



机器之心

Broadview®  
www.broadview.com.cn

# 迁移学习发展现状及案例探究

主讲人：王晋东



电子工业出版社  
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY  
http://www.phei.com.cn



# 嘉宾介绍

- 现任微软亚洲研究院研究员
- 2019年于中国科学院计算技术研究所获得博士学位
- 研究方向：迁移学习、机器学习
- 博士论文《面向普适计算环境的自适应迁移学习方法研究》
  - 获中国科学院优秀博士学位论文奖
- 《迁移学习简明手册》
- 知乎&微信公众号：王晋东不在家





机器之心

Broadview®  
www.broadview.com.cn

# 书籍介绍



顶级AI专家推荐：  
“实践性强” “通俗易懂” “深入浅出”

## 好评来袭

本书配有用于实践的源码和数据集，增加了动手练习的环节，提高了趣味性。作为长期耕耘在迁移学习这一人工智能领域的学者和业界首本迁移学习著作的作者，我强烈力荐这本书给有志于从事迁移学习研究的同学，更快地入门和学习！

——杨强

微众银行首席人工智能官、香港科技大学讲席教授ACM/AAAI/IEEE Fellow

本书深入浅出、系统性地介绍了主要的迁移学习方法，并结合多个领域的应用进行示例分析，为从事相关技术的研究人员提供了非常有益的参考。

——陶建华

中国科学院自动化研究所研究员  
模式识别国家重点实验室副主任

陈益强和王晋东两位老师通俗易懂地介绍了迁移学习的来龙去脉——不仅涵盖了基本的理论脉络、具体的方法和技术，还介绍了其广泛的应用案例和未来的发展方向和前沿问题，为人工智能初学者提供了一份难得的、快速入门的学习和研究资料。

——汪军

伦敦大学学院计算机系教授

迁移学习是机器学习的一个重要研究分支，有广泛的应用价值。该书叙述简洁明了、内容丰富详实，对希望了解并应用迁移学习的读者很有帮助！

——周志华

南京大学教授，ACM/AAAI/IEEE Fellow

迁移学习已成为人工智能的一项核心技术，在计算机视觉、自然语言语音处理、强化学习中得到了广泛的应用。本书语言简洁，内容丰富，相信可以启发读者举一反三、触类旁通，帮助读者更好地解决手头的问题。

——秦涛

微软亚洲研究院首席研究经理  
中国科技大学兼职教授

本书的一大亮点，是对“两头”的把握：一是源头，抓问题和场景，做到“师出有名”，讲清楚针对什么问题、用在哪里；二是笔头，抓代码与实践，做到“落地结果”，在实战中巩固和深化对技术的理解。相信这本书能带给读者思考与实践的双重乐趣，在算力爆炸的时代反思机器学习的高效之道！

——陈光

北京邮电大学副教授  
新浪微博@爱可可-爱生活

## 《迁移学习简明手册》读者评论

• Frank 2018-04-12

我第一时间联系了印刷社，打印出来读！！

• 五百三十三号2018-04-10

楼主好人有好报哇

16回复推荐折叠踩 举报删除

• 赵欣2018-04-10

感谢作者，非常赞的工作，非常幸运能在刚准备入门迁移学习就能阅读这么好的资料！

6回复推荐折叠踩 举报删除

• lingay03-02

对我的研究很有帮助，好人一生平安

• 科学甜美机器2019-03-28

为大佬打call，刚拜读完，感动到哭泣😭

• KwanHeung2018-04-13

真的是纯干货，希望有机会能交流学习！

1回复推荐折叠踩 举报删除

• 剑飞来自火星2018-04-12

大作，不久的将来一定会被翻译成英文的哈哈哈哈哈



电子工业出版社  
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY  
http://www.phei.com.cn

# 书籍主要内容



<https://item.jd.com/13283188.html>



# 主题分享

## 迁移学习发展现状及案例探究

1. 迁移学习简介
2. 迁移学习基本方法
3. 迁移学习最新进展
4. 迁移学习在语音识别中的应用



# Artificial intelligence?

- More artificial, more intelligence... (有多少人工，就有多少智能)



## 数据标注实习生 (中文文职)

工作内容: 听取 Apple 用户和 Siri 的对话, 修改其中的错别字, 修改标点。对相关文档进行标注和解释。

要求:

大专以上学历; 细心耐心, 接受重复性工作;

基本计算机操作。短期项目: 2018.02.26-2018.06月中旬

工作时间: 8:30-5:30 周末双休薪资待遇: 2200-2700

工作地点: 大连软件园12号楼

简历投递: [tingtingsun@wistronits.com](mailto:tingtingsun@wistronits.com) [收起全文](#)

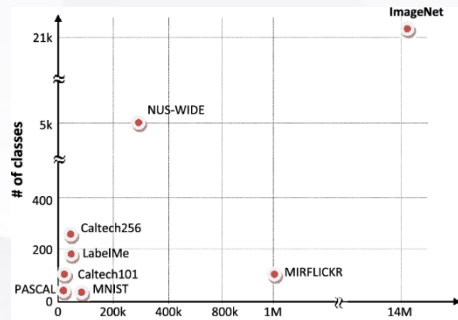
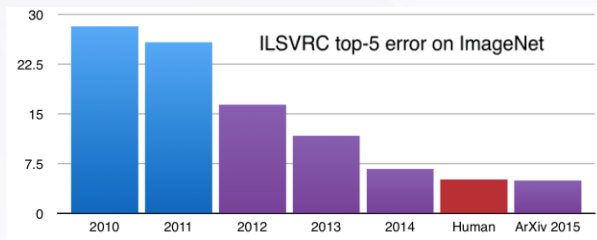
大约向北300英里的河北南宮市, 侯夏梦经营着一家数据工厂, 由她婆家拥有的一间旧水泥厂改建而来。她在大学毕业后第一份工作是给旷视科技做面部标记, 这是一家价值20亿美元的中国面部识别公司, 其最著名的技术平台名为Face++。到目前, 一些面部识别系统可以比识别她的朋友更快地识别她, 她说这是因为“我的脸在原始数据库里。”

