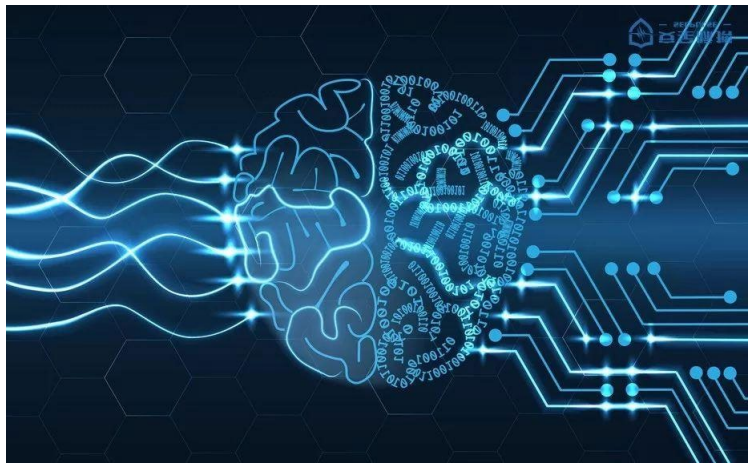


机器学习

Machine Learning

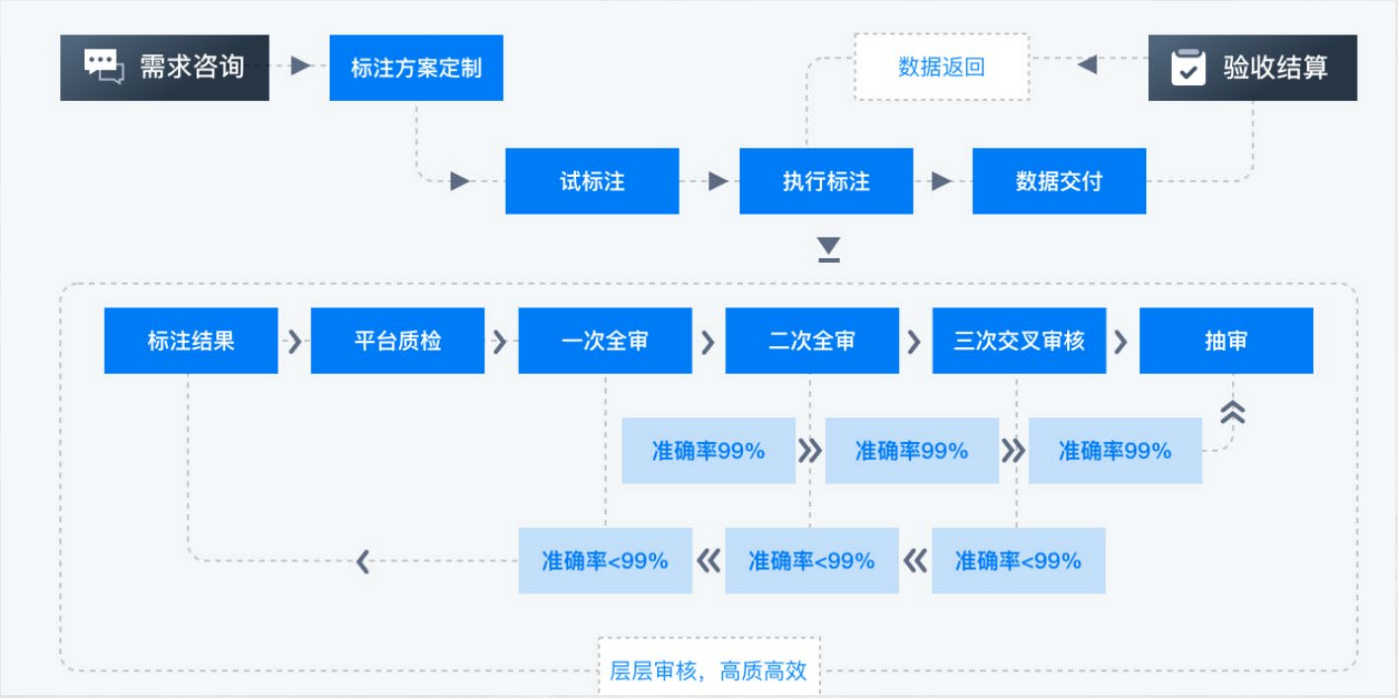
第十一章、迁移学习

目录
CONTENTS

为人师表

1.迁移学习背景

有多少人工，就有多少智能!



数据的标注是一个耗时且昂贵的操作!

1.迁移学习背景

TABLE 3: Compare AlexNet training with different approaches.

	Batch Size	Processor	GPU Interconnect	Time	Top-1 Accuracy
You et al. [41]	512	DGX-1 station	NVLink	6 hours 10 mins	58.8%
You et al. [41]	32K	CPU x 1024	-	11 mins	58.6%
Jia et al. [18]	64K	Pascal GPU x 512	100 Gbps	5 mins	58.8%
Jia et al. [18]	64K	Pascal GPU x 1024	100 Gbps	4 mins	58.7%
This Work (DenseCommu)	64K	Volta GPU x 512	56 Gbps	2.6 mins	58.7%
This Work (SparseCommu)	64K	Volta GPU x 512	56 Gbps	1.5 mins	58.2%

TABLE 4: Compare ResNet-50 training with different approaches.

	Batch Size	Processor	GPU Interconnect	Time	Top-1 Accuracy
Goyal et al. [3]	8K	Pascal GPU x 256	56 Gbps	1 hour	76.3%
Smith et al. [42]	16K	Full TPU Pod	-	30 mins	76.1%
Codreanu et al. [43]	32K	KNL x 1024	-	42 mins	75.3%
You et al. [41]	32K	KNL x 2048	-	20 mins	75.4%
Akiba et al. [38]	32K	Pascal GPU x 1024	56 Gbps	15 mins	74.9%
Jia et al. [18]	64K	Pascal GPU x 1024	100 Gbps	8.7 mins	76.2%
Jia et al. [18]	64K	Pascal GPU x 2048	100 Gbps	6.6 mins	75.8%
Mikami et al. [44]	68K	Volta GPU x 2176	200 Gbps	3.7 mins	75.0%
This Work (DenseCommu)	64K	Volta GPU x 512	56 Gbps	7.3 mins	75.3%

Sun P, Feng W, Han R, et al. Optimizing network performance for distributed dnn training on gpu cluster Imagenet/alexnet training in 1.5 minutes[J]. arXiv preprint arXiv:1902.06855, 2019.

Solver	batch size	steps	F1 score on dev set	TPUs	Time
Baseline	512	1000k	90.395	16	81.4h
LAMB	512	1000k	91.752	16	82.8h
LAMB	1k	500k	91.761	32	43.2h
LAMB	2k	250k	91.946	64	21.4h
LAMB	4k	125k	91.137	128	693.6m
LAMB	8k	62500	91.263	256	390.5m
LAMB	16k	31250	91.345	512	200.0m
LAMB	32k	15625	91.475	1024	101.2m
LAMB	64k/32k	8599	90.584	1024	76.19m

