CCAI 2017
中国人工智能大会

深度学习的迁移模型

杨强

香港科技大学， 新明工程学讲座教授
计算机与工程系主任, 大数据研究所主任

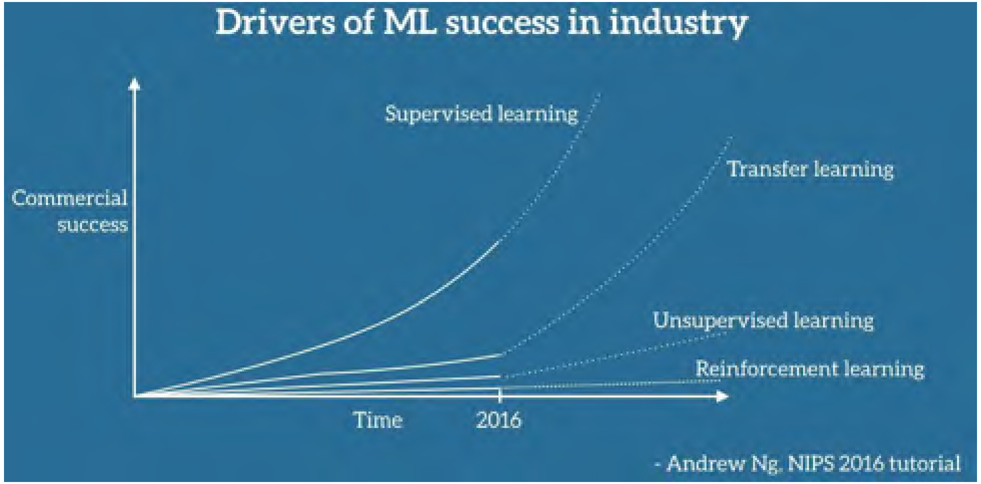
AlphaGo 还不会做什么？举一反三

19x19

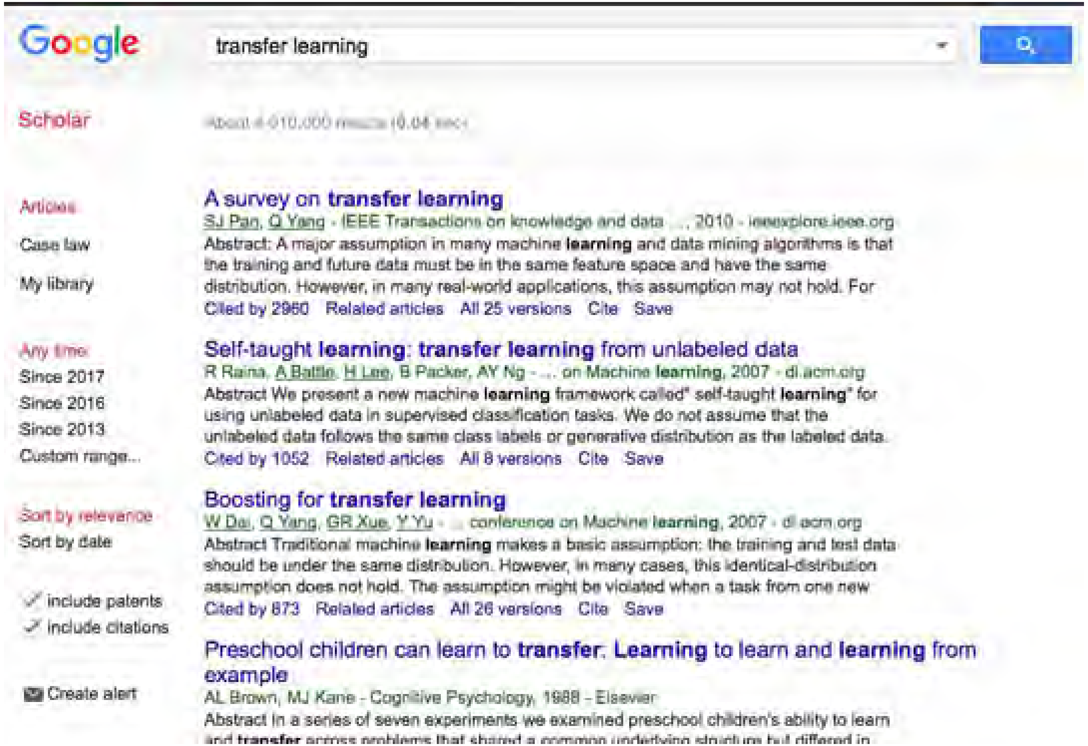
21X21

迁移学习 Transfer Learning

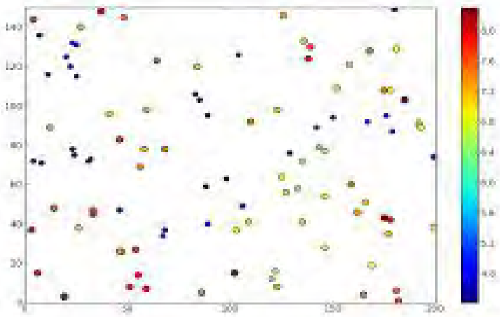
下一个热点：迁移学习 Transfer Learning



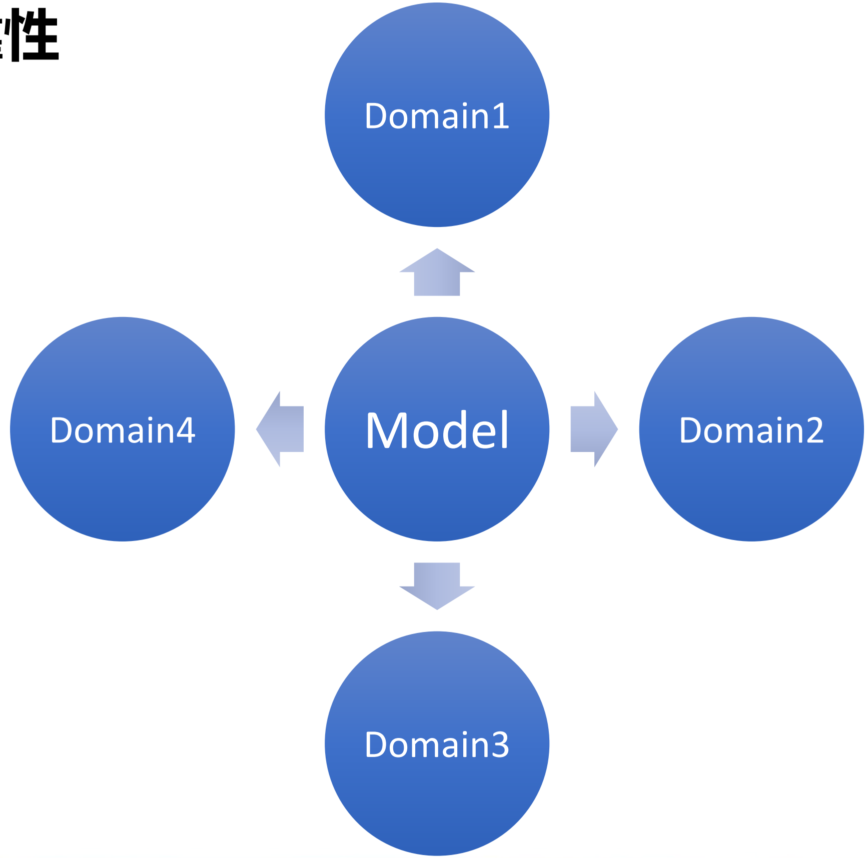
迁移学习 Transfer Learning



迁移学习的优点 1: 小数据



迁移学习的优点 2: 可靠性



迁移学习的优点 3: 个性化

迁移学习的难点

迁移学习本质：找出不变量

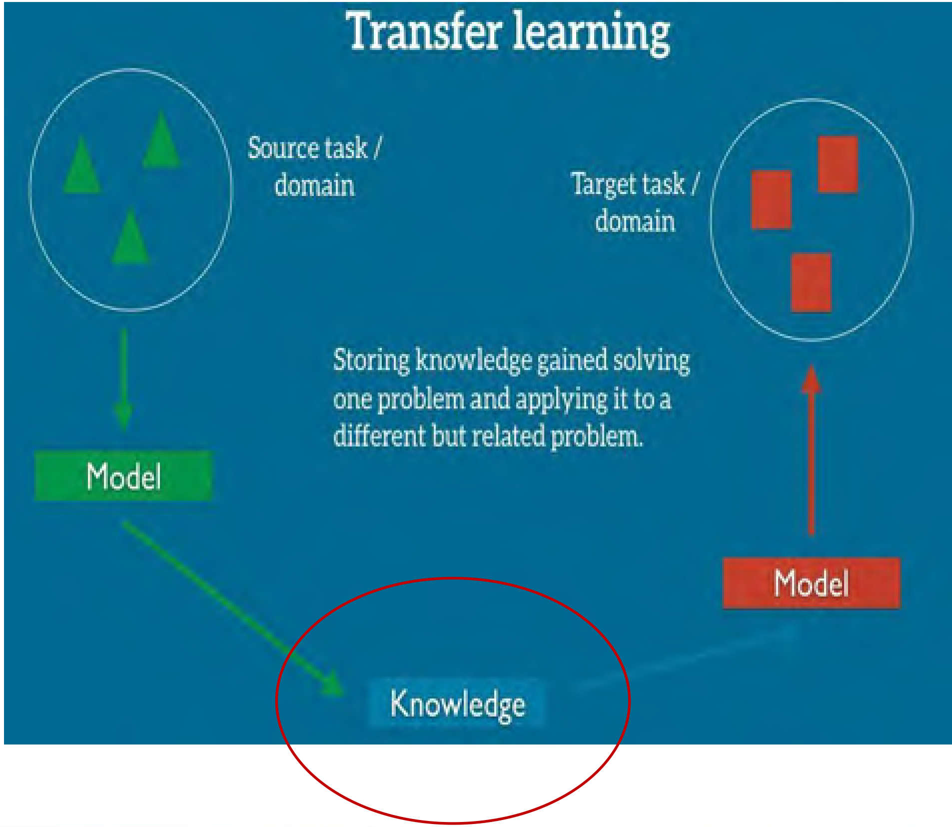
One Knowledge, Two Domains

Driving in Mainland China

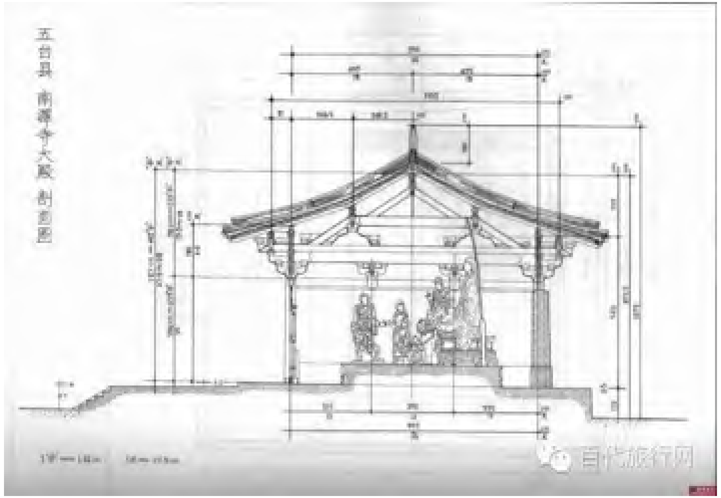
Driving in Hong Kong, China

迁移学习 Transfer Learning

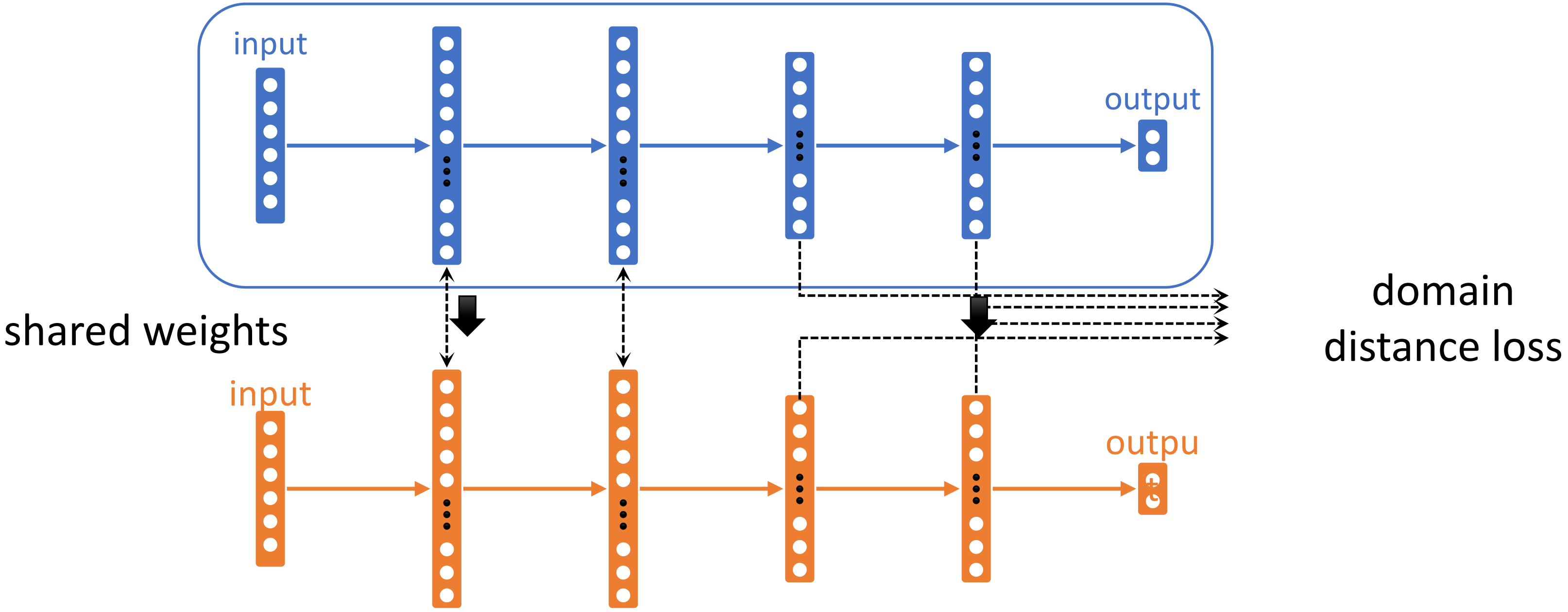
- Yann LeCun: 机器学习的热力学模型?
  - （百度百科）热力学主要是从能量转化的观点来研究物质的热性质，它提示了能量从一种形式转换为另一种形式时遵从的宏观规律，总结了物质的宏观现象而得到的热学理论。



深度学习 + 迁移学习: 多层次的特征学习



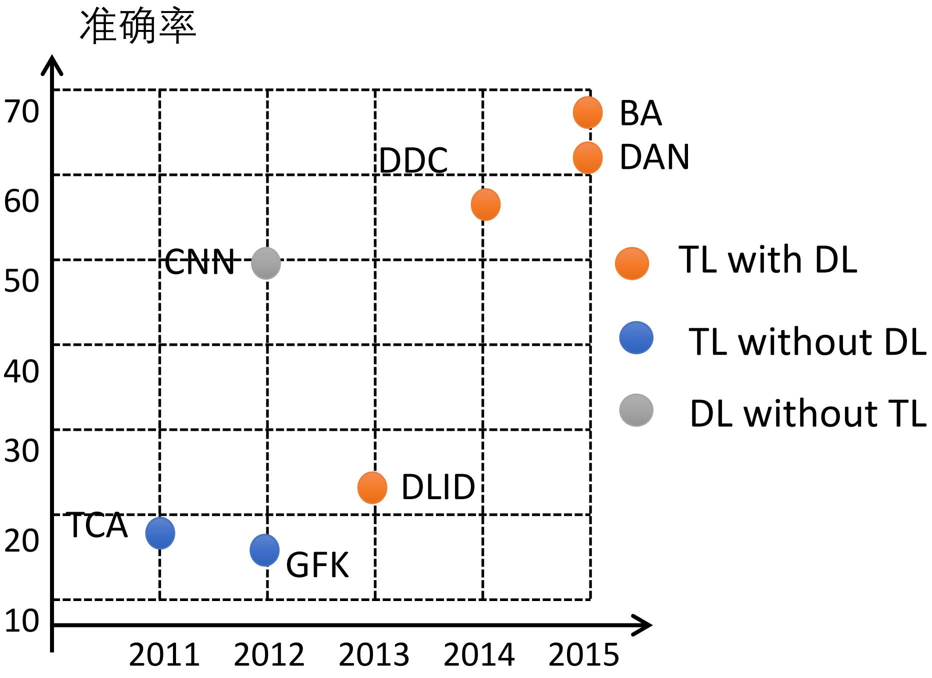
深度学习的迁移模型: 定量分析



- Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

深度学习模型的迁移： 定量分析

- TCA: Transfer Component Analysis: Pan, Sinno Jialin, Ivor W. Tsang, James T. Kwok, and Qiang Yang. "Domain adaptation via transfer component analysis." *IEEE Transactions on Neural Networks* 22, no. 2 (2011): 199-210.
