学术分享

迁移学习问题与方法

王晋东

中国科学院计算技术研究所

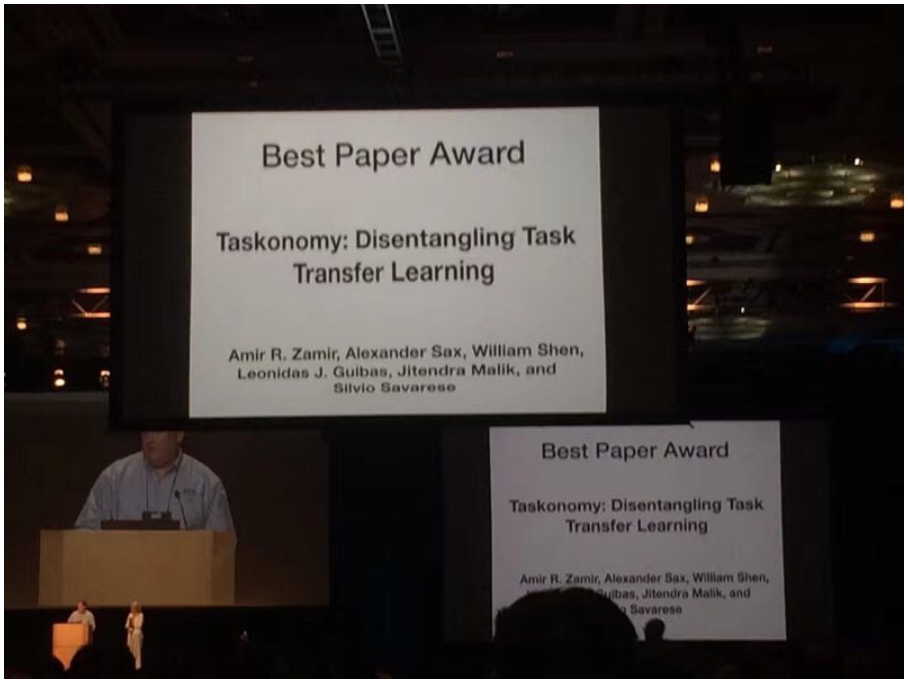
2018年07月

个人简介

- 中国科学院计算技术研究所 2014级直博生
- 主要研究迁移学习及其应用
- 在国际顶级或权威会议ACM MM、ICDM、UbiComp、PerCom等发表若干文章
- IEEE TPAMI, Neurocomputing等期刊审稿人
- 知乎ID：王晋东不在家，乐于在知乎上分享相关知识
- 微博：@秦汉日记
- jindongwang@outlook.com
- 个人主页：http://jd92.wang
- ***不是大牛，仅为分享***

目 录

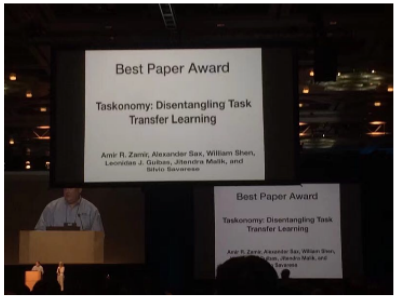
CONTENTS



中国移动 17:45

完成 详情

朋友圈盗图：今年CVPR best paper是迁移学习🙄加油



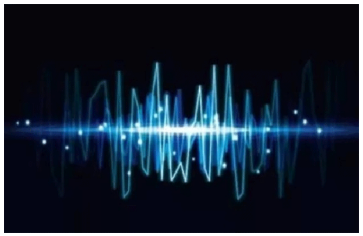
2018年6月19日 23:37 删除

1 迁移学习的背景

- **时代背景**
  - 数据量，以及数据类型不断增加
  - 对机器学习模型的要求：快速构建和强泛化能力
  - 虽然数据量多，但是大部分数据往往**没有标注**
  - 收集标注数据，或者从头开始构建每一个模型，**代价高昂且费时**
  - 缺乏大量标注数据，无法构建具有**强泛化能力**的模型

文本

图片及视频



音频

行为

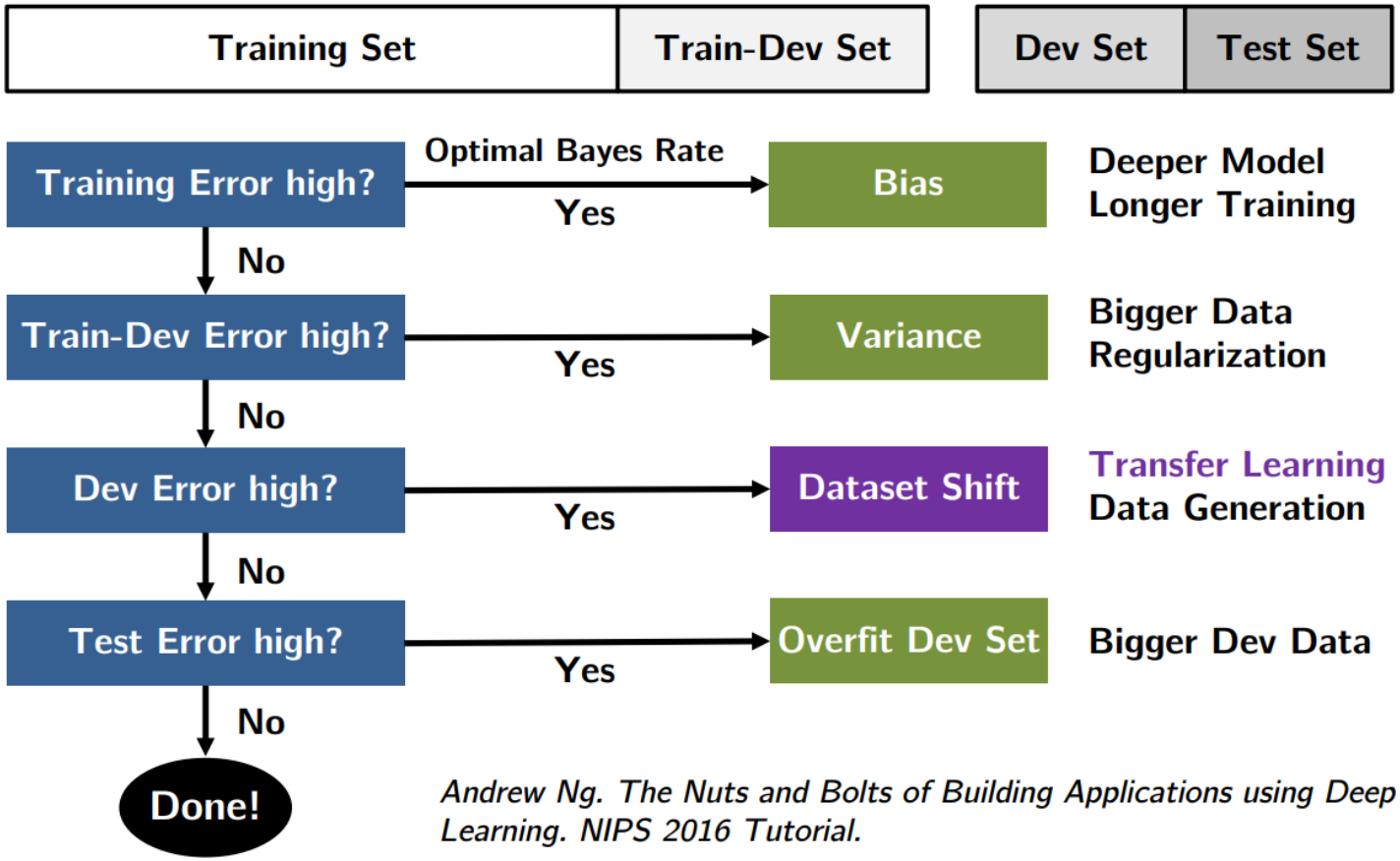
  - 对已有标签的数据和模型进行**重用**成为了可能
    - 传统机器学习方法通常假定这些数据服从**相同分布**，不再适用

如何基于已有的不同分布数据，快速构建强鲁棒性模型，是一个重要问题

1 迁移学习的背景

- **机器学习背景**

$$\Rightarrow f: x \rightarrow y \Rightarrow \epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{complexity}}{n}}$$



Andrew Ng. The Nuts and Bolts of Building Applications using Deep Learning. NIPS 2016 Tutorial.

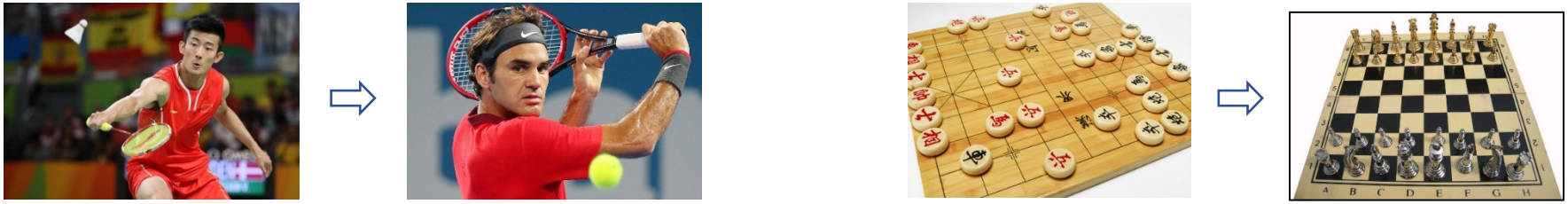
1 迁移学习简介

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

- **迁移学习**
  - 通过减小源域(辅助领域)到目标域的**分布差异**，进行**知识迁移**，从而完成学习**任务**。

- 数据标定
- 跨领域模型
- 强泛化模型
- ……

- **核心思想**
  - 找到不同任务之间的**相关性**
  - **“举一反三”、“照猫画虎”，但不要“东施效颦”（负迁移）**



1 迁移学习基本概念

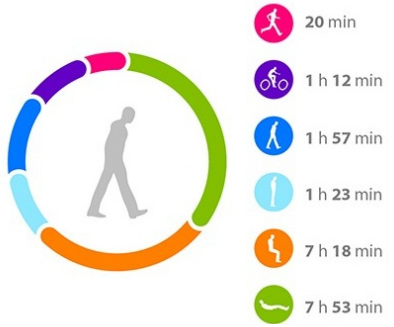
- **迁移学习基本概念**
  - **域(Domain)**：由数据特征和特征分布组成，是学习的主体
    - Source domain (源域)：已有知识的域
    - Target domain (目标域)：要进行学习的域
  - **任务(Task)**：由目标函数和学习结果组成，是学习的结果
- **形式化**
  - 条件：给定一个源域 $\mathcal{D}_S$ 和源域上的学习任务$\mathcal{T}_S$，目标域 $\mathcal{D}_T$ 和目标域上的学习任务$\mathcal{T}_T$
  - 目标：利用 $\mathcal{D}_S$ 和 $\mathcal{T}_S$ 学习在目标域上的预测函数 $f(\cdot)$。
  - 限制条件：$\mathcal{D}_S \neq \mathcal{D}_T$ 或 $\mathcal{T}_S \neq \mathcal{T}_T$
- 人工智能、机器学习中的一个重要问题
  - 每年发表大量相关论文: CVPR、ICCV、ICML、NIPS、IJCAI、AAAI

1 迁移学习应用场景

- **应用前景广阔**
  - 模式识别、计算机视觉、语音识别、自然语言处理、数据挖掘...