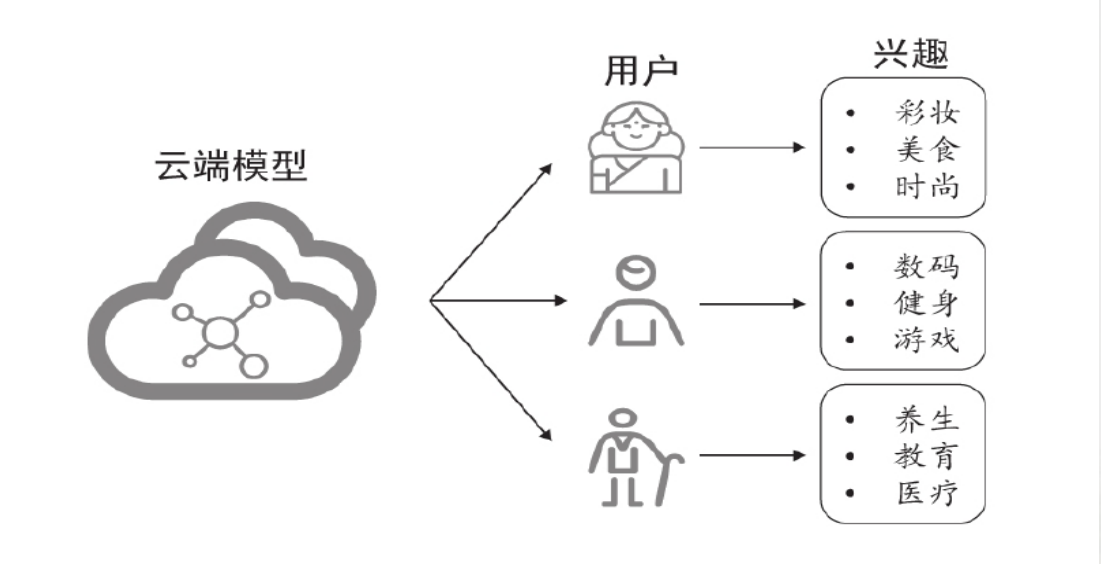
You Y, Li J, Reddi S, et al. Large batch optimization for deep learning: Training bert in 76 minutes[J]. arXiv preprint arXiv:1904.00962, 2019.

大数据需要强计算能力的设备来进行存储和计算。近些年深度学习的数据量和模型大小在迅速增加，高昂的计算成本是普通个人用户无法长期负担的

迁移学习提供了一种基于大数据“预训练”的模型在自己的特定数据集上进行“微调”的技术，普通人也能利用这些数据和模型

1.迁移学习背景

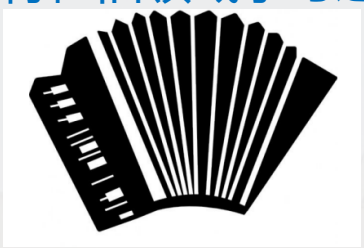
普适化模型→个性化需求

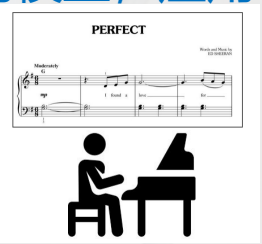


利用迁移学习的思想，我们可以将那些大公司在大数据上训练好的模型，迁移到我们的任务中，针对于我们的任务进行微调，从而也可以拥有在大数据上训练好的模型。更进一步，可以将这些模型针对具体任务进行自适应更新，取得更好的效果。

2.迁移学习概念

- 概念：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。
  - **球类运动相似：乒乓球高手可很快学会打网球**
  - **乐器相通：手风琴和钢琴也有相通的地方**

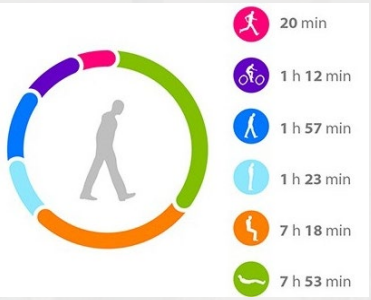




应用领域

语义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

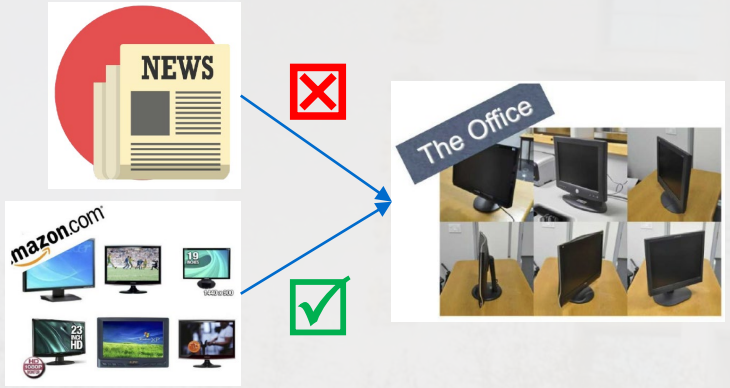
不同用户、不同接口、不同情境下的人机交互

不同场景、不同时间、不同设备的室内定位

2.迁移学习概念

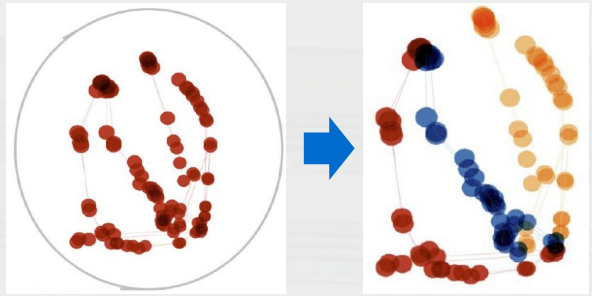
迁移学习的三个基本问题

- **何时迁移**



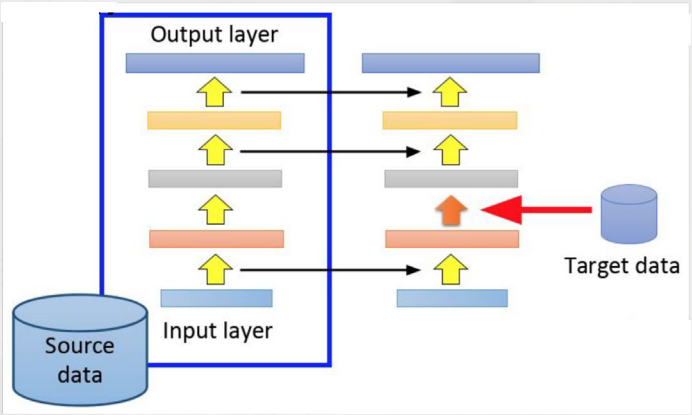
原问题和目标问题具有相关性

- **何处迁移**



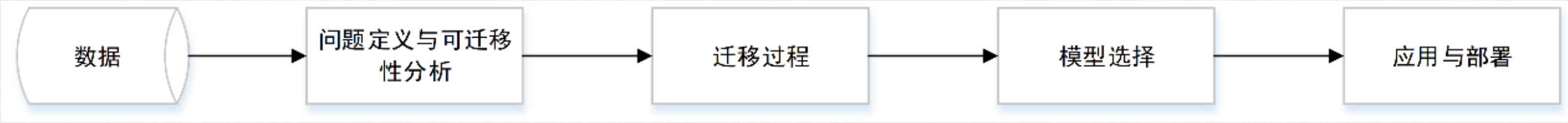
要迁移的知识往往来源于相似的模式或特征

- **如何迁移**



建立原问题和目标问题的关联

迁移学习的一般过程



2.迁移学习概念

迁移学习的重要概念：

领域：是机器学习的主体，主要由两个部分构成：数据和生成这些数据的概率分布。

通常用D表示一个领域，可被表示为$D = \{\mathcal{X}, \mathcal{Y}, P(x, y)\}$。其中$\mathcal{X}, \mathcal{Y}$,分别为数据所处的特征空间和标签空间，P为数据服从的概率分布。

