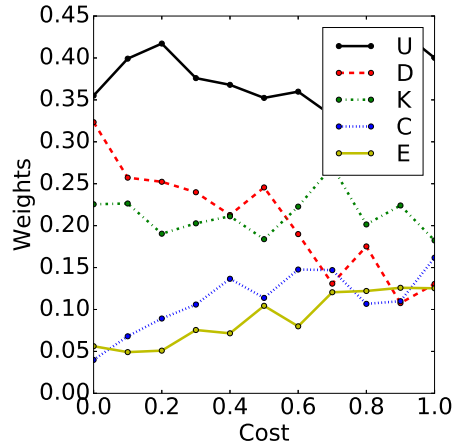
## 6.6 Summary

We proposed a new approach that dynamically adjusts the optimal ensemble proactive learning strategy based on the past annotation history. While conventional active learning approaches aim to optimize only for the selection of samples estimated given a static strategy, we optimize for the near-optimal selection or ensemble of multiple strategies adaptive to the time-varying progress of the active learner. In order to achieve this, we designed alternating optimization over SVM problems where an optimal weight vector combining multiple strategies can be learned from past annotation history. We demonstrated that the proposed approach outperforms or matches the performance of other baselines over several datasets by dynamically adjusting weights according to desirable behaviors at each phase.

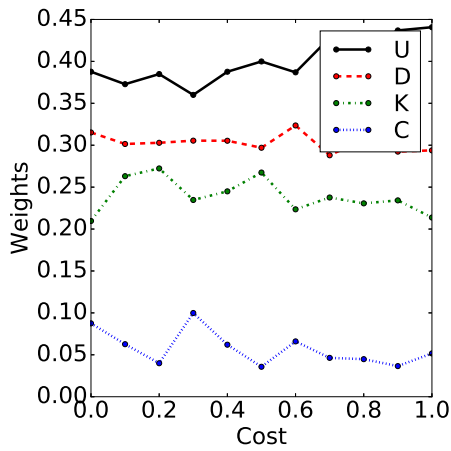
We note that while the proposed approach learns the optimal strategy for the immediately-next sampling, the myopic approximation does not always guarantee a global optimum. Future work will explore different strategies or reward functions that favor more globally-optimal strategies in order to further improve the performance.



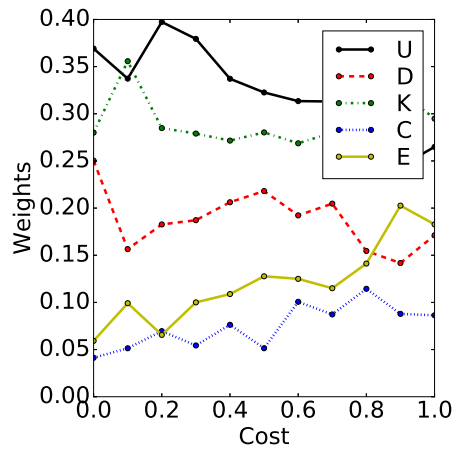
(a) 20 Newsgroups



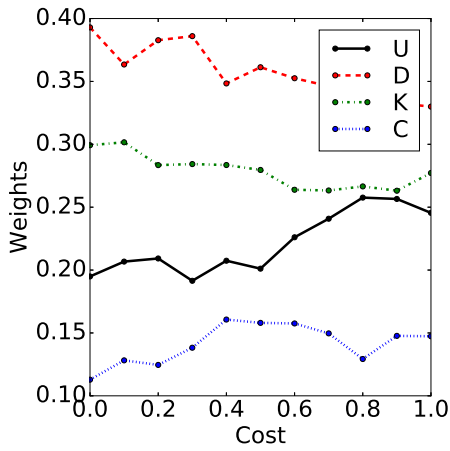
(b) 20 Newsgroups (w/E)



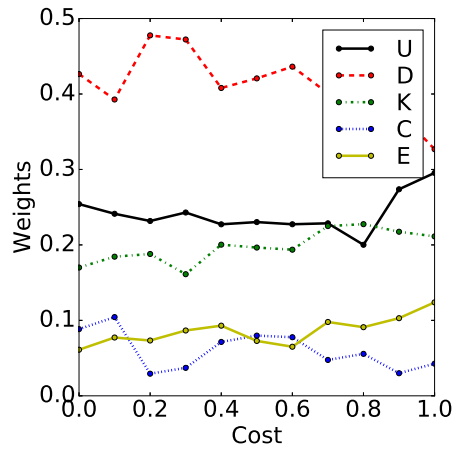
(c) MNIST



(d) MNIST (w/E)



(e) Coverttype



(f) Coverttype (w/E)

Figure 6.4: Normalized DPAL weight transitions for (a), (c), (e): a single noiseless labeler scenario and (b), (d), (f): multiple noised labelers scenario.



# Chapter 7

## Learn to Active Transfer: Dynamic Proactive Transfer Learning

### 7.1 Introduction

Given a low-resource target task, a learning agent typically considers two different approaches: transfer learning and active learning. While these two approaches are complementary in nature, very limited work have studied the combination of the two approaches. Examples include [15, 32, 39, 49, 95, 124], which aim to combine transfer learning with the active learning framework by conditioning transferred knowledge as priors for optimized selection of target instances, overcoming the common *cold-start* problem at the beginning phase of active learning with zero-shot class-relation priors.

In our work, we take a very different approach of combining active learning and transfer learning approaches given the following observation. We observe that the quality of transfer essentially depends on how well we can uncover the *bridge* in the projected space where the two datasets are semantically linked. Intuitively, if the two datasets describe completely different concepts, very little information can be transferred from one to the other. We therefore propose proactive transfer learning approaches which expands the labeled target data to actively *mine* transferable knowledge and to progressively improve the transfer accuracy. Note this objective is conceptually different from conventional active learning strategies, which evaluates utilities within the target domain only, without considering their relation knowledge to source domain.

