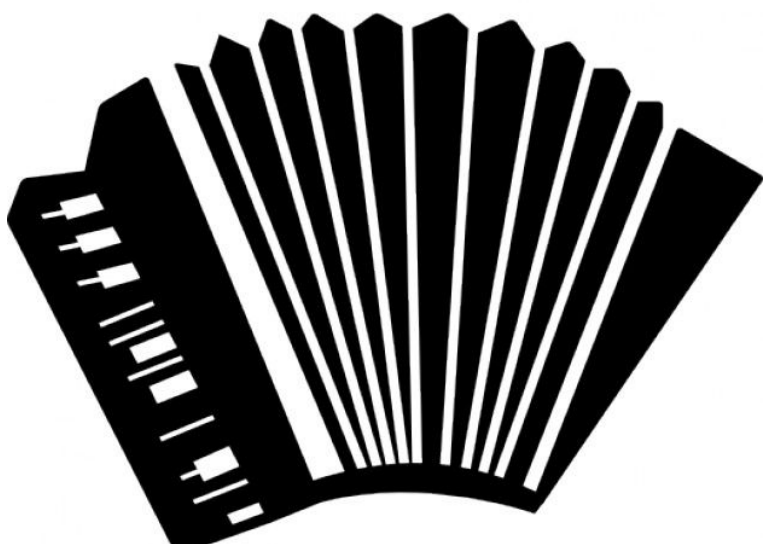
ANLP:
Transfer Learning

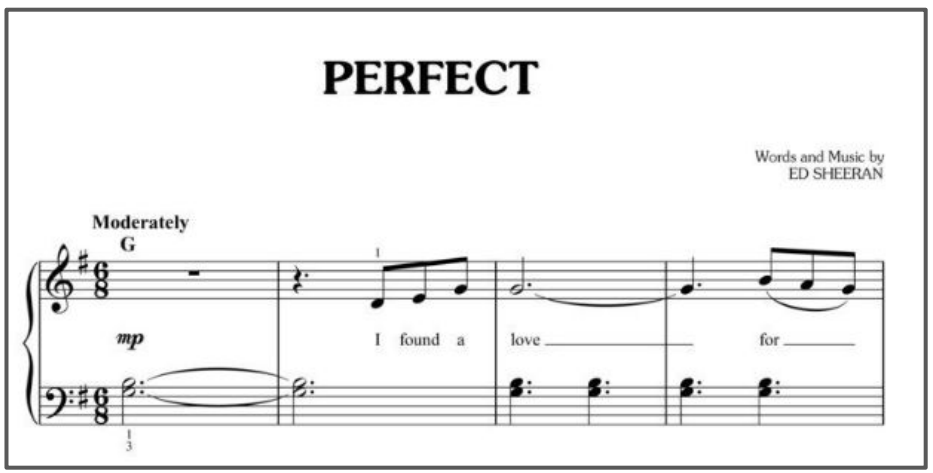
Irene Li

13 Sept. 2018

Learning the Piano…

From accordion to piano…





Motivation: lack of training data

Well-trained Models
NEWS ONLY

New
Domain?

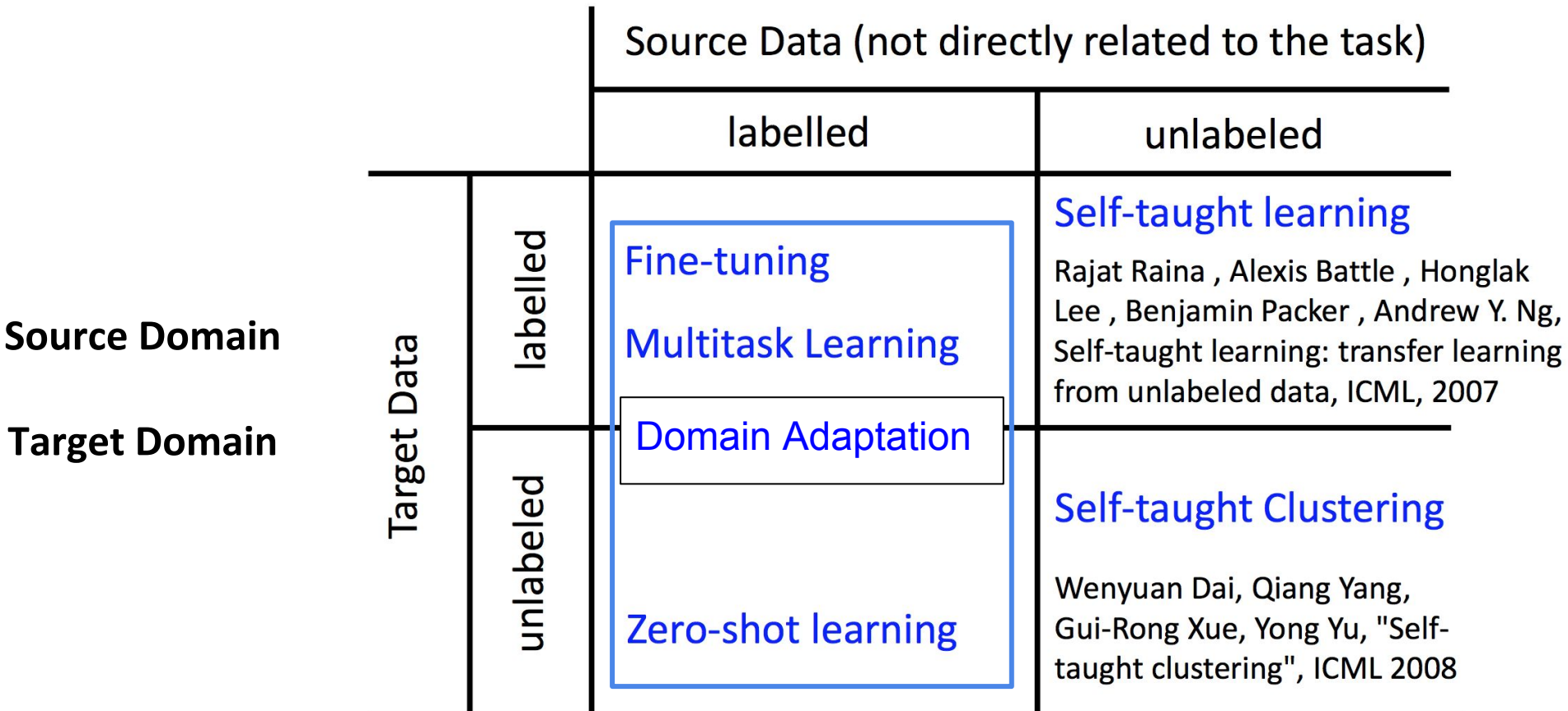
Motivation: performance drop

TRAIN → TEST

Fail

[An Introduction to Transfer Learning and Domain Adaptation]

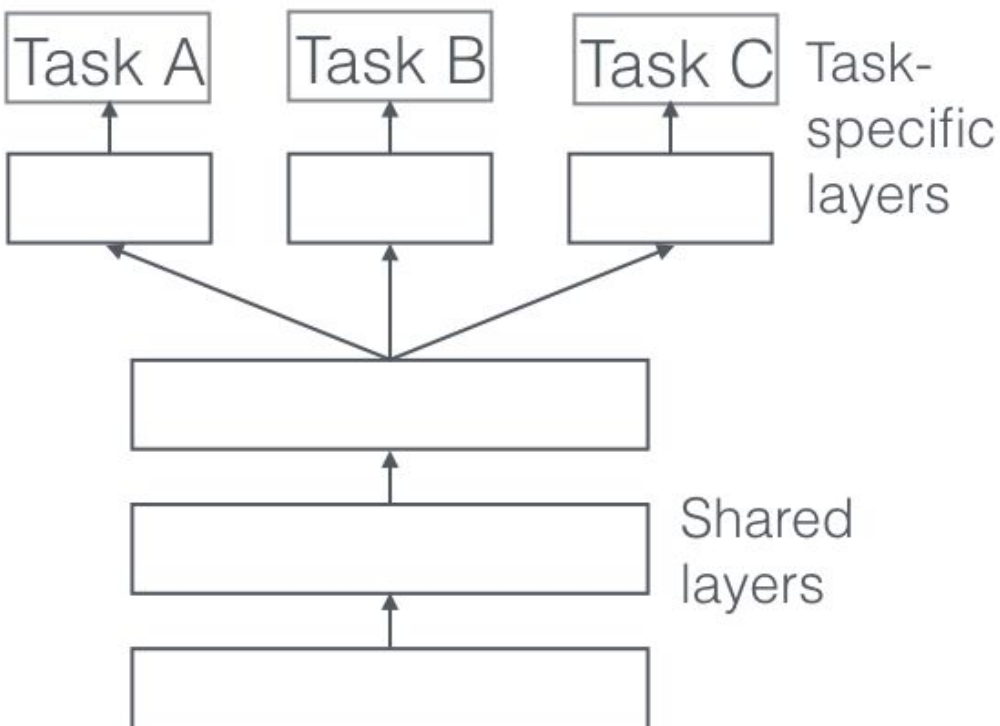
Transfer Learning: Overview



Transfer Learning Tutorial (Hung-yi Lee)

