Learning the Piano...

From accordion to piano...
Motivation: lack of training data
Motivation: performance drop
Transfer Learning: Overview

<table>
<thead>
<tr>
<th>Source Data (not directly related to the task)</th>
</tr>
</thead>
<tbody>
<tr>
<td>labelled</td>
</tr>
<tr>
<td>Unlabelled</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source Domain</th>
<th>Target Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>labelled</td>
<td>labelled</td>
</tr>
<tr>
<td>Unlabelled</td>
<td>Unlabelled</td>
</tr>
</tbody>
</table>

- **Fine-tuning**
- **Multitask Learning**
- **Domain Adaptation**
- **Self-taught learning**
  - Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, Andrew Y. Ng, Self-taught learning: transfer learning from unlabeled data, ICML, 2007
- **Self-taught Clustering**
  - Wenyuan Dai, Qiang Yang, Gui-Rong Xue, Yong Yu, "Self-taught clustering", ICML 2008
- **Zero-shot learning**
Multi-task Learning

Share layers, change task-specific layers

Multi-Task Learning for Multiple Language Translation
Multi-task Learning for Multiple Language Translation
Zero-shot learning

Unseen training samples

- Same distribution:
  Share similar features!

- Different tasks

In my training dataset: Chimp, Dog

In my testing dataset: Fish

http://speech.ee.ntu.edu.tw/~tlkagk/
A toy example...

**Training**

```
<table>
<thead>
<tr>
<th>furry</th>
<th>4 legs</th>
<th>tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```

**Database**

```
attributes

<table>
<thead>
<tr>
<th>furry</th>
<th>4 legs</th>
<th>tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

sufficient attributes for one to one mapping

http://speech.ee.ntu.edu.tw/~tlkagk/
A toy example...

Testing

Find the class with the most similar attributes

attributes

<table>
<thead>
<tr>
<th></th>
<th>furry</th>
<th>4 legs</th>
<th>tail</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>Fish</td>
<td>X</td>
<td>X</td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>Chim</td>
<td>O</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sufficient attributes for one to one mapping

http://speech.ee.ntu.edu.tw/~tlkagk/
Zero-shot Learning with Machine Translation

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Training (blue):

EN <-> JP

EN <-> KO

Goal (orange):

JP <-> KO
Google’s Multilingual Neural Machine Translation Model

https://vimeo.com/238233299
Simple idea...

Everything is shared: encoder, decoder, attention model.

Prepend source to indicate target language:

32,000 word pieces each (smaller language pairs get oversampled)

https://vimeo.com/238233299
Sentence T-SNE Visualization

The stratosphere extends from about 10km to about 50km in altitude.

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation
Cases in NLP? Different Distributions...

10 hours ago
Edward Priz replied:
You know, this isn't the first time that "States Rights" has been used as a cover for racist policies. In fact, the whole "States Rights" thing has become a sort of code for heavy-handed racist policies, hasn't it? And it does provide a sort of contextual...

10 hours ago
RICH HIRTH replied:
The issue here is probable cause. A police officer can question if he has probable cause, and he can document it. This law can be abused if being Latino is probable cause. That is license to harass for the police. As long as the law is applied fairly there.

2 hours ago
Julia Gomez replied:
The Arizona law is so clearly unconstitutional that I do not think it will ever reach the point of being enforced. The article did not say so, but the Republican governor is afraid of a GOP primary electorate that is even more reactionary than usual. That is why she signed the bill, not because she thinks it is legally defensible.

https://drive.google.com/file/d/1jEWXZzayUIXj0TDZZGX9iKsDIFx0JKUw/view
# Cases in NLP? Different Distributions...

<table>
<thead>
<tr>
<th>Electronics</th>
<th>Video games</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) <strong>Compact</strong>; easy to operate; very good picture quality; looks <strong>sharp</strong>!</td>
<td>(2) A very good game! It is action packed and full of excitement. I am very much <strong>hooked</strong> on this game.</td>
</tr>
<tr>
<td>(3) I purchased this unit from Circuit City and I was very <strong>excited</strong> about the quality of the picture. It is really <strong>nice</strong> and <strong>sharp</strong>.</td>
<td>(4) Very <strong>realistic</strong> shooting action and good plots. We played this and were <strong>hooked</strong>.</td>
</tr>
<tr>
<td>(5) It is also quite <strong>blurry</strong> in very dark settings. I will <strong>never_buy</strong> HP again.</td>
<td>(6) It is so boring. I am extremely <strong>unhappy</strong> and will probably <strong>never_buy</strong> UbiSoft again.</td>
</tr>
</tbody>
</table>