- GFK: Geodesic Flow Kernel: Gong, Boqing, Yuan Shi, Fei Sha, and Kristen Grauman. "Geodesic flow kernel for unsupervised domain adaptation." In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2066-2073. IEEE, 2012.
- DLID: Deep Learning for domain adaptation by Interpolating between Domains: Chopra, Sumit, Suhrid Balakrishnan, and Raghuraman Gopalan. "Dlid: Deep learning for domain adaptation by interpolating between domains." *ICML workshop on challenges in representation learning*. Vol. 2. 2013.
- DDC: Deep Domain Confusion: Tzeng, Eric, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. "Deep domain confusion: Maximizing for domain invariance." *arXiv preprint arXiv:1412.3474* (2014).
- DAN: Deep Adaptation Networks: Long, Mingsheng, Yue Cao, Jianmin Wang, and Michael Jordan. "Learning transferable features with deep adaptation networks." In *International Conference on Machine Learning*, pp. 97-105. 2015.
- BA: Backpropagation Adaptation: Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." In *International Conference on Machine Learning*, pp. 1180-1189. 2015.



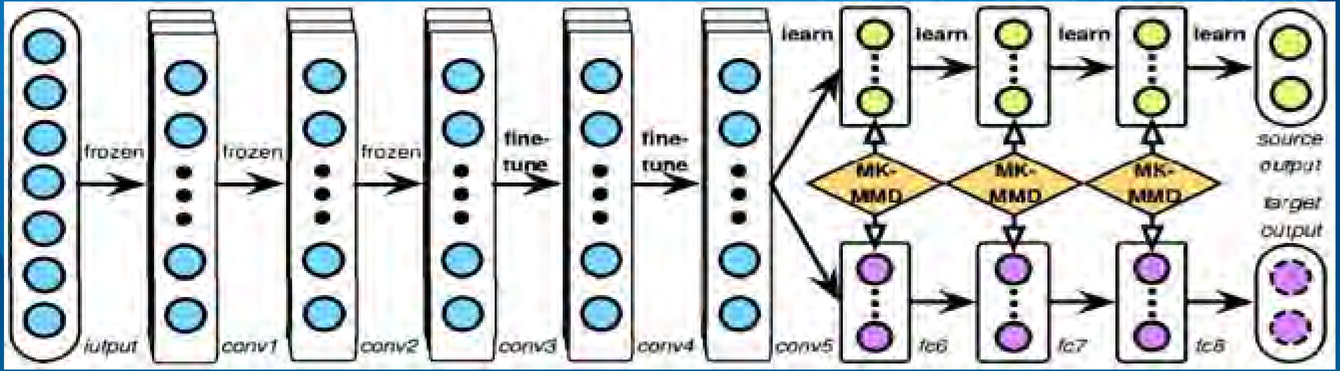
Deep Adaptation Networks (DAN) [Long et al. 2015]