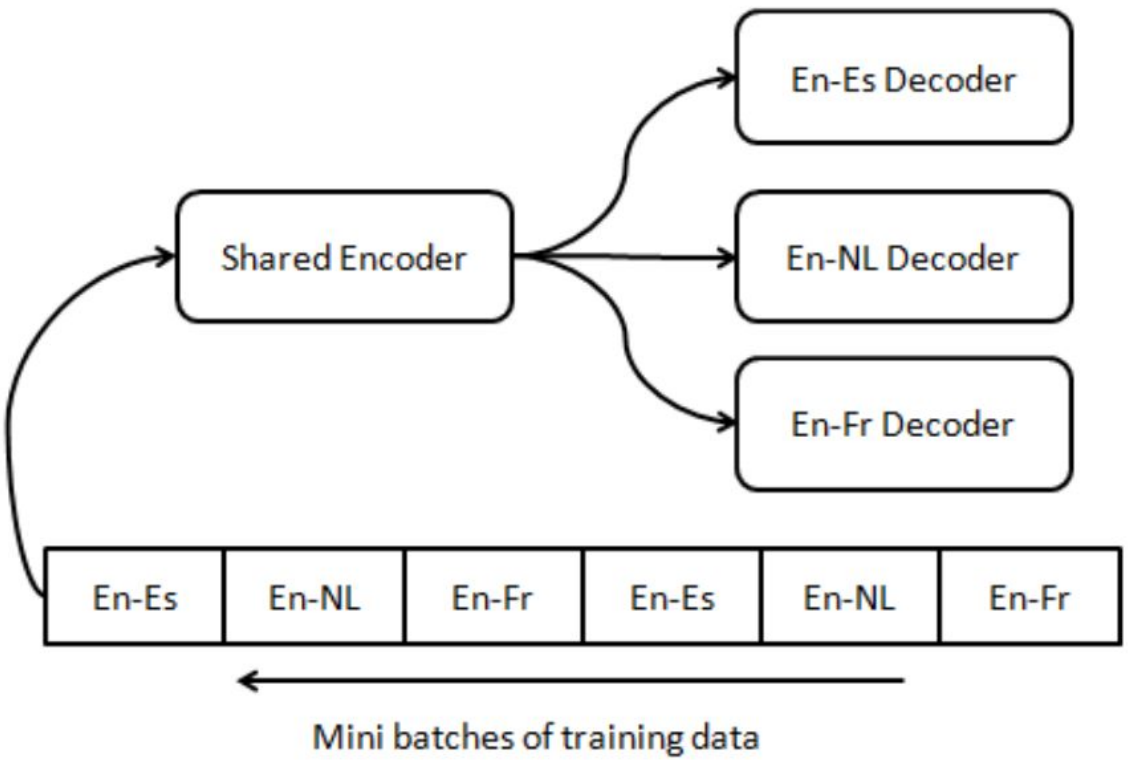
Multi-task Learning

Share layers, change task-specific layers



Multi-Task Learning for Multiple Language Translation

Multi-task Learning



Multi-Task Learning for Multiple Language Translation

Zero-shot learning

Unseen traning samples

-Same distribution:

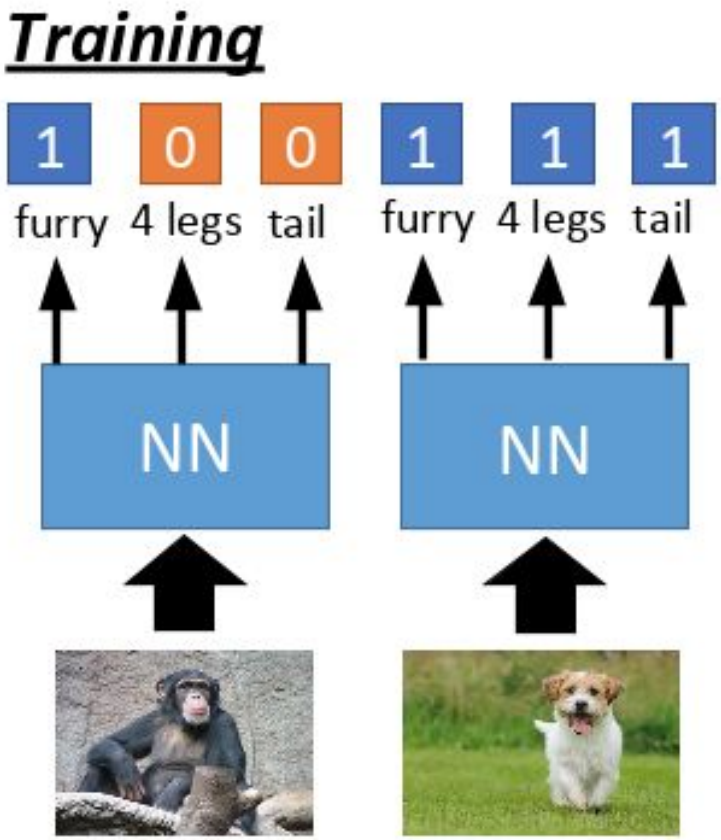
Share similar features!

-Different tasks

In my training dataset:
Chimp , Dog

In my testing dataset:
Fish

A toy example...



Database

attributes

class

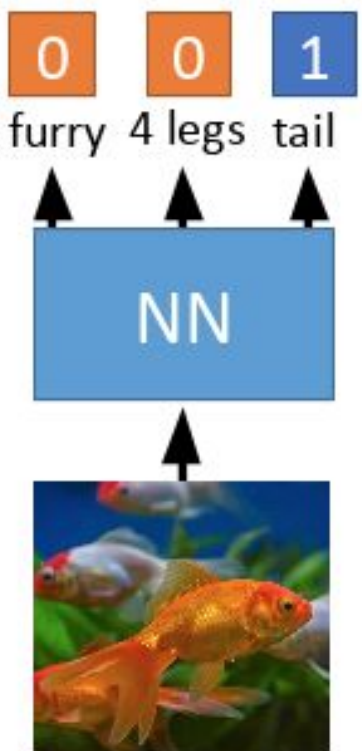
	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

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

class

attributes

	furry	4 legs	tail	...
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
...				

sufficient attributes for one to one mapping

http://speech.ee.ntu.edu.tw/~tlkagk/

Zero-shot Learing with Machine Translation

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

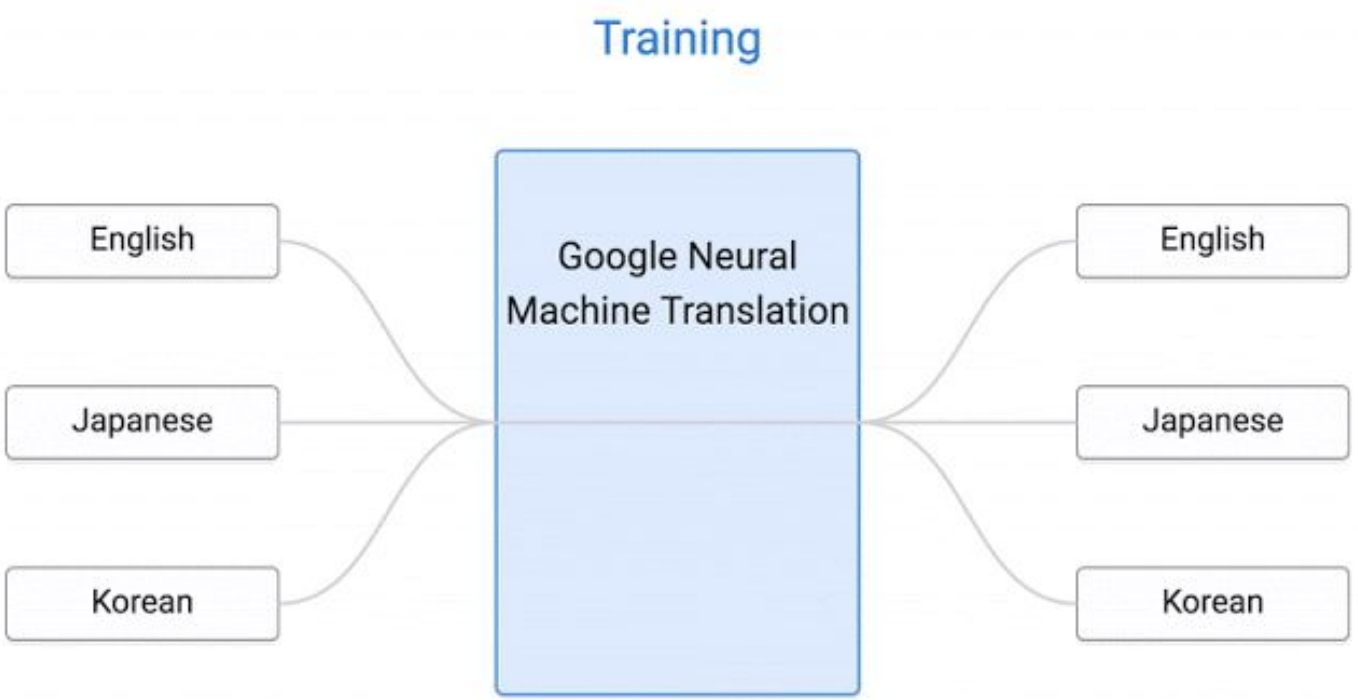
Training (blue):

