CLIP

- Contrastive Language-Image Pre-Training
  (利用文本的监督信号训练一个迁移能力强的视觉模型)
  - 这家伙有什么用呢？想象一个咱们训练图像分类的场景
  - 训练1000个类别，预测就是这1000个类别的概率，无法拓展
  - 新增类别还得重新训练重新标注太麻烦了，能不能一劳永逸呢
  - 这就是CLIP要解决的问题，预训练模型直接zero-shot

CLIP

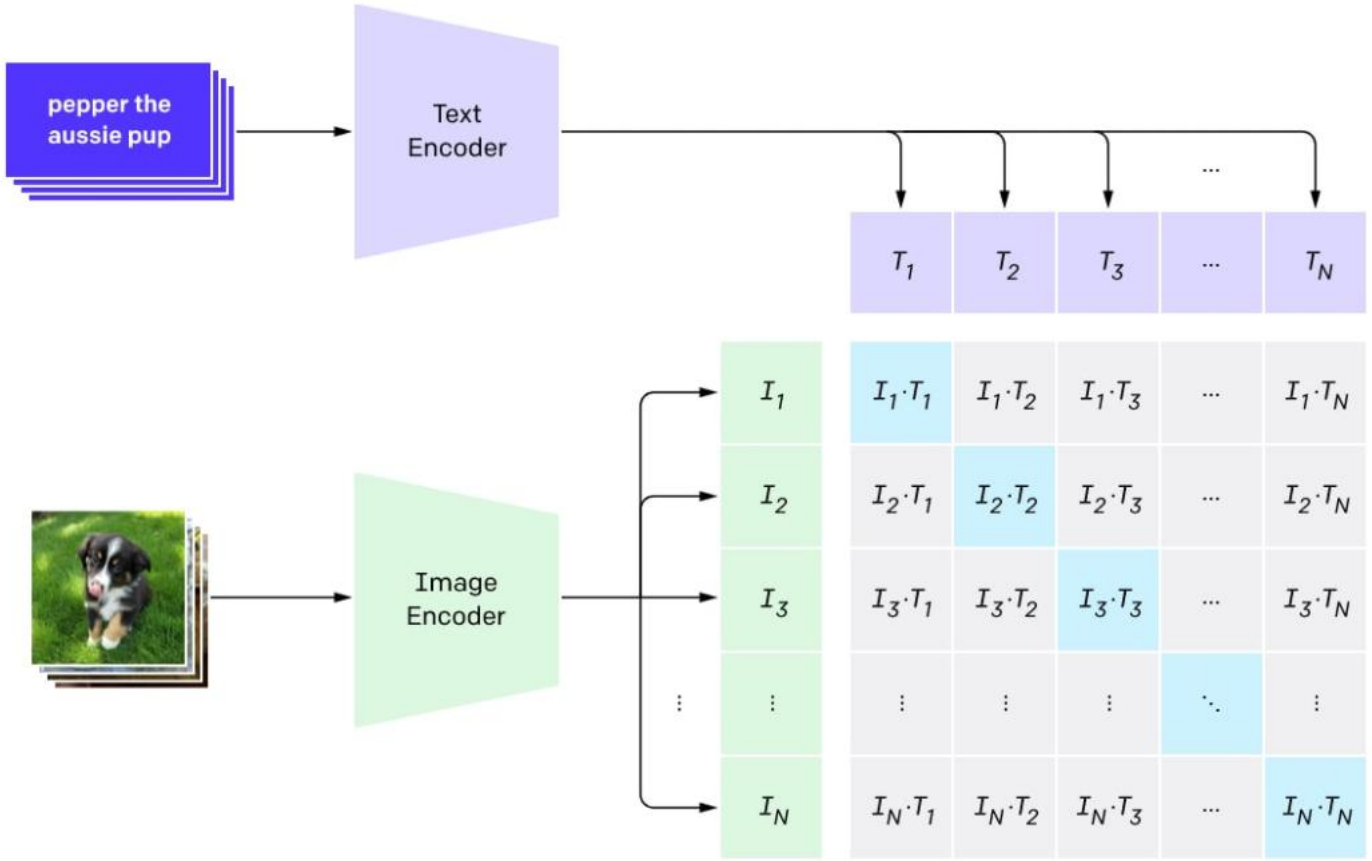
- 与前人工作对比
  - CLIP论文指出，17年就已经开始有这些方法了，但是没获得太多关注
  - 17年类似方法Imagenet上的效果才十几个点，根本就不行
  - 然后OpenAi说了。。。不是方法不行，是资源没到位
  - 一个648解决不了的事，十个648就解决了，这就是CLIP

CLIP

- 成名一战
  - CLIP在完全不使用ImageNet中所有数据训练的前提下
  - 直接Zero-shot得到的结果与Resnet在128W Imagenet数据训练后效果一样
  - 传闻使用4亿个配对的数据和文本来进行训练，不标注直接爬取的
  - 现在CLIP下游任务已经很多了，GAN，检测，分割，检索等都能玩了

