

第四章：图像表达-深度表征学习

授课老师：李厚强，胡洋，周文罡，李礼

研究背景

- 深度学习已成为学习视觉表征的主流范式
 - 将高维的输入信号转化为低维的特征表达，便于后续任务的使用
 - 深度模型可提取多层次、包含丰富语义信息的特征表达



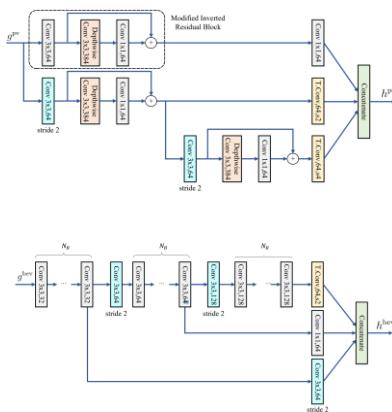
深度表征学习：分类概览

□ 深度表征学习

- 包括：**全监督学习、自监督学习、半监督学习、弱监督学习等**
- 此外还包含一类特殊的**深度生成模型**

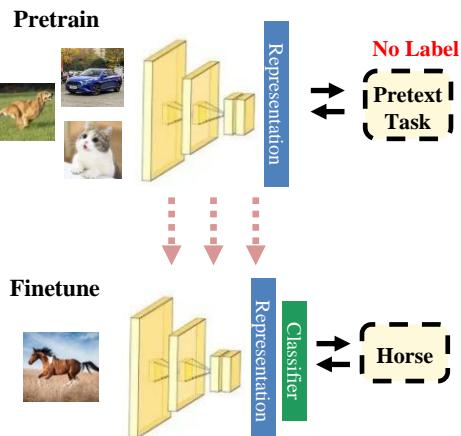
全监督表征学习

侧重于网络模型的设计



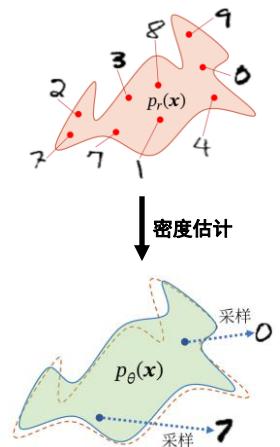
自监督表征学习

侧重于Pretext Task设计



深度生成模型

隐式学习联合概率密度分布

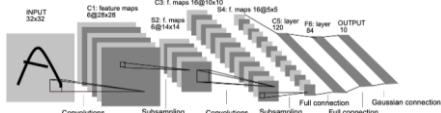


全监督表征学习 I



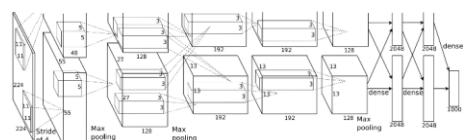
□ 基于CNN的表征学习

卷积神经网络结构设计探索历程



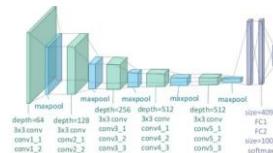
LeNet, 1998

早期利用CNN进行字符识别



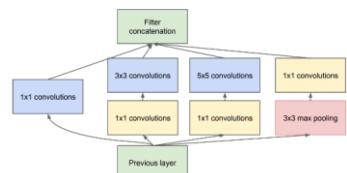
AlexNet, 2012

大规模图像分类任务中展现了CNN的实力



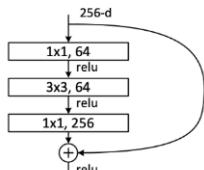
VGG, 2014

19层的CNN网络



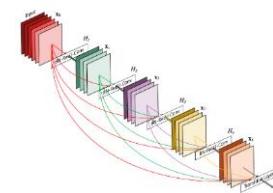
GoogLeNet, 2014

22层的多尺度CNN网络



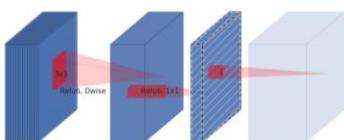
ResNet, 2015

引入short cut机制搭建152层CNN网络



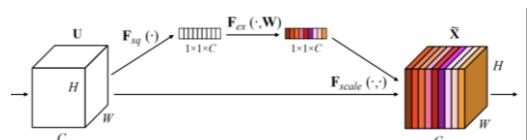
DenseNet, 2017

引入dense connection搭建264层CNN网络

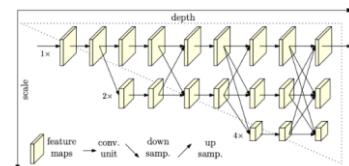


MobileNet, 2017

利用深度可分离卷积设计**轻量化**网络



SENet, 2017
在网络设计中引入注意力机制



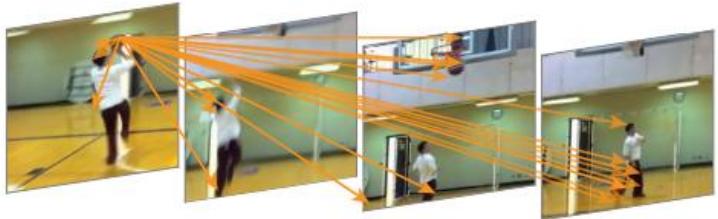
HRNet, 2019

在网络设计中保持高分辨率表征

全监督表征学习 I

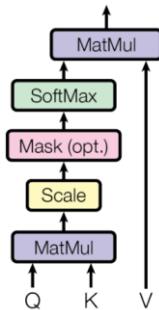
- 基于Transformer的表征学习
 - Transformer的核心技术：注意力机制