Therefore, we provide the **dynamic proactive transfer learning** combining the transfer learning and active learning in a unique way, which leverages the heterogeneous source knowl-

edge given the available source and target data, and also actively mines the unknown knowledge in target domain with two objectives in mind: (1) conventional active learning objectives, which aim at reducing the overall target task uncertainty by querying information *within* the target domain, and (2) the proposed proactive transfer learning approaches which aim at improving the transfer accuracy between source and target, thereby ultimately improving the target task performance. We also study a time-dependent optimality for each transfer task, and build a separate active learning network to find such balance with the dynamic proactive learning (DPAL) framework (Chapter 6).

## 7.2 Methods

### 7.2.1 Proactive Transfer Learning Strategies

The quality of the learned parameters for the target task depends on the available labeled target training samples ( $L_T$ ). As such, we propose to expand  $L_T$  by querying a near-optimal subset of the unlabeled pool  $UL_T$ , which once labeled will improve the performance of the transfer accuracy and ultimately the target task, assuming the availability of unlabeled data and (limited) annotators. In particular, we relax this problem with a greedy pool-based active learning framework, where we iteratively select a small subset of unlabeled samples that maximizes the expected utility to the target model:

$$\hat{\mathbf{x}}_T = \underset{\mathbf{x}_T \in \mathcal{U}_{UL_T}}{\operatorname{argmax}} U(\mathbf{x}_T) \quad (7.1)$$

where  $U(\mathbf{x}_T)$  is a utility function that measures the value of a sample  $\mathbf{x}_T$  defined by a choice of the query sampling objective. In traditional active learning, the uncertainty-based sampling [63, 93] and the density-weighted sampling strategies [79, 123] are often used for the utility function  $U(\mathbf{x}_T)$  in the target domain only. However, the previous approaches in active learning disregard the knowledge that we have in the source domain, thus being prone to query samples of which the information can be potentially redundant to the transferable knowledge. In addition, these approaches only aim at improving the target classification performance, whereas querying *bridge* instances to maximally improve the transfer accuracy instead can be more effective by allowing more information to be transferred in bulk from the source domain. Therefore, we propose the following two proactive transfer learning objectives for sampling in the target domain that utilize the source knowledge in various ways:

**Maximal Marginal Distribution Overlap (MD):** We hypothesize that the overlapping projected region is where the heterogeneous source and target data are semantically related, thus a good candidate for a *bridge* that maximizes the information transferable from the source data. We therefore propose to select unlabeled target samples ( $\mathbf{x}_T$ ) in regions where the marginal distributions of projected source and target samples have the highest overlap:

$$U_{\text{MD}}(\mathbf{x}_T) = \min \left( \hat{P}_T(\mathbf{x}_T | \mathbf{W}_h, \mathbf{W}_f), \hat{P}_S(\mathbf{x}_T | \mathbf{W}_g, \mathbf{W}_f) \right) \quad (7.2)$$

where  $\hat{P}_T$  and  $\hat{P}_S$  are the estimated marginal probability of the projected target and source instances, respectively. Specifically, we estimate each density with the non-parametric kernel method:

$$\begin{aligned} \hat{P}_T(\mathbf{x}_T | \mathbf{W}_h, \mathbf{W}_f) &= \frac{1}{N_T} \sum_{i=1}^{N_T} K_h(\mathbf{f}(\mathbf{h}(\mathbf{x}_T)) - \mathbf{f}(\mathbf{h}(\mathbf{x}_T^{(i)}))) \\ \hat{P}_S(\mathbf{x}_T | \mathbf{W}_g, \mathbf{W}_f) &= \frac{1}{N_S} \sum_{j=1}^{N_S} K_h(\mathbf{f}(\mathbf{g}(\mathbf{x}_T)) - \mathbf{f}(\mathbf{g}(\mathbf{x}_S^{(j)}))) \end{aligned} \quad (7.3)$$

where  $K_h$  is a scaled Gaussian kernel with a smoothing bandwidth  $h$ . In essence, solving  $\max_{\mathbf{x}_T} \min(\hat{P}_T(\mathbf{x}_T), \hat{P}_S(\mathbf{x}_T))$  finds such instance  $\mathbf{x}_T$  whose projection lies in the highest density overlap between source and target instances.

**Maximum Projection Entropy (PE)** aims at selecting an unlabeled target sample that has the maximum entropy of dot product similarities between a *projected* instance and its possible label embeddings:

$$U_{\text{PE}}(\mathbf{x}_T) = - \sum_{\mathbf{z} \in Z_T} \log(\mathbf{f}(\mathbf{h}(\mathbf{x}_T)) \cdot \mathbf{y}_z) \cdot \mathbf{f}(\mathbf{h}(\mathbf{x}_T)) \cdot \mathbf{y}_z \quad (7.4)$$

The projection entropy utilizes the information transferred from the source domain (via  $\mathbf{W}_f$ ), thus avoiding information redundancy between source and target. After samples are queried via the maximum projection entropy method and added to the labeled target data pool, we re-train the weights such that projections of the target samples have less uncertainty in label assignment.

To reduce the active learning training time at each iteration, we query a small fixed number of samples ( $= Q$ ) that have the highest utilities. Once the samples are annotated, we re-train the model with Eq.2.2, and select the next batch of samples to query with Eq.7.1. The overall process is summarized in Algorithm 5.

---

**Algorithm 5** Proactive Transfer Learning

---

**Input:** source data  $\mathbf{S}$ , target data  $\mathbf{T}$ , active learning policy  $U(\cdot)$ , budget  $B$ , query size per iteration  $Q$

Randomly initialize  $\mathbf{a}$ ,  $\mathbf{W}$  (truncated normal)

**for**  $iter = 1$  **to**  $B$  **do**

1. Learn  $\mathbf{a}$ ,  $\mathbf{W}$  by solving

$$\min_{\mathbf{a}, \mathbf{W}} \mu \sum_{k=1}^K \alpha_k \cdot \mathcal{L}_{\text{HR:K}}(\mathbf{S}_k, \mathbf{W}_g, \mathbf{W}_f) + \mathcal{L}_{\text{HR}}(\mathbf{T}, \mathbf{W}_h, \mathbf{W}_f) + \mathcal{L}_{\text{AE}}(\mathbf{S}, \mathbf{T}, \mathbf{W}) + \mathcal{R}(\mathbf{W})$$

2. Query  $Q$  new samples

**for**  $q = 1$  **to**  $Q$  **do**

$$\hat{\mathbf{i}} = \operatorname{argmax}_{\mathbf{i} \in UL_T} U(\mathbf{x}_T^{(i)})$$

$$UL_T := UL_T \setminus \{\hat{\mathbf{i}}\}, L_T := L_T \cup \{\hat{\mathbf{i}}\}$$

**end for**

Update DPAL learning policy  $U(\cdot)$  (Algorithm 4) if conditions are met.