Figure credit: <https://cn.nytimes.com/technology/20181126/china-artificial-intelligence-labeling/>

# 背景

## • 大数据&大模型

— 理想: 大量数据, 好的标注 → 好的模型



— 现实: 绝大多数数据是无标注的

• 收集大量标注并从头训练模型: 耗时且费力

# 迁移学习

- 利用辅助领域已有的知识，快速构建待学习领域的模型。



	3		?
	2	5	?
		3	?

源域  
数据

减小差异  
知识迁移



	3		1
	2	5	3
		3	5

- 数据标定
- 跨领域模型
- 强泛化模型
- .....

- 核心思想

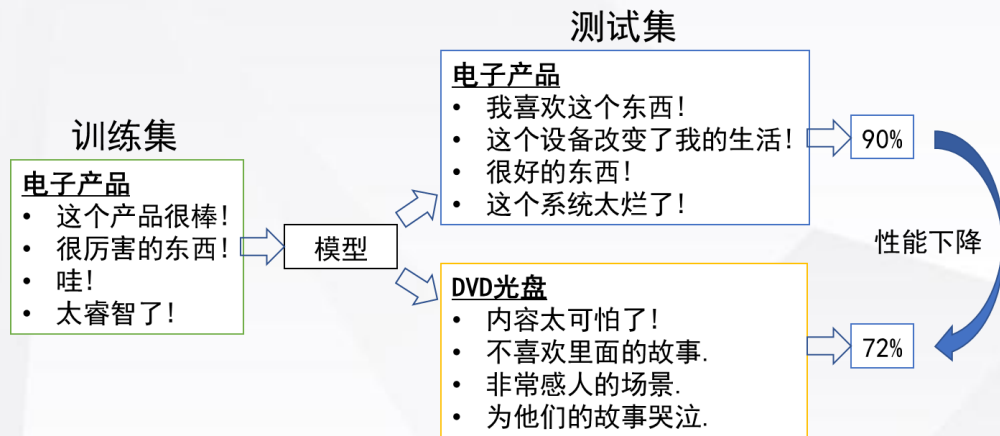
- 找到不同任务的相关性
- “举一反三”，但不要“东施效颦”（负迁移）





# 举例

## • 商品评论



## • 图像识别



图像数据集1

图像数据集2

# 迁移学习的广泛应用



语料匮乏条件下不同语言的相互翻译学习



不同视角、不同背景、不同光照的图像识别



不同用户、不同设备、不同位置的行为识别



不同领域、不同背景下的文本翻译、舆情分析



不同用户、不同接口、不同情境的人机交互

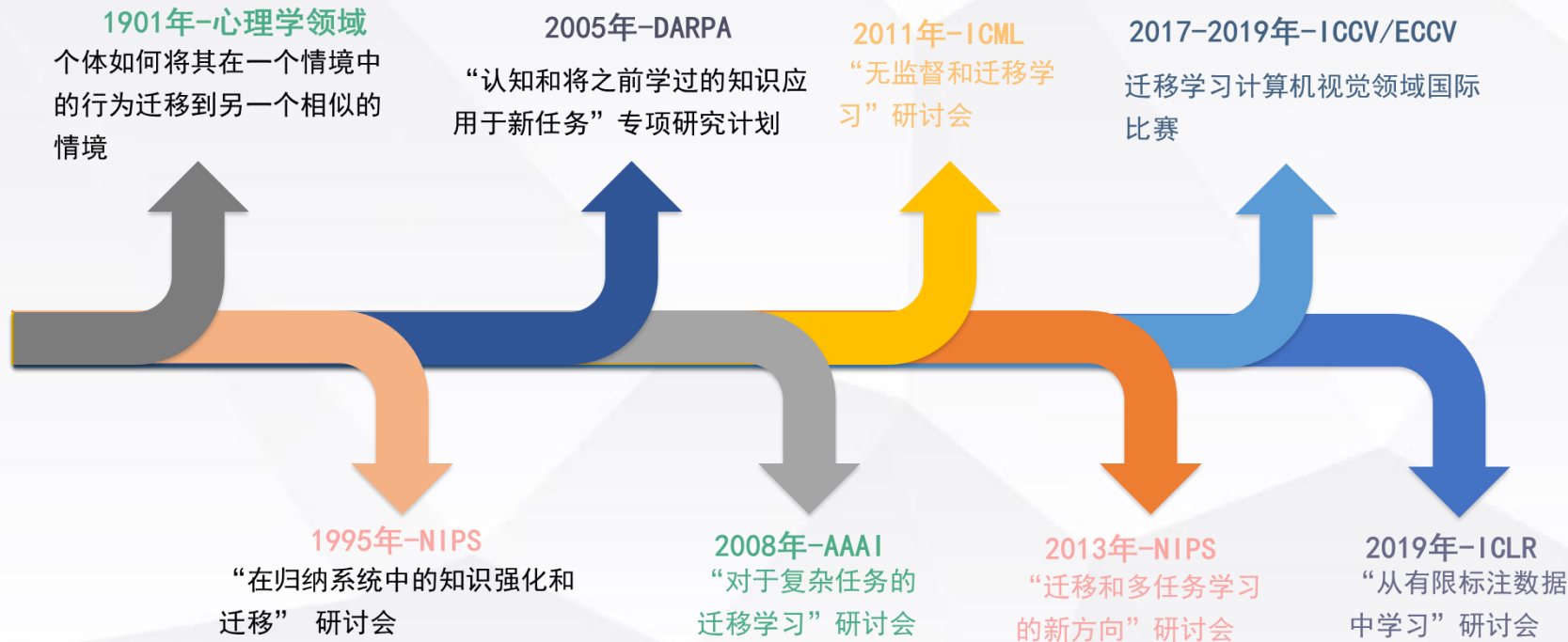


不同场景、不同设备、不同时间的室内定位

对应《迁移学习导论》第1章第5节。



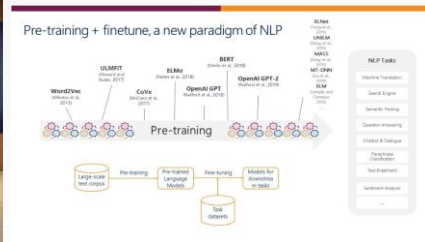
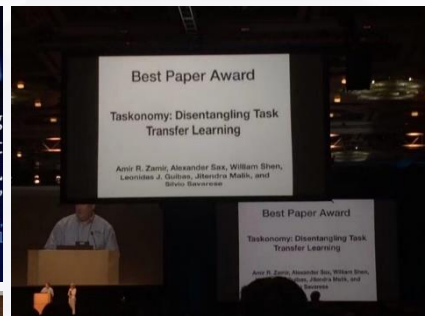
# 迁移学习历史



对应《迁移学习导论》第1章第6节。

# 最新进展

Important milestone



NIPS 2016  
Tutorial

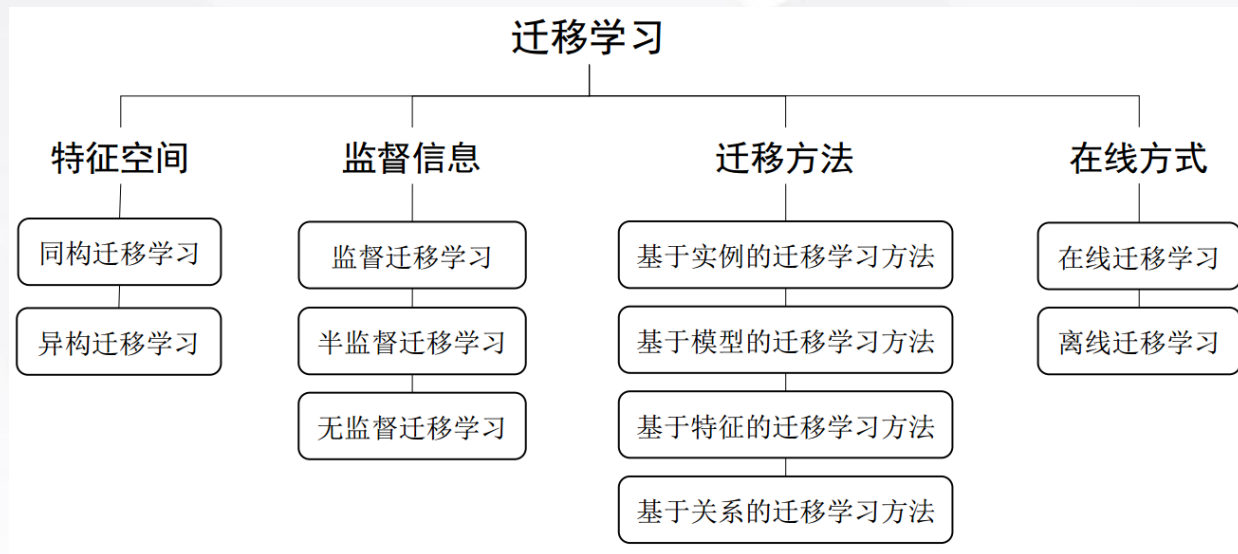
CVPR 2018  
Best paper

