AI007 《人工智能》Week09

迁移学习

Transfer Learning

许岩岩

上海交通大学 · 人工智能研究院

2021年11月11日

Major Ref: 迁移学习简明手册 By 王晋东

提纲

1 迁移学习概述

2 迁移学习基本方法

3 深度迁移学习

4 迁移学习应用案例

迁移学习（Transfer Learning）将已经学习过的知识迁移应用于新的问题中。其核心问题是，找到新问题和原问题之间的相似性，才可以顺利地实现知识的迁移。

迁移学习，是指利用数据、任务、或模型之间的相似性，将在旧领域学习过的模型，应用于新领域的一种学习过程。

Pan, Sinno Jialin, and Qiang Yang. “A survey on transfer learning.” *IEEE Transactions on knowledge and data engineering* 22.10 (2009): 1345-1359. Cited by 14752

为什么需要迁移学习？

1. 大数据与少标注之间的矛盾。

文本

图片及视频

音频

行为

大数据时代，每天每时，社交网络、智能交通、视频监控、行业物流等，都产生着海量的图像、文本、语音等各类数据。数据的增多，使得机器学习和深度学习模型可以依赖于如此海量的数据，持续不段地训练和更新相应的模型，使得模型的性能越来越好，越来越适合特定场景的应用。然而，这些大数据带来了严重的问题：总是缺乏完善的数据标注。

为什么需要迁移学习？

2. 大数据与弱计算之间的矛盾。

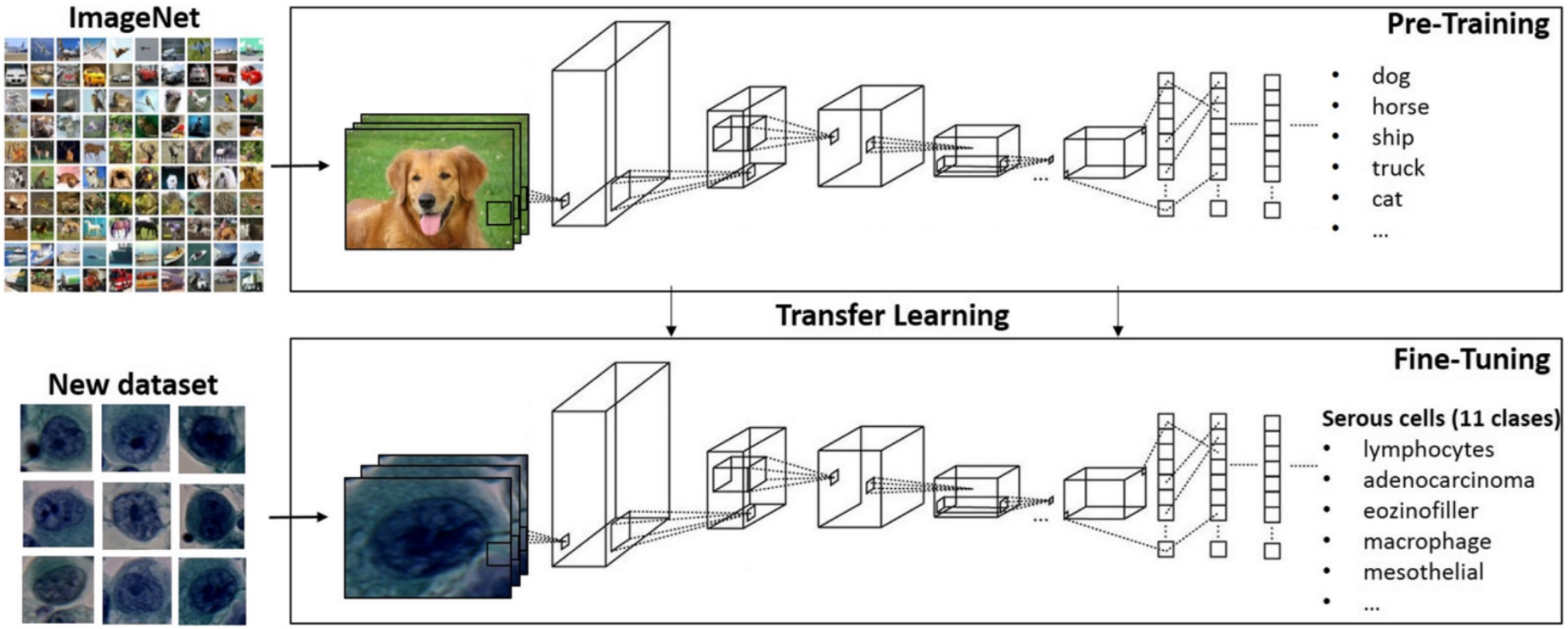
大数据，就需要大设备、强计算能力的设备来进行存储和计算。然而，只有如 Google， Facebook，Microsoft，这些巨无霸公司有着雄厚的计算能力去利用这些数据训练模型。

大数据与强计算能力

为什么需要迁移学习？

3. 普适化模型与个性化需求之间的矛盾。

机器学习的目标是构建一个尽可能通用的模型，但是具体到每个个体、每个需求，都存在其唯一性和特异性，一个普适化的通用模型根本无法满足。那么，能否将这个通用的模型加以改造和适配，使其更好地服务于人们的个性化需求？



为什么需要迁移学习？

4. 特定应用的需求。

一些特定的应用，比如推荐系统的冷启动问题。一个新的推荐系统，没有足够的用户数据，如何进行精准的推荐？一个崭新的图片标注系统，没有足够的标签，如何进行精准的服务？

PANDORA NETFLIX amazon JD.京东.COM

特定需求：冷启动

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

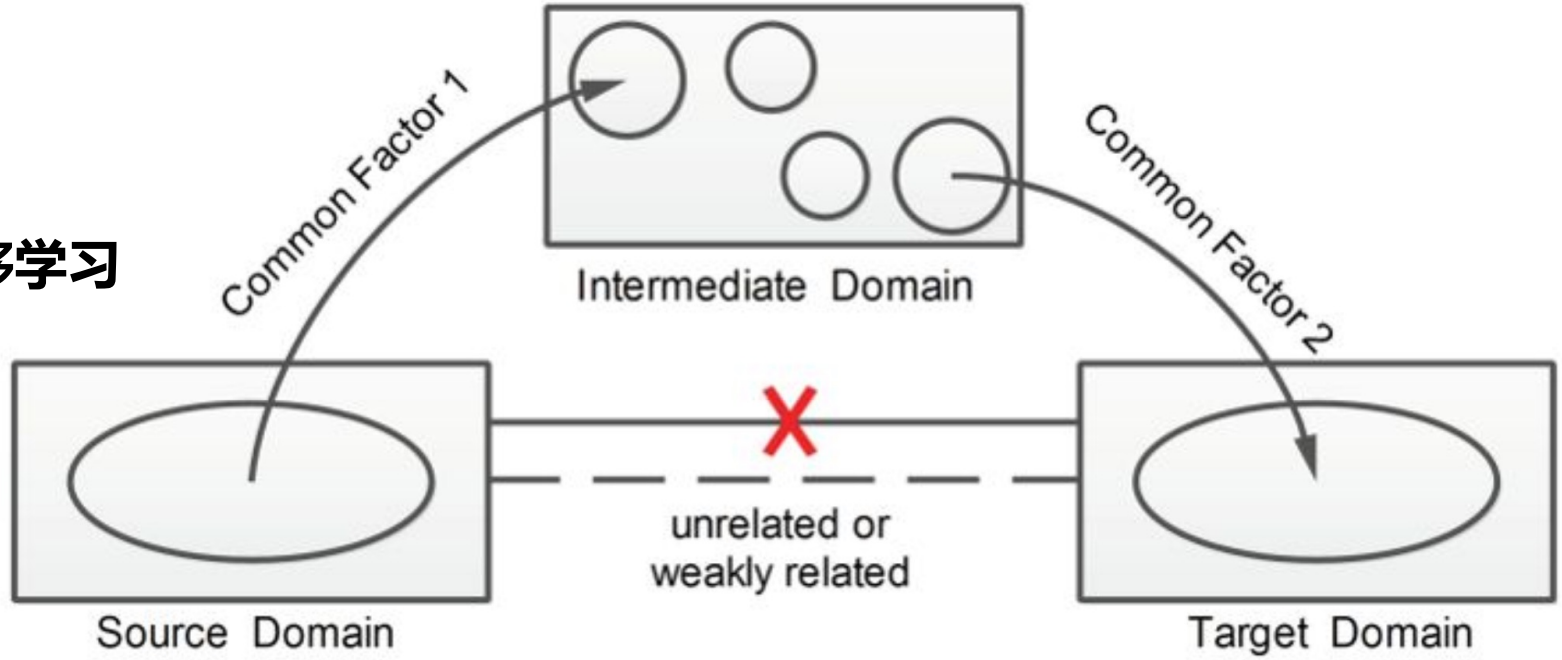
表 2: 传统机器学习与迁移学习的区别

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

负迁移（Negative Transfer）：在源域上学习到的知识，对于目标域上的学习产生负面作用。

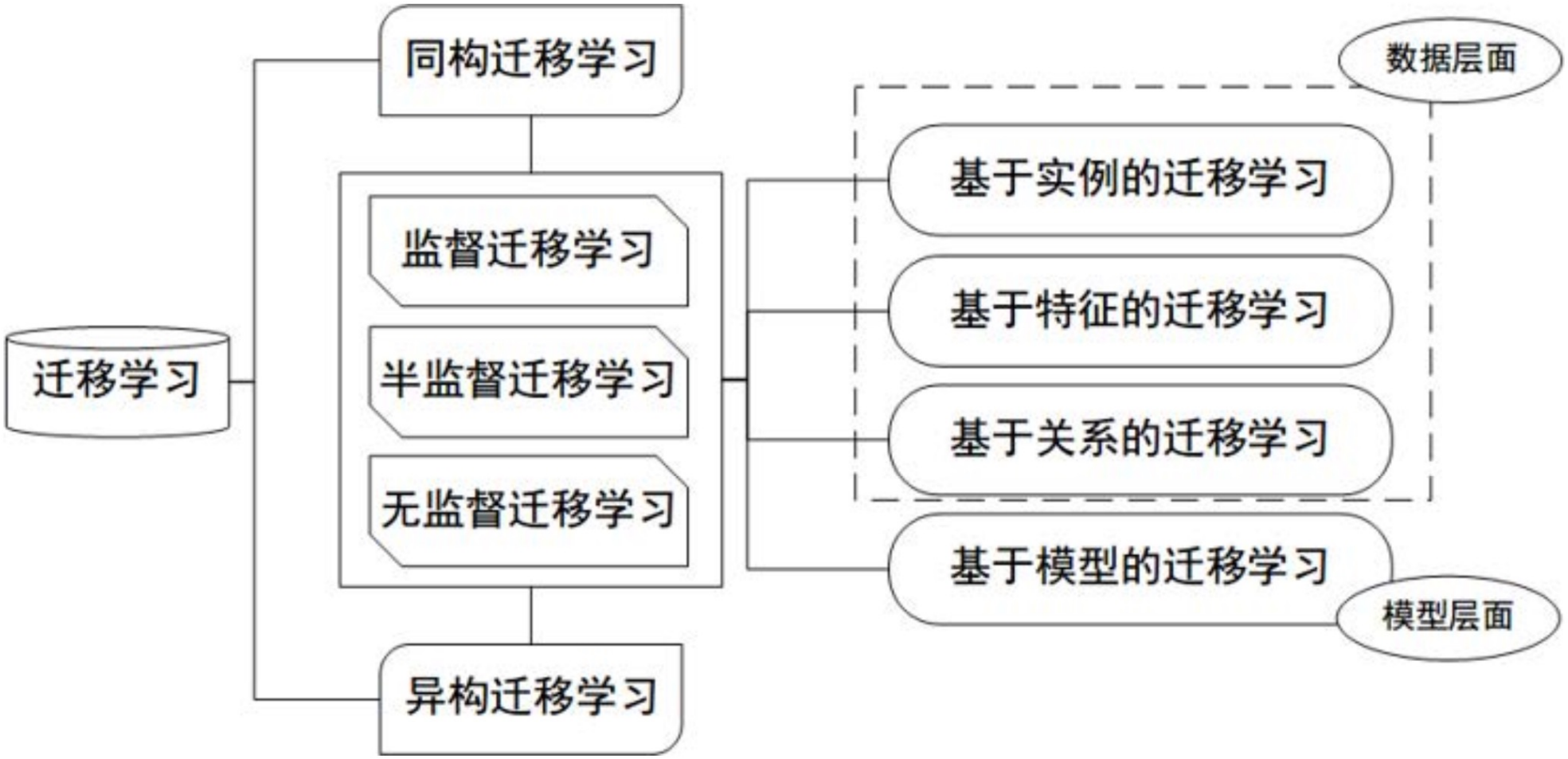
- 数据问题：源域和目标域压根不相似，谈何迁移？
- 方法问题：源域和目标域是相似的，但是，迁移学习方法不够好，没找到可迁移的成分。

传递式迁移学习



Ref: Tan, Ben, et al. "Transitive transfer learning." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015.

迁移学习研究领域



迁移学习的研究领域与研究方法分类

按目标域标签分

这种分类方式最为直观。类比机器学习，按照目标领域有无标签，迁移学习可以分为以下三个大类：

1. 监督迁移学习 (Supervised Transfer Learning)
2. 半监督迁移学习 (Semi-Supervised Transfer Learning)
3. 无监督迁移学习 (Unsupervised Transfer Learning)

按学习方法分类

按学习方法的分类形式，最早在迁移学习领域的权威综述文章 [Pan and Yang, 2010] 给出定义。它将迁移学习方法分为以下四个大类：

1. 基于样本的迁移学习方法 (Instance based Transfer Learning)
2. 基于特征的迁移学习方法 (Feature based Transfer Learning)
3. 基于模型的迁移学习方法 (Model based Transfer Learning)
4. 基于关系的迁移学习方法 (Relation based Transfer Learning)

按特征分类

按照特征的属性进行分类，也是一种常用的分类方法。这在最近的迁移学习综述 [Weiss et al., 2016] 中给出。按照特征属性，迁移学习可以分为两个大类：

1. 同构迁移学习 (Homogeneous Transfer Learning)

2. 异构迁移学习 (Heterogeneous Transfer Learning)

这也是一种很直观的方式：如果特征语义和维度都相同，那么就是同构；反之，如果特征完全不相同，那么就是异构。举个例子来说，不同图片的迁移，就可以认为是同构；而图片到文本的迁移，则是异构的。

按离线与在线形式分类

按照离线学习与在线学习的方式，迁移学习还可以被分为：

1. 离线迁移学习 (Offline Transfer Learning)

2. 在线迁移学习 (Online Transfer Learning)

目前，绝大多数的迁移学习方法，都采用了离线方式。即，源域和目标域均是给定的，迁移一次即可。这种方式的缺点是显而易见的：算法无法对新加入的数据进行学习，模型也无法得到更新。与之相对的，是在线的方式。即随着数据的动态加入，迁移学习算法也可以不断地更新。

迁移学习的应用

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

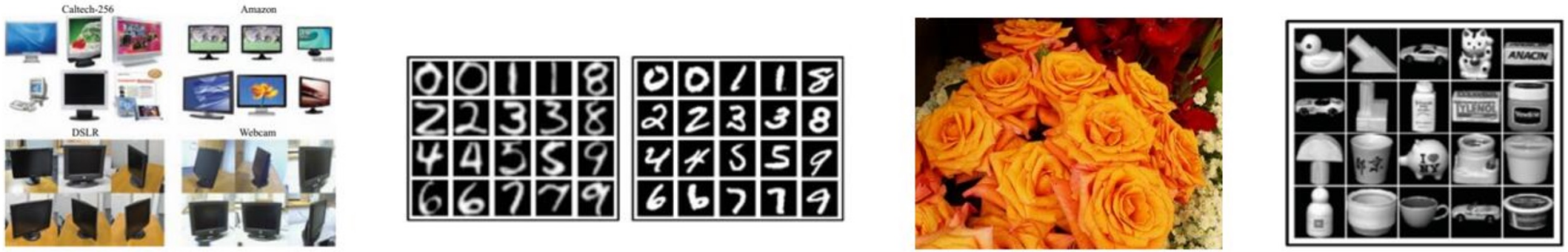
不同场景、不同设备、不同时间的室内定位

迁移学习的应用领域概览

计算机视觉

迁移学习已被广泛地应用于计算机视觉的研究中。特别地，在计算机视觉中，迁移学习方法被称为 Domain Adaptation。Domain adaptation 的应用场景有很多，比如图片分类、图片哈希等。

