【专题11:用自监督学习方法解决计算机视觉问题】1.概述和DINOv2

爱科技勤折腾的苹果二 2023-7-1

讨论内容

0 计算机视觉存在的问题

1 用自监督学习代替监督学习

1.1 计算机视觉中的自监督学习、Pretext任务、下游任务、实现步骤

1.2 自监督学习的类型

1.3 自监督学习在计算机视觉应用的常用算法模型

1.4 自监督学习的缺点

1.5 平衡自监督学习的训练效率和模型泛化能力的方法

2 DINO与DINOv2原理

3 使用DINOv2的方法

3.1 模块说明

3.2 实验检查单

4 参考文献

0 计算机视觉：看得清、对和懂

0 计算机视觉存在怎样的问题?

- 有标记的样本太少，如何有效地利用未标记的数据进行预训练，从而提高模型的泛化能力和性能?
- 如何从未标记的数据学习视觉特征?
- 如何评估从未标记数据学到的视觉特征质量?

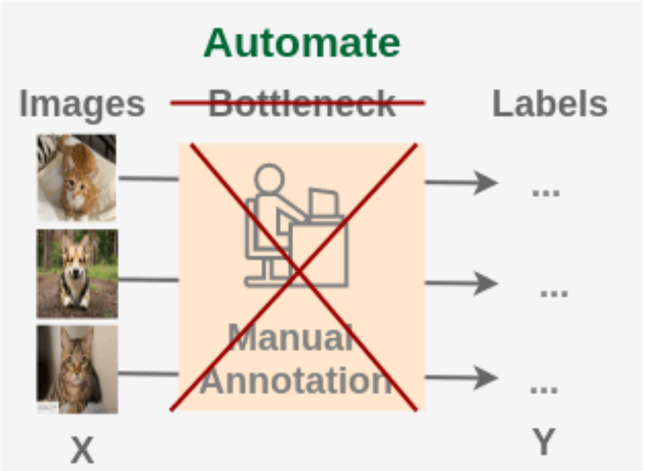
1 用自监督学习代替监督学习

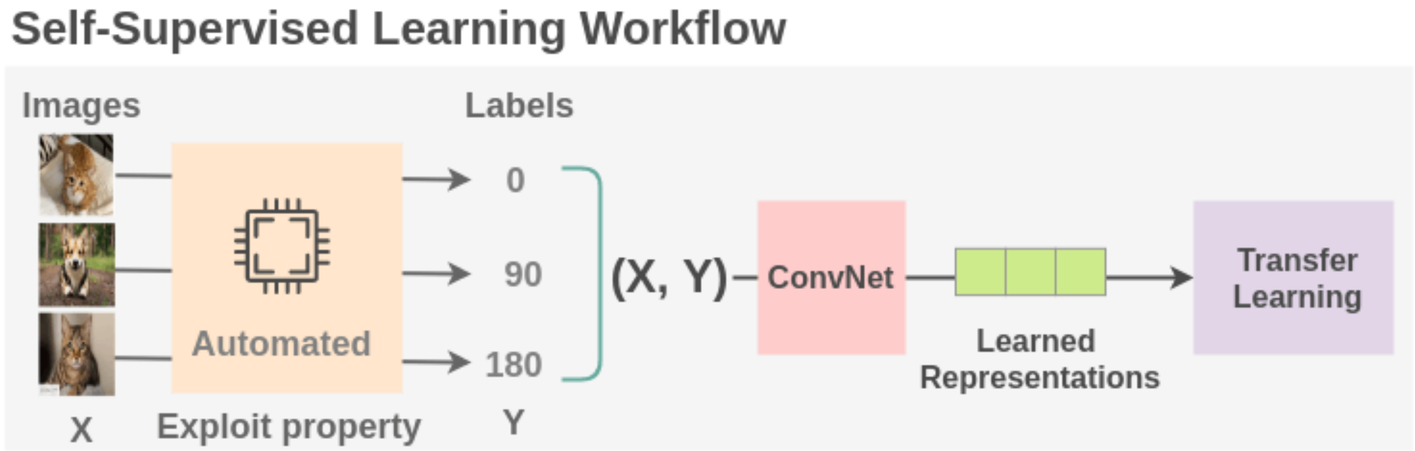
自监督学习相关视频