multi-layer adaptation

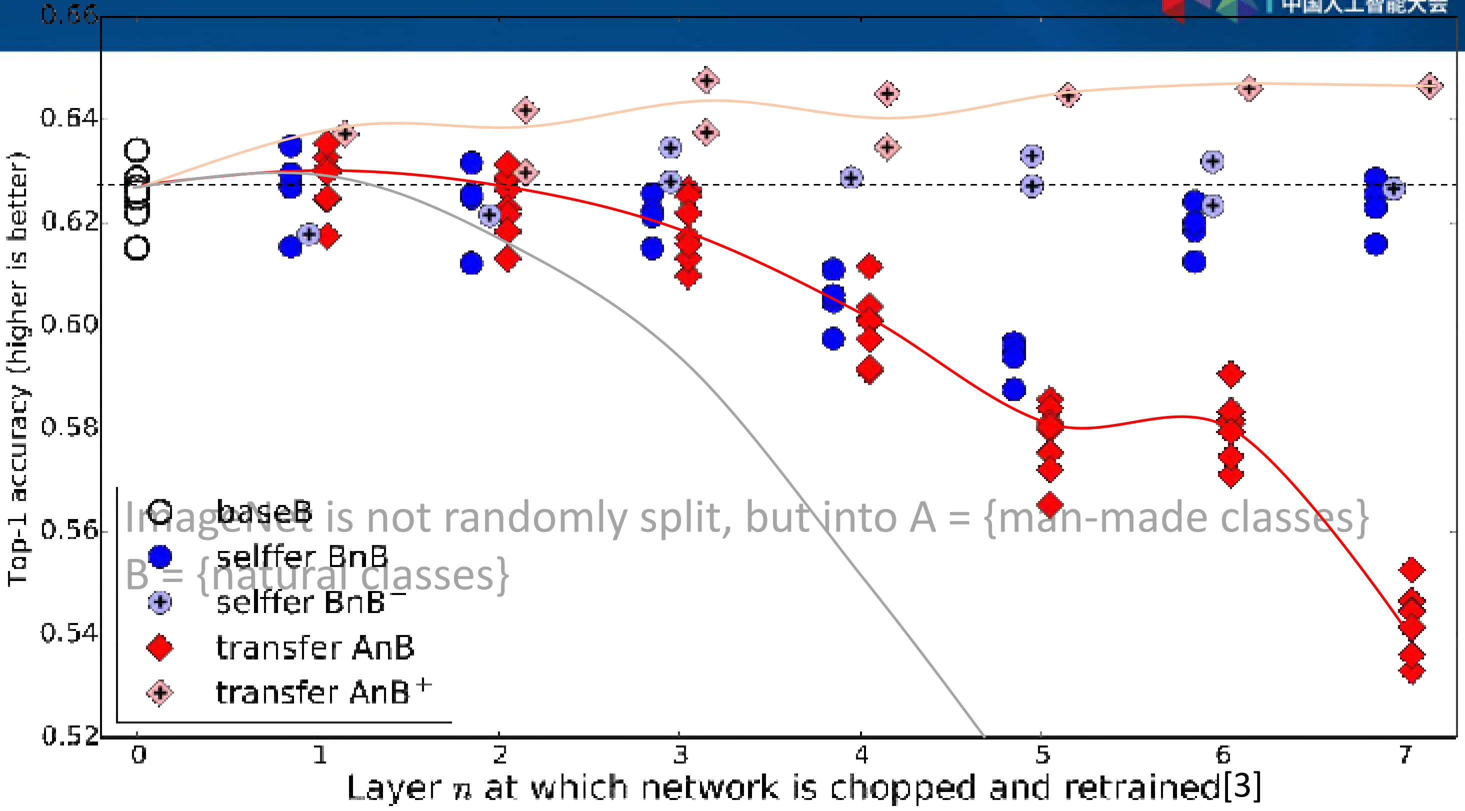
$$\mathcal{L}_D(\mathbf{X}^s,\mathbf{X}^t) = \sum_{l=l_1}^{l_2} \mathrm{MK-MMD}\left(\mathbf{X}^{s(l)},\mathbf{X}^{t(l)}\right) = \left\| \frac{1}{n^s}\sum_{i=1}^{n^s}\phi\left(\mathbf{x}_i^{s(l)}\right) - \frac{1}{n^t}\sum_{j=1}^{n^t}\phi\left(\mathbf{x}_j^{t(l)}\right) \right\|_{\mathcal{H}_k}^2$$

combination of m PSD kernels

$$k\left(\mathbf{x}_i^{s(l)},\mathbf{x}_i^{s(l)}\right) = \left\langle \phi\left(\mathbf{x}_i^{s(l)}\right), \phi\left(\mathbf{x}_i^{s(l)}\right)\right\rangle = \sum_{u=1}^{m}\beta_u k_u(\mathbf{x}_i^{s(l)},\mathbf{x}_i^{s(l)})$$



深度迁移学习的量化分析



ImageNet is not randomly split, but into A = {man-made classes} B = {natural classes}

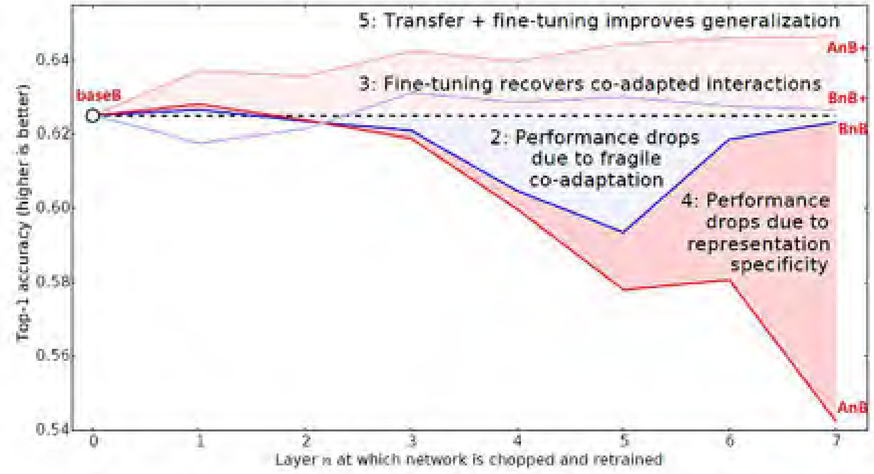
Conclusion 1: lower layer features are more general and transferrable, and higher layer features are more specific and non-transferrable.

Conclusion 2: transferring features + fine-tuning always improve generalization.

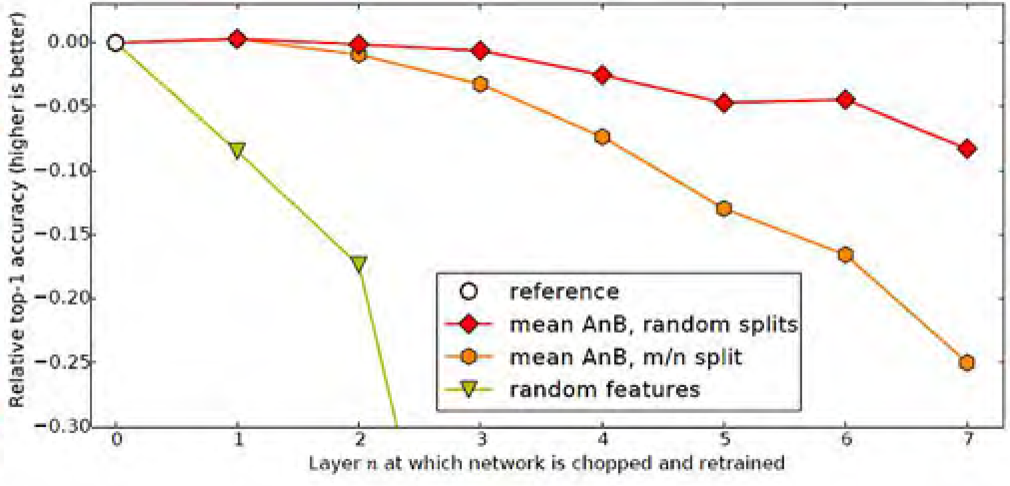
What happens if the source and target domain are very dissimilar?

Transferability of Layer-wise Features

varying four transfer strategies



varying similarity between domains



Conclusions

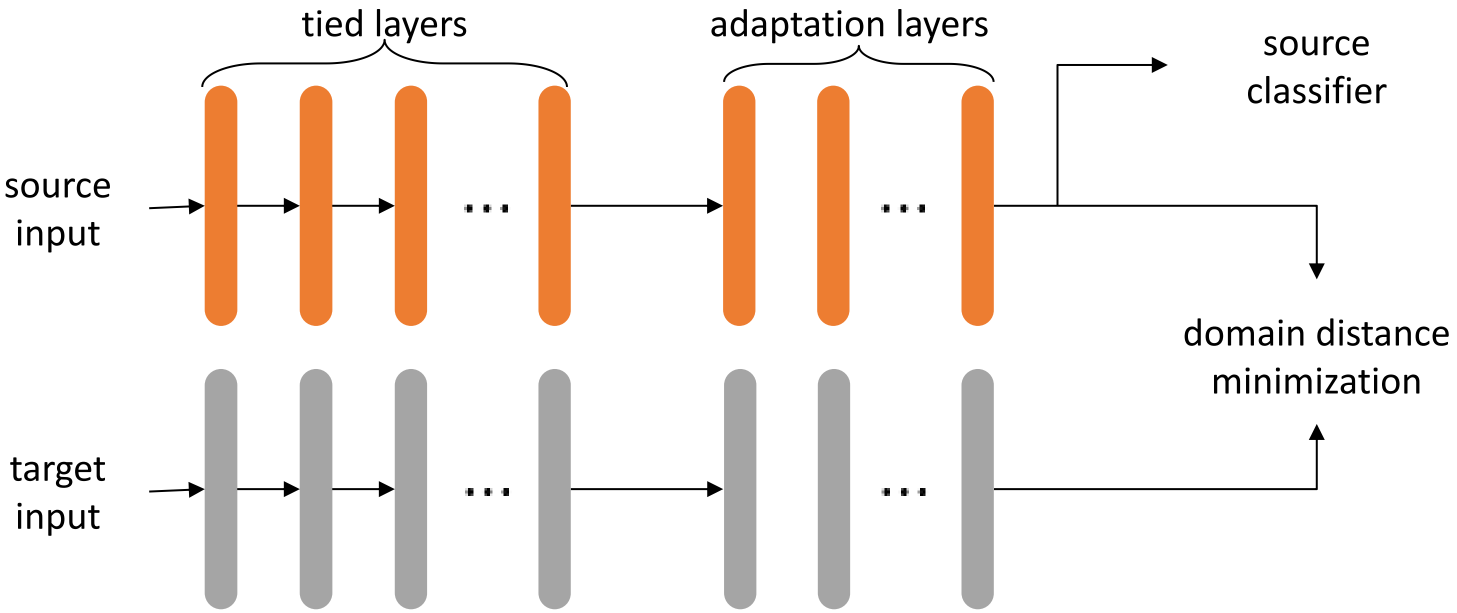
- Fine-tuning with labeled data in a target domain always helps.
- Transition from general to specific in a deep neural network.
- Performance drops when two domains are very dissimilar.