EN <->JP

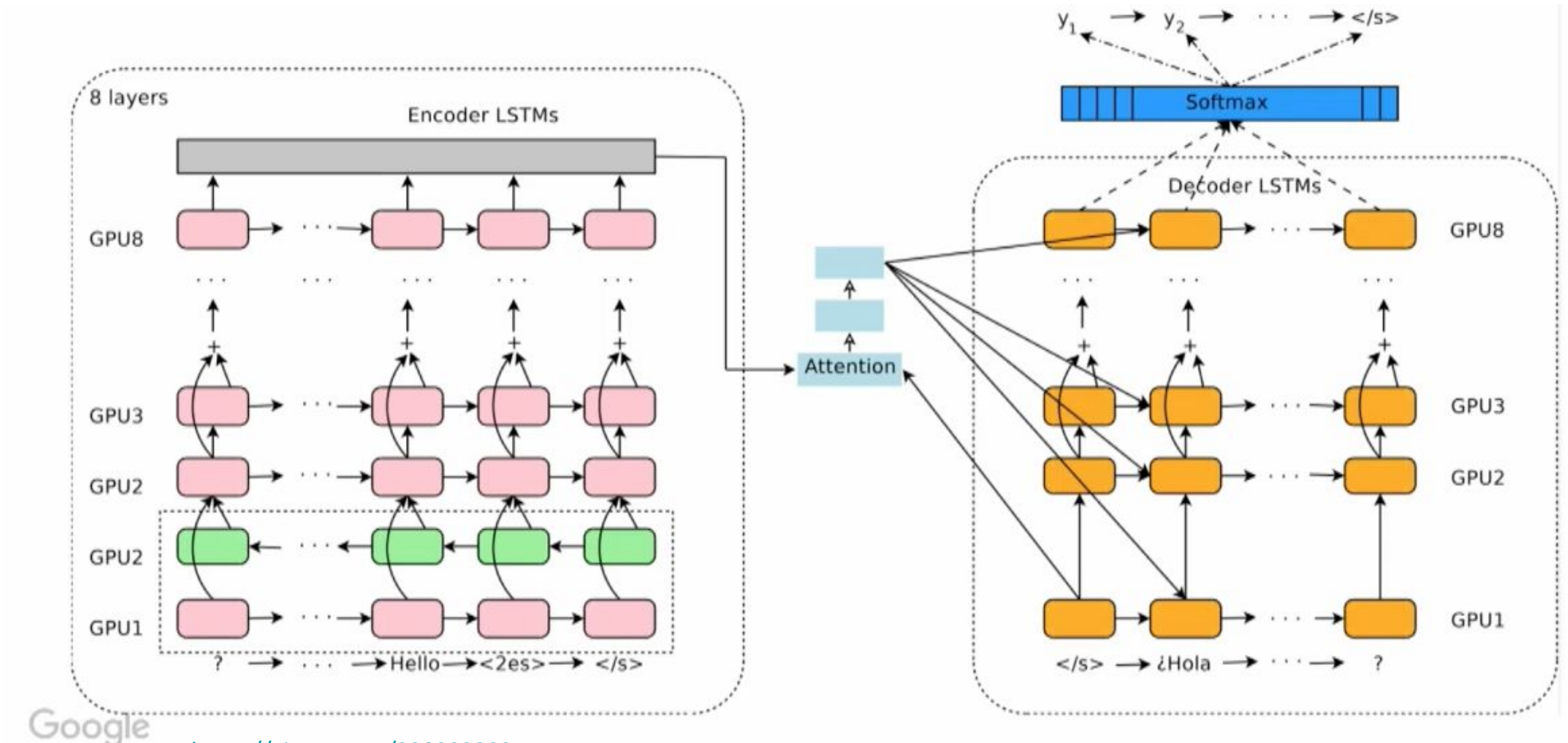
EN <->KO

Goal (orange):

JP <->KO



Google's Multilingual Neural Machine Translation Model

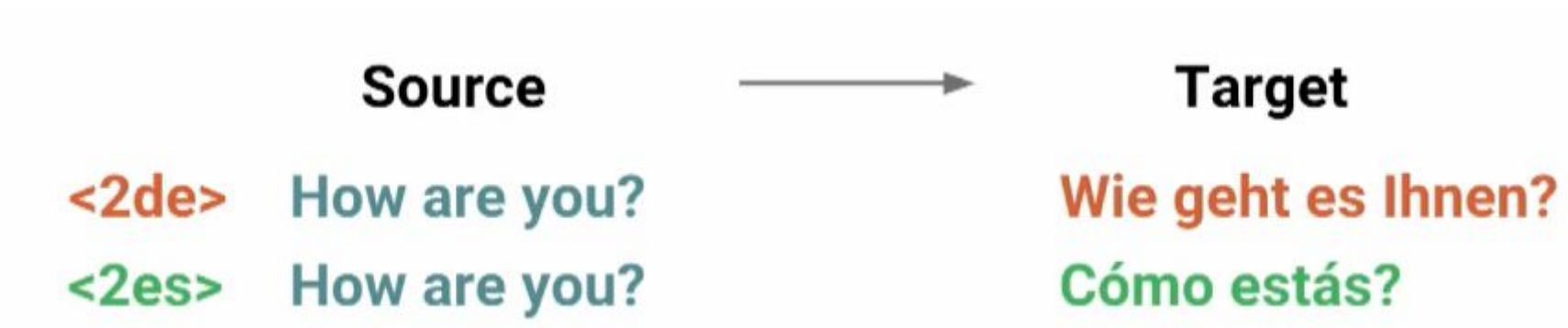


https://vimeo.com/238233299

Simple idea...

Everything is shared: encoder, decoder, attention model.

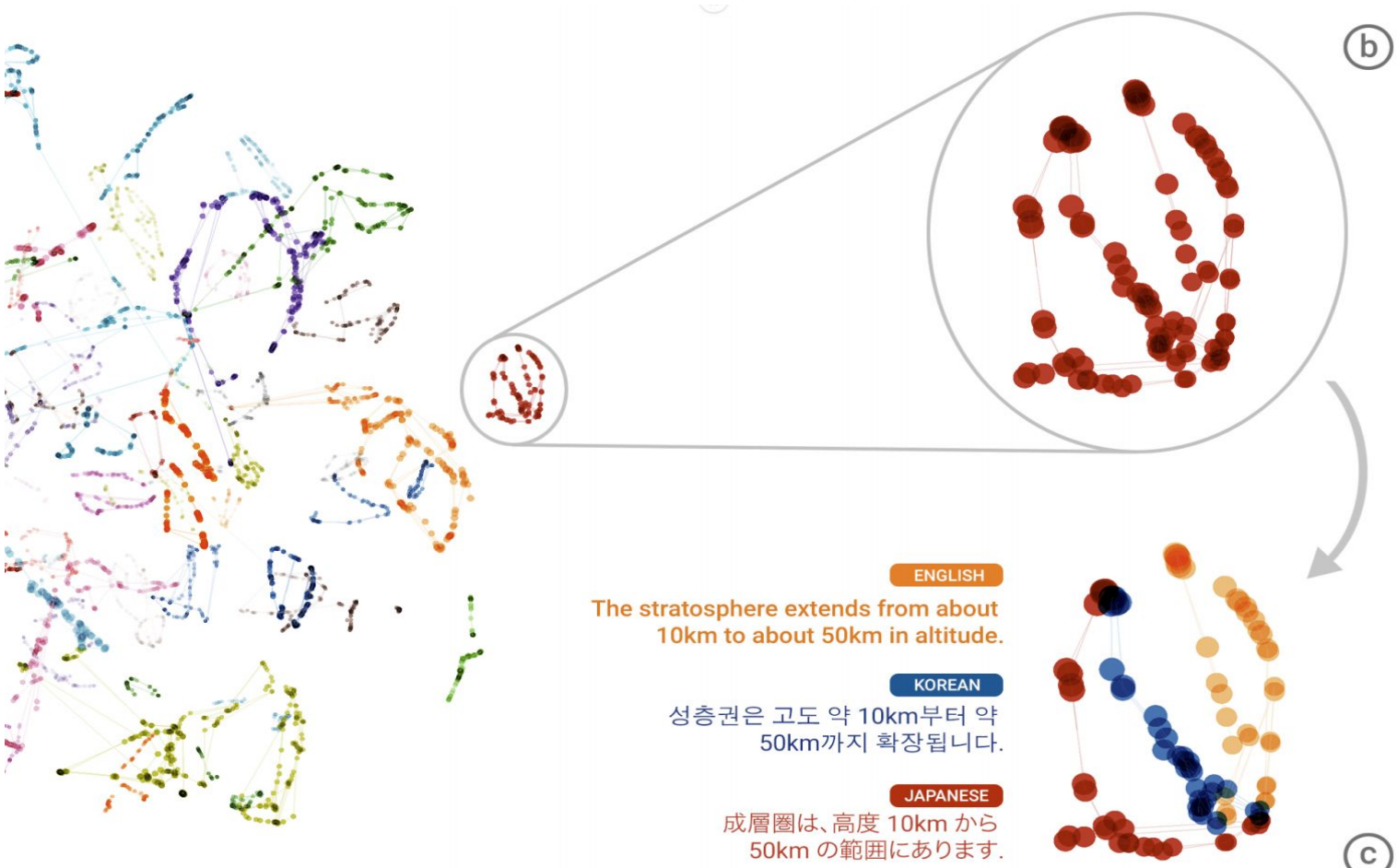
Prepend source to indicate target language:



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

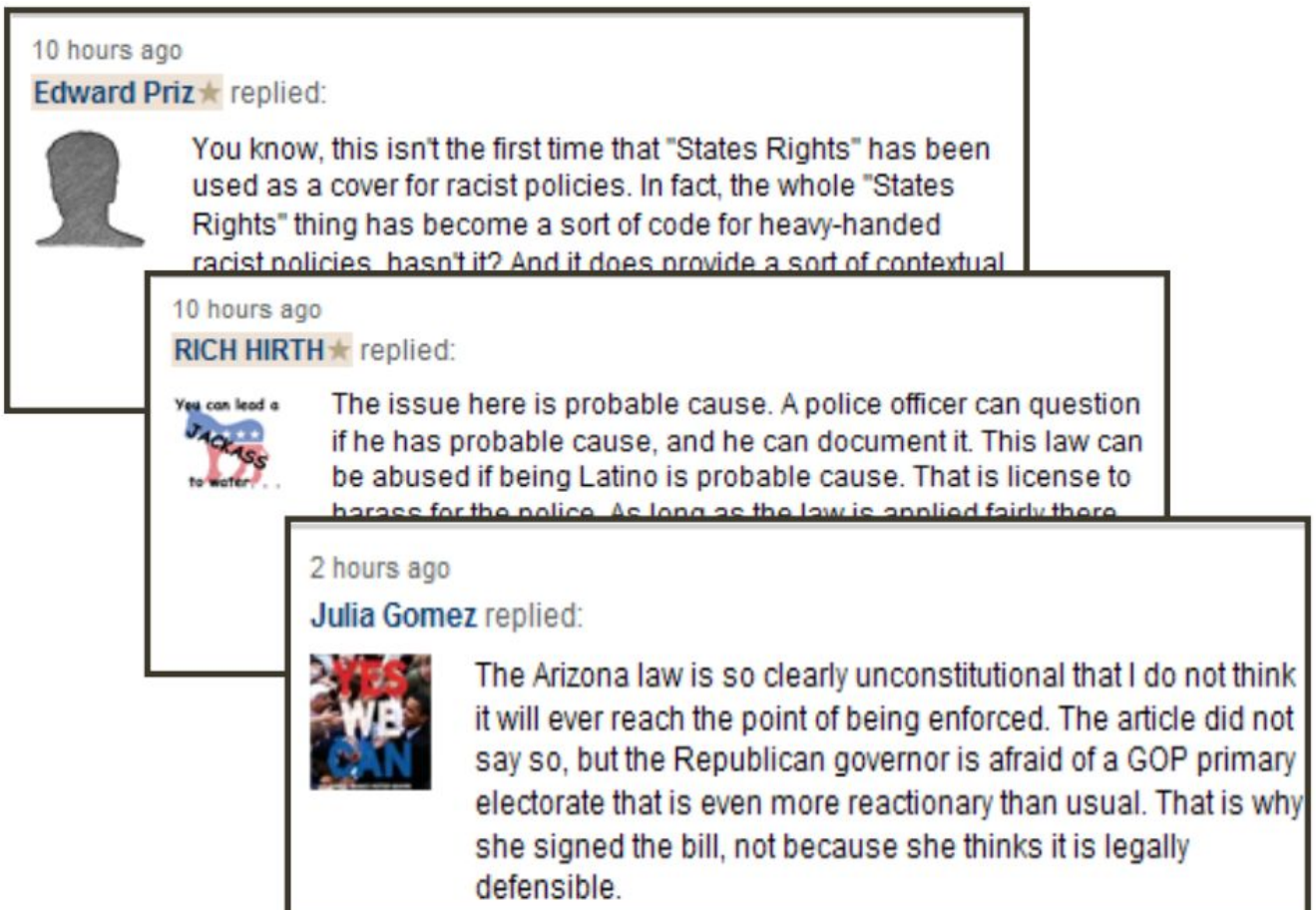
https://vimeo.com/238233299

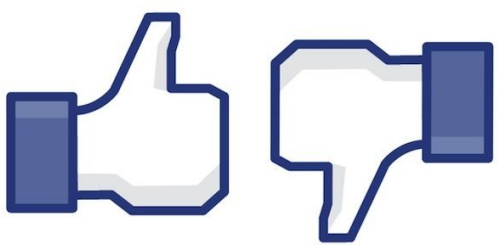
Sentence T-SNE Visualization



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

Cases in NLP? Different Distributions…





https://drive.google.com/file/d/1jEWXZzayUlXj0TDZZGX9iKsDIFxoJKUw/view

Cases in NLP? Different Distributions...

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

- Source specific: *compact, sharp, blurry*.
- Target specific: *hooked, realistic, boring*.
- Domain independent: *good, excited, nice, never_buy, unhappy*.

Other cases in NLP: Q&A

	Source Dataset	Target Dataset	
	MovieQA	TOEFL	MCTest
S	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 ...	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 ...	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 ...
Q	What does Snape tell Harry before he dies?	Why does the professor meet with the student?	What did James do after he ordered the fries?
$\mathbf{C}_1$	To bury him in the forest	To find out if the student is interested in taking part in a genetics project	went to the grocery store
$\mathbf{C}_2$	That he always respected him	To discuss the student's experiment on the taste perception	**went home without paying**
$\mathbf{C}_3$	To remember to him for the good deeds	**To determine if the student has selected an appropriate topic for his class project**	ate them
$\mathbf{C}_4$	**To take his memories to the Pensieve**	To explain what the student should focus on for his class presentation	made up his mind to be a better turtle
$\mathbf{C}_5$	To write down his memories in a book		

Supervised and Unsupervised Transfer Learning for Question Answering ACL 2018

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...

Reference Abstract
joanne starkey reviews her dozen favorite italian restaurants on long island .

Generated Abstract
laura [UNK] reviews italian restaurant in rockville centre, ny .

Domain shifts:

Vocabulary

Wrong Info

Full Article: https://www.nytimes.com/2000/11/12/nyregion/in-an-ocean-of-marinara-sauce-12-places-where-dining-is-more-than-moltobene.html

Domain Adaptation In NLP?

- Setting
  - Source domain: $D_S=\{(\boldsymbol{x}_S,y_S)\}$
  - Target domain: $D_T=\{(\boldsymbol{x}_T)\}$ y_T
- Problems in NLP
  - Frequency bias: $P(\boldsymbol{x}_S)\neq P(\boldsymbol{x}_T)$
    - Different frequencies: same word in different domains
  - Context feature bias: $P(y_S|\boldsymbol{x}_S)\neq P(y_T|\boldsymbol{x}_T)$
    - “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)

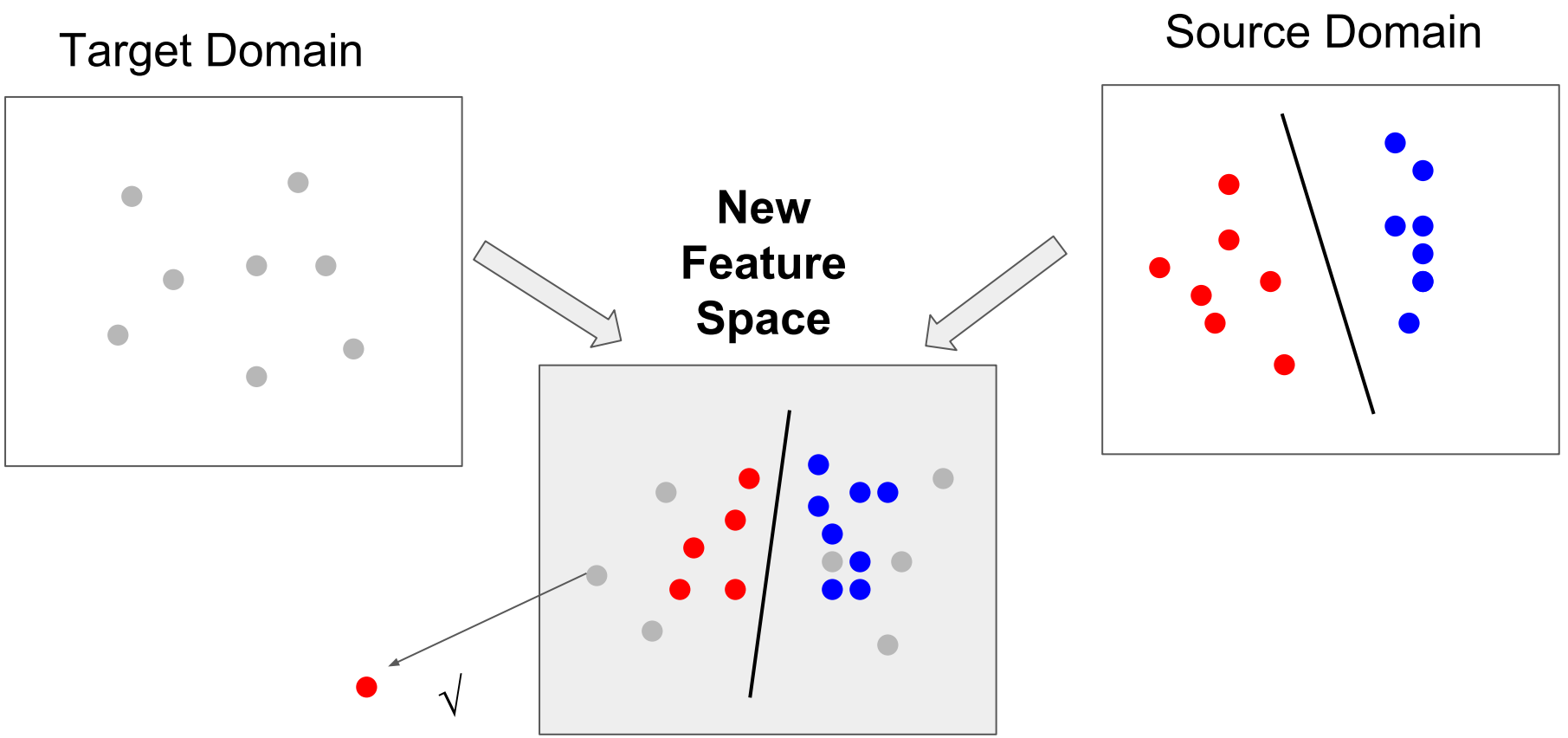
Model-based methods:

- Paramter 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

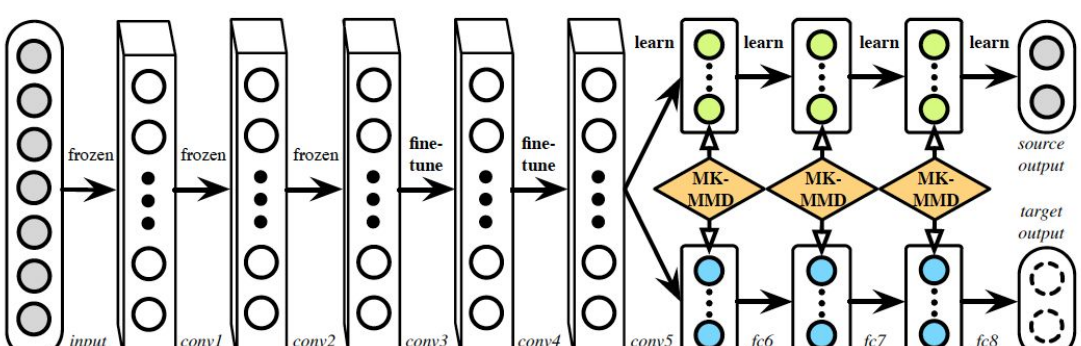
Setting:

Source domain with labels

Target domain **without** labels

Model:

VGG net loss + domain loss



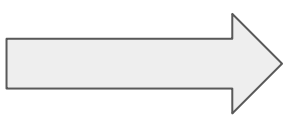
Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." ICML (2015).

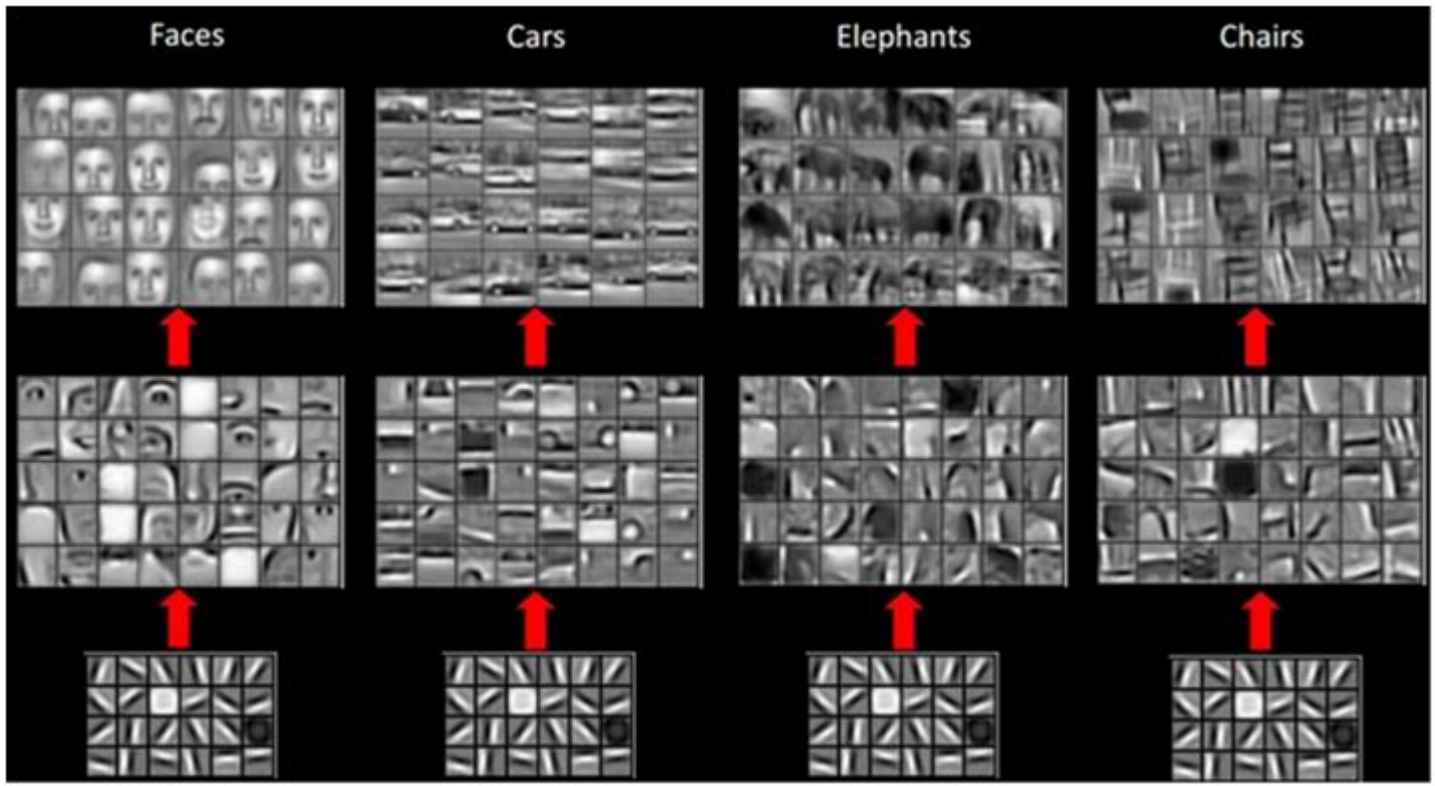
CNN Features

Lower level

fetures

are shared..



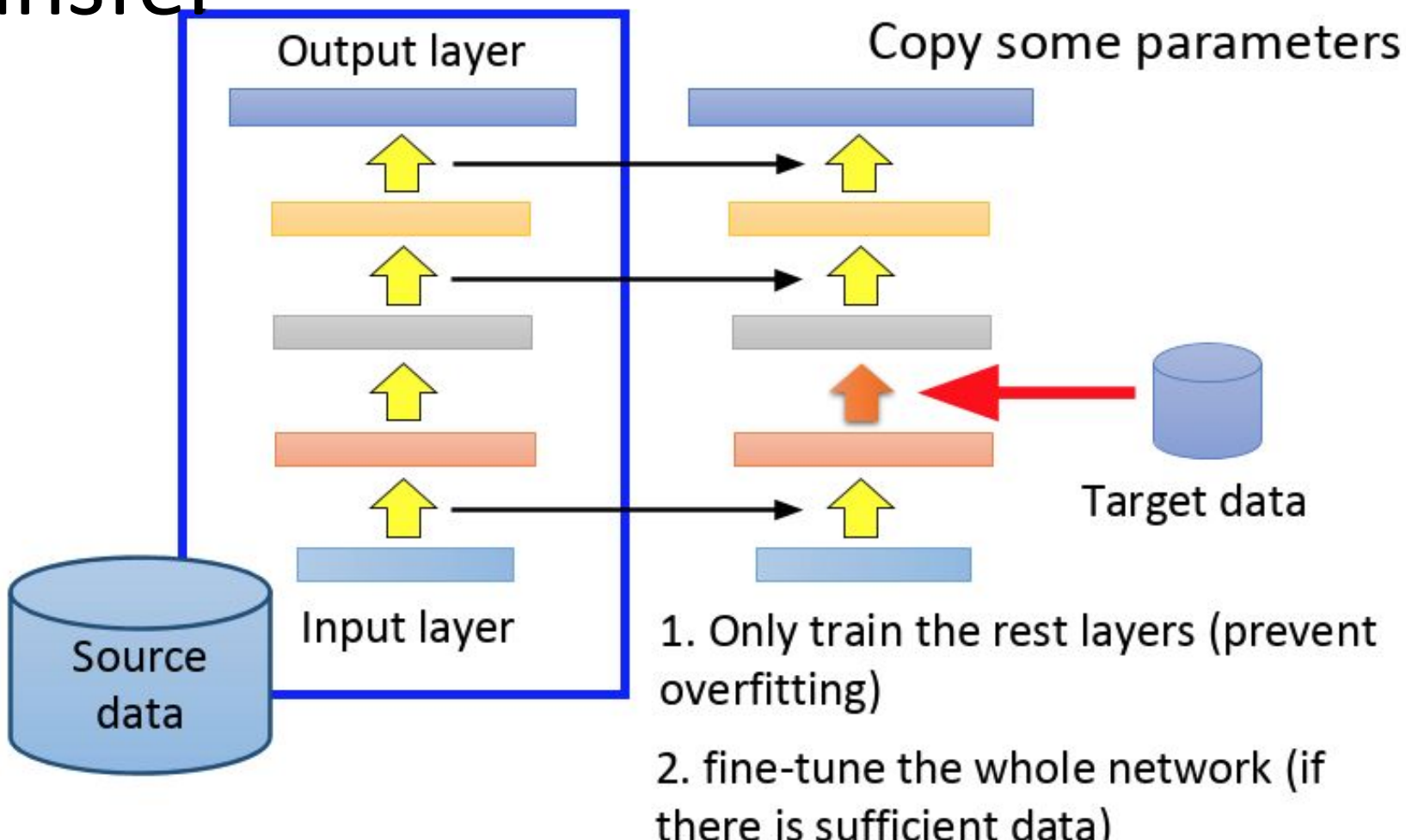


http://cs231n.github.io/convolutional-networks/
https://stats.stackexchange.com/questions/146413/why-convolutional-neural-networks-belong-to-deep-learning

Layer Transfer

Deep

Models...