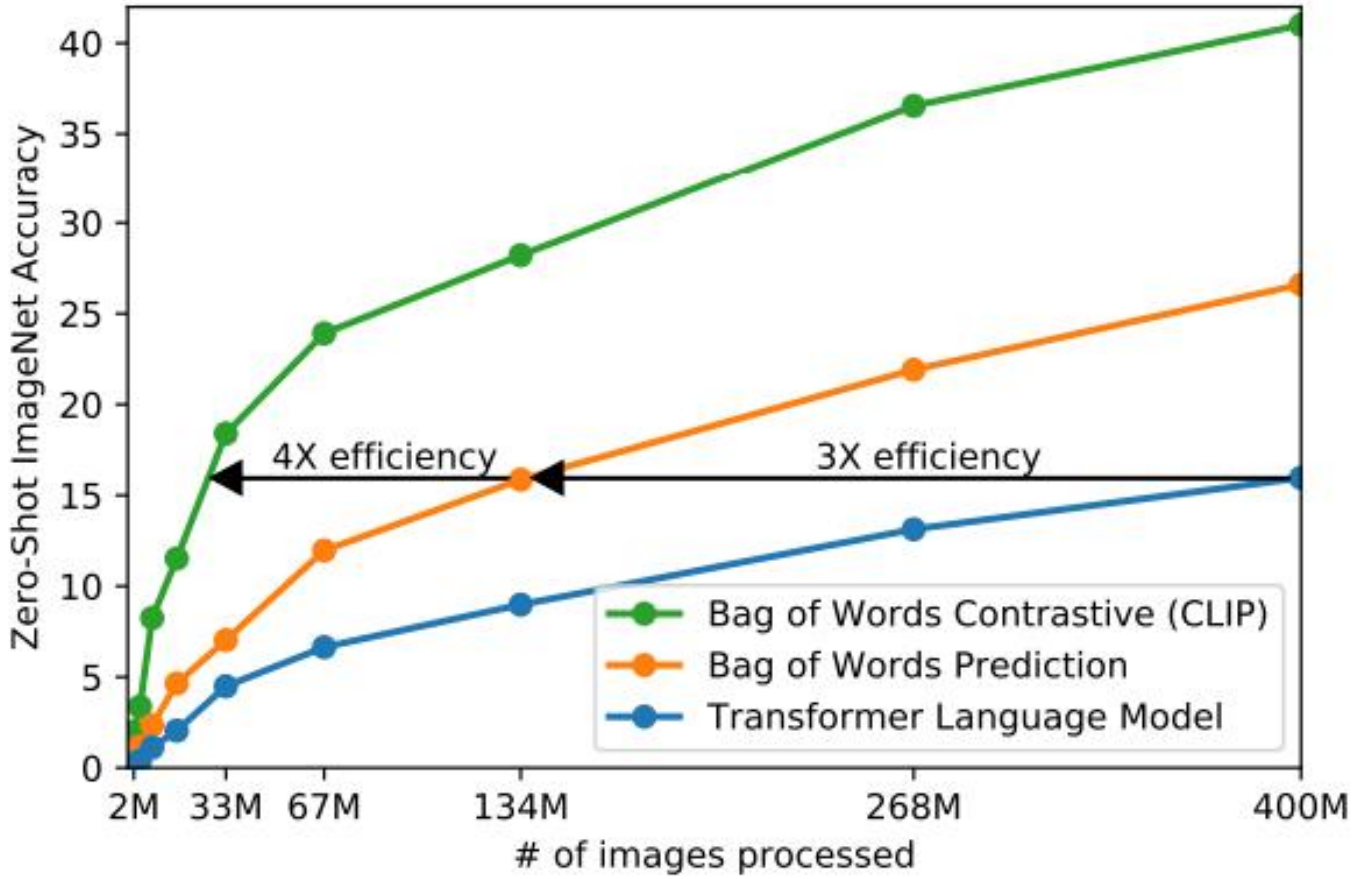
CLIP

- 如何训练模型
  - 图像编码器->特征
  - 文本编码器->特征
  - 计算余弦相似度
  - 对角线的是正样本



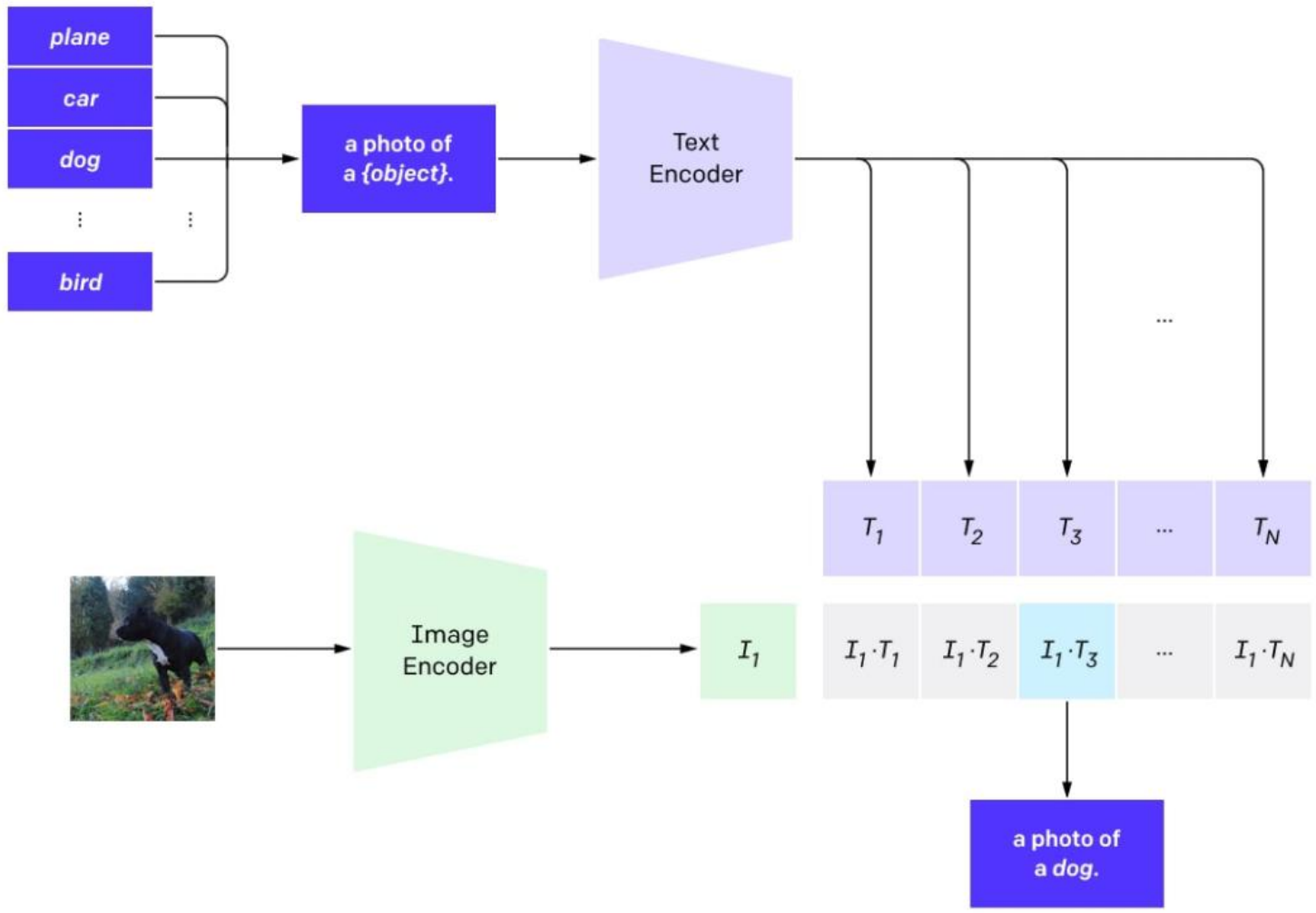
CLIP

- 训练策略
  - 这种规模的训练要想想招了
  - 论文里说了好多GPU YEAR
  - 4X就是对比学习的效率提升
  - 只看配对不，不预测具体词



CLIP

- 如何进行推理
  - 给一些提示文本(任意个)
  - 提示要好好说话
  - 然后每种提示算相似度
  - 找到概率最高的即可



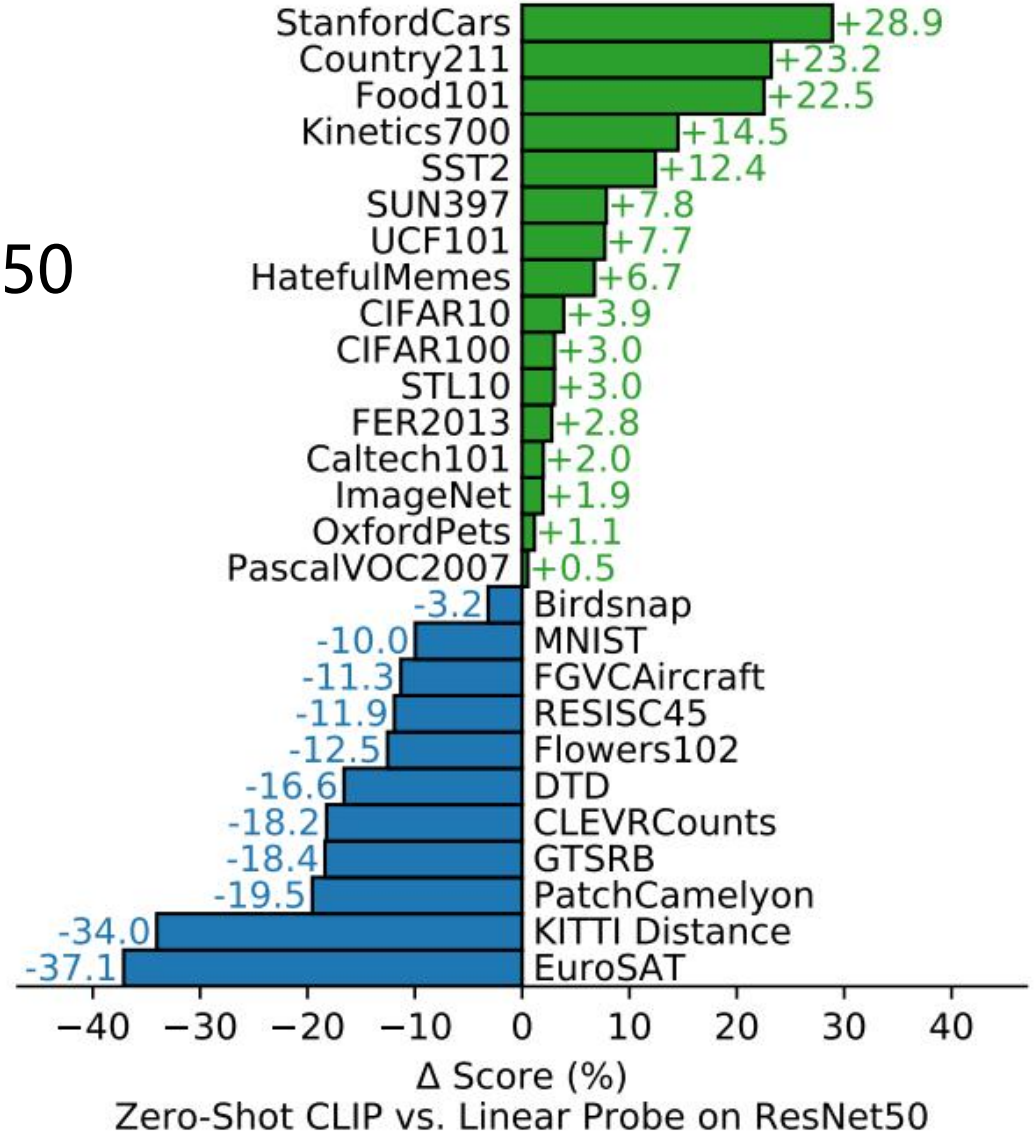
CLIP

- 合理的提示
  - 预测的时候提示也很重要
  - 首先得是一句话来描述
  - 而且最好跟你要预测的场景相关
  - 提示的也全面，结果也会提升

specify the category. For example on Oxford-IIIT Pets, using "`A photo of a {label}, a type of pet.`" to help provide context worked well. Likewise, on Food101 specifying *a type of food* and on FGVC Aircraft *a type of aircraft* helped too. For OCR datasets, we found that putting quotes around the text or number to be recognized improved performance. Finally, we found that on satellite image classification datasets it helped to specify that the images were of this form and we use variants of "`a satellite photo of a {label}.`".

CLIP

- 实验对比
  - 基础模型是Imagenet预训练的Resnet50
  - 不同数据集只训练最后一层FC来微调
  - 绿色的表示CLIP比微调的Resnet强的
  - 蓝色的是有所下降的数据集
    （非常规场景的数据都有所下降）



CLIP

- CLIP的局限
  - 跟Resnet50当成平手，但是Resnet50它不是老大啊，老大比他强10来个点呢
  - OpenAi说如果要跟人家老大能打成平手，数据集需要1000X
  - 只能玩常规的，Mnist它效果一般因为这种数据是人工合成的
  - 面向测试集调参？就跟Imagenet干上了，参数都适合它的

DALL-E

- 得先熟悉下VQGAN这个家伙（Vector Quantized）
- 它跟CLIP有啥关系呢？我们现在希望生成一些图像结果
- VQGAN就相当于生成器，CLIP就相当于判断器（看生成结果与描述相同不）