IJCAI 2018  
Ads challenge winner

ACL 2019  
Opening talk

ACL 2020 Best paper honorable mention: don't stop pretraining...

# 迁移学习研究领域





# 为什么要用迁移学习？

矛盾	传统机器学习	迁移学习
大数据与少标注	增加人工标注，但是昂贵且耗时	数据的迁移标注
大数据与弱计算能力	只能依赖强大计算能力，但是受众少	模型迁移
有限数据与模型泛化能力	无法满足泛化要求	领域泛化、元学习等
普适化模型与个性化需求	通用模型无法满足个性化需求	模型自适应调整
特定应用的需求	冷启动问题无法解决	数据迁移



# 数据和计算

- 大数据 vs. 少标注
- 大数据 vs. 弱计算能力
- 如何“四两拨千斤”：
  - 利用已有的大数据、预训练模型完成自己的任务？

## Money Is All You Need

Nick Debu  
Tokyo Institute of Bamboo Steamer

### Abstract

Transformer-based models routinely achieve state-of-the-art results on a number of tasks but training these models can be prohibitively costly, especially on long sequences. We introduce one technique to improve the performance of Transformers. We replace NVIDIA P100s by TPUs, changing its memory from huge GB to piyo GB. The resulting model performs on par with Transformer-based models while being much more "TSUYO TSUYO".