- Source specific: compact, sharp, blurry.
- Target specific: hooked, realistic, boring.
- Domain independent: good, excited, nice, never_buy, unhappy.
## Other cases in NLP: Q&A

<table>
<thead>
<tr>
<th>Source Dataset</th>
<th>Target Dataset</th>
<th>Target Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MovieQA</strong></td>
<td><strong>TOEFL</strong></td>
<td><strong>MCTest</strong></td>
</tr>
<tr>
<td><strong>S</strong></td>
<td><strong>Q</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>After entering the boathouse, the trio witness Voldemort telling Snape that the elder Wand cannot serve Voldemort until Snape dies ... Before dying, Snape tells Harry to take his memories to the Pensieve ...</td>
<td>I just wanted to take a few minutes to meet with everyone to make sure your class presentations for next week are all in order and coming along well. And as you know, you’re supposed to report on some areas of recent research on genetics ...</td>
<td>James the Turtle was always getting in trouble. Sometimes he’d reach into the freezer and empty out all the food ... Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn’t pay, and instead headed home ...</td>
</tr>
<tr>
<td><strong>C_1</strong></td>
<td><strong>Q</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>To bury him in the forest</td>
<td>Why does the professor meet with the student?</td>
<td>To find out if the student is interested in taking part in a genetics project</td>
</tr>
<tr>
<td><strong>C_2</strong></td>
<td><strong>Q</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>That he always respected him</td>
<td>To discuss the student’s experiment on the taste perception</td>
<td></td>
</tr>
<tr>
<td><strong>C_3</strong></td>
<td><strong>Q</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>To remember to him for the good deeds</td>
<td>To determine if the student has selected an appropriate topic for his class project</td>
<td></td>
</tr>
<tr>
<td><strong>C_4</strong></td>
<td><strong>Q</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td><strong>To take his memories to the Pensieve</strong></td>
<td>To explain what the student should focus on for his class presentation</td>
<td></td>
</tr>
<tr>
<td><strong>C_5</strong></td>
<td><strong>Q</strong></td>
<td><strong>C</strong></td>
</tr>
<tr>
<td>To write down his memories in a book</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Other cases in NLP: Summarization

**Article (part)**

_In an Ocean of Marinara Sauce, 12 Places Where Dining Is More Than Molto Bene_

By **JOANNE STARKEY**

ITALY reigns supreme on _Long Island_, from Montauk to Malverne. Because of that, a list of the top Italian restaurants is in reality the best of the best. Many that we regard as just ordinary would be hailed as outstanding elsewhere. Here are 12 that could be considered an Italian hall of fame...

**Domain shifts:**
- Vocabulary
- Wrong Info

---

**Reference Abstract**


**Generated Abstract**

Domain Adaptation In NLP?

- Setting
  - Source domain: $D_S = \{(x_S, y_S)\}$
  - Target domain: $D_T = \{(x_T)\}$

- Problems in NLP
  - Frequency bias: $P(x_S) \neq P(x_T)$
    - Different frequencies: same word in different domains
  - Context feature bias: $P(y_S|x_S) \neq P(y_T|x_T)$
    - “monitor” in Wall Street Journal and Amazon reviews
General Methods in Transfer Learning

**Feature-based methods:**

- Transfer the features into the same feature space!
- Multi-layer feature learning (representation learning)

**Model-based methods:**

- Parameter init + fine-tune (a lot!)
- Parameter sharing

**Instance-based methods (traditional, not going to cover):**

- Re-weighting: make source inputs similar with target inputs
- Pseudo samples for target domain
Feature-based method: Intuition
First Paper: Feature-based method-Deep Adaptation Network

Learning transferable features with deep adaptation networks ICML, 2015

Task: image classification

Setting:
Source domain with labels
Target domain without labels

Model:
VGG net loss + domain loss

CNN Features

Lower level features are shared..
Layer Transfer

Deep Models...

1. Only train the rest layers (prevent overfitting)
2. fine-tune the whole network (if there is sufficient data)

Source data

Copy some parameters

Target data

http://speech.ee.ntu.edu.tw/~tlkagk/
Layer transfer in CNNs...

Freeze the first few layers, they are shared...