什么是注意力机制？



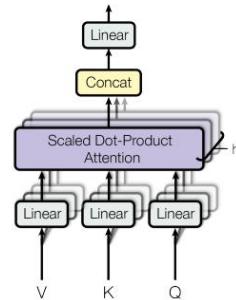
视频动作识别

如何建立注意力机制？



Self-Attention

采用 $\text{Softmax}\left(\frac{QK}{\sqrt{d_k}}\right)V$, 聚合**感兴趣**的信息



Multi-Head Self Attention

采用多头机制，丰富注意力的**多样性**

这头鹿没有横穿公路因为它有些害怕。

文本语义分析

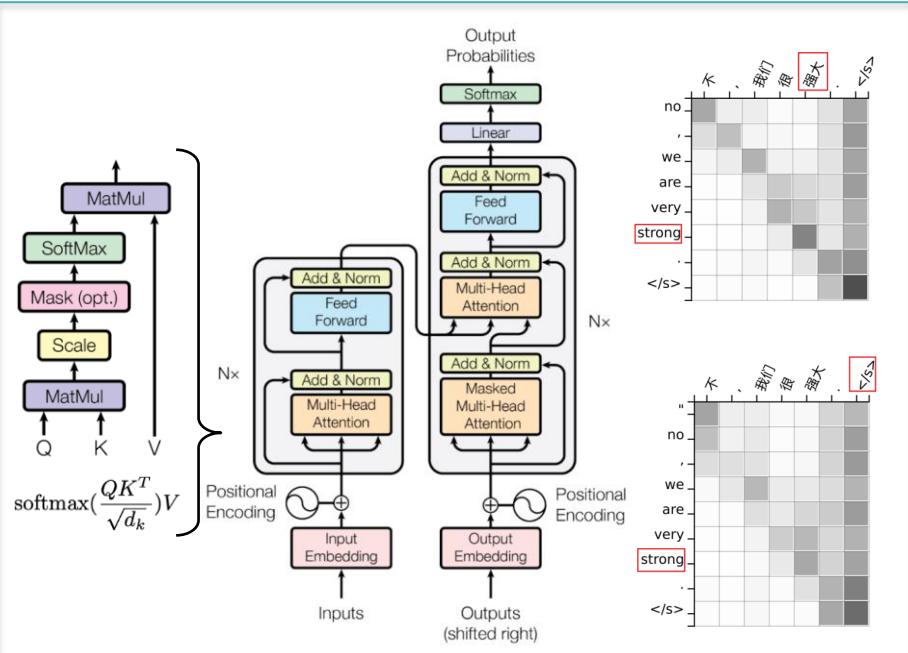
全监督表征学习 I

□ 基于Transformer的表征学习

■ 对文本和图像的表征学习统一为Transformer模型

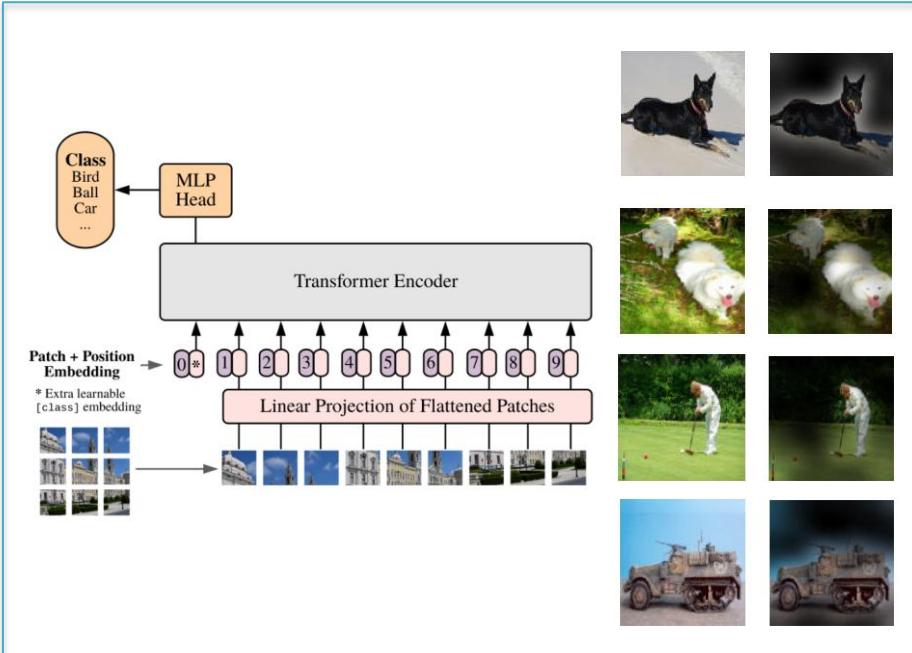
Transformer, 2017

利用全注意力机制网络处理NLP任务



Vision Transformer, 2020

首次在大规模图像分类中使用Transformer结构



自监督表征学习

□ 自监督学习的研究动机

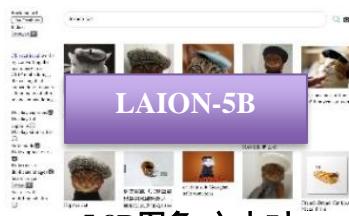
- 深度神经网络参数规模大，依赖大规模数据进行训练优化
- 监督学习范式**依赖大规模标注数据，费时费力**



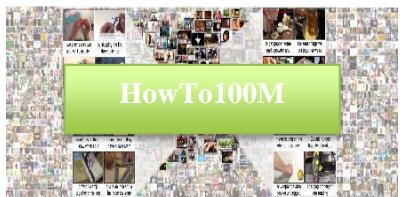
15M图象, 22K类



1.9M图象, 16M边框



5.8B图象-文本对



136M视频片段

大规模的图像视频数据库

人工标注



**“有多少智能，
就要有多少人工”**

监督学习

视觉任务：分类, 检测, ...

