中国科学技术大学
University of Science and Technology of China

《人工智能数学原理与算法》

第6章 自监督学习

6.1 自监督学习概述

凌震华

zhling@ustc.edu.cn

本章的内容组成

6.1 自监督学习概述

6.2 word2vec与BERT模型(非自回归)

6.3 自回归语言建模

6.4 大语言模型

目录

从一个例子开始：如何表示一个单词

- 自然语言处理任务：**情感分析(sentiment analysis)**

  - 输入(input)：句子/单词序列，例如，**I like this movie**
  - 输出(output)：句子情感极性类别，正向(positive)、负向(negative)、中立(neutral)
  - 算法/模型(f)：对输入进行计算，得到分类结果
  - **问题：如何在计算机中表示输入句子每个单词，使其能够被 f 有效处理?**

从一个例子开始：如何表示一个单词

- **算法/模型对于输入数据形式的基本要求**
  - 数值，可计算
  - 对于不同的单词，具有统一的形式，例如：固定个数的一组数值
- 自然语言的基本形式**“字符串”并不满足以上要求**
  - 没有可测量的物理量
  - 在计算机中以字符编码(例如ASCII编码)方式存储
  - 单词对应字符串的长度不统一

二进制	十进制	十六进制	图形	二进制	十进制	十六进制	图形	二进制	十进制	十六进制	图形
0010 0000	32	20	(space)	0100 0000	64	40	@	0110 0000	96	60	`
0010 0001	33	21	!	0100 0001	65	41	A	0110 0001	97	61	a
0010 0010	34	22	"	0100 0010	66	42	B	0110 0010	98	62	b
0010 0011	35	23	#	0100 0011	67	43	C	0110 0011	99	63	c

从一个例子开始：如何表示一个单词

- 表示单词的最基本方式：**独热(one-hot)向量**
  - 假设使用的词汇表V大小为|V|
  - 使用一个维度为 |V| 的向量表示每个单词
  - 对于词汇表中的第i个单词，该向量中位置i为 1，其他位置为 0
  - 例如，如果 **like** 是词汇表中的第 5 个单词，那么独热向量为

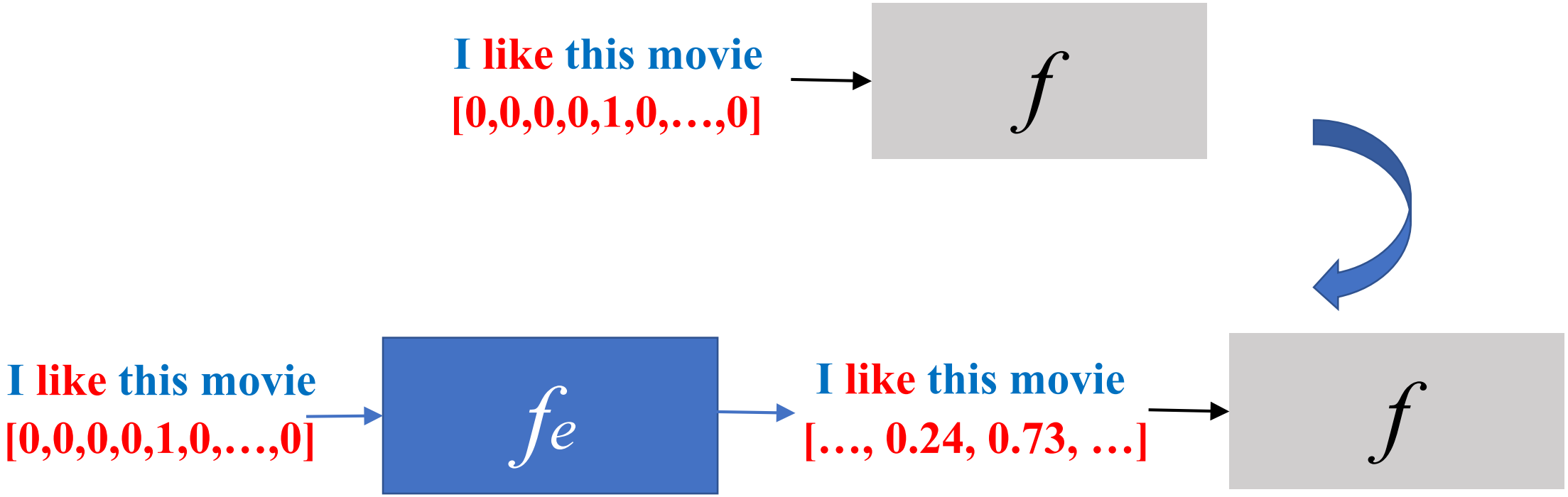
$$[0,0,0,0,1,0,\ldots,0]$$

- 独热向量的缺陷
  - **维度高**，|V|通常较大 (例如对于英文通常为数万)，造成f的参数量大
  - 不同单词的独热向量是**正交**的，无法度量不同单词之间的相似程度

I like this movie ⟷? **I love this movie**

从一个例子开始：如何表示一个单词

- **从原始数据到数据表征**



- *fe* 函数/算法/模型，将**原始数据**转换成为能够**被 *f* 有效利用的形式（数据表征）**

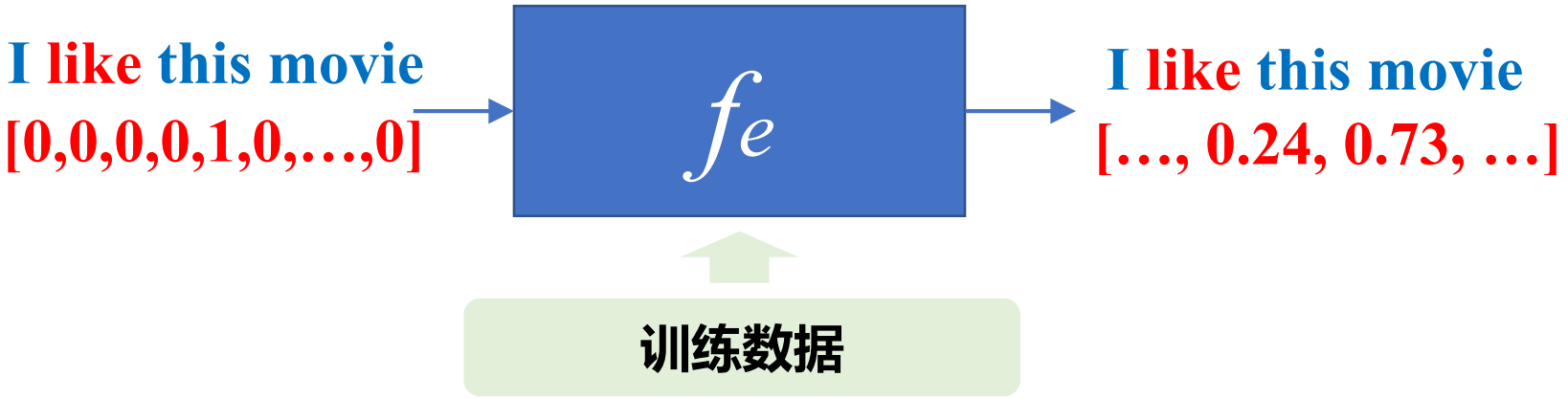
从一个例子开始：如何表示一个单词

- **如何构建 f_e**
  - 人工设计——特征工程(feature engineering)

- 当前单词是否为名词？
- 当前单词是否有情感极性？

… …

  - 学习得到——**表征学习(representation learning)**

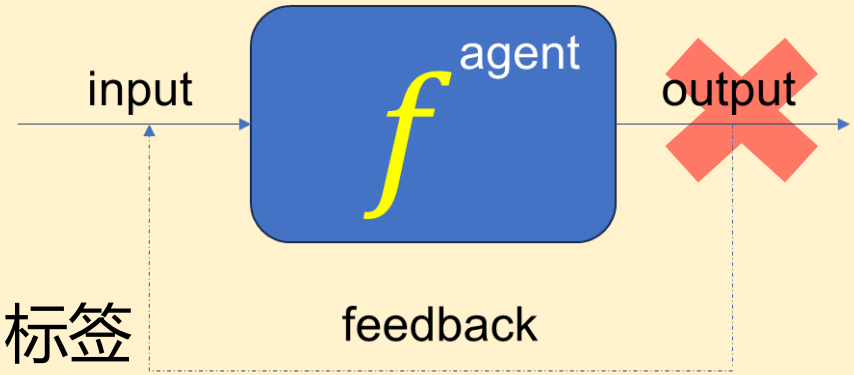


自监督学习的定义

- **自监督学习(Self-Supervised Learning, SSL)**
  - 是一种特殊的表征学习，能够从无标签数据集中学习良好的数据表征

1. 什么是“无标签数据集”？

- 数据中没有进行有监督训练所需的输出信息
- 例如，
  - 文本：只有单词序列，没有情感分析的类别标签
  - 图像：只有像素矩阵，没有人脸识别的身份编号
  - 语音：只有语音波形，没有语音识别的文本转写

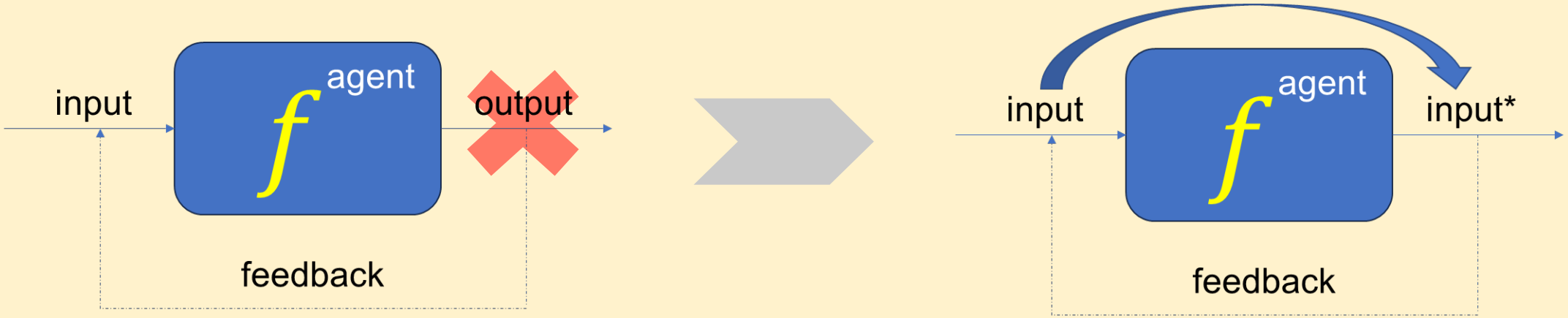


自监督学习的定义

- **自监督学习(Self-Supervised Learning, SSL)**
  - 是一种特殊的表征学习，能够从无标签数据集中学习良好的数据表征

2. 如何进行学习?