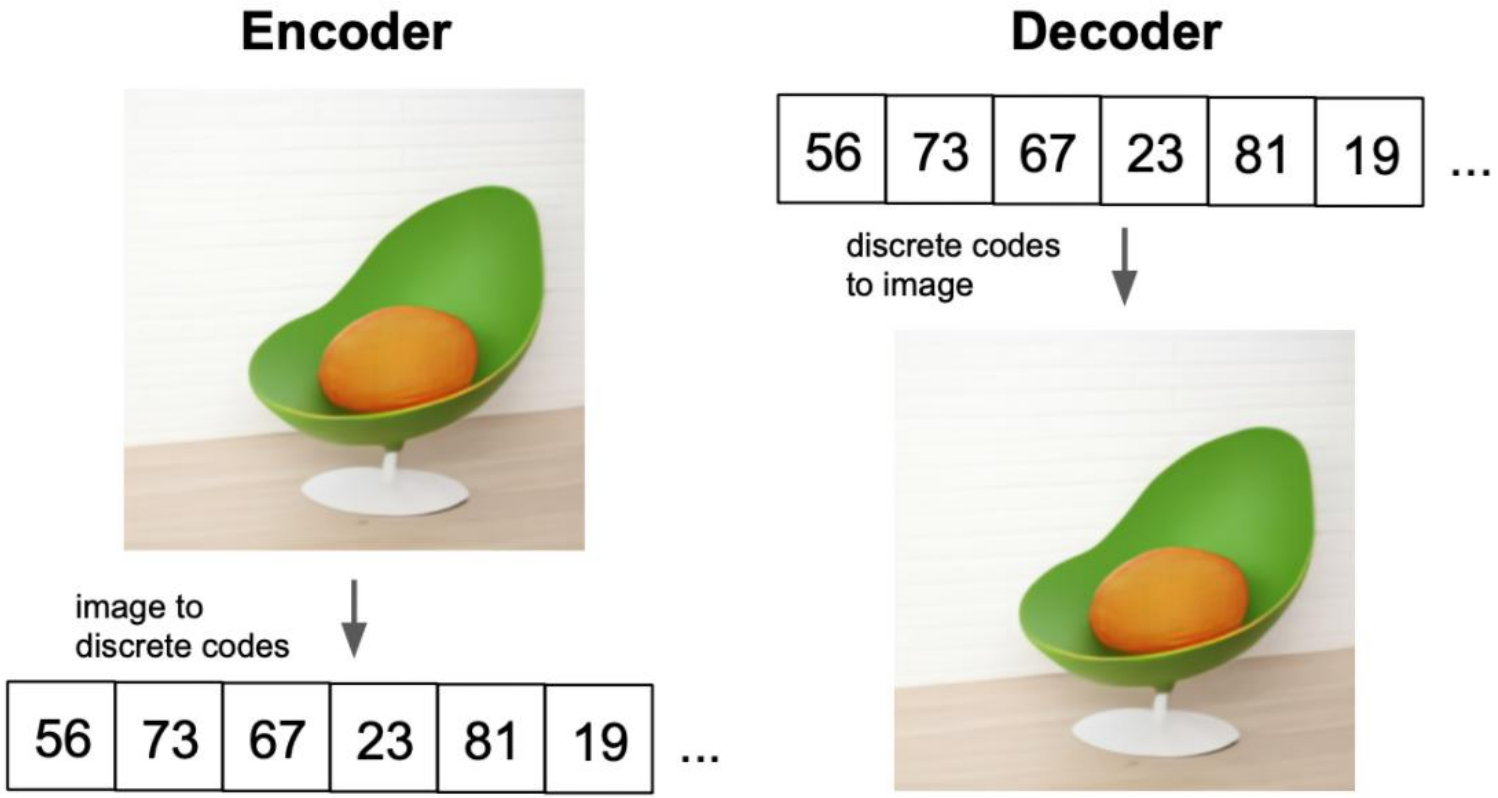
DALL-E

- 如何描述这个小家伙呢?
  - 白色的，俩耳朵，瞅我呢
  - 咱们脑子里可能先想到一些关键词
  - 然后咱们再寻思怎么组成一句话描述呢
  - 一只白色的俩耳朵猫在瞅我呢

DALL-E

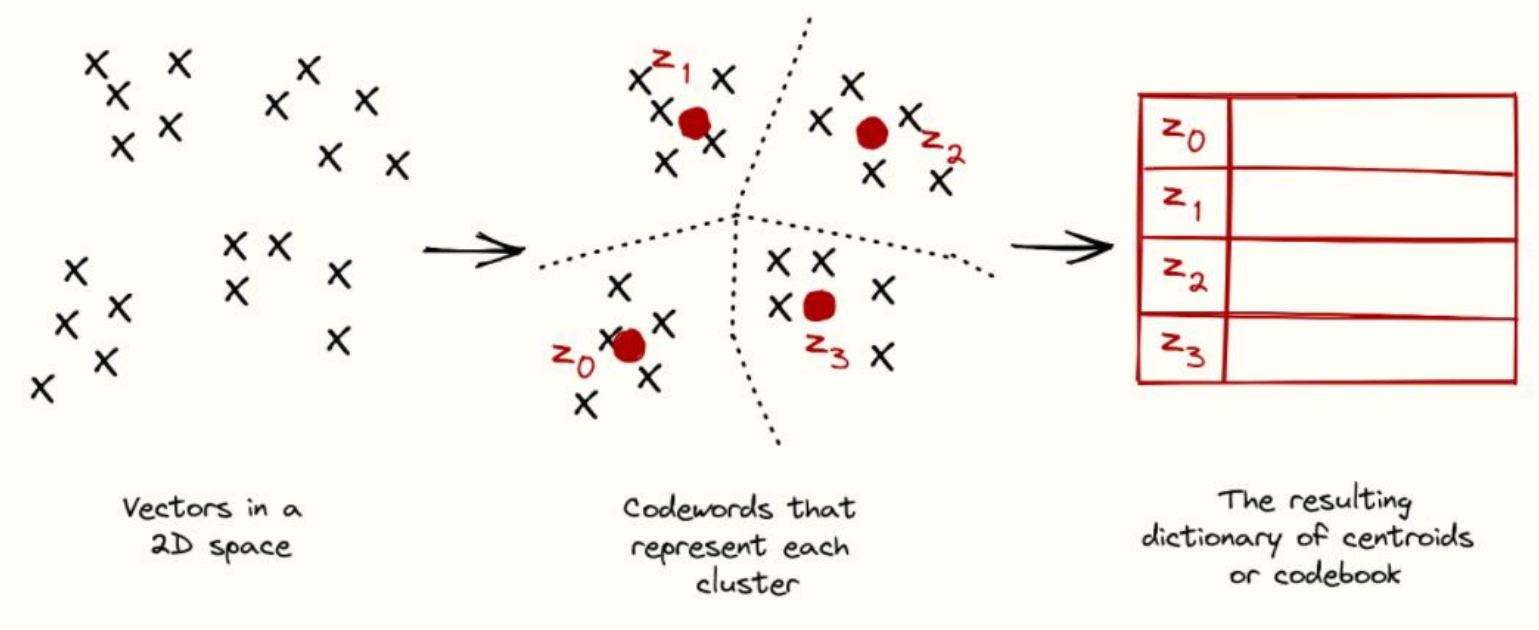
- 图像如何表示

- NLP中我们讲文本向量化，这里我们能否对图像离散向量化呢?



