Transfer Learning for Speech and Language Processing

Dong Wang and Thomas Fang Zheng

CSLT@Tsinghua

2015/12/19

Content

- Transfer learning review
  - Transfer learning methods
  - Transfer learning in deep era
- Transfer learning in speech processing
  - Cross-lingual transfer
  - Speaker adaptation
  - Model transfer
- Transfer learning in language processing
  - Cross-lingual transfer
  - Cross-domain transfer
  - Model transfer
- Perspective and conclusions

Goal of the overview

- ML perspective
  - Categorize existing TL methods
  - Highlight TL in Deep learning
- SLA perspective
  - Highlight TL applications in SLA
  - Promote further research & application
- Associated paper
  - http://arxiv.org/abs/1511.06066

Content

- Transfer learning review
  - Transfer learning methods
  - Transfer learning in deep era
- Transfer learning in speech processing
  - Cross-lingual transfer
  - Speaker adaptation
  - Model transfer
- Transfer learning in language processing
  - Cross-lingual transfer
  - Cross-domain transfer
  - Model transfer
- Perspective and conclusions

Learning to learn from NIPS*95

NIPS*95 Post-Conference Workshop

"Learning to Learn: Knowledge Consolidation and Transfer in Inductive Systems"

[Motivation and Goals], [Submissions], [Organizers], [Schedule], [Talk Abstracts], [For More Information]

Length: 2 days

Organizers: Jonathan Baxter, Rich Caruana, Tom Mitchell, Lorien Y. Pratt, Daniel L. Silver, Sebastian Thrun

Invited Talks:

- Leo Breiman (Berkeley)
- Tom Mitchell (CMU)
- Tomaso Poggio (MIT)
- Noel Sharkey (Sheffield)
- Jude Shavlik (Wisconsin)

Motivation:

The power of tabula rasa learning is limited. As these limits become apparent, interest has increased in developing methods that capitalize on previously acquired domain knowledge. Examples of these methods include:

- using symbolic domain theories to bias connectionist networks
- using unsupervised learning on a large corpus of unlabelled data to learn features useful for subsequent supervised learning on a smaller labelled corpus
- using models previously learned for other problems as a bias when learning new, but related, problems
- using extra outputs on a connectionist network to bias the hidden layer representation towards more predictive features
- updating belief(s) from a set of priors with Bayes rule

The methods used go by many names: hints, knowledge-based artificial neural nets (KBANN), explanation-based neural nets (EBNN), multitask learning (MTL), lifelong learning, knowledge consolidation, etc. What they all have in common is the attempt to transfer knowledge from other sources to benefit the current inductive task. Potential benefits include better generalization, faster learning, and a bias towards representations that are more robust or more broadly applicable.

Goals:

http://plato.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html

Workshop on Unsupervised and Transfer Learning

and results of the UTL challenge
Saturday, July 2, 2011
Bellevue, Washington, USA

DL PROJECT

ICML 2011
BELLEVUE, WASHINGTON
Jun28 - Jul2

- ICML 2011
- Home
- Competition
- Participation
- Schedule
- Links
- Contact
- Credits

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Motivation
Intelligent beings commonly transfer previously learned "knowledge" to new domains, making them capable of learning new tasks from very few examples. In contrast, many recent approaches to machine learning have been focusing on "brute force" supervised learning from massive amount of labeled data. While this last approach makes a lot of sense practically when such data are available, it does not apply when the available training data are unlabeled for the most part. Further, even when large amounts of labeled data are available, some categories may be more depleted than others. For instance, for Internet documents and images the abundance of examples per category typically follows a power law. The question is whether we can exploit similar data (labeled with different types of labels or completely unlabeled) to improve the performance of a learning machine. This workshop will address a question of fundamental and practical interest in machine learning: the assessment of methods capable of generating data representations that can be reused from task to task. To pave the ground for the workshop, we organized a challenge on unsupervised and transfer learning.

Competition
The **unsupervised and transfer learning challenge** just started and will end April 15, 2011. The results of the challenge will be discussed at the workshop and we will invite the best entrants to present their work. Further, we intend to launch a second challenge on supervised transfer learning whose results will be discussed at NIPS 2011.
This workshop is **not limited to the competition program** that we are leading. We encourage researchers to submit papers on the topics of the workshop.

Participation
We invite contributions relevant to unsupervised learning and transfer learning (UTL), including:
- Algorithms for UTL, in particular addressing
  - Learning from unlabeled or partially labeled data.
  - Learning from few examples per class, and transfer learning

New Directions in Transfer and Multi-Task: Learning Across Domains and Tasks

News

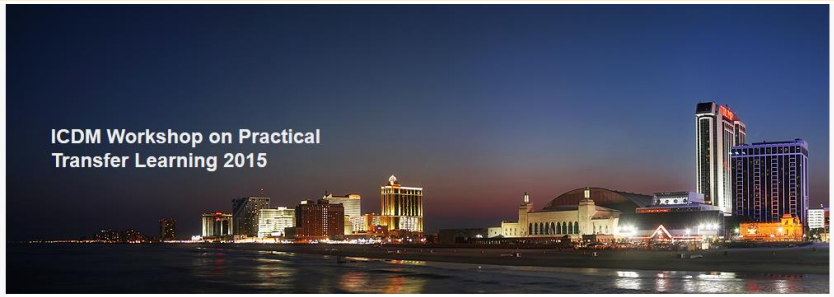
- December 8, detailed schedule online. The Workshop takes place in Harrah's Fallen+Marla room.
- November 27, accepted papers are online.
- Poster size. We follow the main conference NIPS instructions and we advice to use A0-landscape size.
- October 24, paper notifications sent to the authors. The deadline for the camera ready version will be announced soon.
- The workshop will be sponsored by EUCog - European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics.
- Two awards will be assigned to the best student papers.
- Submission deadline approaching: October 9, 2013.
  Authors' names and affiliations should be included, as the review process will not be double blind.

Description

The main objective of the workshop is to document and discuss the recent rise of new research questions on the general problem of learning across domains and tasks. This includes the main topics of transfer and multi-task learning, together with several related variants as domain adaptation and dataset bias.

In the last years there has been an increasing boost of activity in these areas, many of them driven by practical applications, such as object categorization. Different solutions were studied for the considered topics, mainly separately and without a joint theoretical framework. On the other hand, most of the existing theoretical formulations model regimes that are rarely used in practice (e.g. adaptive methods that store all the source samples).

This NIPS 2013 workshop will focus on closing this gap by providing an opportunity for theoreticians and practitioners to get together in one place, to share and debate over current theories and empirical results. The goal is to promote a fruitful exchange of ideas and methods between the different communities, leading to a global advancement of the field.