- https://www.bilibili.com/video/BV1iP41127jL 【专题7:生产管理之异常检测】1 异常检测案例及自监督学习

用自监督学习代替监督学习

我们能否以这样的方式设计任务，即我们可以从现有的图像中生成几乎无限的标签，并利用这些标签来学习表征?



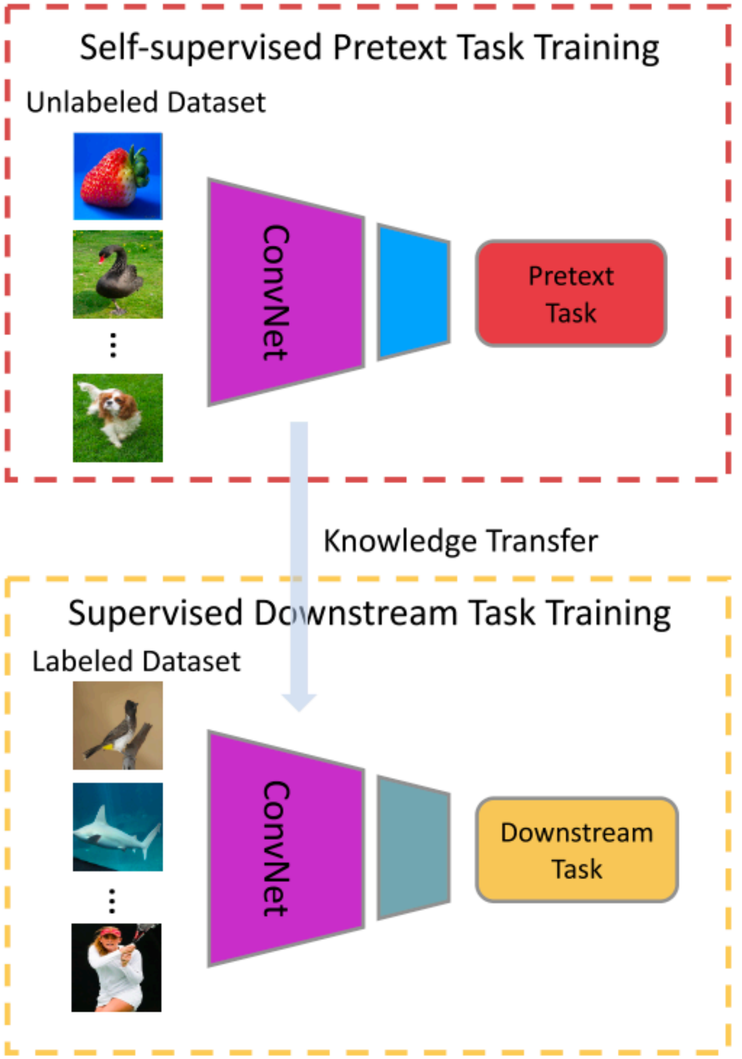


通过创造性地利用数据的某些属性来设置一个伪监督任务，从而取代人类的标签。

图片来源:https://amitness.com/2020/02/illustrated-self-supervised-learning/

自监督学习的Pretex和Downstream任务

- 目标：在没有足够的带标记的数据样本情况下，完成任何图像分类、语义分割、对象检测、动作识别等任务。【下游任务：Downstream Task】
- 基础工作（Pretext Task）：学习到一般的特征
  - 学习视觉表征：从无监督的未标注数据中挖掘自身的监督信息，进而学习到对下游任务有价值的表征；
  - 在每个Pretext任务中，都有部分可见和部分隐藏的数据，任务目标：预测隐藏数据或隐藏数据的某些属性。