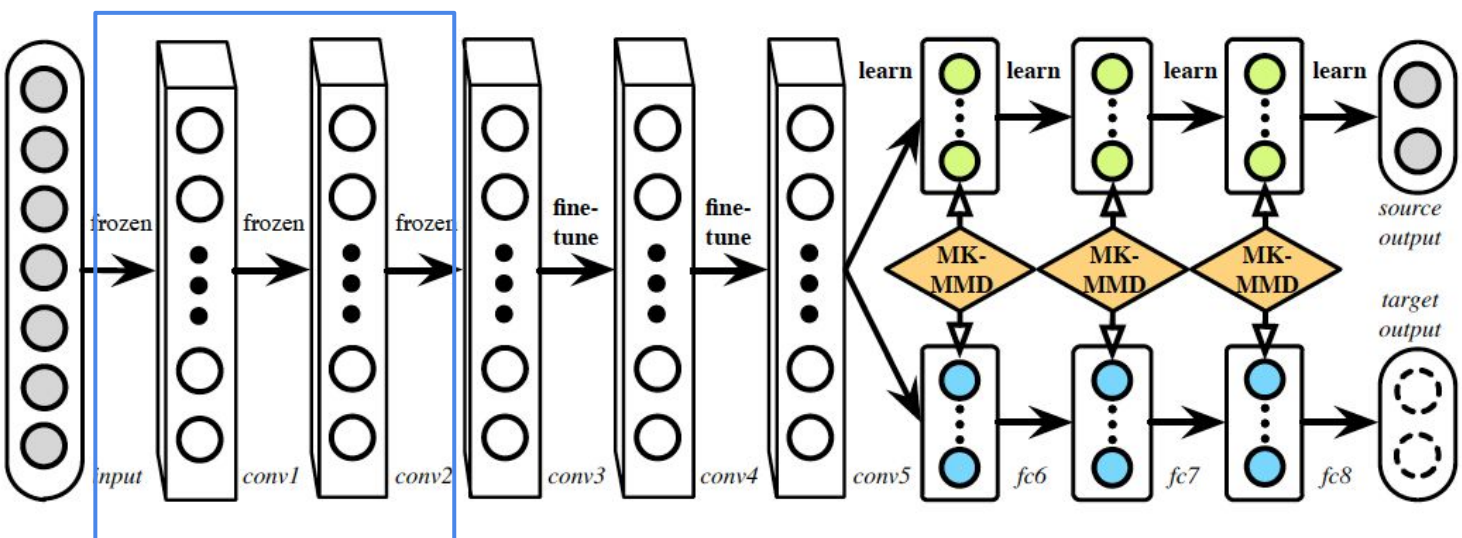


http://speech.ee.ntu.edu.tw/~tlkagk/

Layer transfer in CNNs...

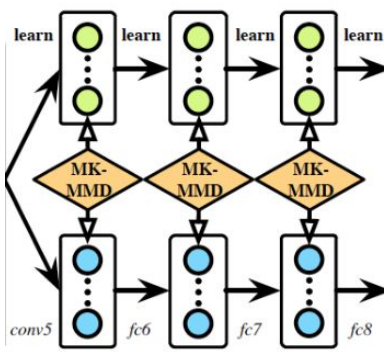
Freeze the first few layers, they are shared…

We train domain-specific layers!



Loss function: discriminativeness and domain invariance

$$\min_{\Theta} \frac{1}{n_a} \sum_{i=1}^{n_a} J\left(\theta\left(\mathbf{x}_i^a\right), y_i^a\right) + \lambda \sum_{\ell=l_1}^{l_2} d_k^2\left(\mathcal{D}_s^\ell, \mathcal{D}_t^\ell\right)$$



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

Multi-kernel Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD)

Two-sample problem (unknown p and q):

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

Maximum Mean Discrepancy (Muller, 1997):

Map the layers into a Reproducing Kernel Hilbert Space H with kernel function k:

$$MMD^2(p,q) = ||p-q||_H^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i,x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i,y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i,y_j)$$

O(n^2)

MK-MMD: Optimization

Unbiased estimation in O(n):

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

$$h(z_i, z_j) \coloneqq k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$$

Kernel:

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

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

Multi-kernel:

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

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?

$$\min_{\Theta} \frac{1}{n_a} \sum_{i=1}^{n_a} J\left(\theta\left(\mathbf{x}_i^a\right), y_i^a\right) + \lambda \sum_{\ell=l_1}^{l_2} d_k^2\left(\mathcal{D}_s^\ell, \mathcal{D}_t^\ell\right)$$

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!

Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)

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)

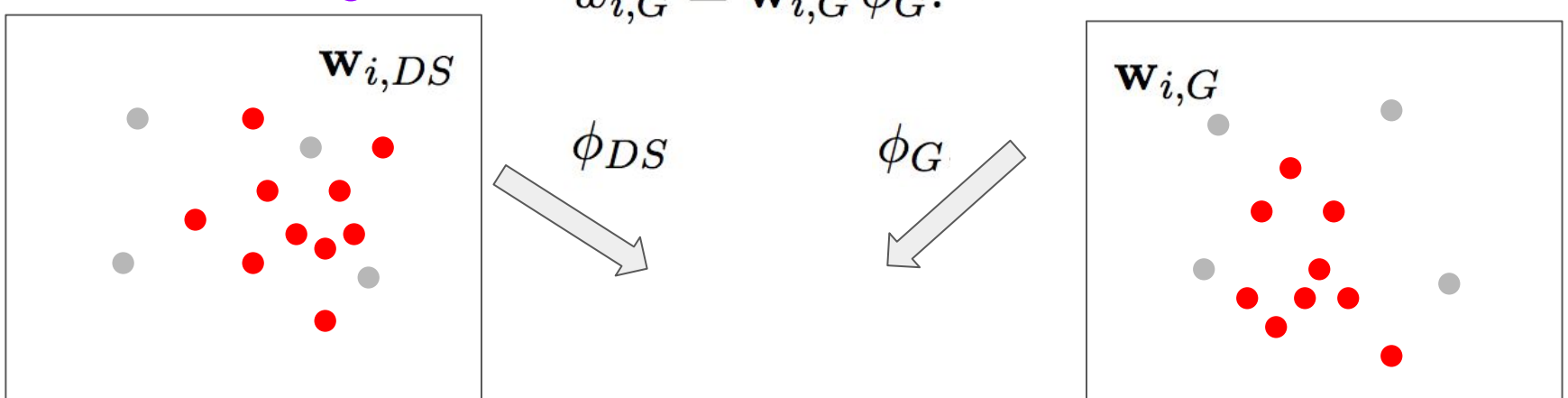
Intuition: Combine two embedding feature space

Domain Specific Embedding

$$\bar{w}_{i,DS} = \mathbf{w}_{i,DS}\,\phi_{DS}$$

$$\bar{w}_{i,G} = \mathbf{w}_{i,G}\,\phi_{G}.$$

Generic Embedding



CCA maximizes the correlation between $\bar{w}_{i,DS}$ and $\bar{w}_{i,G}$ to obtain ϕ_{DS} and ϕ_G such that

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

Find out ϕ_{DS} ϕ_G

Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)

Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA): X = (X1, ..., Xn) and Y = (Y1, ..., Ym) of 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.

$$\bar{w}_{i,DS} = \mathbf{w}_{i,DS}\phi_{DS}$$

LSA Embedding * Mapping

$$\bar{w}_{i,G} = \mathbf{w}_{i,G}\phi_{G}.$$

GloVe Embedding * Mapping

Domain-Adapted Embedding:

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

$$\hat{\mathbf{w}}_{i,DA} = \frac{1}{2}\bar{\mathbf{w}}_{i,DS} + \frac{1}{2}\bar{\mathbf{w}}_{i,G}$$

Final Embedding

Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)

Feature-based Method: share word embeddings

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

Result on Yelp Dataset

Yelp: 1000 balanced

Restaurant reviews

Tokens: 2049

[Domain Adapted Word Embeddings for Improved Sentiment Classification ACL 2018 (short)]

Feature-based Method: share word embeddings

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

Result on A-CHESS Dataset

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?

Paper 3: Model-based Method - pre-train and fine tune

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)

	MovieQA
S	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 ...
Q	What does Snape tell Harry before he dies?
$\mathbf{C}_1$	To bury him in the forest
$\mathbf{C}_2$	That he always respected him
$\mathbf{C}_3$	To remember to him for the good deeds
$\mathbf{C}_4$	**To take his memories to the Pensieve**
$\mathbf{C}_5$	To write down his memories in a book

Dataset example

Supervised and Unsupervised Transfer Learning for Question Answering. NAACL, 2018

Model-based Method: pre-train and fine-tune

Datasets:

(Source) MovieQA

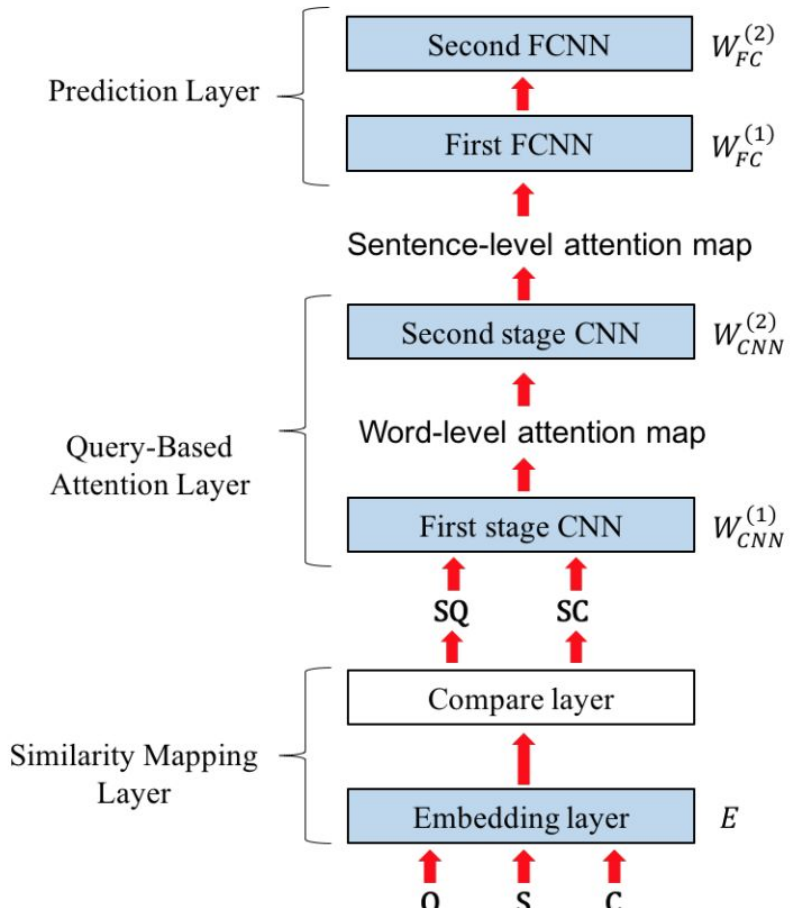
(Target 1) TOEFL listening comprehension