OVERVIEW
CALL FOR PAPERS
IMPORTANT DATES
SCHEDULE
PEOPLE
SUBMISSION

Overview

In many real-world applications, it is often expensive and time-consuming to collect sufficient labeled data in a new domain of interest. Instead of spending huge labeling efforts from scratch, one may prefer to effectively utilize existing well-explored data from other domains, which are referred to as "auxiliary domains" or "source domains", to help the learning task in the new domain (referred to as the "target domain"). However, traditional learning methods cannot be directly applied to learn a precise model for the target domain from the source-domain data because the data from different sources may have different statistical properties. Transfer Learning (TL), as a promising solution on the other hand, has attracted growing attention in the last two decades. Particularly, it has been successfully applied to many applications, such as text mining, video event recognition, sensor-based prediction problems, software engineering, image categorization and so forth.

One of the most challenging problems in TL is about how to reduce the difference in data distributions between domains. In the literature, many works have been proposed along this direction. For instance, some works have been focused on the domain adaptation problem where the source and target domains have data under different marginal distributions but share the same conditional distribution. Moreover, some other works have been focused on the inductive transfer learning or multi-task learning problem where the conditional distributions of the data or the predictive tasks of the source and target domains are usually different. Besides, there are also other works proposed to deal with other TL scenarios, including multi-source domain adaptation, one-shot learning, zero-shot learning, etc.

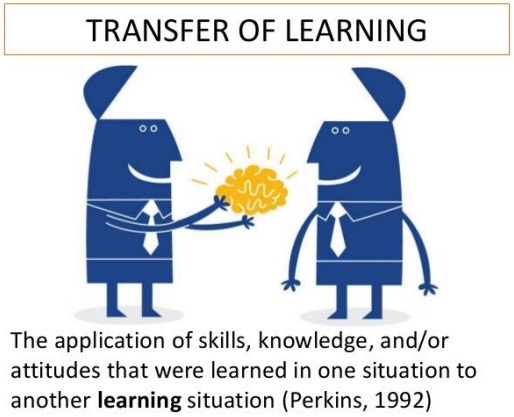
Nowadays, because of the advance of data storage and Internet technology, data become more massive, noisier and more complex. For instance, Internet itself is a very rich and huge database. The Internet data may be associated with certain structure (e.g., social networks data), may be only weakly labeled (e.g., the video and images crawled with search engine), and may be very large scale. Moreover, it is also desirable to exploit data of different formats and structures from multiple sources to further improve the learning tasks in the target domain (e.g., jointly using web images, web videos and social networks data to categorize consumer videos or images). Such new environments bring

The basic idea …

- Knowledge/statistics are ‘general’
  - Conditions, domains, languages, tasks…
- Therefore they should be re-used
  - What to re-use (e.g., data,label,model,…)?
  - What structure (e.g., prior,NN)?
  - What approach (e.g., supervised,unsupervised)?
- Advantage
  - Faster convergence
  - Less data requirement
  - Increased generalizability

Transfer learning

- The application of skills, knowledge, and/or attitudes that were learned in one situation to another learning situation – Perkins, 1992
- Methods that capitalize on previously acquired domain knowledge" – NIPS*95
- Transfer learning … refer to the situation where what has been learned in one setting is exploited to improve generalization in another setting – Bengio, 2015

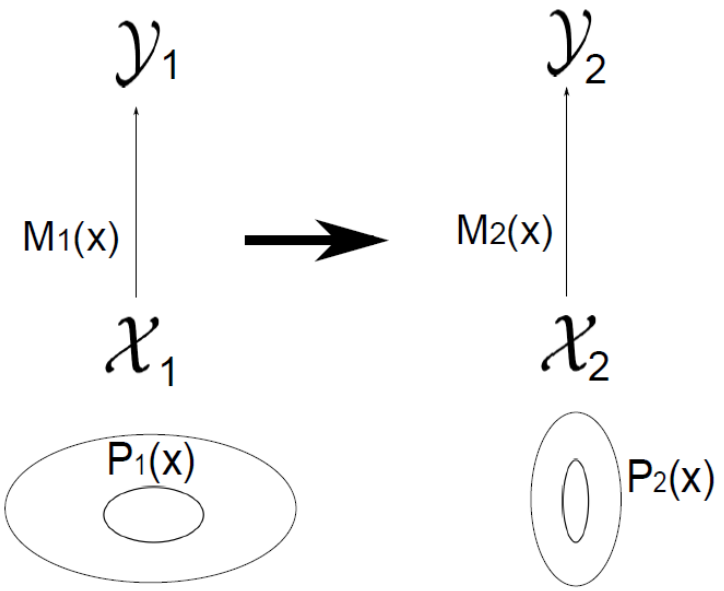


Are they the same?

- A big TL family
  - multitask learning, lifelong learning, knowledge transfer, knowledge consolidation, model adaptation, concept drift, covariance shift …
- Different authors hold different views
  - All are multitask learning (Caruana 1997).
  - Transfer learning should really transfer something (Pan and Yang, 2010).
  - Transfer learning and multitask learning are no difference (Bengio, 2015).
  - Jargon in different domains: ASR, SID, NLP

