表示学习

什么因素决定了一种表示比另一种表示更好呢?

- 一个好的表示可以使后续的学习任务更容易
- 选择什么表示通常取决于后续的学习任务

目标

- 利用第一个情景下的数据，提取那些在第二种情景中学习时或直接进行预测时可能有用的信息

核心思想

- 相同的表示可能在不同的场景之中都是有用的

贪心逐层无监督预训练

每一层使用无监督学习预训练

输入：前一层的输出

输出：数据的新的表示

贪心逐层无监督预训练

算法 15.1 贪心逐层无监督预训练的协定

给定如下：无监督特征学习算法 $\mathcal{L}$，$\mathcal{L}$ 使用训练集样本并返回编码器或特征函数 f。原始输入数据是 $\boldsymbol{X}$，每行一个样本，并且 $f^{(1)}(\boldsymbol{X})$ 是第一阶段编码器关于 $\boldsymbol{X}$ 的输出。在执行精调的情况下，我们使用学习者 $\mathcal{T}$，并使用初始函数 f，输入样本 $\boldsymbol{X}$（以及在监督精调情况下关联的目标 $\boldsymbol{Y}$），并返回细调好函数。阶段数为 m。

$f \leftarrow$ 恒等函数
$\tilde{\boldsymbol{X}} = \boldsymbol{X}$
for $k = 1, \ldots, m$ **do**
  $f^{(k)} = \mathcal{L}(\tilde{\boldsymbol{X}})$
  $f \leftarrow f^{(k)} \circ f$
  $\tilde{\boldsymbol{X}} \leftarrow f^{(k)}(\tilde{\boldsymbol{X}})$
end for
if *fine-tuning* ***then***
  $f \leftarrow \mathcal{T}(f, \boldsymbol{X}, \boldsymbol{Y})$
end if
Return f