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

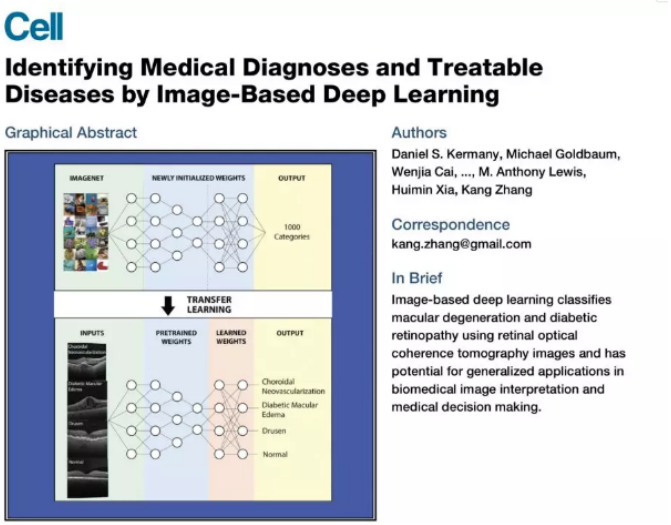
不同领域、不同背景下的文本翻译、舆情分析

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

1 迁移学习应用场景

- **迁移学习+医疗**
  - EEG心跳分类
  - ICU病房中的病人分配调控
  - 前列腺图识别
  - 硬化症检测
  - 胸片X光检测



Cell

Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning

Graphical Abstract

Authors

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, ..., M. Anthony Lewis, Huimin Xia, Kang Zhang

Correspondence

kang.zhang@gmail.com

In Brief

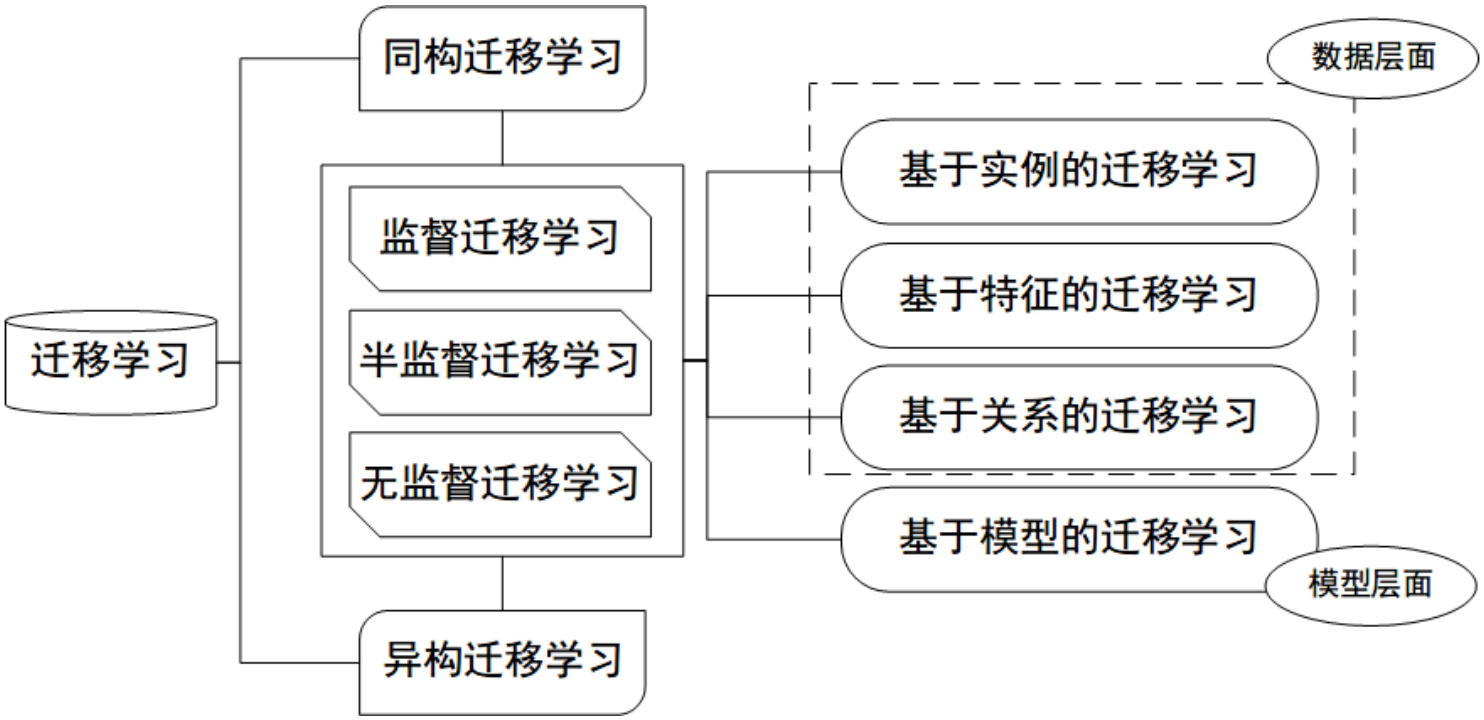
Image-based deep learning classifies macular degeneration and diabetic retinopathy using retinal optical coherence tomography images and has potential for generalized applications in biomedical image interpretation and medical decision making.

- **其他有意义的应用**
  - 犯罪现场图像匹配、工业生产中的错误推断
  - 肢体语言识别、情感分类、银行系统人脸认证

更多应用请见github.com/jindongwang/transferlearning

1 迁移学习简介：迁移学习方法研究领域

- **常见的迁移学习研究领域与方法分类**



目 录

CONTENTS

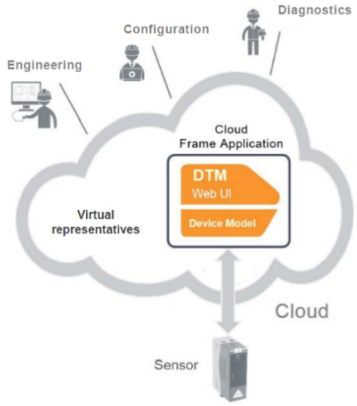
2 迁移学习的必要性

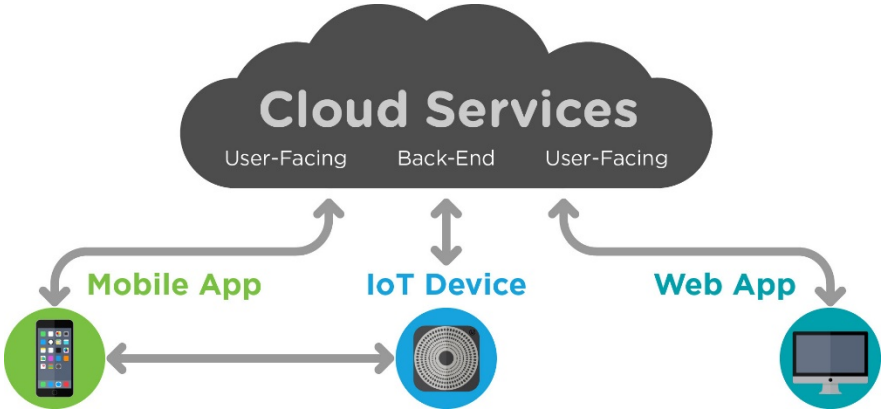
- **数据为王，计算是核心**
  - **大数据与少标注**之间的矛盾
  - **大数据与弱计算**之间的矛盾
- **但是**
  - 大数据、多标注、强计算能力无法被普通研究人员获得

普通研究人员无法获取足够标定数据，并且没有足够的计算资源

2 迁移学习的必要性

- **普适化模型与个性化需求之间的矛盾**
  - 通常需要对设备、环境、用户作具体优化
  - 个性化适配通常很复杂、很耗时
  - 对于不同用户，需要不同的隐私处理方式

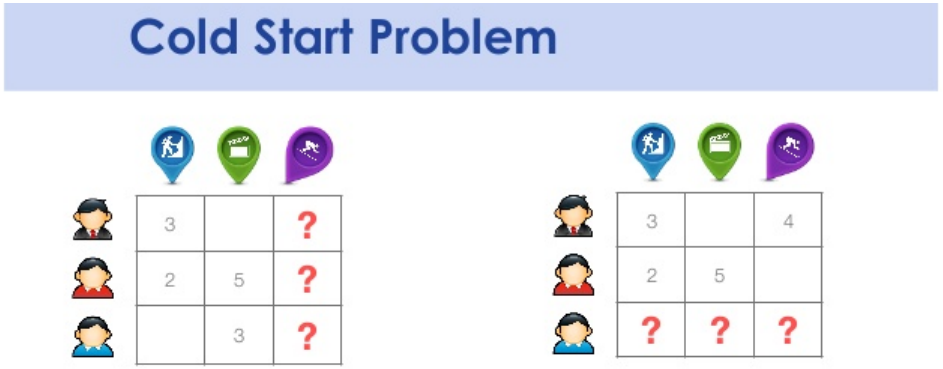




如何针对新用户、新设备、新环境，快速构建模型？

2 迁移学习的必要性

- **特定的机器学习应用**
  - 推荐系统中的冷启动问题：没有数据，如何作推荐？



PANDORA NETFLIX amazon JD.COM 京东

没有足够的用户数据，如何构建模型？

2 迁移学习的必要性

- **为什么需要迁移学习**
  - 迁移学习的必要性

表 1: 迁移学习的必要性

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	只能依赖强大计算能力，但是受众少	模型迁移
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	冷启动问题无法解决	数据迁移

  - 迁移学习与传统机器学习的区别

表 2: 传统机器学习与迁移学习的区别

比较项目	传统机器学习	迁移学习
数据分布	训练和测试数据服从相同的分布	训练和测试数据服从不同的分布
数据标注	需要足够的数据标注来训练模型	不需要足够的数据标注
模型	每个任务分别建模	模型可以在不同任务之间迁移

目 录

CONTENTS

3 迁移学习基本方法

- **常见的迁移学习方法分类**

基于实例的迁移 (instance based TL)

- 通过权重重用源域和目标域的样例进行迁移

基于特征的迁移 (feature based TL)

- 将源域和目标域的特征变换到相同空间

基于模型的迁移 (parameter based TL)

- 利用源域和目标域的参数共享模型

基于关系的迁移 (relation based TL)

- 利用源域中的逻辑网络关系进行迁移

3 迁移学习基本方法

- **基于实例的迁移学习方法**

源领域（图像）






目标领域（图像）


- 假设：源域中的一些数据和目标域会共享很多共同的特征
- 方法：对源域进行instance reweighting，筛选出与目标域数据相似度高的数据，然后进行训练学习