**end for**

**Output:**  $\mathbf{a}$ ,  $\mathbf{W}$

---

### 7.2.2 Integration with DPAL

We integrate the proposed proactive transfer learning framework with DPAL, with the objective of actively improving transfer accuracy as well as enhancing the target task performance. In order to successfully assess trade-off between transfer learning approaches and traditional active learning approaches, we consider several configurations of the DPAL algorithms (Figure 7.1) that allow for efficient learning of optimized strategy creation. Essentially, we frame the DPAL problem as ‘learning to active-transfer’ with historical active learning selection samples as input data, and explore various architectures that can map input to ground-truth performance improvement (as output). We consider the following architectures for ablation study:

- **Linear-DPAL:** takes as input a vector of utility values of selection strategies, and learns an optimal linear weight vector, element-wise composition of which predicts the ground-truth weighted utility (as defined in Section 6.3).
- **DNN-DPAL:** takes as input a vector of utility values of selection strategies as well, but replaces the linear weights with a deep neural network which allows for more complex composition of strategies. As a downside, it loses the interpretability with regards to weight

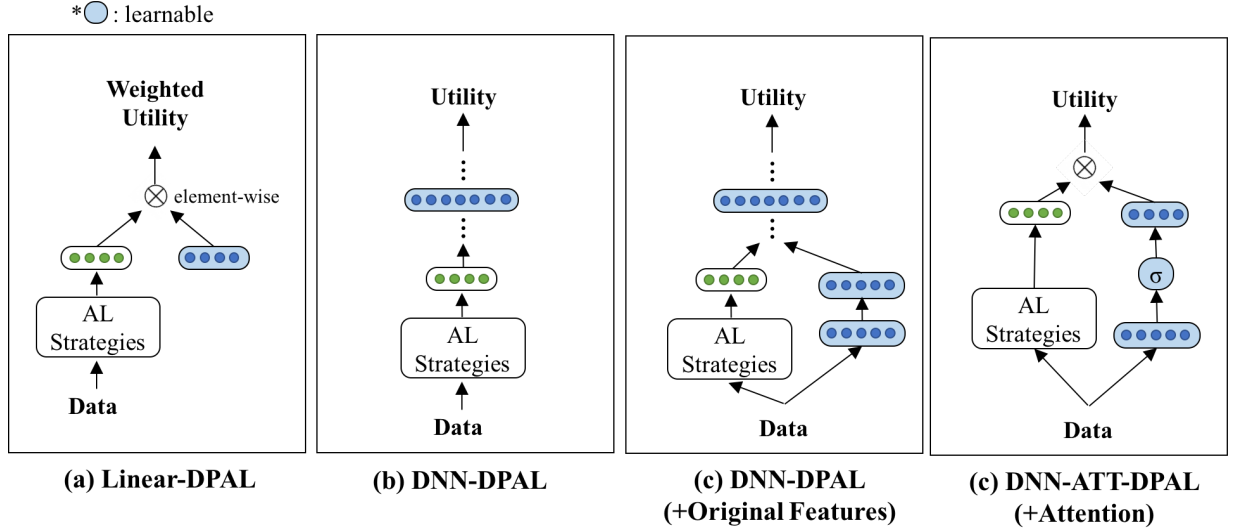


Figure 7.1: Ablation studies of DPAL network architecture. (a) Utility is measured for each active learning strategy, and an optimal linear weight is learned for element-wise composition of weighted utility. (b) A deep neural network replaces the linear weights, allowing for more complex composition of strategies (with less interpretability). (c) Encoded data features are appended as input to the DPAL network. (d) We employ an attention module to determine importance weight of each active learning strategy.

importance of each selection strategy.

- **DNN-DPAL+Original Features:** takes as input original features as well appended to utility values, and builds a deep network.
- **DNN-ATT-DPAL:** takes as input original features to produce attention vectors (via sigmoid operations) of which element-wise composition with utility vectors produce the predicted weighted utility.

We learn the optimal parameters for each DPAL algorithm with the proactive transfer learning strategies proposed in Section 7.2.1, as well as traditional active learning strategies such as entropy and density-based methods.

### 7.3 Empirical Evaluation

We consider a proactive transfer learning scenario, where we expand the labeled target set by querying an oracle given a fixed budget. We compare the proposed proactive transfer learning

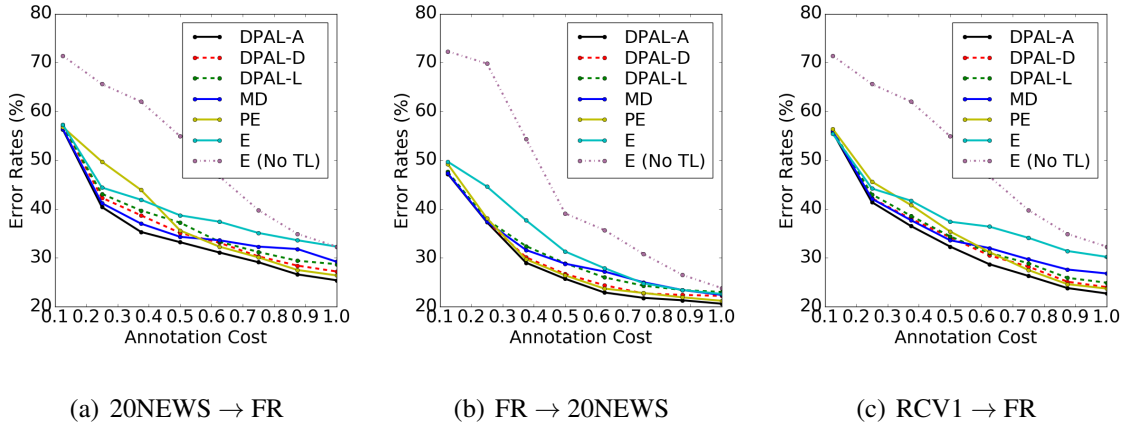


Figure 7.2: Proactive transfer learning results for various source and target dataset pairs.  $X$ -axis: % of queried samples,  $Y$ -axis: classification error rates. DPAL approaches combine multiple strategies to compose an optimal strategy.

strategies (Section 7.2.1) as well as the DPAL-integrated strategies (Section 7.2.2) against the conventional uncertainty-based sampling methods.