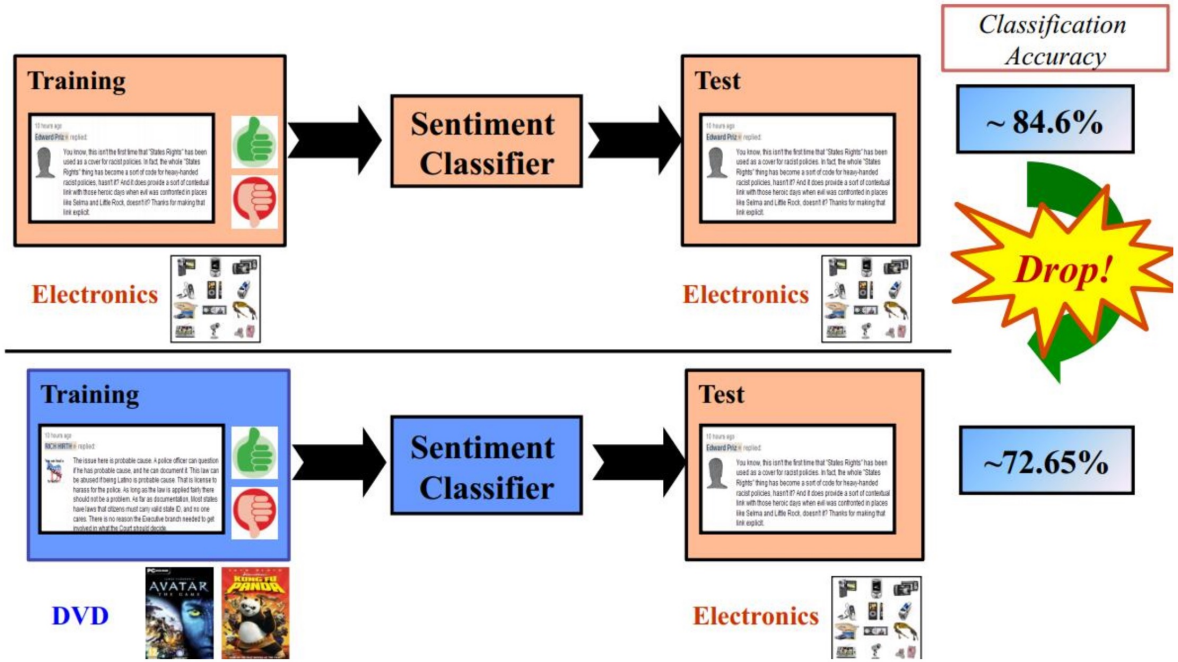
图 10展示了不同的迁移学习图片分类任务示意。同一类图片，不同的拍摄角度、不同光照、不同背景，都会造成特征分布发生改变。因此，使用迁移学习构建跨领域的鲁棒分类器是十分重要的。



迁移学习图片分类任务

文本分类

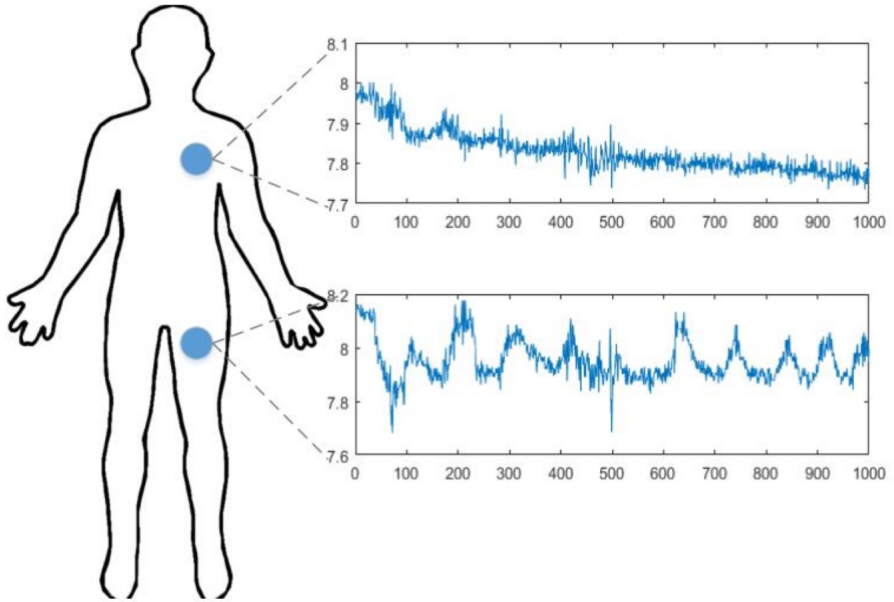
由于文本数据有其领域特殊性，因此，在一个领域上训练的分类器，不能直接拿来作用到另一个领域上。这就需要用到迁移学习。例如，在电影评论文本数据集上训练好的分类器，不能直接用于图书评论的预测。这就需要进行迁移学习。图 11是一个由电子产品评论迁移到 DVD 评论的迁移学习任务。



迁移学习文本分类任务

时间序列

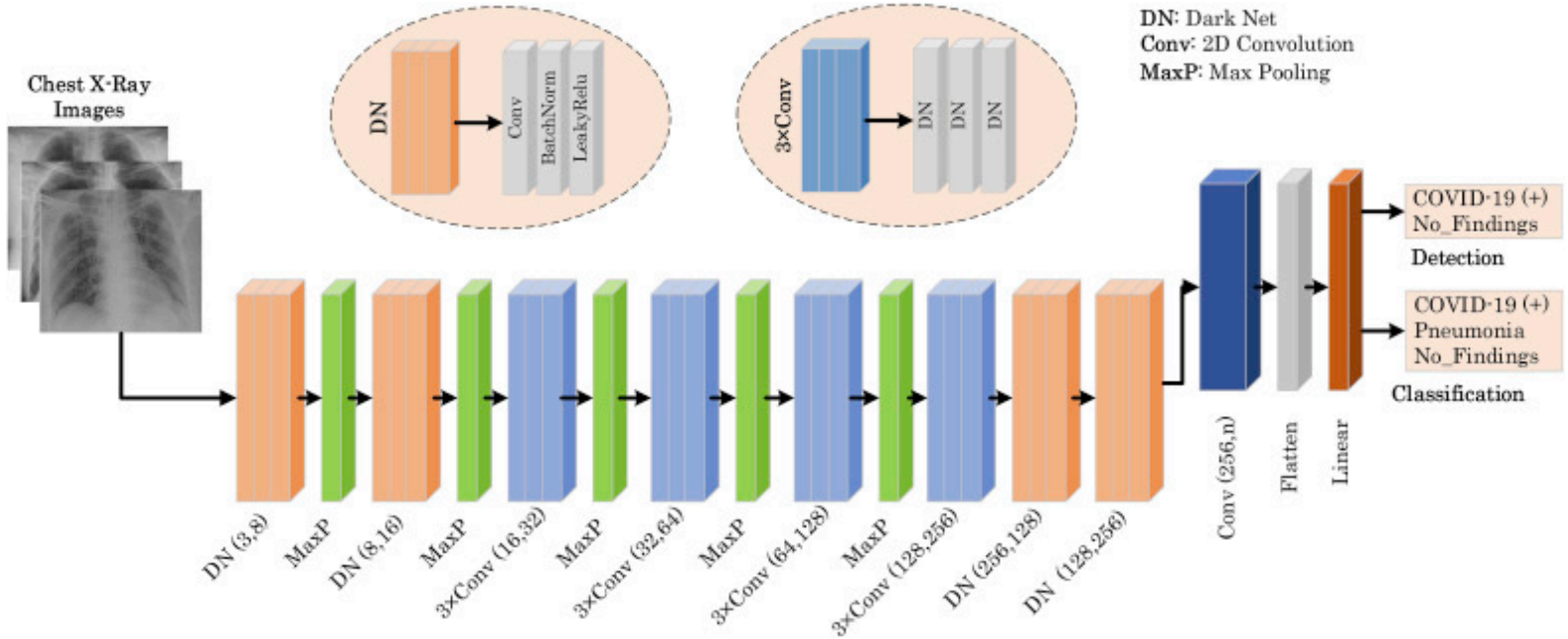
行为识别 (Activity Recognition) 主要通过佩戴在用户身体上的传感器，研究用户的行为。行为数据是一种时间序列数据。不同用户、不同环境、不同位置、不同设备，都会导致时间序列数据的分布发生变化。此时，也需要进行迁移学习。图 12展示了同一用户不同位置的信号差异性。在这个领域，华盛顿州立大学的 Diane Cook 等人在 2013 年发表的关于迁移学习在行为识别领域的综述文章 [Cook et al., 2013] 是很好的参考资料。



不同位置的传感器信号差异示意图

医疗健康

医疗健康领域的研究正变得越来越重要。不同于其他领域，医疗领域研究的难点问题是，无法获取足够有效的医疗数据。在这一领域，迁移学习同样也变得越来越重要。



新冠肺炎判别

问题形式化　基本概念：领域 (Domain) 和 任务 (Task)

领域 (Domain): 是进行学习的主体。领域主要由两部分构成：数据和生成这些数据的概率分布。通常我们用花体 $\mathcal{D}$ 来表示一个 domain，用大写斜体 P 来表示一个概率分布。

特别地，因为涉及到迁移，所以对应于两个基本的领域：**源领域 (Source Domain)** 和 **目标领域 (Target Domain)**。这两个概念很好理解。源领域就是有知识、有大量数据标注的领域，是我们要迁移的对象；目标领域就是我们最终要赋予知识、赋予标注的对象。知识从源领域传递到目标领域，就完成了迁移。

任务 (Task): 是学习的目标。任务主要由两部分组成：标签和标签对应的函数。通常我们用花体 $\mathcal{Y}$ 来表示一个标签空间，用 $f(\cdot)$ 来表示一个学习函数。

相应地，源领域和目标领域的类别空间就可以分别表示为 $\mathcal{Y}_s$ 和 $\mathcal{Y}_t$。我们用小写 y_s 和 y_t 分别表示源领域和目标领域的实际类别。

迁移学习 (Transfer Learning): 给定一个有标记的源域 $\mathcal{D}_s = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ 和一个无标记的目标域 $\mathcal{D}_t = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$。这两个领域的数据分布 $P(\mathbf{x}_s)$ 和 $\mathrm{P}(\mathbf{x}_t)$ 不同，即 $P(\mathbf{x}_s) \neq P(\mathbf{x}_t)$。迁移学习的目的就是要借助 $\mathcal{D}_s$ 的知识，来学习目标域 $\mathcal{D}_t$ 的知识 (标签)。

领域自适应 (Domain Adaptation): 给定一个有标记的源域 $\mathcal{D}_s = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ 和一个无标记的目标域 $\mathcal{D}_t = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$，假定它们的特征空间相同，即 $\mathcal{X}_s = \mathcal{X}_t$，并且它们的类别空间也相同，即 $\mathcal{Y}_s = \mathcal{Y}_t$。但是这两个域的边缘分布不同，即 $P_s(\mathbf{x}_s) \neq P_t(\mathbf{x}_t)$，条件概率分布也不同，即 $Q_s(y_s|\mathbf{x}_s) \neq Q_t(y_t|\mathbf{x}_t)$。迁移学习的目标就是，利用有标记的数据 $\mathcal{D}_s$ 去学习一个分类器 $f : \mathbf{x}_t \mapsto \mathbf{y}_t$ 来预测目标域 $\mathcal{D}_t$ 的标签 $\mathbf{y}_t \in \mathcal{Y}_t$.

距离度量

1. 欧氏距离

定义在两个向量 (两个点) 上：点 $\mathbf{x}$ 和点 $\mathbf{y}$ 的欧氏距离为：

$$d_{Euclidean} = \sqrt{(\mathbf{x}-\mathbf{y})^\top(\mathbf{x}-\mathbf{y})}$$

2. 闵可夫斯基距离

Minkowski distance，两个向量（点）的 p 阶距离：

$$d_{Minkowski} = (|\mathbf{x}-\mathbf{y}|^p)^{1/p}$$

当 $p=1$ 时就是曼哈顿距离，当 $p=2$ 时就是欧氏距离。

3. 马氏距离

定义在两个向量 (两个点) 上，这两个点在同一个分布里。点 $\mathbf{x}$ 和点 $\mathbf{y}$ 的马氏距离为：

$$d_{Mahalanobis} = \sqrt{(\mathbf{x}-\mathbf{y})^\top\Sigma^{-1}(\mathbf{x}-\mathbf{y})}$$

其中，Σ 是这个分布的协方差。

相似度度量

1. 余弦相似度

衡量两个向量的相关性 (夹角的余弦)。向量 $\mathbf{x}, \mathbf{y}$ 的余弦相似度为:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|}$$

2. 互信息

定义在两个概率分布 X, Y 上，$x \in X, y \in Y$。它们的互信息为:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

3. 皮尔逊相关系数

衡量两个随机变量的相关性。随机变量 X, Y 的 Pearson 相关系数为:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

理解：协方差矩阵除以标准差之积。

范围：$[-1, 1]$，绝对值越大表示（正/负）相关性越大。

KL散度与JS距离

1. KL 散度

Kullback–Leibler divergence，又叫做相对熵，衡量两个概率分布 $P(x), Q(x)$ 的距离：

$$D_{KL}(P||Q) = \sum_{i=1} P(x) \log \frac{P(x)}{Q(x)}$$

这是一个非对称距离：$D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

2. JS 距离

Jensen–Shannon divergence，基于 KL 散度发展而来，是对称度量：

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

其中 $M = \frac{1}{2}(P + Q)$。

最大均值差异MMD

最大均值差异是迁移学习中使用频率最高的度量。Maximum mean discrepancy，它度量在再生希尔伯特空间中两个分布的距离，是一种核学习方法。两个随机变量的距离为

$$MMD(X,Y) = \left\| \sum_{i=1}^{n_1} \phi(\mathbf{x}_i) - \sum_{j=1}^{n_2} \phi(\mathbf{y}_j) \right\|_{\mathcal{H}}^2$$

其中 $\phi(\cdot)$ 是映射，用于把原变量映射到再生核希尔伯特空间 (Reproducing Kernel Hilbert Space, RKHS) [Borgwardt et al., 2006] 中。什么是 RKHS？形式化定义太复杂，简单来说就是这个空间是对于函数的内积完备的。就是比欧几里得空间更高端的。

理解：就是求两堆数据在 RKHS 中的均值的距离。

Principal Angle

也是将两个分布映射到高维空间 (格拉斯曼流形) 中，在流形中两堆数据就可以看成两个点。Principal angle 是求这两堆数据的对应维度的夹角之和。

对于两个矩阵 $\mathbf{X}, \mathbf{Y}$，计算方法：首先正交化 (用 PCA) 两个矩阵，然后：

$$PA(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{\min(m,n)} \sin\theta_i$$

其中 m, n 分别是两个矩阵的维度，θ_i 是两个矩阵第 i 个维度的夹角，$\Theta = \{\theta_1, \theta_2, \cdots, \theta_t\}$ 是两个矩阵 SVD 后的角度：

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{U}(\cos\Theta)\mathbf{V}^\top$$

A-distance

$\mathcal{A}$-distance 是一个很简单却很有用的度量。文献 [Ben-David et al., 2007] 介绍了此距离，它可以用来估计不同分布之间的差异性。$\mathcal{A}$-distance 被定义为建立一个线性分类器来区分两个数据领域的 hinge 损失 (也就是进行二类分类的 hinge 损失)。它的计算方式是，我们首先在源域和目标域上训练一个二分类器 h，使得这个分类器可以区分样本是来自于哪一个领域。我们用 $err(h)$ 来表示分类器的损失，则 $\mathcal{A}$-distance 定义为：