What if

- No or limited labeled data
-
- Two dissimilar domains

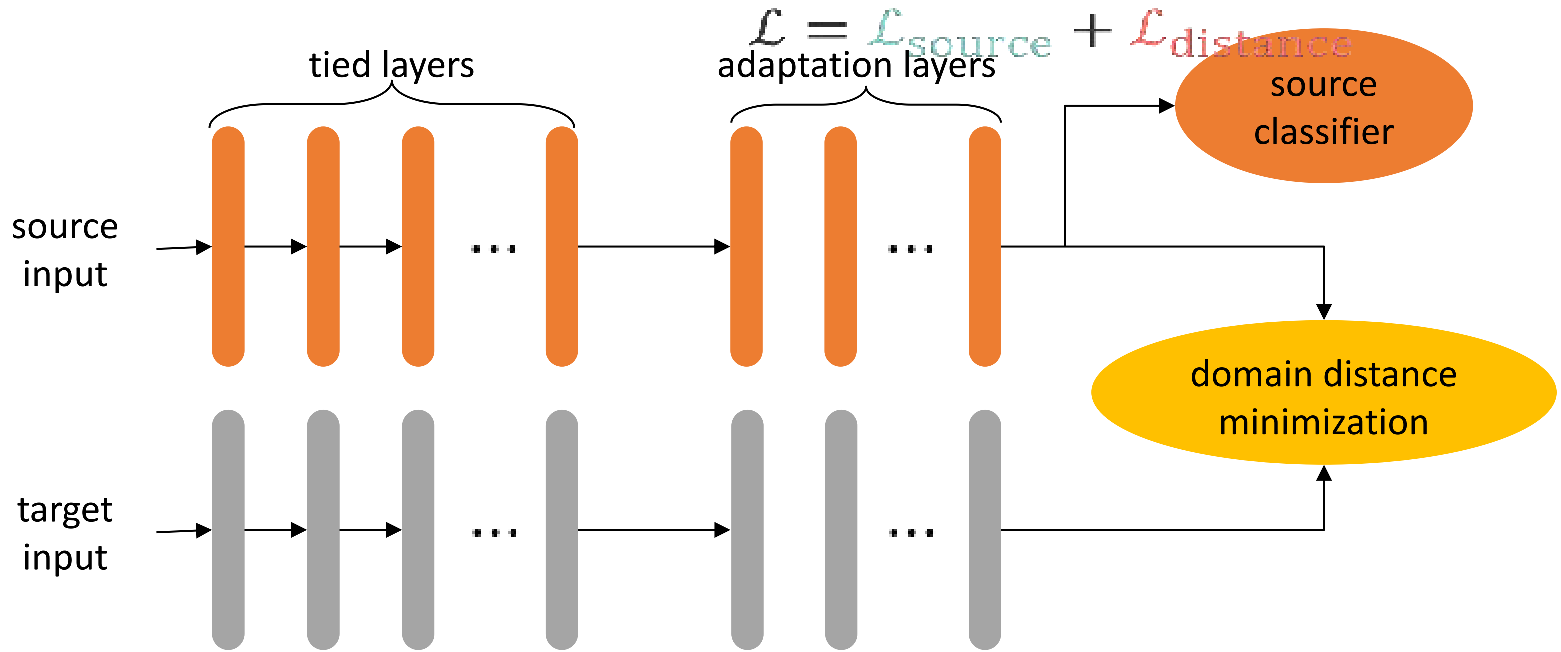
Unsupervised Deep Transfer Learning

- Goal: learn a classifier or a regressor for a target domain which is unlabeled and dissimilar to a source domain.
- General architecture: Siamese architecture



Unsupervised Deep Transfer Learning

- Objective



Unsupervised Deep Transfer Learning

$\mathcal{L}_{\text{distance}}$

discrepancy loss

Directly minimizes the difference between two domains. [Tzeng et al. 2014, Long et al. 2015, Long et al. 2017]

adversarial loss

Encourages a common feature space through an adversarial objective with respect to a domain discriminator. [Ganin et al. 2015, Tzeng et al. 2015, Liu and Tuzel 2016, Tzeng et al. 2017]

reconstruction loss

Combines both unsupervised and supervised training. [Ghifary et al. 2016, Bousmalis et al. 2016]

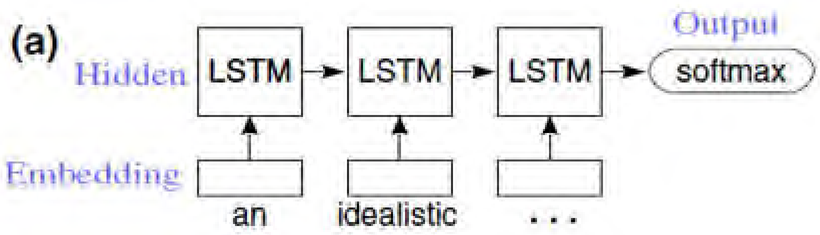
Discrepancy Based Methods

- A source domain's parameters = a target domain's parameters
- Overall objective

source domain classification loss $\mathcal{L} =$ domain distance loss $\mathbf{x}^s, \mathbf{x}^t)$

method	where to adapt	distance between	distance metric
Tzeng et al. 2014	a specific layer	marginal distributions	Maximum Mean Discrepancy (MMD)
Long et al. 2015	multiple layers	marginal distributions	Multi-kernel MMD (MK-MMD)
Long et al. 2017	multiple layers	joint distributions	Joint Distribution Discrepancy (JDD)

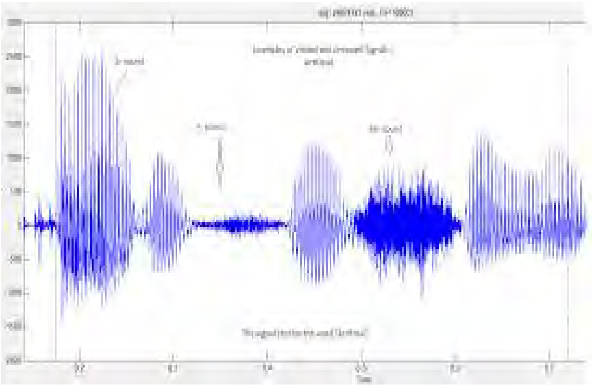
Similarly in RNN for NLP

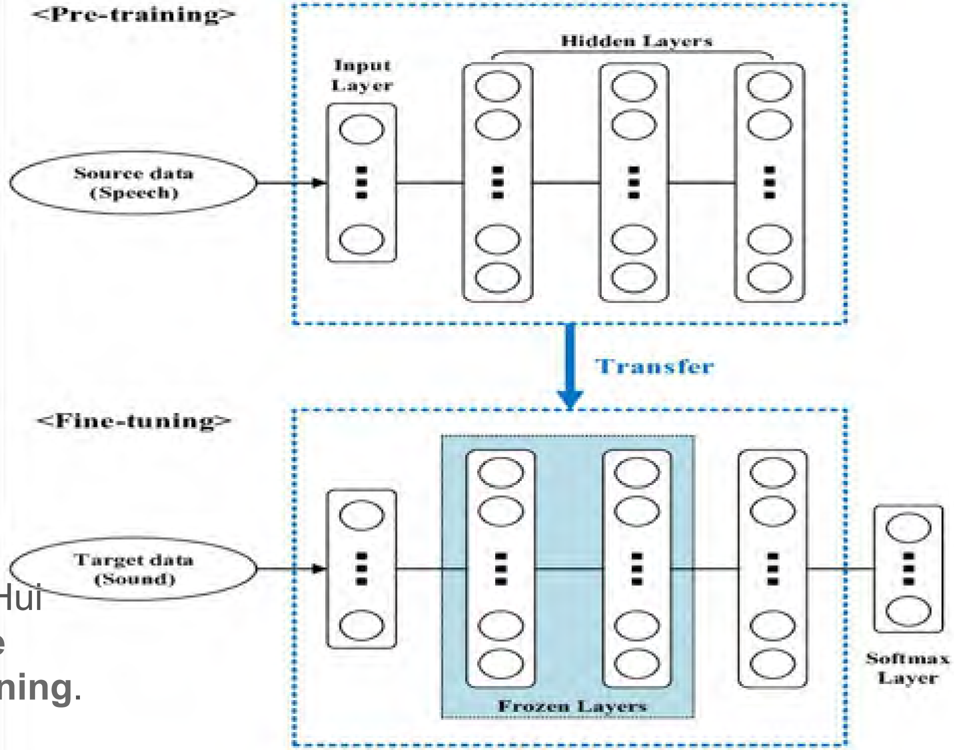


Setting	IMDB→MR	IMDB→QC
Majority	50.0	22.9
E⊠ H□ O□	75.1	90.8
E🔒 H□ O□	78.2	93.2
E🔒 H🔒 O□	78.8	55.6
E🔒 H🔒 O🔒	73.6	–
E🔓 H□ O□	78.3	92.6
E🔓 H🔓 O□	81.4	90.4
E🔓 H🔓 O🔓	80.9	–

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. **How transferable are neural networks in NLP applications**? In EMNLP 2016

语音识别中的迁移学习：口音迁移

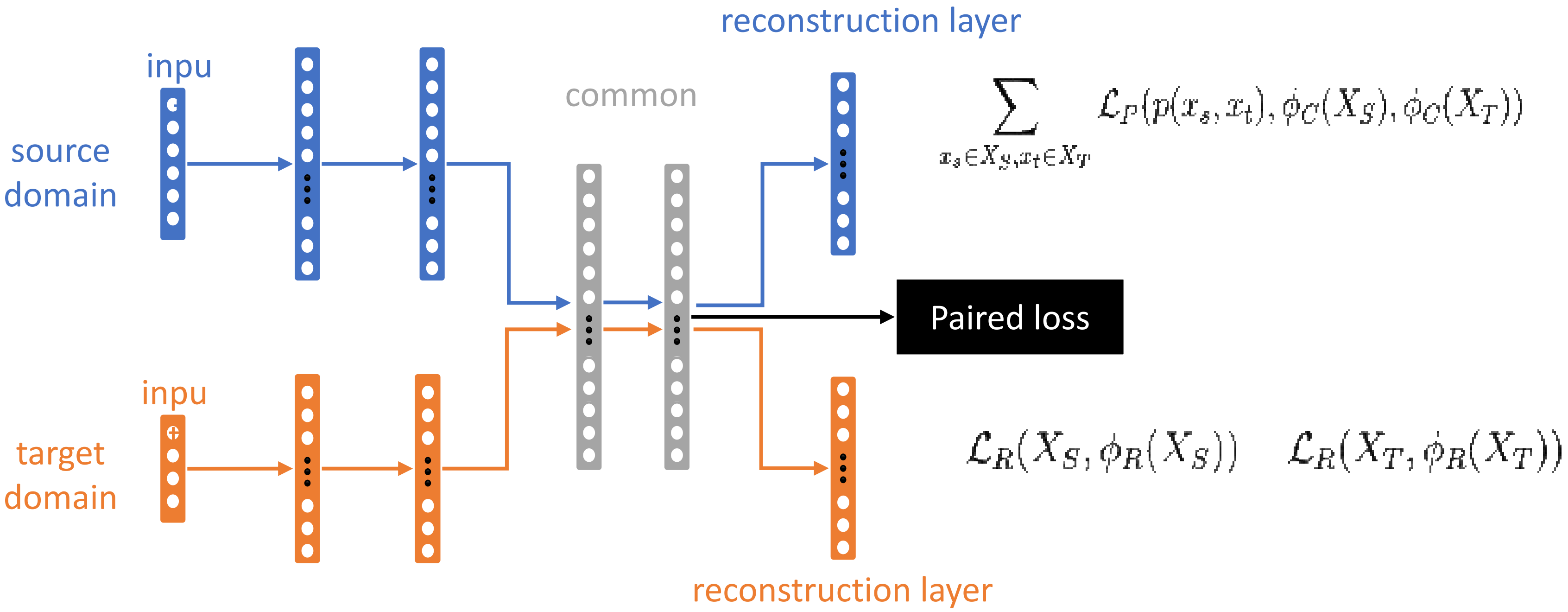




Yue Zhao, Yan M. Xu, Mei J. Sun, Xiao N. Xu, Hui Wang, Guo S. Yang, Qiang Ji: **Cross-language transfer speech recognition using deep learning.** ICCA 2014