3 迁移学习基本方法

- **基于实例的迁移学习方法**
  - **代表工作：**
    - TrAdaBoost [Dai, ICML-07]
    - Kernel Mean Matching (KMM) [Smola, ICML-08]
    - Density ratio estimation [Sugiyama, NIPS-07]
  - **优点：**
    - 方法较简单，实现容易
  - **缺点：**
    - 权重选择与相似度度量依赖经验
    - 源域和目标域的数据分布往往不同

3 迁移学习基本方法

- **基于特征的迁移学习方法**

源域和目标域特征空间一致

源域（图像）→ 目标域（图像）

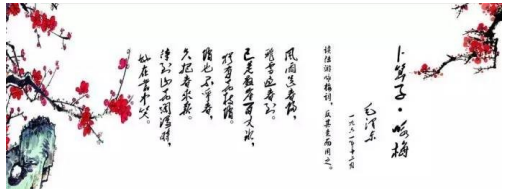
共同特征

源域和目标域特征空间不一致

源域（文本）→ 目标域（图像）

待到山花烂漫时，
她在丛中笑

有文本标记的图片



- 假设：源域和目标域含有一些公共的交叉特征
- 方法：通过特征变换，将两个域的数据变换到同一特征空间，然后进行传统的机器学习

3 迁移学习基本方法

- **基于特征的迁移学习方法**
  - **代表工作：**
    - Transfer component analysis (TCA) [Pan, TKDE-10]
    - Spectral Feature Alignment (SFA) [Pan, WWW-10]
    - Geodesic flow kernel (GFK) [Duan, CVPR-12]
    - Transfer kernel learning (TKL) [Long, TKDE-15]
  - **优点：**
    - 大多数方法采用
    - 特征选择与变换可以取得好效果
  - **缺点：**
    - 往往是一个优化问题，难求解
    - 容易发生过适配

3 迁移学习基本方法

- **基于模型的迁移学习方法**

- 特点：模型相同部分直接进行迁移
- 不需要数据训练

3 迁移学习基本方法

- **基于模型的迁移学习方法**
  - **代表工作：**
    - TransEMDT [Zhao, IJCAI-11]
    - TRCNN [Oquab, CVPR-14]
    - TaskTrAdaBoost [Yao, CVPR-10]
  - **优点：**
    - 模型间存在相似性，可以被利用
  - **缺点：**
    - 模型参数不易收敛

特征+模型=深度迁移方法

3 迁移学习基本方法

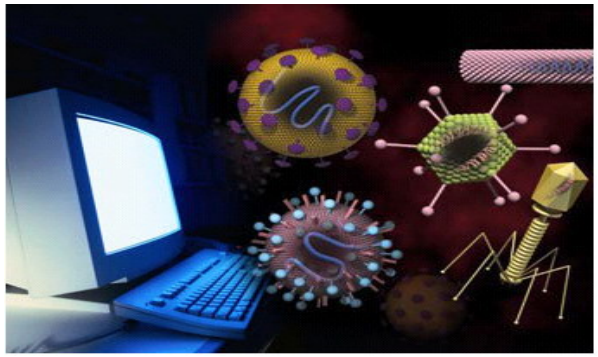
- **基于关系的迁移学习方法**

师生关系

上下级关系

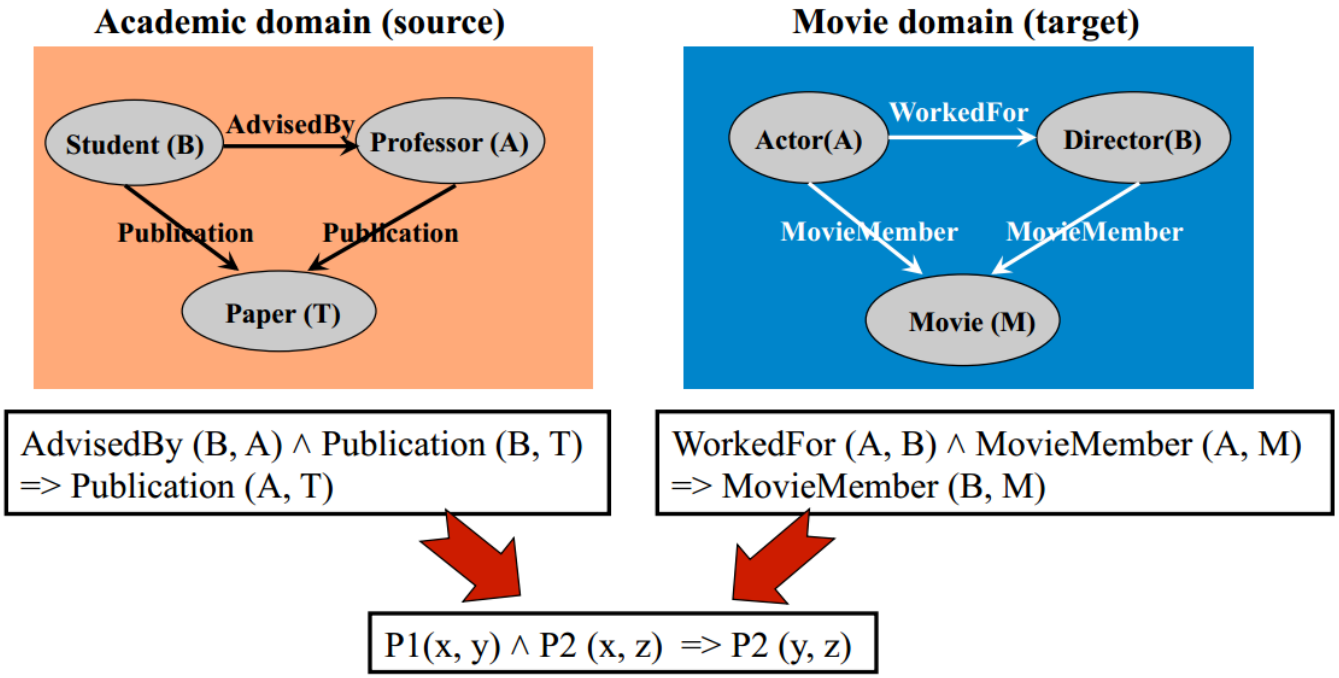
生物病毒

计算机病毒



3 迁移学习基本方法

- **基于关系的迁移学习方法**
  - 假设：如果两个域是相似的，那么它们会共享某种相似关系
  - 方法：利用源域学习逻辑关系网络，再应用于目标域上
  - 代表工作：
    - Predicate mapping and revising [Mihalkova, AAAI-07],
    - Second-order Markov Logic [Davis, ICML-09]

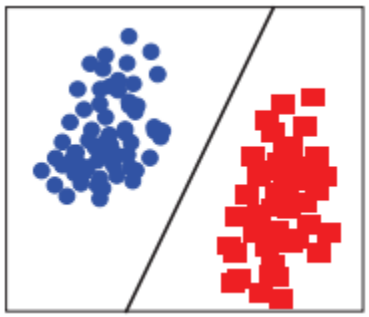


目 录

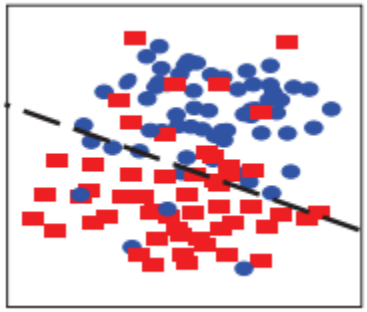
CONTENTS

4 深度迁移学习

- **基本假设**
  - 数据分布与特征变换角度：源域和目标域的**概率分布相似**
    - **最小化**概率分布距离
- **方法**
  - 边缘分布适配 (Marginal distribution adaptation)
    - 假设：$P(\mathbf{X}_s) \neq P(\mathbf{X}_t)$
  - 条件分布适配 (Conditional distribution adaptation)
    - 假设：$P(y_s|\mathbf{X}_s) \neq P(y_t|\mathbf{X}_t)$
  - 联合分布适配 (Joint distribution adaptation)
    - 假设：$P(\mathbf{X}_s, y_s) \neq P(\mathbf{X}_t, y_t)$

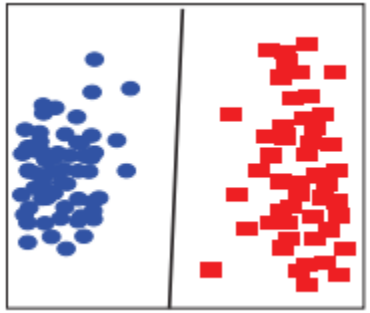


源域数据



目标域数据(1)

优先考虑边缘分布



目标域数据(2)

优先考虑条件分布

4 深度迁移学习

- 深度出现前，传统方法是怎么做的？

4 深度迁移学习

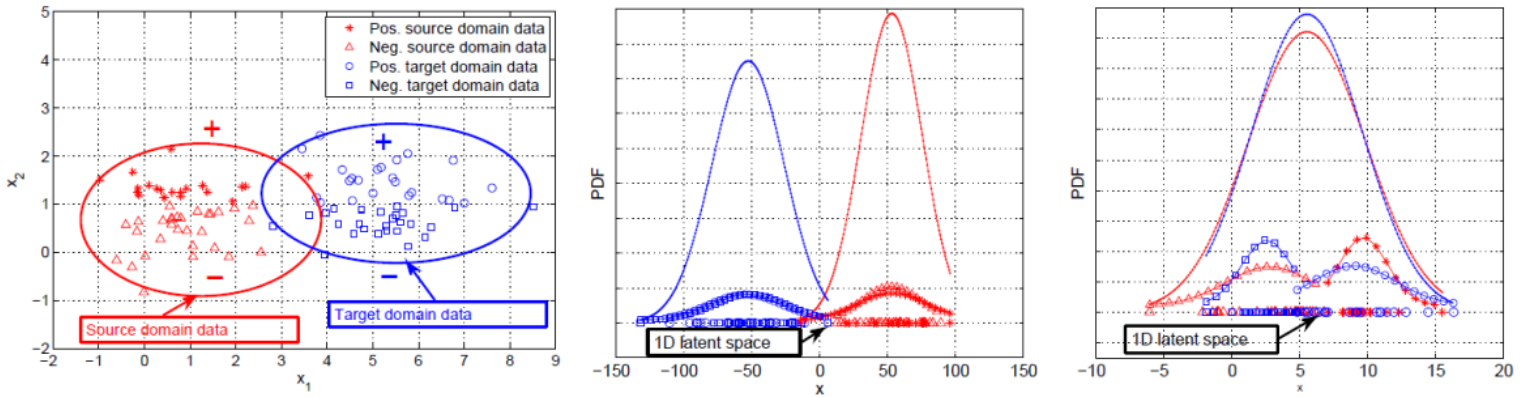
- **边缘分布适配 (1)**
  - 迁移成分分析 (Transfer Component Analysis, TCA) [Pan, TNN-11]
    - 优化目标：

$$\min_{\varphi} \quad \mathrm{Dist}(\varphi(\mathbf{X}_S), \varphi(\mathbf{X}_T)) + \lambda\Omega(\varphi)$$

$$\text{s.t.} \quad \text{constraints on } \varphi(\mathbf{X}_S) \text{ and } \varphi(\mathbf{X}_T)$$