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

$\mathcal{A}$-distance 通常被用来计算两个领域数据的相似性程度，以便与实验结果进行验证对比。

提纲

1 迁移学习概述

2 迁移学习基本方法

3 深度迁移学习

4 迁移学习应用案例

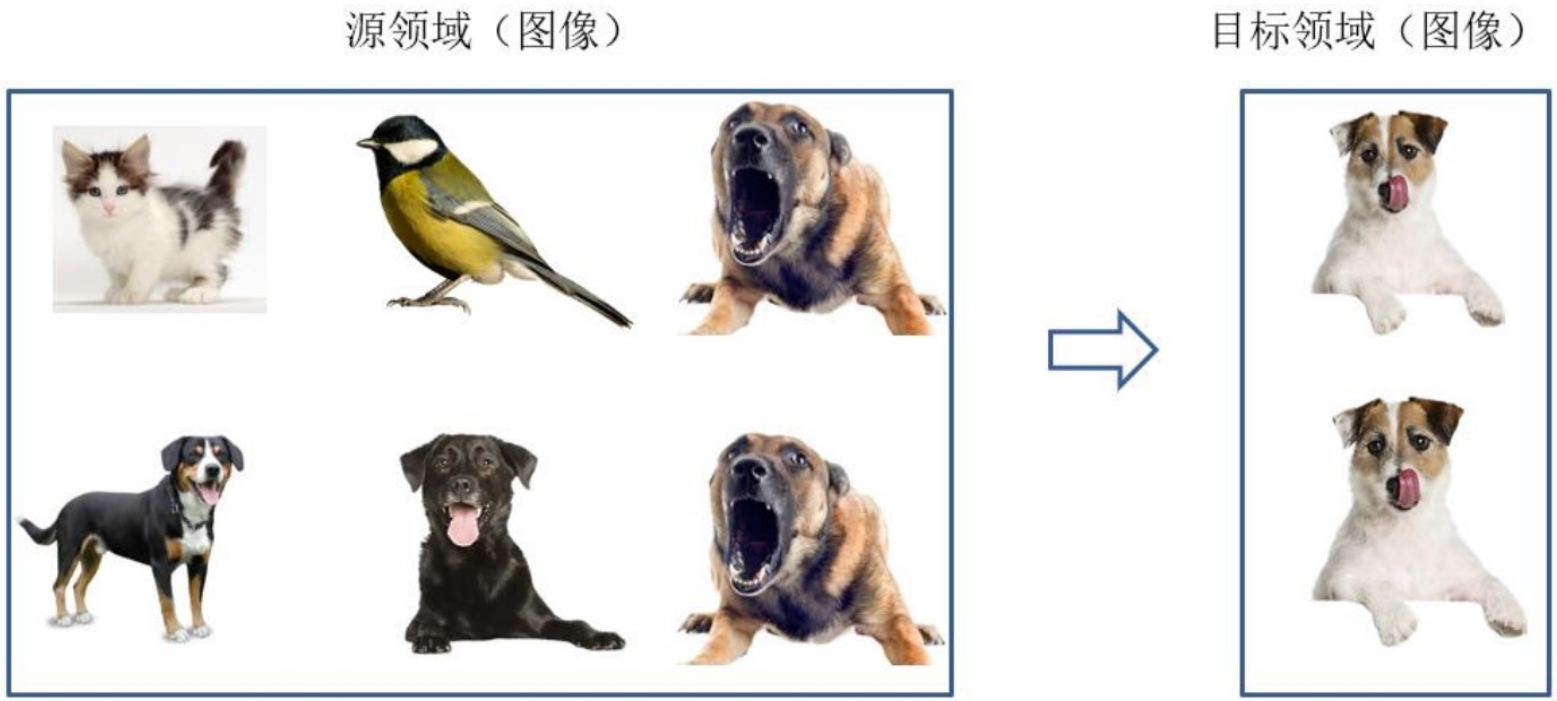
迁移学习基本方法分类

按学习方法的分类形式，最早在迁移学习领域的权威综述文章 [Pan and Yang, 2010] 给出定义。它将迁移学习方法分为以下四个大类：

1. 基于样本的迁移学习方法 (Instance based Transfer Learning)
2. 基于特征的迁移学习方法 (Feature based Transfer Learning)
3. 基于模型的迁移学习方法 (Model based Transfer Learning)
4. 基于关系的迁移学习方法 (Relation based Transfer Learning)

基于样本迁移

基于样本的迁移学习方法 (Instance based Transfer Learning) 根据一定的权重生成规则，对数据样本进行重用，来进行迁移学习。下图形象地表示了基于样本迁移方法的思想。源域中存在不同种类的动物，如狗、鸟、猫等，目标域只有狗这一种类别。在迁移时，为了最大限度地和目标域相似，我们可以人为地提高源域中属于狗这个类别的样本权重。



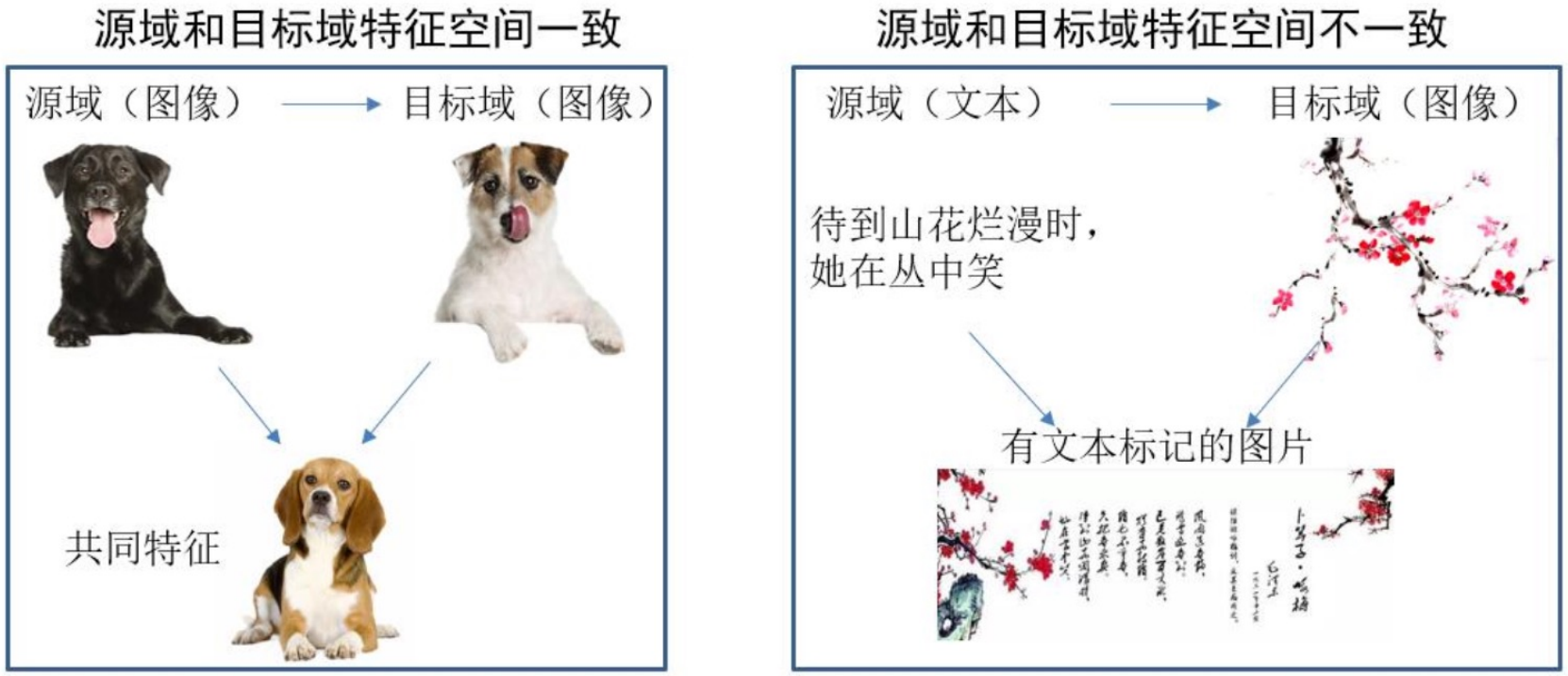
基于样本的迁移学习方法示意图

具体实现：

- **TrAdaboost** [Dai et al., 2007]：将AdaBoost 的思想应用于迁移学习中，提高有利于目标分类任务的实例权重、降低不利于目标分类任务的实例权重，并基于 PAC 理论推导了模型的泛化误差上界。
- **核均值匹配方法** (Kernel Mean Matching, KMM) [Huang et al., 2007] 对于概率分布进行估计，目标是使得加权后的源域和目标域的概率分布尽可能相近。
- **传递迁移学习方法** (Transitive Transfer Learning, TTL) [Tan et al., 2015] 和**远域迁移学习** (Distant Domain Transfer Learning, DDTL) [Tan et al., 2017]，利用联合矩阵分解和深度神经网络，将迁移学习应用于多个不相似的领域之间的知识共享，取得了良好的效果。

基于特征迁移

基于特征的迁移方法 (Feature based Transfer Learning) 是指将通过特征变换的方式互相迁移，来减少源域和目标域之间的差距；或者将源域和目标域的数据特征变换到统一特征空间中，然后利用传统的机器学习方法进行分类识别。根据特征的同构和异构性，又可以分为同构和异构迁移学习。下图展示了两种基于特征的迁移学习方法。



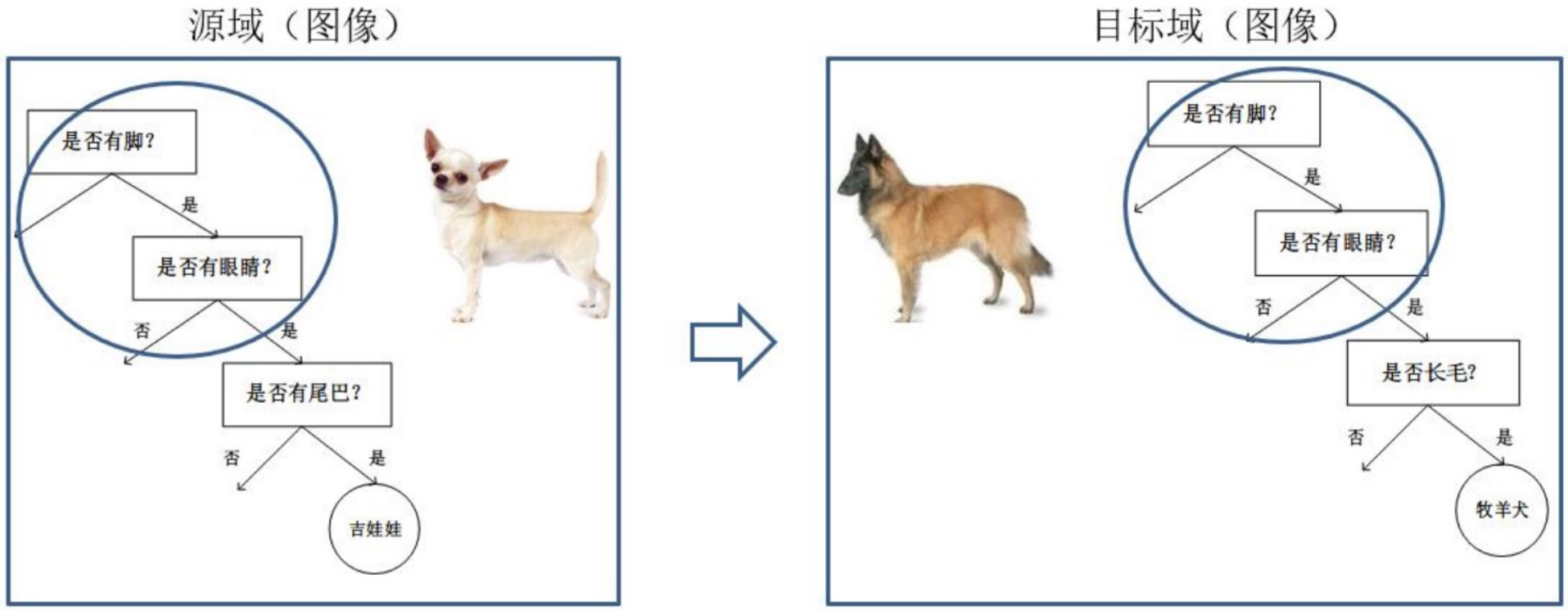
基于特征的迁移学习方法示意图

具体实现：

- **迁移成分分析方法** (Transfer Component Analysis, TCA) [Pan et al., 2011] 是其中较为典型的一个方法。该方法的核心内容是以最大均值差异 (Maximum Mean Discrepancy, MMD) 作为度量准则，将不同数据领域中的分布差异最小化。
- **基于结构对应的学习方法** (Structural Corresponding Learning, SCL) [Blitzer et al., 2006]，该算法可以通过映射将一个空间中独有的一些特征变换到其他所有空间中的轴特征上，然后在该特征上使用机器学习的算法进行分类预测。
- **迁移联合匹配** (Tranfer Joint Matching, TJM) [Long et al., 2014] 提出在最小化分布距离的同时，加入实例选择的迁移联合匹配方法，将实例和特征迁移学习方法进行了有机的结合。

基于模型迁移

基于模型的迁移方法 (Parameter/Model based Transfer Learning) 是指从源域和目标域中找到他们之间共享的参数信息，以实现迁移的方法。这种迁移方式要求的假设条件是： 源域中的数据与目标域中的数据可以共享一些模型的参数。



基于模型的迁移学习方法示意图

具体实现：

- **TransEMDT** [Zhao et al., 2011] 首先针对已有标记的数据，利用决策树构建鲁棒性的行为识别模型，然后针对无标定数据，利用 K-Means 聚类方法寻找最优化的标定参数。
- 目前绝大多数基于模型的**迁移学习方法都与深度神经网络进行结合** [Long et al., 2015a, Long et al., 2016, Long et al., 2017, Tzeng et al., 2015, Long et al., 2016]。这些方法对现有的一些神经网络结构进行修改，在网络中加入领域适配层，然后联合进行训练。因此，这些方法也可以看作是基于模型、特征的方法的结合。

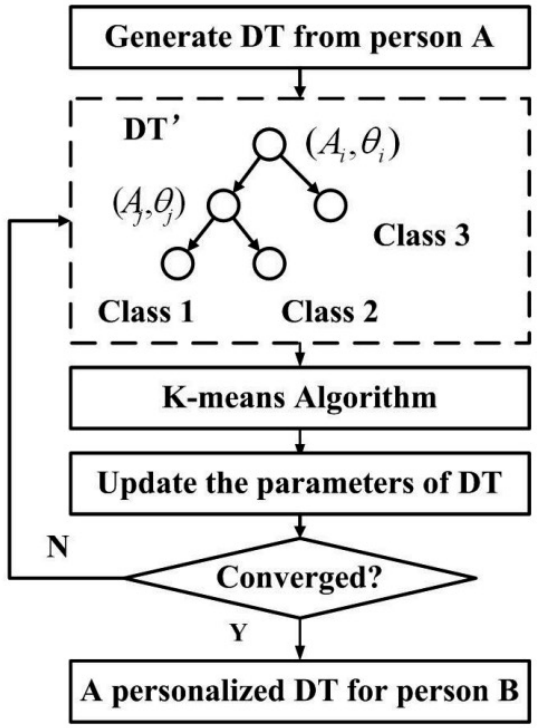
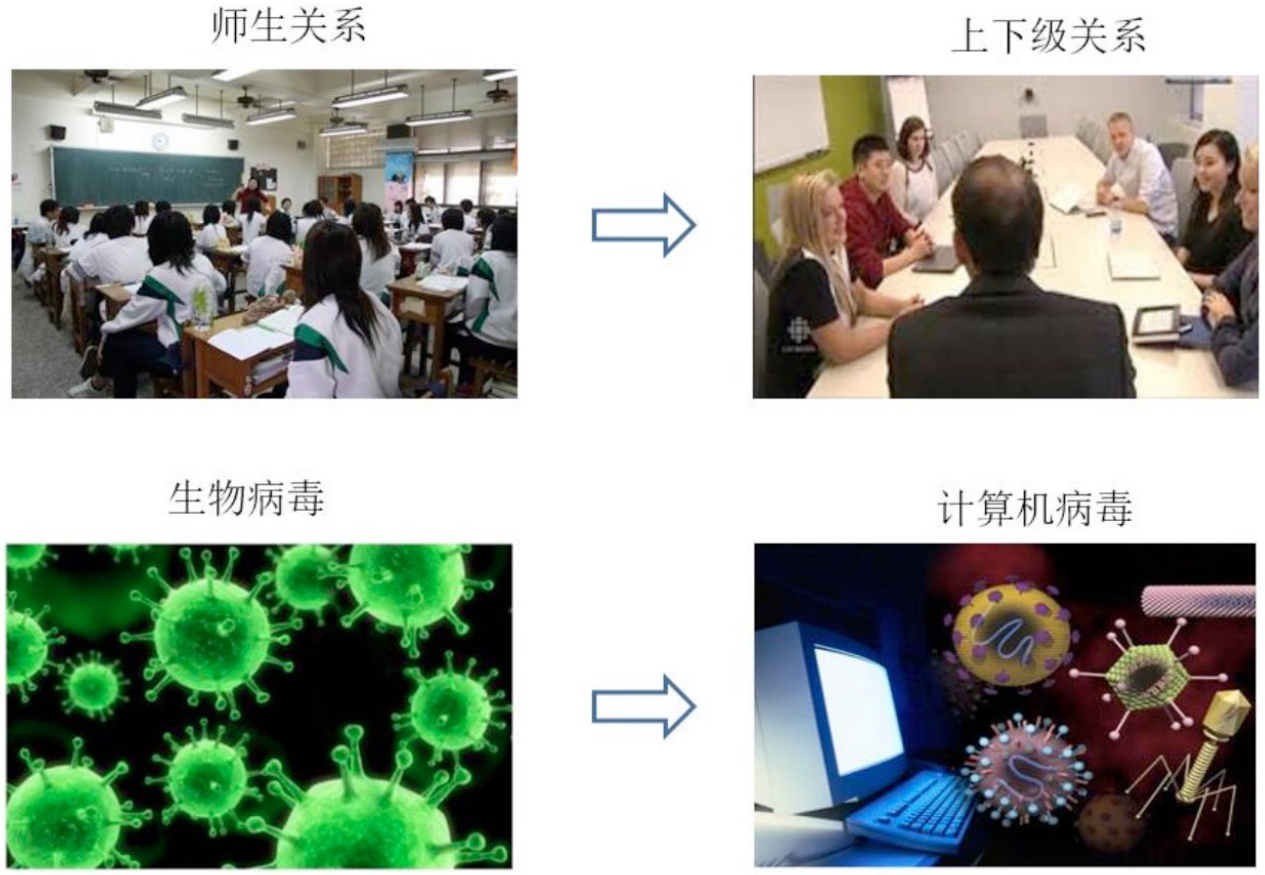