We train domain-specific layers!
Loss function: discriminativeness and domain invariance

\[
\min_{\Theta} \frac{1}{n_a} \sum_{i=1}^{n_a} J(\theta(x_i^a), y_i^a) + \lambda \sum_{\ell=l_1}^{l_2} d^2_{k_\ell}(\mathcal{D}_s^\ell, \mathcal{D}_t^\ell)
\]

Source error (CNN loss) + domain discrepancy (MK-MMD)

Multi-kernel Maximum Mean Discrepancy
Maximum Mean Discrepancy (MMD)

Two-sample problem (unknown p and q):

\[ X := \{x_1, x_2, \ldots, x_m\} \sim p \text{ and } Y := \{y_1, y_2, \ldots, y_n\} \sim q, \text{ test whether } p = q \]

Maximum Mean Discrepancy (Muller, 1997):
Map the layers into a Reproducing Kernel Hilbert Space H with kernel function k:

\[
MMD^2(p, q) = \|p - q\|_H^2 = \frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j) \leq O(n^2)
\]
MK-MMD: Optimization

Unbiased estimation in $O(n)$:

$$MMD_{u}^{2}[F, X, Y] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h(z_i, z_j)$$

$$h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$$

Kernel:

Gaussian Kernel (RBF), bandwidth sigma could be estimated.

$$K(x, y) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right)$$

Multi-kernel:

$$\mathcal{K} \triangleq \left\{ k = \sum_{u=1}^{m} \beta_u k_u : \sum_{u=1}^{m} \beta_u = 1, \beta_u \geq 0, \forall u \right\}$$

Optimize the beta
About this method...

Competitive performance!

Loss function:

need to learn lambda from the validation set;

hard to control (optimization), when to plug in the domain loss?

$$\min_{\Theta} \frac{1}{n_a} \sum_{i=1}^{n_a} J(\theta(x_i^a), y_i^a) + \lambda \sum_{\ell=l_1}^{l_2} d_{k}^2(D_s^\ell, D_t^\ell)$$

Few research on applying it into NLP applications.
Paper 2: Feature-based Method -- word embeddings

**Task:** sentiment classification (pos or neg)

**Motivation:** reviews or healthcare domain sentiment classifier?

I’ve been clean for about 7 months but even now I still feel like maybe I won’t make it….

I feel like I am getting my life back…

Samples from A-CHESS dataset: a study involving users with alcohol addiction.

**Method:** improve the word embeddings

**Domain Adapted (DA) embeddings**

[Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)]
Why word embeddings?

**Word2vec and GloVe are old news!**

--- Alex Fabbri

**Generic (G) embeddings:**

- word2vec, GloVe trained from Wikipedia, WWW;
- general knowledge

**Domain Specific (DS) embeddings:**

- trained from domain datasets (small-scale);
- domain knowledge

**Domain Adapted (DA) embeddings:** combine them!
Why word embeddings?

Domain-Adapted Embeddings:
- Canonical Correlation Analysis (CCA)
- Kernel CCA (KCCA, nonlinear version of CCA, using RBF kernel)

Word embeddings to sentence encoding:
  i.e. a weighted combination of their constituent word embeddings.

Use a Logistic Regressor to do classification (pos or neg).

Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)
Intuition: Combine two embedding feature space

Domain Specific Embedding

\[ \bar{w}_{i,DS} = w_{i,DS} \phi_{DS} \]

\[ \bar{w}_{i,G} = w_{i,G} \phi_{G}. \]

Generic Embedding

CCA maximizes the correlation between \( \bar{w}_{i,DS} \) and \( \bar{w}_{i,G} \) to obtain \( \phi_{DS} \) and \( \phi_{G} \) such that

\[
\rho(\phi_{DS}, \phi_{G}) = \max_{\phi_{DS}, \phi_{G}} \frac{\mathbb{E}[\langle \bar{w}_{i,DS}, \bar{w}_{i,G} \rangle]}{\sqrt{\mathbb{E}[\bar{w}_{i,DS}^2] \mathbb{E}[\bar{w}_{i,G}^2]}}
\]

\[ (2) \]
Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA): Let $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_m)$ be random variables, and there are correlations among the variables, then canonical-correlation analysis will find linear combinations of $X$ and $Y$ which have maximum correlation with each other.