2.迁移学习概念

领域$D = \{X, Y, P(x, y)\}$

- 域不同的三种情形：

➢ **特征空间不同，即$X_s \neq X_t$：**

例如，当源域为RGB彩色图像、目标域为黑白二值图像

源域样本：

目标域样本：

➢ **标签空间不同，即$Y_s \neq Y_t$：**

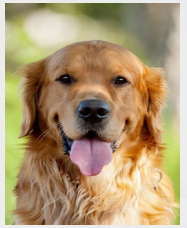
例如，在分类问题中，源域和目标域的类别不完全相同

源域只有猫和狗两种类别：

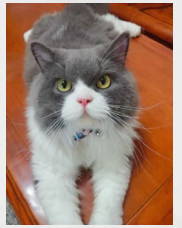
猫

狗

目标域除了猫和狗外还包含品种



金毛犬



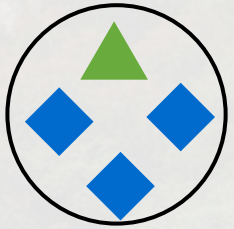
英国短毛猫

➢ **概率分布不同，即$P_s(x, y) \neq P_t(x, y)$：　(重点介绍)**

特指即使两个领域的特征空间和类别空间都相同，其联合概率分布也会存在不匹配的问题

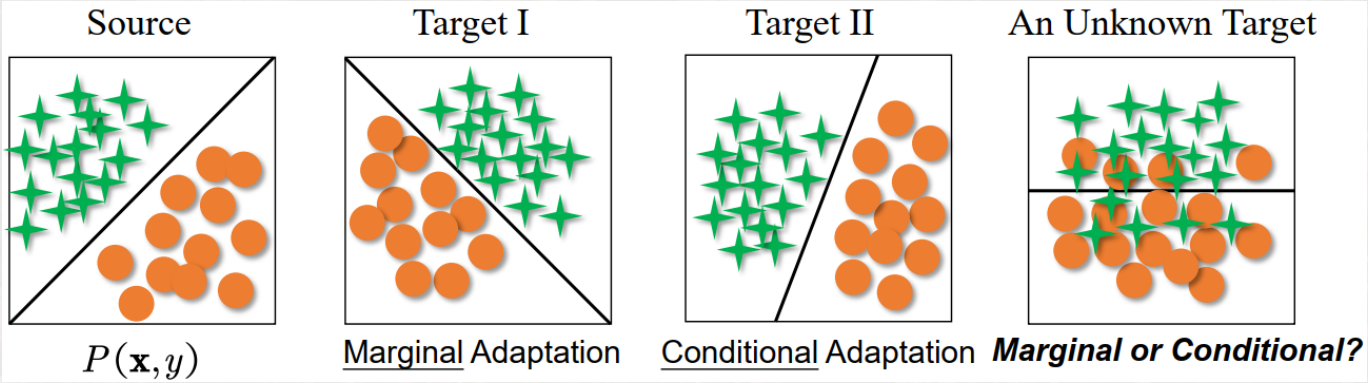
源域：

目标域：



2.迁移学习概念

- $P_s(x,y) \neq P_t(x,y)$ 说明了什么?
  - 贝叶斯定理：$P(x,y) = P(y|x)\,P(x)$
  - 若 $P_s(y|x) = P_t(y|x)$，且边缘分布不同：
    - $P_s(x) \neq P_t(x)$，采用边缘分布自适应方法：
  - 若$P_s(x) = P_t(x)$，且条件分布不同：
    - $P_s(y|x) \neq P_t(y|x)$，采用条件分布分布自适应方法：



(Yu et al. Transfer learning with dynamic adversarial adaptation network. ICDM 2019.)

2.迁移学习概念

迁移学习的核心思路：

- **找到源域和目标域之间的差异性，并将整体工作归结为两方面**
  - 一是度量两个领域的相似性，不仅定性地告诉我们它们是否相似，更定量地给出相似程度
  - 二是以度量为准则，通过我们所要采用的学习手段，增大两个领域之间的相似性，完成迁移学习。

3.迁移学习方法

迁移学习的统一数学表征：

- **从机器学习的最小化经验风险准则（ERM）出发**

$f^* = \arg\min_f \frac{1}{m}\sum_{i=1}^{m} L(f(x_i), y_i)$，$L(\cdot,\cdot)$是损失函数

- **使用一些方法来减少域差异，得到迁移学习的统一表征**

$$f^* = \arg\min_{f\in\mathcal{H}} \frac{1}{N_s}\sum_{i=1}^{N_s} L(v_i f(x_i), y_i) + \lambda R\big(T(D_s), T(D_t)\big)$$

➢ **其中：**

- $v_i \in \mathbb{R}^{N_s}$，$v_i \in [0,1]$。N_s为源域样本的数量
- T为作用于源域和目标域上的特征变换函数
- $R\big(T(D_s), T(D_t)\big)$为迁移正则化项，用于定量衡量源域和目标域之间的分布距离

3.迁移学习方法

$$f^* = \arg\min_{f\in\mathcal{H}} \frac{1}{N_s}\sum_{i=1}^{N_s} L(v_i f(x_i), y_i) + \lambda R\big(T(D_s), T(D_t)\big)$$

迁移学习方法分类：

- 三种基本的迁移学习类型

基于样本的迁移学习

- 目标为根据源域和目标域的相似度来学习源域样本的权重 v_i

基于特征的迁移学习

- 目标为学习一个特征变换 T 来减小迁移正则化项 $R(\cdot,\cdot)$

基于模型的迁移学习

- 目标为学习如何将源域的判别函数$f(\cdot,\cdot)$对目标域数据进行正则化和微调

3.迁移学习方法

基于样本的迁移方法：

- **核心思想：增大源域样本中与目标域样本差异度小的比重**

根据 v_i 取值范围划分

- 样本选择法： $v_i \in \{0,1\}$
- 权重自适应法： $v_i \in [0,1]$

3.迁移学习方法

基于特征的迁移方法：

- **将源域和目标域的差异性视为两个域内样本数据的概率分布的差异性进行研究，根据采用显式或隐式的距离度量分为两类**

  - 显式度量：由预定义好的距离公式产生的度量，具有特定形式，常用的距离度量有欧式距离、闵可夫斯基距离、马氏距离、余弦相似度等

  - 隐式度量：并非预先定义好的，而是可以在数据中动态学习的、更适合数据分布的度量

3.迁移学习方法

基于特征的迁移方法：

- **迁移学习中使用最广泛的距离度量之一 —— 最大均值差异（Maximum Mean Discrepancy，MMD）**

$$\mathrm{MMD}^2(A,B)=\left\|\sum_{i=1}^{n1}\phi(a_i)-\sum_{j=1}^{n2}\phi(b_j)\right\|^2$$

A, B: 两个样本集合　a_i: 集合A中的样本　b_j: 集合B中的样本

ϕ: 特征映射函数　$\phi(b_j)$: 样本b_j的特征

$\phi(a_i)$: 样本a_i的特征

直观解释：计算两个集合中样本的特征变换后分布之前的距离

3.迁移学习方法

基于特征的迁移方法：

➢ 基于特征的迁移方法优化目标：

$$f^* = \arg\min_{f\in\mathcal{H}} \frac{1}{N_s}\sum_{i=1}^{N_s} L(f(x_i), y_i) + \lambda R\big(T(D_s), T(D_t)\big)$$ 这里$R(\cdot,\cdot)$即为选择的距离度量MMD

➢ 迁移学习中的概率分布差异方法与假设：

方法	假设	优化目标
边缘分布自适应	$P_s(y\|x) = P_t(y\|x)$	$\min MMD(P_s(x), P_t(x), f)$
条件分布自适应	$P_s(x) = P_t(x)$	$\min MMD(P_s(y\|x), P_t(y\|x), f)$
联合分布自适应	$P_s(x,y) \neq P_t(x,y)$	$\min MMD(P_s(x), P_t(x), f) + MMD(P_s(y\|x), P_t(y\|x), f)$
动态分布自适应	$P_s(x,y) \neq P_t(x,y)$	$\mathrm{in}(1-\mu)\, MMD(P_s(x), P_t(x), f) + \mu MMD(P_s(y\|x), P_t(y\|x), f$

（f 为特征变换函数）

第四种为最一般的形式，前三种为第四种取特值得退化形式！

3.迁移学习方法

动态分布自适应问题：

- **分布自适应因子μ的计算方法**

➢ 随机猜测法：任意从[0，1]区间内选择一个μ的值，然后进行动态迁移学习。重复此过程t次，则随机猜测法最终的迁移结果为t次迁移学习结果的均值。

➢ 最大最小平均法：在[0,1]区间内，以固定步长取等距的μ，多样重复学习多次，最终的迁移结果为多次迁移学习的均值。

有没有一种精确计算且具有理论依据的计算方法?