**Setup:** We choose various source-target dataset pairs to study: (a) 20NEWS $\rightarrow$ FR, (b) FR $\rightarrow$ 20NEWS, and (c) RCV1 $\rightarrow$ FR. The lines **MD** (maximal marginal distribution overlap; solid blue) and **PE** (maximum projection entropy; solid yellow) refer to the proposed proactive learning strategies in Section 7.2.1, respectively, where the weights are learned with the CHTL network. The baseline active learning strategies **E** (entropy; solid cyan) and **E (No CHTL)** (entropy; dotted purple) select target samples that have the maximum class-posterior entropy given the original target input features only, which quantifies the uncertainty of samples in multiclass classification. **DPAL-A** (DPL with the attention module; solid black), **DPAL-D** (DNN-DPAL; dashed red), **DPAL-L** (Linear DPAL; dashed green) refer to the proposed dynamic proactive learning approaches that learn the optimal strategy combining the above. The uncertainty-based sampling strategies are widely used in conventional active learning [63, 93], however these strategies do not utilize any information from the source domain. Once the samples are queried, we re-train the CHTL classifier (except E (No TL); it refers to the baseline that does not employ transferred knowledge at all, and learns a softmax classifier instead). Each experiment is conducted over 10 folds.

**Main results:** Figure 7.2 shows the target task performance improvement over iterations with various active learning strategies. We observe that both of the proposed active learning strategies (**MD**, **PE**) outperform the baselines on all of the source-target dataset pairs. Specifically, **PE** outperforms **E** on most of the cases, which demonstrates that reducing entropy in the projected space

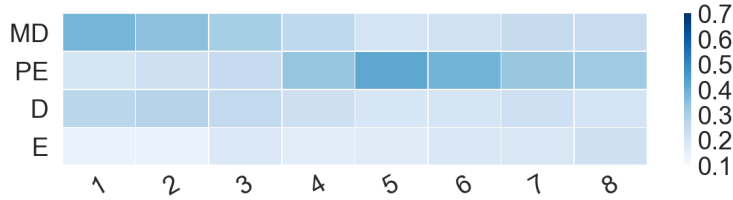


Figure 7.3: **Visualization of DPAL selection strategy attention** on an example case of RCV1  $\rightarrow$  FR datasets pair. For each DPAL update step (column), the strategy attention module amplifies the most informative selection strategy (darker) while attenuating less important or noisy selection strategies (lighter). The model makes final predictions of the weighted utility based on weighted signals from all selection strategies. Strategies used - MD: Maximal marginal distribution overlap, PE: Projected embeddings entropy, D: target sample density, E: target sample class-conditional entropy.

is significantly more effective than reducing class-posterior entropy given the original features. Because we re-train the joint network after each query batch, avoiding information redundancy between source and target while reducing target entropy is critical. Note that **MD** outperforms **PE** generally at the beginning, while the performance of **PE** improves faster as it gets more samples annotated. This result indicates that selecting samples with the maximal source and target density overlap (**MD**) helps in building a *bridge* for transfer of knowledge initially, while this information may eventually get redundant, thus the decreased efficacy. **DPAL-ATT**, which combines multiple selection strategies to compose the optimal strategy, outperforms all other baselines at all phases, showing the efficacy of the proposed approach. Other DPAL approaches show comparable results that also outperform other static strategies, but the maximum performance is achieved with the attention model, demonstrating the need for a model with a bigger set of parameters. Note also that all of the projection-based methods (**MD**, **PE**, **E**) significantly outperform **E (No TL)**, which measures the entropy and learns the classifier at the original feature space. This result demonstrates that the learned projections effectively encode input target features, from which we can build a robust classifier efficiently even with a small number of labeled instances.

**Visualization of the selection strategy attention:** Figure 7.3 visualizes the DPAL selection strategy attention module at each DPAL weight update step (each column), where amplified selection strategy is represented with darker color, and attenuated strategy is represented with lighter color. It can be seen that in the successful DPAL case of RCV1  $\rightarrow$  FR datasets pair,

we confirm that the strategy attention module successfully attenuates the density-related metrics (MD: maximal marginal distribution overlap between source and target, and D: target sample density) towards the beginning phase of active learning, allowing for batch update of knowledge on a large number of samples. Towards the end, PE (Maximal projected embeddings entropy) strategy starts to dominate, which allows for efficient fine-tuning of challenging class boundaries, leading to better classification results.

## 7.4 Related Work

**Active learning** provides an alternative solution to the label scarcity problem, which aims at reducing sample complexity by iteratively querying the most informative samples with the highest utility given the labeled sampled thus far [71, 93].

**Transfer active learning** approaches [15, 32, 49, 95, 124] aim to combine transfer learning with the active learning framework by conditioning transferred knowledge as priors for optimized selection of target instances. Specifically, [39] overcomes the common *cold-start* problem at the beginning phase of active learning with zero-shot class-relation priors. However, many of the previously proposed transfer active learning methods do not apply to our setting because they require source and target data to be in either homogeneous feature space or the same label space or both. Therefore, we propose a *proactive transfer learning* approach for heterogeneous source and target datasets, where the objective is to progressively find and query *bridge* instances that allow for more accurate transfer, given a sampling budget.