Table 1 : Training time and top-1 1-crop validation accuracy with ImageNet/ResNet-50

	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [7]	256	Tesla P100 x8	Caffe	29 hours	75.3%
Goyal et al. [1]	8K	Tesla P100 x256	Caffe2	1 hour	76.3%
Smith et al. [4]	8K→16K	full TPU Pod	TensorFlow	30 mins	76.1%
Akiba et al. [5]	32K	Tesla P100 x1024	Chainer	15 mins	74.9%
Jia et al. [6]	64K	Tesla P40 x2048	TensorFlow	6.6 mins	75.8%
This work	34K→68K	Tesla V100 x2176	NNL	224 secs	75.03%



# 普适化模型

- 普适化模型 vs. 个性化需求
  - 通常需要对设备、环境、用户作具体优化
  - 对于不同用户，需要不同的隐私处理方式
- 如何用一个普适化模型来更好地“普惠众人”？

2021年6月10日通过《中华人民共和国数据安全法》

对应《迁移学习导论》第1章第3节。







# 迁移学习问题定义

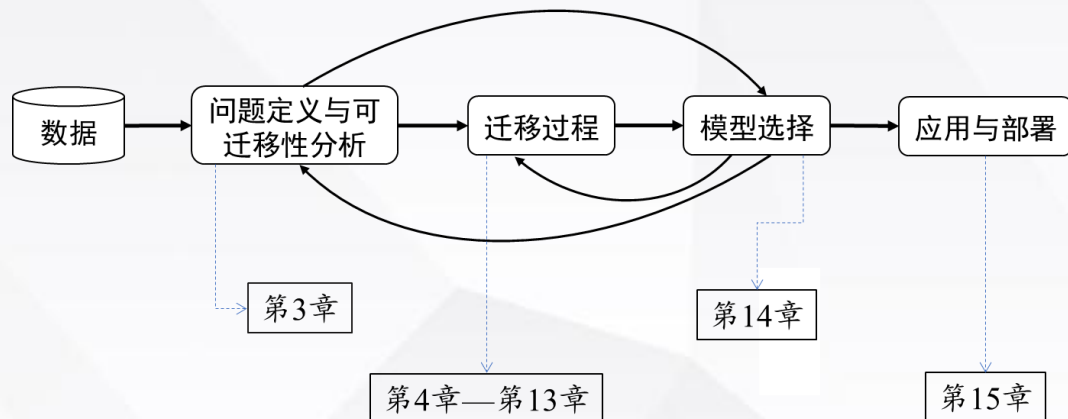
**定义 3** 给定一个源域  $\mathcal{D}_s = \{\mathbf{x}_i, y_i\}_{i=1}^{N_s}$  和目标域  $\mathcal{D}_t = \{\mathbf{x}_j, y_j\}_{j=1}^{N_t}$ , 其中  $\mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$ 。迁移学习的目标是当以下三种情形:

1. 特征空间不同, 即  $\mathcal{X}_s \neq \mathcal{X}_t$ ;
2. 标签空间不同, 即  $\mathcal{Y}_s \neq \mathcal{Y}_t$ ;
3. 特征和类别空间均相同、概率分布不同, 即  $P_s(\mathbf{x}, y) \neq P_t(\mathbf{x}, y)$ ,

至少有一种成立时, 利用源域数据去学习一个目标域上的预测函数  $f: \mathbf{x}_t \mapsto y_t$ , 使得  $f$  在目标域上拥有最小的预测误差 (用  $\epsilon$  来衡量):

$$f^* = \arg \min_f \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}_t} \epsilon(f(\mathbf{x}), y). \quad (2.5.1)$$

# 迁移学习流程





# 迁移学习基本方法

- 三大类基本方法

## 样本权重迁移法

- 选择或重新给不同样本以适当的权重

## 特征变换迁移法

- 学习更好的特征表示

## 预训练迁移方法

- 重用已有的预训练模型

# 通用计算框架

$$f^* = \arg \min_{f \in \mathcal{H}} \frac{1}{N_s} \sum_{i=1}^{N_s} \ell(v_i f(\mathbf{x}_i), y_i) + \lambda R(T(\mathcal{D}_s), T(\mathcal{D}_t))$$

方法大类	问题设定	求解目标
样本权重迁移法	$T(\mathcal{D}_s), T(\mathcal{D}_t) = \mathcal{D}_s, \mathcal{D}_t$	$v_i$
特征变换迁移法	$v_i = 1, \forall i$	$T$
模型预训练迁移法	$v_i = 1, \forall i, R(T(\mathcal{D}_s), T(\mathcal{D}_t)) := R(\mathcal{D}_t; f_s)$	SRM

**定理 1** 基于  $\mathcal{H}$ -divergence 的目标域误差界 令  $\mathcal{H}$  表示一个 VC 维为  $d$  的假设空间, 给定从源域上以  $iid$  (Independent and identically distributed) 方式采样的大小为  $m$  的样本集, 则至少以  $1 - \delta$  的概率, 对于任意一个  $h \in \mathcal{H}$  有