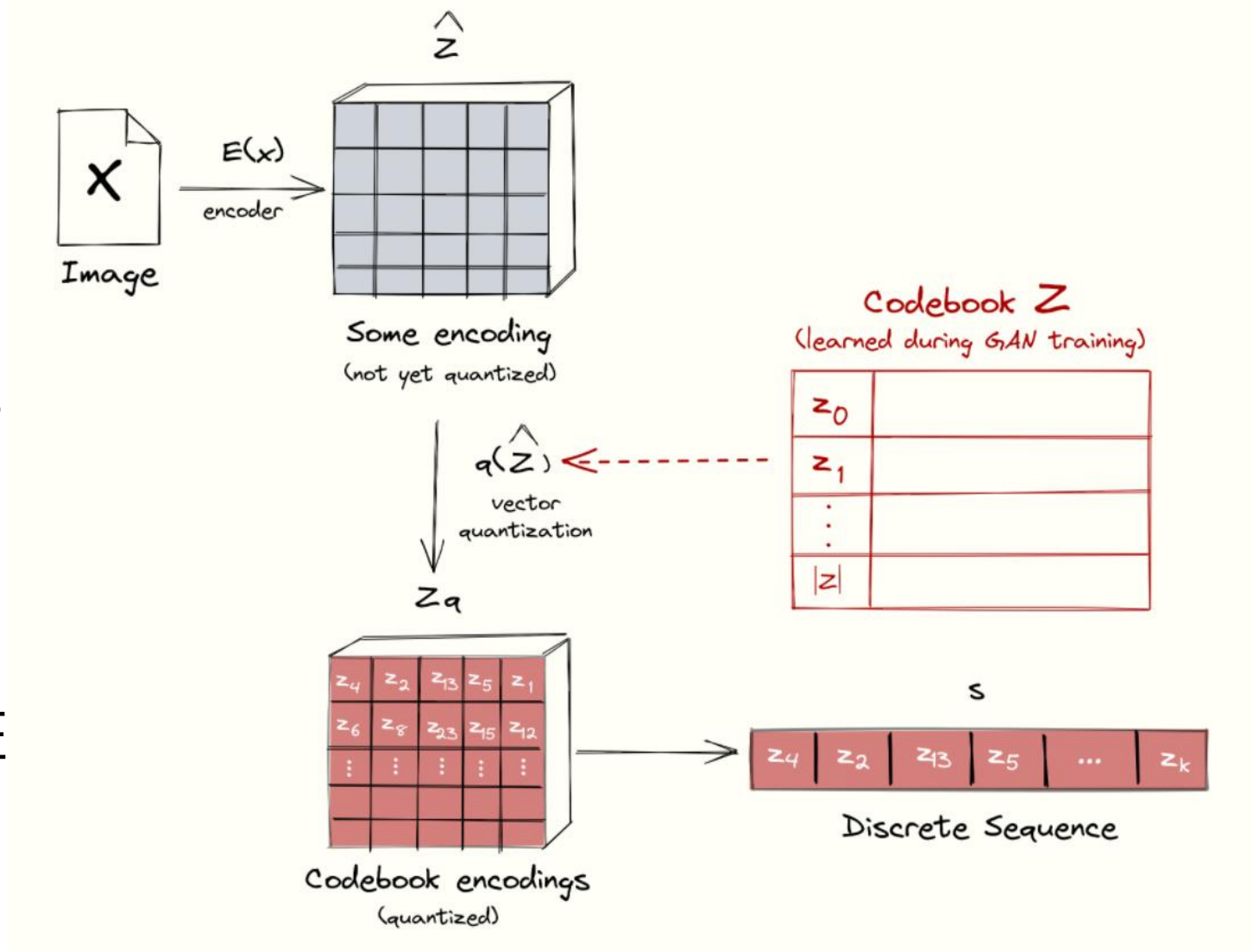
DALL-E

- codebook
  - 按照咱们刚才想法，那得先把特征离散化再整合啊
  - 如何离散化特征呢，咱们得有章可循，那不如查表吧（有限个离散特征）



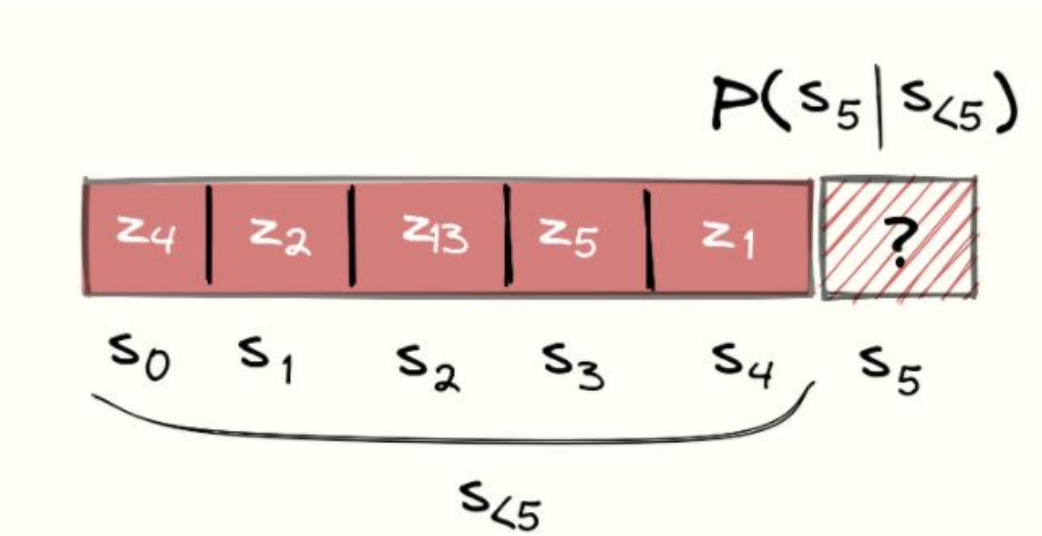
DALL-E

- 大致流程
  - 先通过编码器得到特征
  - 然后选相似度得到离散特征
  - 其实就是查codebook表
  - 最后得到实际编码后的特征

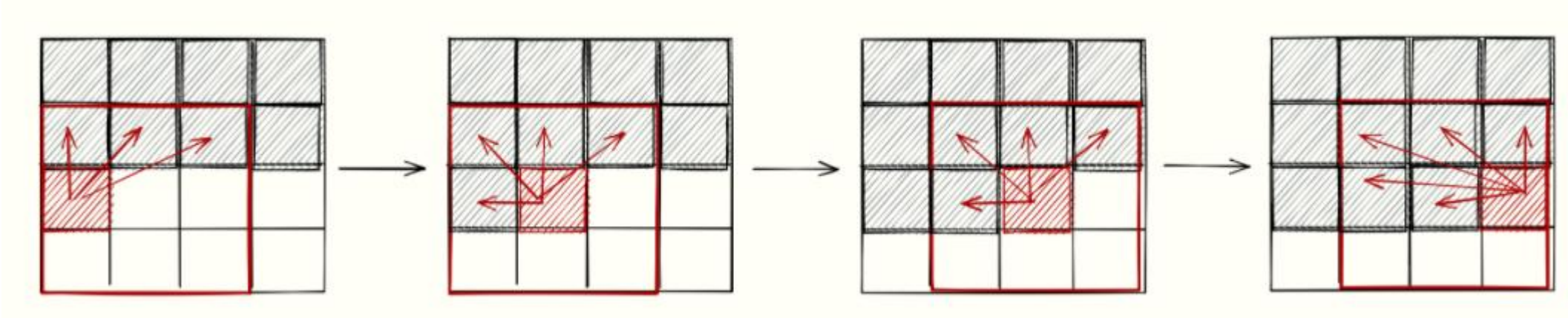


DALL-E

- 根据特征序列逐步生成
  - 类似transformer的感觉，一点点生成
  - 生成的过程中还要它们之间的考虑关系