    - 最大均值差异(Maximum Mean Discrepancy, MMD)

$$\mathrm{Dist}(P(X_S), P(X_T)) = \left\| \frac{1}{n_S}\sum_{i=1}^{n_S}\Phi(x_{S_i}) - \frac{1}{n_T}\sum_{j=1}^{n_T}\Phi(x_{T_j}) \right\|_{\mathcal{H}}$$



Original feature space　　PCA　　TCA

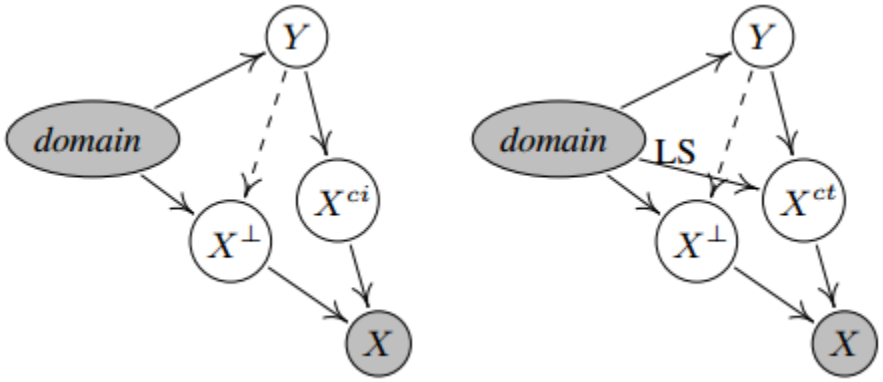
4 深度迁移学习

- **边缘分布适配 (2)**
  - 迁移成分分析 (TCA)方法的一些扩展
    - Adapting Component Analysis (ACA) [Dorri, ICDM-12] $\mathit{maximize} \quad \frac{tr(HK_XHL_\Phi)}{tr(HL_MHL_\Phi)}$
      - 最小化MMD，同时维持迁移过程中目标域的结构
    - Domain Transfer Multiple Kernel Learning (DTMKL) [Duan, PAMI-12]
      - 多核MMD
    - Deep Domain Confusion (DDC) [Tzeng, arXiv-14]
      - 把MMD加入到神经网络中
    - Deep Adaptation Networks (DAN) [Long, ICML-15]
      - 把MKK-MMD加入到神经网络中
    - Distribution-Matching Embedding (DME) [Baktashmotlagh, JMLR-16]
      - 先计算变换矩阵，再进行映射
    - Central Moment Discrepancy (CMD) [Zellinger, ICLR-17]
      - 不只是一阶的MMD，推广到了k阶

$$k = \sum_{m=1}^{M} d_m k_m$$

4 深度迁移学习

- **条件分布适配**
  - Domain Adaptation of Conditional Probability Models via Feature Subsetting [Satpal, PKDD-07]
    - 条件随机场+分布适配
    - 优化目标：$$\operatorname{argmax}_{\mathbf{w},S} \sum_{(\mathbf{x},\mathbf{y})\in D} \sum_{k\in S} w_k f_k(\mathbf{x},\mathbf{y}) - \log z_{\mathbf{w}}(\mathbf{x})$$ $$\text{such that} \quad \operatorname{dist}(\mathcal{D},\mathcal{D}'|S,D,D') \leq \epsilon.$$
  - Conditional Transferrable Components (CTC) [Gong, ICML-15]
    - 定义条件转移成分，对其进行建模



4 深度迁移学习

- **联合分布适配 (1)**
  - 联合分布适配 (Joint Distribution Adaptation, JDA) [Long, ICCV-13]
    - 直接继承于TCA，但是加入了条件分布适配
    - 优化目标：

$$D(\mathcal{D}_s, \mathcal{D}_t) \approx D(P(\mathbf{x}_s), P(\mathbf{x}_t)) + D(P(y_s|\mathbf{x}_s), P(y_t|\mathbf{x}_t))$$

    - 问题：如何获得估计条件分布？
      - 充分统计量：用类条件概率近似条件概率
      - 用一个弱分类器生成目标域的初始软标签
    - 最终优化形式

$$\min_{\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{H}\mathbf{K}^{\mathrm{T}}\mathbf{A}=\mathbf{I}} \sum_{c=0}^{C} \mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{K}\mathbf{M}_c\mathbf{K}^{\mathrm{T}}\mathbf{A}\right) + \lambda \|\mathbf{A}\|_F^2$$

    - **联合分布适配的结果普遍优于比单独适配边缘或条件分布**

4 深度迁移学习

- **联合分布适配 (2)**
  - 联合分布适配(JDA)方法的一些扩展
    - Adaptation Regularization (ARTL) [Long, TKDE-14]
      - 分类器学习+联合分布适配
    - Visual Domain Adaptation (VDA) [Tahmoresnezhad, KIS-17]
      - 加入类内距、类间距
    - Joint Geometrical and Statistical Alignment (JGSA) [Zhang, CVPR-17]
      - 加入类内距、类间距、标签适配
    - [Hsu, TIP-16]：加入结构不变性控制
    - [Hsu, AVSS-15]：目标域选择
    - Joint Adaptation Networks (JAN) [Long, ICML-17]
      - 提出JMMD度量，在深度网络中进行联合分布适配

4 深度迁移学习

- **联合分布适配 (3)**
  - 平衡分布适配 (Balanced Distribution Adaptation, BDA) [Wang, ICDM-17]
    - 仅仅适配条件分布和边缘分布就够了吗？
      - 联合分布适配的问题：两种分布同等重要
      - 真实环境：两种分布**不一定**同等重要
    - 加入**平衡因子**动态衡量两种分布的重要性

$$D(\mathcal{D}_s, \mathcal{D}_t) \approx (1-\mu) D(P(\mathbf{x}_s), P(\mathbf{x}_t)) + \mu D(P(y_s|\mathbf{x}_s), P(y_t|\mathbf{x}_t))$$

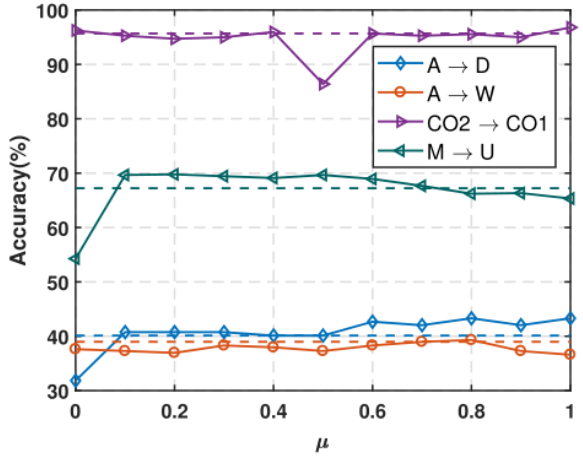
平衡因子

      - 当$\mu \to 0$，表示边缘分布更占优，应该优先适配
      - 当$\mu \to 1$，表示条件分布更占优，应该优先适配
    - 最终表示形式

$$\min \ \operatorname{tr}\left(\mathbf{A}^\top \mathbf{X}\left((1-\mu)\mathbf{M}_0 + \mu \sum_{c=1}^{C} \mathbf{M}_c\right)\mathbf{X}^\top \mathbf{A}\right) + \lambda \|\mathbf{A}\|_F^2$$
$$\text{s.t. } \mathbf{A}^\top \mathbf{X}\mathbf{H}\mathbf{X}^\top \mathbf{A} = \mathbf{I}, \quad 0 \le \mu \le 1$$

4 深度迁移学习

- **联合分布适配 (4)**
  - 平衡分布适配 (BDA)：平衡因子的重要性
    - 对于不同的任务，边缘分布和条件分布并不是同等重要，因此，BDA方法可以**有效衡量**这两个分布的权重，从而达到最好的结果



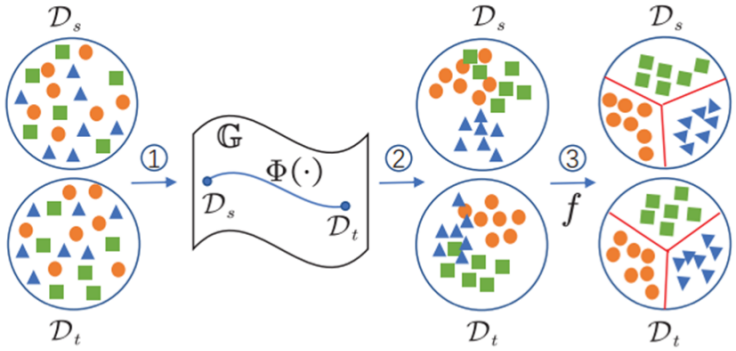
  - 平衡分布适配 (BDA)：平衡因子的求解与估计
    - 目前尚无精确的计算方法; 我们采用A-distance来进行估计
      - 求解源域和目标域整体的A-distance
      - 对目标域聚类，计算源域和目标域每个类的A-distance
      - 计算上述两个距离的比值，则为平衡因子

$$\mathcal{A}(\mathcal{D}_s, \mathcal{D}_t) = 2(1 - 2\epsilon(h))$$

$$\hat{\mu} \approx 1 - \frac{\mathcal{A}_M}{\mathcal{A}_M + \sum_{c=1}^{C} \mathcal{A}_c}$$

4 深度迁移学习

- **基于流形学习的分布适配方法(MEDA)**
  - 以**结构风险最小化**原则作为准则，在**流形空间**中进行分布适配，最终学习一个领域不变的**分类器**
  - 三个步骤：流形变换、分布适配、分类器构建



流形学习

$$\langle \mathbf{z}_i, \mathbf{z}_j \rangle = \int_0^1 (\Phi(t)^T \mathbf{x}_i)^T (\Phi(t)^T \mathbf{x}_j)\, dt = \mathbf{x}_i^T \mathbf{G} \mathbf{x}_j$$

结构风险最小化

$$f = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} (y_i - f(\mathbf{z}_i))^2 + \eta \|f\|_K^2 + \lambda \overline{D_f}(\mathcal{D}_s, \mathcal{D}_t) + \rho R_f(\mathcal{D}_s, \mathcal{D}_t)$$

分布适配　　图拉普拉斯

- **流形学习**：得到数据的非扭曲映射
- **动态分布适配**：动态衡量边缘和条件分布差异，减小数据的分布差异
- **图拉普拉斯**：充分利用流形假设(流形中距离近的点，性质也相似)
- **结构风险最小化**：学习准则，统一学习分类器

4 深度迁移学习

- **基于流形学习的分布适配方法(MEDA)**
  - 采用**表示定理**，将分类器变换为

$$f(\mathbf{z}) = \sum_{i=1}^{n+m} \beta_i K(\mathbf{z}_i, \mathbf{z})$$

  - 问题最终变换成

$$f = \arg\min_{f \in \mathcal{H}_K} ||(\mathbf{Y} - \boldsymbol{\beta}^T \mathbf{K})\mathbf{A}||_F^2 + \eta \operatorname{tr}(\boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta}) + \operatorname{tr}\left(\boldsymbol{\beta}^T \mathbf{K}(\lambda \mathbf{M} + \rho \mathbf{L})\mathbf{K}\boldsymbol{\beta}\right)$$