- 从无监督（无标签）数据集中构建有监督学习任务



目录

自监督学习的意义

- 数据标注成本高昂，因此高质量的标注数据集数量有限

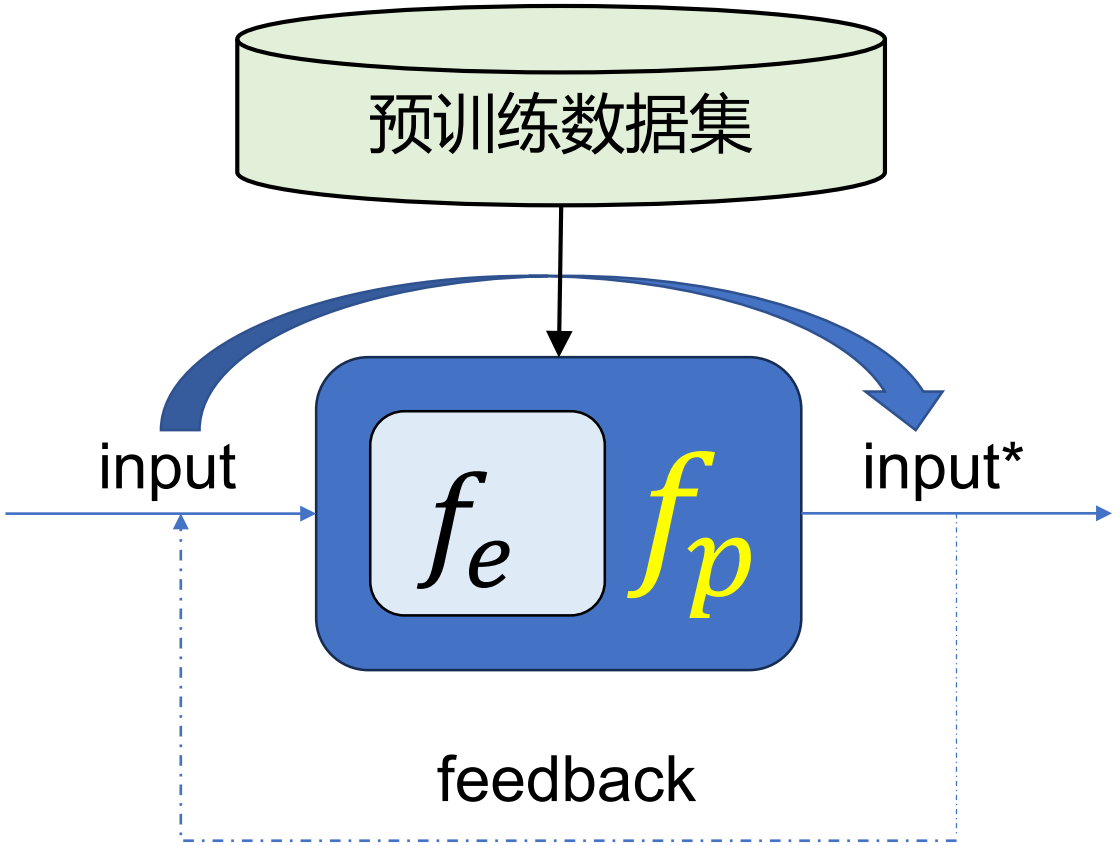
- 学习良好的表征有助于将有用信息迁移到各种下游任务中，例如：
  - 某个下游任务只有少量样本
  - 零样本 (zero-shot) 迁移到新任务

- 自监督学习任务通常也被称为前置任务 (pretext task)

自监督学习的应用范式

Step 1 预训练 (Pre-training)

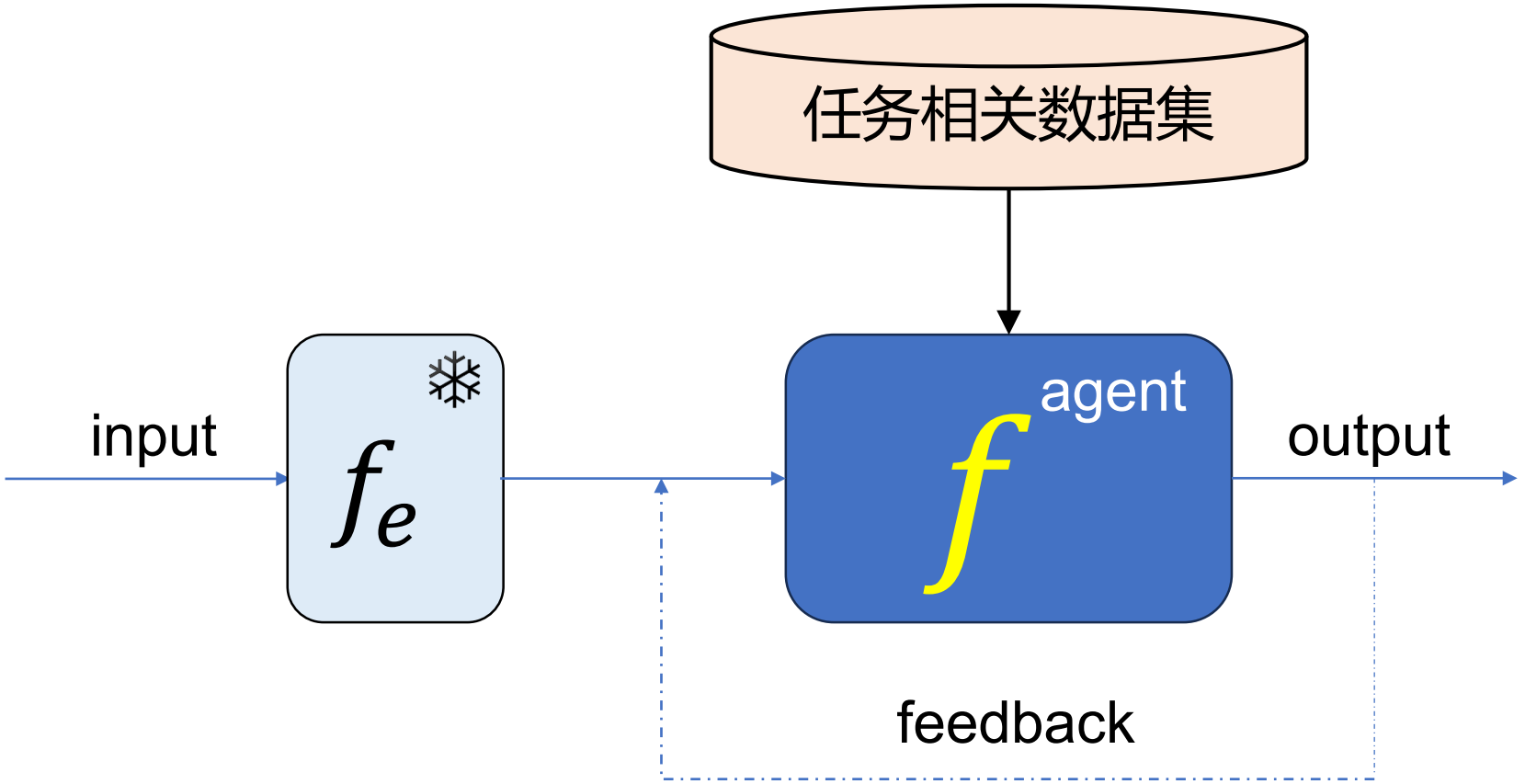
- 利用大量无标签预训练数据，通过自监督学习任务，训练 fp 模型
- 表征模型 f_e 通常为 fp 中靠近输入端的一部分



自监督学习的应用范式

Step 2 迁移

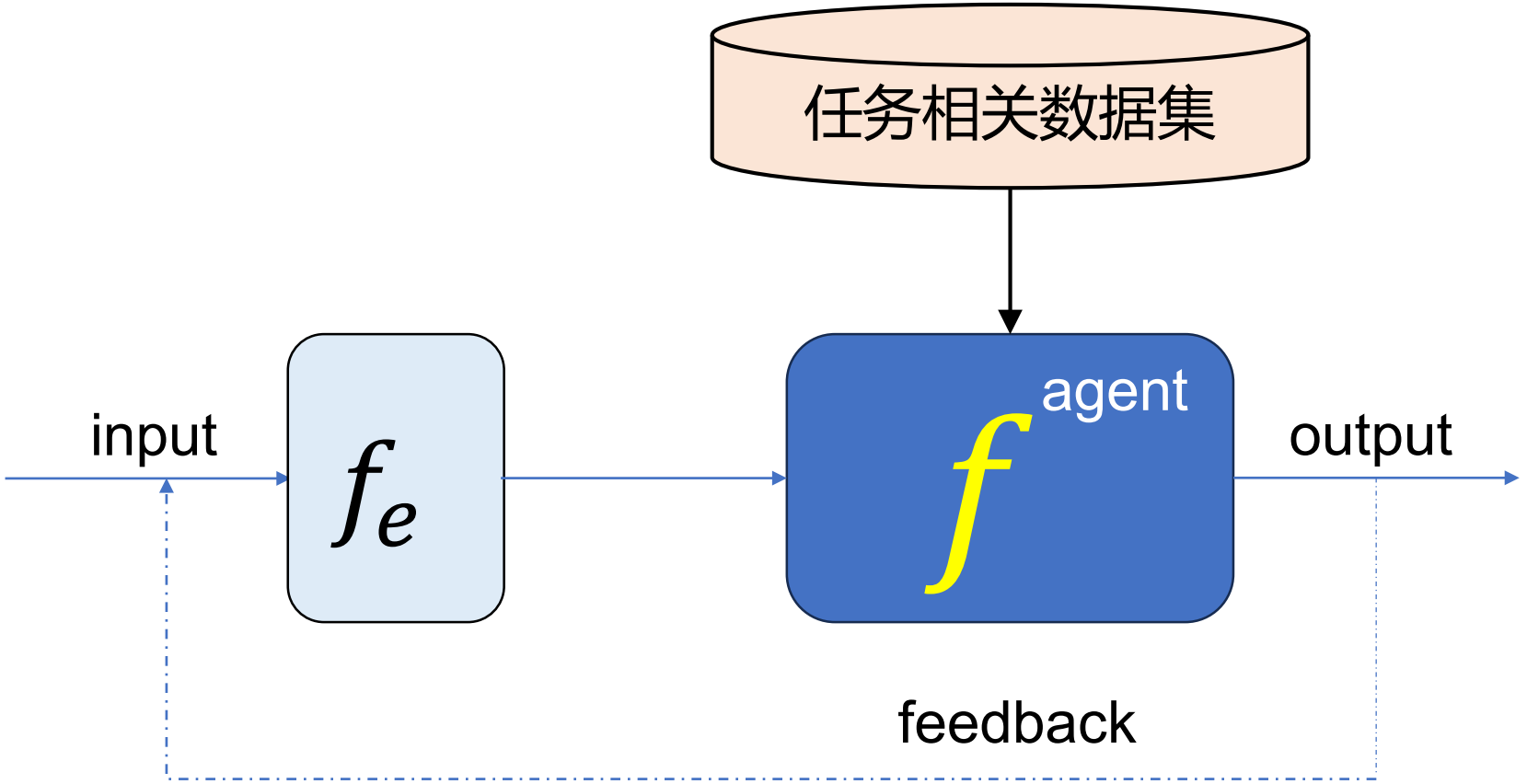
- Case 1: 固定表征模型f_e，利用任务相关有标签数据，对任务模型f进行有监督的**微调(fine-tuning)**