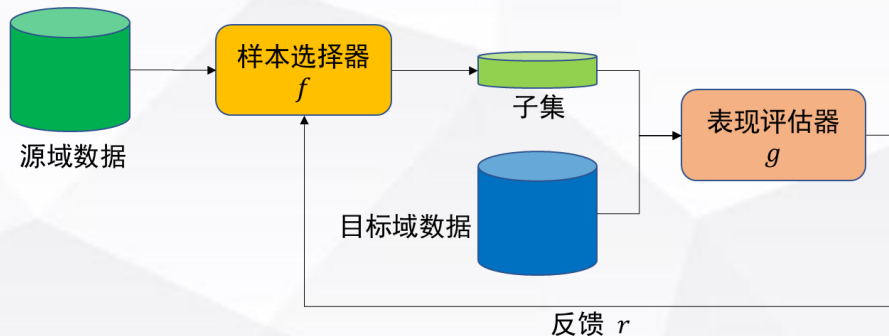
$$\epsilon_t(h) \leq \hat{\epsilon}_s(h) + d_{\mathcal{H}}(\hat{D}_s, \hat{D}_t) + \lambda^* + \sqrt{\frac{4}{m} (d \log \frac{2em}{d} + \log \frac{4}{\delta})}, \quad (4.5.5)$$

其中,  $e$  是自然底数,  $\lambda^* = \epsilon_s(h^*) + \epsilon_t(h^*)$  是理想联合误差,  $h^* = \arg \min_{h \in \mathcal{H}} \epsilon_s(h) + \epsilon_t(h)$  是在源域和目标域上的最优分类器。

对应《迁移学习导论》第4章第5节“迁移学习理论”。

# 样本权重迁移法

- 样本选择
  - 样本选择器  $f: f(D_s) \rightarrow D'_s$
  - 表现评估器  $g: g(D'_s, D_t)$
  - 反馈  $r$ : 选择的是不是好样本?



# 特征变换迁移法

- 显式距离:  $R(\cdot, \cdot) \approx \text{Distance}(D_s, D_t)$

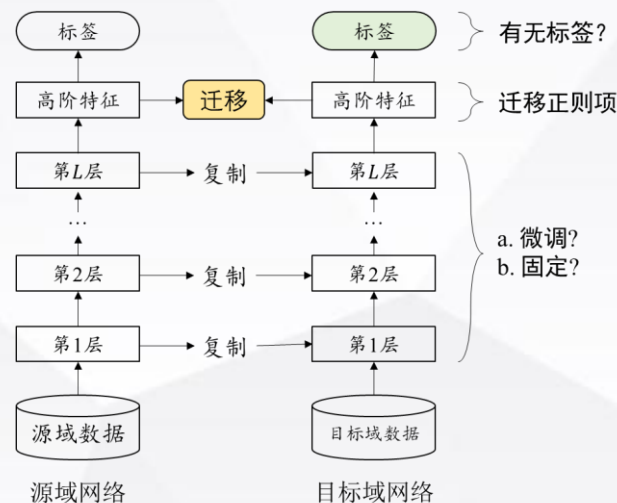
- Goal:  $f^* = \arg \min_f \frac{1}{m} \sum_{i=1}^m L(f(x_i), y_i) + \lambda \cdot \text{Distance}(D_s, D_t)$

- Kernel-based distance

- Maximum mean discrepancy (MMD) [1]
  - KL-divergence

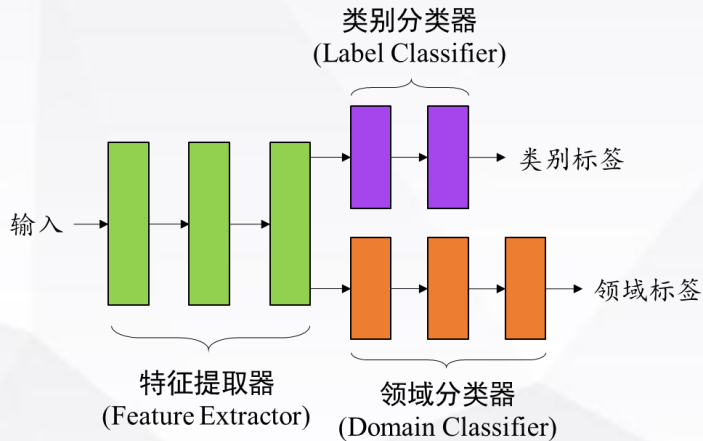
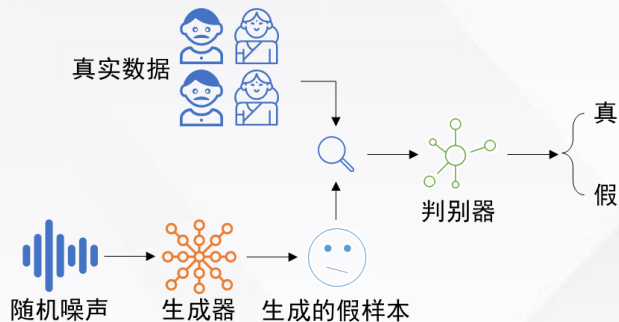
- Geometrical distance

- Geodesic flow kernel (GFK) [2]
  - Correlation alignment (CORAL) [3]
  - Riemannian manifold [4]