深度
表征
学习

高层特征

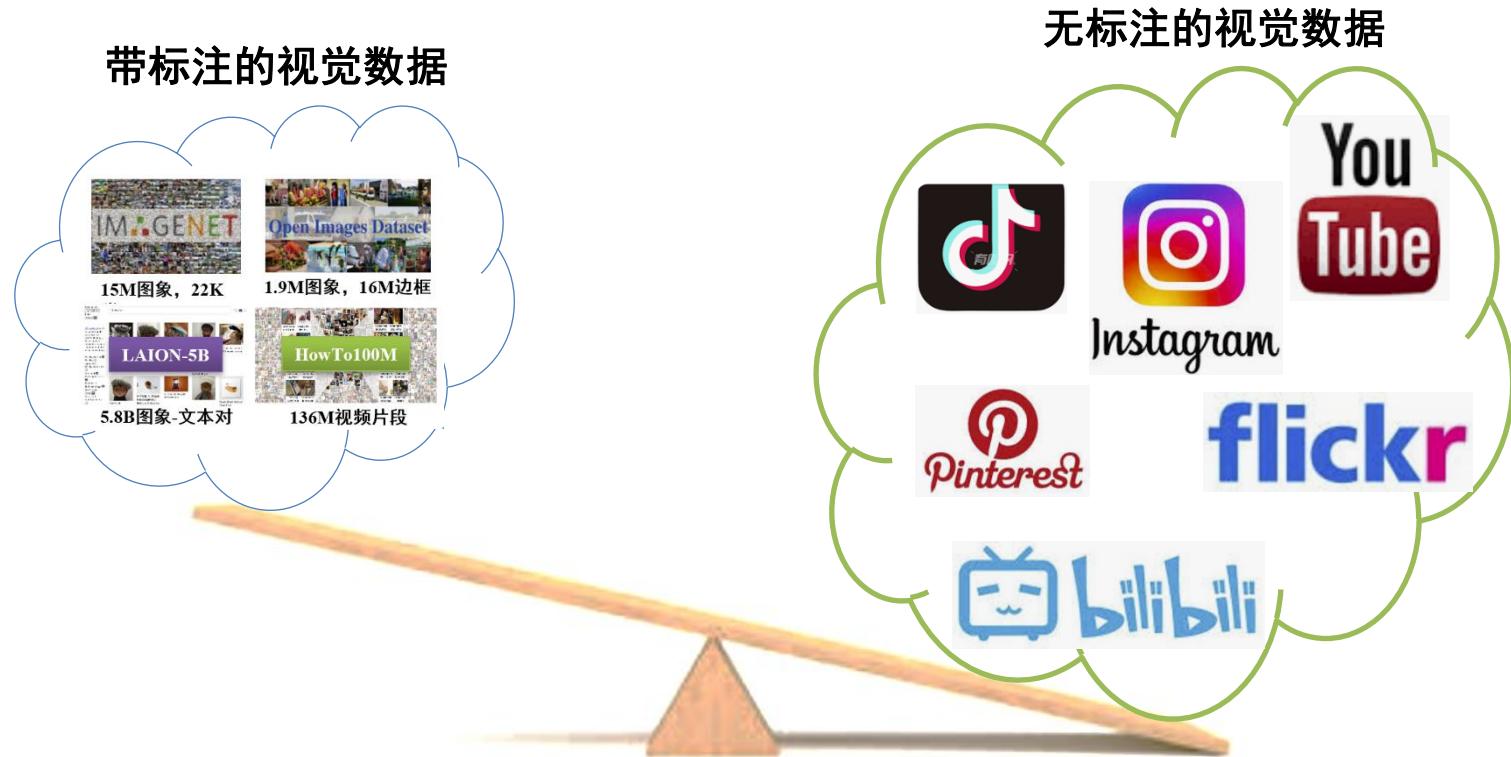
中层特征

底层特征

自监督表征学习

□ 自监督学习的研究动机

- 深度神经网络参数规模大，依赖大规模数据进行训练优化
- 无标注的图象/视频数据在互联网上唾手可得



科学问题：如何设计代理任务，自监督地高效训练网络模型？

自监督表征学习

□ 典型的视觉自监督表征学习

■ 基于图像变换的方法

- ✓ 代理任务常为图像变换：显式或者隐式地预测具体的变换形式

■ 基于对比学习的方法

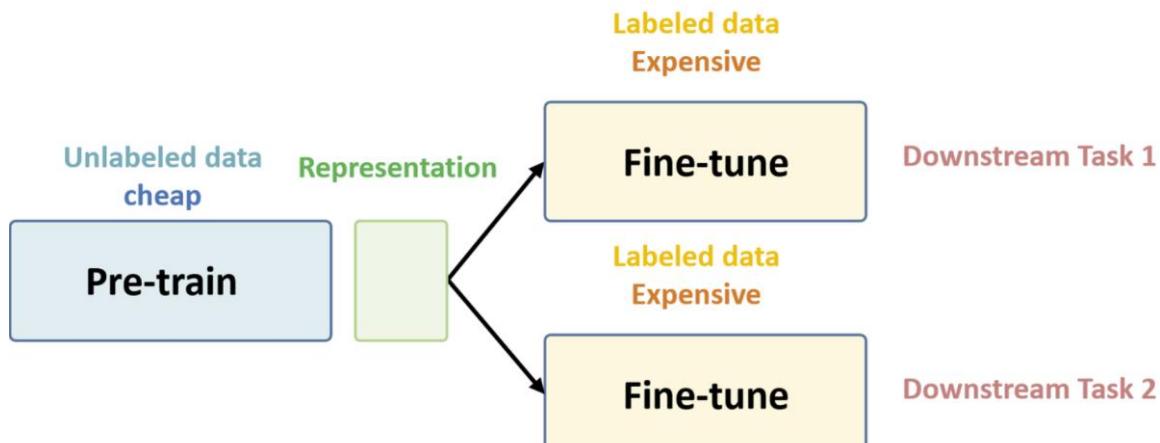
- ✓ 构建正负样本对，在语义空间缩小正样本对的距离，扩大负样本对的距离