Our opinion for TL

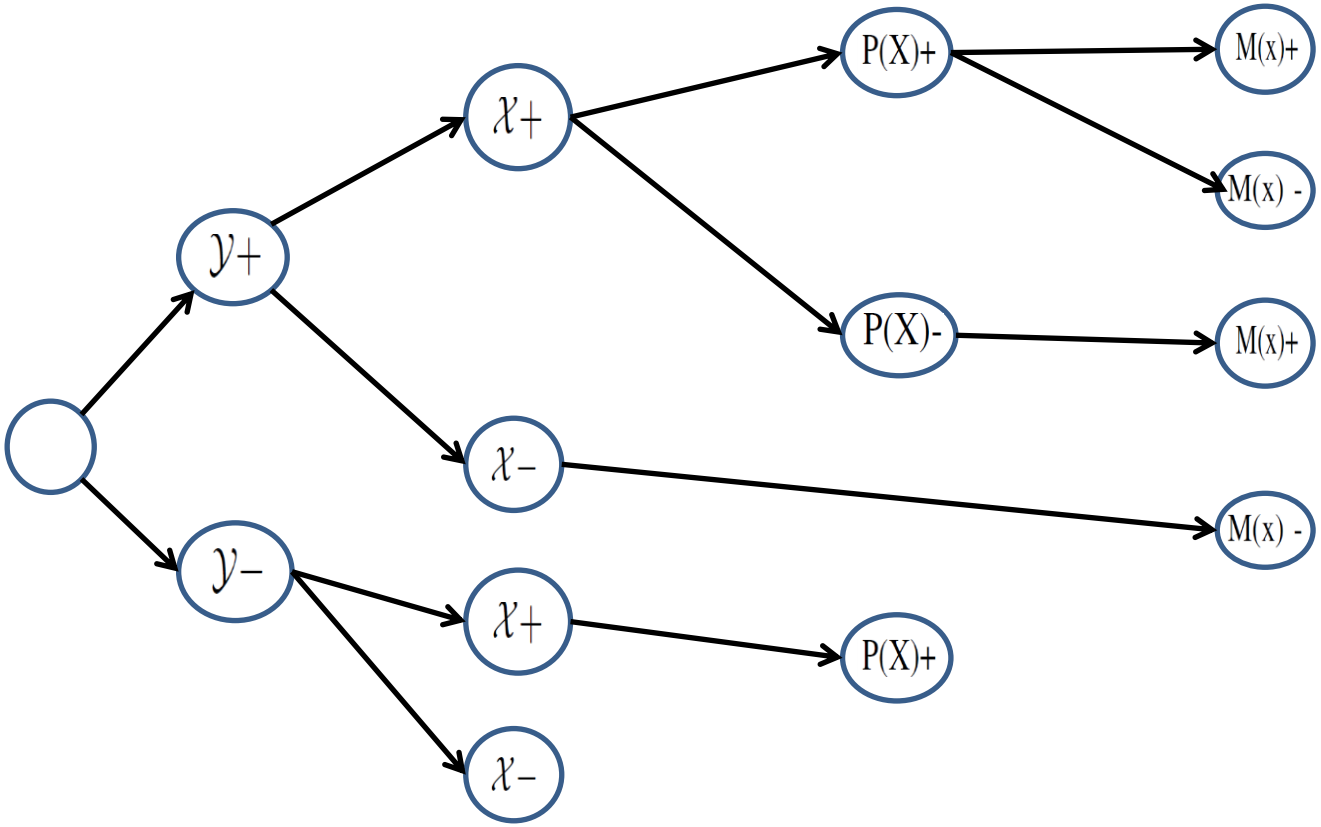
- Transfer learning is a general framework.
- Implementations in different conditions or by different ways lead to different methods.



- **Condition = Data + Task**
- **Data = Feature + Distribution**
- **Task = Label + Model**

Categorization of TL

		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]



Conventional ML

Model transfer[10]

Model Adaptation
incremental learning

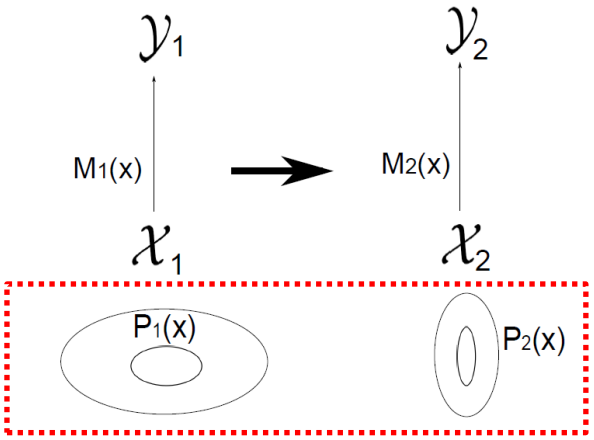
Co-training
Heterogeneous
transfer learning

Multitask learning

Analogy learning

TL method (1): Model adaptation

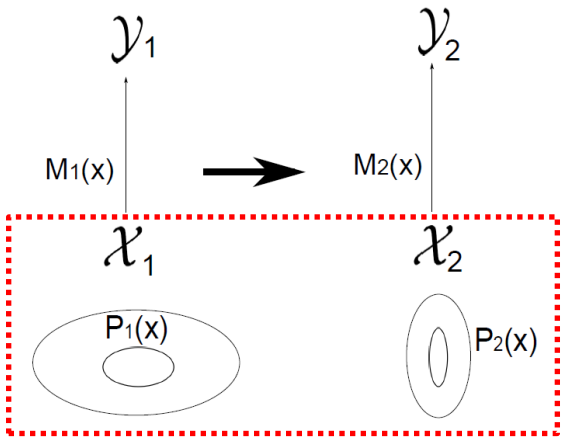
- The same task and feature, different distributions
- MAP, MLLR
- Incremental/online learning
- Unsupervised adaptation
  - Semi-supervised learning
  - Feature transform (e.g., TCA)
  - Self-taught learning



		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

TL method (2): Heterogeneous transfer learning

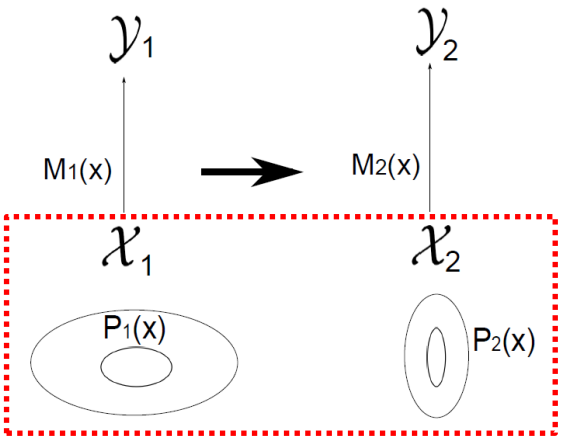
- The same task, different features
- Establish cross-domain correspondence
  - cross-domain transfer
  - Common representation
    - MF, RBM, Joint transfer
    - Deep representation



		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

TL method (3) Co-training

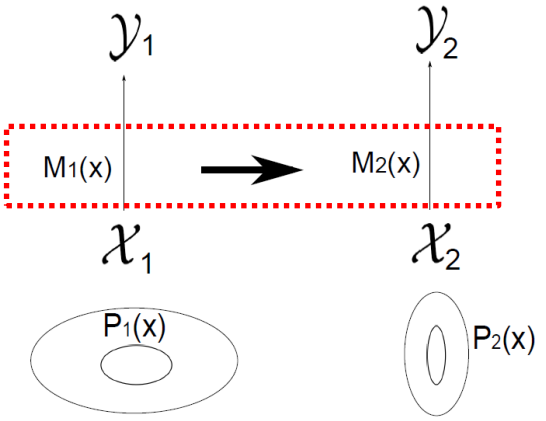
- A special heterogeneous TL
- Multi-view data for training, single-view data at run-time
- Semi-supervised learning by co-supervision