图片来源：Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey https://arxiv.org/pdf/1902.06162.pdf

1.1.2 Pretext任务：从学习内容分

- **学习图像的基本特征和结构**
  - 图像重构：通过对输入图像进行某种形式的变换（如旋转、缩放、裁剪等），然后训练模型去预测原始图像。
- **学习图像的上下文信息**
  - 图像填充：将输入图像的一部分遮挡或移除，然后训练模型去预测被遮挡或移除的部分。
- **学习图像的不变性特征**
  - 对比学习：从输入图像中生成两个或多个变换版本，然后训练模型去判断它们是否来自同一原始图像。

1.1.3 Pretext任务：从数据属性分

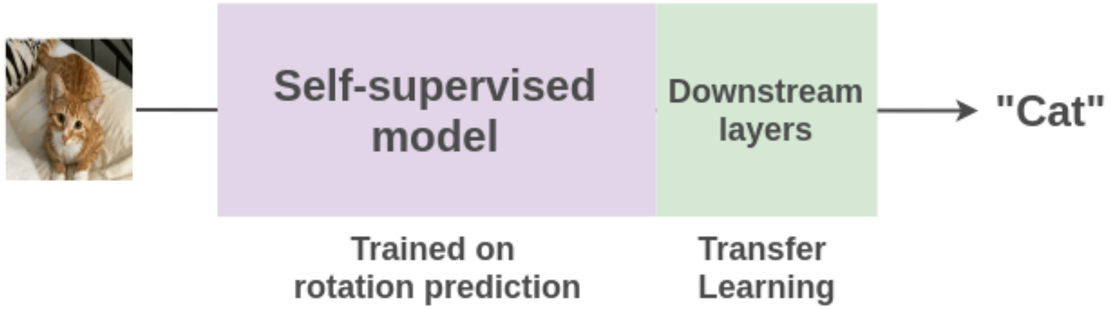
- **基于生成**：通过解决图像或视频生成任务学习视觉特征，如图像着色、图像补全、视频彩色化和视频预测等。
- **基于上下文**：基于上下文的空间语境结构或时间语境结构的相似性学习视觉特征，如图像拼图、上下文预测、视频输入的帧序列是否正确、识别帧序列的顺序等。
- **基于自动生成语义标签**：借助传统的硬编码算法或视频游戏引擎自动预测初始语义标签，常用标签预测任务包括物体分割、轮廓检测、深度预测、法向估计等。
- **基于跨模态**：通过验证两个不同通道的输入数据是否相互对应进行网络训练，如视觉与听觉的对应性验证、RGB (Red, Green and Blue)与光流对应性验证，物体和场景与自我运动的响应验证等。

1.1.4 下游任务

- 通常，只将预训练模型的前几层特征迁移到下游任务中去。
- 下游任务基于自监督学习学到的预训练模型，进行微调，可以从图像或视频中捕捉**高级**视觉特征。
- 评估任务：
  - 学习的图像特征的泛化能力：图像分类、语义分割和对象检测
  - 学习的视频特征的质量：视频中的人类动作识别

1.1.5 自监督学习的步骤

- 根据对数据的理解，以编程方式从未标注的数据中生成输入数据和标签。
- 预训练：用上一步的数据/标签来训练模型。
- 微调：使用预训练的模型作为初始权重，对感兴趣的任务进行训练。



图片来源：https://amitness.com/2020/02/illustrated-self-supervised-learning/

1.2.1 自监督学习的类型

类型	学习的目标	代表算法模型
基于生成模型的方法	学习数据的真实分布，以便生成新的、与真实数据相似的样本	自编码器（AutoEncoder）、变分自编码器（Variational AutoEncoder，VAE）和生成对抗网络（Generative Adversarial Network，GAN
基于预测模型的方法	预测未标记数据的某些属性或特征	自回归模型（AutoRegressive Model）和自监督对比学习（Self-Supervised Contrastive Learning）
基于重构模型的方法	通过重构输入数据来学习其内在的表示	去噪自编码器（Denoising AutoEncoder）和稀疏自编码器（Sparse AutoEncoder）