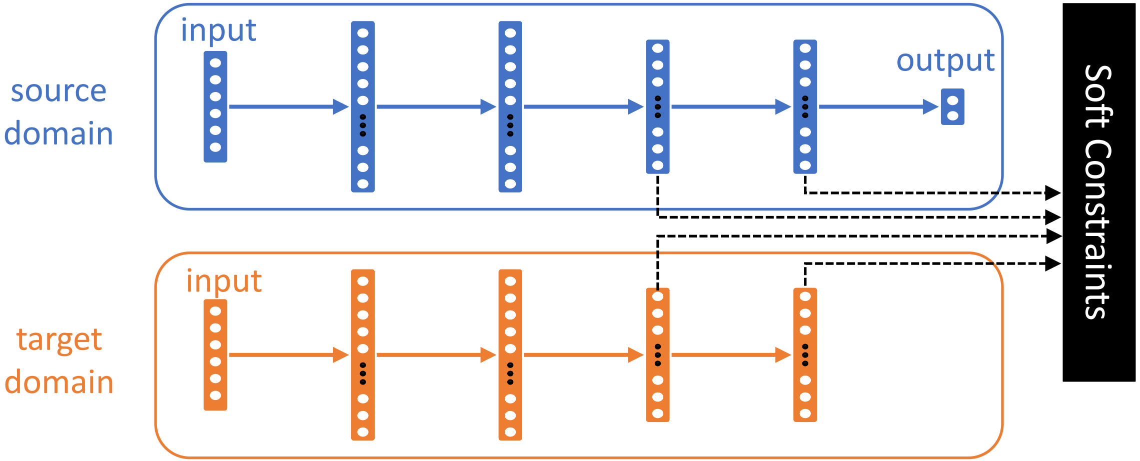
多模态学习和迁移学习

Multimodal Transfer Deep Learning with Applications in Audio-Visual Recognition,Seungwhan Moon, Suyoun Kim, Haohan Wang, arXiv:1412.3121



加入正则化 Regularization

$$L = L_c(X_s, Y_s) + \lambda L_D(X_s, X_t)$$



1. Determinative Distance MMD
2. Learn to align: fool the domain classifier
   - Reverse Gradient: reverse the domain classifier gradient for CNN[7] and RNN[8] representation layers
   - ADDA[9]: Alternatively Optimize Domain classifier layer or the common feature by fixing the other
3. Auxiliary Task Loss
   - Clustering[10]: add interpretability and enable zero-shot learning

$$MMD(X_S, X_T) = \left\| \frac{1}{\|X_S\|} \sum_{x_s \in X_S} \phi(x_s) - \frac{1}{\|X_T\|} \sum_{x_t \in X_T} \phi(x_t) \right\|_2^2$$

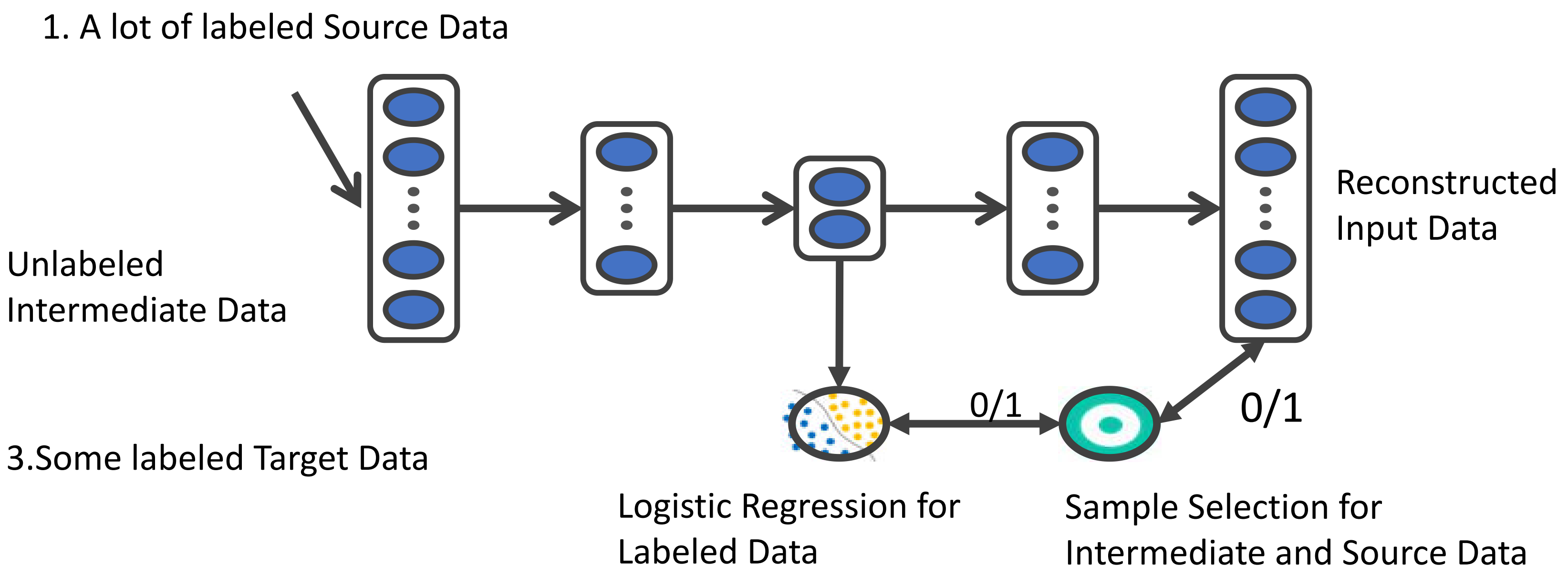
$$\min_{\theta_e, \theta_y, \theta_y} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y(\mathbf{x}^i; \theta_y, \theta_e) + \lambda \max_{\theta_d} \left[-\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_d(\mathbf{x}^i; \theta_d, \theta_e) - \frac{1}{n'} \sum_{i=n+1}^{N} \mathcal{L}_d(\mathbf{x}^i; \theta_d, \theta_e) \right]$$

传递式的迁移学习 Transitive Transfer Learning

- Ben Tan, Yu Zhang, Sinno Jialin Pan, Qiang Yang: Distant Domain Transfer Learning. AAAI 2017

- Ben Tan, Yangqiu Song, Erheng Zhong, Qiang Yang: **Transitive Transfer Learning.** KDD 2015

传递式迁移学习


1. A lot of labeled Source Data
Unlabeled
Intermediate Data
Reconstructed
Input Data
0/1
0/1
3.Some labeled Target Data
Logistic Regression for
Labeled Data
Sample Selection for
Intermediate and Source Data

Parameter Initialization + Fine-tune

- Transfer Learning for Poverty prediction on satellite image[4]
- VGG-Net: initialize the parameter with last domain and then finetune

Closely related

ImageNet
Massive data

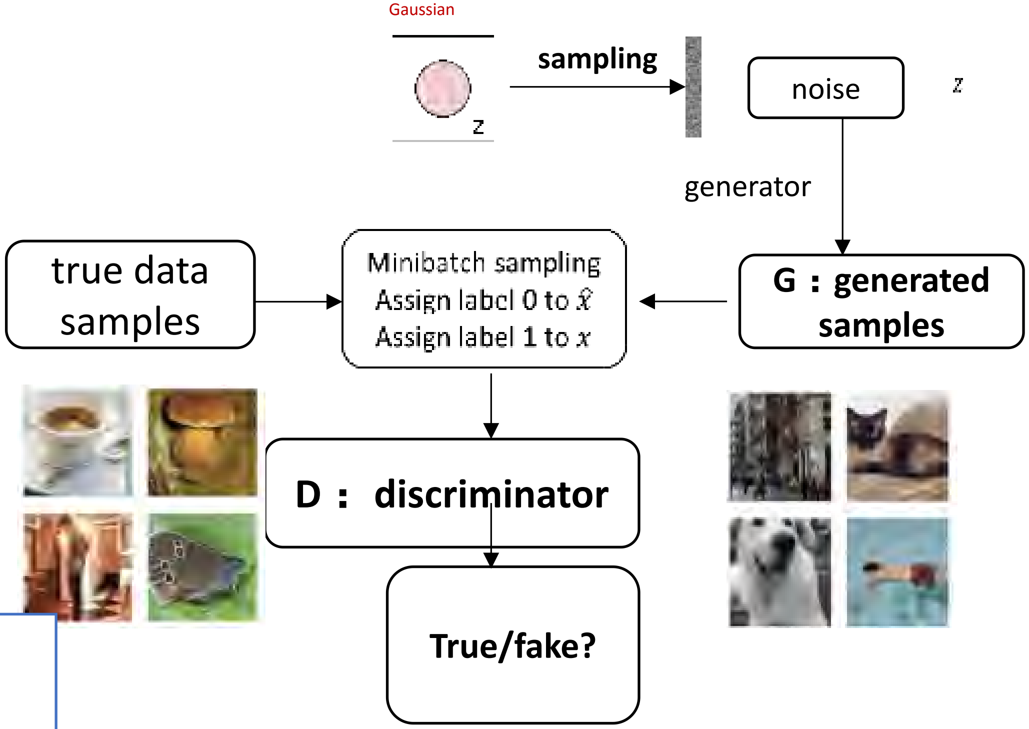
Night Light
Less data

Poverty
Scarce data

	Survey	ImgNet	Lights	ImgNet +Lights	Transfer
Accuracy	0.754	0.686	0.526	0.683	**0.716**
F1 Score	0.552	0.398	0.448	0.400	**0.489**
Precision	0.450	0.340	0.298	0.338	**0.394**
Recall	0.722	0.492	0.914	0.506	0.658
AUC	0.776	0.690	0.719	0.700	**0.761**

生成对抗网络 GAN

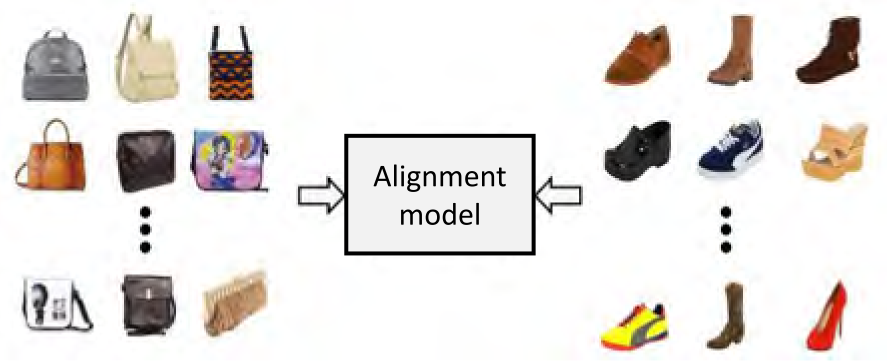
- G: 生成模型 generator
- D: 判别模型 discriminator



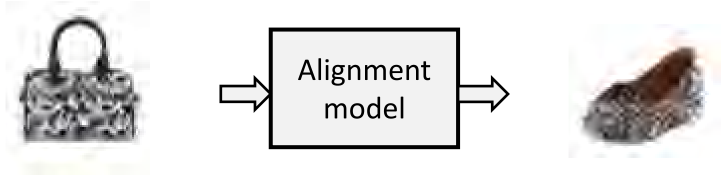
Goodfellow, Ian, et al. “Generative adversarial nets.” *NIPS* 2014.