		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

TL method (4) Model transfer

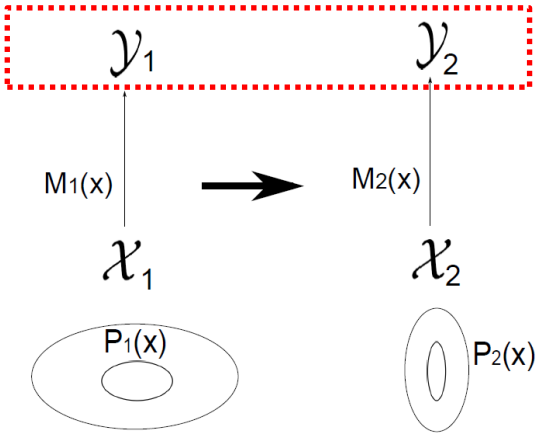
- The same feature and task, different models
- Homogeneous transfer
  - Dark knowledge distiller
- Heterogeneous transfer
  - GMM to DNN
  - LDA to DNN



		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

TL method (5) Multitask learning

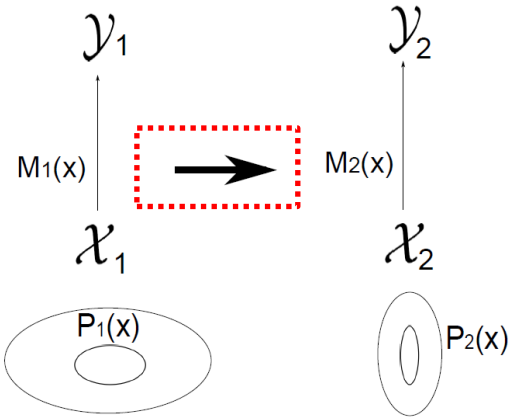
- The same feature, different tasks
- Auxiliary task helps primary task
- A related-task view
- A regularization view



		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

TL method (6) Analogy learning

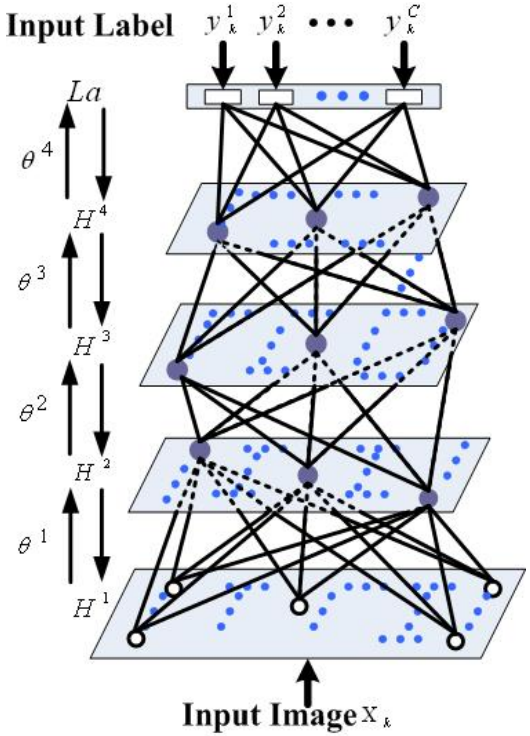
- Different feature and task, similar mapping
- Cross-lingual concepts
  - apple-orange=苹果-橘子
- Cross-domain concepts
  - disease-drug=ignorance-book
- Far from human-level performance
- Deep learning offers new possibilities

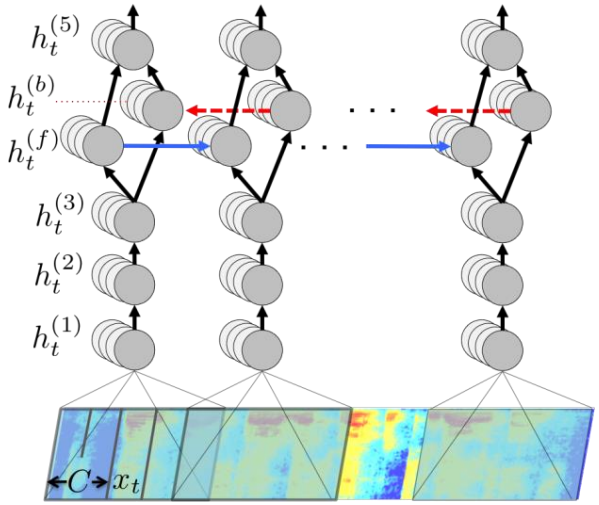


		$\mathcal{Y}+$		$\mathcal{Y}-$
		M(x)+	M(x) -	
$\mathcal{X}+$	P(X)+	Conventional ML	Model transfer[10]	Multitask learning[11]
	P(X)-	Model Adaptation[12], [13], incremental learning[14]		
$\mathcal{X}-$			Co-training[15] Heterogeneous transfer learning[16], [17]	Analogy learning [18]

Deep learning and representation learning

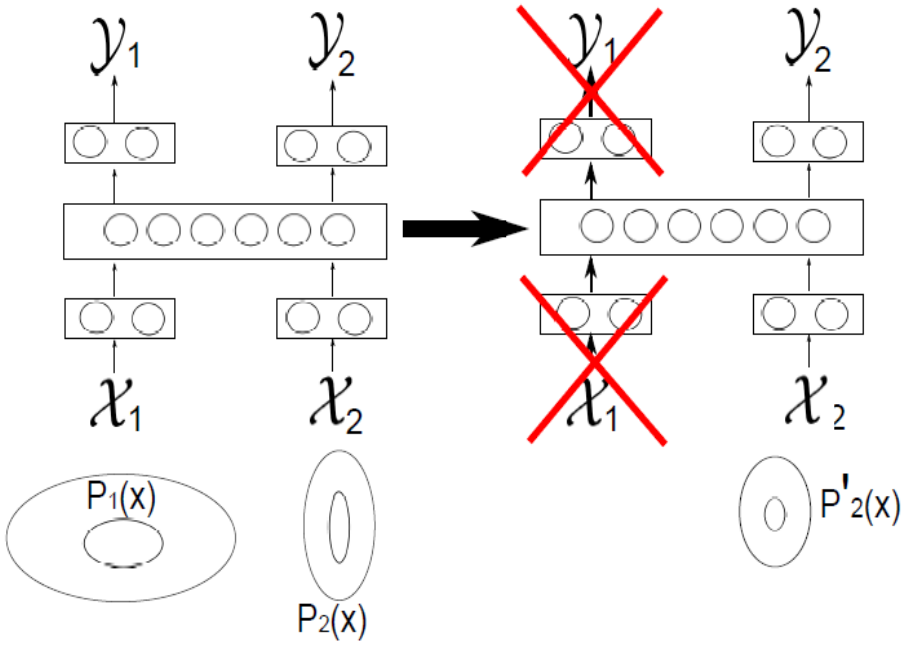




TL in deep learning era

- Learn representations shared by various inputs and various tasks.
- Training use all possible features and task.
- Adaptation change the model to suite the target domain.