自监督学习的应用范式

Step 2 迁移

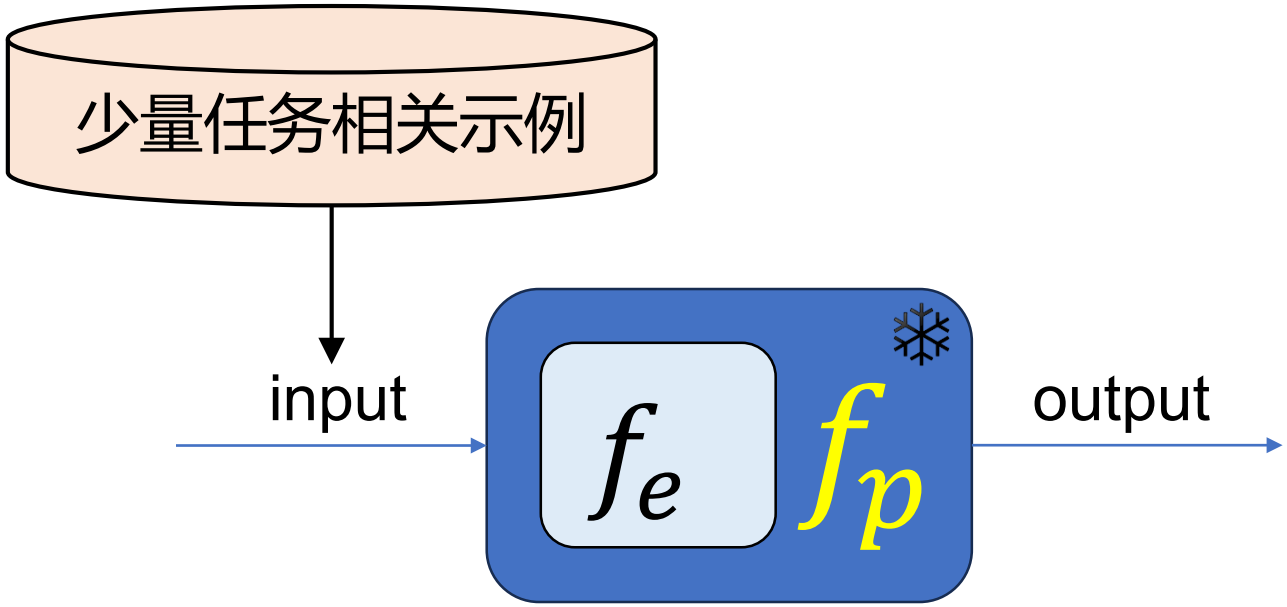
- Case 2: 利用任务相关有标签数据，对表征模型f_e和任务模型f进行有监督的**联合微调**



自监督学习的应用范式

Step 2 迁移

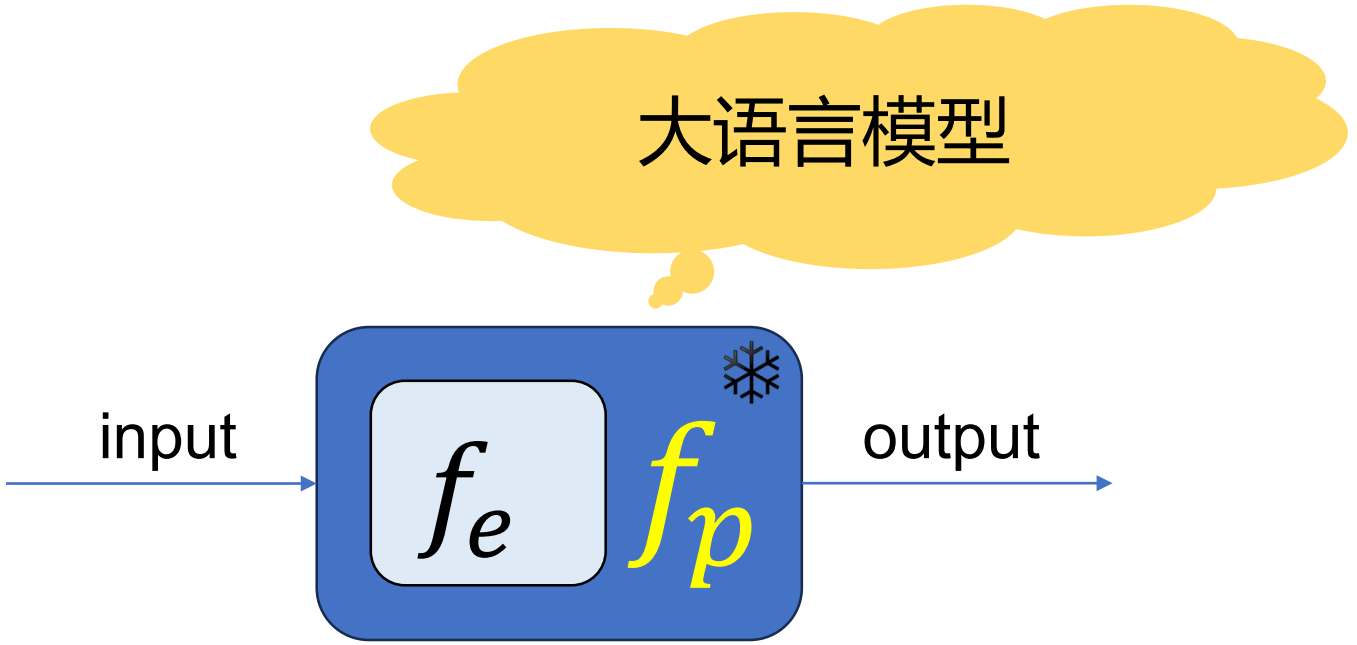
- Case 3: **上下文学习(in-context learning)**，少量任务相关示例作为模型输入，预模型 f_p 无需显式的梯度更新即可识别和执行新的任务



自监督学习的应用范式

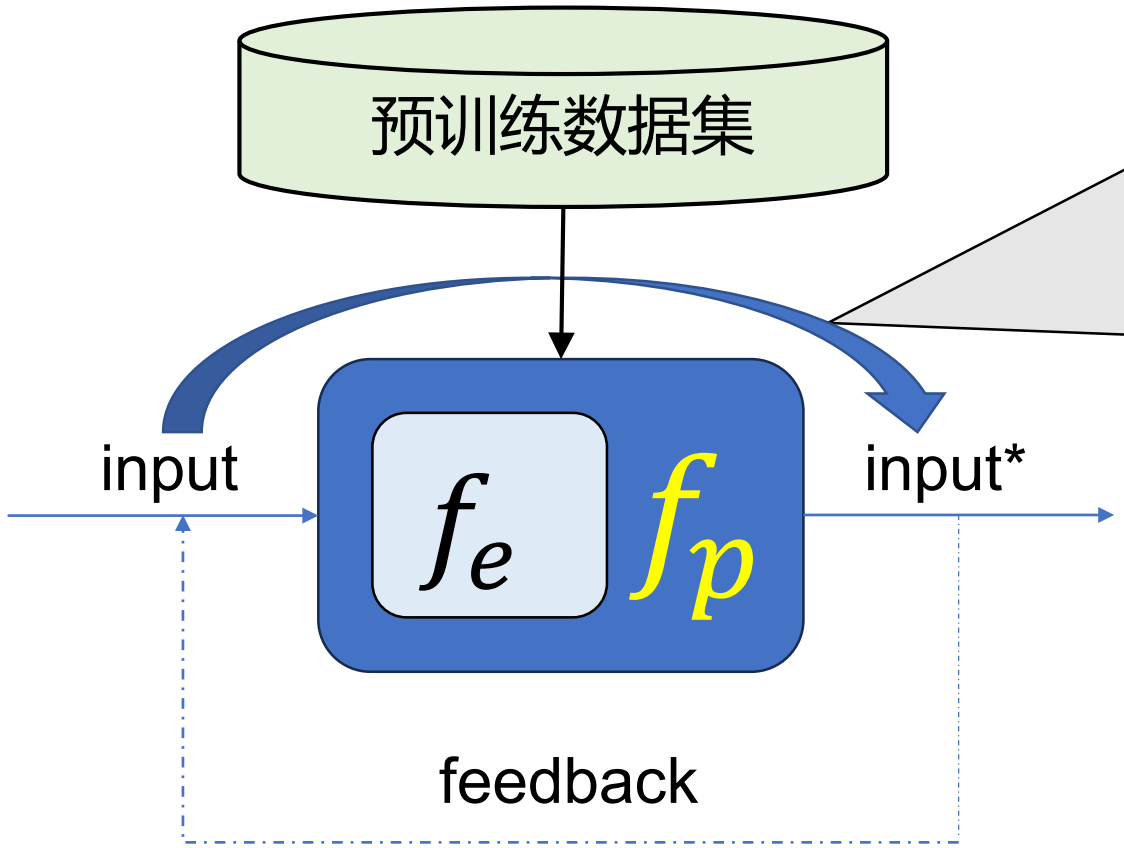
Step 2 迁移

- Case 4: **零样本学习(zero-shot learning)**，模型在没有任务相关训练样例的情况下有能力执行多种任务，不需要经过任务相关的有监督参数更新



目录

自监督学习的主要框架



如何从无监督（无标签）数据集中构建有监督学习任务?

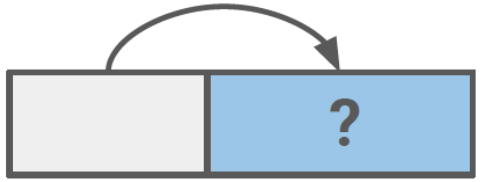
当前两种主要框架

- 自预测 (self-prediction)
- 对比学习（contrastive learning)

自监督学习的主要框架

1. 自预测

- 给定单个数据样本，任务是根据样本的一部分预测另一部分
- 被预测的部分被假定为缺失
- 这是一种 “**样本内 (intra-sample)**” 预测



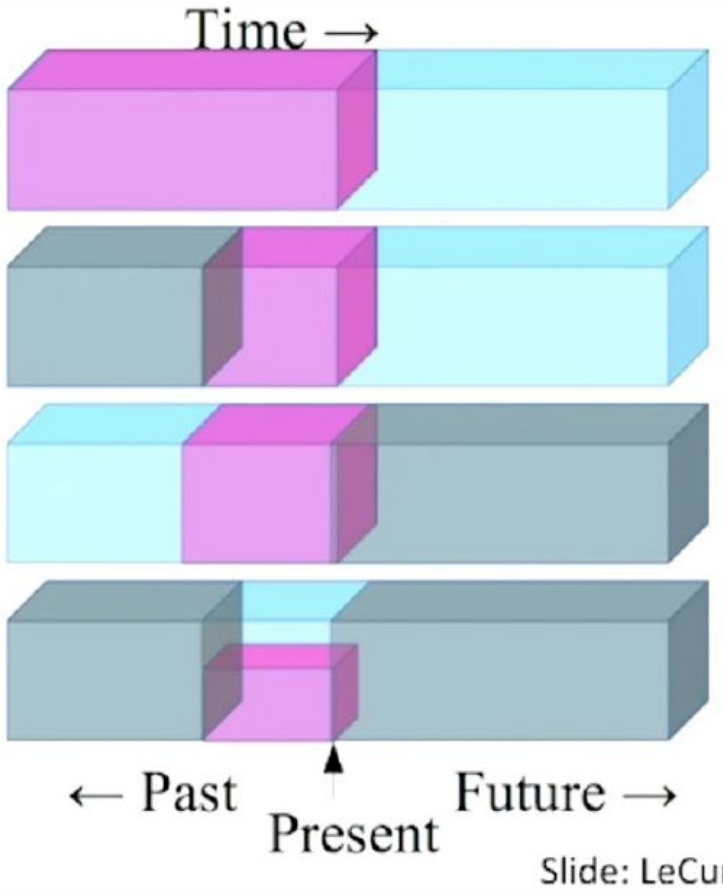
2. 对比学习

- 给定多个数据样本，任务是预测它们之间的关系
- 这些样本可以根据已知逻辑从数据集中选择，也可以通过改变原始数据生成
- 这是一种 “**样本间 (inter-sample)**” 预测