贪心逐层无监督预训练

贪心

- 贪心算法：独立地优化解决方案的每一个部分

逐层

- 独立的解决方案就是网络的层

无监督

- 每一层用无监督表示学习算法训练

预训练

- 只是在联合训练算伐精调所有层之前的第一步

无监督预训练

无监督预训练结合了两种不同的想法：

- 深度神经网络对初始化参数的选择，可以对模型有着显著的正则化效果
- 学习输入分布有助于学习从输入到输出之前的映射

无监督预训练

作为学习一个表示

- 无监督预训练在初始表示较差的情况下更有效

作为正则化项

- 无监督预训练添加的信息来源于未标注数据
- 无监督学习不偏向学习一个简单函数，只是学习对无监督学习任务有用的特征函数

无监督预训练

缺点

- 没有明确的方法来调整无监督阶段正则化的强度
- 需要两个单独的训练阶段，每一个阶段都有各自的超参数

迁移学习和领域自适应

两者都指利用一个情景中已经学习到的内容去改善另一个情景中的泛化情况

迁移学习

- 学习器必须执行两个或更多个不同的任务，假设能够解释P_1变化的许多因素和学习P_2需要抓住的变化相关
- 概念漂移

领域自适应

- 每个情景之间任务（和最优的输入到输出的映射）都是相同的，只是输入分布稍有不同

迁移学习

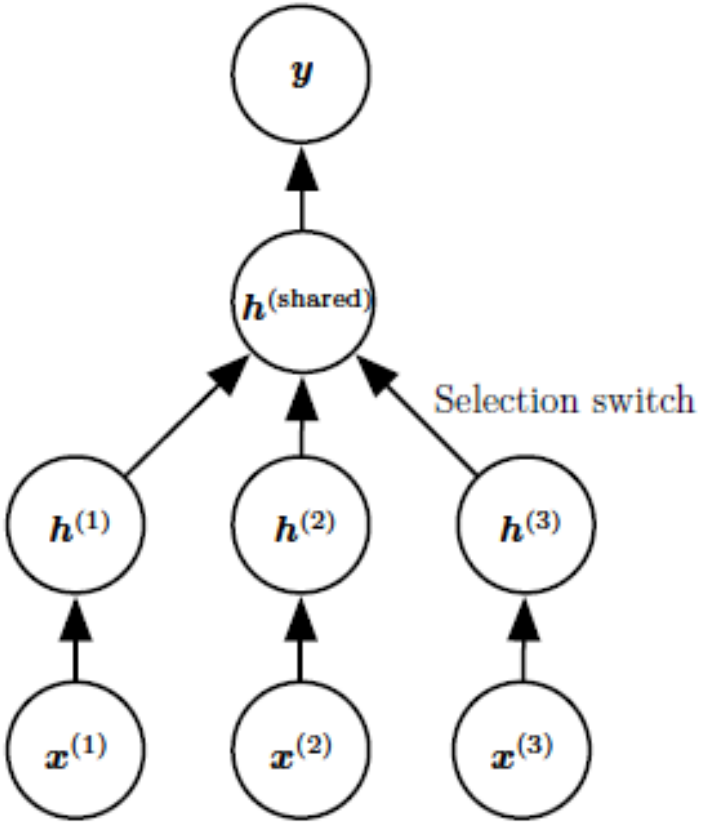
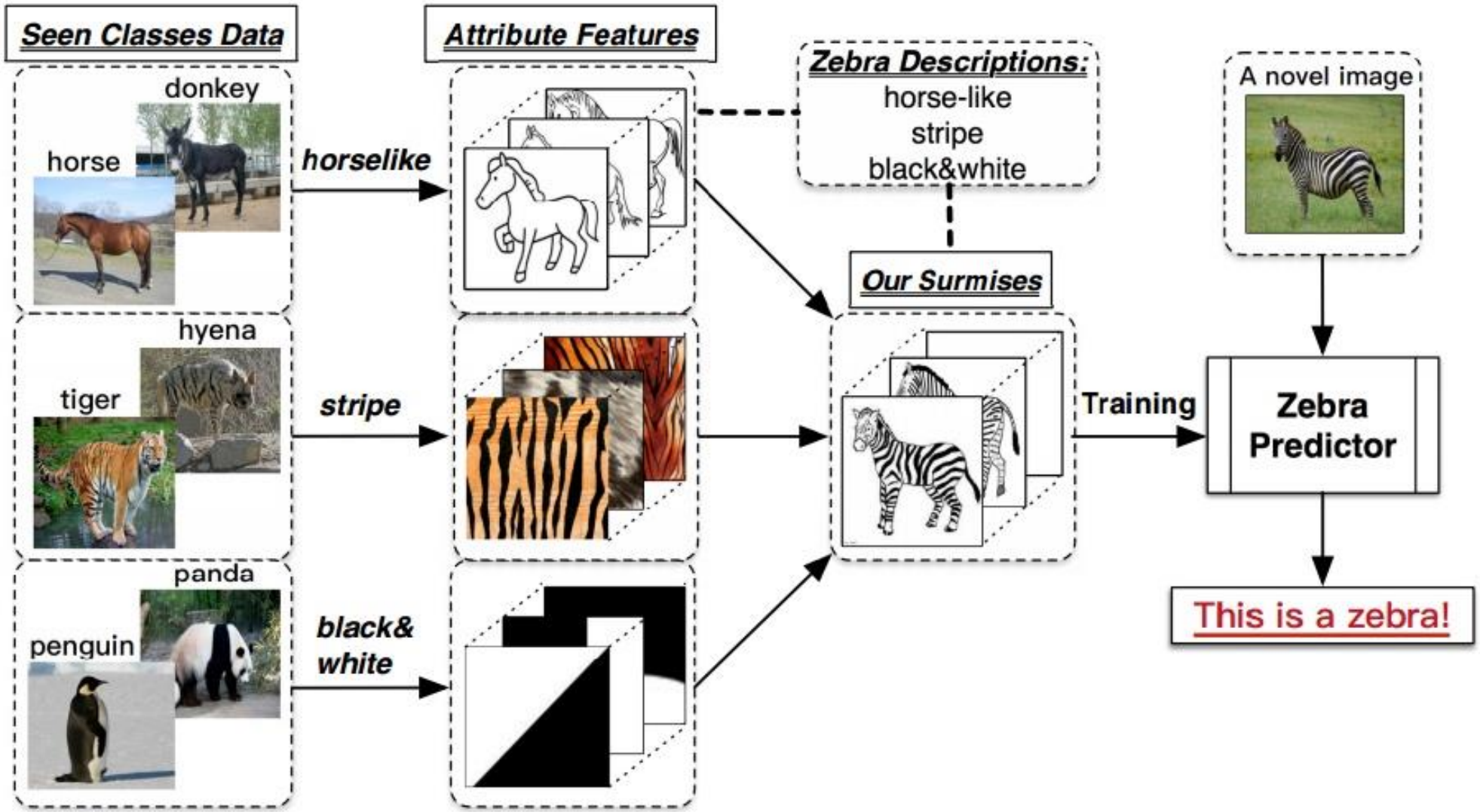