Training Adaptation

Y. Bengio, “Deep learning of representations for unsupervised and transfer learning,” in ICML Unsupervised and Transfer Learning, 2012

TL in deep learning era (2)

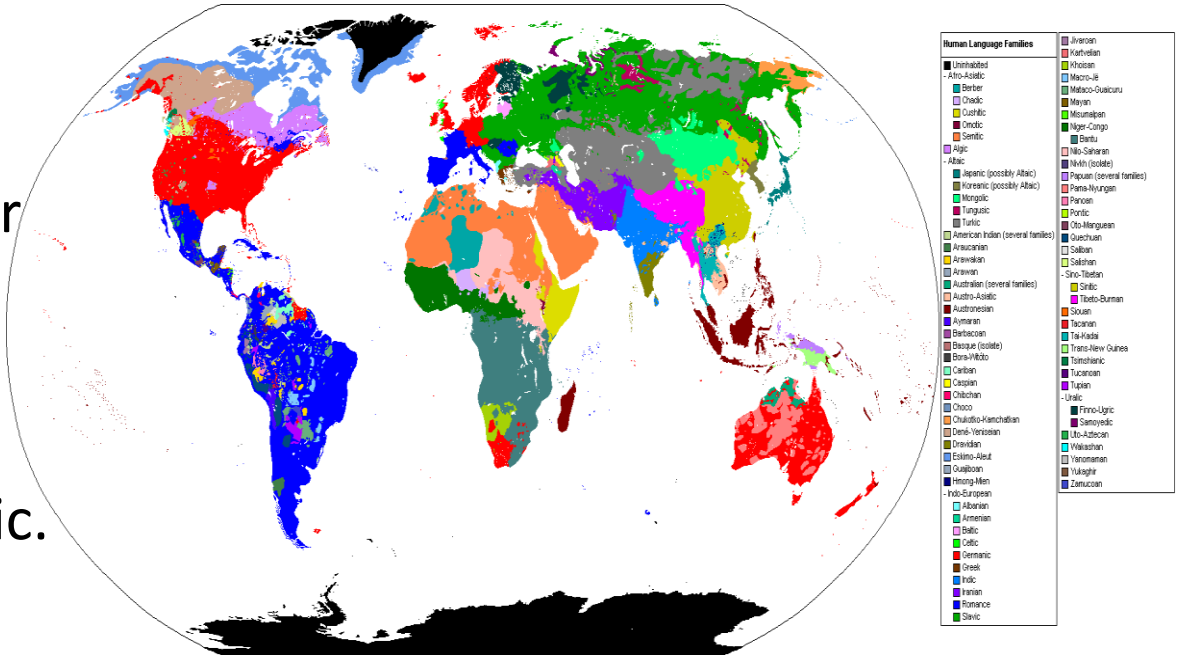
- Some representative work
  - Cross-domain migration, from book review to DVD review . Glorot 2011.
  - Cross-database migration, from PASCAL to VOC. Oquab 2014.
  - One/zero-shot learning. Lee 2006, Larochelle 2008, Socher 2013.

Content

- Transfer learning review
  - Transfer learning methods
  - Transfer learning in deep era
- Transfer learning in speech processing
  - Cross-lingual transfer
  - Speaker adaptation
  - Model transfer
- Transfer learning in language processing
  - Cross-lingual transfer
  - Cross-domain transfer
  - Model transfer
- Perspective and conclusions

Cross-lingual transfer

- More than 5000 languages, 389 languages accounts for 94% population.
- Most languages are minor.
- Languages are dynamic.
- Make use of the similarity!



https://www.ethnologue.com/statistics

Naïve transfer

- Transfer via IPA or phone pairs. Schultz 01, Vu 11.

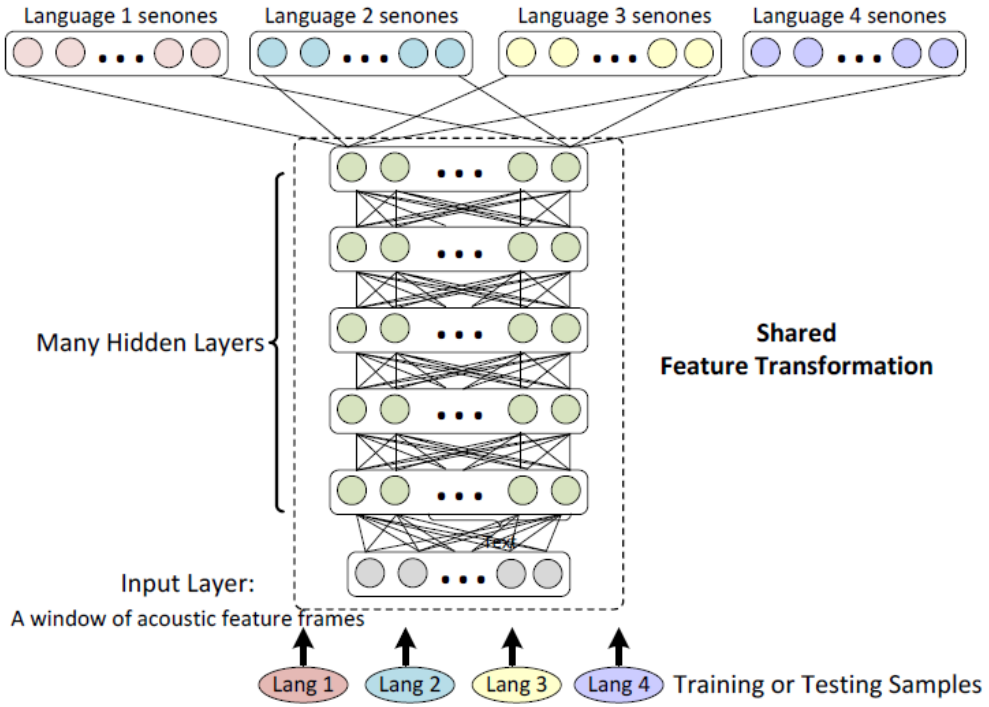
the international phonetic alphabet (2005)

consonants (pulmonic)	LABIAL		CORONAL				DORSAL				RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ		ɲ	ŋ	ɴ			
Plosive	p b			t d		ʈ ɖ		c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ		j	ɰ				
Tap, flap		ⱱ		ɾ		ɽ							
Trill	ʙ			r						ʀ		ʀ̠	
Lateral fricative				ɬ ɮ									
Lateral approximant				l		ɭ		ʎ	ʟ				
Lateral flap				ɺ		ɭ̆							