  - 解决方案为

$$\boldsymbol{\beta}^\star = ((\mathbf{A} + \lambda \mathbf{M} + \rho \mathbf{L})\mathbf{K} + \eta \mathbf{I})^{-1} \mathbf{A} \mathbf{Y}^T$$

相比现有方法，MEDA的优势：

- 通过流形变换减小了数据的**扭曲性**，极大地减小了分布之间的差异
- **直接学习**目标域的分类函数，而现有方法只是学习特征变换，还需依赖于进一步的分类器构建

4 深度迁移学习

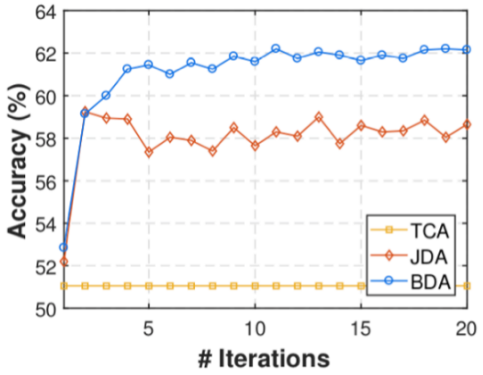
- **精度对比**
  - 我们的MEDA方法比当前最新最好的CVPR-17方法
    - 精度提升**3.5%**
    - 错误率降低**13.88%**
    - 结果标准差降低**50%**以上

Method	A → D	A → W	D → A	D → W	W → A	W → D	Average
SVM	55.7	50.6	46.5	93.1	43.0	97.4	64.4
TCA [10]	45.4	40.5	36.5	78.2	34.1	84.0	53.1
GFK [6]	52.0	48.2	41.8	86.5	38.6	87.5	59.1
SA [3]	46.2	42.5	39.3	78.9	36.3	80.6	54.0
DANN [5]	34.0	34.1	20.1	62.0	21.2	64.4	39.3
CORAL [12]	57.1	53.1	51.1	94.6	47.3	98.2	66.9
AlexNet [7]	63.8	61.6	51.1	95.4	49.8	99.0	70.1
DDC [14]	64.4	61.8	52.1	95.0	52.2	98.5	70.6
DAN [8]	67.0	68.5	54.0	96.0	53.1	99.0	72.9
RTN [9]	71.0	**73.3**	50.5	**96.8**	51.0	99.6	73.7
DCORAL [13]	66.4	66.8	52.8	95.7	51.5	99.2	72.1
DUCDA [15]	68.3	68.3	53.6	96.2	51.6	**99.7**	73.0
MEDA	**69.5**	69.9	**58.0**	94.0	**56.0**	96.8	**74.0**

Jindong Wang, Wenjie Feng, et al. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. ACM MM 2018. ORAL (Top 10 papers).

4 深度迁移学习

- **概率分布适配：总结**
  - 方法
    - 基础：大多数方法基于MMD距离进行优化求解
    - 分别进行边缘 / 条件 / 联合概率适配
    - **效果**：平衡 (BDA) > 联合 (JDA) > 边缘 (TCA) > 条件
  - 使用
    - 数据整体差异性大 (相似度较低)，边缘分布更重要
    - 数据整体差异性小 (协方差漂移)，条件分布更重要
  - 最新成果
    - **深度学习+分布适配**往往有更好的效果 (DDC、DAN、JAN)



BDA、JDA、TCA精度比较

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
AlexNet (Krizhevsky et al., 2012)	61.6±0.5	95.4±0.3	99.0±0.2	63.8±0.5	51.1±0.6	49.8±0.4	70.1
TCA (Pan et al., 2011)	61.0±0.0	93.2±0.0	95.2±0.0	60.8±0.0	51.6±0.0	50.9±0.0	68.8
GFK (Gong et al., 2012)	60.4±0.0	95.6±0.0	95.0±0.0	60.6±0.0	52.4±0.0	48.1±0.0	68.7
DDC (Tzeng et al., 2014)	61.8±0.4	95.0±0.5	98.5±0.4	64.4±0.3	52.1±0.6	52.2±0.4	70.6
DAN (Long et al., 2015)	68.5±0.5	96.0±0.3	99.0±0.3	67.0±0.4	54.0±0.5	53.1±0.5	72.9
RTN (Long et al., 2016)	73.3±0.3	**96.8**±0.2	**99.6**±0.1	71.0±0.2	50.5±0.3	51.0±0.1	73.7
RevGrad (Ganin & Lempitsky, 2015)	73.0±0.5	96.4±0.3	99.2±0.3	72.3±0.3	53.4±0.4	51.2±0.5	74.3
JAN (ours)	74.9±0.3	96.6±0.2	99.5±0.2	71.8±0.2	**58.3**±0.3	55.0±0.4	76.0
JAN-A (ours)	**75.2**±0.4	96.6±0.2	**99.6**±0.1	**72.8**±0.3	57.5±0.2	**56.3**±0.2	**76.3**

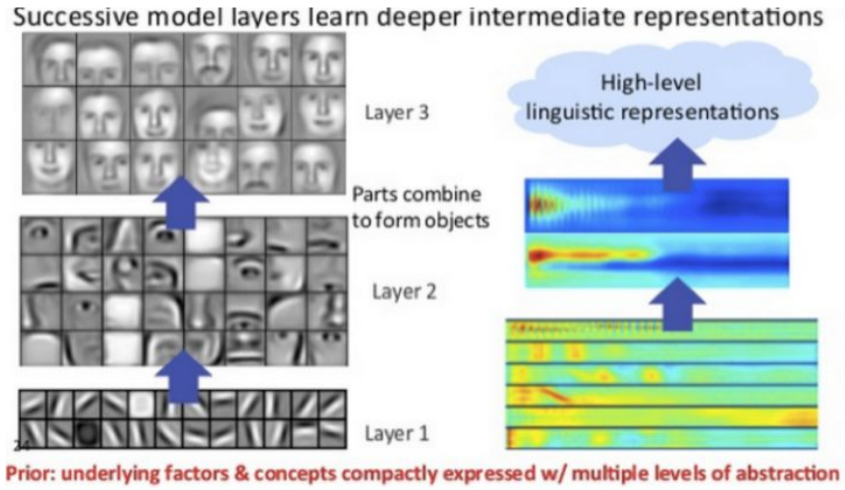
DDC、DAN、JAN与其他方法结果比较

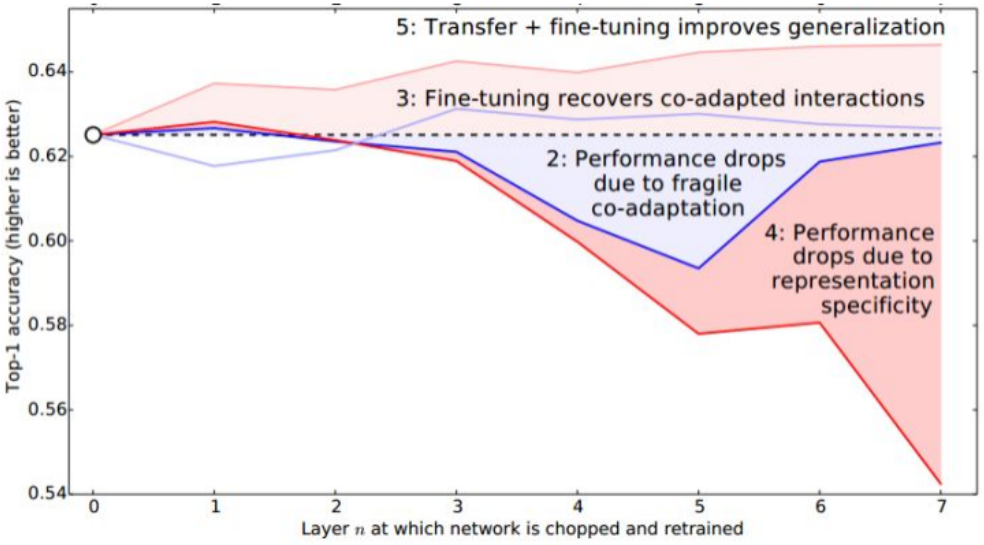
4 深度迁移学习

- **为什么要用深度网络进行迁移？**
  - 深度**端到端** vs 传统依赖特征
  - 深度网络能学习到**更好**的特征表示
  - 深度网络有很多预训练好的模型

4 深度迁移学习

- **深度迁移学习的可行性**
  - How transferrable are features in deep neural networks? [Yosinski, NIPS-14]

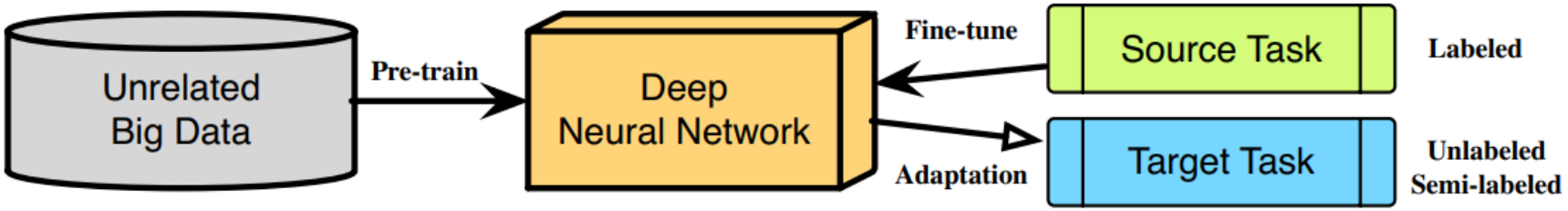




  - 深度网络提取的特征具有**层次性**
  - 网络从浅到深，特征从**通用**到**特定**

4 深度迁移学习

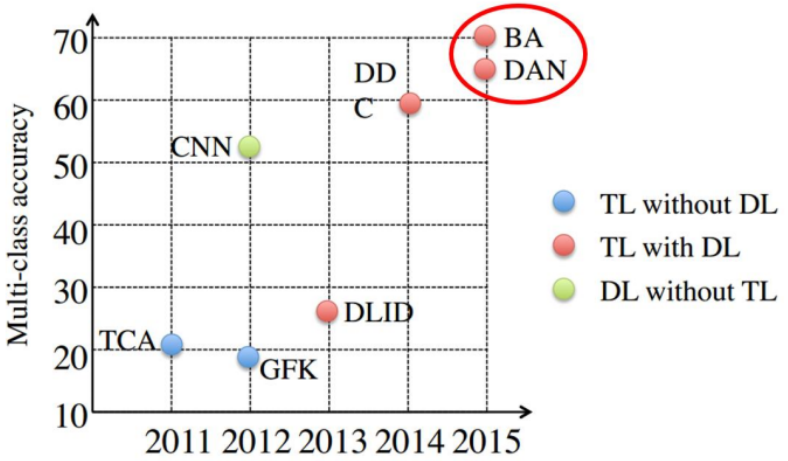
- **深度迁移学习的基本方法**



$$\ell = \ell_c(\mathcal{D}_s, \mathbf{y}_s) + \lambda \ell_A(\mathcal{D}_s, \mathcal{D}_t)$$