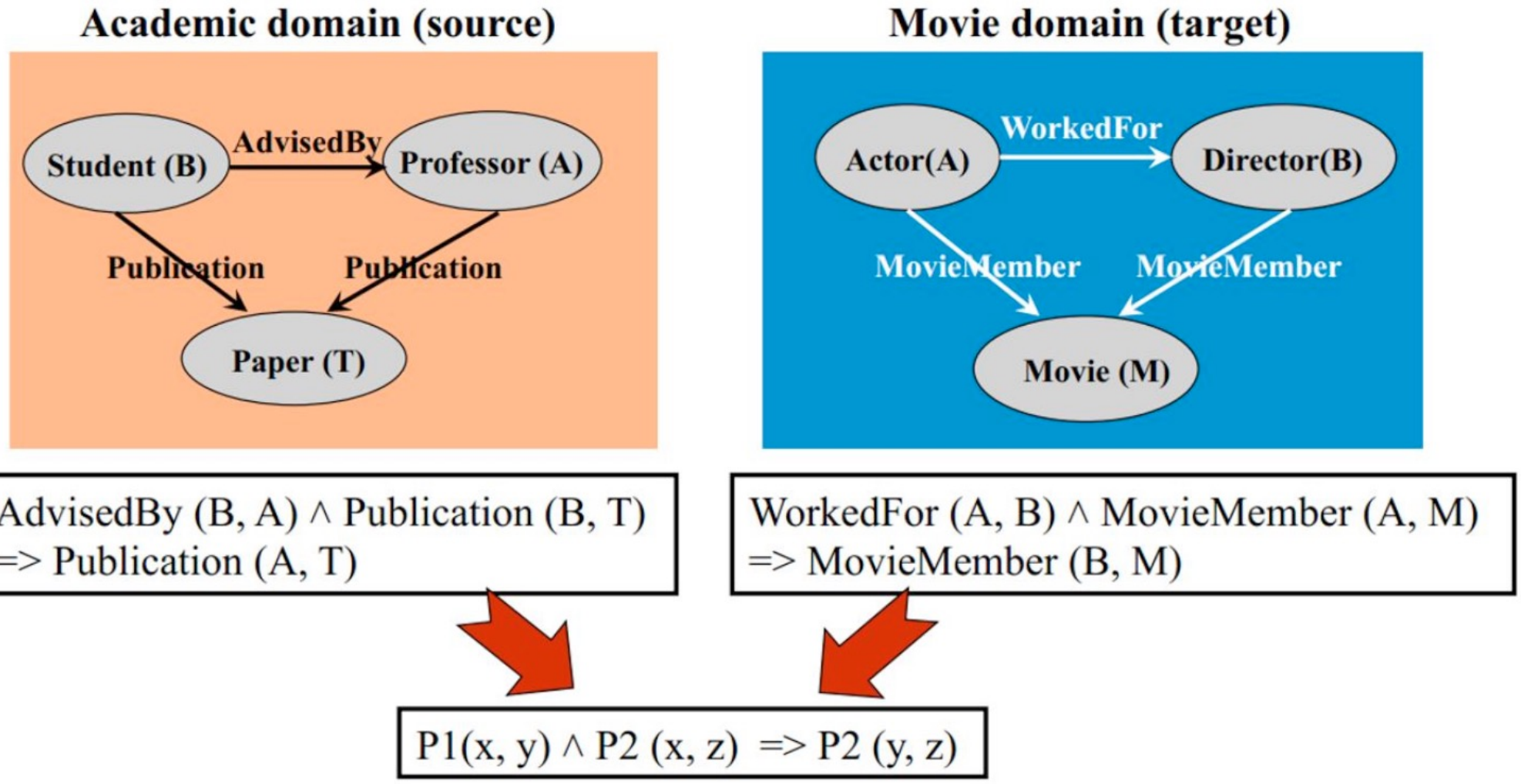
Figure 4: the TransEMDT Framework

基于关系的迁移学习方法 (Relation Based Transfer Learning) 与上述三种方法具有截然不同的思路。这种方法比较关注源域和目标域的样本之间的关系。下图展示了不同领域之间相似的关系。



基于关系的迁移学习方法示意图

基于关系迁移



基于马尔科夫逻辑网的关系迁移

提纲

1 迁移学习概述

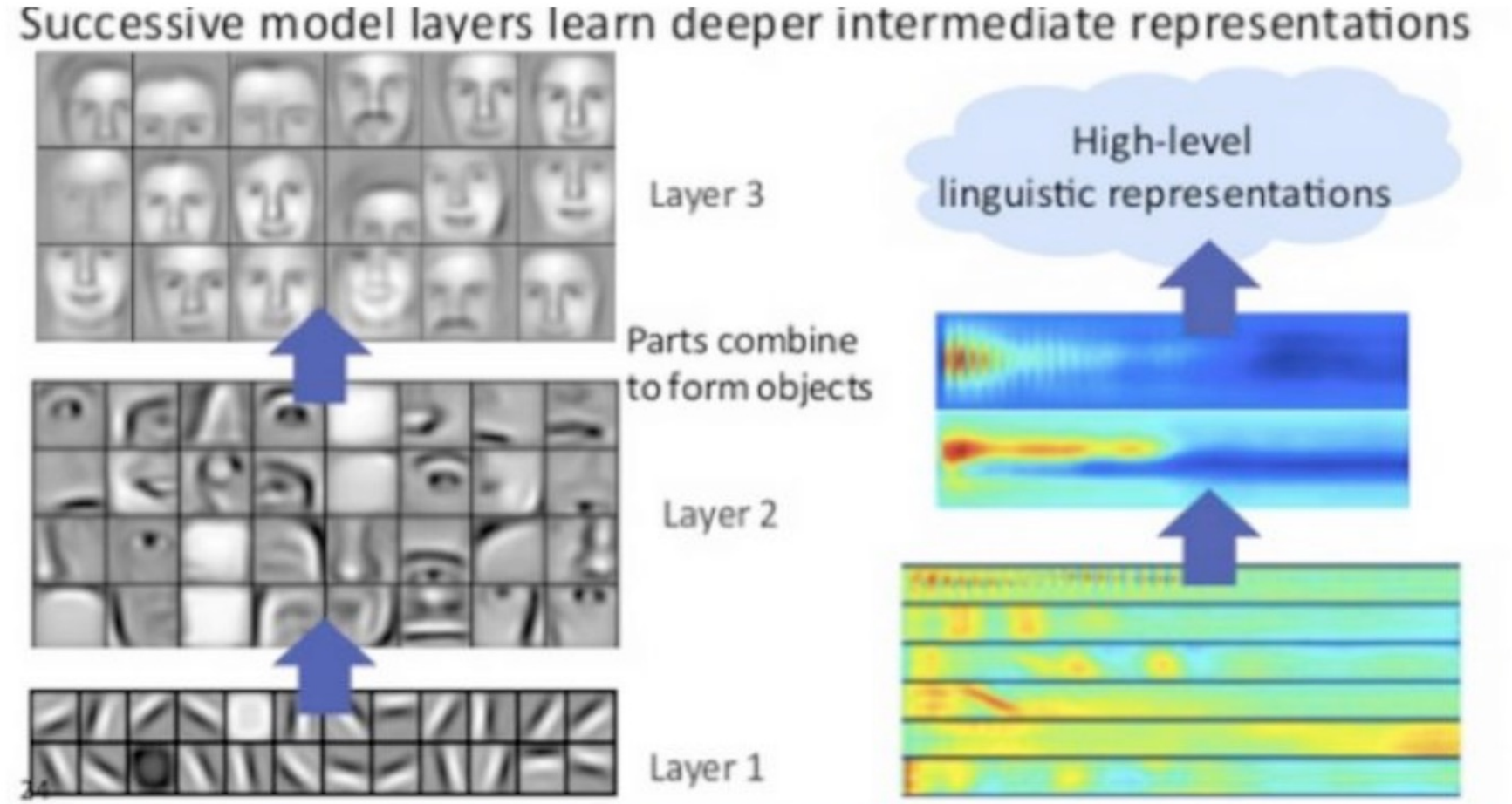
2 迁移学习基本方法

3 深度迁移学习

4 迁移学习应用案例

深度网络的可迁移性

为什么深度网络可以迁移？



特定任务的特征

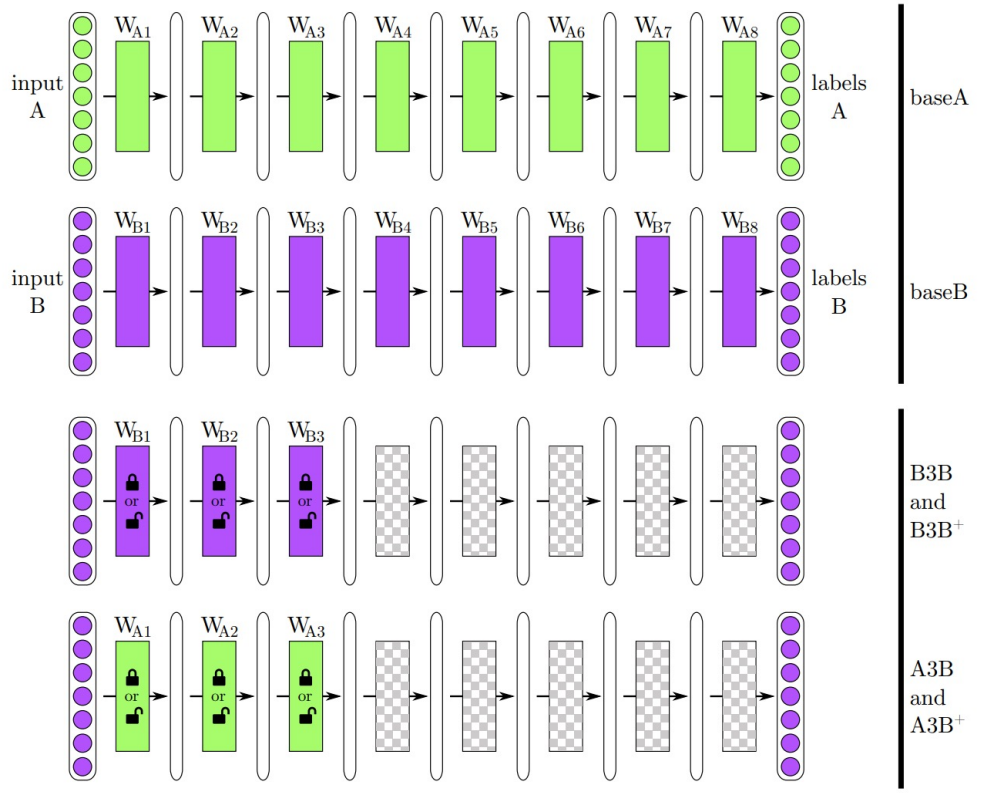
通用特征

深度神经网络进行特征提取到分类的简单示例

深度网络的可迁移性

如何得知哪些层能够学习到 general feature，哪些层能够学习到 specific feature 。更进一步： 如果应用于迁移学习，如何决定该迁移哪些层、固定哪些层？

创建A和B数据集时，对ImageNet 1000类的物体随机分成两组，每组包含500类和大约一半的数据，每个类大约645000张图片。。我们在任务A上训练一个八层卷积网络，在B上训练另一个八层卷积网络。这些网络，我们称之为baseA和baseB，如图1的前两行。



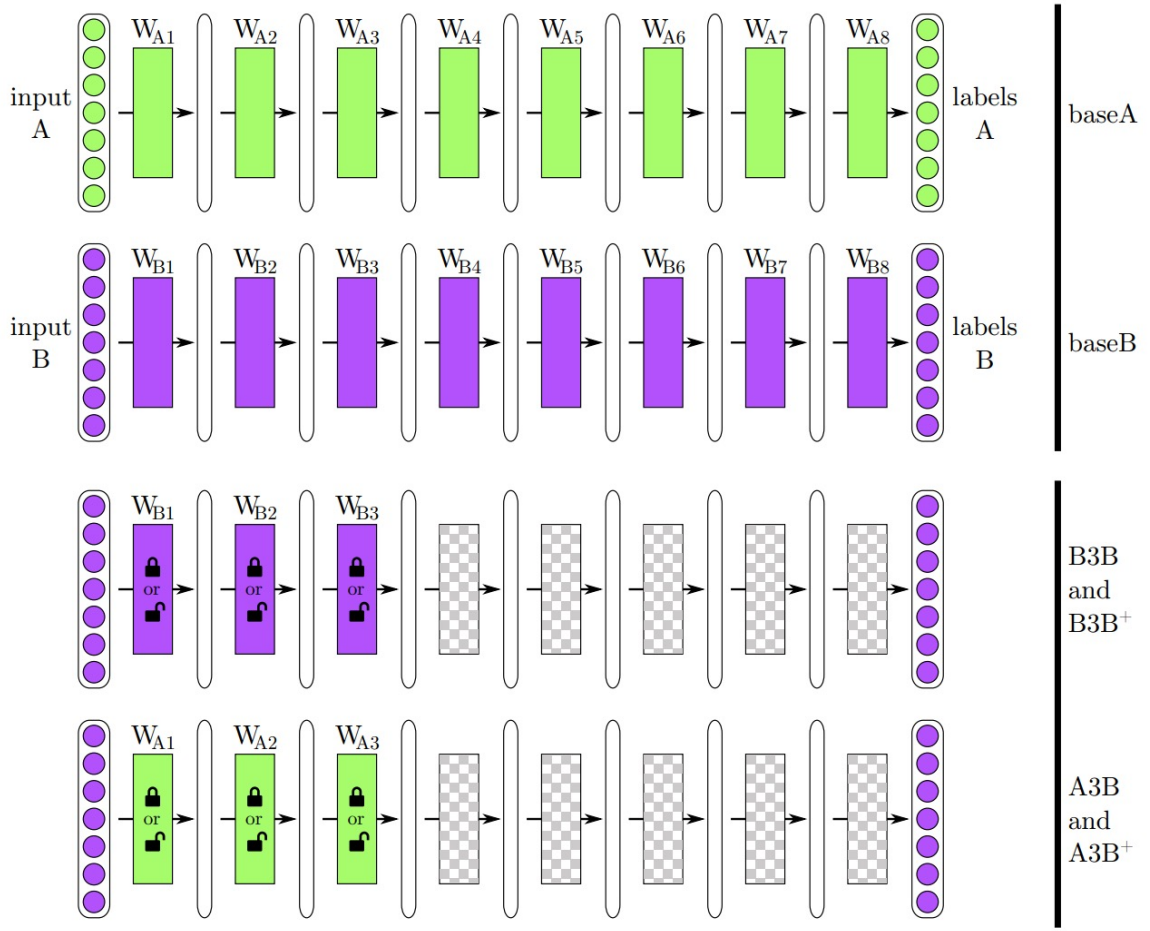
How transferable are features in deep neural networks?
J Yosinski, J Clune, Y Bengio, H Lipson
arXiv preprint arXiv:1411.1792