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

DNN-based transfer

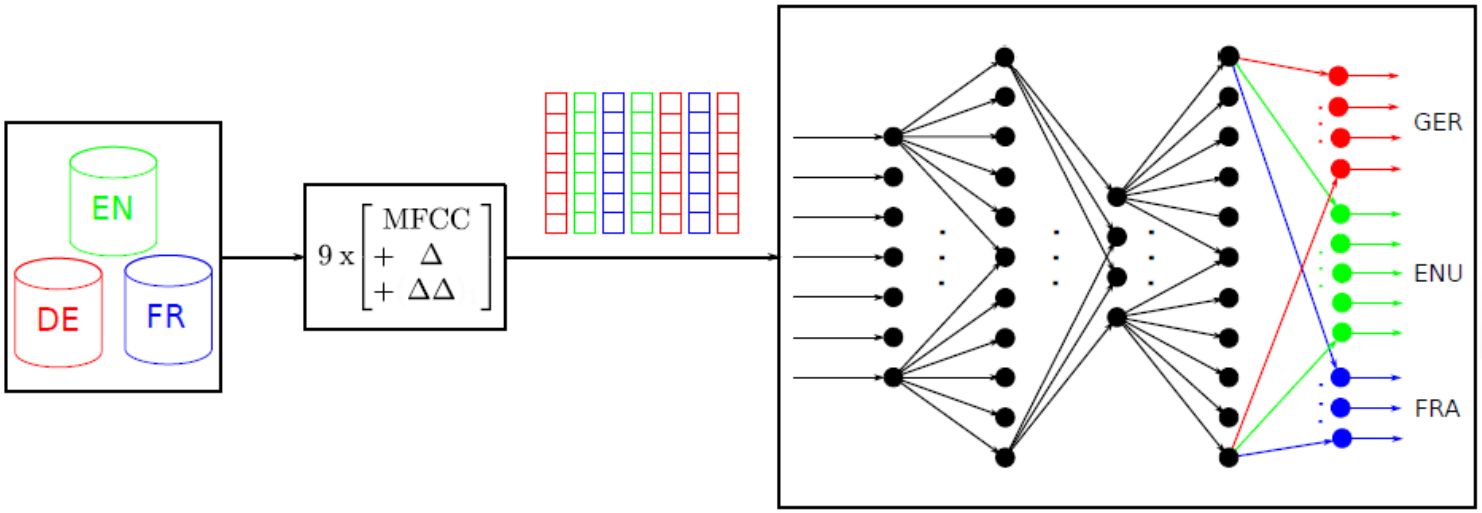
- Multilingual data DNN initialization. (Swietojanski 2012)
- Multilingual hybrid DNN. (Huang 2013. Heigold 2013, Ghoshal 2013)



J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," ICASSP 2013.

DNN-based transfer (2)

- Multilingual tandem feature. (Vesely 12, Thomas13, Tuske 13, Knill 14)



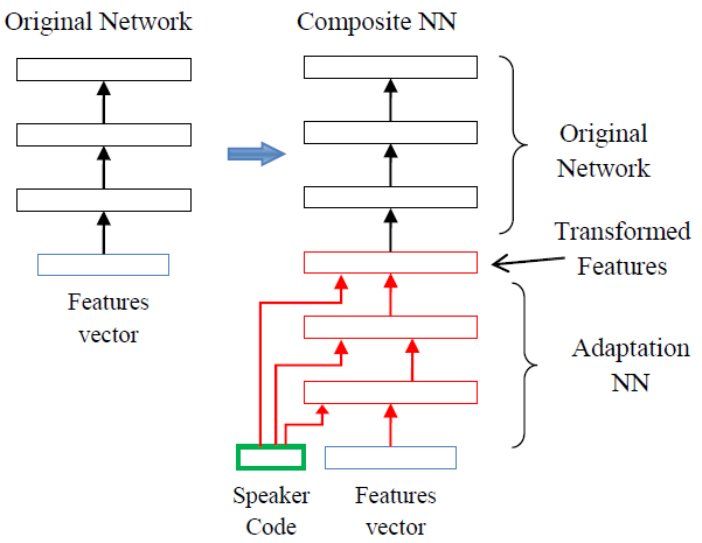
Z. Tuske, J. Pinto, D. Willett, and R. Schluter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," ICASSP 2013.

Other multitask learning

- Phone + grapheme (Chen 2014)
- Phone + Language (Tang 2015)
- Phrase + speaker (Chen 2015)
- Multilingual language recognition (Fer 2015)

Speaker adaptation

- Basic adaptation in GMM era: MAP, MLLR
- DNN is not easy
  - Compact and global
- Involving speaker vector
  - Speaker code (Ossama 13)
  - i-vector (Saon 13)



Ossama and Jiang, Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code, ICASSP 13.

Speaker adaptation (2)

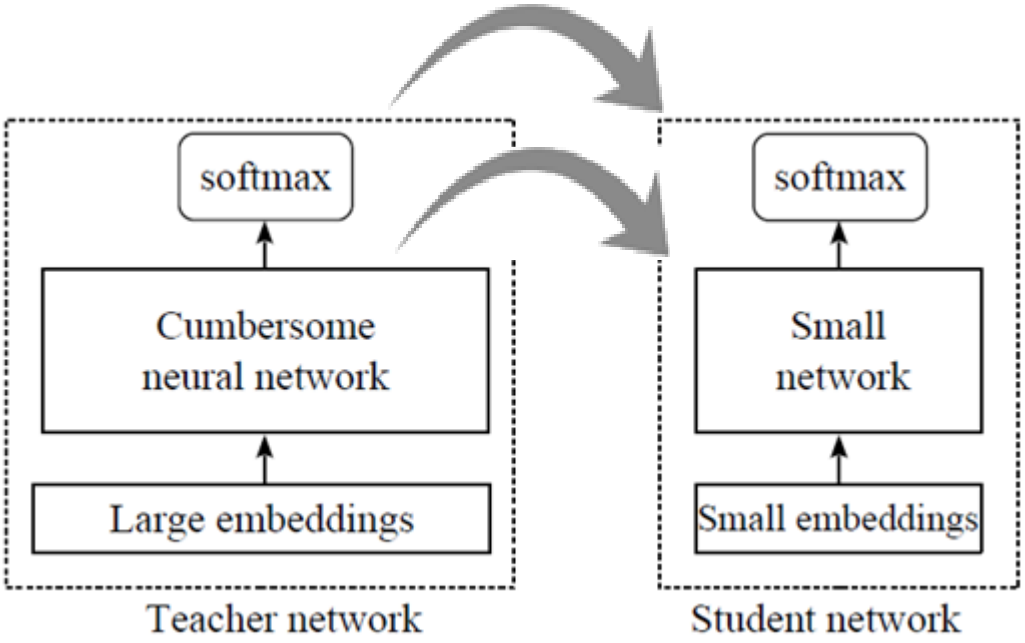
- Adapt DNN by structure constraint
  - Input layer (Neto95, Yao12)
  - Hidden layer (Siniscalchi13, Swietojanski14)
  - Output layer (Yao12)
  - Singular value (Xue 2014)
  - Adaptation by prior constraint (Yu 13)
- RNN adaptation
  - Speaker-adaptive front-end (Miao2015)
- Speaker adaptation is effective for only small network (Liao 13)