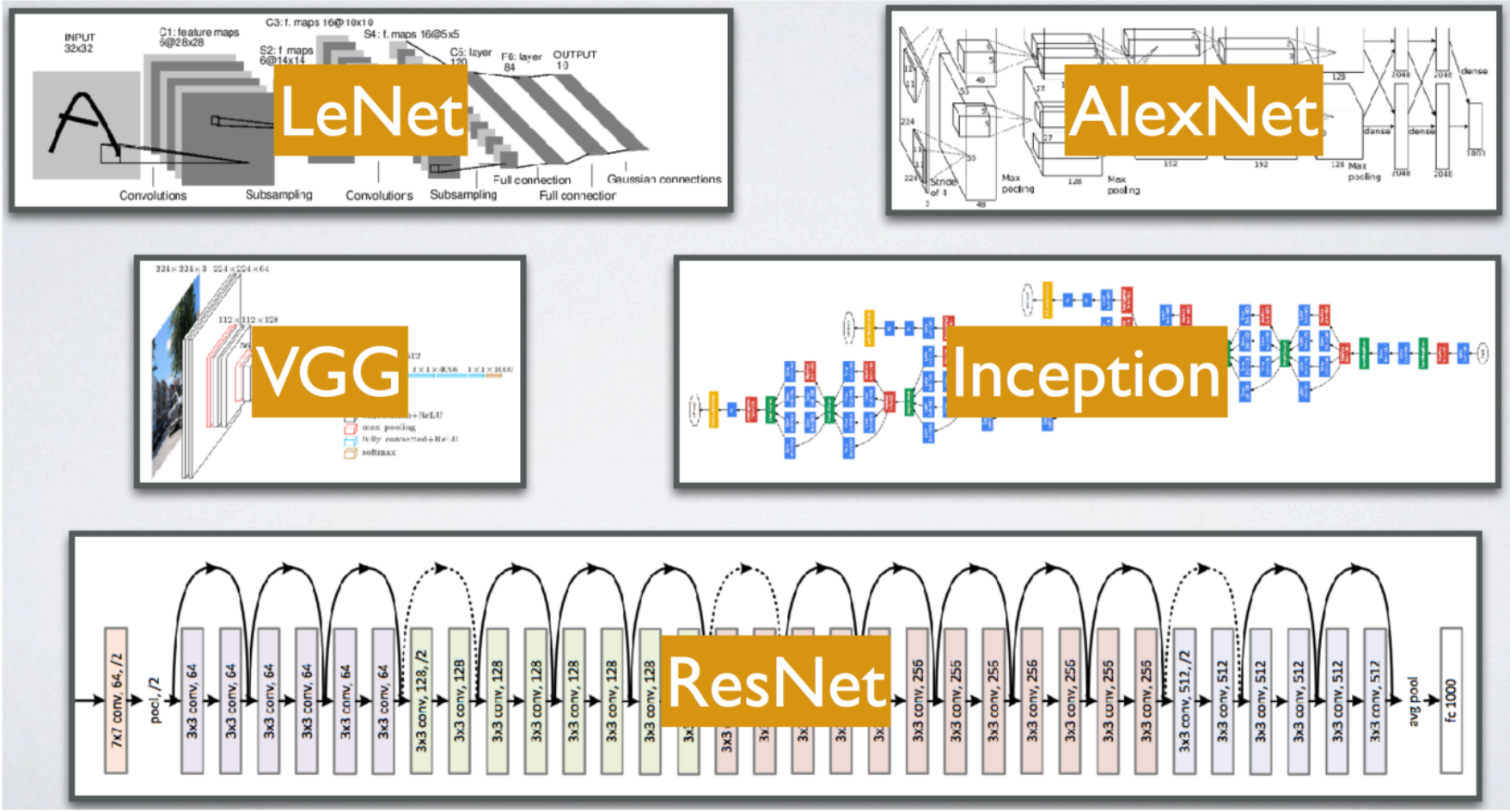
- 最简单的深度迁移：Pre-train + Fine-tune
- 深度网络特征的自适应迁移
- 与对抗网络的结合

通常，深度迁移网络的效果都要好于传统方法以及fine-tune



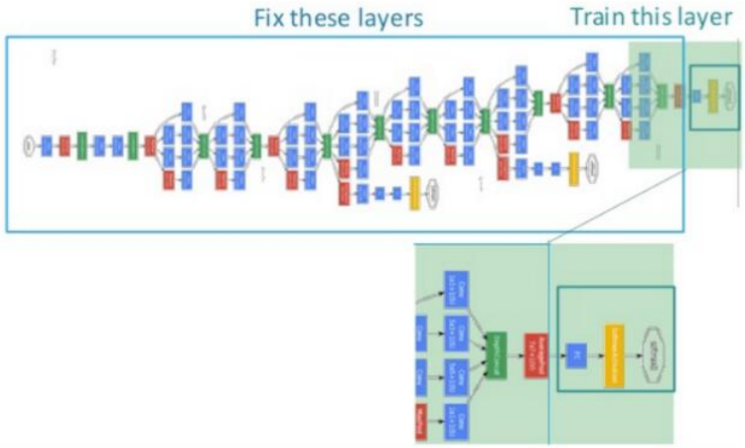
4 深度迁移学习

- **深度迁移学习的基本方法**
  - 常用来做迁移学习的深度网络



4 深度迁移学习

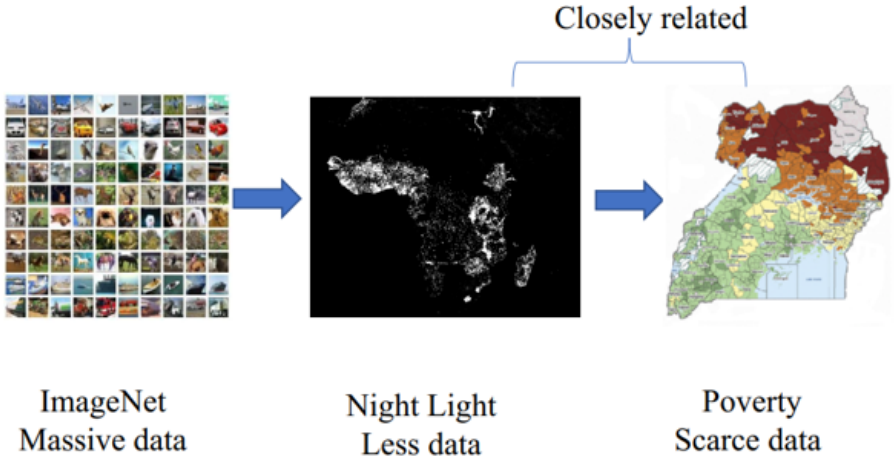
- **最简单的深度迁移学习：fine-tune**



- 不需要针对新任务从头开始训练网络，节省了时间成本
- 预训练好的模型通常都是在大数据集上进行的，无形中扩充了我们的训练数据，使得模型更鲁棒、泛化能力更好
- Finetune 实现简单，使得我们只关注自己的任务即可

4 深度迁移学习

- **Fine-tune的直接应用**
  - 斯坦福大学利用深度网络进行非洲贫困统计[Xie, AAAI-16]



	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**

Xie M, Jean N, Burke M, et al. Transfer learning from deep features for remote sensing and poverty mapping. AAAI 2016.

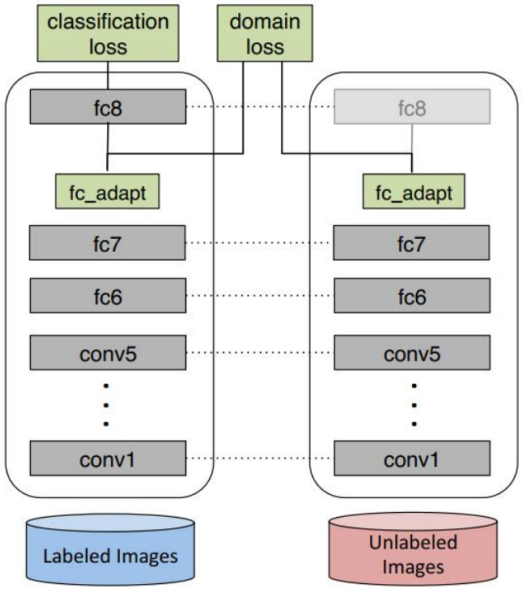
4 深度迁移学习

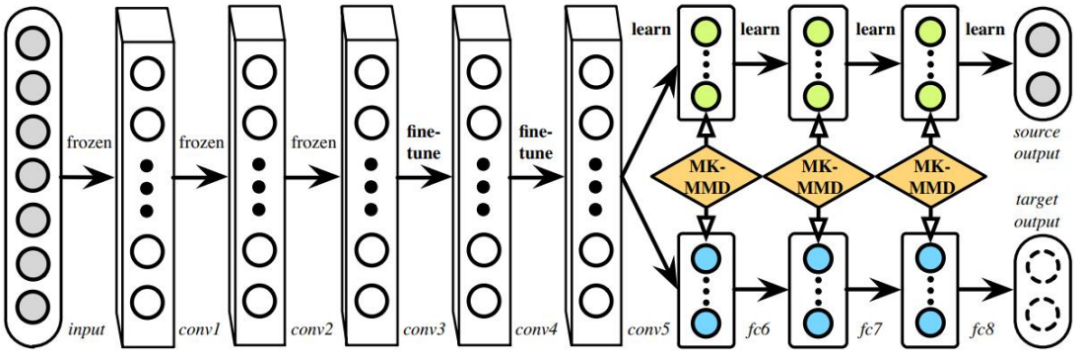
- **深度迁移学习代表方法**
  - DDC: Deep domain confusion [Tzeng, arXiv-14]

$$\ell = \ell_c(\mathcal{D}_s, \mathbf{y}_s) + \lambda MMD^2(\mathcal{D}_s, \mathcal{D}_t)$$

  - DAN: Deep adaptation network [Long, ICML-15]