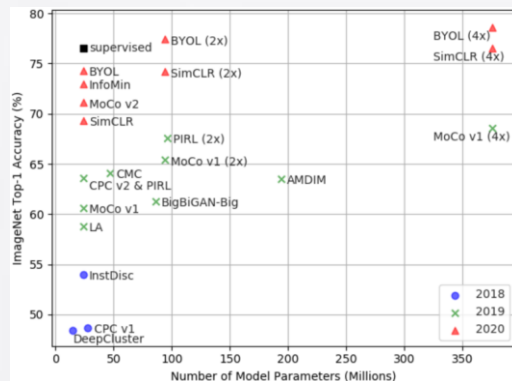
# 特征变换迁移法

- 隐式距离：对抗

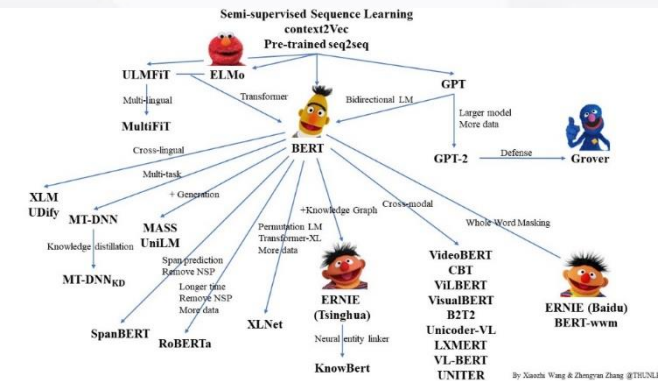


# 预训练迁移法

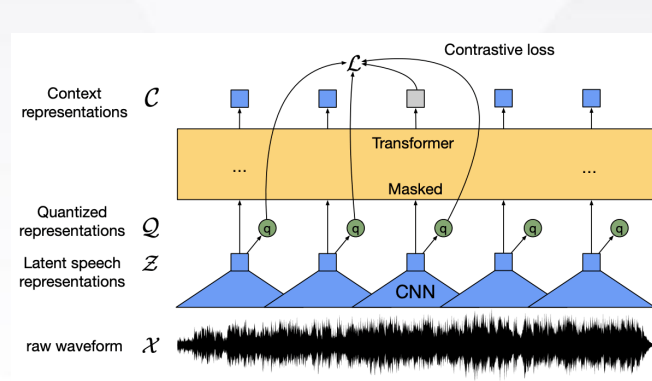
## • 预训练@CV



## • 预训练@NLP

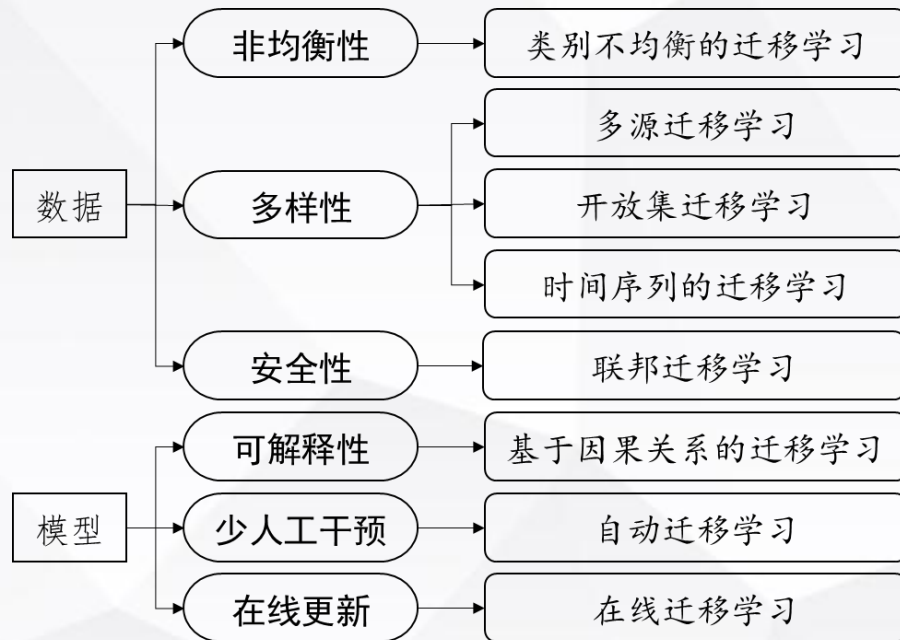


## • 预训练@Speech





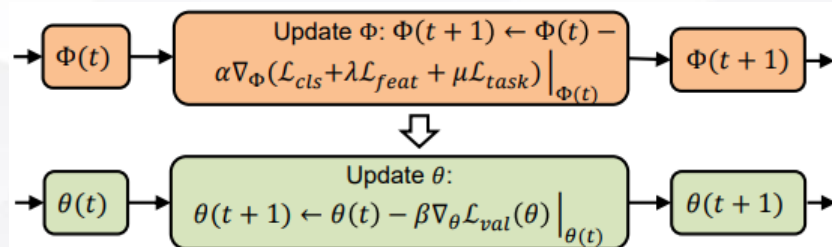
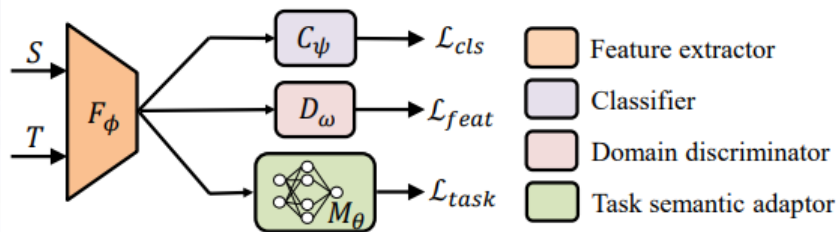
# 迁移学习最新进展