6504 2014

深度网络的可迁移性

测试四种模型迁移方式：

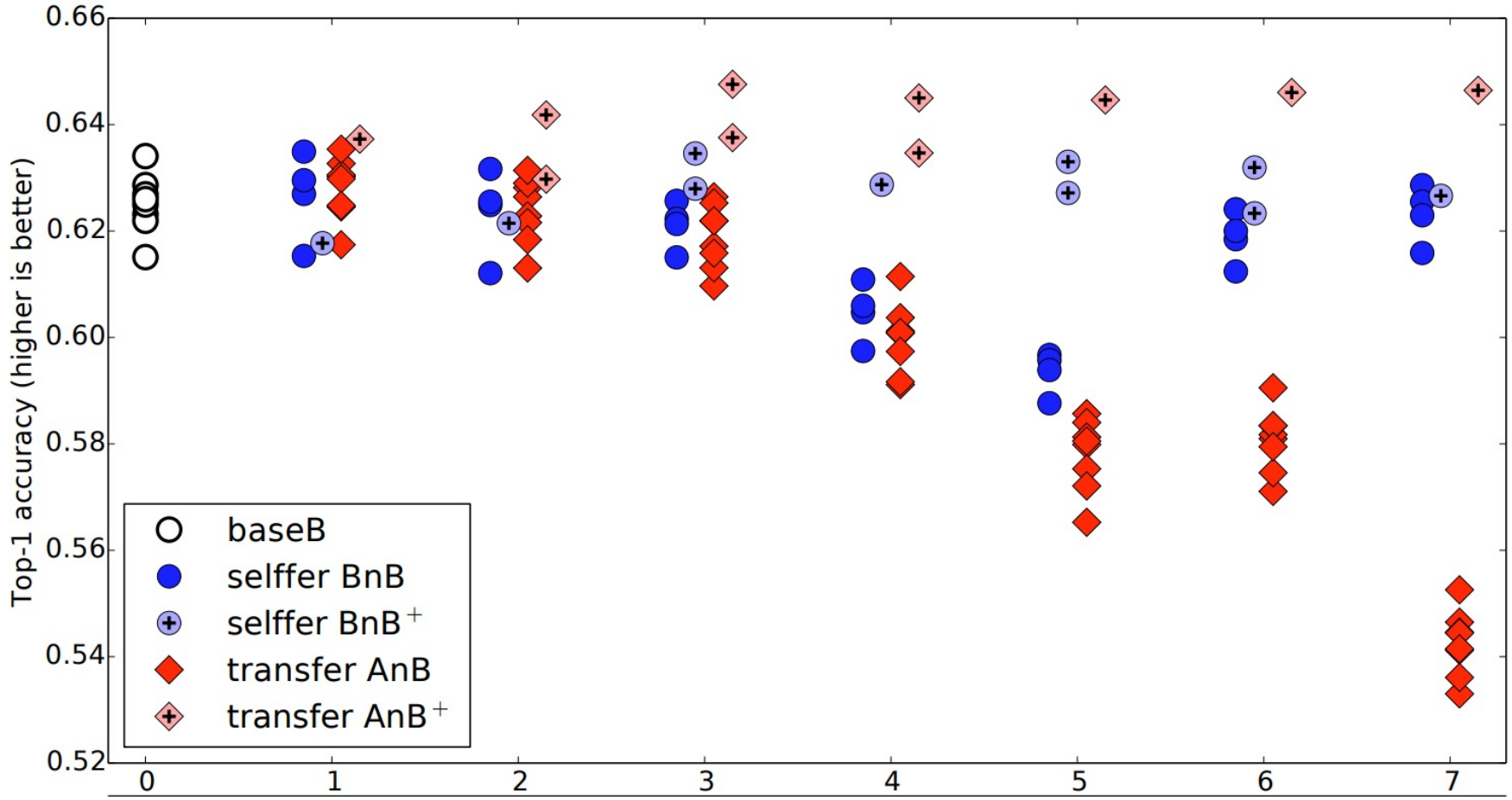
1. **B3B（图1第3行）**：前三层的参数从baseB中迁移而来，并且固定不变。第四到第八层的参数随机初始化，然后在数据集B上进行训练。
2. **A3B（图1第4行）**：前三层的参数从baseA中迁移而来，并且固定不变。第四到第八层的参数随机初始化，然后在数据集B上进行训练。
3. **B3B+**：同B3B，只不过迁移的参数需要微调。
4. **A3B+**：同A3B，只不过迁移的参数需要微调。



How transferable are features in deep neural networks? 6504 2014
J Yosinski, J Clune, Y Bengio, H Lipson
arXiv preprint arXiv:1411.1792

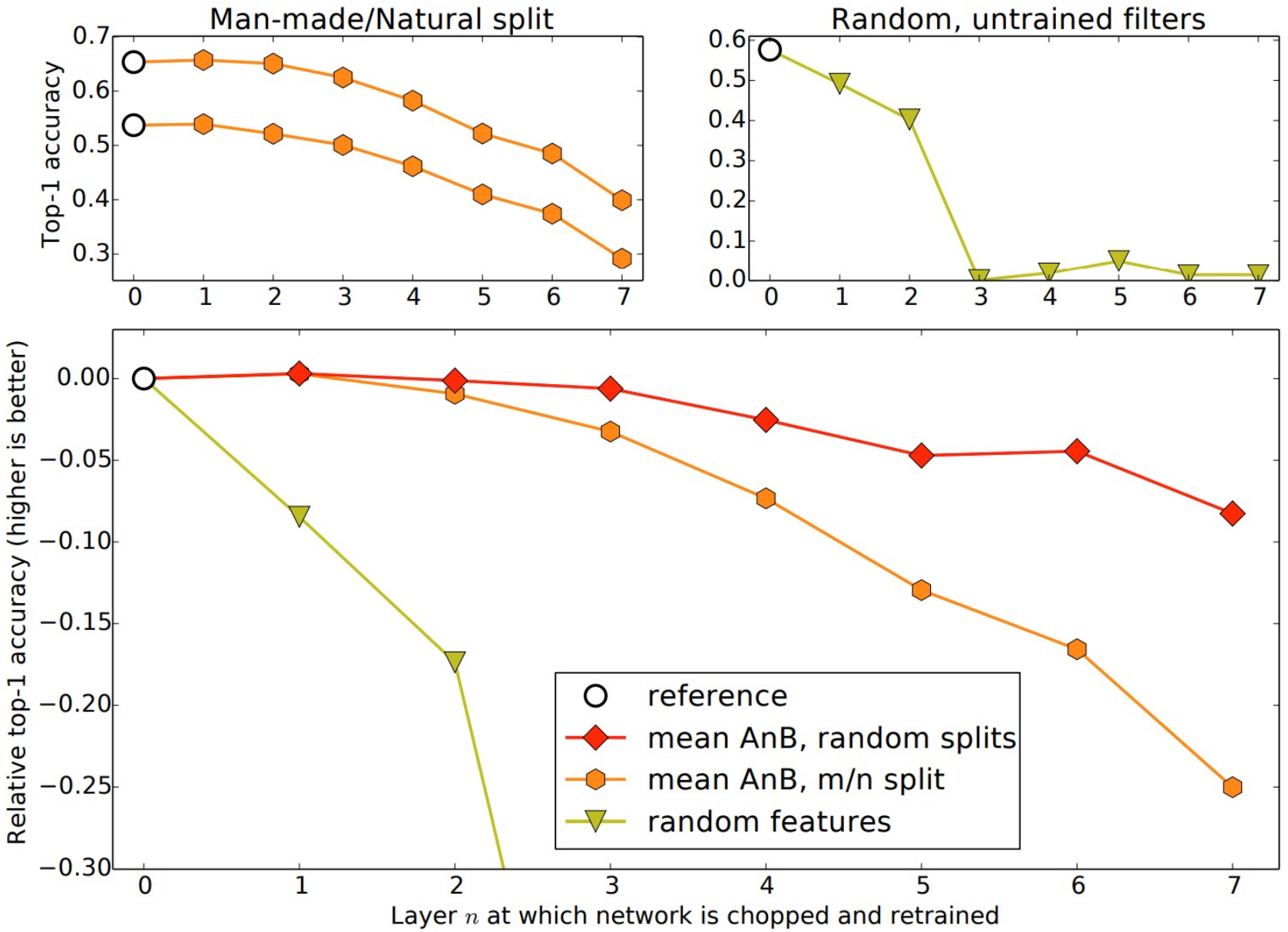
深度网络的可迁移性

1. **B3B**：前三层的参数从baseB中迁移而来，并且固定不变。第四到第八层的参数随机初始化，然后在数据集B上进行训练。
2. **A3B**：前三层的参数从baseA中迁移而来，并且固定不变。第四到第八层的参数随机初始化，然后在数据集B上进行训练。
3. **B3B+**：迁移的参数需要微调。
4. **A3B+**：迁移的参数需要微调。



How transferable are features in deep neural networks? 6504 2014
J Yosinski, J Clune, Y Bengio, H Lipson
arXiv preprint arXiv:1411.1792

深度网络的可迁移性


Man-made/Natural split
Top-1 accuracy
Random, untrained filters
Relative top-1 accuracy (higher is better)
Layer n at which network is chopped and retrained
reference
mean AnB, random splits
mean AnB, m/n split
random features

深度网络的可迁移性

如何得知哪些层能够学习到 general feature，哪些层能够学习到 specific feature。更进一步： 如果应用于迁移学习，如何决定该迁移哪些层、固定哪些层？

- 神经网络的前 3 层基本都是 general feature，进行迁移的效果会比较好；
- 深度迁移网络中加入 fine-tune，效果会提升比较大，可能会比原网络效果还好；
- Fine-tune 可以比较好地克服数据之间的差异性；
- 深度迁移网络要比随机初始化权重效果好；
- 网络层数的迁移可以加速网络的学习和优化

How transferable are features in deep neural networks? 6504 2014
J Yosinski, J Clune, Y Bengio, H Lipson
arXiv preprint arXiv:1411.1792

最简单的深度迁移 —— Finetune

1. 为什么需要已经训练好的网络？

在实际的应用中，我们通常不会针对一个新任务，就去从头开始训练一个神经网络。这样的操作显然是非常耗时的。尤其是，我们的训练数据不可能像 ImageNet 那么大，可以训练出泛化能力足够强的深度神经网络。即使有如此之多的训练数据，我们从头开始训练，其代价也是不可承受的

2. 为什么需要 finetune？

因为别人训练好的模型，可能并不是完全适用于我们自己的任务。可能别人的训练数据和我们的数据之间不服从同一个分布；可能别人的网络能做比我们的任务更多的事情；可能别人的网络比较复杂，我们的任务比较简单。

3. Finetune 的优势

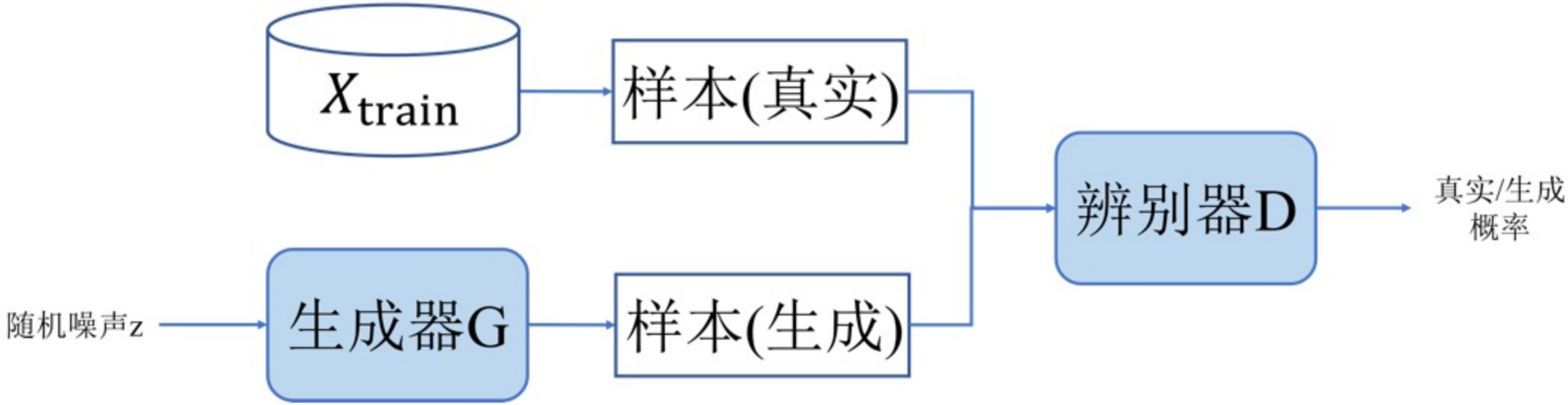
- 不需要针对新任务从头开始训练网络，节省了时间成本；
- 预训练好的模型通常都是在大数据集上进行的，无形中扩充了我们的训练数据，使得模型更鲁棒、泛化能力更好；
- Finetune 实现简单，使得我们只关注自己的任务即可。

4. Finetune 的扩展

Finetune 并不只是针对深度神经网络有促进作用，对传统的非深度学习也有很好的效果。例如，可以使用深度网络对原始数据进行训练，依赖网络提取出更丰富更有表现力的特征。然后，将这些特征作为传统机器学习方法的输入，既避免了繁复的手工特征提取，又能自动地提取出更有表现力的特征。

深度对抗网络迁移

GAN 受到自博弈论中的二人零和博弈 (two-player game) 思想的启发而提出。它一共包括两个部分：一部分为生成网络 (Generative Network)，此部分负责生成尽可能地以假乱真的样本，这部分被成为生成器 (Generator)；另一部分为判别网络 (Discriminative Network)，此部分负责判断样本是真实的，还是由生成器生成的，这部分被成为判别器 (Discriminator)。生成器和判别器的互相博弈，就完成了对抗训练。



由于在迁移学习中，天然地存在一个源领域，一个目标领域，因此，我们可以免去生成样本的过程，而直接将其中一个领域的数据（通常是目标域）当作是生成的样本。此时，生成器的职能发生变化，不再生成新样本，而是扮演了特征提取的功能：不断学习领域数据的特征，使得判别器无法对两个领域进行分辨。这样，原来的生成器也可以称为特征提取器（Feature Extractor）。