Unsupervised cross-domain instance alignment

- Goal: Transfer style from source to target
- No pair-wise correspondence (CycleGAN, DiscoGAN and DualGAN)



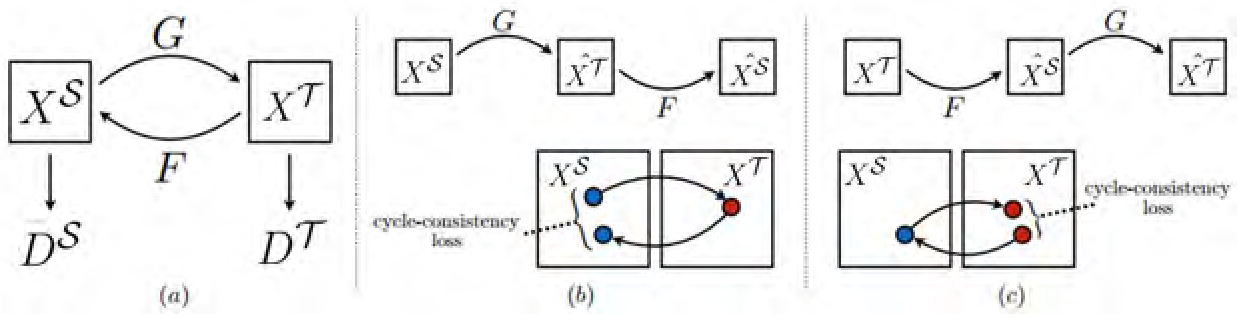
First, learn relations between handbags and shoes



DiscoGAN (Kim et al., 2017)

Then, generate a shoe while retaining key attributes of handbags

Cycle GAN Model architecture



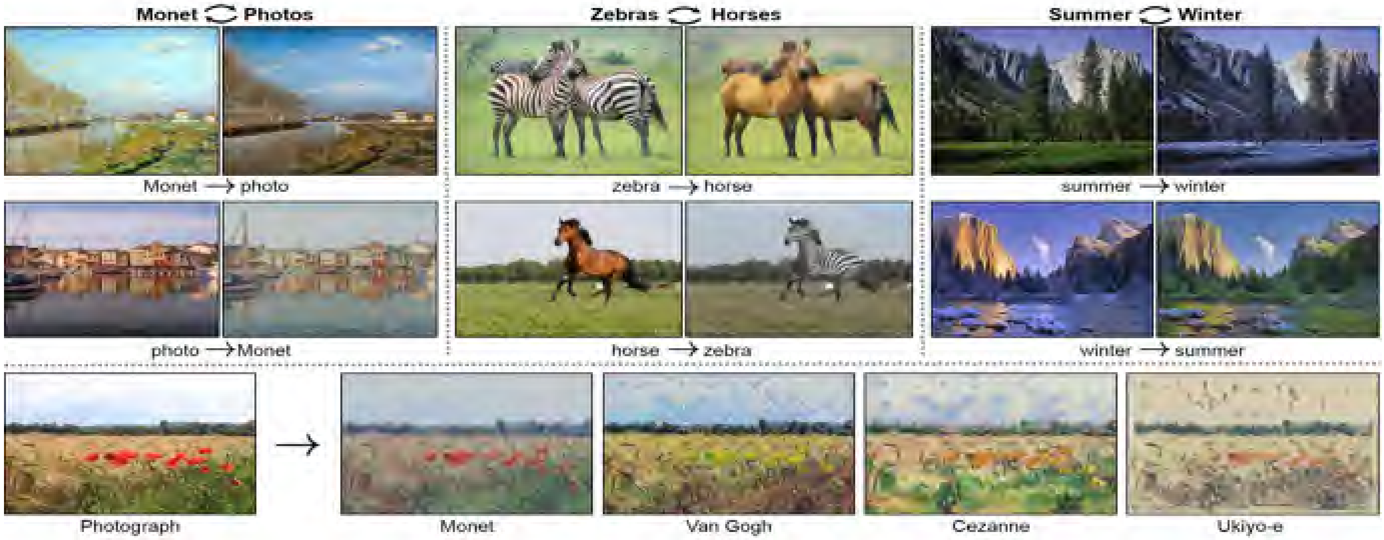
$$\mathcal{L}(G,F,D^S,D^T) = \underbrace{\mathcal{L}_{GAN}(G,D^T,X^S,X^T) + \mathcal{L}_{GAN}(F,D^S,X^S,X^T)}_{\text{Adversarial loss}} + \lambda\mathcal{L}_{cyc}(G,F)$$

- G: mapping from the source to the target, F: inverse mapping
- Total loss = Adversarial loss + cycle-consistency loss

Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *arXiv preprint arXiv:1703.10593* (2017).

Alignment results

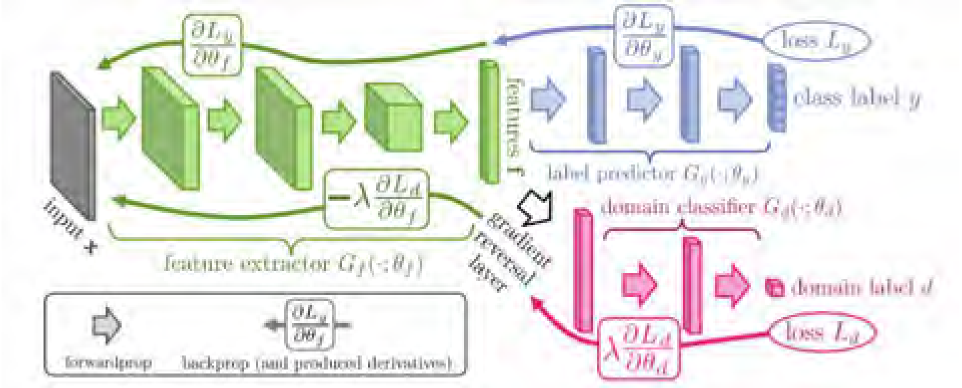
- CycleGAN can fool human annotators on 25% of trials



More image translation results produced by CycleGAN (Zhu et al., 2017)

Adversarial domain adaptation

- target domain has no labels; find **common** feature space between the source and target by formulating a min-max game. Two constraints:
  - Helpful for the source domain classification task
  - indistinguishable between the source and target domain



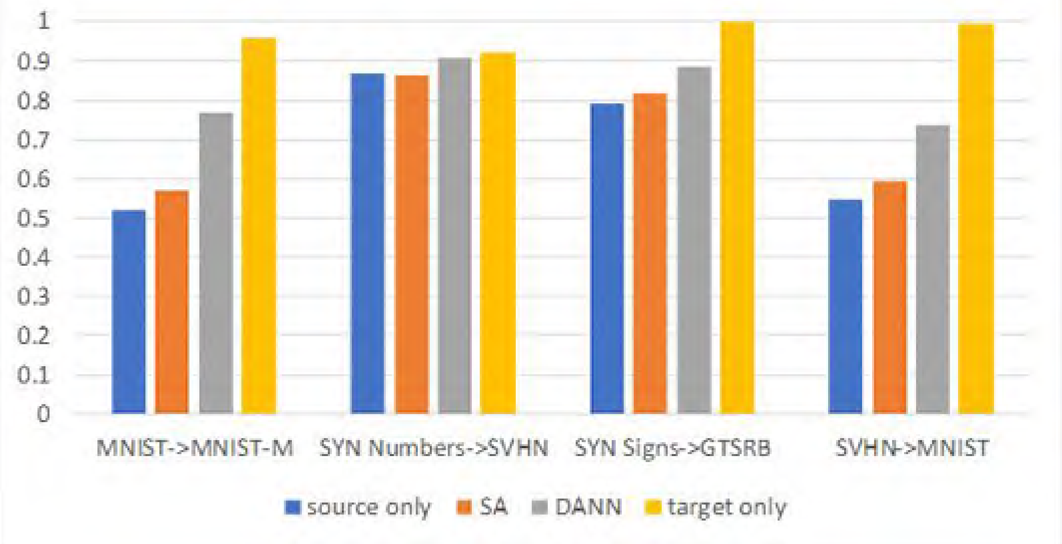
Minimize source label classification error

Maximize domain classification error

Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *Journal of Machine Learning Research* 17.59 (2016): 1-35.

Classification accuracies for multiple domain adaptation pairs

- Four source-target domain adaptation



- Source only: lower bound performance, no adaptation is performed
- Target only: upper bound performance, train the classifier with known target domain labels
- Subspace Alignment (SA) (Fernando *et al*., 2013)
- **Domain Adversarial Neural Networks (DANN) (Ganin, Yaroslav, *et al*., 2016)**

迁移学习应用案例1: 解决大额消费金融的困境 (第四范式)

在千万量级微信公众号客户中，挖掘近期有购车意向的客户，通过微信营销购车分期业务。客户可点击其中链接提交申请。

难点：新渠道，成功办理客户<100

方法：基于全渠道营销数据（成功次数>1亿），帮助汽车分期贷款模型学习

效果：与SAS相比，营销响应率提升200%+

Dai, Wenyuan et al. 2017

跨领域舆情分析：IJCAI 2017: Zheng Li, Yu Zhang, et al.

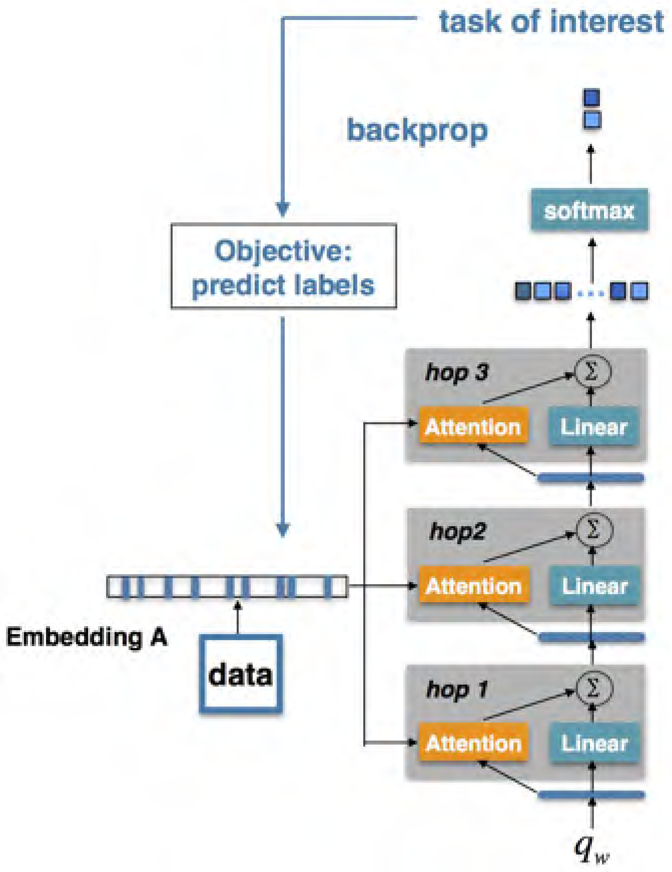
"End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification", IJCAI 2017, Zheng Li, et al.