# 自动迁移学习

- Learning to match (L2M)
  - 用哪个分布距离? MMD, adversarial, coral...
  - L2M: 不用. 我们可以让模型自动学习分布适配

$$\mathcal{L}_{val} = \mathbb{E}_{\mathbf{x} \sim \mathcal{P}} \sigma(M_{\theta}(\mathbf{x}; \Phi(t+1)) - M_{\theta}(\mathbf{x}; \Phi(t))),$$



Wang et al. Learning invariant representations across domains and tasks. <https://arxiv.org/pdf/2103.05114.pdf>

# 子结构迁移

- Substructural domain adaptation
  - 自适应地发现数据中的子结构
  - 高性能、高准确性: 5x faster

Table 1: Comparison between different matching schemes

Type	Assumption	Formulation	Limitations
Domain-level	domain-invariant features	$p(\mathbf{x}_s) \longleftrightarrow p(\mathbf{x}_t)$	too coarse matching
Class-level	class-invariant features	$p(\mathbf{x}_s y_s) \longleftrightarrow p(\mathbf{x}_t y_t)$	coarse matching
Substructure-level	substructure-invariant features	$p(\mathbf{x}_s, y_s c_s) \longleftrightarrow p(\mathbf{x}_t c_t)$	affected by noise, low-efficiency
Sample-level	no strict restrictions	$(\mathbf{x}_s, y_s) \longleftrightarrow \mathbf{x}_t$	



Target Figure 3: The learning process of SSDA

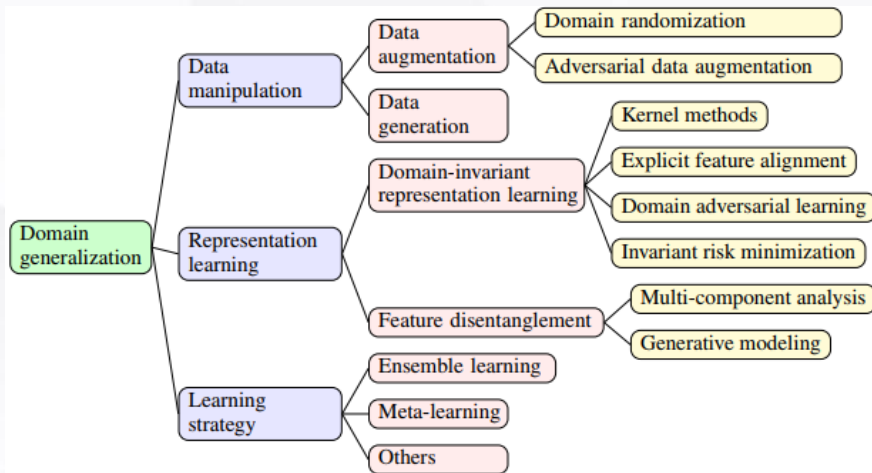
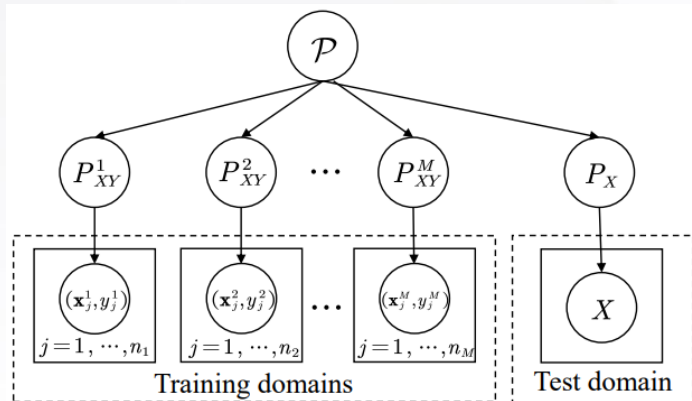
Lu et al. Cross-domain activity recognition via substructural optimal transport.

<https://arxiv.org/pdf/2102.03353.pdf>

# 领域泛化

## • Out-of-distribution generalization

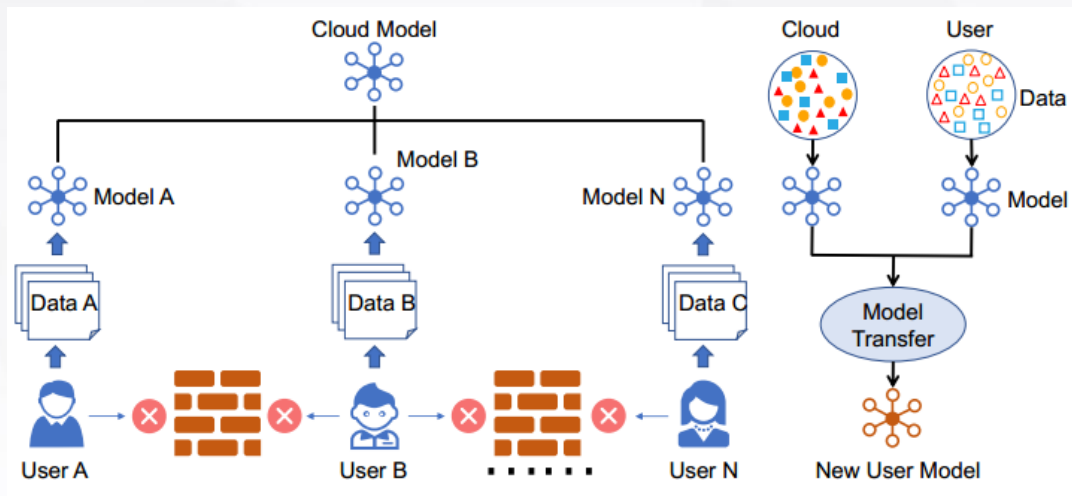
- 超越domain adaptation: 不需要访问目标数据, 学习一个泛化能力强的模型



Wang et al. Generalizing to Unseen Domains: A Survey on Domain Generalization. <https://arxiv.org/pdf/2103.03097.pdf>, IJCAI-21.