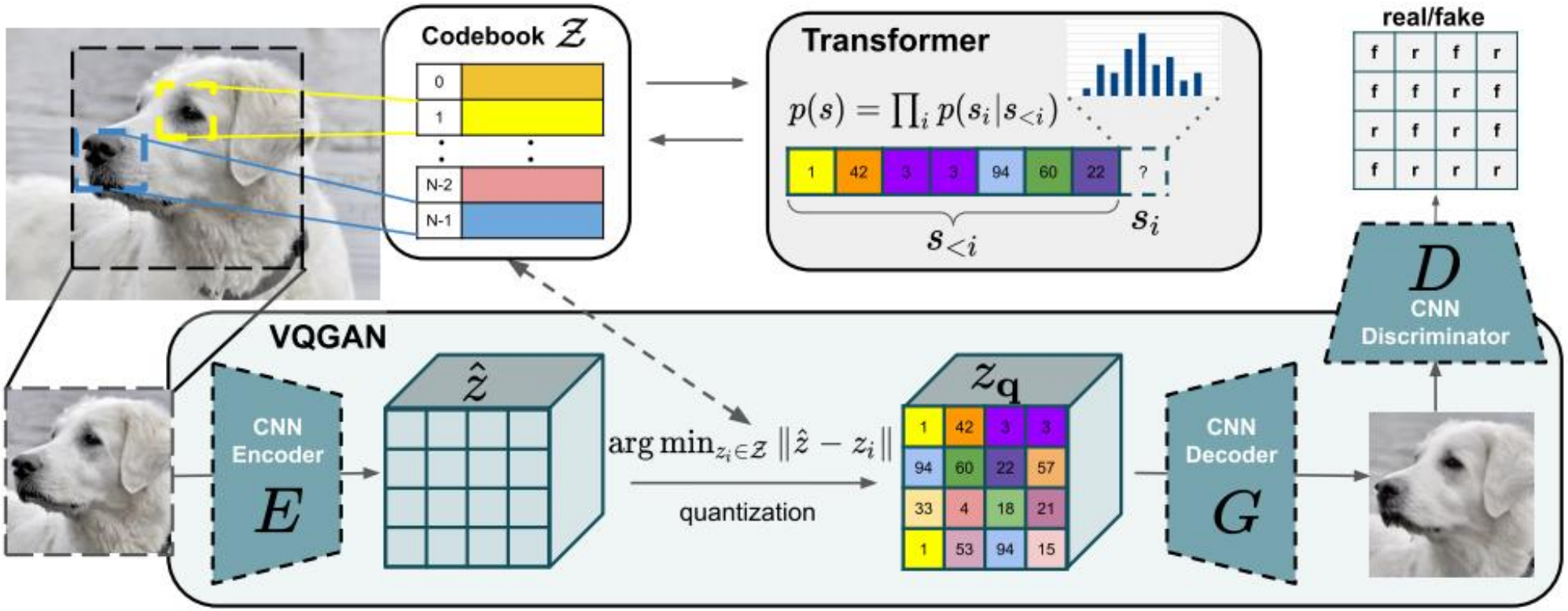
  - 但是并不是全局特征都会用到，论文貌似指的是邻居特征



DALL-E

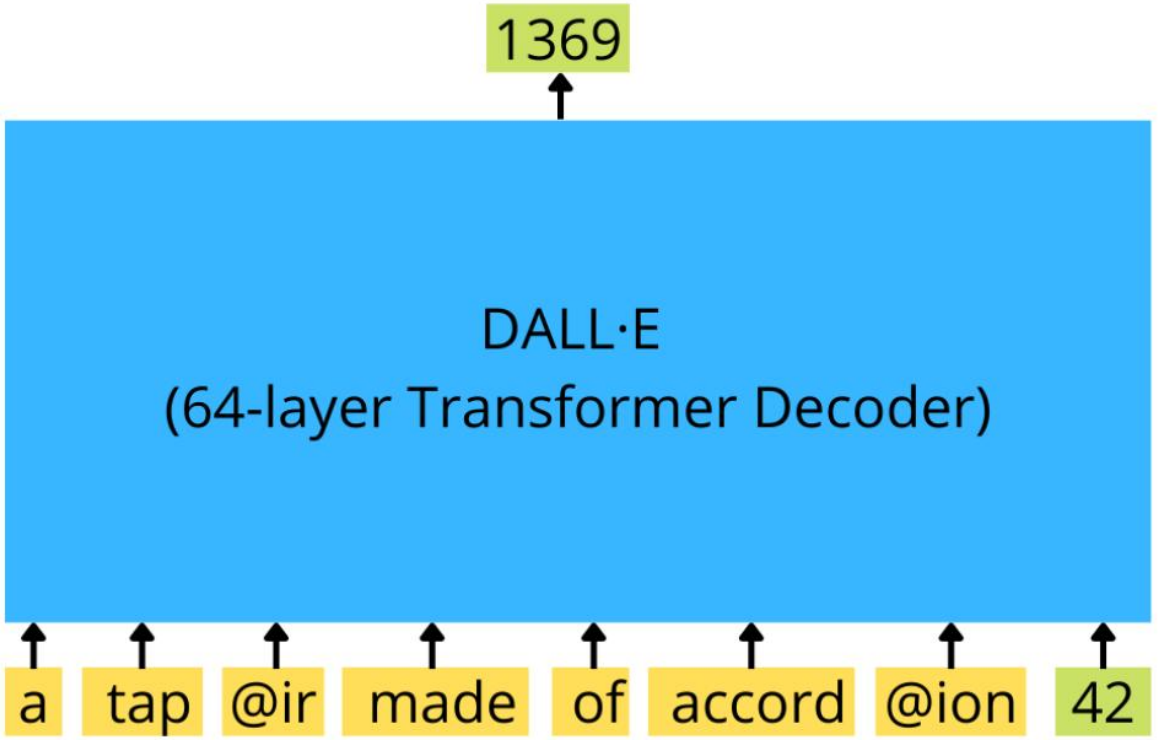
- VQGAN相当于给DALL-E提供了基本出发点

- 整理流程，别忘了还有transformer更新codebook



DALL-E

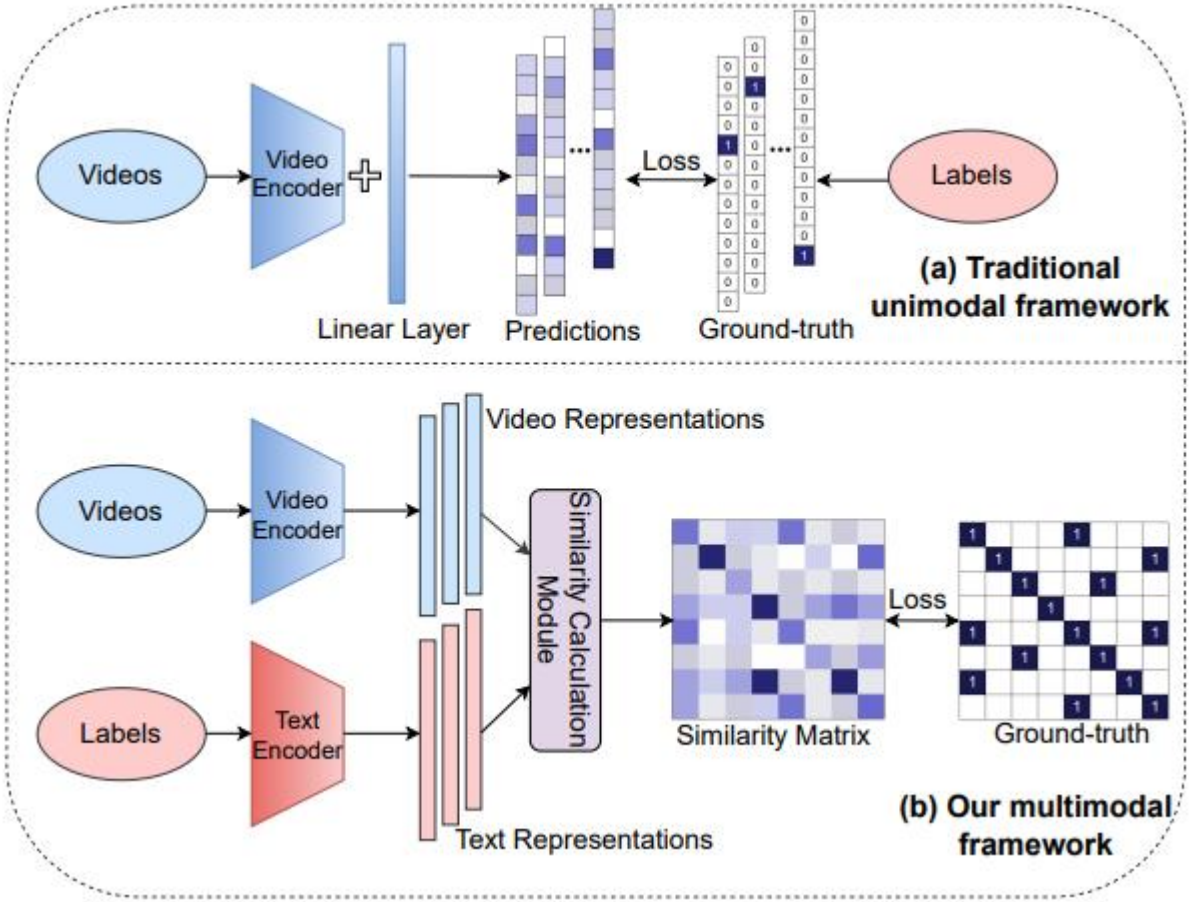
- DALL-E中如何生成结果呢
  - 类似GPT，只不过输入多了图像特征
  - 3种注意力（还有文本和图像的）
  - 42就是图像特征编码，预测输出图像
  - 例如输出的图像编码是1369(万物GPT化)