1.2.2 自监督学习的类型

A Cookbook of Self-Supervised Learning

Randall Balestriero*, Mark Ibrahim*, *Vlad Sobal**, *Ari Morcos**, *Shashank Shekhar**, *Tom Goldstein*†, *Florian Bordes**‡, *Adrien Bardes**, *Gregoire Mialon**, *Yuandong Tian**, *Avi Schwarzschild*†, *Andrew Gordon Wilson***, *Jonas Geiping*†, *Quentin Garrido**§, *Pierre Fernandez***, *Amir Bar**, *Hamed Pirsiavash*+, *Yann LeCun** and *Micah Goldblum***

*Meta AI, FAIR
**New York University
†University of Maryland
+University of California, Davis
‡Universite de Montreal, Mila
§Univ Gustave Eiffel, CNRS, LIGM
*Univ. Rennes, Inria, CNRS, IRISA
italicEqual contributions, randomized ordering

图片来源：https://arxiv.org/pdf/2304.12210.pdf

对比学习的特点

- **对比学习**：让模型学习如何区分不同的图像或图像的不同部分。通过比较和区分数据样本学习数据的有用表示，模型被训练来识别来自同一数据源的样本（正样本）并区分来自不同数据源的样本（负样本）。
- 缺点：忽视了对语境表征的学习。
- 优点：模型可以学习到数据的内在结构和特性。
- 核心要点
  - 如何构造样本Pair?
    - 采用图像增强方式：随机裁剪、随机噪音、高斯模糊、抖动、颜色通道转化、灰度化、对比度调节、亮度调节
  - 如何设计Loss

对比学习算法模型比较

- SimCLR（Simple Contrastive Learning of Visual Representations）：通过比较来自同一图像的两个增强版本（正样本）并区分来自不同图像的样本（负样本）来学习视觉表示。
- MoCo（Momentum Contrast）：使用动量编码器和一个大的队列来存储和更新负样本，从而提高对比学习的效果。
- BYOL（Bootstrap Your Own Latent）：不需要负样本，通过比较同一图像的两个增强版本的表示来学习视觉表示。
- SwAV（Swapping Assignments between multiple Views of the same image）：一种通过交换聚类分配来进行自监督学习的方法，可以在**没有负样本**的情况下进行有效的对比学习。
- InfoNCE（Information Noise Contrastive Estimation）：基于信息理论，通过最大化正样本对和负样本对之间的互信息来学习数据表示。

1.3.1 自监督学习的应用

- 从图像中进行自监督学习
  - **重构**：图像着色、图像超分辨率、图像修复、跨渠道预测
  - **常识性任务**：图像拼图、语境预测、几何变换识别
  - **自动标签生成**：图像聚类、合成图像
- 从视频中进行自监督学习
  - **帧顺序验证**

1.3.2 自监督学习在计算机视觉应用的常用算法模型

任务类型	任务内容	准备的训练对	常见算法
图像重构	图像着色Image Colorization	（灰度化、彩色化）	GAN、Autoencoders、U-Net、注意力机制（Attention Mechanisms）
	提高图像分辨率Image Superresolution	（低分辨率、高分辨率）	SRGAN、SRCNN、Autoencoder
	图像修复Image Inpainting	（损坏的、正常的）	Contextual Attention GANs和Generative Multi-Adversarial Networks (GMAN) Partial Convolutional Autoencoders，Context Encoders（CE）
	跨通道预测Cross-Channel Prediction	（输入的、预测的）	GAN、Autoencoder、Transformer、 Colorful Image Colorization、Pix2Pix、Split-Brain Autoencoder、Vision Transformer (ViT)
图像常识性任务	图像拼图Image Jigsaw Puzzle	（洗牌的，有序的）	Jigsaw Puzzle Solver Context-Free Network (CFN) DeepCluster