- 问题：如何自动找出 Pivot 关键词？

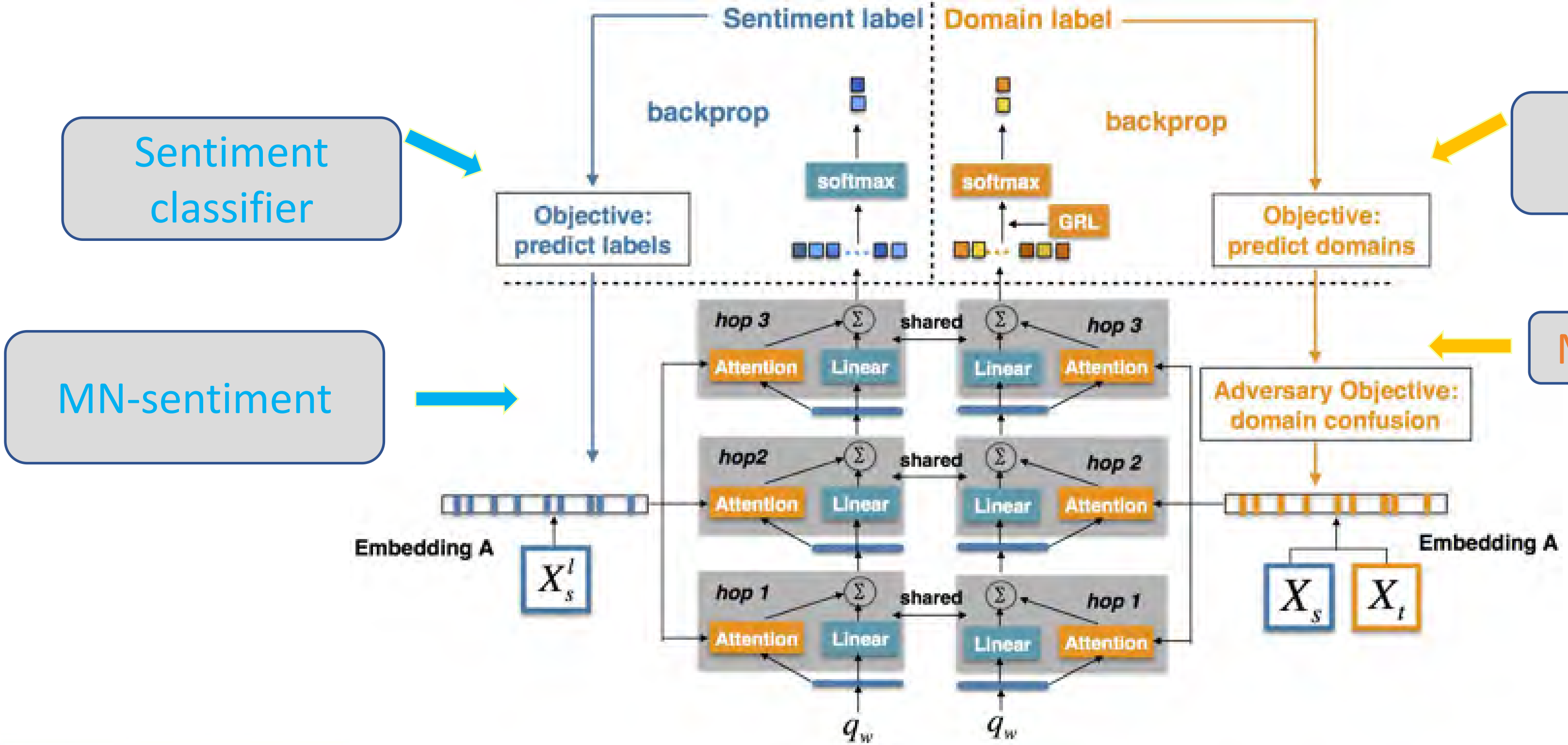
舆情	Books (源领域)	Restaurant (目标领域)	舆情
	**Great** books. His characters are **engaging**.	The food is **great**, and the drinks are **tasty** and **delicious**.	
	It is a very **nice** and **sobering** novel.	The food is very **nice** and **tasty**, and we'll go back again.	
	A **awful** book and it is a little **boring**.	**Shame** on this place for the **rude** staff and **awful** food.	

Memory Networks

- Capture evidence (sentences, words) by **interest** via attention mechanism



同时使用Memory Network 和 GAN



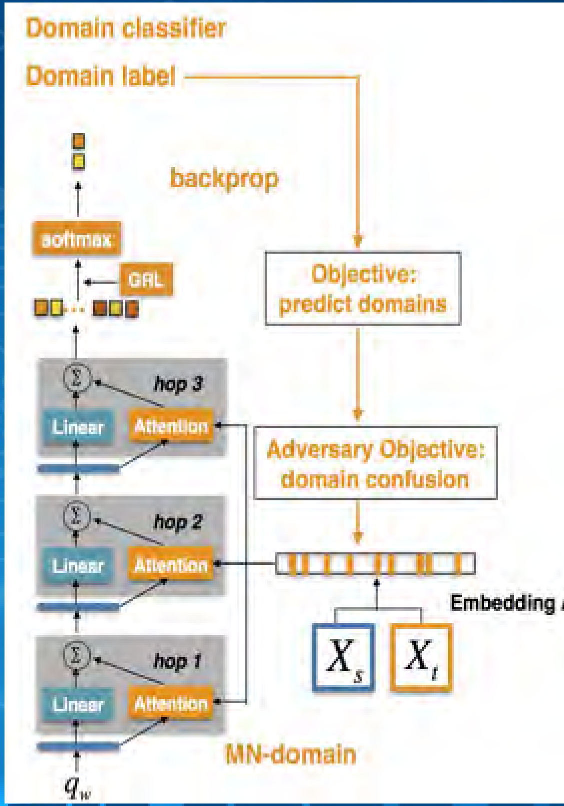
- Feed the output vector $\hat{v}_d$ of GRL to the softmax layer for domain classification:

$$\mathrm{d} = softmax(W_d \hat{v}_d + b_d)$$

- Minimize the cross-entropy for all data in source and target domains, except adversarial part:

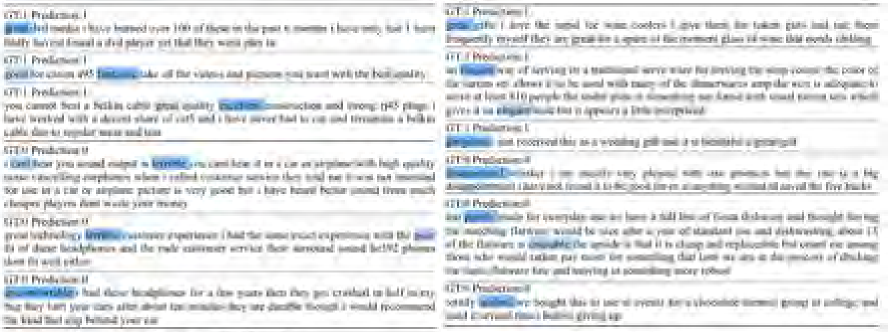
-

$$L^{dom} = -\frac{1}{N_s + N_t} \sum_{i=1}^{N_s+N_t} \hat{\mathrm{d}}_i \log(\mathrm{d}_i) + (1 - \hat{\mathrm{d}}_i) \log(1 - \mathrm{d}_i)$$



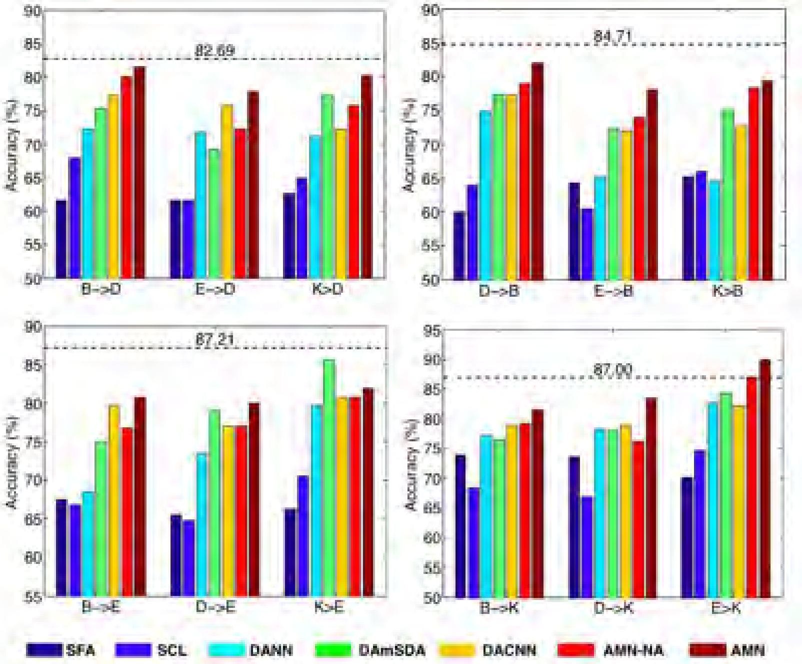
跨领域舆情分析结果

Domain	#Train	#Test	#Unlab.	% Neg.
Books	1600	400	6000	13.45%
DVD	1600	400	34741	21.47%
Electronics	1600	400	13153	11.92%
Kitchen	1600	400	16785	17.82%