■ 基于时序回路一致性的方法

- ✓ 利用视频帧的前向和后向一致性约束，学习视觉表征

■ 基于生成学习的方法

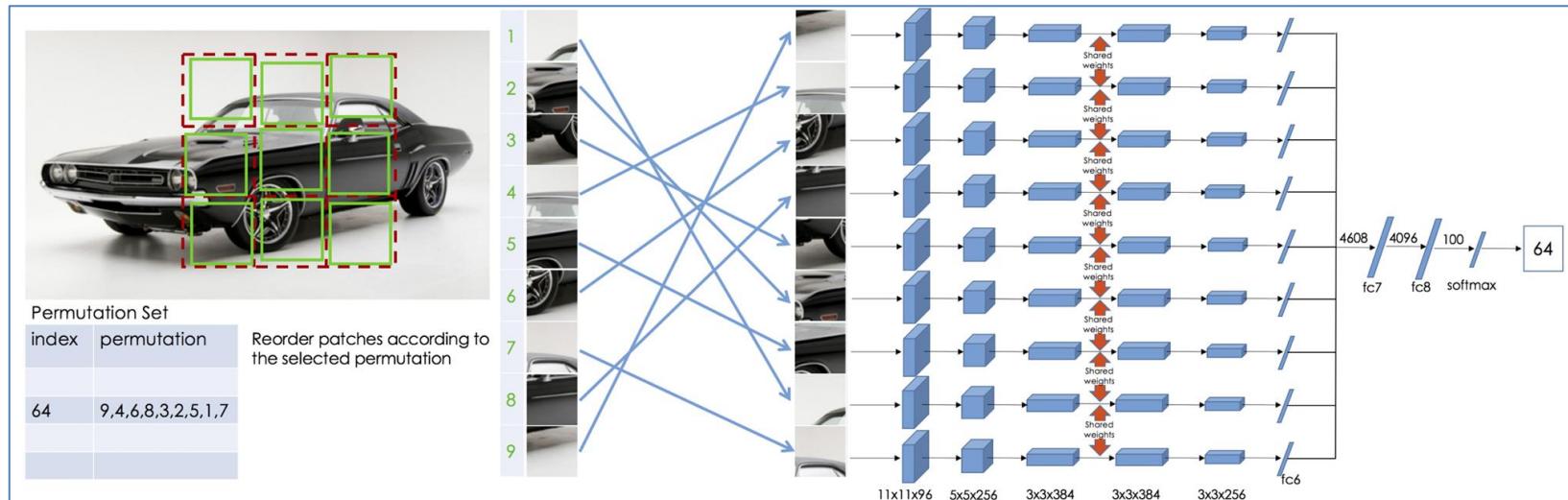
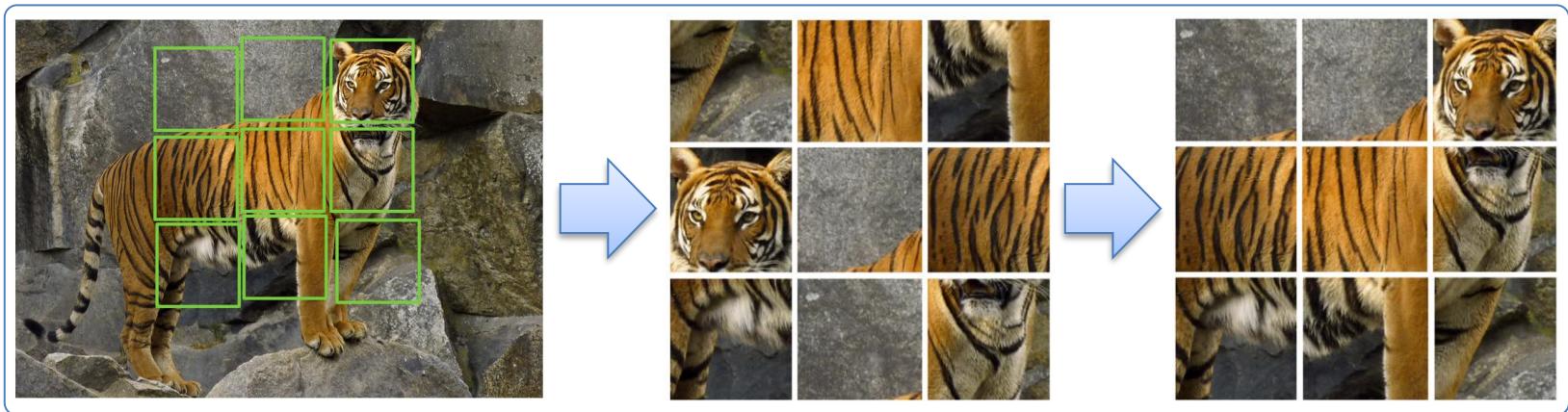
- ✓ 通过掩码重建，学习视觉表征



基于图像变换的自监督表征学习

□ Jigsaw拼图重排 (ECCV 2016)

- 通过求解拼图游戏，学习CNN上下文中对象的视觉空间表示



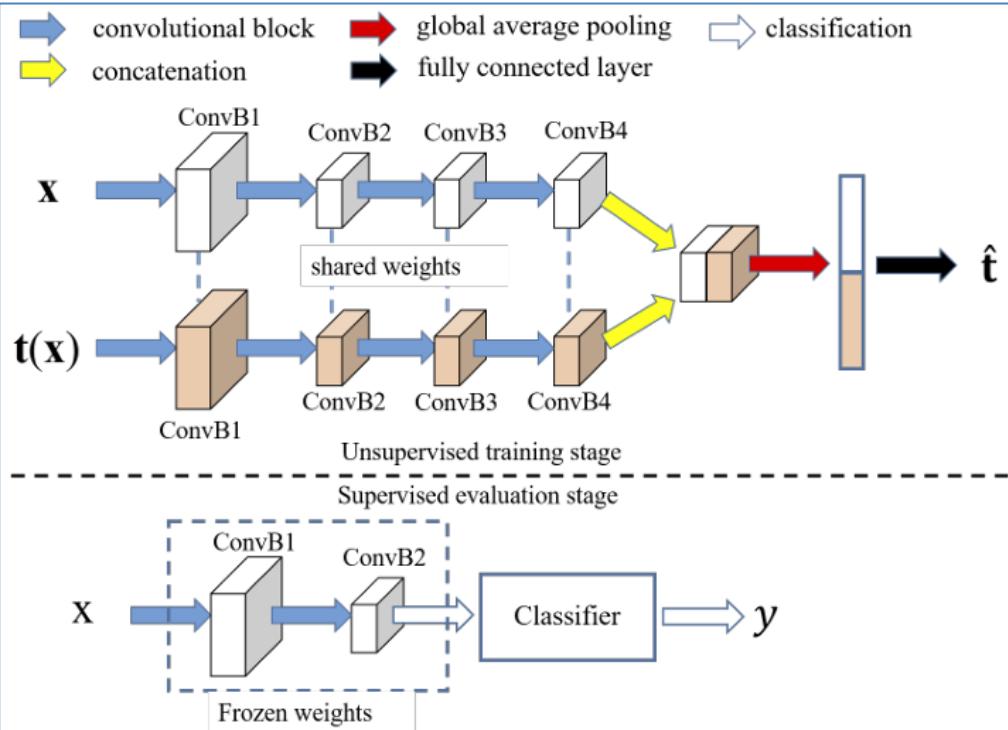
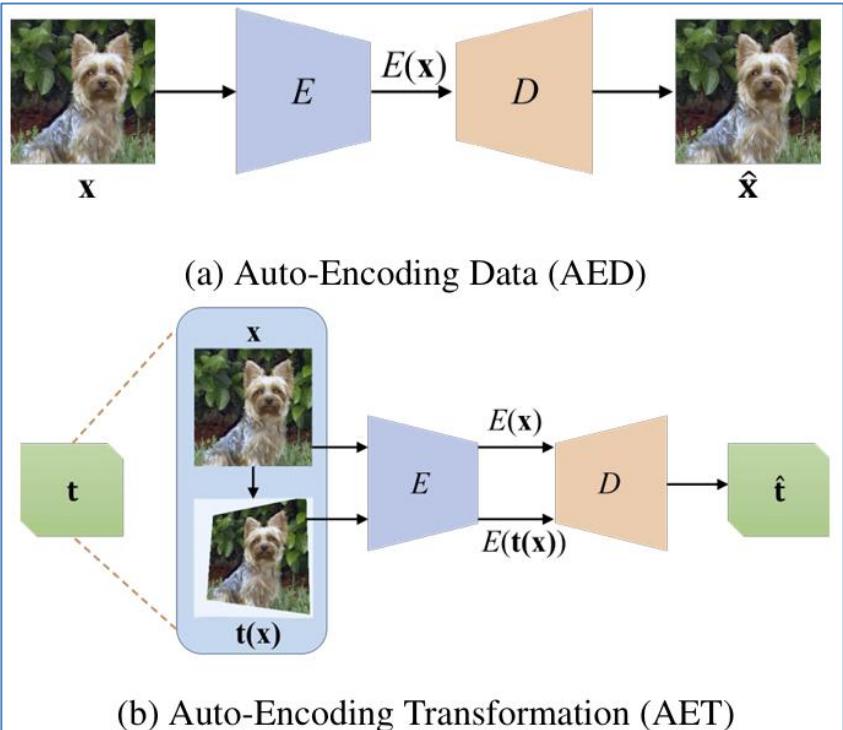
基于图像变换的自监督表征学习