ActionCLIP

- ActionCLIP: A New Paradigm for Video Action Recognition

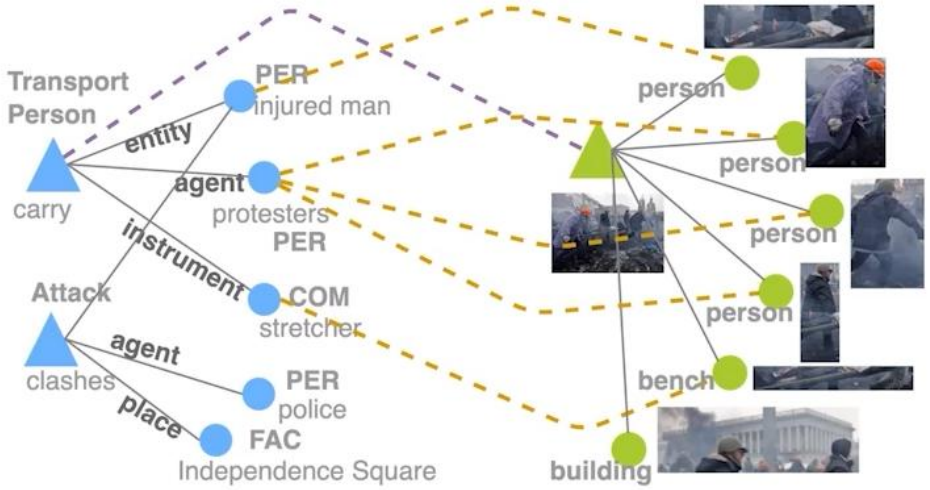
  - 视频分类，行为识别也类似
  - 其实本质也是构建特征提取器
  - 同样是zero-shot来预测
  - 那感觉CV又可以通吃了

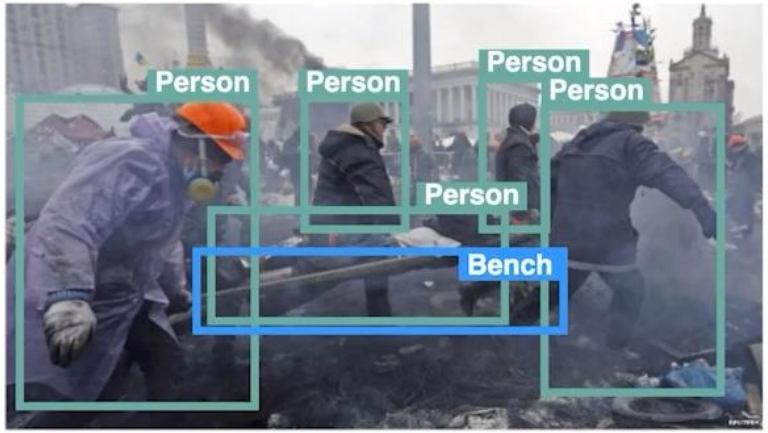


CLIP-Event

- CLIP-Event: Connecting Text and Images with Event Structures
  - 这个还挺有意思的，能将事件中的人与动作链接起来
  - 相当于先通过文本时间抽取得到一些关系组合，再与图像进行配对

Antigovernment protesters carry an injured man on a stretcher after clashes with riot police on Independence Square in Kyiv on February 20, 2014.





CLIP-Event

- 对图像多了一层描述
- 不仅仅是检测任务
- 还要推断检测到的东西咋地了
- 将事件信息与图像结合到一起
- 还要能区分角色，谁干啥了

Event	Wearing
Item	mask
Agent	person

Event	Treatment
Agent	doctor
Target	patient

Event	Researching
Agent	researcher
Target	dropper

Event	Sanitizing
Agent	person
Tool	sprayer

Event	Testing
Agent	woman
place	car

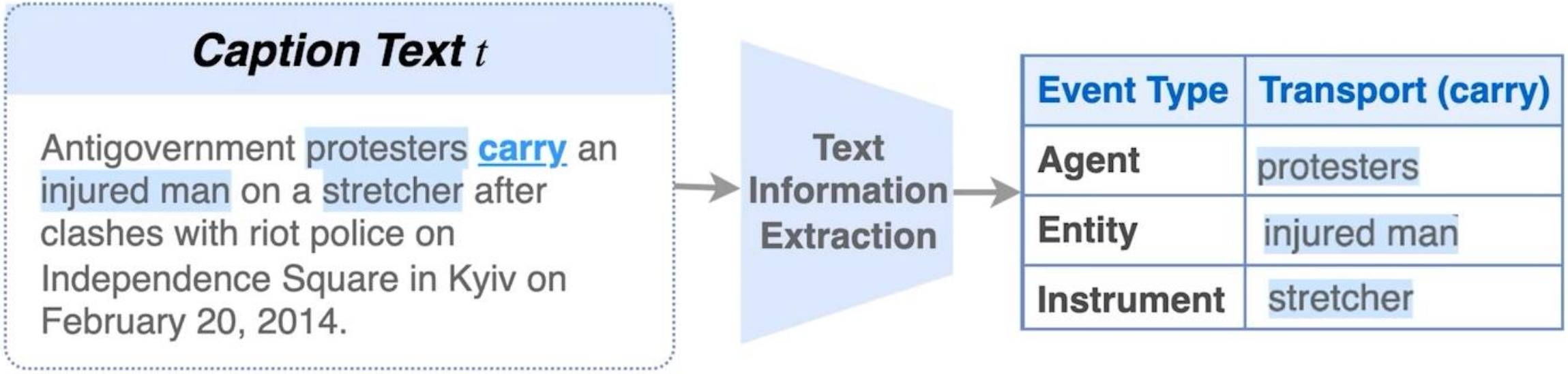
Event	Vaccination
Agent	woman
Target	girl