$$\tilde{w}_{i, DS} = \underbrace{w_{i, DS}} \phi_{DS}$$  
LSA Embedding * Mapping

$$\tilde{w}_{i, G} = \underbrace{w_{i, G}} \phi_{G}.$$  
GloVe Embedding * Mapping

Domain-Adapted Embedding:

$$\min_{\alpha, \beta} \| \tilde{w}_{i, DS} - (\alpha \tilde{w}_{i, DS} + \beta \tilde{w}_{i, G}) \|_2^2 + \| \tilde{w}_{i, G} - (\alpha \tilde{w}_{i, DS} + \beta \tilde{w}_{i, G}) \|_2^2.$$  

$$\hat{w}_{i, DA} = \frac{1}{2} \tilde{w}_{i, DS} + \frac{1}{2} \tilde{w}_{i, G}$$  
Final Embedding

Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)
Feature-based Method: share word embeddings

Result on Yelp Dataset

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Embedding</th>
<th>Avg Precision</th>
<th>Avg F-score</th>
<th>Avg AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>KCCA(Glv, LSA)</td>
<td>85.36± 2.8</td>
<td>81.89± 2.8</td>
<td>82.57± 1.3</td>
<td></td>
</tr>
<tr>
<td>CCA(Glv, LSA)</td>
<td>83.69± 4.7</td>
<td>79.48± 2.4</td>
<td>80.33± 2.9</td>
<td></td>
</tr>
<tr>
<td>KCCA(w2v, LSA)</td>
<td>87.45± 1.2</td>
<td>83.36± 1.2</td>
<td>84.10± 0.9</td>
<td></td>
</tr>
<tr>
<td>CCA(w2v, LSA)</td>
<td>84.52± 2.3</td>
<td>80.02± 2.6</td>
<td>81.04± 2.1</td>
<td></td>
</tr>
<tr>
<td>KCCA(GlvCC, LSA)</td>
<td>88.11± 3.0</td>
<td>85.35± 2.7</td>
<td>85.80± 2.4</td>
<td></td>
</tr>
<tr>
<td>CCA(GlvCC, LSA)</td>
<td>83.69± 3.5</td>
<td>78.99± 4.2</td>
<td>80.03± 3.7</td>
<td></td>
</tr>
<tr>
<td>KCCA(w2v, DSw2v)</td>
<td>78.09± 1.7</td>
<td>76.04± 1.7</td>
<td>76.66± 1.5</td>
<td></td>
</tr>
<tr>
<td>CCA(w2v, DSw2v)</td>
<td>86.22± 3.5</td>
<td>84.35± 2.4</td>
<td>84.65± 2.2</td>
<td></td>
</tr>
<tr>
<td>concSVD(Glv, LSA)</td>
<td>80.14± 2.6</td>
<td>78.50± 3.0</td>
<td>78.92± 2.7</td>
<td></td>
</tr>
<tr>
<td>concSVD(w2v, LSA)</td>
<td>85.11± 2.3</td>
<td>83.51± 2.2</td>
<td>83.80± 2.0</td>
<td></td>
</tr>
<tr>
<td>concSVD(GlvCC, LSA)</td>
<td>84.20± 3.7</td>
<td>80.39± 3.7</td>
<td>80.83± 3.9</td>
<td></td>
</tr>
<tr>
<td>GloVe</td>
<td>77.13± 4.2</td>
<td>72.32± 7.9</td>
<td>74.17± 5.0</td>
<td></td>
</tr>
<tr>
<td>GloVe-CC</td>
<td>82.10± 3.5</td>
<td>76.74± 3.4</td>
<td>78.17± 2.7</td>
<td></td>
</tr>
<tr>
<td>word2vec</td>
<td>82.80± 3.5</td>
<td>78.28± 3.5</td>
<td>79.35± 3.1</td>
<td></td>
</tr>
<tr>
<td>LSA</td>
<td>75.36± 5.4</td>
<td>71.17± 4.3</td>
<td>72.57± 4.3</td>
<td></td>
</tr>
<tr>
<td>word2vec</td>
<td>73.08± 2.2</td>
<td>70.97± 2.4</td>
<td>71.76± 2.1</td>
<td></td>
</tr>
</tbody>
</table>

Yelp: 1000 balanced reviews
Tokens: 2049

Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)
## Feature-based Method: share word embeddings