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





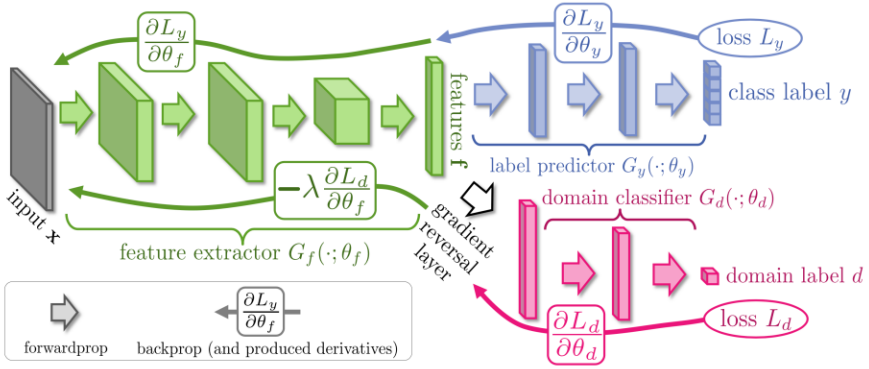
4 深度迁移学习

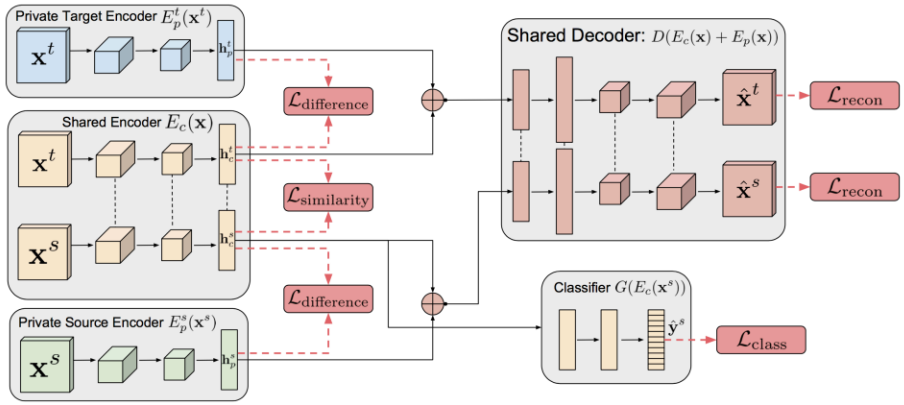
- **深度迁移学习代表方法**
  - DANN: Domain-adversarial Neural Network [Ganin, ICML-15]

$$E(\theta_f,\theta_y,\theta_d)=\sum_{\mathbf{x}_i\in\mathcal{D}_s}L_y\left(G_y\left(G_f(\mathbf{x}_i)\right),y_i\right)-\lambda\sum_{\mathbf{x}_i\in\mathcal{D}_s\cup\mathcal{D}_t}L_d\left(G_d\left(G_f(\mathbf{x}_i)\right),d_i\right)$$

  - DSN: Domain separation networks [Bousmalis, NIPS-16]

$$\ell=\ell_{task}+\alpha\ell_{recon}+\beta\ell_{difference}+\gamma\ell_{similarity}$$

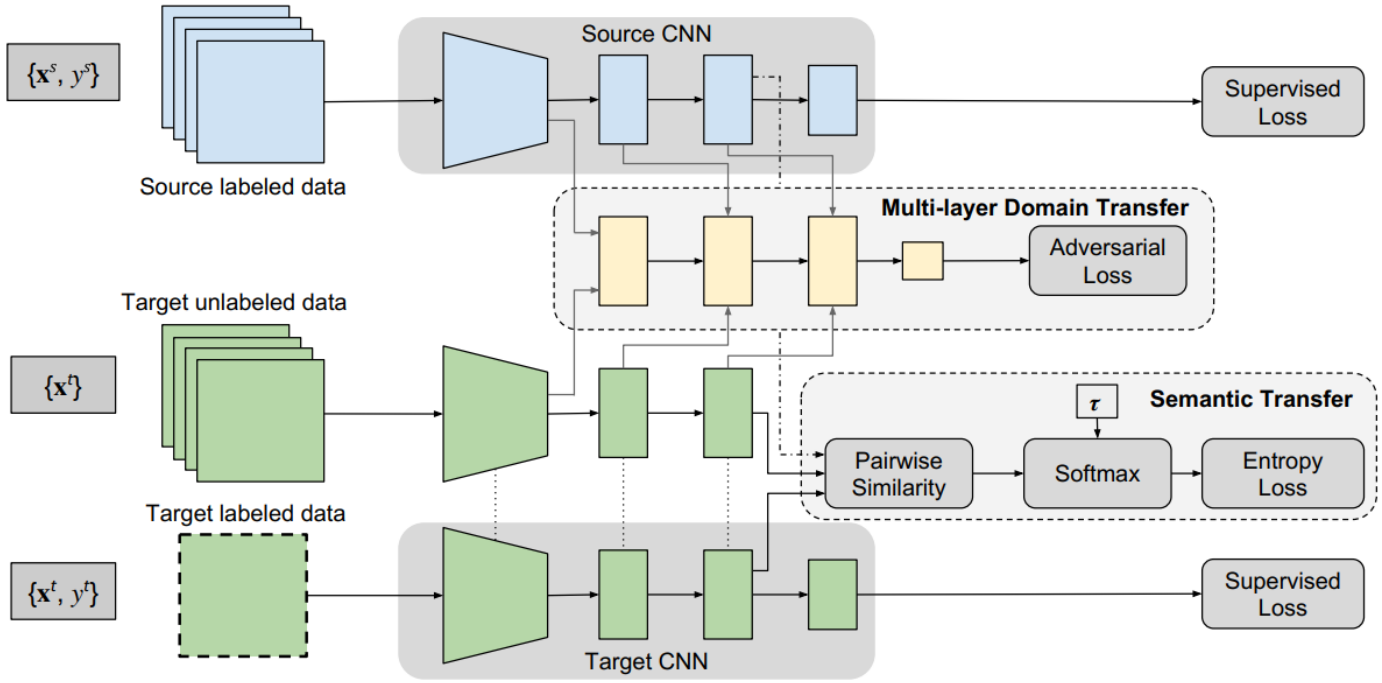




4 深度迁移学习

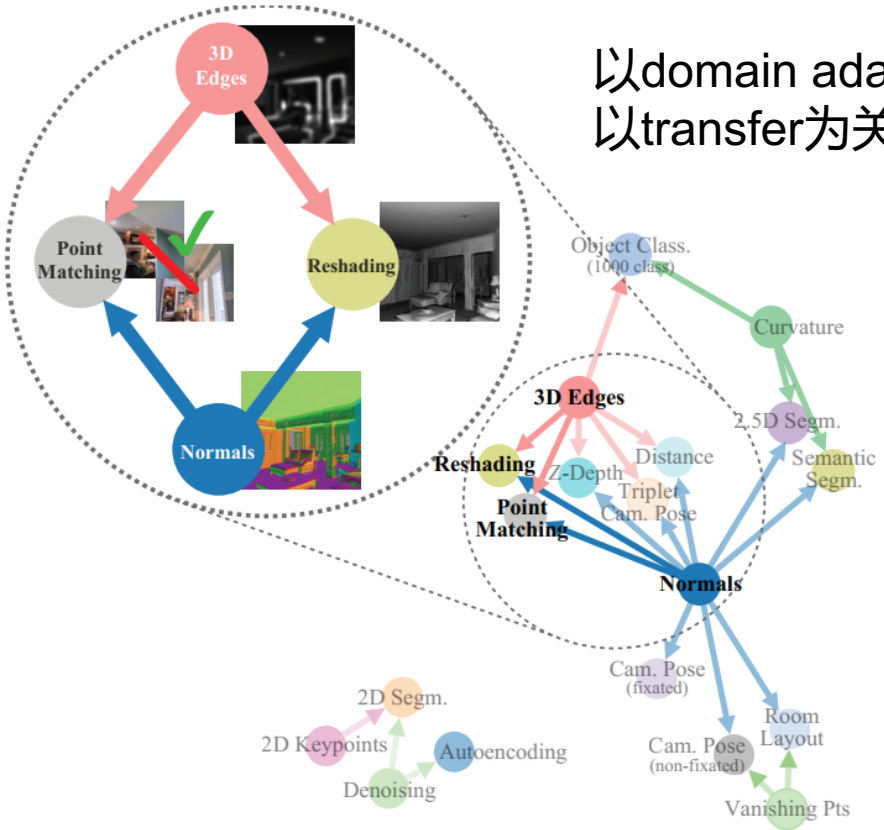
- **深度迁移学习代表方法**
  - 特征和任务同时迁移 [Luo, NIPS-17]

$$\mathcal{L}(\mathcal{X}^{\mathcal{S}}, \mathcal{Y}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}}, \tilde{\mathcal{X}}^{\mathcal{T}}) = \mathcal{L}_{\text{sup}}(\mathcal{X}^{\mathcal{T}}, \mathcal{Y}^{\mathcal{T}}) + \alpha \mathcal{L}_{DT}(\mathcal{X}^{\mathcal{S}}, \tilde{\mathcal{X}}^{\mathcal{T}}) + \beta \mathcal{L}_{ST}(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}}, \tilde{\mathcal{X}}^{\mathcal{T}})$$



4 深度迁移学习

- **深度迁移学习 @ CVPR 2018**
  - Taskonomy [Zamir, CVPR-18] (**Best paper)**
    - 探秘不同任务之间的可迁移性
  - Maximum Classifier Discrepancy for Unsupervised Domain Adaptation



以domain adaptation为关键字，CVPR 18 有16篇
以transfer为关键字，CVPR 18 有30篇

4 深度迁移学习

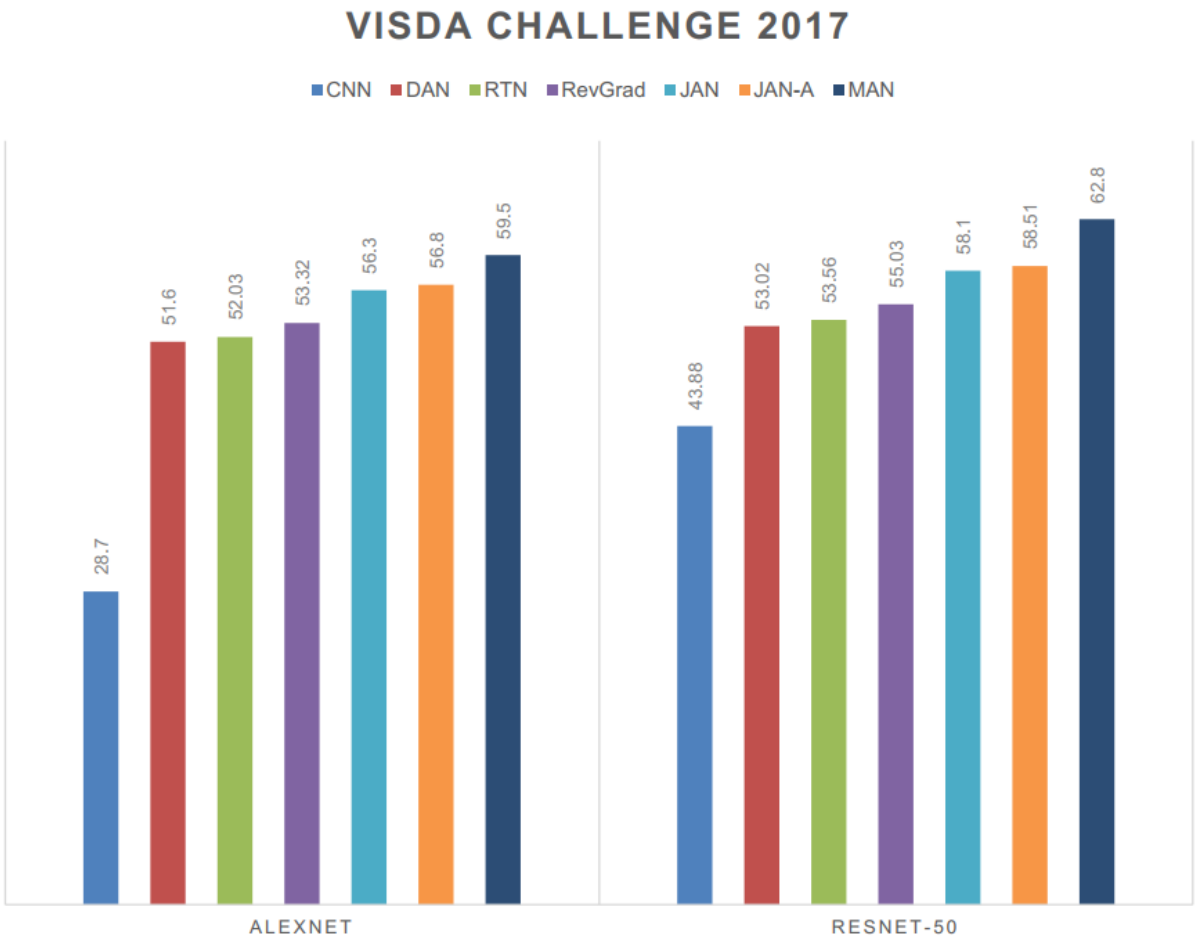
- **深度迁移学习 @ ICML 2018**
  - CyCADA: Cycle Consistent Adversarial Domain Adaptation
    - 在CycleGAN中进行feature和classifier的adaptation
  - L2T: Learning to transfer
    - 利用迁移学习中的经验进行迁移
  - MSTN: Semantic Alignment
    - 在深度网络中进行各个类之间的对齐
  - Knowledge Transfer with Jacobian Matching
    - 用kernel方法进行transfer