□ AET: 预测图像变换算子 (CVPR 2019)

- 采样一些参数化算子来变换图像，通过训练自动编码器，学习原始图像和转换图像的特征表示，从而预测这些参数化算子

$$\hat{\mathbf{t}} = D [E(\mathbf{x}), E(\mathbf{t}(\mathbf{x}))]$$

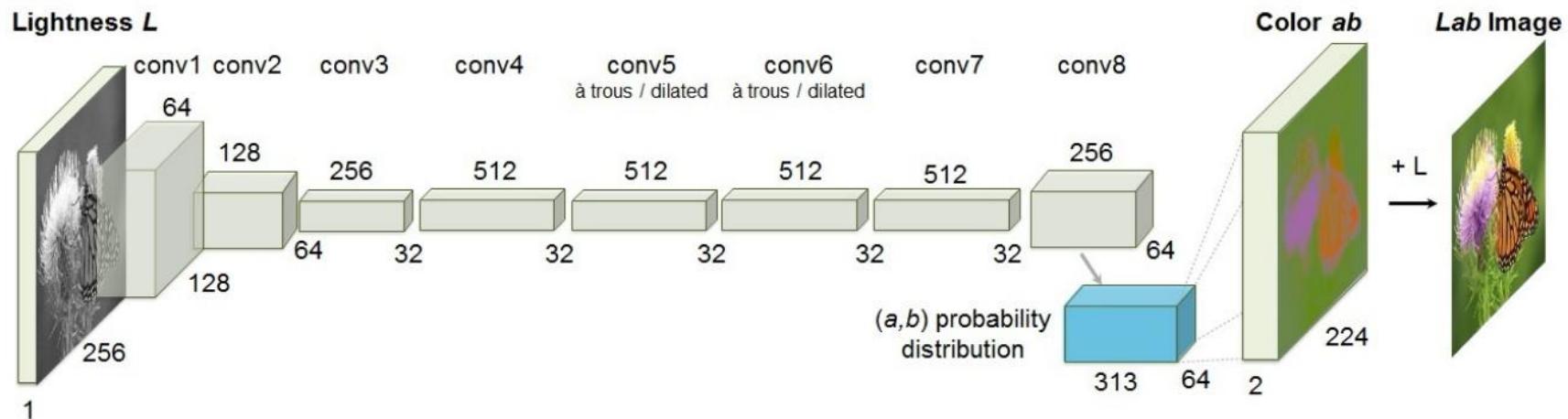
$$\min_{E,D} \mathbb{E}_{\mathbf{t} \sim \mathcal{T}, \mathbf{x} \sim \mathcal{X}} \ell(\mathbf{t}, \hat{\mathbf{t}})$$



基于图像变换的自监督表征学习

□ Colorization: 灰度图像上色

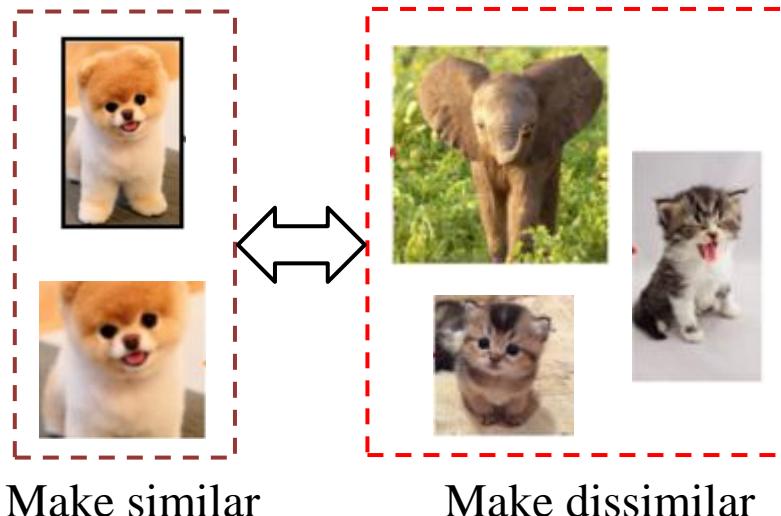
- 将彩色图像灰度化，得到“灰度图像-彩色图像”的样本对
- 构建CNN模型，学习从灰度图像到彩色图像的映射
 - ✓ L-a-b颜色模型：从L通道到a-b通道的映射



基于对比学习的自监督表征学习

□ 基本思想

- 通过对不同样本之间的相似性和差异性来学习数据的有效表征
- ## □ 不依赖于大量标注数据，而是通过构造正样本对（相似样本）和负样本对（不相似样本）的方式来训练模型
- 通过构建正负样本对，在语义空间里缩小正样本对之间的距离，扩大负样本对之间的距离，从而使得编码器学习到对相似样本的一致性表征