1. 自预测

- 在每个数据样本内构建预测任务；在假装不知道数据某部分的情况下，根据其余部分预测该部分

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



Slide: LeCun

1. 自预测

- 在每个数据样本内构建预测任务；在假装不知道数据某部分的情况下，根据其余部分预测该部分

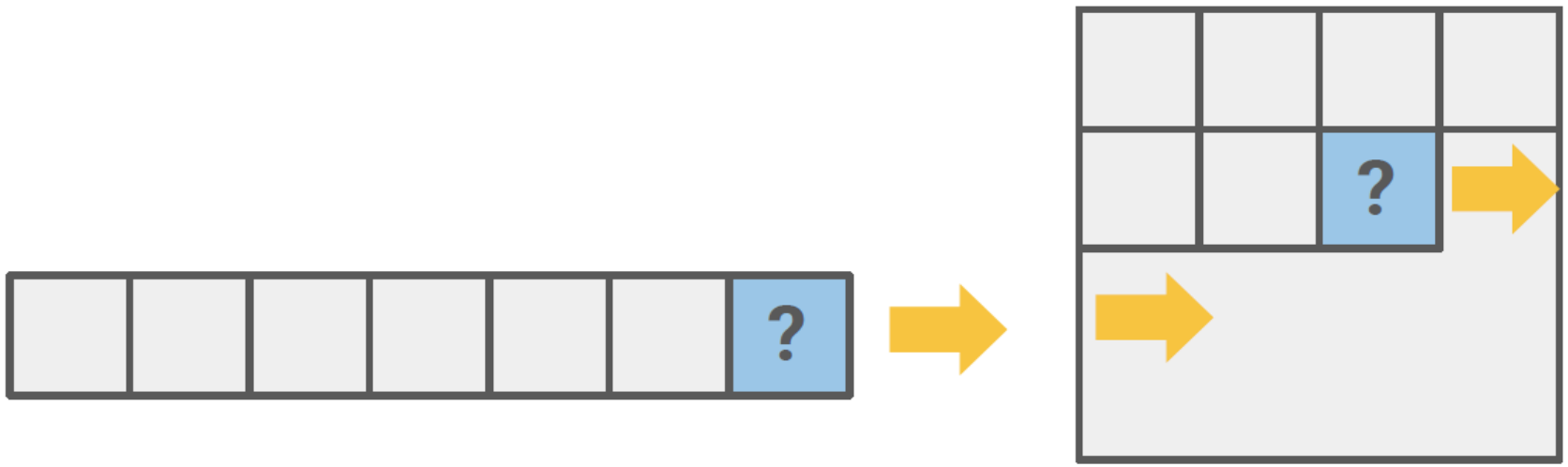
（1）自回归生成 (autoregressive generation)

（2）掩码生成 (masked generation)

（3）内在关系预测 (innate relationship prediction)

1. 自预测——（1）自回归生成

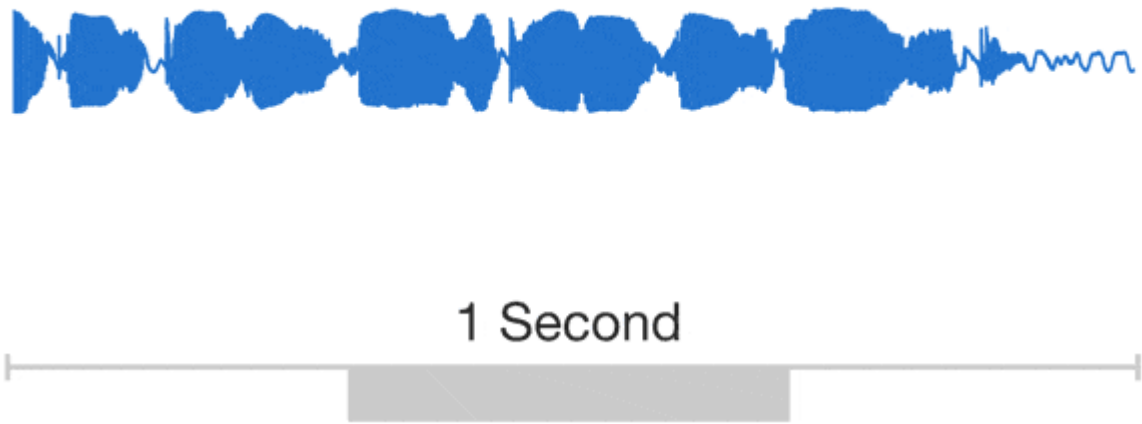
- 根据过去的数据预测未来的数据
- 任何具有固有顺序的数据都可以用回归模型进行建模，例如：
  - 音频（WaveNet、WaveRNN）
  - 语言（GPT、XLNet）
  - 图像（PixelCNN、PixelRNN）

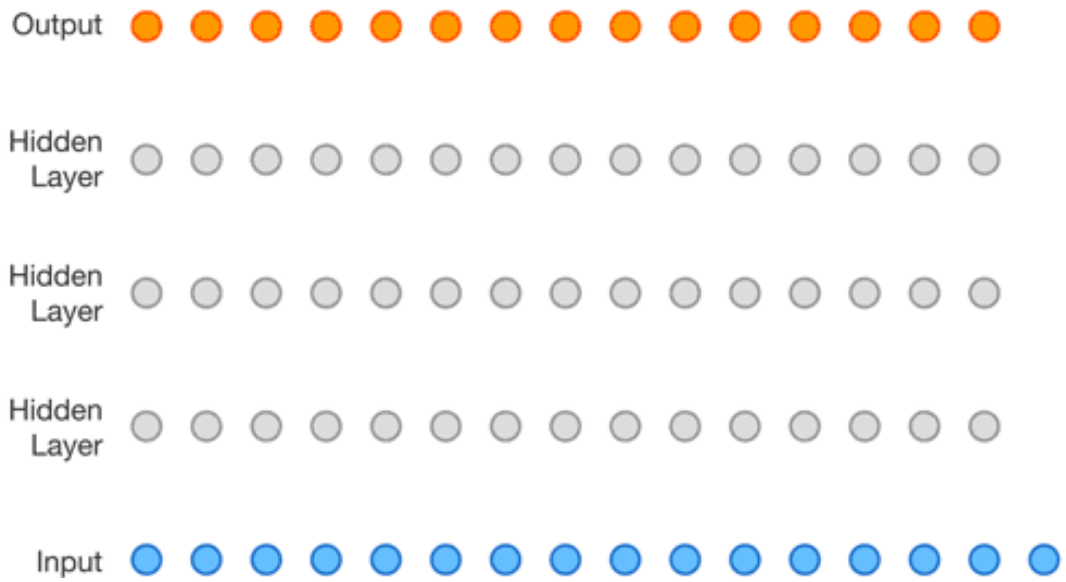


1. 自预测——（1）自回归生成

- 示例：用于音频建模的WavNet模型 [van den Oord et al. 2016]

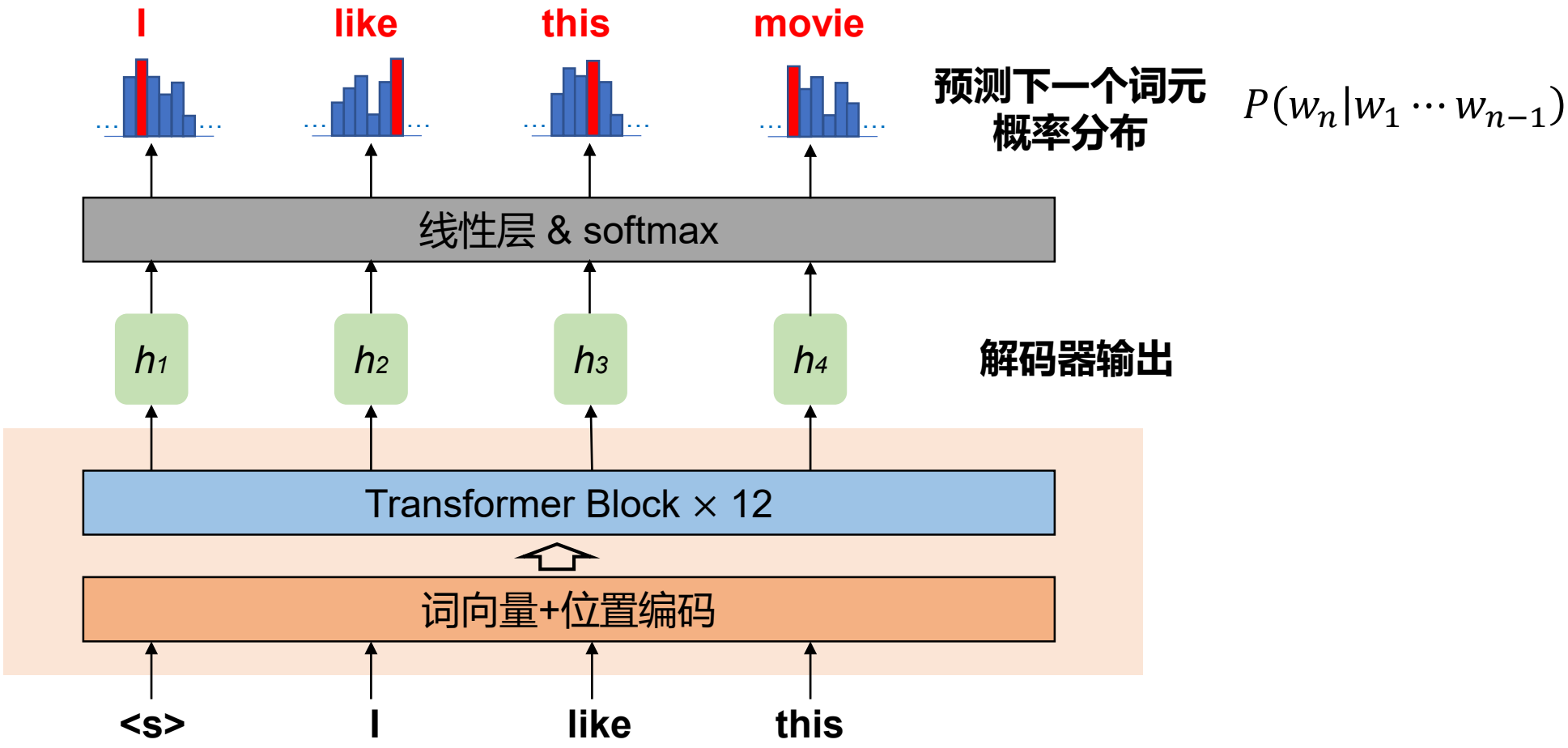
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$





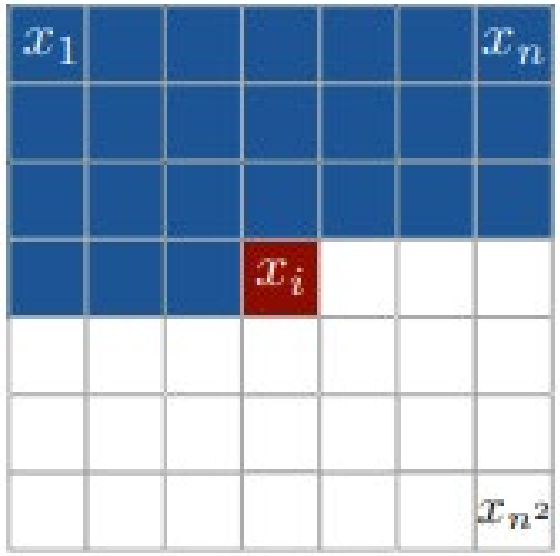
1. 自预测——（1）自回归生成

- 示例：用于语言建模的GPT模型 [Radford et al. 2018]