图 15.2: 多任务学习或者迁移学习的架构示例。输出变量 $\mathbf{y}$ 在所有的任务上具有相同的语义；输入变量 $\mathbf{x}$ 在每个任务（或者，比如每个用户）上具有不同的意义（甚至可能具有不同的维度），图上三个任务为 $\mathbf{x}^{(1)}$，$\mathbf{x}^{(2)}$，$\mathbf{x}^{(3)}$。底层结构（决定了选择方向）是面向任务的，上层结构是共享的。底层结构学习将面向特定任务的输入转化为通用特征。

零次学习


Seen Classes Data
donkey
horse
hyena
tiger
panda
penguin
horselike
stripe
black&
white
Attribute Features
Zebra Descriptions:
horse-like
stripe
black&white
Our Surmises
A novel image
Training
Zebra
Predictor
This is a zebra!

多模态学习

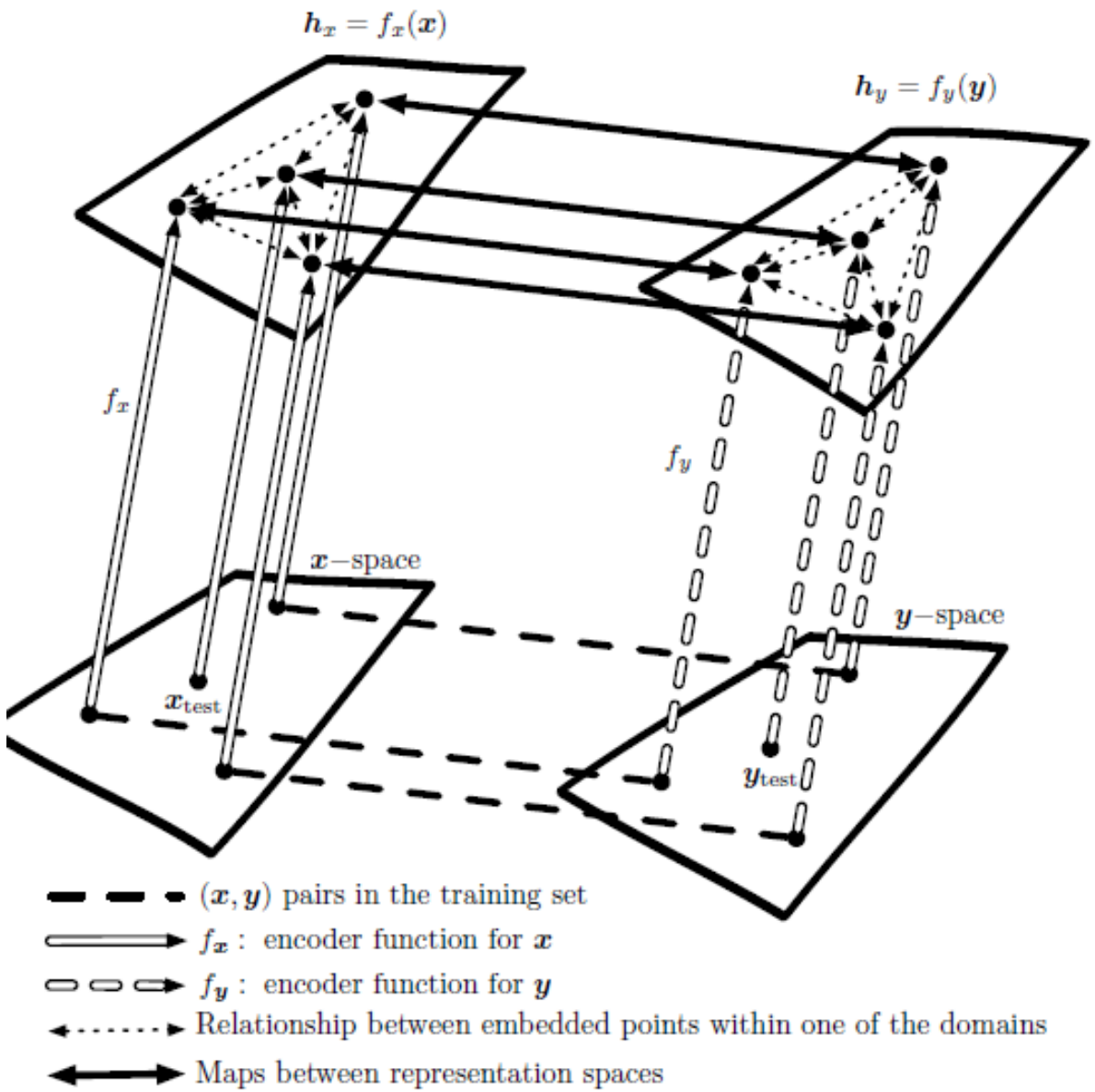
学习两种模态的表示，和(x, y)之间的关系

- x和y分别是两种模态的观察结果

通过学习**三组参数**，将一个表示中的概念锚定到另一个表示之中

- 从x到它的表示
- 从y到它的表示
- 两个表示之间的关系

多模态学习


$\boldsymbol{h}_x = f_x(\boldsymbol{x})$
$\boldsymbol{h}_y = f_y(\boldsymbol{y})$
f_x
f_y
$\boldsymbol{x}$−space
$\boldsymbol{y}$−space
$\boldsymbol{x}_{\text{test}}$
$\boldsymbol{y}_{\text{test}}$
$(\boldsymbol{x}, \boldsymbol{y})$ pairs in the training set
$f_{\boldsymbol{x}}$: encoder function for $\boldsymbol{x}$
$f_{\boldsymbol{y}}$: encoder function for $\boldsymbol{y}$
Relationship between embedded points within one of the domains
Maps between representation spaces

半监督解释因果关系

对于很多人工智能任务而言，有两个相随的特点

- 一旦我们能够获得观察结果基本成因的解释
- 那么将会很容易分离出个体属性

半监督解释因果关系

什么能将$p(y|x)$和$p(x)$关联在一起

- 如果y与x的成因之一非常相关，那么$p(x)$和$p(y|x)$也会紧密关联
- 如果y是x的成因之一，则根据贝叶斯公式

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}.$$

- 边缘概率$p(x)$和条件概率$p(y|x)$密切相关

半监督解释因果关系

什么能将$p(\mathrm{y}|\mathrm{x})$和$p(\mathrm{x})$关联在一起

- 令h表示x的成因

$$p(\mathbf{h}, \mathbf{x}) = p(\mathbf{x} \mid \mathbf{h})p(\mathbf{h}).$$

- 则x的边缘概率：

$$p(\boldsymbol{x}) = \mathbb{E}_{\mathbf{h}} p(\boldsymbol{x} \mid \boldsymbol{h}).$$