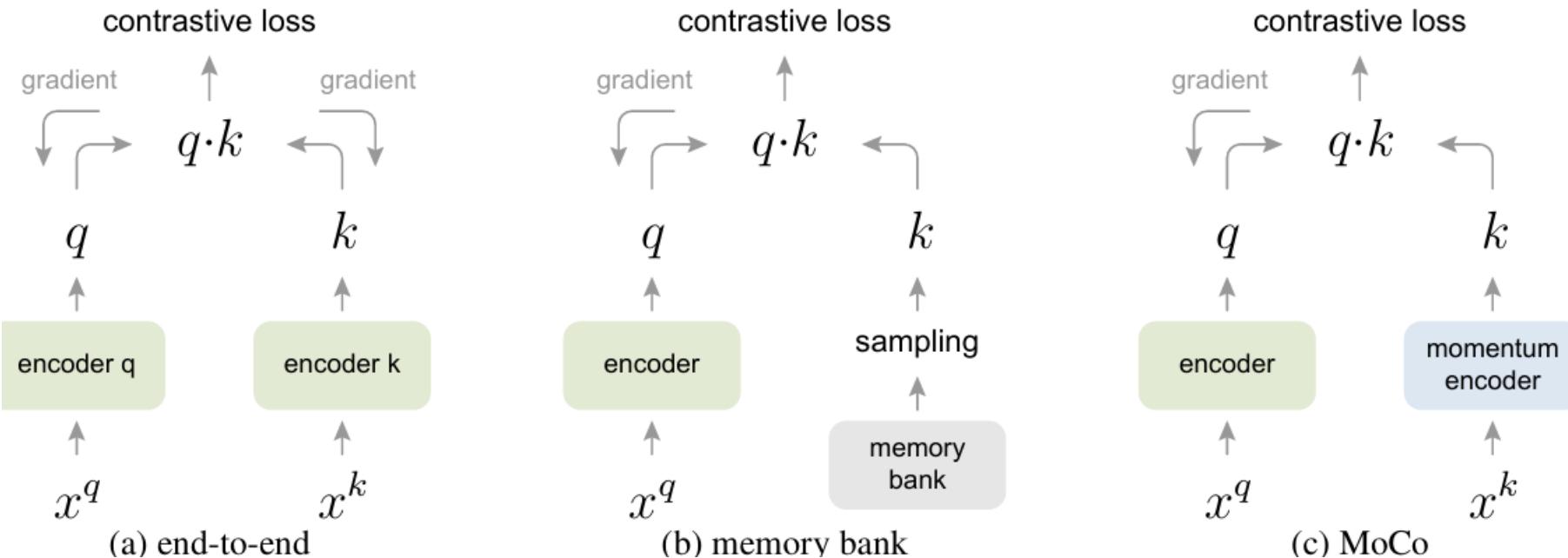
基于对比学习的自监督表征学习

□ 动量(MOCO, Momentum Contrast)对比学习

- 对比损失： q 与 k_+ 相似而与所有负样本不相似时，则对比损失越小

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^K \exp(q \cdot k_i / \tau)}$$

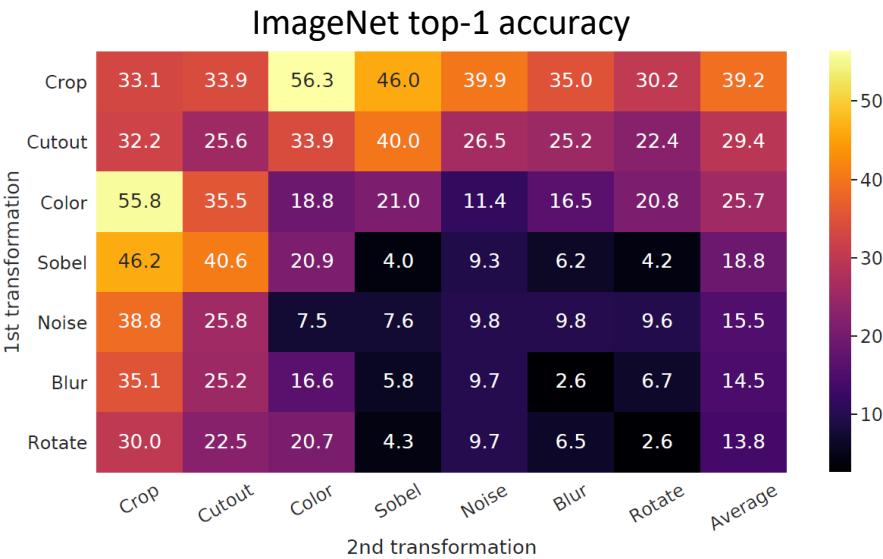
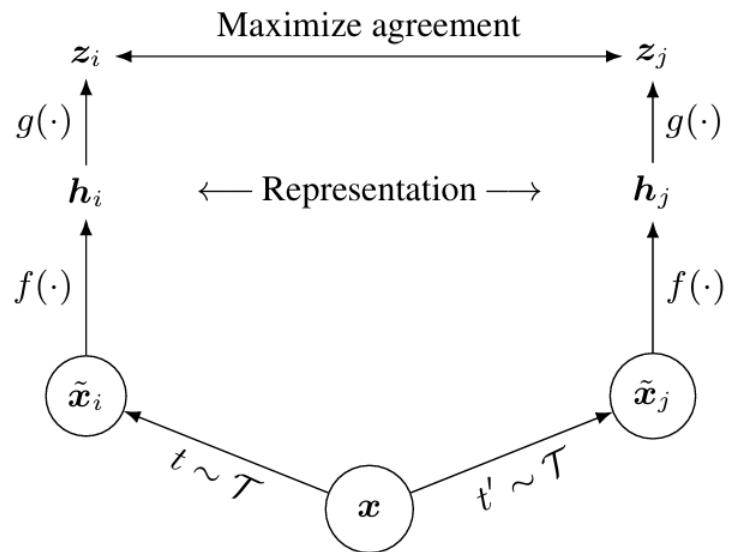
- 动量编码器：对模型参数移动加权平均 $\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$



基于对比学习的自监督表征学习

□ SimCLR:

- 给定输入图像 x , 通过不同的数据增强方式 \mathcal{T} , 得到两幅图像 \tilde{x}_i 和 \tilde{x}_j
 - ✓ \mathcal{T} : 随机裁剪再resize到原来尺寸, 随机色彩失真, 随机高斯模糊
- 将 \tilde{x}_i 和 \tilde{x}_j 输入到共享参数的编码器 $f(\cdot)$, 得到视觉表征 h_i 和 h_j
- 视觉表征 h_i 和 h_j 在经过project head得到 z_i 和 z_j , 做对比学习优化