1.3.3 自监督学习在计算机视觉应用的常用算法模型（续）

任务类型	任务名称	准备的训练对	常用算法模型
图像常识性任务	语境预测Context Prediction：理解图像的上下文关系	（图像补丁，邻居）	Context Encoders,Jigsaw Puzzles,Colorization,Relative Position Prediction
	几何变换识别Geometric Transformation Recognition:理解图像的空间结构和变换	（旋转的图像，旋转角度）	Spatial Transformer Networks,Relative Position Prediction,Unsupervised Learning of Depth and Ego-Motion,DPC-Net
图像：自动生成标签	图像聚类Image Clustering	（图像，聚类数）	DeepCluster,SeLa(Self-labelling),SwAV
	合成图像Synthetic Imagery	（图像，属性）	GAN,VAE,CycleGAN,StyleGAN
视频	帧顺序验证Frame Order Verification	（视频帧，正确/错误的顺序）	Time-Contrastive Networks (TCN),Shuffle and Learn,OPN (Odd-One-Out Networks)

特别推荐查论文和代码神器

- https://paperswithcode.com/task/self-supervised-learning

1.4 自监督学习的缺点

- **监督信号的质量**： 因为监督信号是来自于数据本身，而不是由人类注释者明确提供的，监督信号的质量可能低于监督学习。监督信号可能是嘈杂的或不完整的，这可能导致任务的性能降低。
- **限于某些类型的任务**： 自我监督学习对于数据更复杂或非结构化的任务可能不那么有效。
- **训练的复杂性**： 一些自我监督学习技术在实施和训练时可能比监督学习技术更复杂。例如，对比学习和无监督表征学习在实现和调整上可能比监督学习方法更具挑战性。
- **特征的泛化能力**：自监督学习的目标是学习到能够泛化到各种下游任务的特征。然而，如何确保学习到的特征具有良好的泛化能力，而不仅仅是对预训练任务过度拟合，是一个需要解决的问题。
- **训练效率**：通常需要大量的未标记数据进行训练，这可能导致训练效率低下。
- **模型的解释性**：模型通常是黑箱模型，其内部工作机制难以理解。

1.5 如何平衡自监督学习的训练效率和模型的泛化能力?

- **数据增强**：通过对训练数据进行各种形式的变换（如旋转、缩放、裁剪等），可以有效地扩大训练数据集，提高模型的泛化能力，同时也可以提高训练效率。
- **模型正则化**：通过在模型的目标函数中添加正则化项，可以防止模型过拟合，提高模型的泛化能力。常见的正则化方法包括L1正则化、L2正则化和Dropout等。
- **批量训练**：通过将训练数据分成多个批次进行训练，可以有效地提高训练效率。同时，批量训练也可以在一定程度上提高模型的泛化能力，因为它可以减少模型对单个样本的过度依赖。
- **早停策略**：通过在验证集上监控模型的性能，当模型的性能不再提高时，就停止训练。可防止模型过拟合，提高模型的泛化能力，同时也可以提高训练效率。
- **学习率调整**：通过动态调整学习率，可以在训练初期快速收敛，训练后期避免震荡，从而提高训练效率。同时，合适的学习率也可以帮助模型更好地泛化。
- 以上每一种策略和技术都有其特点和适用场景，具体选择哪一种需要根据实际问题和数据情况来决定。