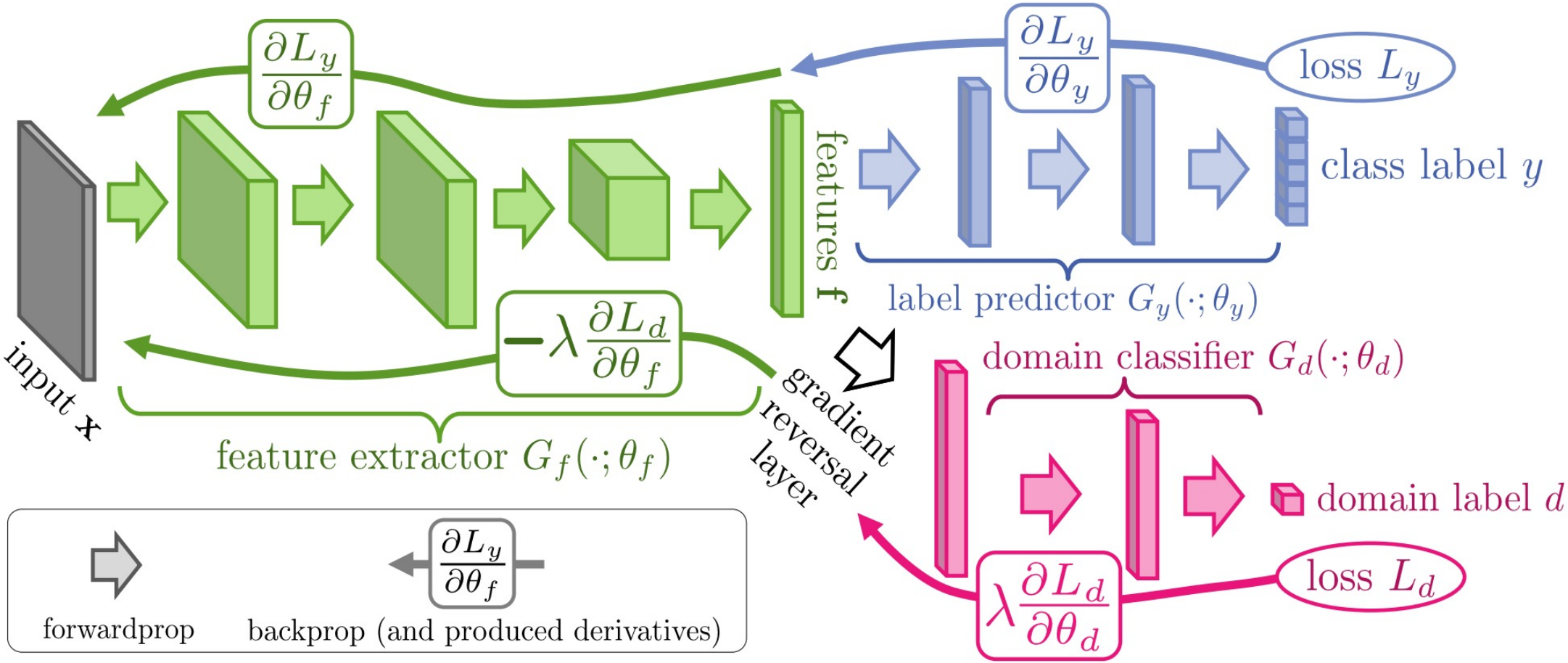
通常用 G_f 来表示特征提取器，用 G_d 来表示判别器。

正是基于这样的领域对抗的思想，深度对抗网络可以被很好地运用于迁移学习问题中。

与深度网络自适应迁移方法类似，深度对抗网络的损失也由两部分构成：网络训练的损失 ℓ_c 和领域判别损失 ℓ_d：

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

深度对抗网络迁移 - DANN

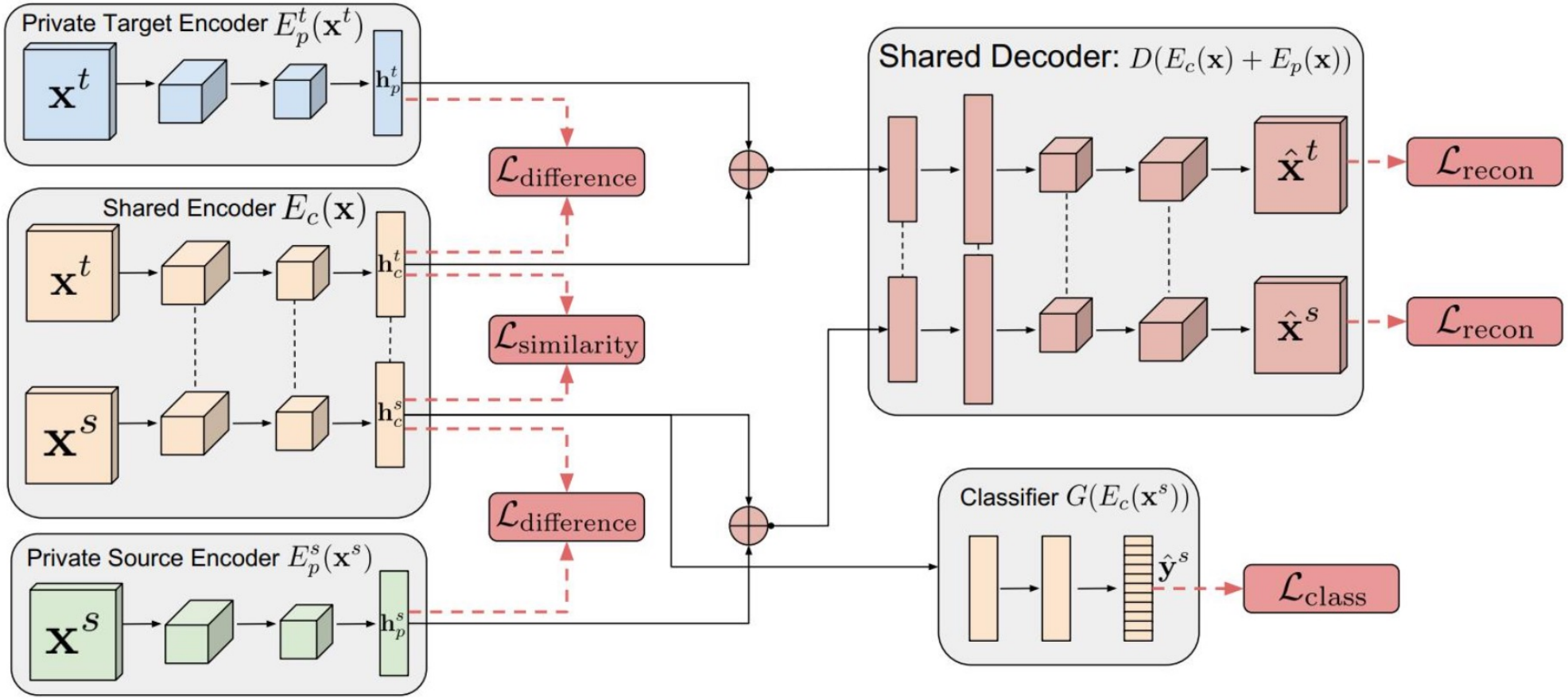


Domain-adversarial training of neural networks 3860 2016

Y Ganin, E Ustinova, H Ajakan, P Germain, H Larochelle, F Laviolette, ...

The journal of machine learning research 17 (1), 2096-2030

深度对抗网络迁移 - DSN



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

Domain separation networks 948 2016

K Bousmalis, G Trigeorgis, N Silberman, D Krishnan, D Erhan

Advances in neural information processing systems 29, 343-351

提纲

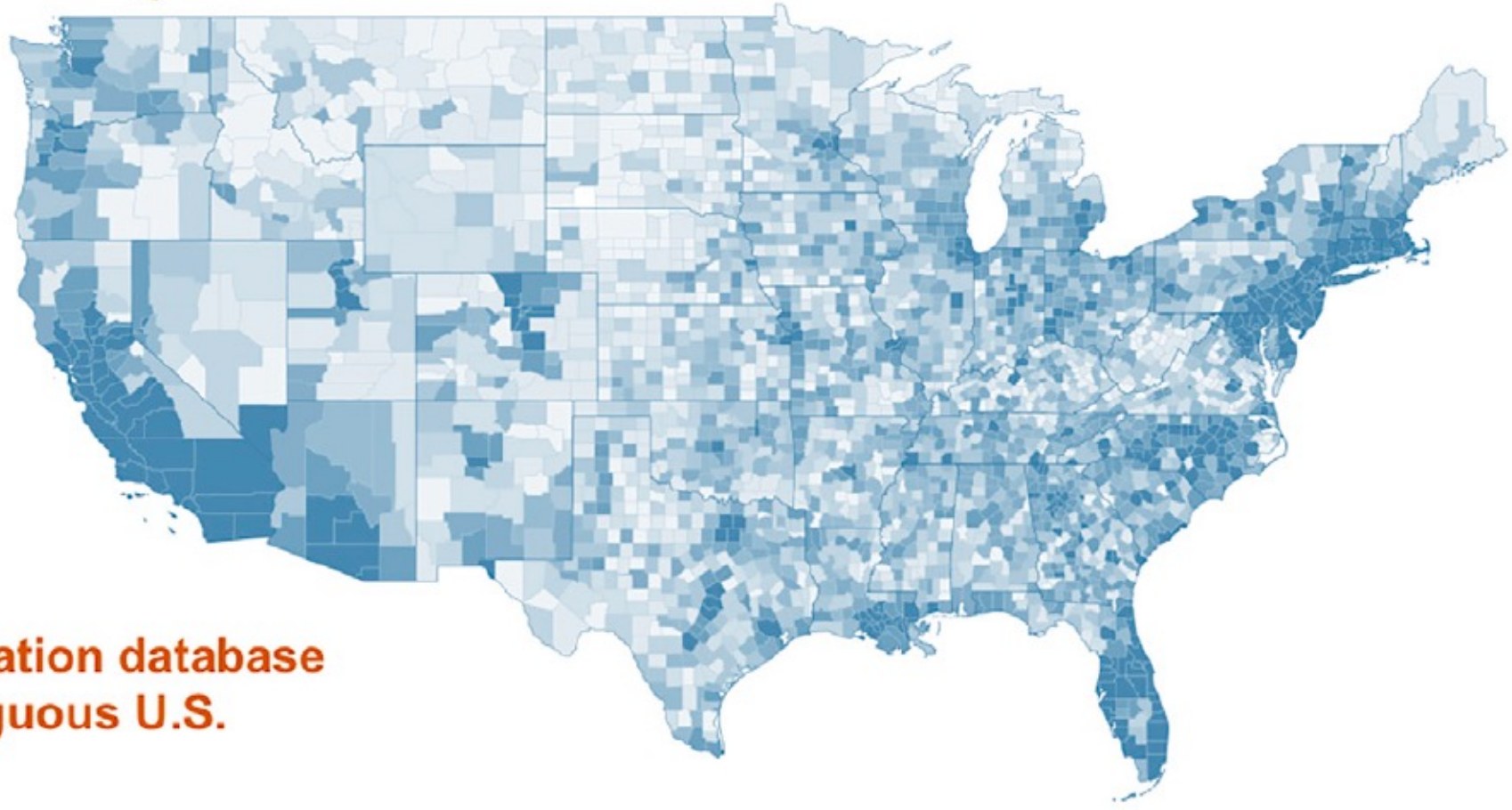
1 迁移学习概述

2 迁移学习基本方法

3 深度迁移学习

4 迁移学习应用案例

DeepSolar

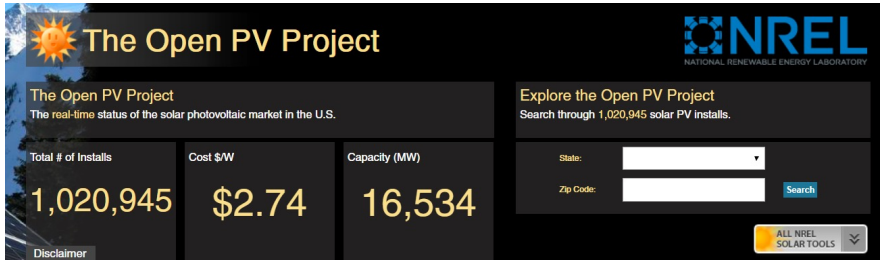


Solar installation database of the contiguous U.S.

Yu, Jiafan, et al. "DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States." *Joule* 2.12 (2018): 2605-2617.

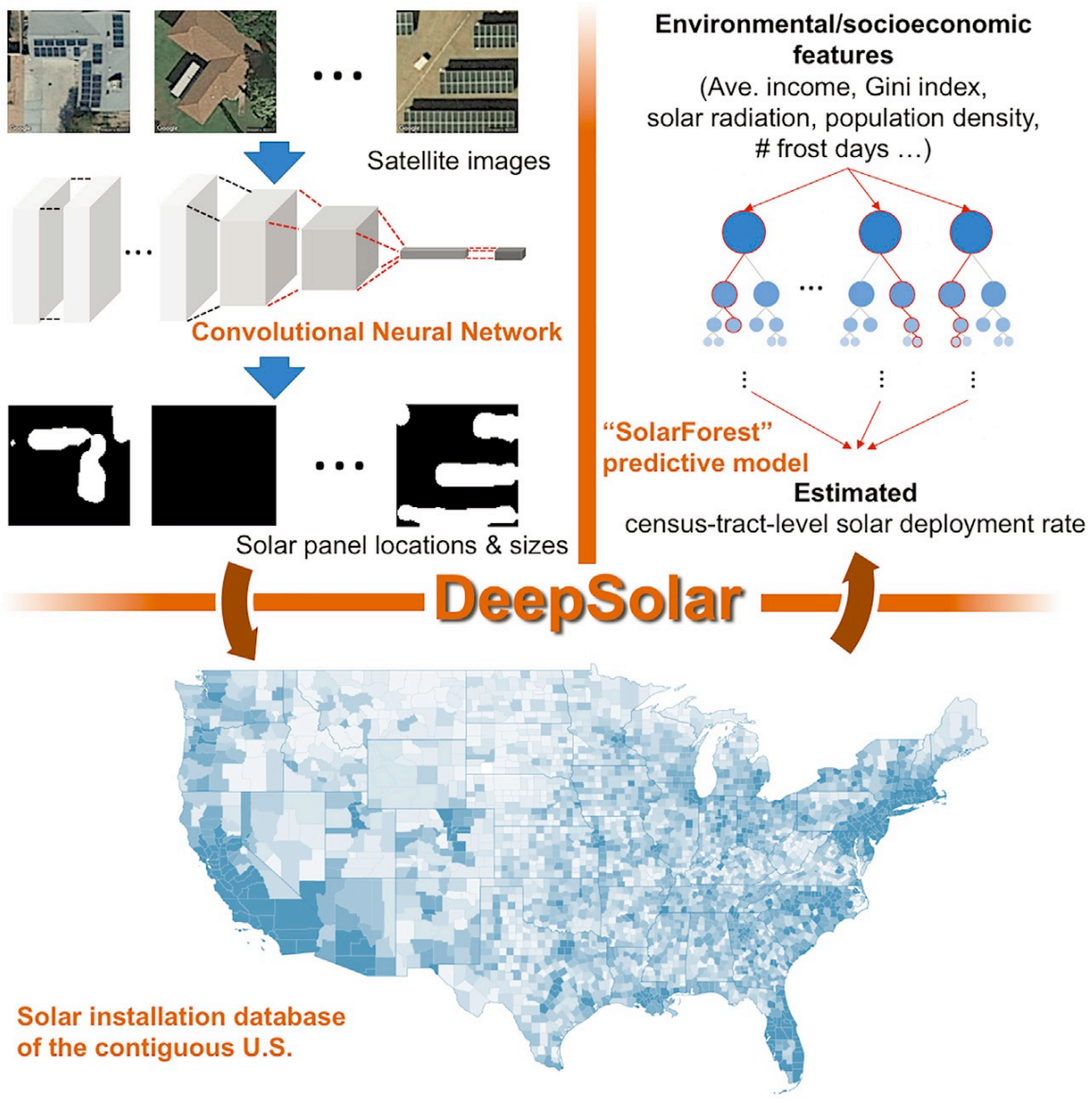
问题描述

- 在全美范围内建立精准的太阳能使用数据库，包括安装位置及面积
- 其他类似方案:
  - Open PV Project (~ 1 million PV systems), relies on voluntary surveys and self-reports, but **runs the risk of being incomplete and outdated**.
  - Google's Project Sunroof (~ 0.67 million PV systems), utilizes machine learning approach to report locations **without any size information**.