(Target2) 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

Model	Training	TOEFL		MCTest	
		manual	ASR	MC160	MC500
QACNN	(a) Target Only	48.9	47.5	57.5	56.4
	(b) Source Only	51.2	49.2	68.1	61.5
	(c) Source + Target	52.5	49.7	72.1	64.6
	(d) Fine-tuned (1)	53.4 (4.5)	51.5 (4.0)	**76.4** (18.9)	68.7 (12.3)
	(e) Fine-tuned (2)	**56.1** (7.2)	**55.3** (7.8)	73.8 (16.3)	**72.3** (15.9)
	(f) Fine-tuned (all)	56.0 (7.1)	55.1 (7.6)	69.3 (11.8)	67.7 (11.3)

Without pre-train on MovieQA

Train only on MovieQA

Train on both!

Results

[Supervised and Unsupervised Transfer Learning for Question Answering. NAACL, 2018]

Paper 4: Model-based Method - domain mixing

Effective Domain Mixing for Neural Machine Translation

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

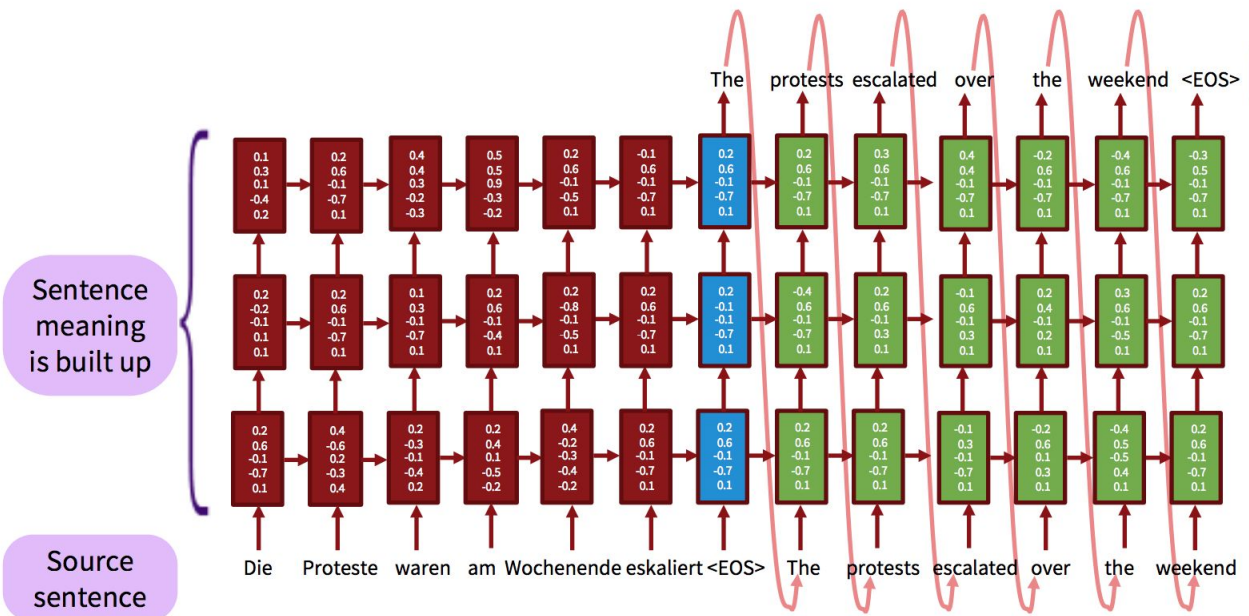
Three translation tasks:

EN-JA, EN-ZH, EN-FR

Heterogeneous corpora: News vs TEDtalks

Pryzant et al Effective Domain Mixing for Neural Machine Translation 2017

Recall NMT...