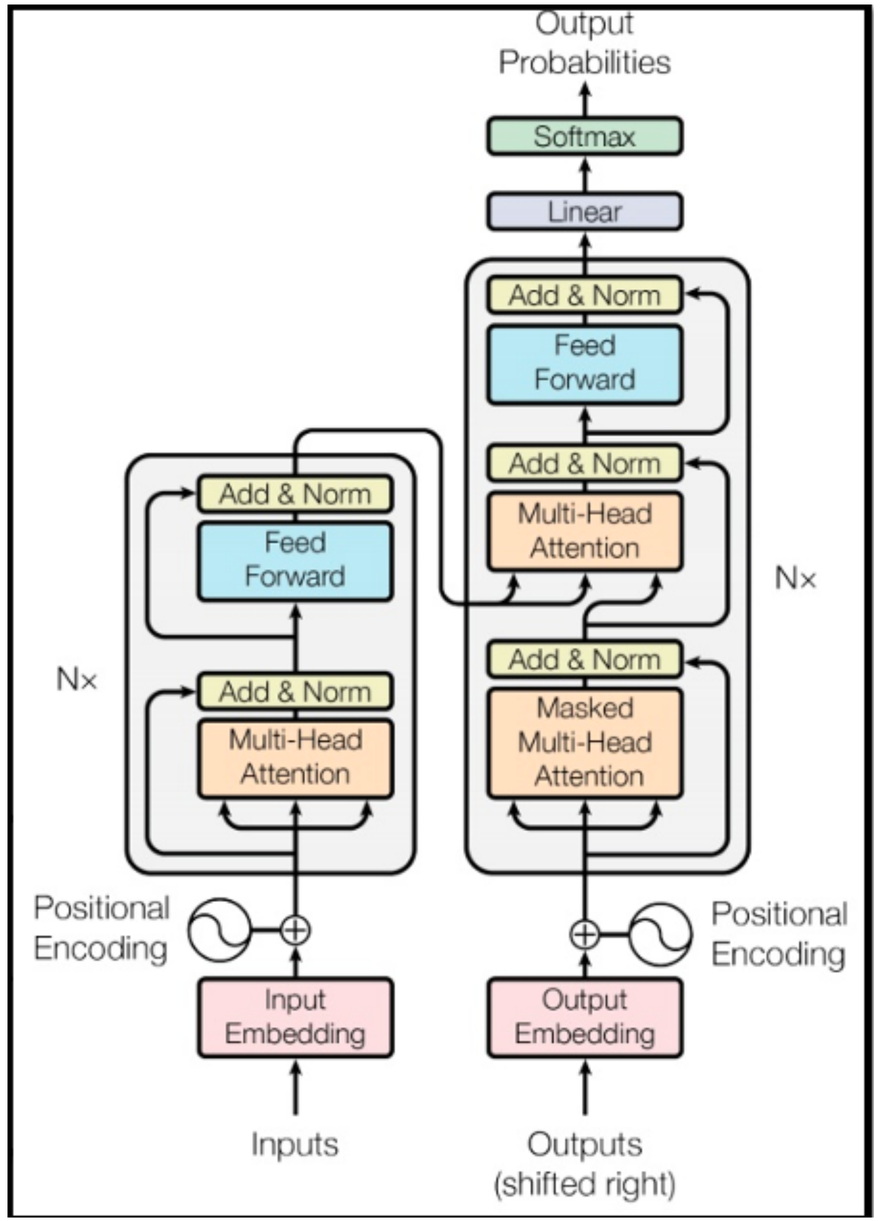
2 无标签知识蒸馏DINO与DINOv2原理

DINOv2的Demo

- DINOV2的Demo网站：https://dinov2.metademolab.com/
- DINO：Facebook AI 的发布的视觉理解领域自监督学习的解决方案。学习丰富的视觉表征。