### Result on A-CHESS Dataset

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Embedding</th>
<th>Avg Precision</th>
<th>Avg F-score</th>
<th>Avg AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-CHESS</td>
<td>KCCA(Glv, LSA)</td>
<td>32.07±1.3</td>
<td>39.32±2.5</td>
<td>65.96±1.3</td>
</tr>
<tr>
<td></td>
<td>CCA(Glv, LSA)</td>
<td>32.70±1.5</td>
<td>35.48±4.2</td>
<td>62.15±2.9</td>
</tr>
<tr>
<td></td>
<td><strong>KCCA(w2v, LSA)</strong></td>
<td>33.45±1.3</td>
<td><strong>39.81±1.0</strong></td>
<td>65.92±0.6</td>
</tr>
<tr>
<td></td>
<td>CCA(w2v, LSA)</td>
<td>33.06±3.2</td>
<td>34.02±1.1</td>
<td>60.91±0.9</td>
</tr>
<tr>
<td></td>
<td>KCCA(GlvCC, LSA)</td>
<td>36.38±1.2</td>
<td>34.71±4.8</td>
<td>61.36±2.6</td>
</tr>
<tr>
<td></td>
<td>CCA(GlvCC, LSA)</td>
<td>32.11±2.9</td>
<td>36.85±4.4</td>
<td>62.99±3.1</td>
</tr>
<tr>
<td></td>
<td><strong>KCCA(w2v, DSw2v)</strong></td>
<td>25.59±1.2</td>
<td>28.27±3.1</td>
<td>57.25±1.7</td>
</tr>
<tr>
<td></td>
<td>CCA(w2v, DSw2v)</td>
<td>24.88±1.4</td>
<td>29.17±3.1</td>
<td>57.76±2.0</td>
</tr>
<tr>
<td></td>
<td>concSVD(Glv, LSA)</td>
<td>27.27±2.9</td>
<td>34.45±3.0</td>
<td>61.59±2.3</td>
</tr>
<tr>
<td></td>
<td>concSVD(w2v, LSA)</td>
<td>29.84±2.3</td>
<td>36.32±3.3</td>
<td>62.94±1.1</td>
</tr>
<tr>
<td></td>
<td>concSVD(GlvCC, LSA)</td>
<td>28.09±1.9</td>
<td>35.06±1.4</td>
<td>62.13±2.6</td>
</tr>
<tr>
<td>W_G</td>
<td>GloVe</td>
<td>30.82±2.0</td>
<td>33.67±3.4</td>
<td>60.80±2.3</td>
</tr>
<tr>
<td></td>
<td>GloVe-CC</td>
<td><strong>38.13±0.8</strong></td>
<td>27.45±3.1</td>
<td>57.49±1.2</td>
</tr>
<tr>
<td></td>
<td>word2vec</td>
<td>32.67±2.9</td>
<td>31.72±1.6</td>
<td>59.64±0.5</td>
</tr>
<tr>
<td>W_DS</td>
<td>LSA</td>
<td>27.42±1.6</td>
<td>34.38±2.3</td>
<td>61.56±1.9</td>
</tr>
<tr>
<td></td>
<td>word2vec</td>
<td>24.48±0.8</td>
<td>27.97±3.7</td>
<td>57.08±2.5</td>
</tr>
</tbody>
</table>

**A_CHESS:** 8% unbalanced

**Total:** 2500 samples

**Tokens:** 3400

---

*Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)*
About this method...

Straightforward: modify the word embeddings;
Easy to implement;
Possible to improve on sentence embeddings as well.

Small datasets (thousands for training and testing).
Improvements on classification, what about other tasks?
Datasets:
(Source) MovieQA
(Target 1) TOEFL listening comprehension
(Target2) MCTest

Task: QA
Read an article + a question, find out a correct answer from 4 or 5 choices.

Models:
MemN2N (End-to-end Memory Network),
QACNN (Query-Based Attention CNN)

Dataset example

<table>
<thead>
<tr>
<th>MovieQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>After entering the boathouse, the trio witness Voldemort telling Snape that the elder Wand cannot serve Voldemort until Snape dies ... Before dying, Snape tells Harry to take his memories to the Pensieve ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q</th>
<th>C_1</th>
<th>C_2</th>
<th>C_3</th>
<th>C_4</th>
<th>C_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>What does Snape tell Harry before he dies?</td>
<td>To bury him in the forest</td>
<td>That he always respected him</td>
<td>To remember to him for the good deeds</td>
<td>To take his memories to the Pensieve</td>
<td>To write down his memories in a book</td>
</tr>
</tbody>
</table>
Model-based Method: pre-train and fine-tune

Datasets:
- (Source) MovieQA
- (Target 1) TOEFL listening comprehension
- (Target 2) MCTest

Pre-train on MovieQA,
Fine-tune using target datasets.
Fine-tune different layers.