3.迁移学习方法

□动态分布自适应问题：

- **分布自适应因子μ的计算方法**

利用领域的整体和局部性质来定量计算 ⟶ 采用$\mathcal{A}-distance$作为基本度量方式

$$d_A(D_s, D_t) = 2\left(1 - 2\epsilon(h)\right)$$

$\epsilon(h)$：线性分类器区分源域D_s和目标域D_t的误差

$$\hat{\mu} = 1 - \frac{d_M}{d_M + \sum_{c=1}^{C} d_c}$$

d_M为边缘分布距离 $d_M = d_A(D_s, D_t)$

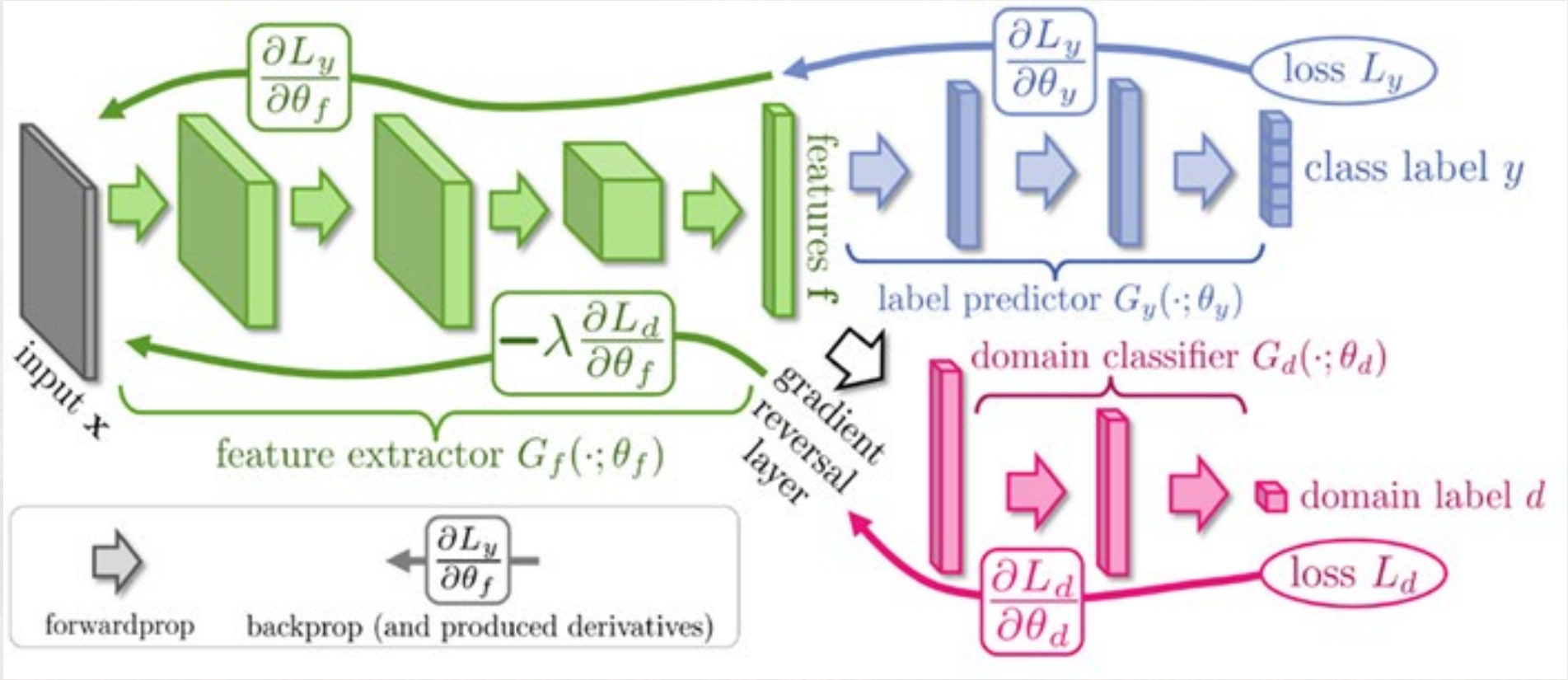
d_c为表示对应于类别c的条件分布距离 $d_c = d_A\left(D_s^{(c)}, D_t^{(c)}\right)$

Wang J, Feng W, Chen Y, et al. Visual domain adaptation with manifold embedded distribution alignment[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 402-410.