解决方案

- 采集海量卫星图像数据及部分标签
- 识别卫星图像中是否包含太阳能电板
- 分割太阳能电板，并计算其面积
- 将太阳能电板使用率与环境和社会经济信息相关联
- 提出预测模型 SolarForest 预测太阳能使用率



A. 输入从Google地图获取的局部卫星图像；

B. 基于CNN的图像分类器

C. 分类结果（是否包含Solar Panel）

D. 使用半监督方法，对包含Solar Panel的图像进行分割操作；

E. 生成包含Solar Panel位置的激活地图

F. 对激活地图进行二值化，计算Solar Panel面积

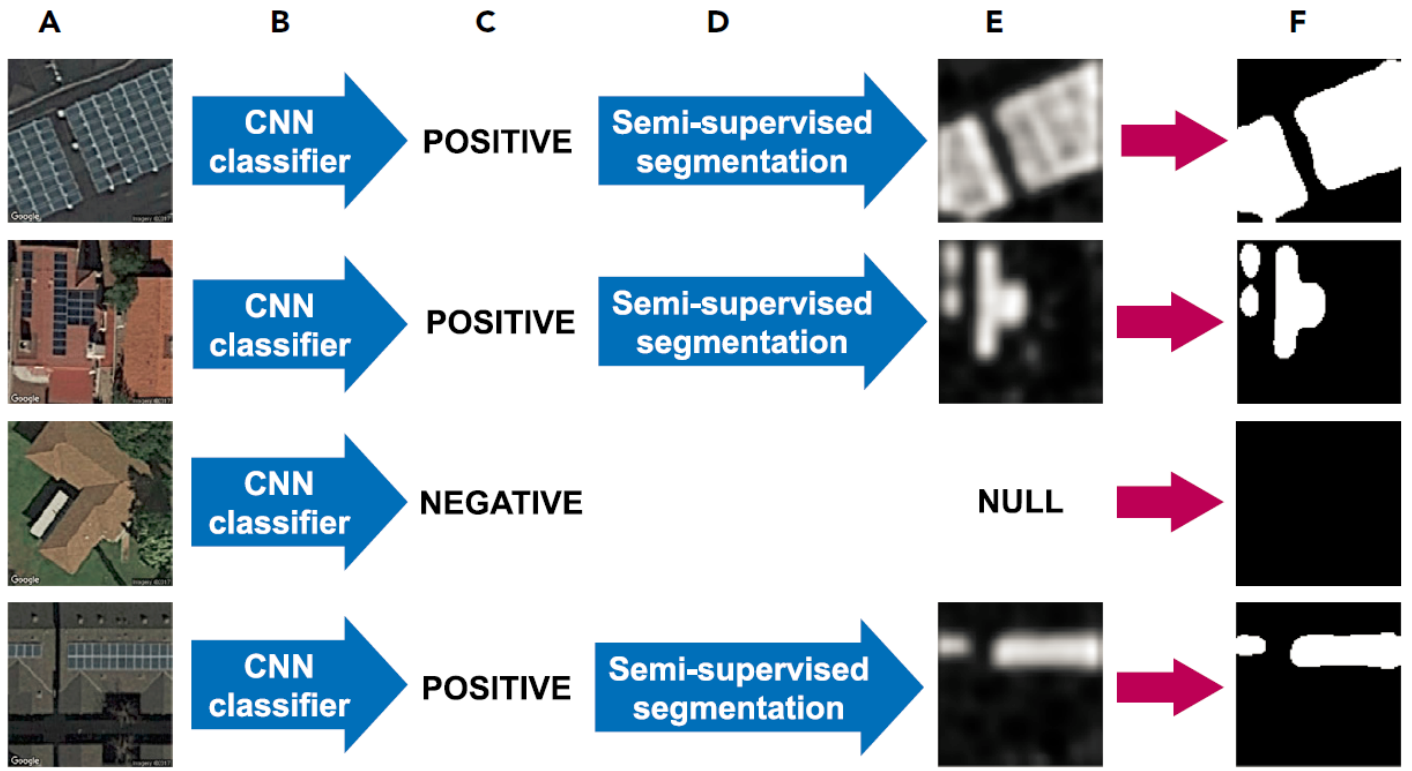


Figure 1. Schematic of DeepSolar Image Classification and Segmentation Framework

数据准备

- 数据源: Google Static Map API (15 cm resolution), width of image tile: 15 – 22m
- 手动标注所有数据几乎不可能.
- 联合人工手动与CNN模型自动标注
  1. manually labelling 320,000 images with Amazon Mechanical Turk (MTurk), a crowdsourcing platform for solving labor intensive task. (Less than 3,000 are positive!)
  2. training a preliminary classification model based on VGG-1637 network with all positive samples and a small fraction of the negative samples, then applying VGG-1637 to larger dataset.
  3. Leveraging MTurk, false positive samples are removed from the positive-predicted samples.

  Finally, authors construct a dataset containing 472,953 samples (0.043% of the total number of images scanned in US), of which 50,507 are actually positive.

太阳能电板分类 – 迁移学习

- 对先进的CNN模型 Inception-v3 [Szegedy, C., et al. CVPR 2016] 进行迁移
- 对模型最后一层参数随机初始化，其他层保留预训练模型参数，然后进行Finetune
- 采取Cost sensitive Learning来更新Loss函数，解决样本不均衡问题.

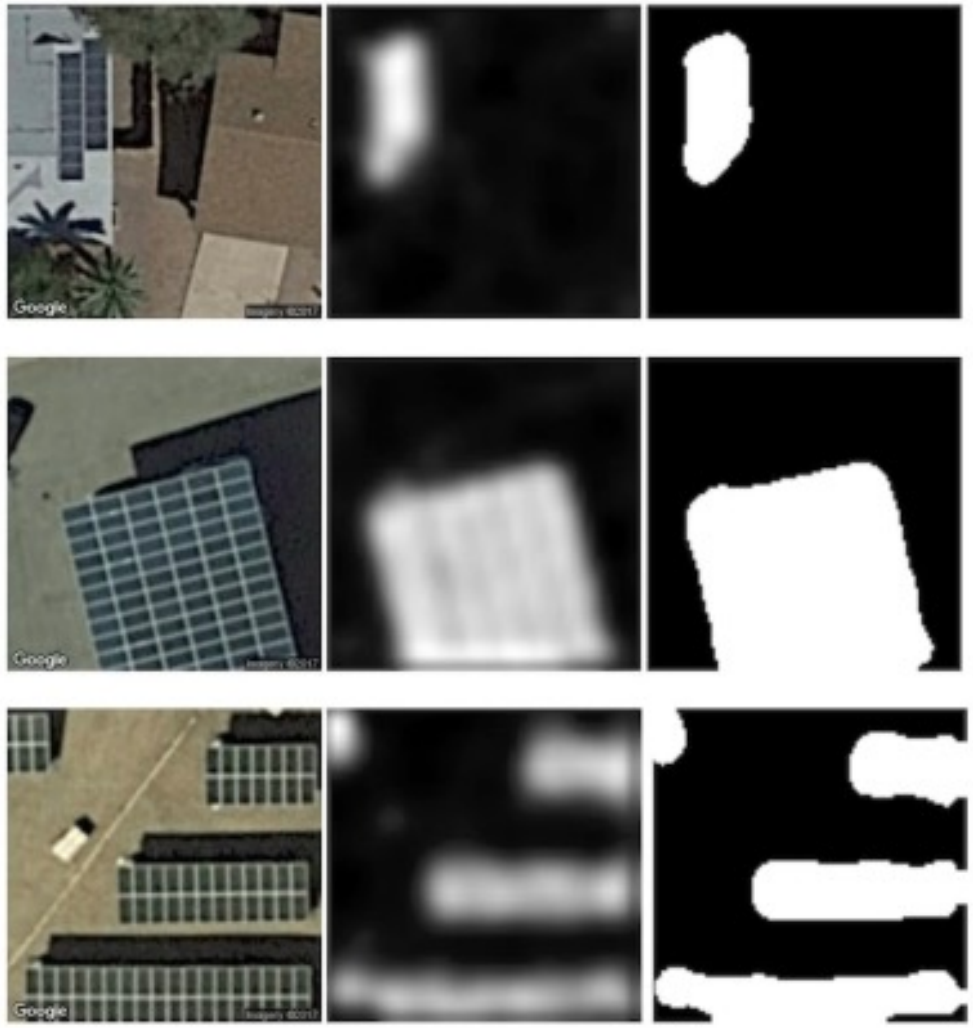
$$L = \sum_i [\alpha \mathbf{1}(y_i = 1)\text{softmax}(y_i, f(x_i)) + \mathbf{1}(y_i = 0)\text{softmax}(y_i, f(x_i))] + \text{regularization term}$$

- 输出两个概率值，分别表示正样本 (containing solar) 和 负样本 (not containing solar).

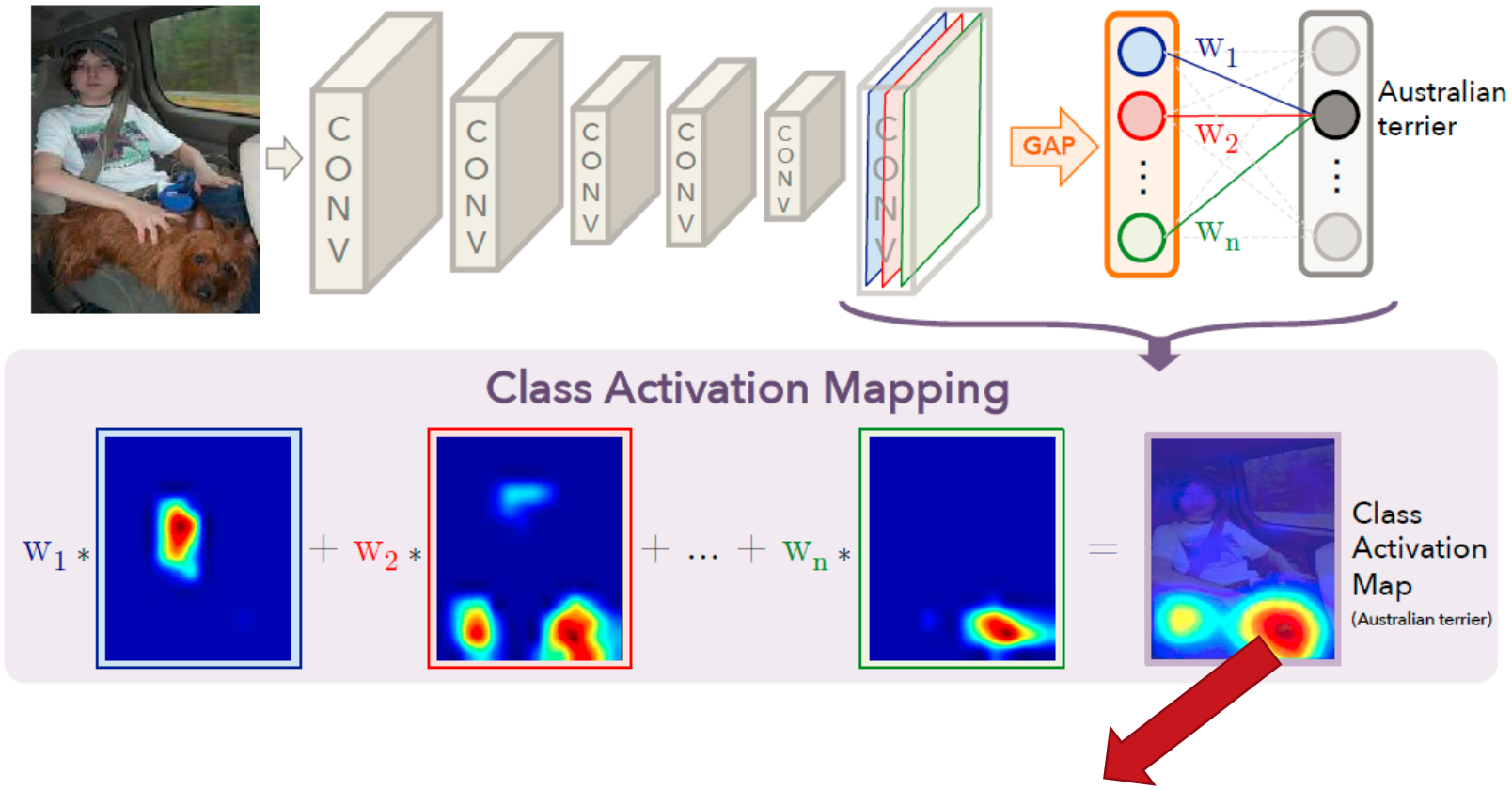
Szegedy, C., et al. Rethinking the inception architecture for computer vision. CVPR 2016, 2818–2826.

半监督方法对 Panel 进行分割

- 目标：准确定位卫星图像中的Solar Panel，并估计面积
- 图像在训练集中只标注了是否包含Solar Panel，没有真实位置和面积信息
- 在训练好的CNN分类器中，作者生成了类别激活地图（ Class Activation Map ），以此辅助面积估计



Class Activation Mapping



Saliency map

Learning deep features for discriminative localization
B Zhou, A Khosla, A Lapedriza, A Oliva, A Torralba
Proceedings of the IEEE conference on computer vision and pattern …

4820 2016

利用CAM估计Panel Size

- CAM 利用特征图像的加权求和来获取激活地图。
- 作者只考虑被判断为Positive的图像（包含Solar Panel）
- 直观上，CAM是将所有可能包含Panel的特征进行融合

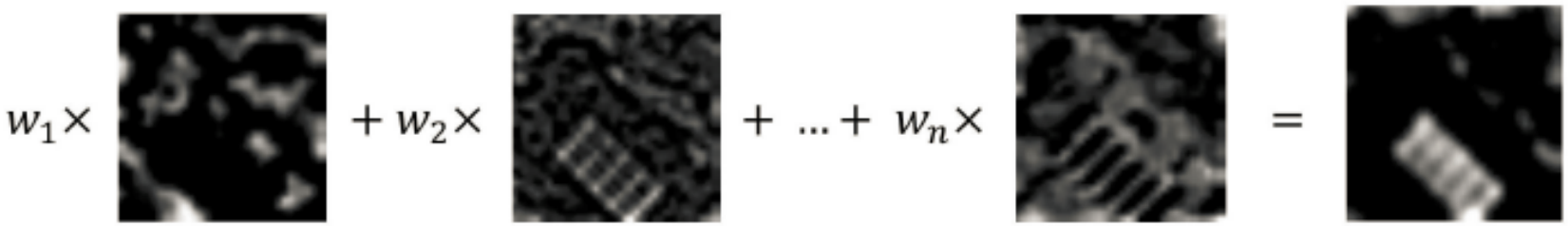


Figure S2: Class Activation Map (CAM)

(directly applying CAM to the retrained Inception-v3 model)

CAM + greedy

- **CAM 低估了Panel面积**: 只有Panel最显著的位置才会被激活

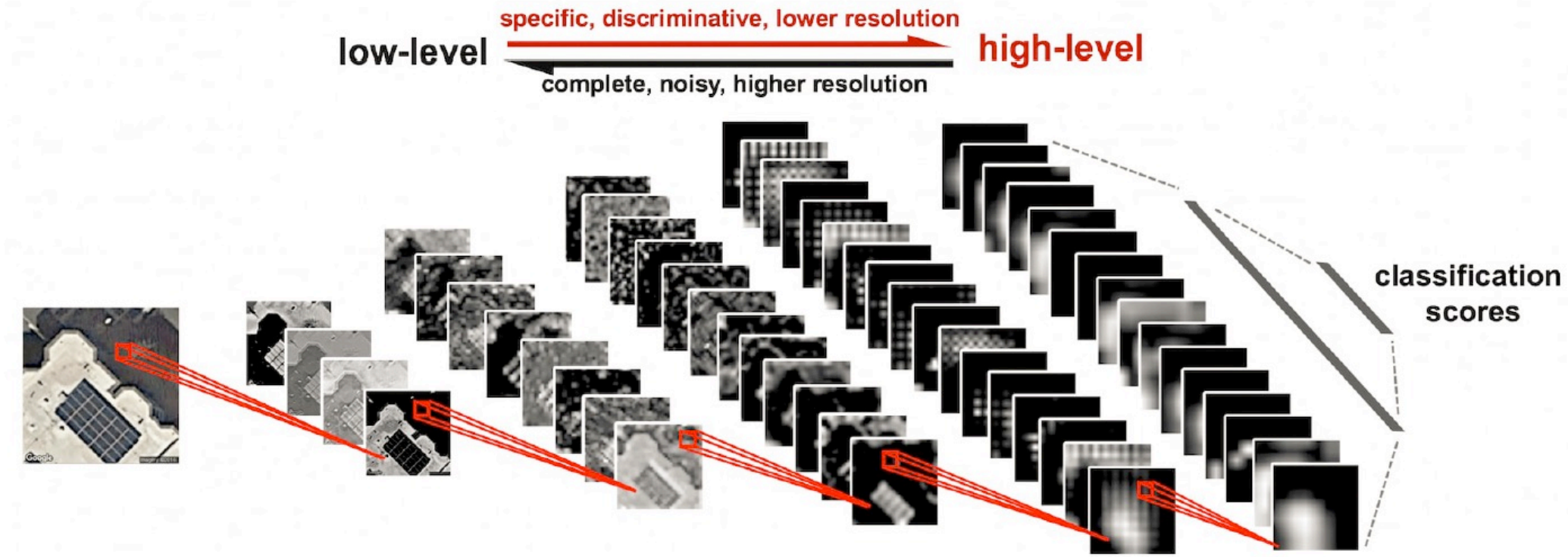


Figure S3: The feature hierarchy of CNN

From raw image input to class score output, different layers extract features of different levels. The feature maps at low-level hierarchy are complete, high-resolution but noisy, while the feature maps at high-level hierarchy are specific, discriminative but low-resolution.