# 联邦迁移学习

- FedHealth: 个性化、隐私保护的迁移学习系统

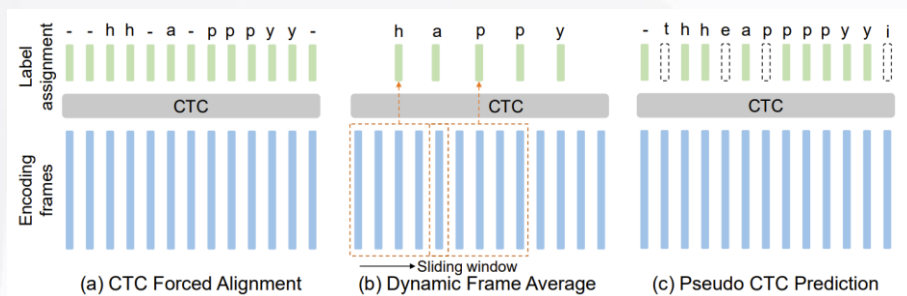


- Chen Y., Wang J., et al. FedHealth: a federated transfer learning framework for wearable health. IJCAI 2019 (Federated learning workshop)
- Chen Y, Lu W, Wang J, et al. FedHealth 2: Weighted Federated Transfer Learning via Batch Normalization for Personalized Healthcare[J]. arXiv preprint arXiv:2106.01009, 2021.

# 迁移学习与语音识别

- 跨设备、跨环境的迁移学习

- CMatch: Character-level distribution matching



Task	Source-only	MMD	ADV	CMatch
M → P	23.87	20.87	21.11	<b>20.38</b>
M → R	25.21	22.21	22.27	<b>21.77</b>
P → M	31.15	27.22	28.29	<b>26.17</b>
P → R	23.99	21.90	21.74	<b>20.43</b>
R → M	32.45	28.27	29.95	<b>27.77</b>
R → P	23.48	21.09	21.23	<b>20.58</b>
Average	26.69	23.59	24.10	<b>22.85</b>

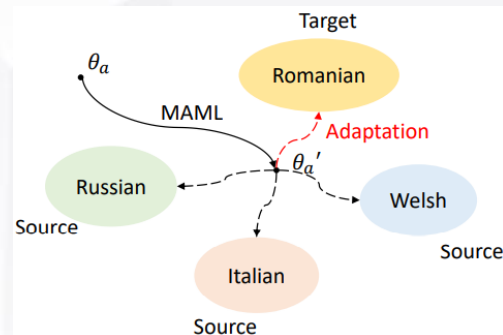
+16.50%

Target	Source-only	MMD	ADV	CMatch
Rain	38.21	33.61	34.65	<b>32.90</b>
Wind	29.70	26.06	26.73	<b>23.12</b>
Laughter	33.36	29.85	30.41	<b>28.55</b>
Average	33.76	29.84	30.60	<b>28.19</b>

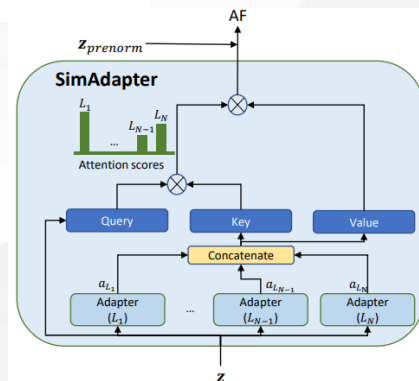
Hou W, Wang J, Tan X, et al. Cross-domain Speech Recognition with Unsupervised Character-level Distribution Matching[J]. arXiv preprint arXiv:2104.07491, 2021.

# 迁移学习与语音识别

- 跨语言、低资源的迁移学习  
– MetaAdapter & SimAdapter

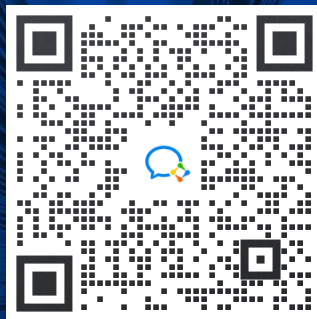


Target Language	Hybrid DNN/HMM	Transformer	Head	Full-FT	Adapter	SimAdapter	MetaAdapter	SimAdapter +
Romanian (ro)	70.14	97.25	63.98	53.90	48.34	47.37	<b>44.59</b>	47.29
Czech (cs)	63.15	48.87	75.12	34.75	37.93	35.86	37.13	<b>34.72</b>
Breton (br)	-	97.88	82.80	61.71	58.77	<b>58.19</b>	58.47	59.14
Arabic (ar)	69.31	75.32	81.70	47.63	47.31	47.23	46.82	<b>46.39</b>
Ukrainian (uk)	77.76	64.09	82.71	<b>45.62</b>	50.84	48.73	49.36	47.41
Average	-	76.68	77.26	48.72	48.64	47.48	47.27	<b>46.99</b>
Weighted Average	-	72.28	77.54	46.72	47.38	46.08	46.12	<b>45.45</b>



在技术进步的道路上  
我们共同努力

Q & A



欢迎进入  
读者交流群