3.迁移学习方法

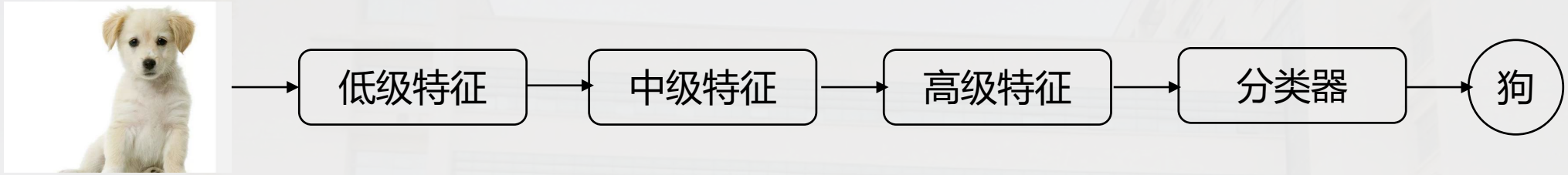
基于特征的迁移方法：

- **隐式度量——例如基于生成对抗网络的迁移学习**



3.迁移学习方法

基于模型的迁移方法：

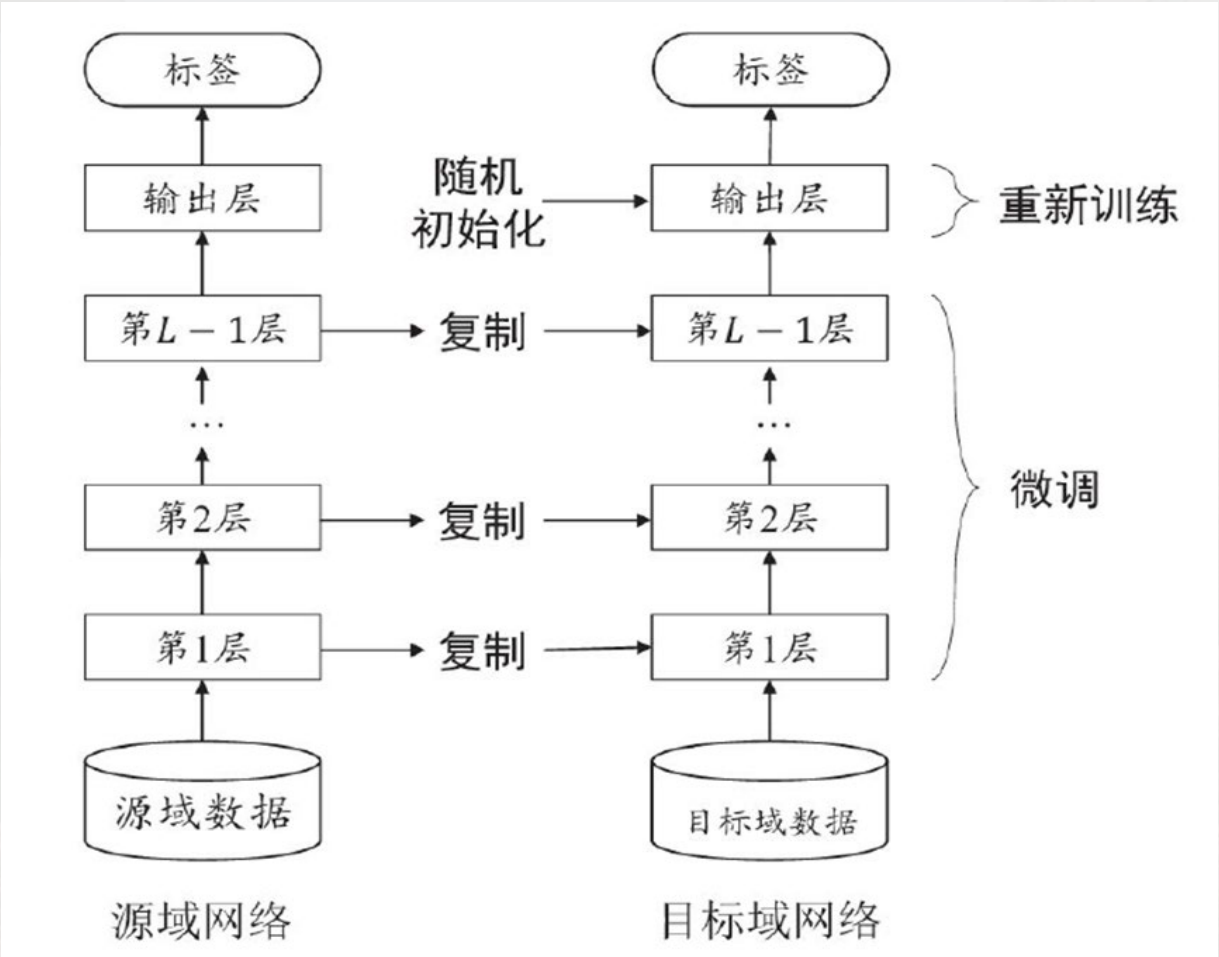


深度网络的可迁移性

- 网络浅层负责学习通用特征（General features，在上图中则为动物边边角角的低级特征）
- 深层则负责学习与任务相关的特殊特征（Specific features，在上图中则为腿、脸等中高级特征）
- 随着层次的加深，网络渐渐从通用特征过渡到特殊特征的学习表征
- 在大数据集上训练得到一个具有强泛化能力的模型（预训练模型）。对学习通用特征的浅层直接保留并微调，对学习特殊特征的深层进行改造，这就是“预训练–微调”

3.迁移学习方法

□基于模型的迁移方法：



3.迁移学习方法

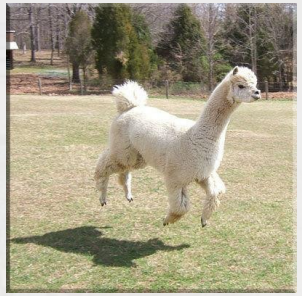
➢零样本学习

- Source data: (x^s, y^s) → Training data
- Target data: (x^t) → Testing data

不同任务

x^s:   ……

y^s: cat dog ……

x^t:  

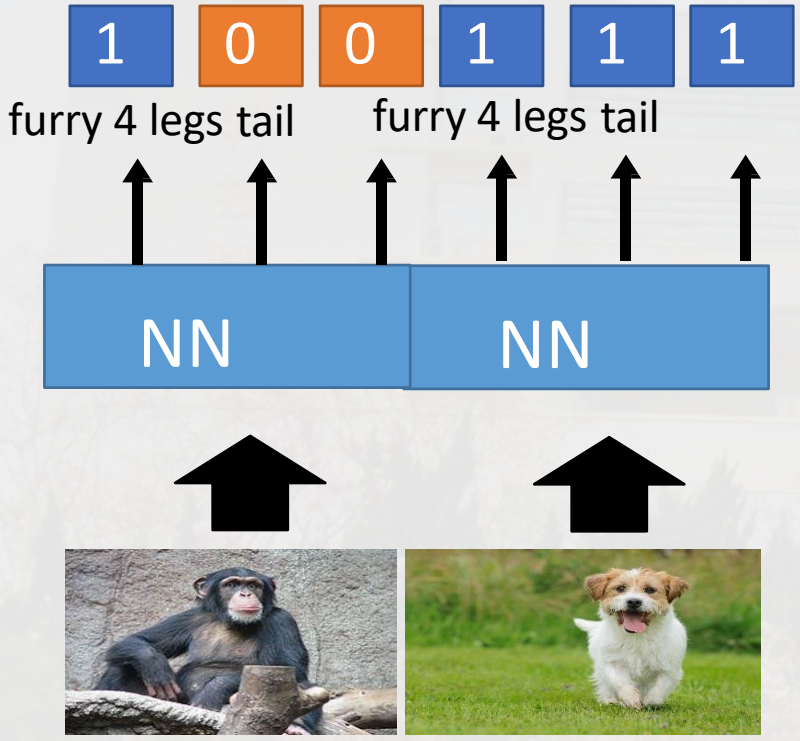
在语音识别中，source(training) data不可能出现所有的词汇。

我们如何在语音识别中解决这个问题?