CAM + greedy

- 另一种迁移思路：从底层的CNN Layer中提取基础特征，生成完整又具备判别力的CAM

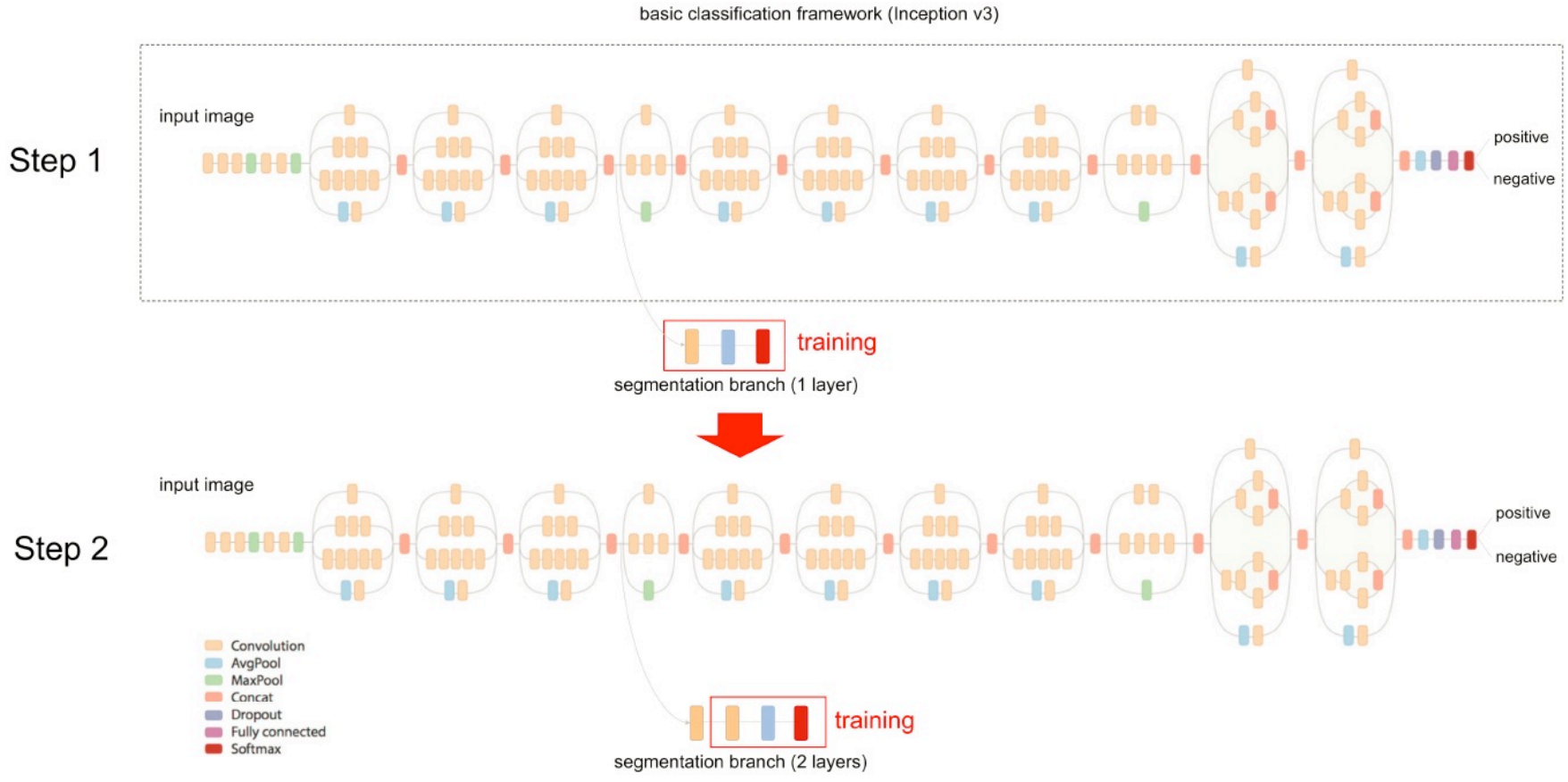


Figure S4: Inception-v3 framework and greedy layer-wise training

CAM + greedy

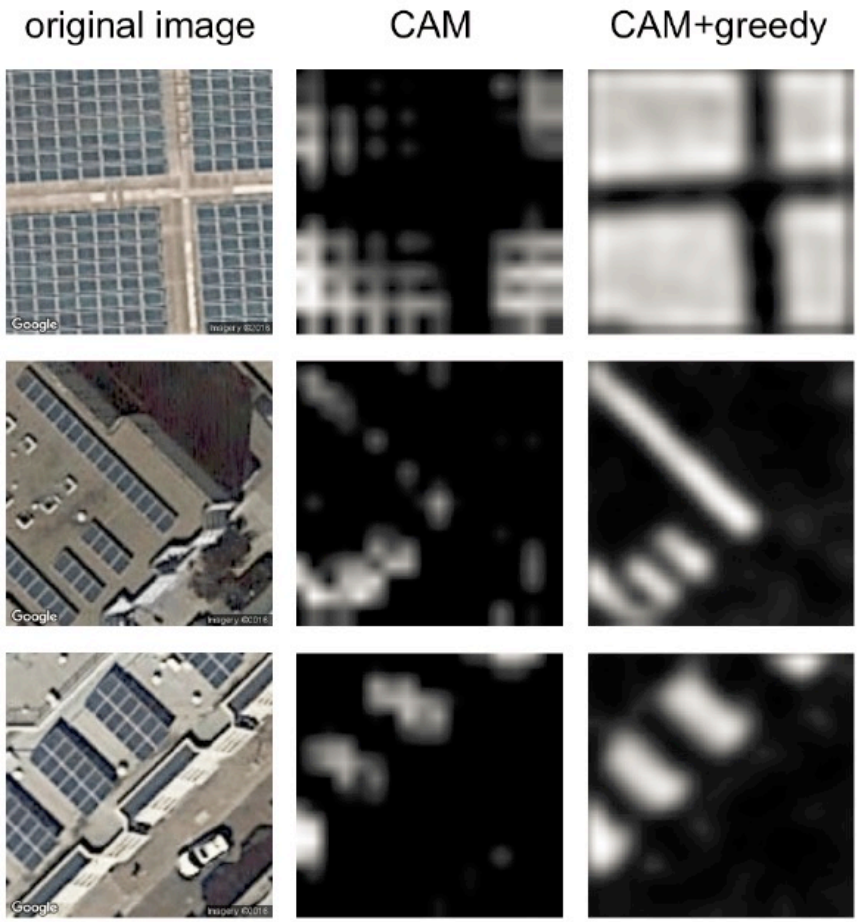


Figure S5: Benefit of greedy layer-wise training

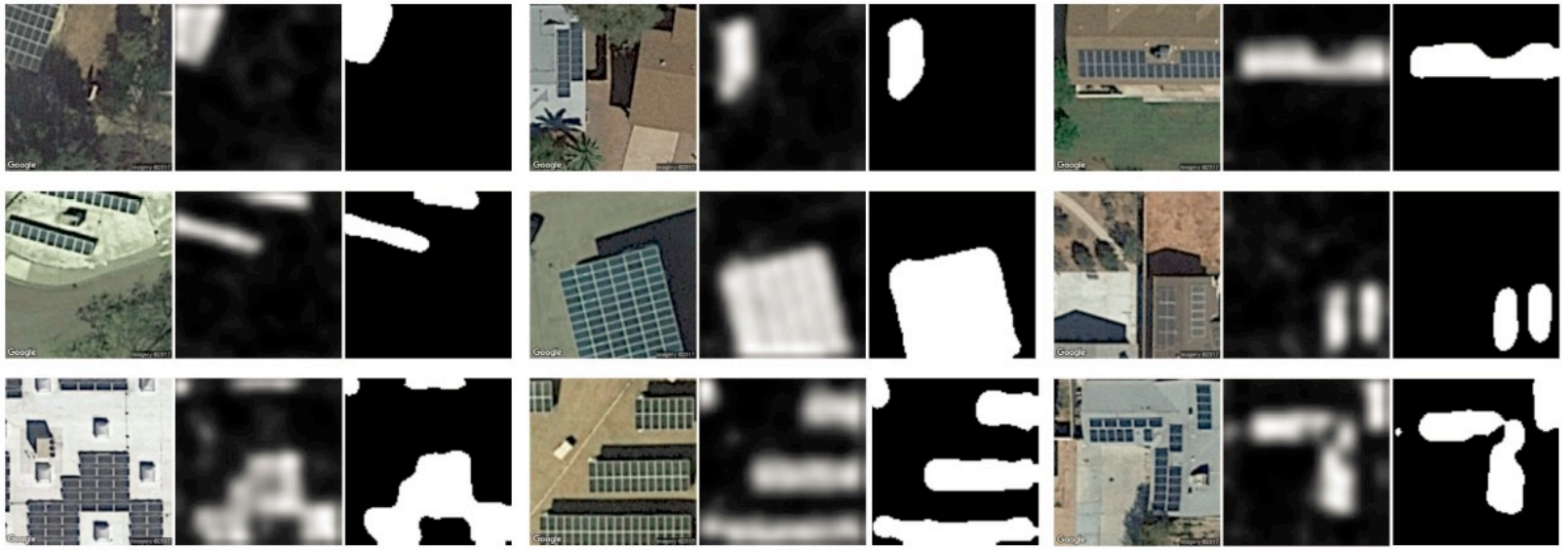


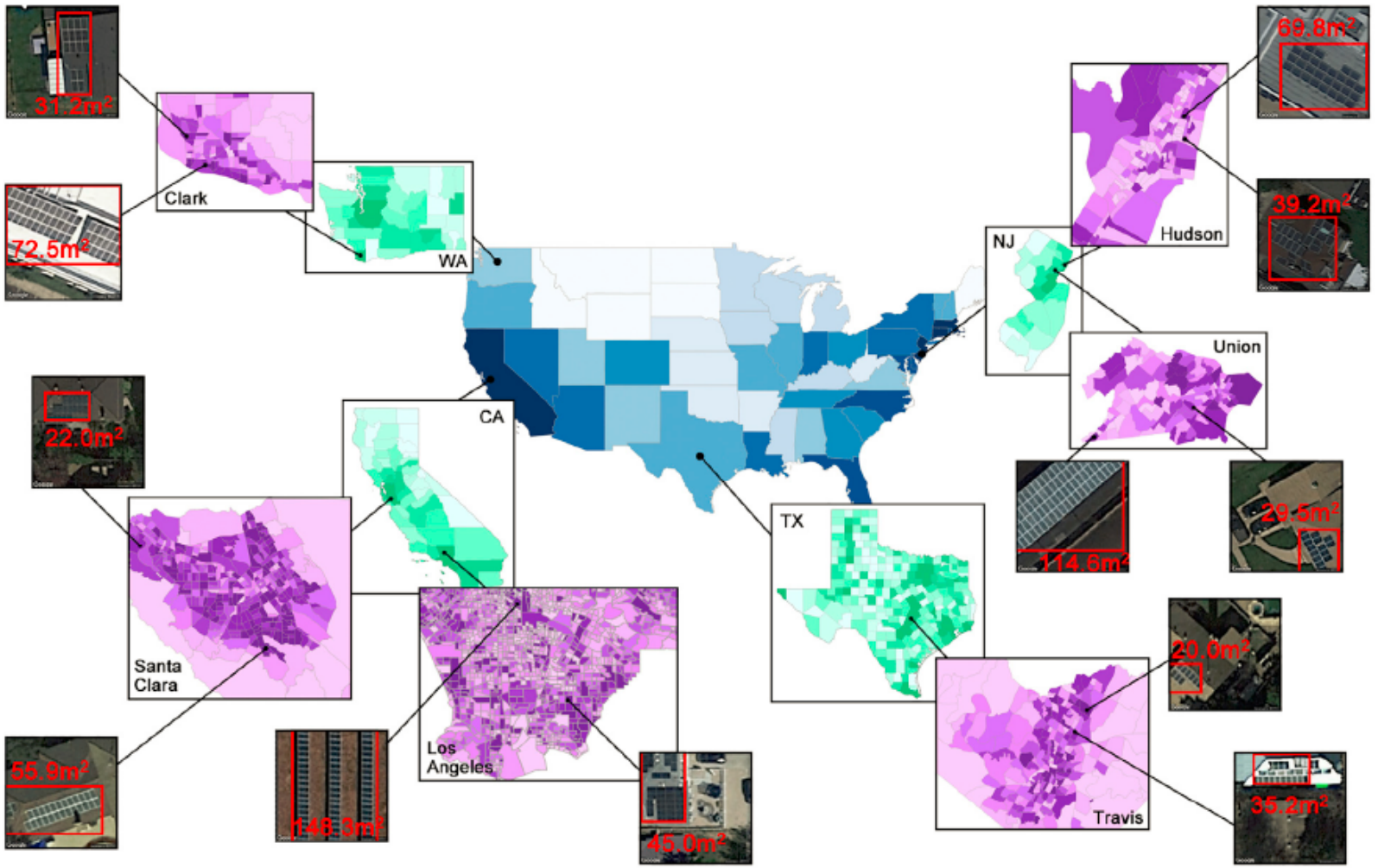
Figure S6: More examples of CAMs and the segmentation results in test set

Note: As a large-scale solar panel can be split to several image tiles, we use a simple recursive algorithm to find adjacent pieces of solar panels and merge them to form a solar system.

Results

Solar Resource Density Map

Solar Panel Area per Unit Area [m^2/$mile^2$]) at State, County, and Census Tract Levels



state level	county level				census tract level					
contiguous U.S.	CA	TX	NJ	WA	Santa Clara	Los Angeles	Travis	Union	Hudson	Clark
<1.34	<0.69	<0.03	<132	<0.11	< 255	< 618	< 70	< 323	< 11	< 0.84
1.34 - 3.45	0.69 - 6.53	0.03 - 0.15	132 -507	0.11 - 0.27	255 - 921	618 - 1181	70 - 205	323 - 598	11 - 327	0.84 - 22.53
3.45 - 4.34	6.53 - 21.18	0.15 - 0.27	507 - 729	0.27 - 0.51	921 - 1526	1181 - 1579	205 - 301	598 - 1092	327 - 743	22.53 - 49.59
4.34 - 7.92	21.18 - 90.44	0.27 - 0.46	729 - 1194	0.51 - 1.33	1526 - 2152	1579 - 2027	301 - 435	1092 - 1499	743 - 1460	49.59 - 79.60
7.92 - 14.24	90.44 - 165.9	0.46 - 1.03	1194 - 1325	1.33 - 1.90	2152 - 2864	2027 - 2531	435 - 647	1499 - 2022	1460 - 3331	79.60 - 95.80
14.24 - 24.12	165.9 - 281.9	1.03 - 1.71	1325 - 1814	1.90 - 5.95	2864 - 3753	2531 - 3176	647 - 842	2022 - 2942	3331 - 6598	95.80 - 131.59
24.12 - 76.74	281.9 - 446.2	1.71 - 3.98	1814 - 3027	5.95 - 11.68	3753 - 5071	3176 - 4079	842 - 1319	2942 - 4398	6598 - 9971	131.59 - 165.94
76.74 - 224.1	446.2 - 1088.1	3.98 - 11.95	3027 - 4148	11.68 - 24.20	5071 - 7490	4079 - 6366	1319 - 2197	4398 - 7960	9971 - 38058	165.94 - 281.21
> 224.1	> 1088.1	> 11.95	> 4148	> 24.20	> 7490	> 6366	> 2197	> 7960	> 38058	> 281.21

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

人工智能研究院
Artificial Intelligence Institute

谢谢！

饮水思源 爱国荣校

http://humnetlab.berkeley.edu/~yxu/