Our proposed dynamic proactive transfer learning strategies improve upon the conventional transfer active learning strategies in the following ways: (1) We explicitly measure the expected utility of a target sample in its expected improvement of transfer accuracy, which potentially better leverages the vast source knowledge, and (2) we provide a dynamic active learning framework which re-evaluates the optimal strategy adaptive to the learning progress, without fixating the strategy nor using the transferred knowledge only at the beginning.

## 7.5 Summary

Given a low-resource target task, two different approaches (transfer learning and active learning) are conventionally employed separately. In our study, we provide a framework called dynamic proactive transfer learning combining these two approaches, which leverages the heterogeneous

source knowledge given the available source and target data, and also actively mines the unknown knowledge in target domain with two objectives in mind: (1) conventional active learning objectives, which aim at reducing the overall target task uncertainty by querying information *within* the target domain, and (2) the proposed proactive transfer learning approaches which aim at improving the transfer accuracy between source and target, thereby ultimately improving the target task performance. These two objectives are synergistic to each other, and hence we also study a time-dependent optimality for each transfer task, and build a separate active learning network to find such balance with the dynamic proactive learning (DPAL) framework. Experimental results show the efficacy of the proposed framework.



# Chapter 8

## Conclusions

In this thesis, we proposed the **Proactive Transfer Learning** (PTL) framework which addresses a low-resource domain learning scenario (scarce labeled data). The proactive transfer learning combines the two main branches of literature in machine learning that address a low resource scenario: transfer learning and active learning. The key contributions are as follows.

First, we presented **Completely Heterogeneous Transfer Learning** (CHTL) methods where source and target datasets are heterogeneous in both feature and label spaces, which extends the reach of conventional transfer learning algorithms to truly heterogeneous settings. Unlike traditional transfer learning approaches, we do not require explicit relations between source and target tasks, such as source-target correspondent parallel dataset or label space homogeneity. For successful heterogeneous knowledge transfer, we presented a novel CHTL architecture that leverages auxiliary label embeddings from a language model or a knowledge graph, as well as a modality-level attention module that avoids negative transfer. In order to compensate for low target resources, we also presented a unsupervised transfer learning method (Transferrable Adversarial Network) which builds a unified set of a generator and a discriminator for all sources and target domains, to share a pathway of network for label prediction. The main CHTL network is trained via min-max training between the generator and the discriminator in addition to supervised training, producing a more robust classifier for label embeddings space. The efficacy of the proposed CHTL network has been demonstrated on several simulation studies and empirical analysis, such as hetero-lingual text classification or image-aided multimodal named entity disambiguation tasks for social media posts, etc.

Second, we presented the **Dynamic Proactive Learning** (DPAL), which extends the practical reach of conventional active learning algorithms, which aims at acquiring optimized subset of

labels given a fixed budget (*e.g.* asking minimal number of questions to oracles while maximizing information gain). The proposed DPAL algorithm studies optimality of sample complexity with regards to (a) multiple annotators with varying expertise and (b) multiple heuristic selection strategies. To accommodate for existence of multiple annotators, we formulate an iteration-wise greedy utility optimization problem where utility is defined for each annotator and sample pair as expected information gain attenuated by its unique cost and estimated expertise. In order to optimize across multiple available selection strategies, we design an adaptive policy that aims at finding exactly what and how much to explore or exploit with a careful balance that can approximate the ground-truth optimal strategy. In our "learn to active-learn" formulation, the ground-truth optimal strategy is estimated from previous active learning history data, where a new strategy is composed as a weighted ensemble of multiple existing strategies. In our simulated active learning experiments on real-world dataset demonstrate that the proposed DPAL algorithm composes an optimal strategy from multiple existing strategies from previous literature, outperforming the state-of-the-art algorithms that adhere to single static heuristic strategy.

Lastly, we presented the dynamic proactive transfer learning (PTL) framework which combines these two options of addressing a low resource domain learning - active learning and transfer learning, which determines *when* to transfer knowledge or query annotations from unlabeled target set. The proposed PTL framework progressively builds *bridges* between target and source domains in order to improve transfer accuracy (hence improving target task accuracy), and (2) exploits or explores target domains to improve target task accuracy as well, where the strategy is balanced and optimized adaptive to dataset pairs and active learning phases. We provide a unique active learning setting where a new target model is built for each iteration with transferred knowledge, which then evaluates each unsupervised sample in target domain by taking into account its relations to available source label samples.

The proposed framework is evaluated across various tasks with low resource assumptions, such as hetero-lingual text classification tasks (leveraging document classification datasets with heterogeneous languages and label space), text-aided image scene classification, multi-modal named entity disambiguation tasks, etc. Our empirical results show that the proposed proactive transfer learning framework can produce a unified representation where adequate subset of heterogeneous source knowledge can be drawn in inferring of a novel task, and optimally decide to acquire more information from oracles to improve transfer accuracy or target performance, or both.

We believe that the proposed approach of combining heterogeneous transfer learning and

active learning leads to several promising future work. Specifically, we identify the following areas of research as potential future directions of the proposed framework.

**Multi-source Proactive Transfer Learning:** While we have investigated multi-source learning for transfer learning (CHTL) approaches, extending the iterative proactive transfer learning for multiple sources scenario remains as an open challenge. Multi-source PTL would have to identify utility of each target sample in relation to its intrinsic value in its marginal distribution for target-domain only (traditional active learning approach), as well as its value for learning relation between each heterogeneous source and target pair, or compositional relations among all sources. This approach is promising in that it accounts for all of the available source knowledge in active learning schemes, however it inevitably complicates the formulation of active learning algorithms, and finding the optimal strategy would become even more challenging. Specifically, we have proposed two new active learning strategies which aim at evaluating a target sample for its estimated transfer accuracy improvement: maximal marginal distribution overlap and maximal label embeddings projection entropy, which, along with the DPAL optimal strategy composition, improves upon the conventional active learning strategies that focus only on target domains. Extending the current proactive transfer learning approaches to multi-source scenario would thus require a new definition of pairwise marginal distribution overlap.