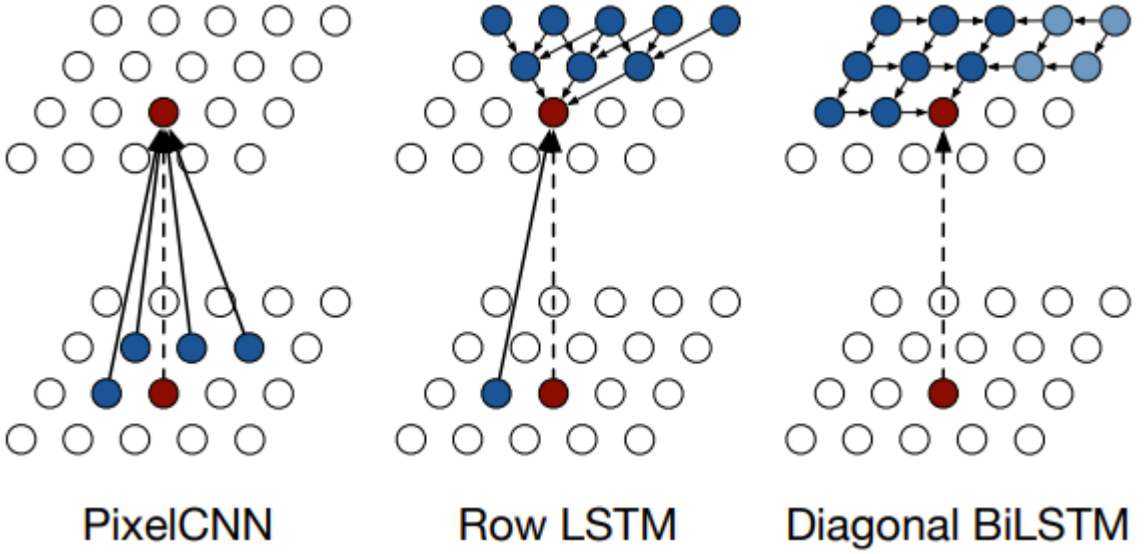


1. 自预测——（1）自回归生成

- 示例：用于图像建模的PixelRNN/PixelCNN模型 [van den Oord et al. 2016]

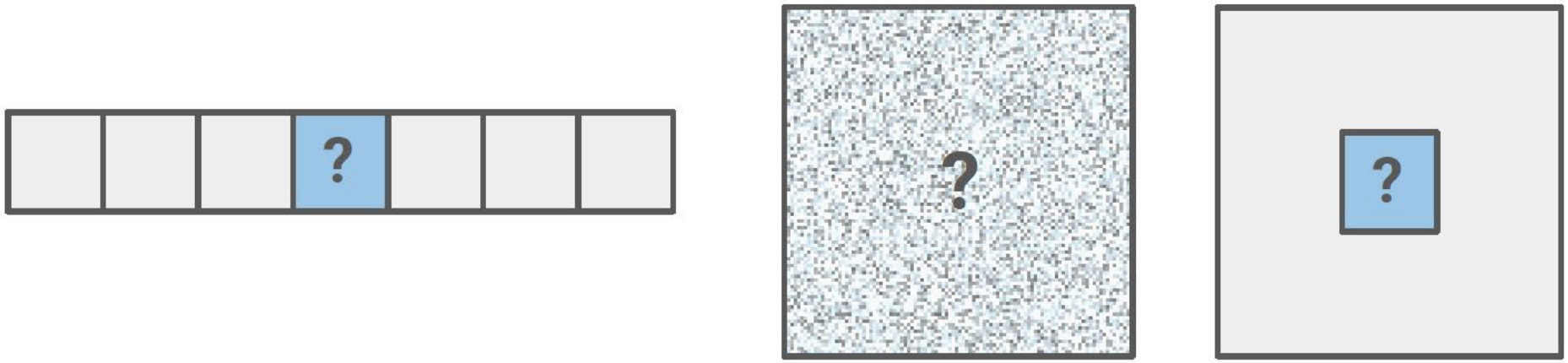


$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1})$$



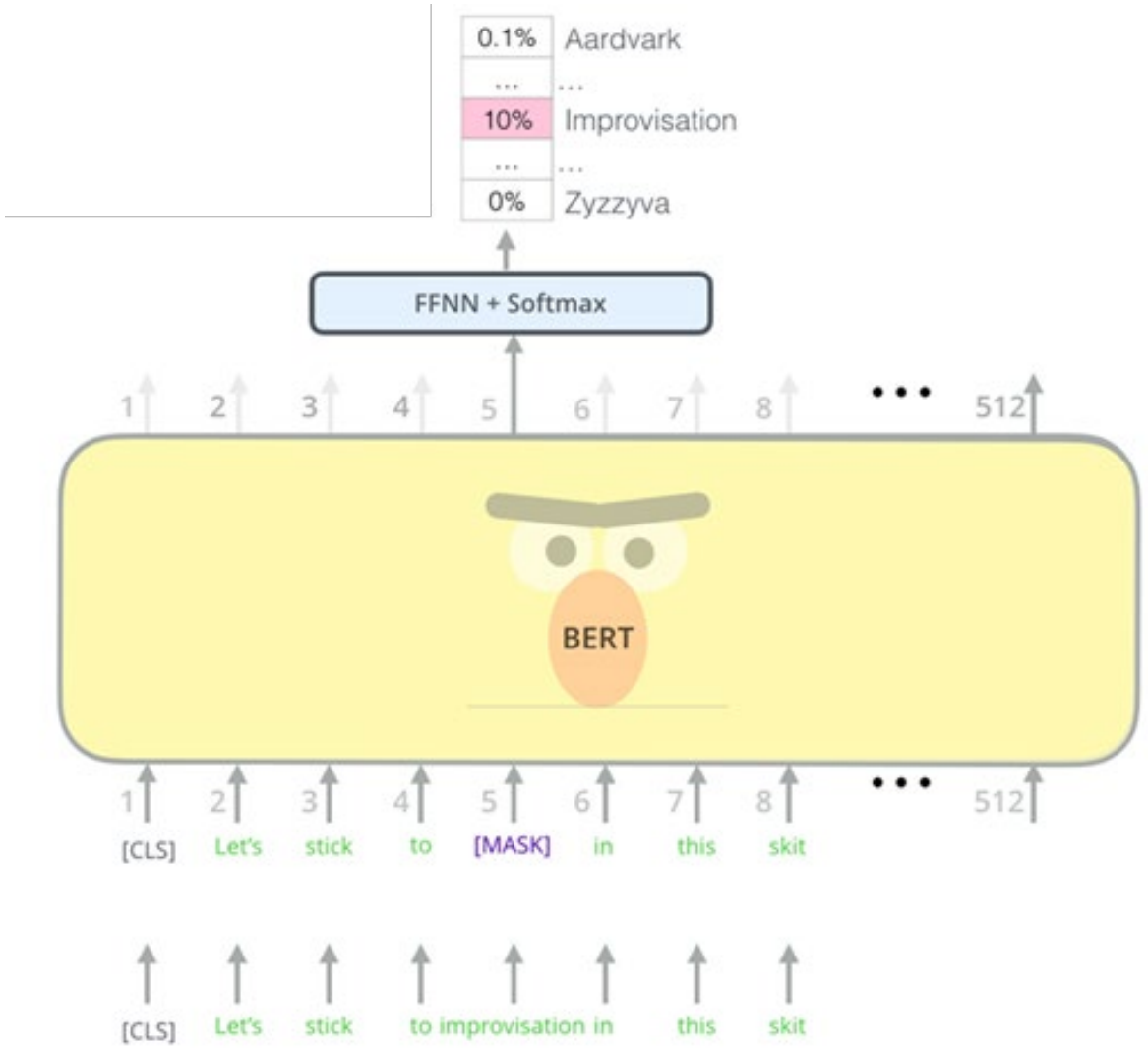
1. 自预测——（2）掩码生成

- 随机掩盖一部分信息并假装其缺失，而不考虑自然顺序
- 模型根据其他未掩盖的信息预测缺失部分，例如：
  - 掩码语言建模（BERT）
  - 掩码图像建模（去噪自动编码器、上下文自动编码器、上色）



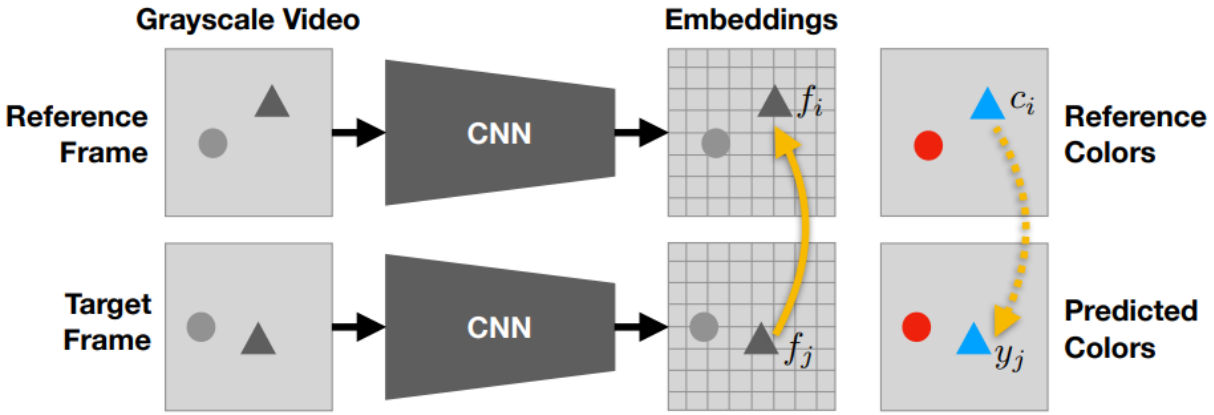
1. 自预测——（2）掩码生成

- 示例：BERT [Devlin et al. 2018]
  - Bidirectional Encoder Representations from Transformers
  - 自然语言的自监督学习模型
  - 掩码语言模型 (masked language model, MLM)是BERT模型的学习任务之一

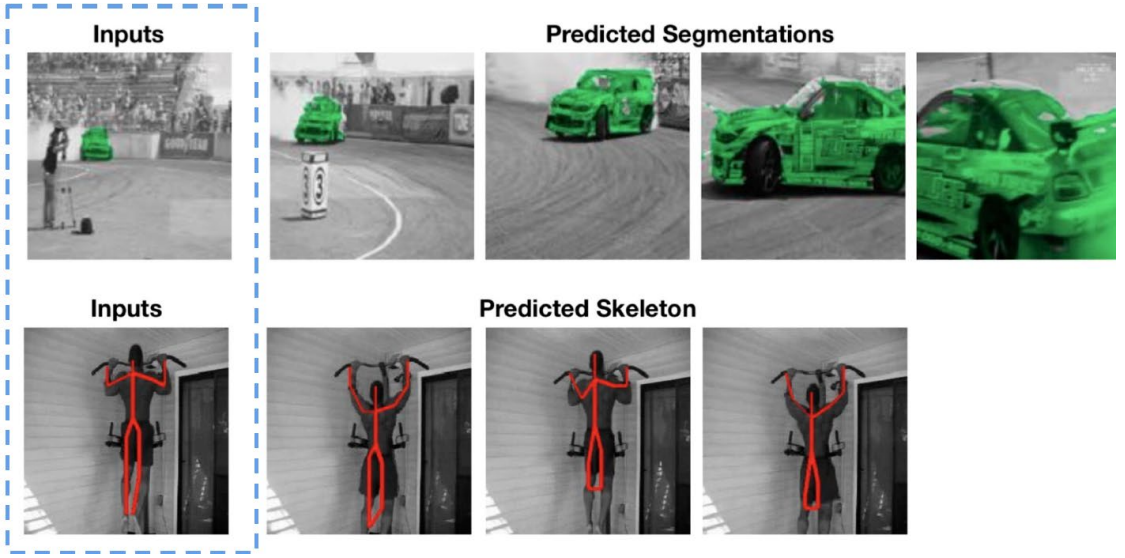


1. 自预测——（2）掩码生成

- 示例：视频上色 (Video colorization) [Vondrick et al. 2018]
  - 从灰度视频预测颜色
  - 作为一种自监督学习方法，能产生丰富的表征
  - 可用于视频分割和无标签视觉区域跟踪，且无需额外微调