Adaptation in other places

- Speaker-specific speech synthesis
  - Tamura01,Wu09, Yamagishi09
- Multilingual speech synthesis
  - Wu09,Liang10,Gibson10
- DNN adaptation for speech synthesis
  - Wu15, Potard15

Model transfer

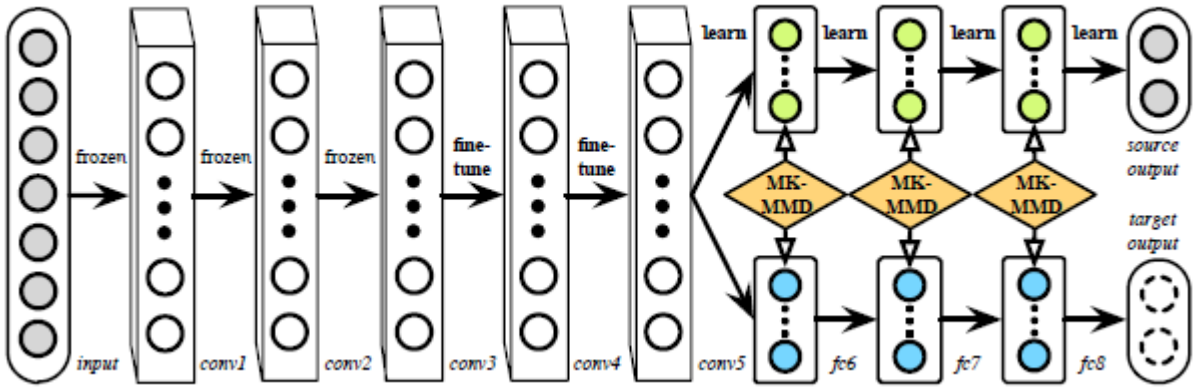
- Using a complex model to supervise simple model (Ba14, Hinton14)



Lili Mou, Ge Li, Yan Xu, Lu Zhang, Zhi Jin, DistillingWord Embeddings: An Encoding Approach, 2015.

Various transfer schemes

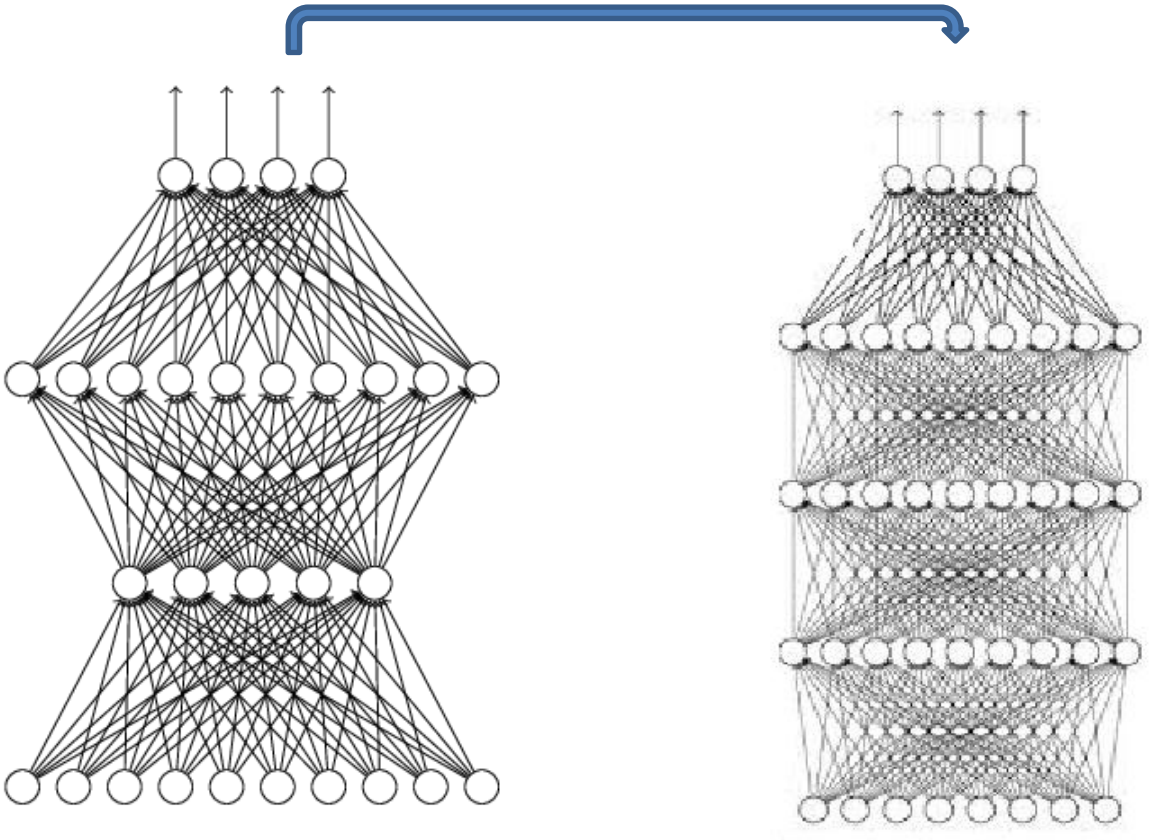
- Using complex DNN to regularize simple DNN in ASR (Li2014)
- Using RNN to supervise DNN (Chan15)
- Fitnet: regularize hidden layers (Romero15)
- Regularize multiple layers (Long 15)



Long, Mingsheng and Wang, Jianmin, Learning Transferable Features with Deep Adaptation Networks, ICML2015

Can stupid teacher supervise smart students?

- Using simple model to teach complex models is possible (Tang 15, Wang 15)



Why it is possible?