更多工作请见ICML官网

以transfer为关键字，ICML 18有12篇

4 深度迁移学习

- **效果比较**



目 录

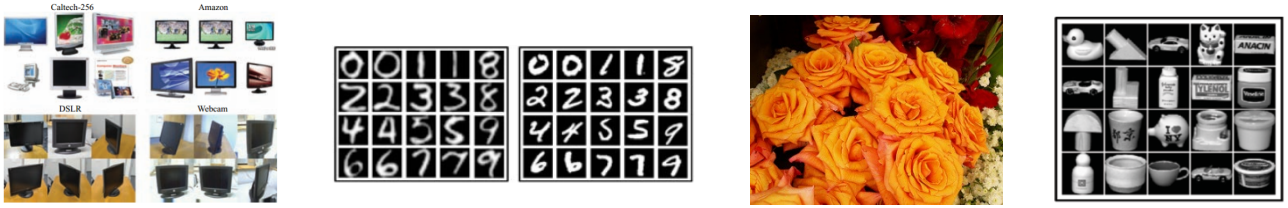
CONTENTS

5 总结、展望与参考资料

- **总结**
  - 迁移学习基本概念与应用
  - 为什么需要迁移学习？
  - 基本方法
  - 深度迁移方法
- **展望**
  - 机器智能与人类经验结合迁移
  - 终身迁移学习
  - 在线迁移学习
  - 强化迁移学习
  - 迁移学习的可解释性

5 总结、展望与参考资料

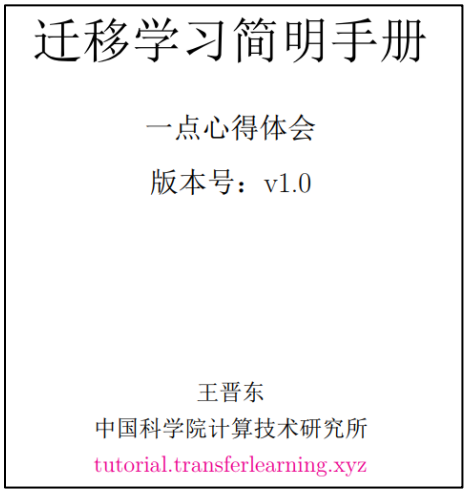
- **参考资料**
  - 迁移学习综述文章
    - A survey on Transfer Learning [Pan and Yang, TKDE-10]
  - (可能是有史以来)最全的**迁移学习资料库**，(文章/资料/代码/数据)
    - http://transferlearning.xyz
  - 迁移学习视频教程
    - https://www.youtube.com/watch?v=qD6iD4TFsdQ
  - 知乎专栏“机器有颗玻璃心”中《小王爱迁移》系列
    - https://zhuanlan.zhihu.com/p/27336930
    - 用浅显易懂的语言深入讲解经典+最新的迁移学习文章
  - 迁移学习与领域自适应论文分享与笔记
    - Paperweekly：http://www.paperweekly.site/collections/231/papers
  - 迁移学习与领域自适应公开**数据集**
    - https://github.com/jindongwang/transferlearning/blob/master/doc/dataset.md

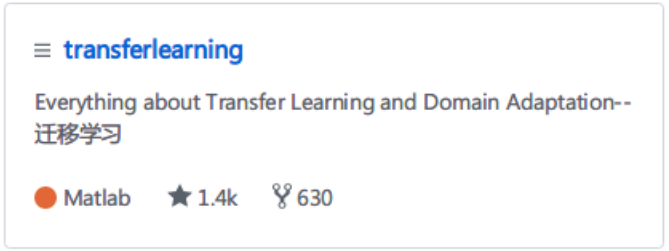


图：Office+Caltech、USPS+MNIST、ImageNet+VOC、COIL20数据集

One More Thing...

- **迁移学习简明手册**





推荐语

看了王晋东同学的“迁移学习小册子”，点三个赞！迁移学习被认为是机器学习的下一个爆点，但介绍迁移学习的文章却很有限。这个册子深入浅出，既回顾了迁移学习的发展历史，又囊括了迁移学习的最新进展。语言流畅，简明通透。应该对机器学习的入门和提高都有很大帮助！

——杨强 (迁移学习权威学者，香港科技大学教授，IJCAI president, AAAI/ACM fellow)

- http://tutorial.transferlearning.xyz

参考文献(1)

- [Pan, TNN-11] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2011, 22(2): 199-210.
- [Dorri, ICDM-12] Dorri F, Ghodsi A. Adapting component analysis[C]//Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012: 846-851.
- [Duan, PAMI-12] Duan L, Tsang I W, Xu D. Domain transfer multiple kernel learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 465-479.
- [Long, ICML-15] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//International Conference on Machine Learning. 2015: 97-105.
- [Baktashmotlagh, JMLR-16] Baktashmotlagh M, Harandi M, Salzmann M. Distribution-matching embedding for visual domain adaptation[J]. The Journal of Machine Learning Research, 2016, 17(1): 3760-3789.
- [Zellinger, ICLR-17] Zellinger W, Grubinger T, Lughofer E, et al. Central moment discrepancy (CMD) for domain-invariant representation learning[J]. arXiv preprint arXiv:1702.08811, 2017.
- [Satpal, PKDD-07] Satpal S, Sarawagi S. Domain adaptation of conditional probability models via feature subsetting[C]//PKDD. 2007, 4702: 224-235.
- [Gong, ICML-15] Gong M, Zhang K, Liu T, et al. Domain adaptation with conditional transferable components[C]//International Conference on Machine Learning. 2016: 2839-2848.
- [Long, ICCV-13] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in ICCV, 2013, pp. 2200–2207.
- [Long, TKDE-14] Long M, Wang J, Ding G, et al. Adaptation regularization: A general framework for transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(5): 1076-1089.
- [Tahmoresnezhad , KIS-17] J. Tahmoresnezhad and S. Hashemi, “Visual domain adaptation via transfer feature learning,” Knowl. Inf. Syst., 2016.
- [Zhang, CVPR-17] Zhang J, Li W, Ogunbona P. Joint Geometrical and Statistical Alignment for Visual Domain Adaptation, CVPR 2017.
- [Hsu, AVSS-15] T. Ming Harry Hsu, W. Yu Chen, C.-A. Hou, and H. T. et al., “Unsupervised domain adaptation with imbalanced cross-domain data,” in ICCV, 2015, pp. 4121–4129.
- [Hsu, TIP-16] P.-H. Hsiao, F.-J. Chang, and Y.-Y. Lin, “Learning discriminatively reconstructed source data for object recognition with few examples,” TIP, vol. 25, no. 8, pp. 3518–3532, 2016.
- [Long, ICML-17] Long M, Wang J, Jordan M I. Deep transfer learning with joint adaptation networks. ICML 2017.
- [Wang, ICDM-17] Wang J, Chen Y, Hao S, Feng W, Shen Z. Balanced Distribution Adaptation for Transfer Learning. ICDM 2017. pp.1129-1134.
- [Blitzer, ECML-06] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning[C]//Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006: 120-128.

参考文献(2)

- [Gu, IJCAI-11] Gu Q, Li Z, Han J. Joint feature selection and subspace learning[C]//IJCAI Proceedings-International Joint Conference on Artificial Intelligence. 2011, 22(1): 1294.
- [Long, CVPR-14] Long M, Wang J, Ding G, et al. Transfer joint matching for unsupervised domain adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1410-1417.
- [Li, IJCAI-16] Li J, Zhao J, Lu K. Joint Feature Selection and Structure Preservation for Domain Adaptation[C]//IJCAI. 2016: 1697-1703.
- [Fernando, ICCV-13] Fernando B, Habrard A, Sebban M, et al. Unsupervised visual domain adaptation using subspace alignment[C]//Proceedings of the IEEE international conference on computer vision. 2013: 2960-2967.
- [Sun, BMVC-15] Sun B, Saenko K. Subspace Distribution Alignment for Unsupervised Domain Adaptation[C]//BMVC. 2015: 24.1-24.10.
- [Sun, AAAI-16] Sun B, Feng J, Saenko K. Return of Frustratingly Easy Domain Adaptation[C]//AAAI. 2016, 6(7): 8.
- [Sun, ECCV-16] Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation[C]//Computer Vision–ECCV 2016 Workshops. Springer International Publishing, 2016: 443-450.
- [Gopalan, ICCV-11] Gopalan R, Li R, Chellappa R. Domain adaptation for object recognition: An unsupervised approach[C]//Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011: 999-1006.
- [Gong, CVPR-12] Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2066-2073.
- [Baktashmotlagh, CVPR-13] Baktashmotlagh M, Harandi M T, Lovell B C, et al. Unsupervised domain adaptation by domain invariant projection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 769-776.
- [Baktashmotlagh, CVPR-14] Baktashmotlagh M, Harandi M T, Lovell B C, et al. Domain adaptation on the statistical manifold[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2481-2488.
- [Ganin, JMLR-16] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine Learning Research, 2016, 17(59): 1-35.
- [Busto, ICCV-17] Panareda Busto P, Gall J. Open Set Domain Adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 754-763.
- [Lu, ICCV-17] Lu H, Zhang L, Cao Z, et al. When unsupervised domain adaptation meets tensor representations. ICCV 2017.
- [Tzeng, arXiv-17] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation[J]. arXiv preprint arXiv:1702.05464, 2017.
- [Wei, arXiv-17] Wei Y, Zhang Y, Yang Q. Learning to Transfer. arXiv 1708.05629, 2017.
- [Xie, AAAI-16] Xie M, Jean N, Burke M, et al. Transfer learning from deep features for remote sensing and poverty mapping. AAAI 2016.

谢谢！
请批评指正