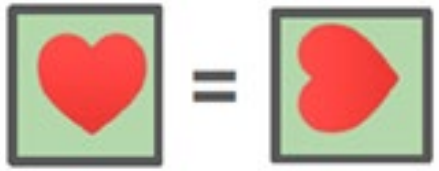
自监督学习任务

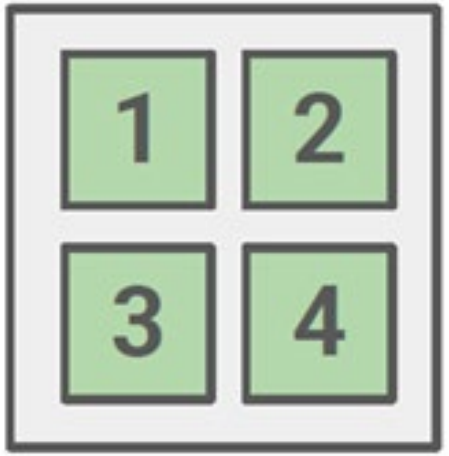


下游任务

1. 自预测——（3）内在关系预测

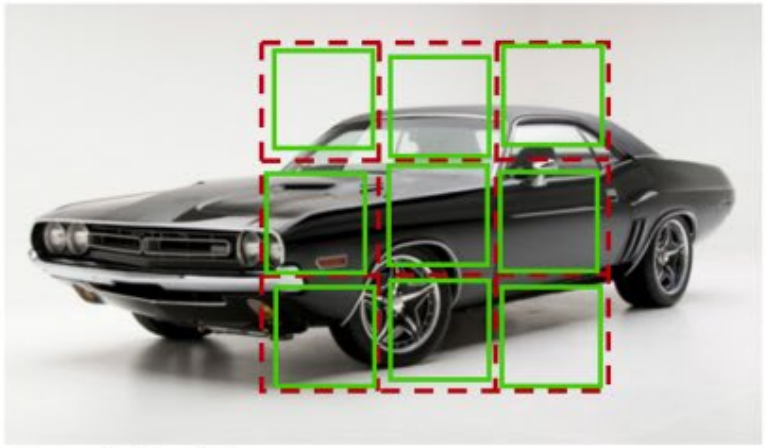
- 对一个数据样本进行的某些变换（例如分割、旋转）应保持原始信息或遵循所需的内在逻辑，例如：
  - 图像旋转
  - 图像块顺序（例如相对位置、拼图）





1. 自预测——（3）内在关系预测

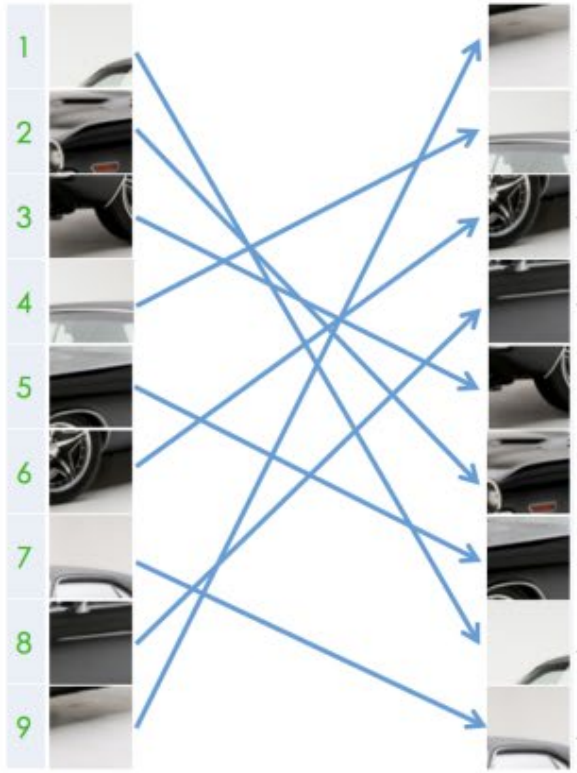
- 示例：基于拼图的视觉表征学习[Noroozi et al. 2016]
  - 将图片分为 9 块，预先定义好 64 种排序方式
  - 模型输入任意一种被打乱的序列，期待能够学习到这种序列的顺序属于哪个类

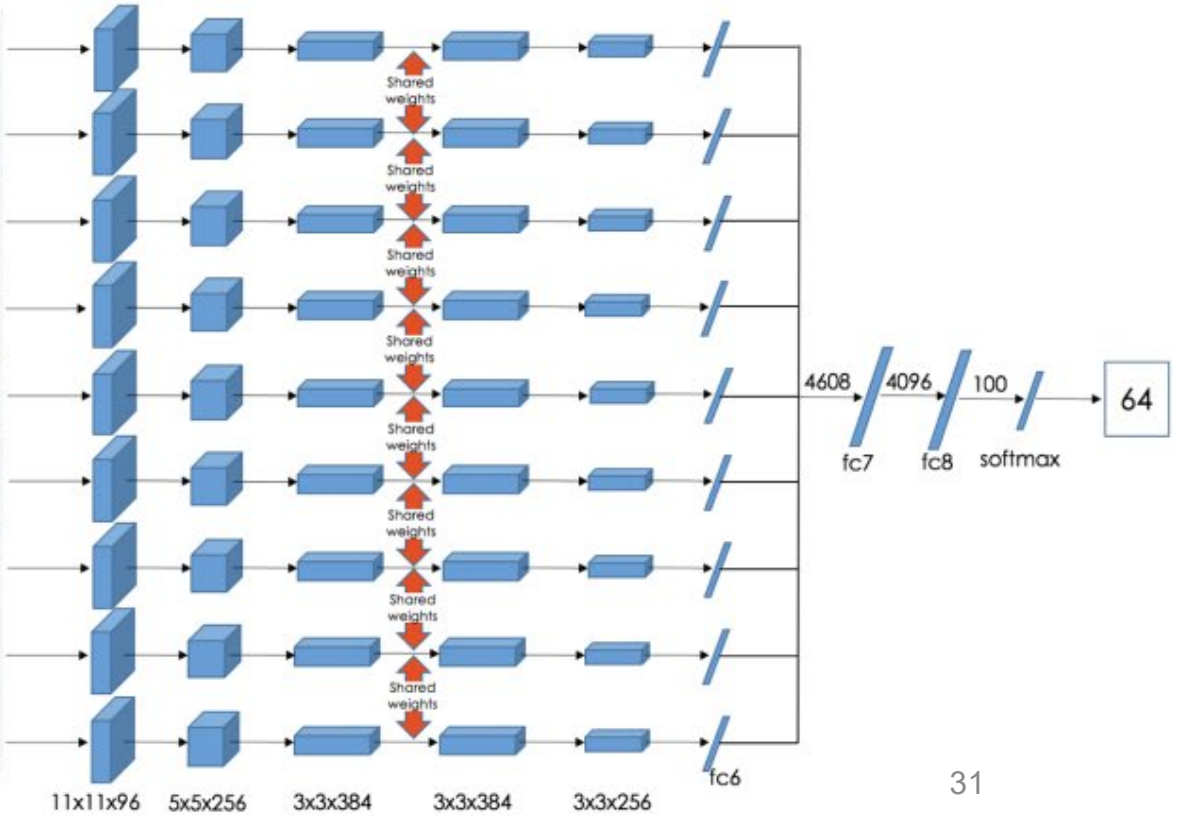


Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

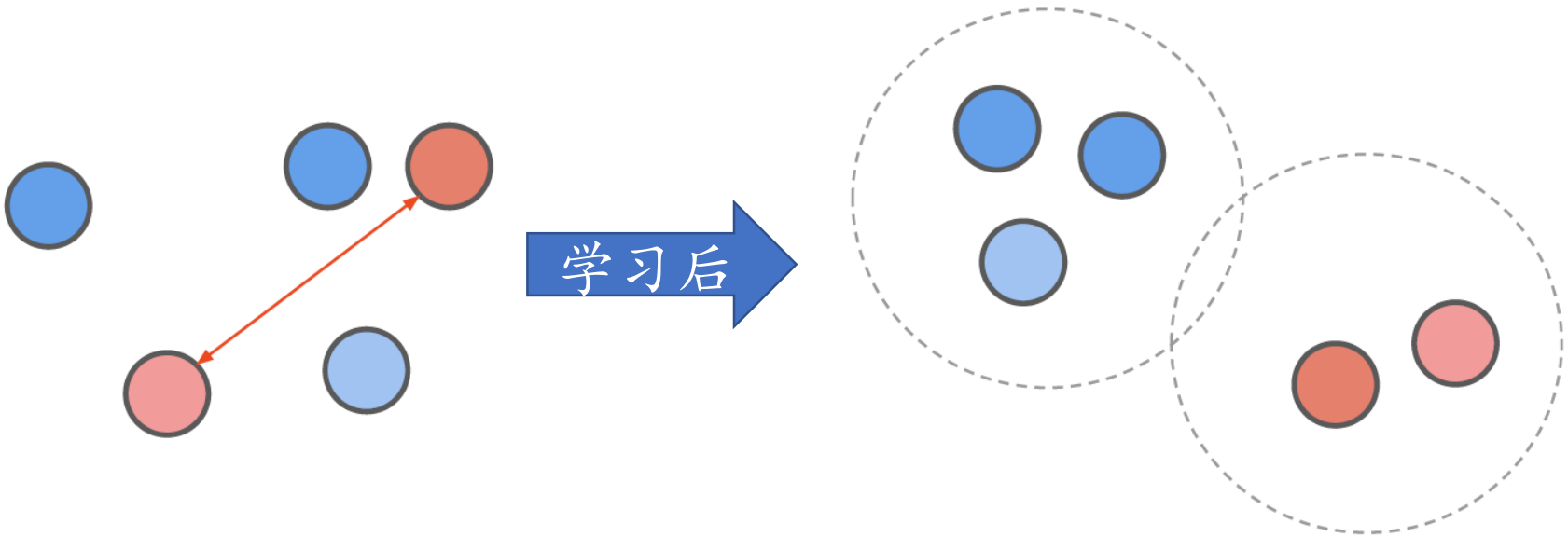
Reorder patches according to the selected permutation





2. 对比学习

- 目标是学习一个嵌入空间 (embedding space)，即数据表征所在空间
- 其中相似的样本对彼此靠近，而不相似的样本对彼此远离



2. 对比学习

- 目标是学习一个嵌入空间 (embedding space)，即数据表征所在空间
- 其中相似的样本对彼此靠近，而不相似的样本对彼此远离

（1）样本间分类 (inter-sample classification)

（2）特征聚类 (feature clustering)

（3）多视角编码 (multiview coding)

2. 对比学习——（1）样本间分类

- 给定与目标数据点（**锚点 anchor 数据**）相似（**正样本 positive**）和不相似（**负样本 negative**）的候选样本集合
- 识别哪些样本与锚点数据相似，是一个**分类任务**

- **问题一：如何构建数据点的正负样本集?**

- **问题二：如何设计分类任务的损失函数?**

2. 对比学习——（1）样本间分类

- 示例：FaceNet [Schroff et al. 2015]