QACNN (Query-Based Attention CNN)

Supervised and Unsupervised Transfer Learning for Question Answering. NAACL, 2018
## Model-based Method: fine-tune

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>TOEFL</th>
<th>MCTest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>manual</td>
<td>ASR</td>
</tr>
<tr>
<td>QACNN</td>
<td>(a) Target Only</td>
<td>48.9</td>
<td>47.5</td>
</tr>
<tr>
<td></td>
<td>(b) Source Only</td>
<td>51.2</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>(c) Source + Target</td>
<td>52.5</td>
<td>49.7</td>
</tr>
<tr>
<td></td>
<td>(d) Fine-tuned (1)</td>
<td>53.4</td>
<td>51.5</td>
</tr>
<tr>
<td></td>
<td>(e) Fine-tuned (2)</td>
<td>56.1</td>
<td>55.3</td>
</tr>
<tr>
<td></td>
<td>(f) Fine-tuned (all)</td>
<td>56.0</td>
<td>55.1</td>
</tr>
</tbody>
</table>

Without pre-train on MovieQA
Train only on MovieQA
Train on both!

**Results**

*Supervised and Unsupervised Transfer Learning for Question Answering. NAACL, 2018*
Effective Domain Mixing for Neural Machine Translation

Model: Sequence-to-sequence model for neural Machine Translation

Three translation tasks:

- EN-JA, EN-ZH, EN-FR

Heterogeneous corpora: News vs TEDtalks
Recall NMT...

A deep recurrent neural network

\[
\log p(y|x)
\]

\[
\mathcal{L}_{gen} = \sum_{(x,y) \in \mathcal{D}} - \log p(y|x)
\]

x: source input
y: target output

Discriminative Mixing

**encoder:** domain-related information

\[
\mathcal{L} = \mathcal{L}_{gen} + \mathcal{L}_{disc}
\]

- **source** → **encoder**
- **decoder** → **target**
- **discriminator** → **domain**

\[
c = \sum_j a_j h_j
\]

\[
a = \text{softmax}(\hat{a})
\]

\[
\hat{a}_i = v_a^T \tanh(W_a h_i)
\]

Pryzant et al. *Effective Domain Mixing for Neural Machine Translation* 2017
Adversarial Discriminative Mixing

\[ \mathcal{L} = \mathcal{L}_{gen} + \mathcal{L}_{disc} \]

such representations lead to better generalization across domains...

Pryzant et al. Effective Domain Mixing for Neural Machine Translation 2017
Target Token Mixing

(prepend a special token … “domain=subtitles”)

Pryzant et al. Effective Domain Mixing for Neural Machine Translation 2017
How similar are two domains?

Intuition: is it easy to distinguish?

Proxy A-Distance (PAD):

- Mix the two datasets. Apply label that indicate each example's origin.
- Train a classifier on these merged data (linear bag-of-words SVM).
- Measure the classifier's error $e$ on a held-out test set.
- Set $PAD = 2 \ (1 - 2e)$

Small PAD: similar domains (when $e$ is large, hard to tell)

Large PAD: dissimilar domains (when $e$ is small, easy to tell)

Tools available at: https://github.com/rpryzant/proxy-a-distance
Results

Baseline is mixing samples...

Mixing will help...

Similar domains

Table 1: Proxy $A$-distances ($\hat{d}_A$) for each domain pair.

<table>
<thead>
<tr>
<th>Language</th>
<th>Domain 1</th>
<th>Domain 2</th>
<th>$\hat{d}_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>ASPEC</td>
<td>SubCrawl</td>
<td>1.89</td>
</tr>
<tr>
<td>Chinese</td>
<td>News</td>
<td>TED</td>
<td>1.73</td>
</tr>
<tr>
<td>French</td>
<td>Europarl</td>
<td>OpenSubs</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Pryzant et al. Effective Domain Mixing for Neural Machine Translation 2017