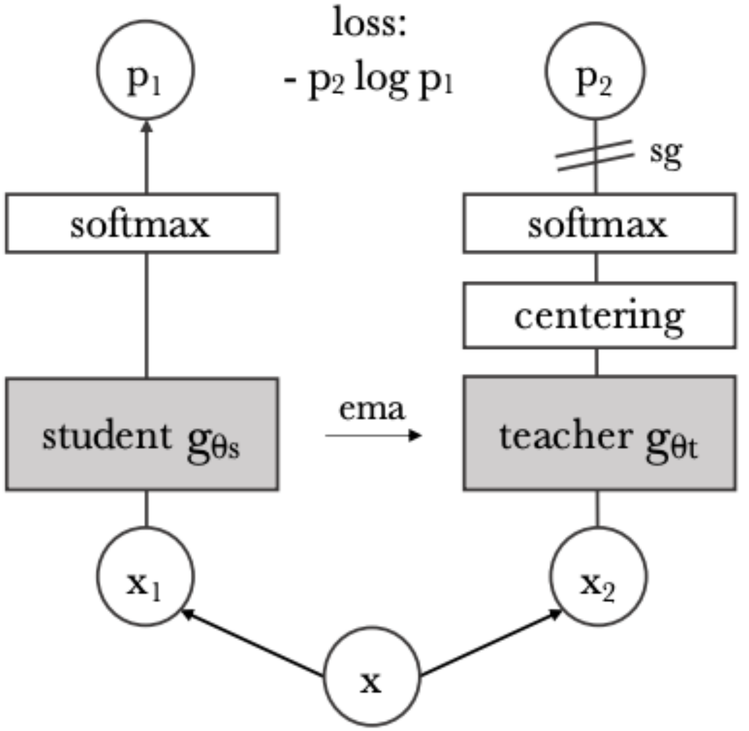
回顾Transformer模型

- Transformer 架构最早是由谷歌在 2017 年的论文《Attention is all you need》中引入，核心是自注意力机制和位置编码。
- 优势在于以下特点：
  - 通过自注意力机制捕获序列中的**长距离依赖关系**：利于完成对象检测和语义分割。
  - **可解释性**：自注意力机制为每个元素的处理提供明确的解释。
  - **并行计算**：可以同时处理序列中的所有元素，具有较高效率。
  - **灵活性强**：易于扩展和修改，可堆叠多个Transformer层，可修改自注意力机制。
- https://www.bilibili.com/video/BV1a24y1F73K 【专题5:生产运营优化与机器翻译原理】2.排程排产优化问题与Seq2seq transformer部分：2:15:42 -2:49:18



DINO（self-distillation with no labels）原理

- **核心**：是不使用标签的知识提炼，**di**stillation with **no** labels。训练了一个学生网络来模仿一个更强大的教师网络的行为，所有这些都不需要在训练数据中有明确的标签。
- **底层**：是 Vision Transformer (ViT) 架构，该设计从自然语言处理 (NLP) 中的Transformer模型中汲取灵感，并将其应用于视觉数据。
- **对比学习方法**：模型学习从图像检索任务中有用的数据中识别相似和不同的例子。DINO 未采用负采样，选择了**全局自注意力机制**，以捕获更全面的数据视图。
- **性能**：优于其他自监督学习方法，甚至可以与一些监督方法相媲美。可用于图像分类、对象检测，甚至实例分割。
- 图片来源：论文Emerging Properties in Self-Supervised Vision Transformers https://arxiv.org/pdf/2104.14294v2.pdf