$$\log p(\boldsymbol{y}|\boldsymbol{x})$$

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

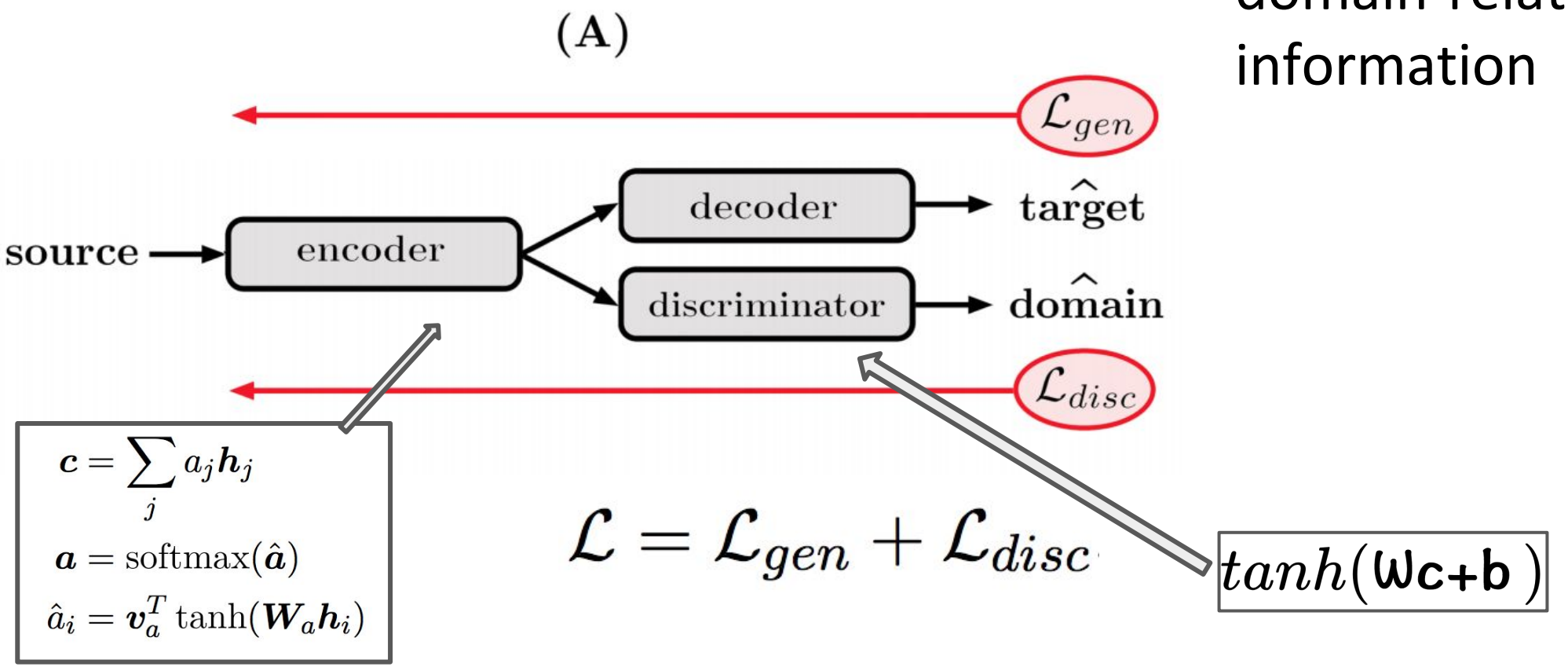
A deep recurrent neural network

x: source input
y: target output

Slide page from: https://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

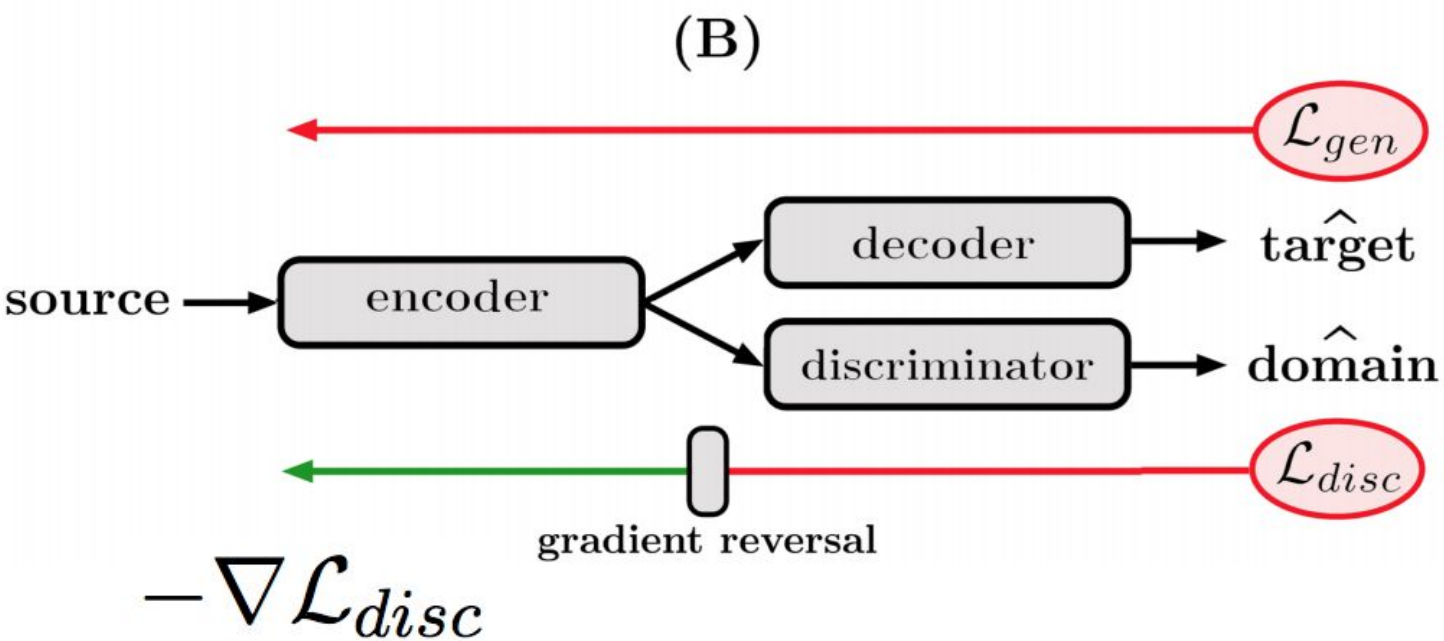
Discriminative Mixing

encoder: domain-related information



Pryzant et al [Effective Domain Mixing for Neural Machine Translation](#) 2017

Adversarial Discriminative Mixing

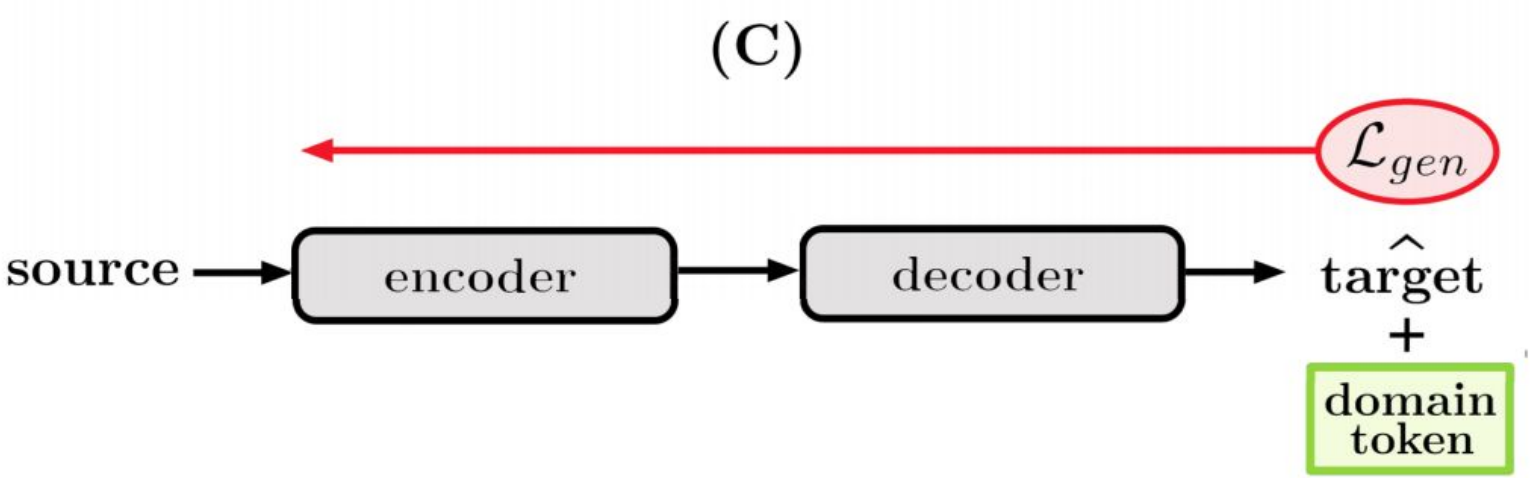


such representations lead to better generalization across domains...

$$\mathcal{L} = \mathcal{L}_{gen} + \mathcal{L}_{disc}$$

Pryzant et al Effective Domain Mixing for Neural Machine Translation 2017

Target Token Mixing



regularizing effect on enc and dec

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 avaliable at: https://github.com/rpryzant/proxy-a-distance

Results

Baseline is mixing samples...

Mixing will help...

Similar domains

Lanuage	Domain 1	Domain 2	$\hat{d}_A$
Japanese	ASPEC	SubCrawl	1.89
Chinese	News	TED	1.73
French	Europarl	OpenSubs	1.23

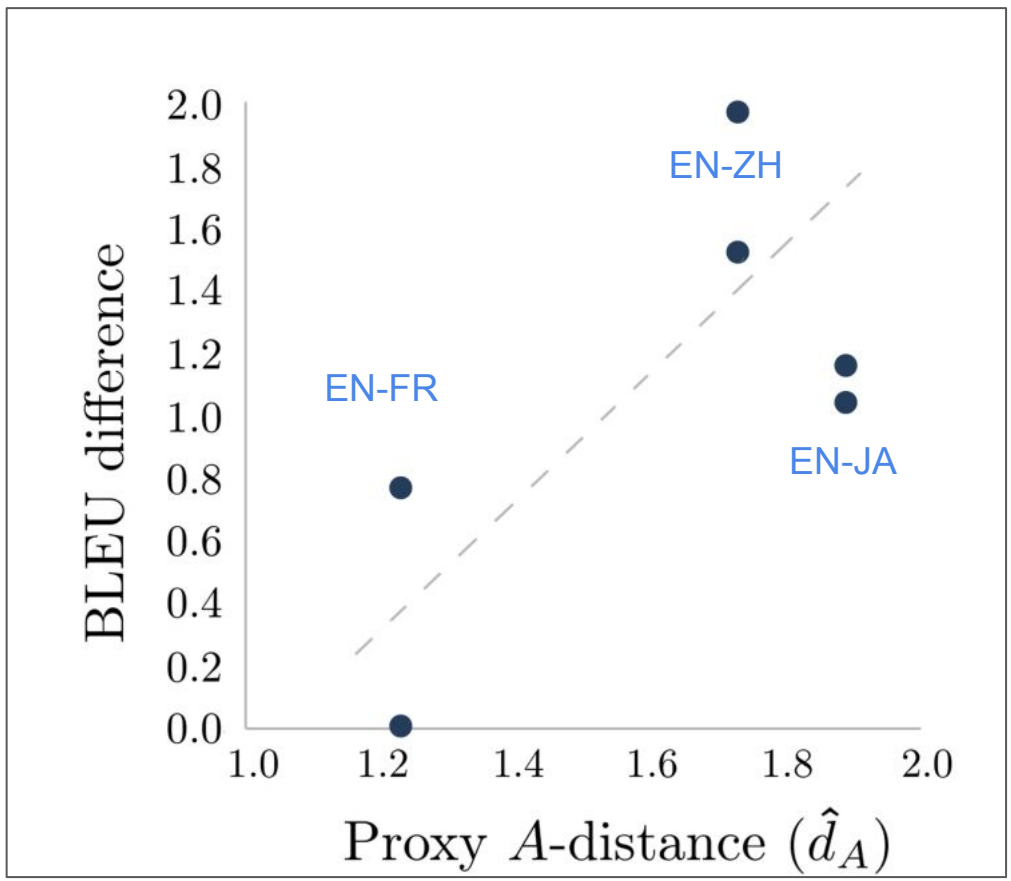
Table 1: Proxy A-distances ($\hat{d}_A$) for each domain pair.

EN-JA Model	ASPEC	SubCrawl
ASPEC	38.87	3.85
SubCrawl	2.74	16.91
ASPEC + SubCrawl	33.85	14.34
Discriminator	35.01	**15.38**
Adv. Discriminator	29.87	13.31
Target Token	**35.05**	14.92
EN-FR Model	**Europarl**	**OpenSubs**
Europarl	34.51	13.36
OpenSubtitles	13.12	15.2
Europarl + OpenSubs	38.26	27.9
Discriminator	39.03	**27.91**
Adv. Discriminator	38.38	25.67
Target Token	**39.1**	25.32
EN-ZH Model	**News**	**TED**
News	12.75	3.12
TED	2.79	8.41
News + TED	11.36	6.67
Discriminator	**12.88**	**8.64**
Adv. Discriminator	12.15	8.16
Target Token	11.98	7.69

BLEU Scores

Pryzant et al Effective Domain Mixing for Neural Machine Translation 2017

BLUE scores and PAD



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

The more diverge the more discriminative model helps.

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.

Current Research

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

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

! 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)?

References

Muandet, Krikamol, et al. "Kernel mean embedding of distributions: A review and beyond." Foundations and Trends® in Machine Learning 10.1-2 (2017): 1-141.

Gretton, Arthur, et al. "A kernel method for the two-sample-problem." Advances in neural information processing systems. 2007.

Mou, Lili, et al. "How transferable are neural networks in nlp applications?." arXiv preprint arXiv:1603.06111 (2016).

Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2010): 1345-1359.

Sun, Baochen, Jiashi Feng, and Kate Saenko. "Return of frustratingly easy domain adaptation." AAAI. Vol. 6. No. 7. 2016.

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