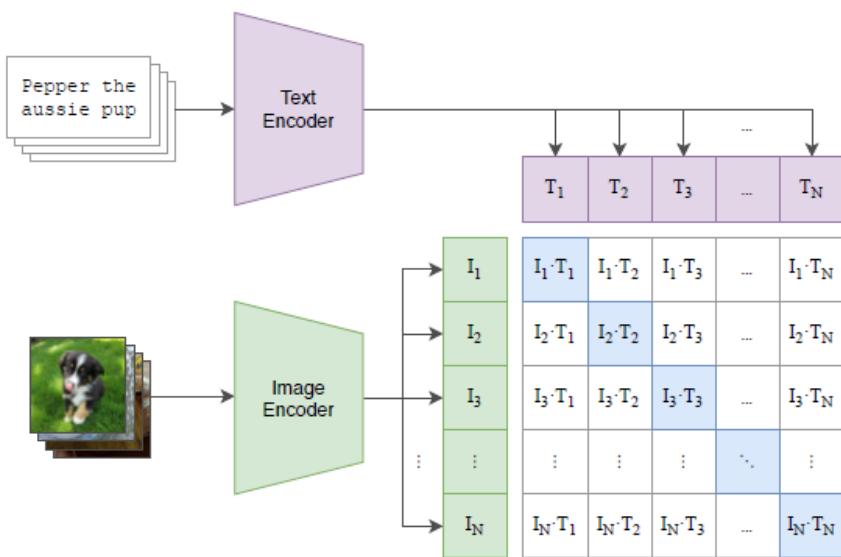


基于对比学习的跨模态表征学习

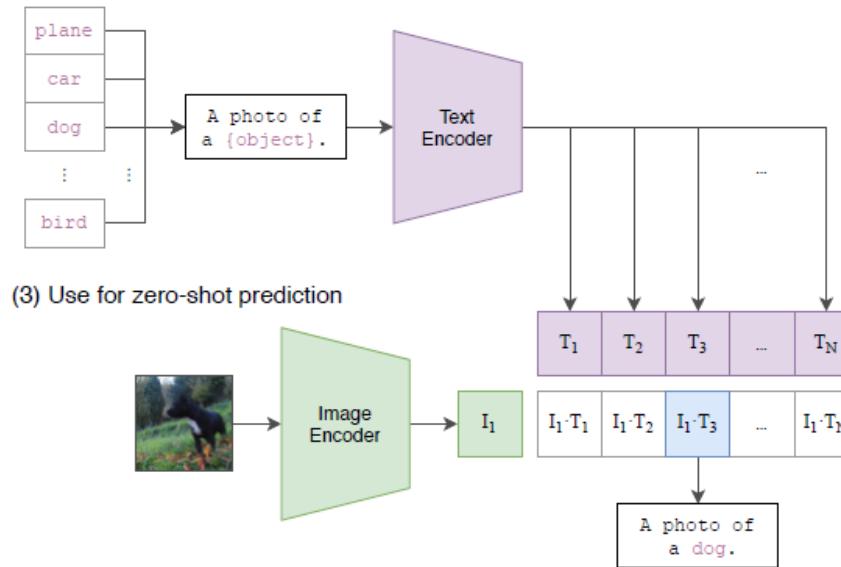
□ CLIP: Contrastive Language Image Pre-training

- 收集大规模（400M）“图像-文本句子描述”样本，通过对比学习，训练文本编码器和图像编码器，**对齐**视觉表征和语言表征
- 自然语言（句子）监督：提供了更广泛、更详细的描述，涵盖了更多的视觉概念
- 学习得到的文本/视觉编码器具有良好的零样本迁移能力

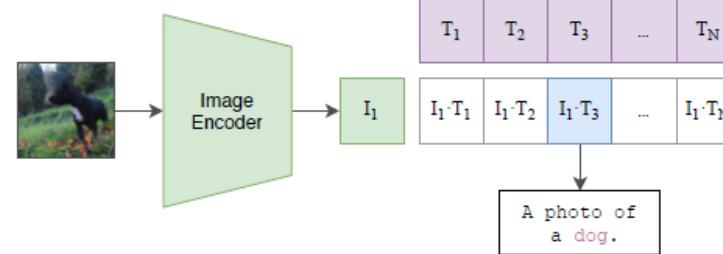
(1) Contrastive pre-training



(2) Create dataset classifier from label text



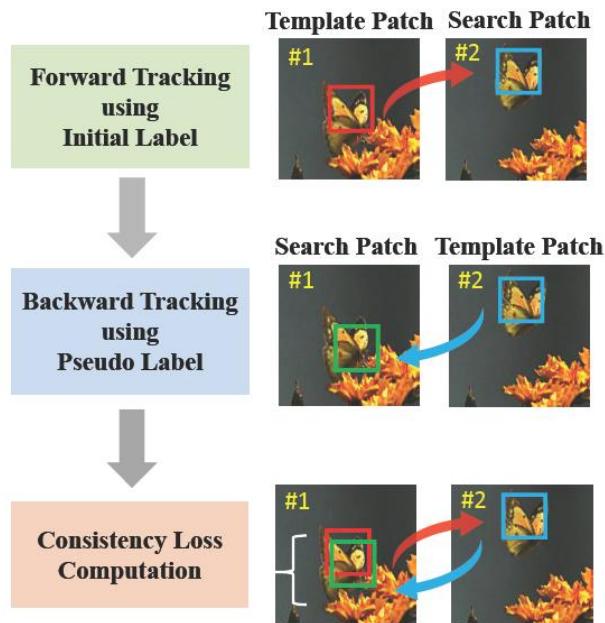
(3) Use for zero-shot prediction



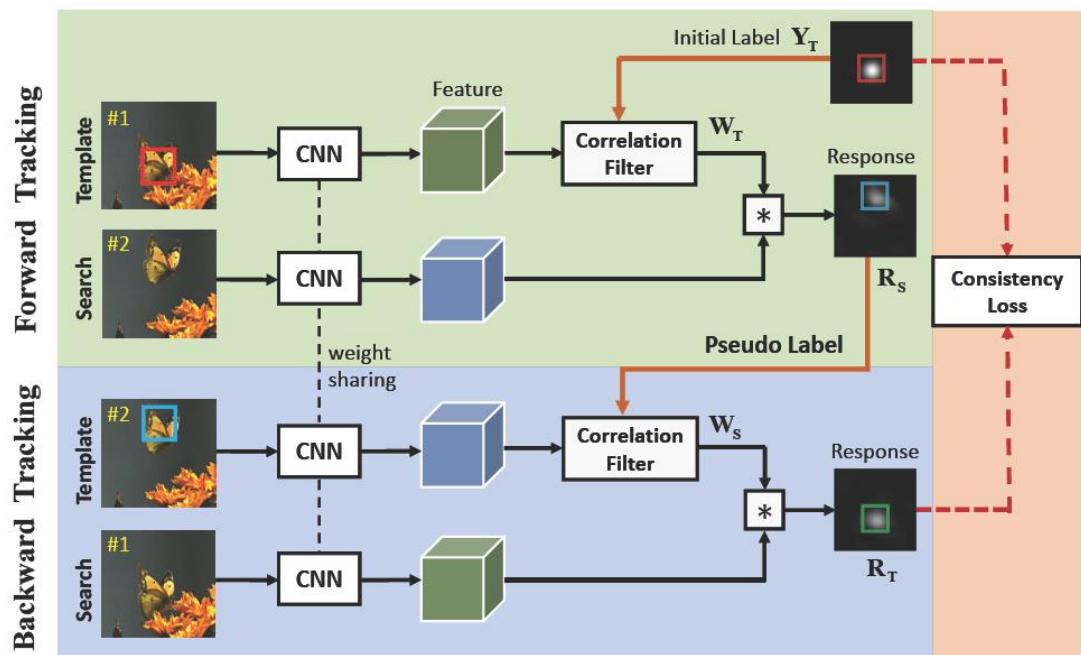
基于回路一致性的自监督表征学习

□ 双向目标跟踪的一致性优化学习

- 利用物体可双向跟踪的特点，随意生成初始标签进行前向跟踪和后向跟踪
- 提出了前向后向轨迹一致性的无监督训练，并扩展到多帧、多轨迹情况
- **无需人工标注**，取得了**媲美经典全监督学习方法**的视频跟踪性能