DINOv2的创新点

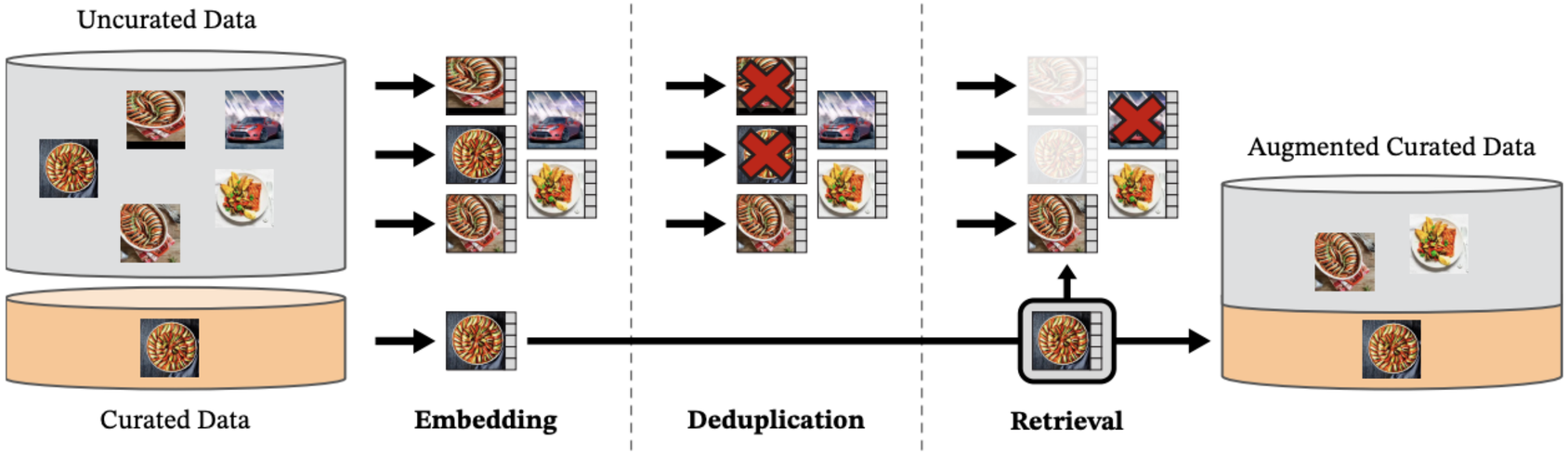


Figure 3: **Overview of our data processing pipeline.** Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting combination augments the initial dataset through a self-supervised retrieval system.

- **Flash Attention机制**: 新的自注意力层，提高了内存使用效率和速度，这对于管理大模型至关重要。每个头部维数是64的倍数，整个嵌入维数是256的倍数时，效果最好。
- **Self-Attention 中的嵌套张量**：在同一前向传播中运行全局裁剪和局部裁剪（具有不同数量的补丁令牌），可以显著提高计算效率。
- **Efficient Stochastic Depth**：跳过了残差下降计算，节省了内存和计算能力。
- **Fully-Shared Data Parallel (FSDP)**：模型跨 GPU 拆分，模型大小不受单个 GPU 内存的限制，而是受所有计算节点上 GPU 显存的总和限制。
- **模型蒸馏**: 对于较小的模型，DINOv2利用最大模型ViT-g的知识蒸馏，而不是从头开始训练，从而提高了性能。这个过程包括将知识从更大、更复杂的模型(教师)转移到更小的模型(学生)。学生模型被训练来模仿教师的输出，从而继承其优越的能力。这个过程提高了小型模型的性能，使它们更有效率。
- DINOv2: Learning Robust Visual Features without Supervision https://arxiv.org/pdf/2304.07193v1.pdf

DINOv2的应用案例

- **实例检索**：使用非参数方法，DINOv2能够在各种数据集(如Paris、Oxford、Met和amsterdam)上优于自监督和弱监督模型。优秀的特征能力体现在可以在不同任务粒度上表现良好的能力。

- **语义分割**：在所有数据集上都表现出强大的性能，使用更简单的预测器也是如此。当使用boosted recipe进行评估时，几乎与 Pascal VOC 上的最新技术水平相匹配。通过冻结主干并调整适配器和头部的权重，模型在 ADE20k 数据集上取得了接近现有技术水平的结果。

- **深度估计**：DINOv2在单目深度估计任务上表现出很好的结果，超过了自监督模型和弱监督模型。SUN-RGBd数据集突出了它在领域外的泛化能力，其中一个在纽约大学室内场景上训练的模块可以泛化到室外场景。