- Teacher's outputs (soft labels) are easier to learn than hard labels
- Can play the role of (1) regularization (2) pre-training

	# LSTM	T	TR FA%	CV FA%	WER%
DNN [4 hidden layers]	0	-	63.0	45.2	11.40
RNN [raw]	1	-	67.3	51.9	13.57
RNN [prt.]	1	1	59.4	49.9	11.46
RNN [prt.+ft.]	1	1	65.5	54.2	10.71
RNN [prt.]	1	2	58.2	49.5	11.32
RNN [prt.+ft.]	1	2	64.6	54.1	10.57
RNN [raw]	2	-	68.8	53.2	12.34
RNN [prt.]	2	1	60.4	50.6	11.11
RNN [prt.+ft.]	2	1	66.6	55.4	10.13
RNN [prt.]	2	2	58.6	49.7	11.26
RNN [prt.+ft.]	2	2	65.8	55.2	**10.10**

Content

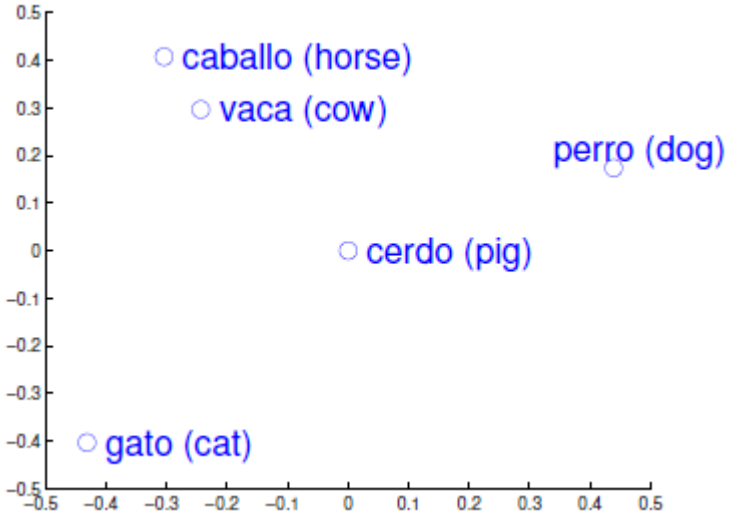
- Transfer learning review
  - Transfer learning methods
  - Transfer learning in deep era
- Transfer learning in speech processing
  - Cross-lingual transfer
  - Speaker adaptation
  - Model transfer
- Transfer learning in language processing
  - Cross-lingual transfer
  - Cross-domain transfer
  - Model transfer
- Perspective and conclusions

Cross-lingual and multilingual transfer

- Transfer by word-to-word dictionary (Shi10, Ma15) or by SMT
- Transfer by learning common latent structures
  - Multilingual LDA (De Smet 11)
  - RBM factor learning (Wei 11)
  - Multilingual cluster NER (Tackstrom2012)
  - Linear projection, tested on TC (Duan 12)

Cross-lingual and multilingual transfer (2)

- Multilingual word embedding by projection
  - Linear projection (Mikolov 13)
  - Orthogonal projection (Xing 15)
  - Canonical correlation analysis (Faruqui2014)
- Multilingual word embedding by changed cost function
  - Klementiev 12
- Deep learning (Zhou14)



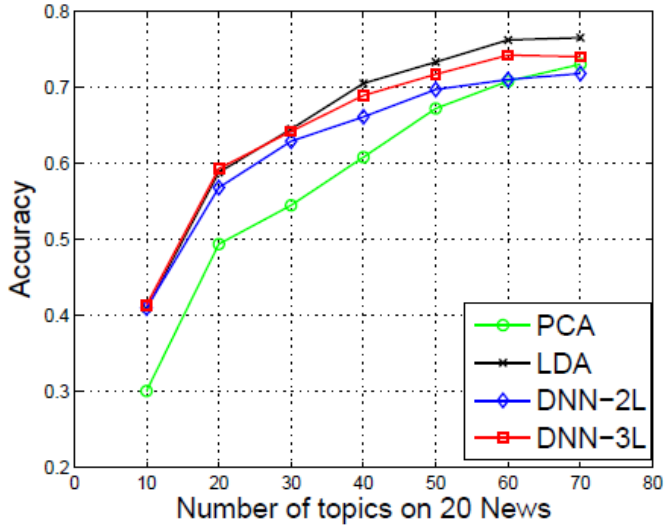
Mikolov et al., Exploiting similarities among languages for machine translation, 2013.

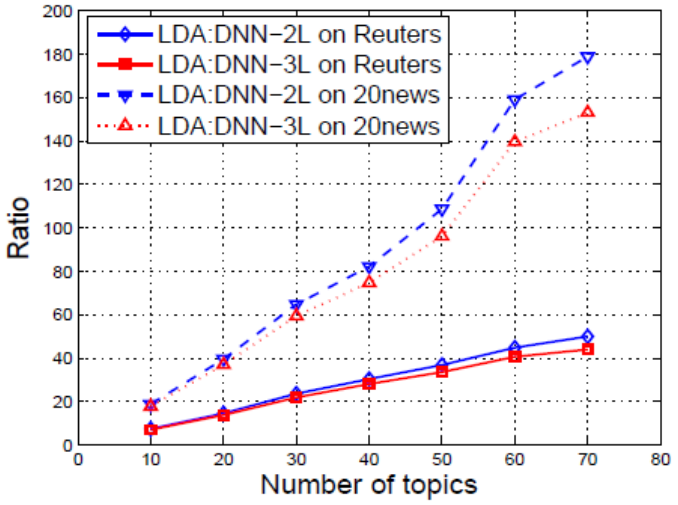
Cross-domain transfer

- Transfer knowledge between domains with different features
  - Estimate correspondence by co-occurrence. Tested on text-aided image classification (Dai 08)
  - Common space approach for image segmentation and labeling (Socher 10)
  - Deep learning
    - Image + Text common space for image classification (Socher13, Frome13)
    - Heterogeneous LM (Kiros2014)
    - RBM (Srivastava 2012) and RNN (Socher2014)

Model transfer

- NN Knowledge distillation for sentiment classification. (Mou 15)
- LDA to supervise NN. Applied to document classification (Zhang 15)





Content

- Transfer learning review
  - Transfer learning methods
  - Transfer learning in deep era
- Transfer learning in speech processing
  - Cross-lingual transfer
  - Speaker adaptation
  - Model transfer
- Transfer learning in language processing
  - Cross-lingual transfer
  - Cross-domain transfer
  - Model transfer
- Perspective and conclusions

Go back to NIPS*95

- What do we mean by related tasks and how can we identify them?
  - Not easy to answer, but seems not very critical.
- How do we predict when transfer will help (or hurt)
  - Again, not simple. But we can do something (e.g., Long 14). Deep learning promises.
- What are the benefits: speed, generalization, intelligibility,...?
  - Seems all
- What should be transferred: internal representations, parameter settings, features,...?
  - Seems all
- How should it be transferred: weight initialization, biasing the error metric,...?
  - All seems fine, though regularization seems more promising.
- How do we look inside to see what has been transferred?
  - Depends on the model

How it works in speech and language processing?

- Very important and has been employed with a long history.
- However in most of time, we didn’t notice what it is.
- Many unexplored aspects, mostly due to the lack of systematic thinking.

Some example questions for future

- How to involve heterogeneous resources including audio, visual, language to solve the most challenging tasks in the respective research fields?
- Can we learn common representations for both speech, language and speaker recognition, and use them for AI?
- How to utilize the large amount of unlabeled data more efficiently in the big-data era?

Acknowlegement

- Thanks to Zhiyuan Tang for reference collection.
- Thanks to the team for the hard work.