(a) Electronics domain

(b) Kitchen domain



迁移学习应用案例2: 上海汽车汽车的互联网汽车分类问题

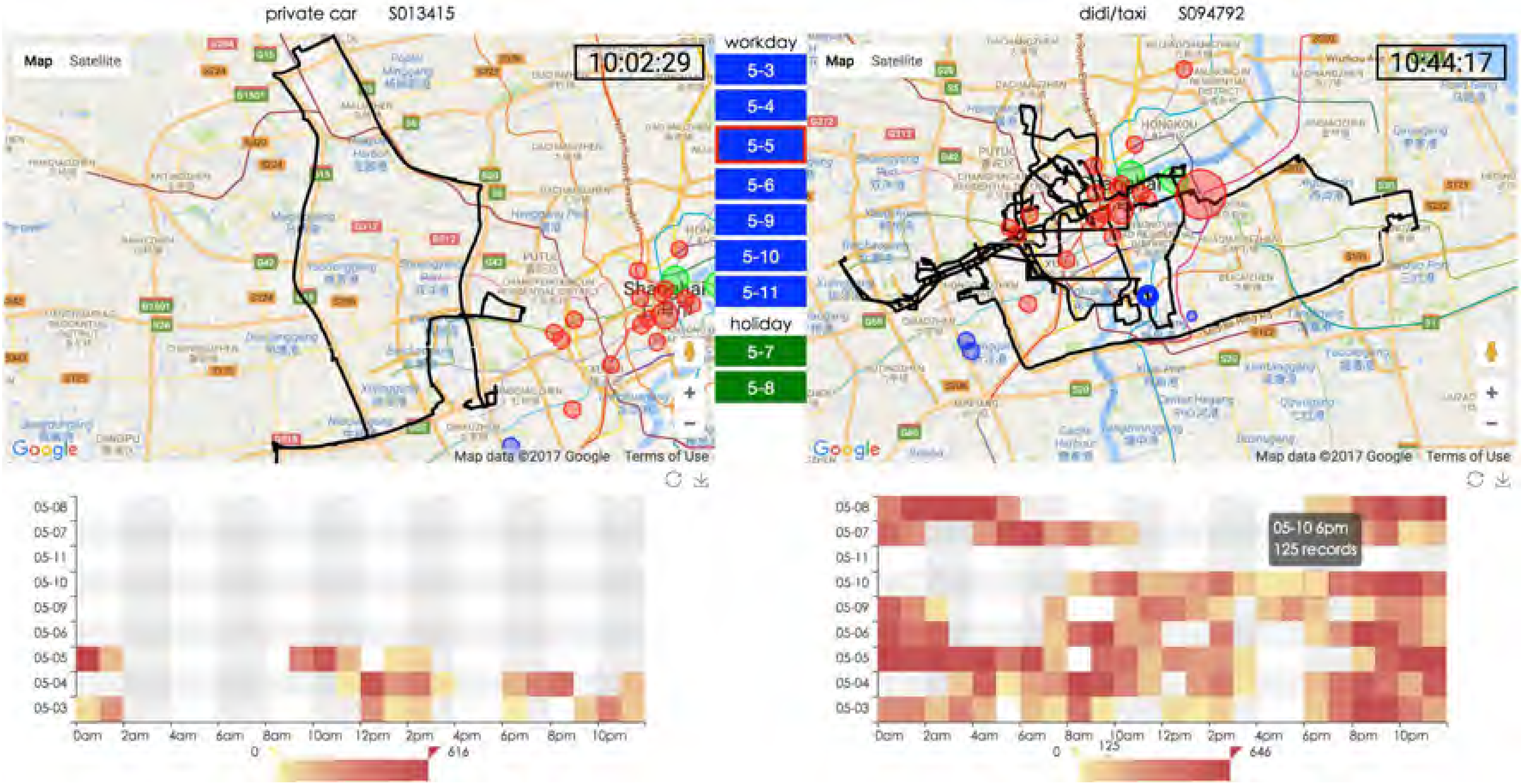
- 共享汽车：公用私用分类
- GPS + Time, 1/15 sec, no labels，7 Days, 10,000 cars

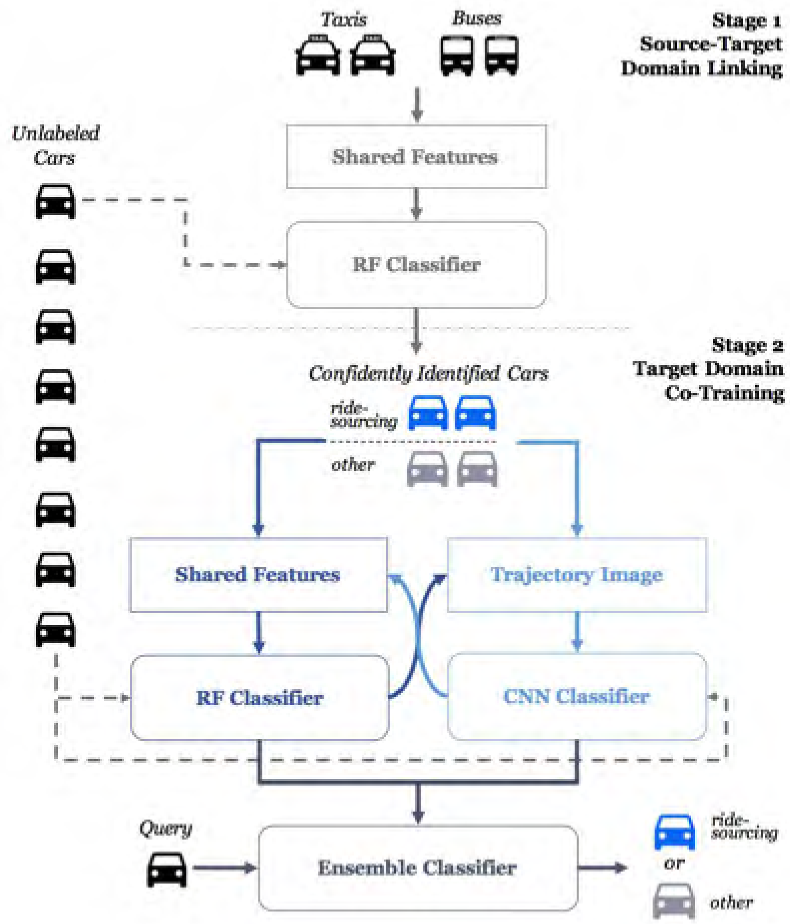
共享

?

私用

Wang, Leye , et al. 2017


private car S013415
didi/taxi S094792
10:02:29
10:44:17
workday
5-3
5-4
5-5
5-6
5-9
5-10
5-11
holiday
5-7
5-8
Map
Satellite
Map data ©2017 Google
Terms of Use
05-08
05-07
05-11
05-10
05-09
05-06
05-05
05-04
05-03
0am
2am
4am
6am
8am
10am
12pm
2pm
4pm
6pm
8pm
10pm
0
616
05-10 6pm
125 records
125
646



CoTrans Framework

Stage 1: Source-Target Domain Linking

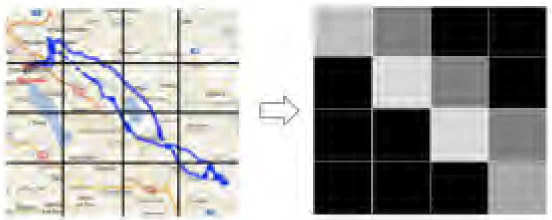
Shared (transferable) Features: **dist., cov.**

Random Forest (RF)

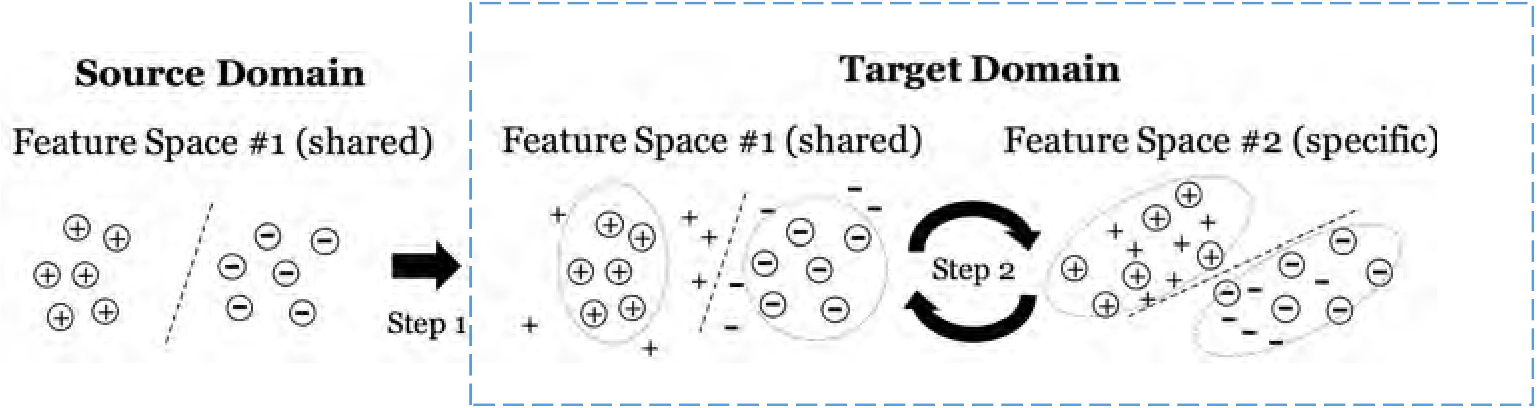
Stage 2: Target Domain Co-training

On RF + CNN (**trajectory image**)

Trajectory image: the brighter color, the longer stay time in that cell.



Stage 2: Co-Training



- 1. In Feature Space 1, train new model M1 and find samples by M1 (First time M1 comes from Source Domain)
- 2. In Feature Space 2, find image features of samples from Step 1, train model M2; Find new samples by M2

迁移学习 + 深度学习 = 深度学习的迁移模型

感谢：李正，王伟彦，魏颖， 谭奔博士，张宇博士，第四范式

References

[1] Bousmalis, Konstantinos, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. "Domain separation networks." In *Advances in Neural Information Processing Systems*, pp. 343-351. 2016.

[2] Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." In *International Conference on Machine Learning*, pp. 1180-1189. 2015.

[3] Ghifary, Muhammad, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. "Deep reconstruction-classification networks for unsupervised domain adaptation." In *European Conference on Computer Vision*, pp. 597-613. 2016.

[4] Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." In *Advances in neural information processing systems*, pp. 469-477. 2016.

[5] Long, Mingsheng, Yue Cao, Jianmin Wang, and Michael Jordan. "Learning transferable features with deep adaptation networks." In *International Conference on Machine Learning*, pp. 97-105. 2015.

[6] Long, Mingsheng, Jianmin Wang, and Michael I. Jordan. "Deep transfer learning with joint adaptation networks." In *International Conference on Machine Learning*, 2017.

[7] Tzeng, Eric, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. "Deep domain confusion: Maximizing for domain invariance." *arXiv preprint arXiv:1412.3474* (2014).

[8] Tzeng, Eric, Judy Hoffman, Trevor Darrell, and Kate Saenko. "Simultaneous deep transfer across domains and tasks." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4068-4076. 2015.

[9] Tzeng, Eric, Judy Hoffman, Kate Saenko, and Trevor Darrell. "Adversarial discriminative domain adaptation." *arXiv preprint arXiv:1702.05464* (2017).