- 不难看出， h的结构信息应该有助于学习y

半监督解释因果关系

什么能将$p(\mathrm{y}|\mathrm{x})$和$p(\mathrm{x})$关联在一起

- x的成因可能有很多，因此h可能维度很高
- 假设$\mathrm{y} = \mathrm{h}_i$，暴力求解在实际情况中不可行
- 所以需要选择一个**更好的确定哪些潜在因素最为关键**的定义
  - 图像：显著地改变大量像素的亮度的影响因素是重要的
  - 生成式对抗网络方法（第20.10节）

半监督解释因果关系

什么因素决定了一种表示比另一种表示更好呢?

(另外一个答案)

- 一个理想的表示能够区分生成数据变化的潜在因果因子

分布式表示

分布式表示非常强大

- 能用具有k个值的n个特征去描述k^n个不同的概念
- 表示空间中的每个方向都对应着一个不同的潜在配置变量的值
- 泛化能力强
  - $O(nd)$个参数能够明确表示输入空间中$O(n^d)$个不同的区域

分布式表示

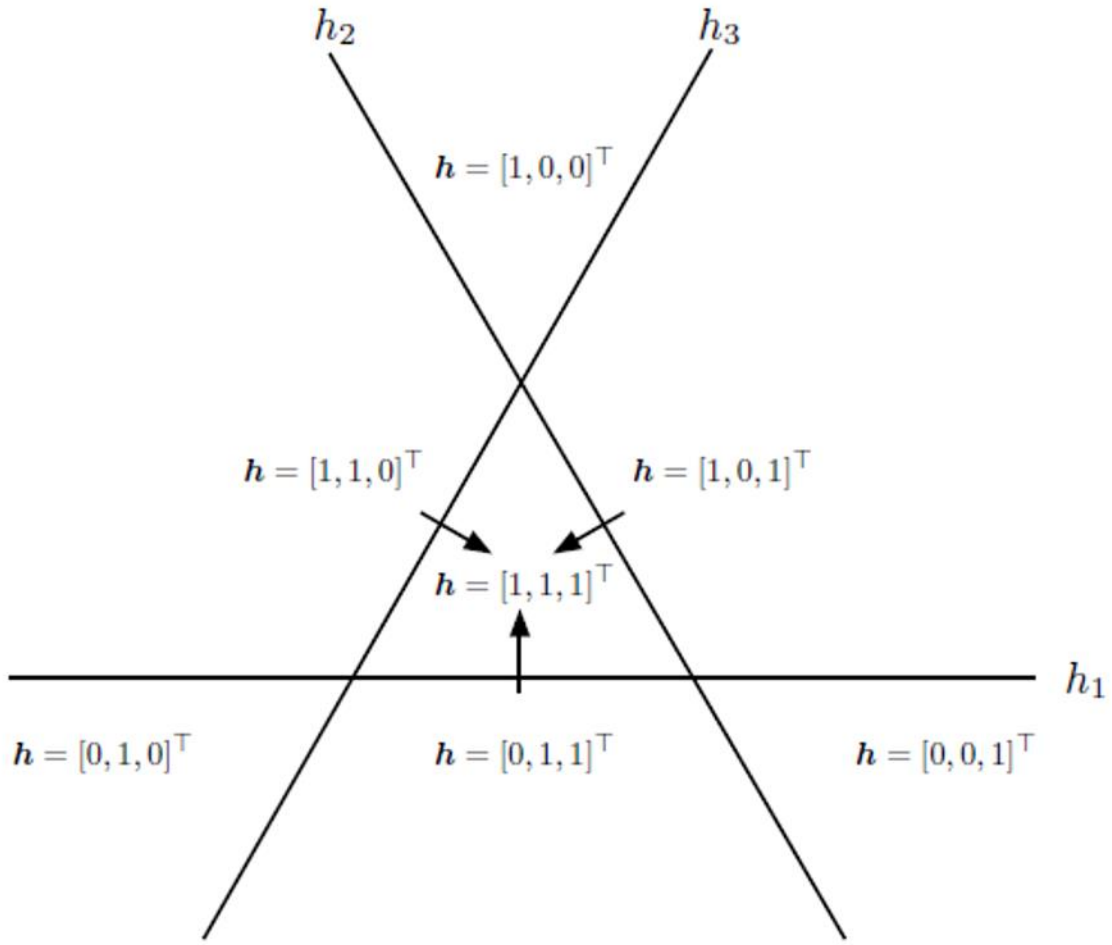


图 15.7: 基于分布式表示的学习算法如何将输入空间分割成多个区域的图示。

分布式表示

图 15.9: 生成模型学到了分布式表示，能够从戴眼镜的概念中区分性别的概念。如果我们从一个戴眼镜的男人的概念表示向量开始，然后减去一个没戴眼镜的男人的概念表示向量，最后加上一个没戴眼镜的女人的概念表示向量，那么我们会得到一个戴眼镜的女人的概念表示向量。生成模型将所有这些表示向量正确地解码为可被识别为正确类别的图像。图片转载许可自 Radford *et al.* (2015)。

深度

深度神经网络

- 层网络例如线性网络不能学习出这些抽象解释因子和图像像素之间的复杂关系
  - 需要学习的因子需要被独立地抽取，而且要对应到有意义输入的因素
  - 输入呈高度非线性的关系
- 需要深度分布式表示
- 需要许多非线性组合来获得较高级的特征（被视为输入的函数）或因子（被视为生成原因）。