<table>
<thead>
<tr>
<th>EN-JA Model</th>
<th>ASPEC</th>
<th>SubCrawl</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASPEC</td>
<td>38.87</td>
<td>3.85</td>
</tr>
<tr>
<td>SubCrawl</td>
<td>2.74</td>
<td>16.91</td>
</tr>
<tr>
<td>ASPEC + SubCrawl</td>
<td>33.85</td>
<td>14.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EN-FR Model</th>
<th>Europarl</th>
<th>OpenSubs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl</td>
<td>34.51</td>
<td>13.36</td>
</tr>
<tr>
<td>OpenSubtitles</td>
<td>13.12</td>
<td>15.2</td>
</tr>
<tr>
<td>Europarl + OpenSubs</td>
<td>38.26</td>
<td>27.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EN-ZH Model</th>
<th>News</th>
<th>TED</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>12.75</td>
<td>3.12</td>
</tr>
<tr>
<td>TED</td>
<td>2.79</td>
<td>8.41</td>
</tr>
<tr>
<td>News + TED</td>
<td>11.36</td>
<td>6.67</td>
</tr>
</tbody>
</table>

Pryzant et al. Effective Domain Mixing for Neural Machine Translation 2017

BLEU Scores
The more diverge the more discriminative model helps.

(c) Comparing the proposed discriminator approach and mixed-domain baseline ($\text{BLEU}_{\text{discriminator}} - \text{BLEU}_{\text{mixed}}$) while varying domain distance. The discriminator always improves over the baseline, and this is accentuated when the merged domains are more distant.

Pryzant et al. Effective Domain Mixing for Neural Machine Translation 2017
Conclusion

Mixing data from heterogeneous domains leads to suboptimal results compared to the single-domain setting;

The more distant these domains are, the more their merger degrades downstream translation quality;

Target Token Mixing: off the shelf method.

Pryzant et al Effective Domain Mixing for Neural Machine Translation 2017
Model-based methods vs. Feature-based methods

Model-based methods:
Explicit, straightforward: add some modules, or fine-tune, etc.
Simple but really works in engineering!

Feature-based methods:
Theoretical: statistics, etc.
Now there are more research works: i.e., better sentence representations.
Transfer Learning works in CV: a lot!

Transfer Learning works in NLP:
- Simple tasks like classification, sentiment analysis, SRL, etc: a lot!
- Other tasks like machine translation, summarization: few!

More efforts:
- Datasets: in ‘domains’-> ‘News’ vs ‘Tweets’; ‘General’ vs ‘Medical’, etc
- Explainable models: how the models are transferred? What are transferred?
## Other Related Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Task</th>
<th>Size</th>
<th>#domains</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-newsgroup</td>
<td>Classification</td>
<td>18,828 in total</td>
<td>6</td>
<td>URL</td>
</tr>
<tr>
<td>Reuters--21578</td>
<td>Classification</td>
<td>4,771</td>
<td>3</td>
<td>URL</td>
</tr>
<tr>
<td>Amazon Reviews</td>
<td>Sentiment</td>
<td>3,685 to 5,945 per domain</td>
<td>4 to 20 (unprocessed)</td>
<td>URL</td>
</tr>
<tr>
<td>New York Times Annotated</td>
<td>Summarization</td>
<td>650k in total</td>
<td>2 main</td>
<td>URL</td>
</tr>
</tbody>
</table>

Do experiments across the datasets: Yelp vs Amazon...
Summary

Why do we need Transfer Learning?
What is Transfer Learning?
Multi-task learning, zero-shot learning
Transfer Learning Methods:
  Feature-based methods
  Model-based methods
  Uncovered: GANs, Reinforcement Learning methods, etc
Transfer Learning: future?
Discussion on open questions...

1. Where TL can help in other scenarios (NLP, CV, Speech Recognition)?

2. CNNs for images VS. seq2seq models for texts:
   - how are models transferred? (CNN: shallow features are learned by first few CNN layers, which are easy to be shared, what about seq2seq models?)

3. Other methods to see how similar of two domains besides PAD(Proxy A-distance)?


http://alex.smola.org/icml2008/
https://github.com/jindongwang/transferlearning
http://cs231n.github.io/transfer-learning/
Suggested readings...

Adversarial Networks:

[Domain-Adversarial Training of Neural Networks](#)

[Aspect-augmented Adversarial Networks for Domain Adaptation](#)

Seq2seq + Transfer:

[How Transferable are Neural Networks in NLP Applications?](#)

And the [Bibliography](#)
THANKS!

Q&A

ireneli.eu