**Unsupervised Transfer Learning:** It is known that unsupervised representation learning greatly helps supervised tasks especially when label resource is scarce. In our work, we have proposed a novel transferrable adversarial network (TAN) which builds a unified generator and a discriminator for all sources and target, the joint training of which builds a more robust representation unifying all sources and target in label embeddings space. While the training objective of TAN is to differentiate among genuine label embeddings representation and generated ones, the same framework can be extended for various other objectives. An example of such objectives is using the TAN structure for differentiating genuine and generated samples for their latent modality matches (testing whether or not two samples come from the same latent modality), which multiplies the unsupervised training pairs. We plan to investigate several alternative objectives to train TAN architecture to aid shared representation learning.

**Beyond classification tasks:** The core of transfer learning approaches is in unified knowledge representation learning. While we have studied and analyzed the applicability of the proposed transfer learning architecture mainly on classification tasks (via embeddings projection to label embeddings space), the learned knowledge representation can also be applied to other tasks, including multi-lingual translation tasks, image caption generation tasks, etc. The extension of

CHTL network in generation tasks would need additional generation layer on the same level as classification layer, which takes as input intermediate shared knowledge representation, formulated as a multi-task learning problem. We plan to investigate the applicability of the learned transfer representation in various tasks.

**Heterogeneous domains applications:** While the empirical evaluation of CHTL was conducted primarily on the text and image domains, our formulation does not restrict the input domain to be of a certain modality, as evident in the domain-free simulation results. We thus believe the approach can be applied broadly, and as future work, we plan to investigate the transferability of knowledge with diverse heterogeneous settings, given suitable source and target data. Examples of such applications include multi-domain medical diagnosis and prescription prediction tasks, etc.

To summarize, the proposed proactive transfer learning framework addresses a unique solution for low-resource task learning with the following key contributions. The proposed framework exhibits an ability to acquire knowledge from already-abundant related source of information, albeit heterogeneous in nature to target tasks, without requiring a heavily curated, task-specific dataset. This contribution significant in that it demonstrates that the proposed algorithms greatly enhance the practical reach of the existing transfer learning algorithms in various settings. In addition, when there does not exist enough relevant information readily available, the proposed framework is able to actively query novel knowledge from existing oracles in an optimized manner to avoid inefficient or redundant information acquisition. By taking into account available source knowledge in the optimization of iterative query selection with regards to its expected transfer learning improvement, the framework allows for faster convergence of learning performance. We believe the proposed framework is an important step towards building an intelligent and independent learning system, addressing a crucial aspect of it in coping with diverse novel tasks with varying availability of resource.

# Bibliography

- [1] V. Ambati, S. Vogel, and J. G. Carbonell. Active learning and crowdsourcing for machine translation. *LREC 10*, 2010. 5.1
- [2] Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *NIPS*, pages 28–36, 2009. 2.4.3
- [3] Mauricio Araya-López, Olivier Buffet, Vincent Thomas, and François Charpillet. Active learning of mdp models. In *Recent Advances in Reinforcement Learning*, pages 42–53. Springer, 2011. 6.2
- [4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007. 2.3.3
- [5] Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *JMLR*, 4, 2003. 2.5
- [6] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *ICML 03*, pages 19 – 26, 2003. 5.1
- [7] Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, 5:255–291, December 2004. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1005332.1005342>. 6.5
- [8] Hannah Bast, Florian Baurle, Bjorn Buchhold, and Elmar Haussmann. Easy access to the freebase dataset. In *WWW*, 2014. 3.2.4, 3.3
- [9] Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, and Kateryna Tymoshenko. Extending english ace 2005 corpus annotation with ground-truth links to wikipedia. In *Proceedings of the 2nd Workshop on The Peoples Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 19–27, 2010. 3.4

- [10] Mustafa Bilgic, Lilyana Mihalkova, and Lise Getoor. Active learning for networked data. *Proceedings of the 27th International Conference on Machine Learning*, 2010. [5.1](#)
- [11] A. Bondu, V. Lemaire, and M. Boulle. Exploration vs. exploitation in active learning : A bayesian approach. In *IJCNN*, pages 1–7, July 2010. doi: 10.1109/IJCNN.2010.5596815. [6.5](#)
- [12] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013. [2.3.1](#), [3.2.4](#)
- [13] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017. [4.2.1](#), [4.4](#)
- [14] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015. [2.3.3](#), [2.5](#)
- [15] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *ICML*, 2013. [2.5](#), [7.1](#), [7.4](#)
- [16] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *ICML*, 2012. [2.5](#)
- [17] Andrew Chisholm and Ben Hachey. Entity disambiguation with web links. *Transactions of the Association of Computational Linguistics*, 3(1):145–156, 2015. [3.4](#)
- [18] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-78019-5. [6.3](#)
- [19] Peng Dai, Mausam, and Daniel S. Weld. Artificial intelligence for artificial artificial intelligence. *AAAI*, 2011. [5.1](#)
- [20] Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008. [2.5](#)
- [21] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, 2017. [3.3.1](#)

- [22] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, pages 199–207, 2011. 2.5
- [23] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 4.2.1, 4.4
- [24] P. Donmez and J. G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008. 5.1, 5.2.1, 5.3.2
- [25] P. Donmez and J. G. Carbonell. From Active to Proactive Learning. *Advances in Machine Learning I*, 262:97 – 120, 2010. 5.1, 5.3.2
- [26] P. Donmez, J. G. Carbonell, and P. Bennett. Dual strategy active learning. In *Proceedings of the European Conference on Machine Learning*, pages 116 – 127, 2007. 5.1, 6.4.2, 6.5
- [27] Lixin Duan, Dong Xu, and Ivor Tsang. Learning with augmented features for heterogeneous domain adaptation. *ICML*, 2012. 2.5
- [28] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 2011. 2.4.3, 3.3.3, 4.3.1, 4.3.2
- [29] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. Transition-based dependency parsing with stack long short-term memory. *ACL*, 2015. 3.1, 3.2.2
- [30] Desmond Elliott, Stella Frank, and Eva Hasler. Multi-language image description with neural sequence models. *CoRR*, *abs/1510.04709*, 2015. 3.4
- [31] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuda Yamada, and Omer Levy. Named entity disambiguation for noisy text. *CoNLL*, 2017. 3.1, 3.3.2, 3.4
- [32] Meng Fang, Jie Yin, and Dacheng Tao. Active learning for crowdsourcing using knowledge transfer. In *AAAI*, 2014. 7.1, 7.4
- [33] Min Fang, Yong Guo, Xiaosong Zhang, and Xiao Li. Multi-source transfer learning based on label shared subspace. *Pattern Recognition Letters*, 51:101–106, 2015. 4.4
- [34] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Journal of Machine learning*, 28(2-3):133–168, 1997. 6.5
- [35] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio

- Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2.1, 2.3.2, 2.3.3, 2.4.1, 2.5, 3.2.4
- [36] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015. 2.5
- [37] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 87–97, 2016. 4.4
- [38] Ravi Ganti and Alexander G Gray. Building bridges: Viewing active learning from the multi-armed bandit lens. *arXiv preprint arXiv:1309.6830*, 2013. 6.5
- [39] E. Gavves, T. E. J. Mensink, T. Tommasi, C. G. M. Snoek, and T Tuytelaars. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In *ICCV*, 2015. 7.1, 7.4
- [40] Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. Collective entity resolution with multi-focal attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 621–631, 2016. 3.4
- [41] Rakesh Gupta and Lev-Arie Ratinov. Text categorization with knowledge transfer from heterogeneous data sources. In *AAAI*, pages 842–847, 2008. 2.5
- [42] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 2.3.1
- [43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4.3.2
- [44] Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. Learning entity representation for entity disambiguation. 2013. 3.1, 3.3.2, 3.4
- [45] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011. 3.4
- [46] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J

- Smola. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007. 2.5
- [47] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010)*, volume 3, pages 3–13, 2010. 3.4
- [48] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural SVMs. *Mach Learn*, 77:27–59, 2009. 6.3
- [49] D. Kale and Yan Liu. Accelerating active learning with transfer learning. In *ICDM*, pages 1085–1090, 2013. 7.1, 7.4
- [50] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 2.5
- [51] Meghana Kshirsagar, Jaime Carbonell, and Judith Klein-Seetharaman. Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks. 2013. 2.5, 4.4
- [52] Tian Lan, Weilong Yang, Yang Wang, and Greg Mori. Image Retrieval with Structured Object Queries Using Latent Ranking SVM. In *ECCV*, 2012. 6.3.2
- [53] Yann Lecun and Corinna Cortes. The MNIST database of handwritten digits. URL <http://yann.l ecun. com/exdb/mni st/>. 6.4
- [54] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989. 3.1
- [55] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. *Proceedings of the International Conference on Machine Learning (ICML) '94*, pages 148 – 156, 1994. 5.2.1, 5.2.3, 6.3.1
- [56] D. Lewis and W. Gale. Training text classifiers by uncertainty sampling. In *Proceedings of ACM-SIGIR Conference on Information Retrieval*, 1994. 5.1
- [57] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr): 361–397, 2004. 2.4.3
- [58] Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, pages 4211–4219, 2015. 2.5

- [59] M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>. (document), 6.1, 6.4
- [60] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 469–477. Curran Associates, Inc., 2016. URL <http://papers.nips.cc/paper/6544-coupled-generative-adversarial-networks.pdf>. 4.4
- [61] Chengjiang Long and Gang Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. *ICCV*, 2015. 6.4.1
- [62] Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. *ICML*, 2015. 2.5
- [63] C.C. Loy, T.M. Hospedales, Tao Xiang, and Shaogang Gong. Stream-based joint exploration-exploitation active learning. In *CVPR*, pages 1560–1567, June 2012. doi: 10.1109/CVPR.2012.6247847. 6.5, 7.2.1, 7.3
- [64] Naoki Abe Hiroshi Mamitsuka. Query learning strategies using boosting and bagging. In *Machine Learning: Proceedings of the Fifteenth International Conference (ICML)*, volume 1. Morgan Kaufmann Pub, 1998. 6.5
- [65] A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. *ICML 98*, pages 359 – 367, 2001. 5.1
- [66] P. Melville and R. Mooney. Diverse ensembles for active learning. *International Conference on Machine Learning (ICML) '04*, 2004. 5.1, 6.5
- [67] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Economical active feature-value acquisition through expected utility estimation. *KDD 05 Workshop on Utility-based data mining*, 2005. 5.1
- [68] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 2.5
- [69] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013. 2.3.1
- [70] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2.3.1
- [71] Seungwhan Moon and Jaime Carbonell. Proactive learning with multiple class-sensitive

- labelers. *International Conference on Data Science and Advanced Analytics (DSAA)*, 2014. [1.2.2](#), [6.2](#), [6.3.1](#), [6.4.1](#), [6.4.2](#), [7.4](#)
- [72] Seungwhan Moon and Jaime Carbonell. Proactive transfer learning for heterogeneous feature and label spaces. *ECML-PKDD*, 2016. [1.2.1](#), [1.2.4](#), [2.4.1](#)
- [73] Seungwhan Moon and Jaime Carbonell. Dynamic active learning with multiple selection strategies. In *arxiv*, 2017. [1.2.2](#)
- [74] Seungwhan Moon and Jaime Carbonell. Completely heterogeneous transfer learning with attention: What and what not to transfer. *IJCAI*, 2017. [1.2.1](#), [1.2.3](#), [1.2.4](#)
- [75] Seungwhan Moon and Jaime Carbonell. Adversarial transfer network for feature learning. *arxiv*, 2018. [1.2.1](#), [1.2.4](#)
- [76] Seungwhan Moon, Calvin McCarter, and Yu-Hsin Kuo. Active learning with partially observed data. *Proceedings of International World Wide Web Conference (WWW)*, 2014. [1.2.2](#), [5.1](#)
- [77] Seungwhan Moon, Leonard Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. *NAACL*, 2018. [1.2.4](#), [3.2.5](#), [3.4](#)
- [78] Seungwhan Moon, Leonard Neves, and Vitor Carvalho. Zeroshot multimodal named entity disambiguation for noisy social media posts. *ACL*, 2018. [1.2.4](#)
- [79] H.T. Nguyen and A. Smeulders. Active learning using pre-clustering. *International Conference on Machine Learning (ICML)*, 2004. [5.1](#), [5.2.3](#), [5.2.3](#), [5.3.1](#), [6.3.1](#), [7.2.1](#)
- [80] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. *AAAI*, 2016. [2.3.1](#), [3.2.4](#), [3.2.4](#)
- [81] T. Osugi, Deng Kim, and S. Scott. Balancing exploration and exploitation: a new algorithm for active machine learning. In *Data Mining, Fifth IEEE International Conference on*, pages 8 pp.–, Nov 2005. doi: 10.1109/ICDM.2005.33. [6.5](#)
- [82] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. Accurate Integration of Crowdsourced Labels Using Workers’ Self reported Confidence Scores. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013. [5.1](#)
- [83] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010. [2.5](#)
- [84] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vec-