潜在原因

什么因素决定了一种表示比另一种表示更好呢?

- 一个好的表示可以使后续的学习任务更容易
- 选择什么表示通常取决于后续的学习任务
- 一个理想的表示能够区分生成数据变化的潜在因果因子

提供发现潜在原因的线索

表示学习的大多数策略都会引入一些有助于学习潜在变差因素的线索

- 这些线索可以帮助学习器将这些观察到的因素与其他因素分开
- 比如：监督学习引入标签

提供发现潜在原因的线索

为了利用丰富的未标注数据，表示学习会使用关于潜在因素的其他不太直接的提示

- 强加的隐式先验信息：正则化策略
- 当不可能找到一个普遍良好的正则化策略时，深度学习的一个目标是**找到一套相当通用的正则化策略**

提供发现潜在原因的线索

正则化策略列表

- **平滑**：允许学习器从训练样本泛化到输入空间中附近的点，不能克服维数灾难
- **线性**：假定一些变量之间的关系是线性的，但是具有很大权重的线性函数在高维空间中可能不是非常平滑的
- **多个解释因子**：数据是由多个潜在解释因子生成的，只要给定每一个因子的状态，大多数任务都能轻易解决
- **因果因子**：学成表示所描述的变差因素是观察数据x的成因

提供发现潜在原因的线索

正则化策略列表

- **深度（解释因子的层次组织）**：高级抽象概念能够通过将简单概念层次化来定义
- **任务间共享因素**：通过共享的中间表示$P(\mathbf{h}|\mathrm{x})$来学习所有的$P(\mathrm{y}_i|\mathrm{x})$能够使任务间共享统计强度
- **流形**：概率质量集中，并且集中区域是局部连通的，且占据很小的体积
- **自然聚类**：输入空间中每个连通流形可以被分配一个单独的类

提供发现潜在原因的线索

正则化策略列表

- **时间和空间的相干性**：慢特征分析和相关的算法假设，最重要的解释因子随时间变化很缓慢
- **稀疏性**：大部分特征和大部分输入不相关
- **简化因子依赖**：在良好的高级表示中，因子会通过简单的依赖相互关联