3.迁移学习方法

➢零样本学习-通过属性来表示每个类

Training



Database

attributes

	furry	4 legs	tail	…
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
…				

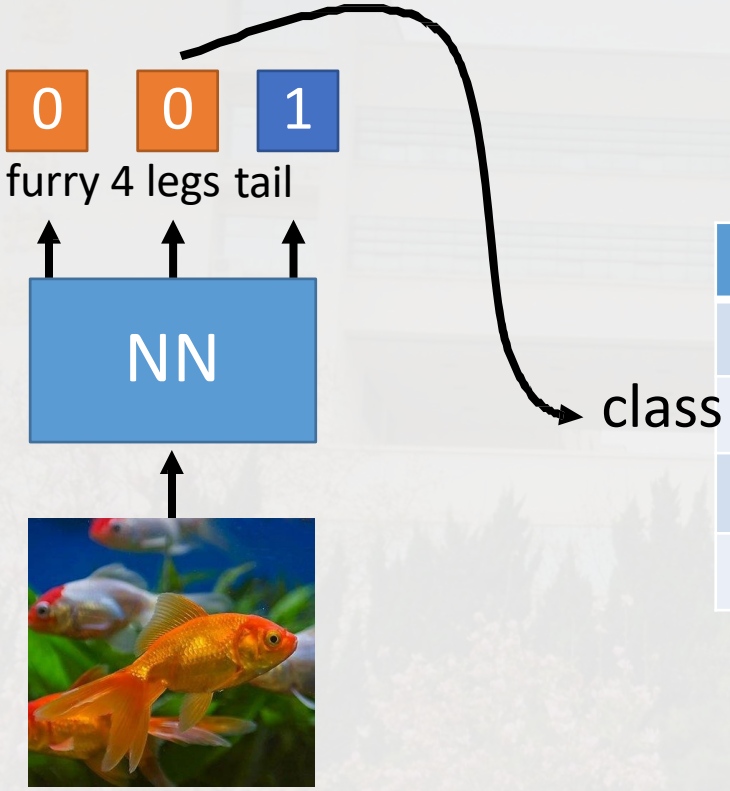
class

sufficient attributes for one to one mapping

3.迁移学习方法

➢零样本学习-通过属性来表示每个类

Testing



Database

attributes

	furry	4 legs	tail	…
Dog	O	O	O	
Fish	X	X	O	
Chimp	O	X	X	
…				

sufficient attributes for one to one mapping

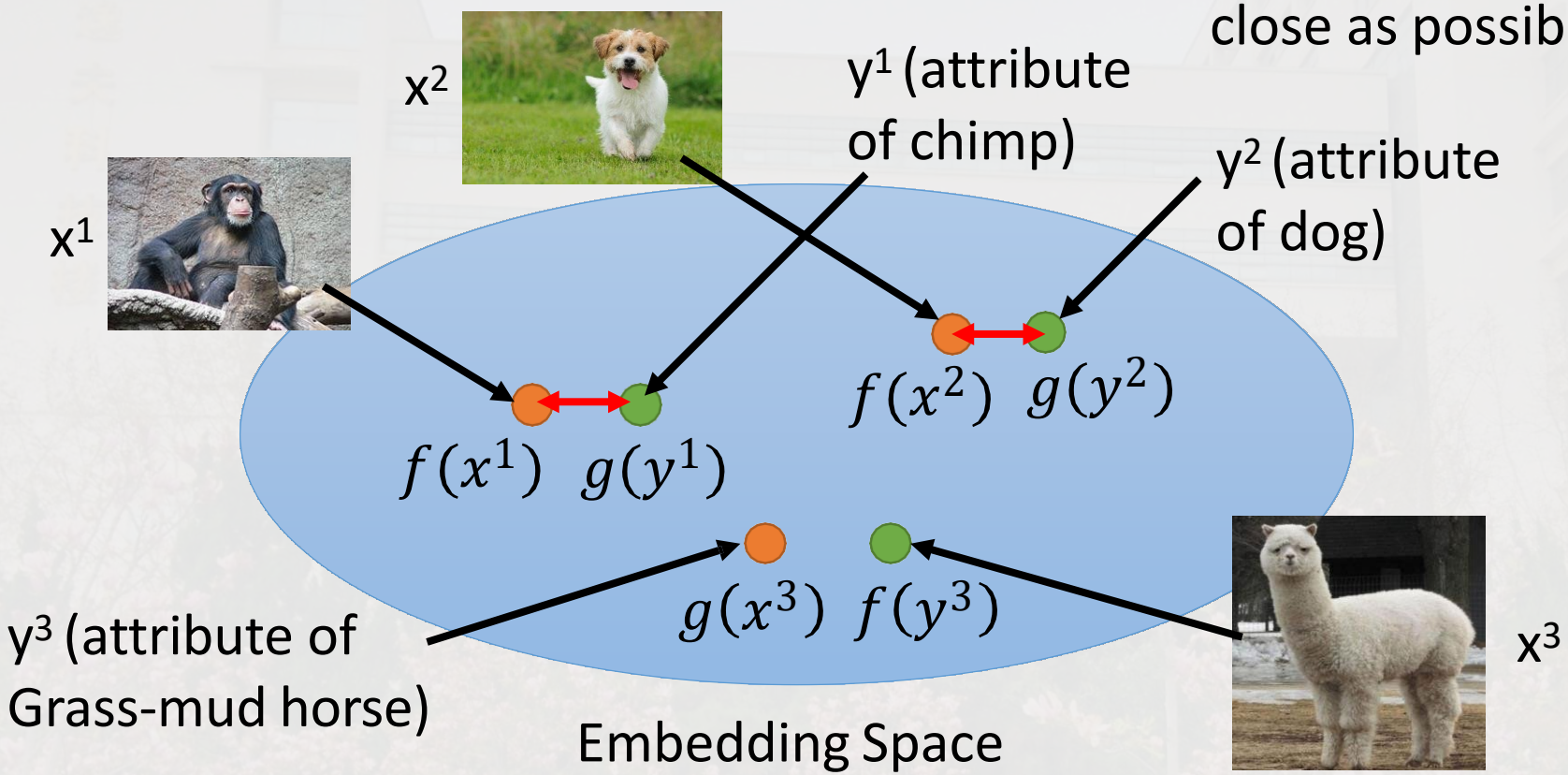
3.迁移学习方法

➢零样本学习-属性嵌入

$f(*)$ and $g(*)$ can be NN.

Training target:

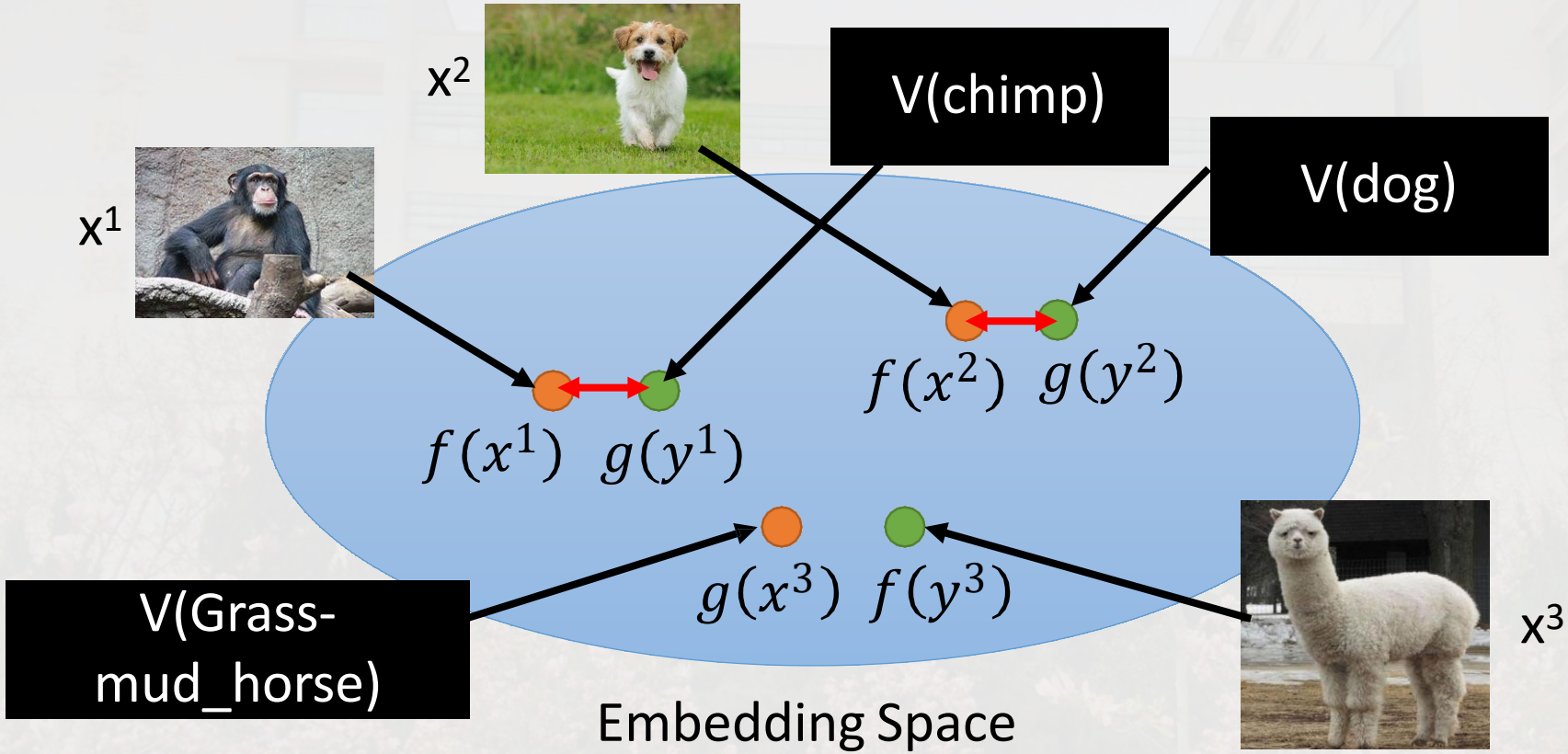
$f(x^n)$ and $g(y^n)$ as close as possible