- tors for word representation. In *EMNLP*, 2014. URL <http://www.aclweb.org/anthology/D14-1162>. 3.2.2, 3.3.1
- [85] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014. 4.3.2
- [86] Maria Pershina, Yifan He, and Ralph Grishman. Personalized page rank for named entity disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–243, 2015. 3.4
- [87] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4.4
- [88] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *NIPS*, pages 46–54, 2013. 2.5
- [89] Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*, volume 2, page 7, 2005. 2.5
- [90] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. pages 441 – 448, 2001. 5.1
- [91] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 3.2.2, 4.3.2
- [92] Marcos Salganicoff and Lyle H Ungar. Active exploration and learning in real-valued spaces using multi-armed bandit allocation indices. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 480–487, 2014. 6.5
- [93] B. Settles and M. Craven. Training text classifiers by uncertainty sampling. *EMNLP*, pages 1069 – 1078, 2008. 5.2.3, 6.3.1, 6.4.2, 7.2.1, 7.3, 7.4
- [94] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 287–294. ACM, 1992. 6.5

- [95] Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In *KDD*, pages 342–357. Springer, 2008. 7.1, 7.4
- [96] Xiaoxiao Shi, Qi Liu, Wei Fan, Qiang Yang, and S Yu Philip. Predictive modeling with heterogeneous sources. In *SDM*, pages 814–825. SIAM, 2010. 2.5
- [97] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero Shot Learning Through Cross-Modal Transfer. In *NIPS*. 2013. 2.5
- [98] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multi-modal machine translation and crosslingual image description. In *WMT*, pages 543–553, 2016. 3.4
- [99] B Strack, J Olmo, DeShazo, C Jennings, KJ Cios, and JN Clore. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014. 5.3
- [100] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NIPS*, pages 2440–2448, 2015. 2.5, 3.4
- [101] Qian Sun, Mohammad Amin, Baoshi Yan, Craig Martell, Vita Markman, Anmol Bhasin, and Jieping Ye. Transfer learning for bilingual content classification. In *KDD*, pages 2147–2156, 2015. 2.5
- [102] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR*, 2015. 3.2.2
- [103] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017. 4.2.1, 4.4
- [104] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008. 2.4.3
- [105] Paolo Viappiani, Sandra Zilles, Howard J. Hamilton, and Craig Boutilier. Learning complex concepts using crowdsourcing: A bayesian approach. *Algorithmic Decision Theory*, 2011. 5.1
- [106] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Who should label what? instance allocation in multiple expert active learning. In *In Proc. of the SIAM International Conference on Data Mining (SDM)*, 2011. 5.1
- [107] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view repre-

- sentation learning. *ICML*, 2015. 2.5
- [108] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119. Citeseer, 2014. 2.3.1, 3.2.4
- [109] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI'11*, 2011. 2.5
- [110] Min Xiao and Yuhong Guo. Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation. In *ECMLPKDD*. 2015. 2.5
- [111] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015. 2.3.3, 2.5, 3.4
- [112] Zhijie Xu and Shiliang Sun. Multi-source transfer learning with multi-view adaboost. In *International Conference on Neural Information Processing*, pages 332–339. Springer, 2012. 4.4
- [113] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. *CoNLL*, 2016. 3.1, 3.3.2
- [114] Yan Yan, Glenn M Fung, Rómer Rosales, and Jennifer G Dy. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168, 2011. 5.1
- [115] Yan Yan, Rmer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327, 2014. ISSN 0885-6125. doi: 10.1007/s10994-013-5412-1. URL <http://dx.doi.org/10.1007/s10994-013-5412-1>. 6.3.1, 6.4.1, 6.4.2
- [116] L. Yang and J. Carbonell. Cost complexity of proactive learning via a reduction to realizable active learning. *Tech report CMU-ML-09-113*, 2010. 5.1
- [117] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015. 2.5
- [118] Jamie Callan Yi Zhang, Wei Xu. Exploration and exploitation in adaptive filtering based on bayesian active learning. 2003. 6.5

- [119] Chun-Nam John Yu and Thorsten Joachims. Learning Structural SVMs with Latent Variables. In *ICML*, 2009. 6.3, 6.3.2, 6.3.2
- [120] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 2.5
- [121] Joey Zhou, Sinno Pan, Ivor Tsang, and Shen-Shyang Ho. Transfer learning for cross-language text categorization through active correspondences construction. *AAAI*, 2016. 2.5
- [122] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W. Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. *AAAI*, 2014. 2.5
- [123] J. Zhu, H. Wang, B. Tsou, and M. Ma. Active Learning With Sampling by Uncertainty and Density for Data Annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18, 2010. 5.1, 5.2.3, 5.3.1, 6.3.1, 6.4.2, 7.2.1
- [124] Zhenfeng Zhu, Xingquan Zhu, Yangdong Ye, Yue-Fei Guo, and Xiangyang Xue. Transfer active learning. In *CIKM*, pages 2169–2172, 2011. 2.5, 7.1, 7.4