CLIP-Event

- 新闻事件抽取

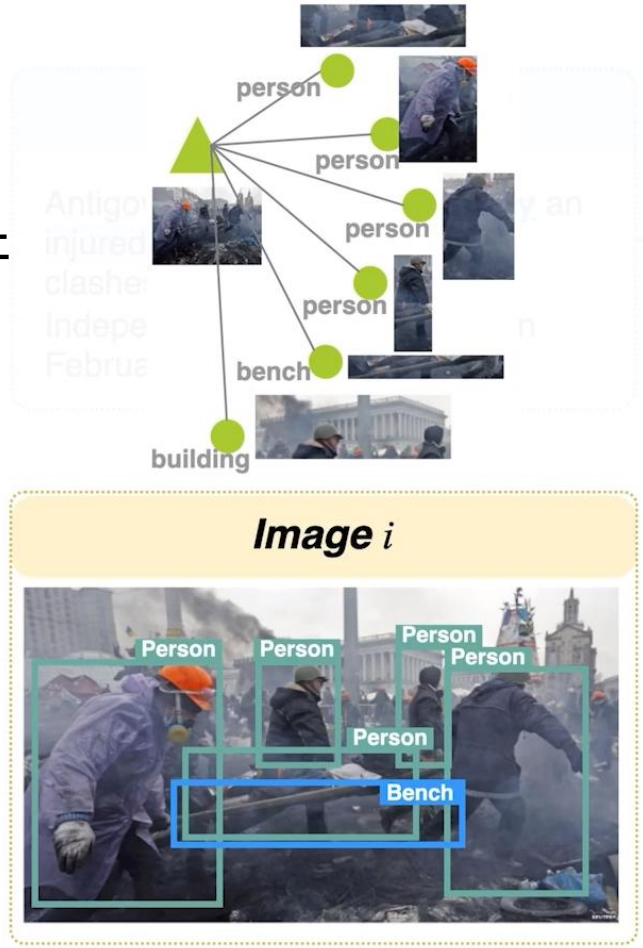
- 感觉跟咱们之前唠的知识图谱的抽取差不多，谁把谁咋地了

- 先把新闻抽取成一个三元组，然后还要再讲他们组合成一句话



CLIP-Event

- 正负样本的制作
  - 正样本就是抽取的事件
  - 负样本可以替换事件
  - 也可以替换主体
  - 类似hard negative



Positive Labels

Event Type	Transport (carry)
Agent	protesters
Entity	injured man
Instrument	stretcher

Negative Labels (events)

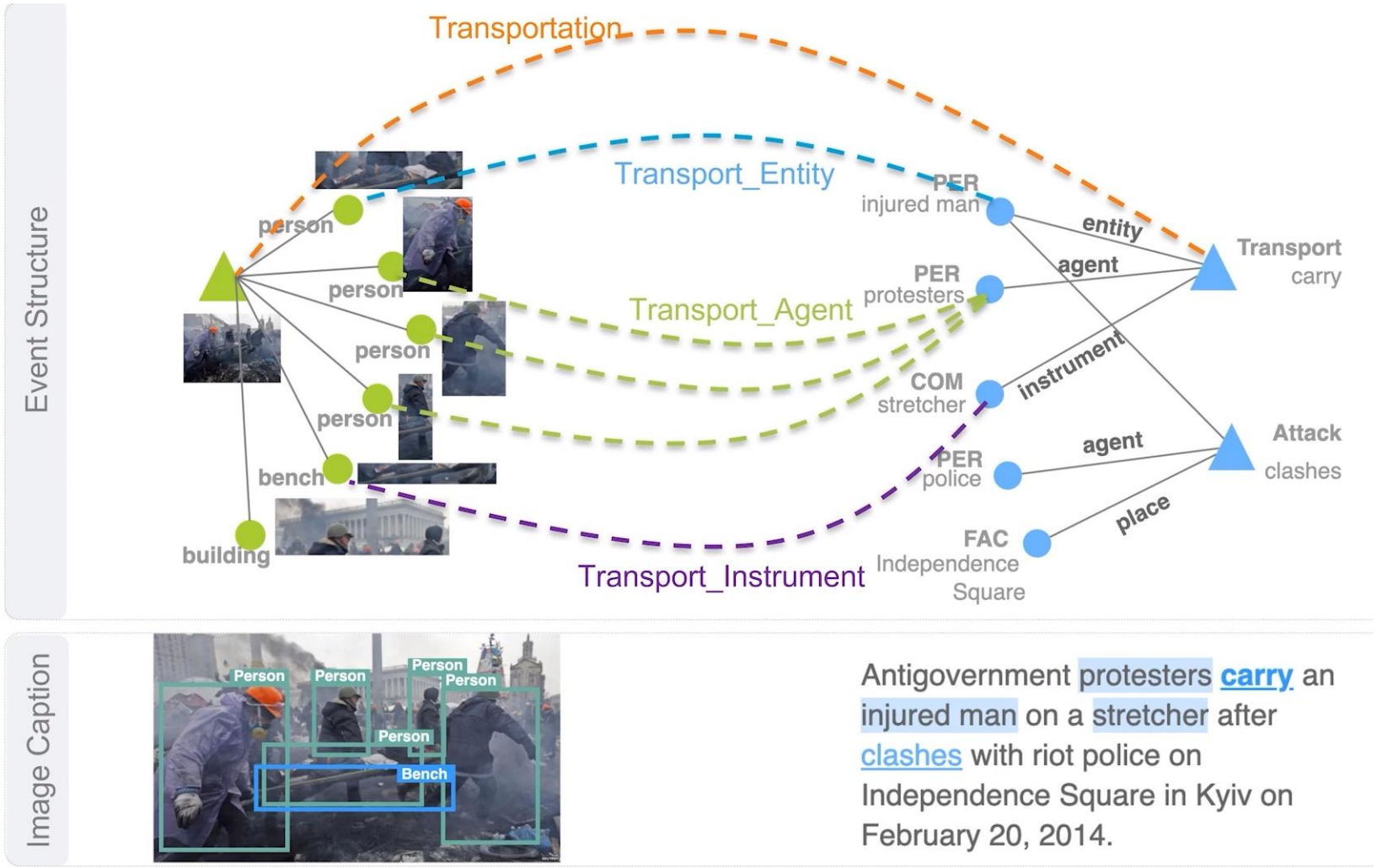
Event Type	Arrest (arrest)
Agent	protesters
Entity	injured man
Instrument	stretcher

Negative Labels (arguments)

Event Type	Transport (carry)
Agent	injured man
Entity	stretcher
Instrument	protesters

CLIP-Event

- 这不就组合好啦
- 不同人，不同事
- 看看谁咋地谁了
- 这个还挺有意思
- 感觉类似推理了



CLIP-Event

- 这个挺有意思的

- 结果还是很有意思的，能区别不用检测目标与事件的关系信息