3.迁移学习方法

- 零样本学习-属性嵌入+词汇嵌入

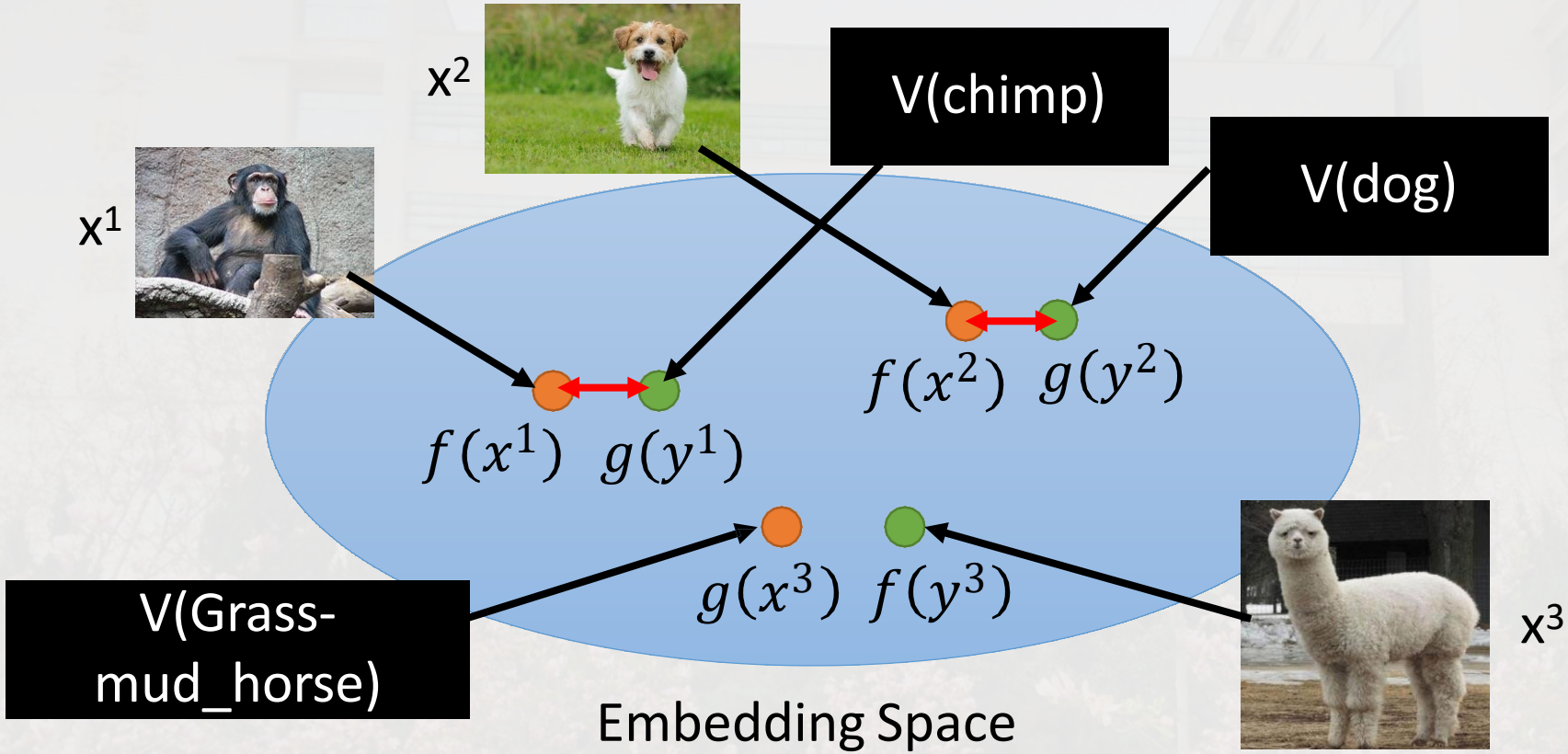
如果我们没有数据库会怎样



3.迁移学习方法

- 零样本学习-属性嵌入+词汇嵌入

如果我们没有数据库会怎样



3.迁移学习方法

➢零样本学习

$$f^*, g^* = arg \min_{f,g} \sum_n ||f(x^n) - g(y^n)||$$ Problem?

$$f^*, g^* = arg \min_{f,g} \sum_n \max(0, k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m)$$

Margin you defined

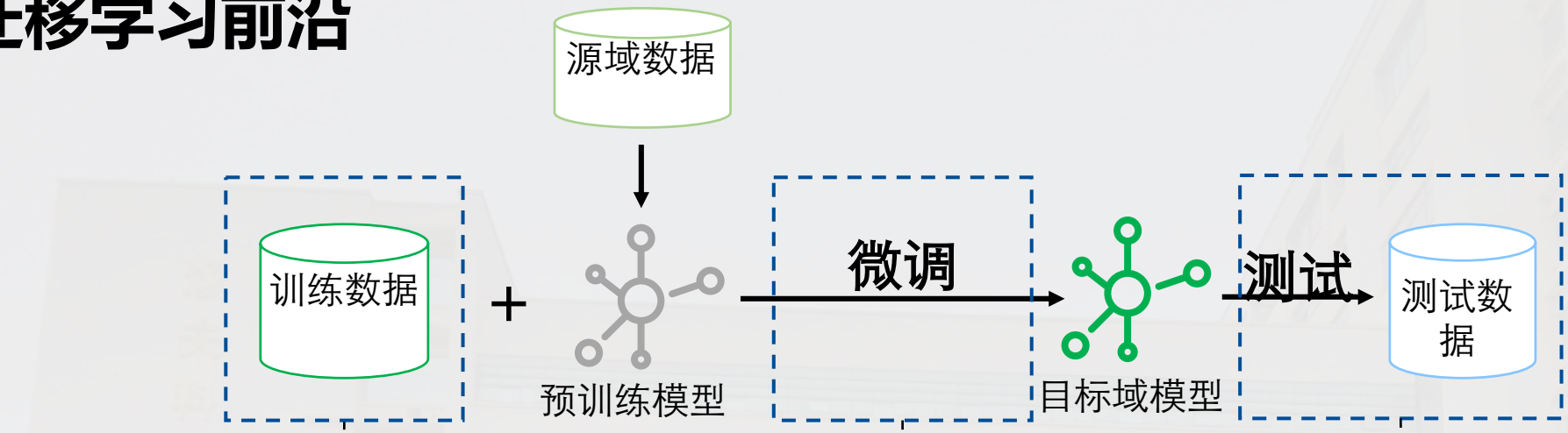
Zero loss: $k - f(x^n) \cdot g(y^n) + \max_{m \neq n} f(x^n) \cdot g(y^m) < 0$

$$f(x^n) \cdot g(y^n) - \max_{m \neq n} f(x^n) \cdot g(y^m) > k$$

$f(x^n)$和$g(y^n)$越近越好

$f(x^n)$和$g(y^n)$要足够远

4.迁移学习前沿



Low-resource learning
(低资源学习)

Domain adaptation
(领域自适应)

Safe transfer
(安全迁移)

Domain generalization
(领域泛化)