样本集构建

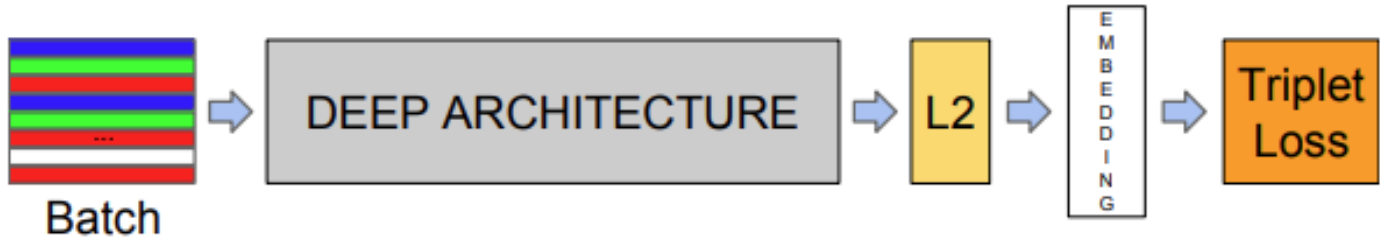
Anchor 一幅人脸图片

Positive 与Anchor来自同一人的其他人脸图片

Negative 来自任意其他人的人脸图片

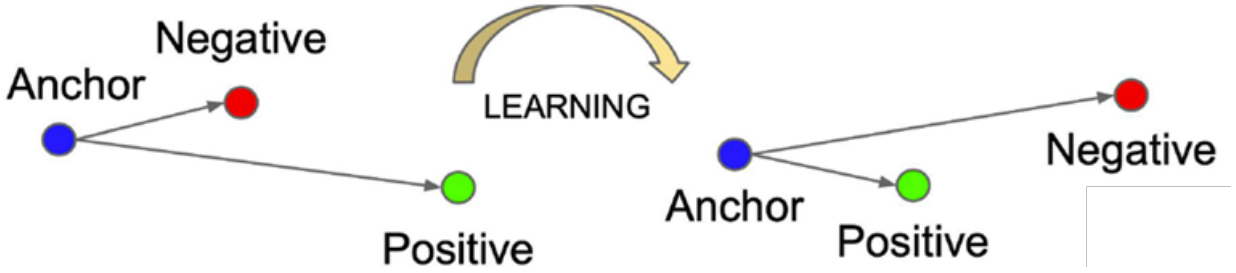
三元组损失函数

- 最小化锚点x与正样本$x+$之间的距离
- 最大化锚点x与负样本$x-$之间的距离



模型结构

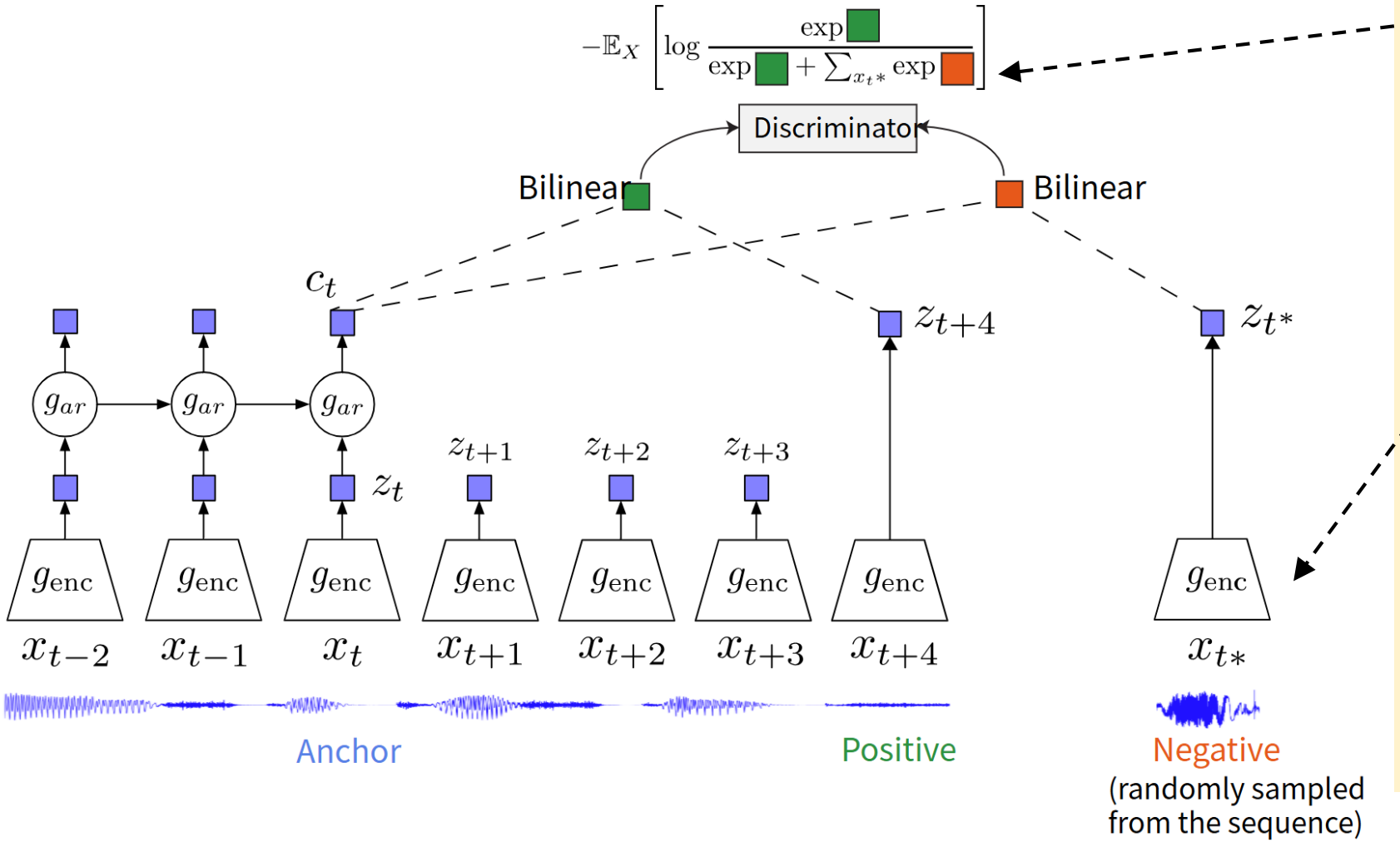
$$\mathcal{L}_{\text{triplet}}(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = \sum_{\mathbf{x}\in\mathcal{X}} \max\left(0, \|f(\mathbf{x}) - f(\mathbf{x}^+)\|_2^2 - \|f(\mathbf{x}) - f(\mathbf{x}^-)\|_2^2 + \epsilon\right)$$



三元组(triplet)损失

2. 对比学习——（1）样本间分类

- 示例：对比预测编码(Contrastive Predictive Coding)[van den Oord et al. 2019]