(a) Unsupervised Learning Motivation

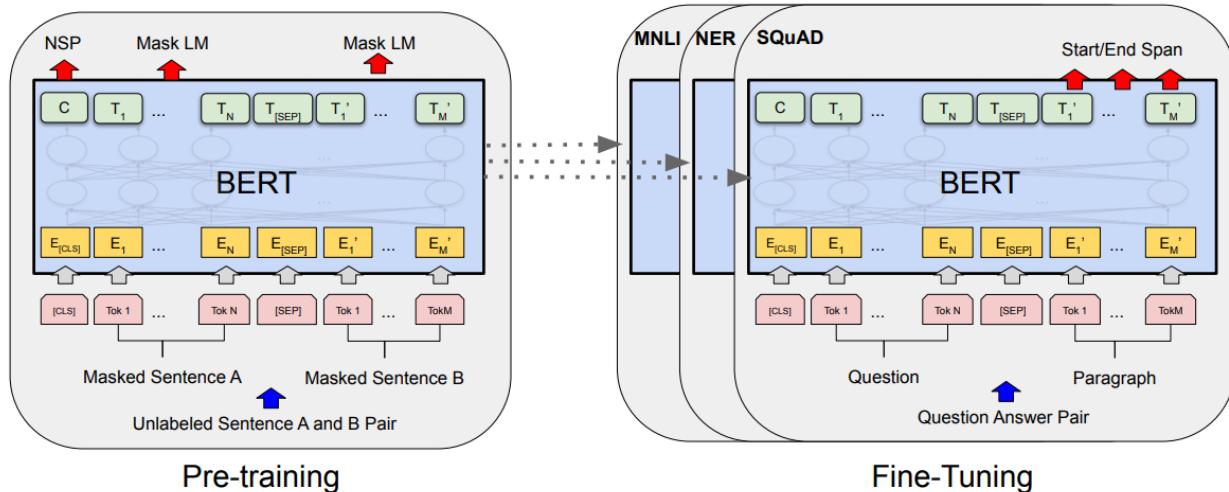


(b) Unsupervised Learning Pipeline using a Siamese Network

基于生成式的自监督表征学习

□ 基于遮罩建模的生成式方法

- 通常为对被遮罩输入标识的重建任务，使得网络从剩余标识中捕获用于重构的上下文信息，进而最大化输入领域信息的条件概率分布
- BERT预训练
 - ✓ 掩码语言模型 (MLM)：完形填空
 - 随机掩盖输入中的一些词元，预测被掩盖的词元
 - ✓ 下一句预测 (NSP)：判断两个句子是否连续，学习句子间关系

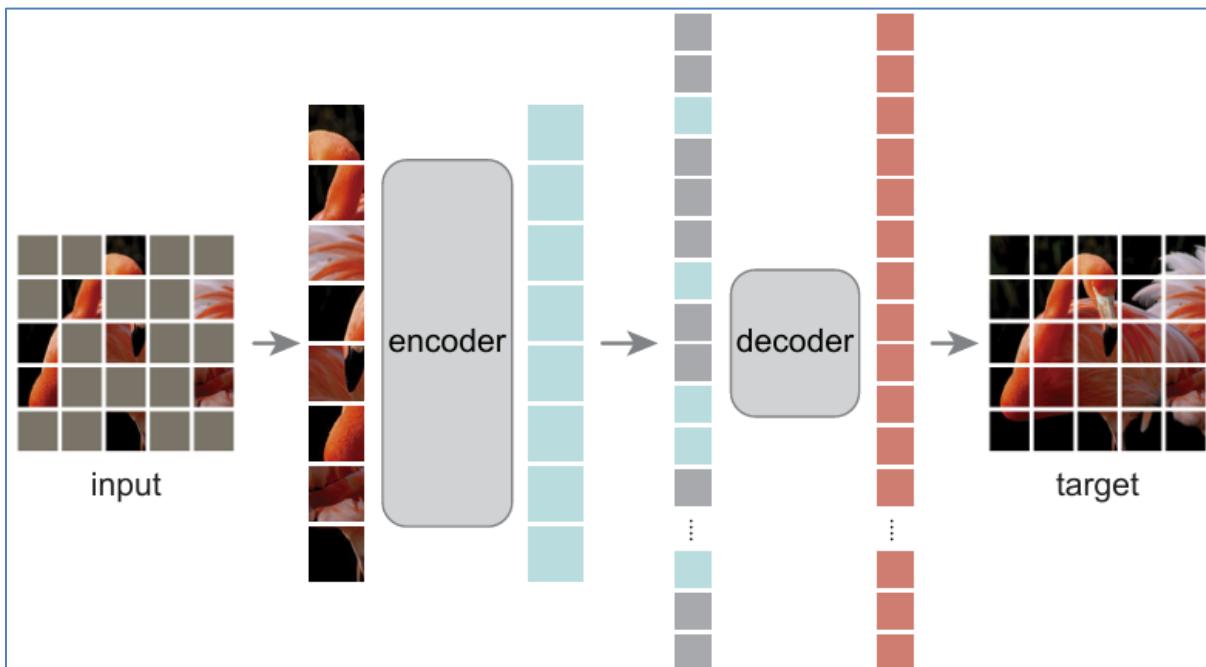


BERT: 自然语言处理领域典型方法

基于生成式的自监督表征学习

□ MAE: Masked autoencoders

- 随机mask掉图片中的一些patch，通过自编码器去重建缺失的patch
- 非对称编码-解码结构
 - ✓ 在编码器阶段，仅将未被mask掉的图片patch作为输入
 - ✓ 在解码器阶段，将编码器输出的隐变量和mask token共同作为输入，去重建完成的图片



基于生成式的自监督表征学习

□ MAE: Masked autoencoders

- 随机mask掉图片中的一些patch，通过自编码器去重建缺失的patch
- 非对称编码-解码结构
 - ✓ 在编码器阶段，仅将未被mask掉的图片patch作为输入
 - ✓ 在解码器阶段，将编码器输出的隐变量和mask token共同作为输入，去重建完成的图片