4.迁移学习前沿

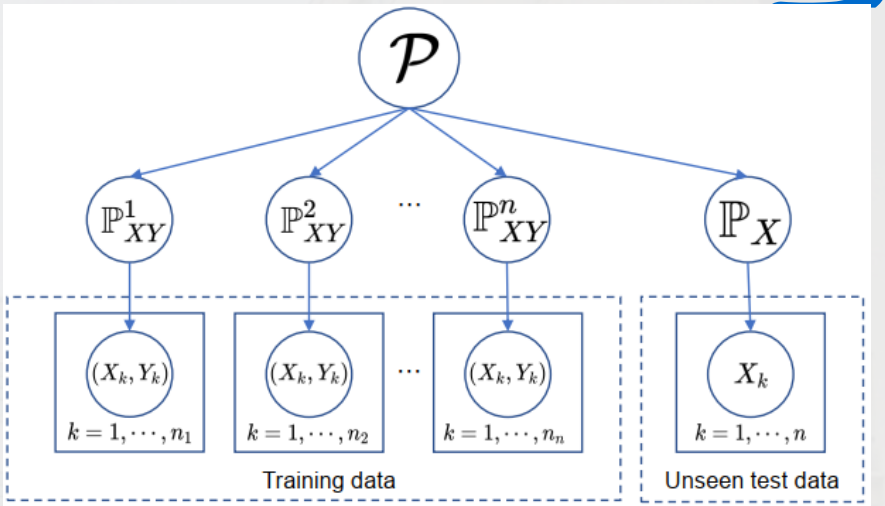
低资源学习：

- 研究背景
  - 依赖少量有标签样本学习泛化性强的模型
- 研究问题
  - 如何保证网络学习知识从有标签的数据运用于无标签的数据时能保持相同效果?
  - 迁移准则: 固定阈值(Google提出的 FixMatch[NeurIPS’20])
- 研究挑战
  - 预定义阈值能否解决半监督问题?
  - 能否为半监督学习设计更好的阈值计算方式?

4.迁移学习前沿

领域泛化：

- 研究背景
  - 对多种分布进行训练来学习在未知域上的泛化模型
- 研究问题
  - 解决数据特性随时间变化的问题
  - 解决数据分布是动态变化的问题
- 研究挑战
  - 如何捕捉获取数据分布的动态变化?
  - 如何对时间序列中的数据分布进行量化?
- Wang et al. Generalizing to unseen domains: a survey on domain generalization. IJCAI 2021 survey track. https://arxiv.org/abs/2103.03097



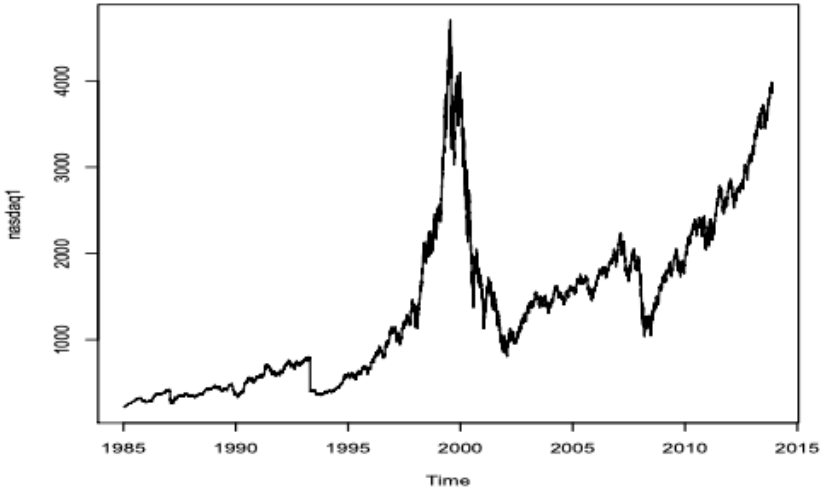
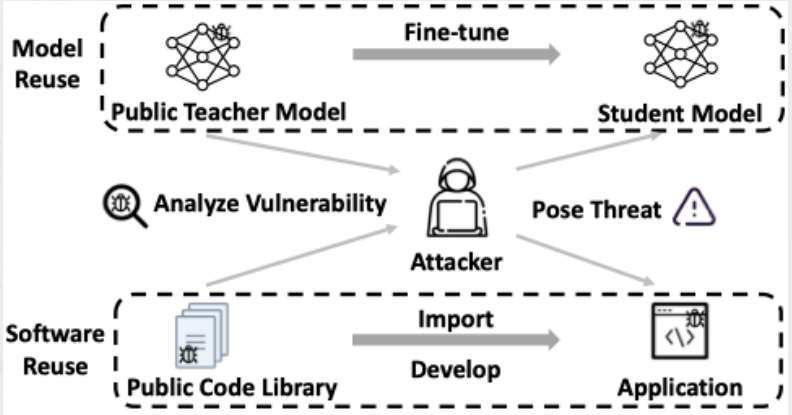


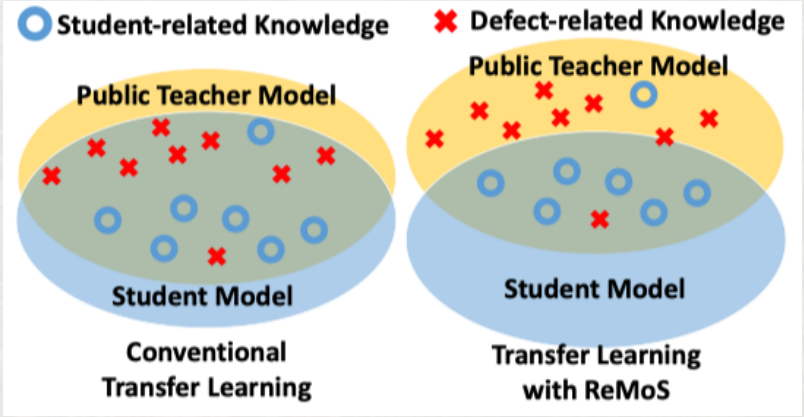
Figure 1.3: Plot of daily closing price of Nasdaq 1985-2014

4.迁移学习前沿

安全迁移：

- 研究动机
  - 软件重复使用在软件工程中非常流行→预训练模型在机器学习中的广泛应用
  - 恶意程序/代码会造成损害→微调后的模型可能会预训练模型中继承漏洞
  - 缺陷很容易从预训练模型传播给微调后模型，遗传率在52.58% ~ 97.85%之间
- 研究问题
  - 减少微调后模型从预训练模型继承缺陷
  - 同时还要保留它的性能优势
- 挑战
  - 对模型的攻击方式位置
  - 深度神经网络模型多样化且缺乏可解释性





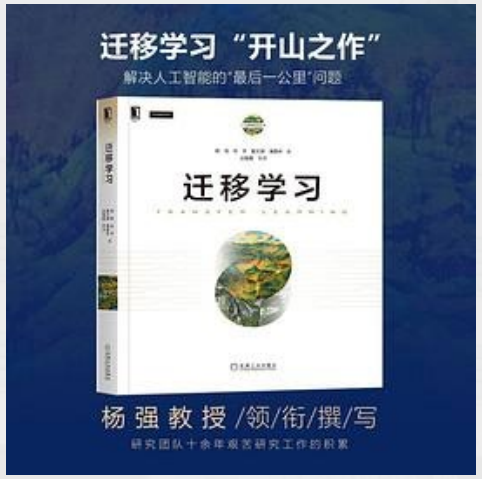
5.相关读物推荐

□迁移学习经典书籍：

书名：迁移学习导论（第二版）

作者：王晋东，陈益强 著

出版社：电子工业出版社



书名：迁移学习

作者：杨强，张宇等 著

出版社：机械工业出版社

□迁移学习网课：http://transferlearning.xyz

□迁移学习开源代码库：https://github.com/jindongwang/transferlearning/tree/master/code

□迁移学习benchmark数据集：https://github.com/jindongwang/transferlearning/tree/master/data