InfoNCE损失函数

- 1 个正例，N-1个负例
- N分类的交叉熵损失，希望锚点数据能够分到正例对应的类别

样本集构建

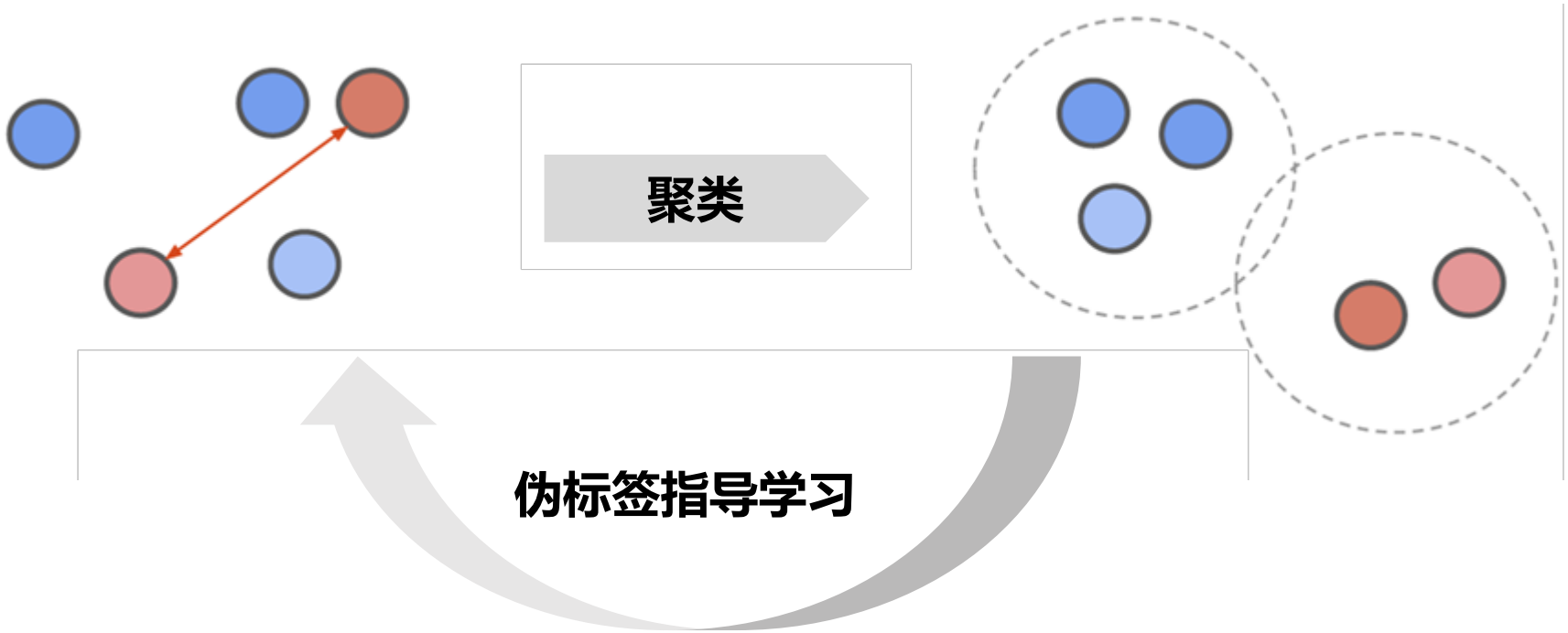
Anchor 语音片段

Positive 与Anchor近邻的未来语音片段

Negative 从当前语句或其他语句中随机采用的语音片段

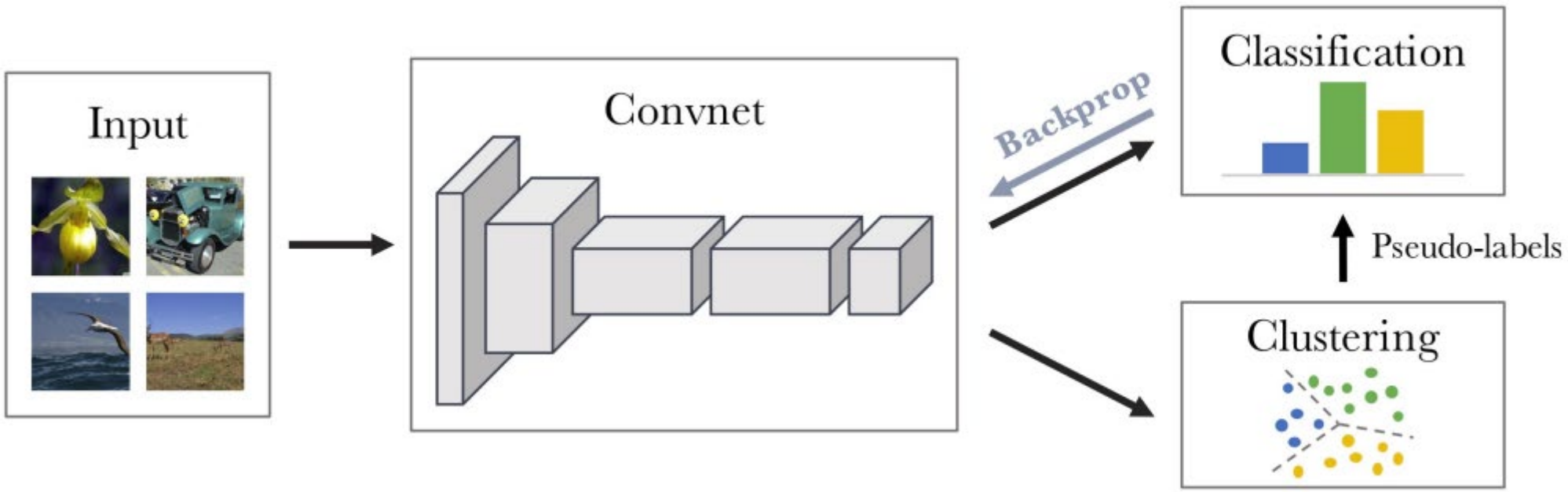
2. 对比学习——（2）特征聚类

- 通过使用学习到的特征对数据样本进行聚类，找到相似的数据样本
- 对聚类后的样本分配伪标签，使用伪标签进一步指导表征模型的学习



2. 对比学习——（2）特征聚类

- 示例：DeepCluster [Caron et al. 2018]
  - 以得到通用的视觉表征为目标
  - 对卷积神经网络得到的图片表征向量进行无监督聚类
  - 使用聚类得到伪标签作为分类目标，指导卷积神经网络更新
  - 以上两步交替进行



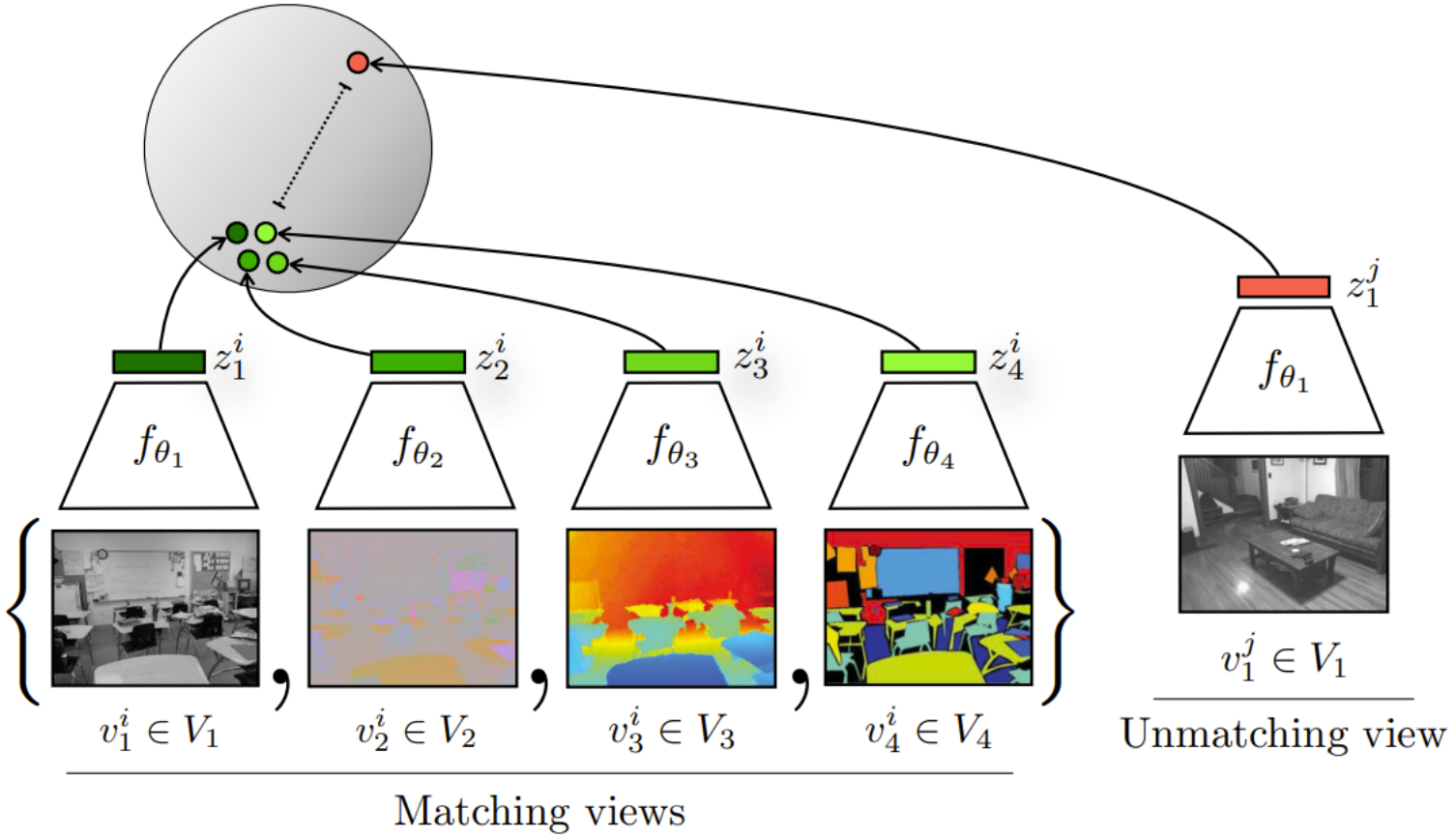
2. 对比学习——（3）多视角编码

- 在现实世界中同一实体是能够通过多个视角(view)的数据来描述的
- 这里的“视角”是广义的概念
  - **同模态**：不同波段的光学图像
  - **跨模态**：对于同一个场景，既可以用图像描述，也可以用文本描述
- 每个视角信息往往都是有噪声、不完整的
- 但是一些重要信息往往被不同视角所共享

- **将对比学习目标应用于输入数据的两个或更多不同视角**

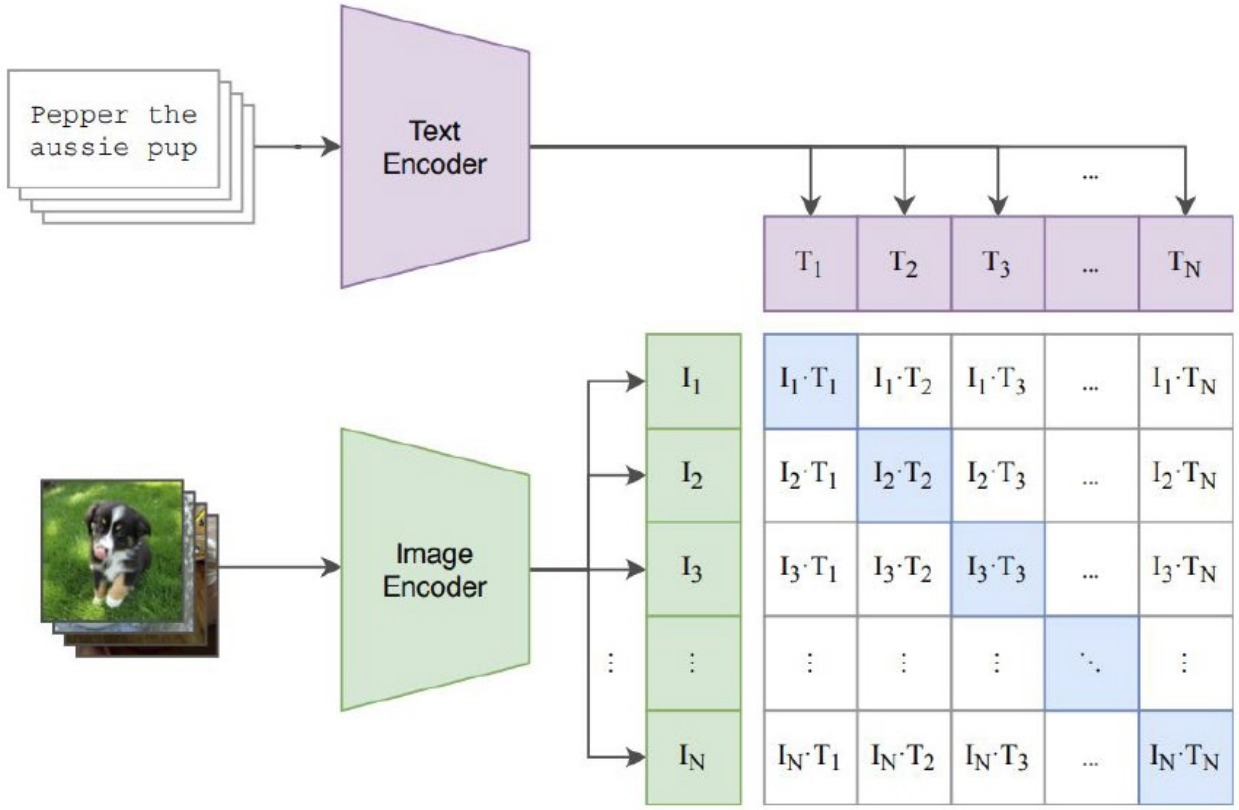
2. 对比学习——（3）多视角编码

- 示例：对比多视角编码
  [Tian et al. 2019]
  - 基于多传感器视角获得的场景图像，学习图像表征
  - 同一场景不同视角的表征尽量接近（正例）
  - 不同场景的表征尽量远离（负例）
  - 使用类似InfoNCE损失函数



2. 对比学习——（3）多视角编码

- 模态间的对比学习
  - “视角” 可以来自两个或更多模态的配对输入
  - CLIP [Radford et al. 2021] ALIGN [Jia et al. 2021]：实现零样本分类、跨模态检索、引导图像生成
  - CodeSearchNet [Husain et al 2019]：文本和代码之间的对比学习



本节小节

- **自监督学习的目的**
  - 将原始数据转化成能够被下游任务模型更好利用的**数据表征**
  - 有监督数据量受限情况下， 将无监督数据中的有用信息迁移到各种下游任务中
- **应用范式**
  - **预训练+迁移**；迁移有多种方式，包括**微调、上下文学习、零样本学习**等
- **核心思想**
  - **无监督（无标签）数据集中构建有监督学习任务**
- **主要框架**
  - **自预测**：根据样本的一部分预测另一部分， “样本内” 预测
  - **对比学习**：预测多个样本之间的关系， “样本间” 预测
- 在**文本、图像、语音**等各模态任务上已经有了广泛应用

讨论：为什么自监督学习有效?

1. **利用海量未标注数据**
   - 自监督学习无需依赖人工标注，能够充分利用大量易获取的未标注数据
2. **任务驱动的特征学习**
   - 通过设计巧妙的预训练任务(如掩码语言建模、图像补全、对比学习等)，模型被迫学习数据的内在结构和语义关系，加强对数据本质的理解
3. **普遍存在的数据冗余**
   - 现实数据(如文本、图像、视频)往往存在冗余，可以提供丰富的自监督信号来源
4. **提升泛化能力避免过拟合**
   - 自监督预训练使模型具备通用特征表示能力，这些特征可通过微调迁移到下游任务
   - 预训练过程中接触的多样化数据增强了模型对新任务、少样本场景的适应能力

课后思考

基于本节课介绍的“自预测”和“对比学习”两种自监督学习框架，除了课件中提到的任务外，你觉得针对不同的数据模态（图像、音频、文本… …）还可以设计什么样的自监督学习任务？