



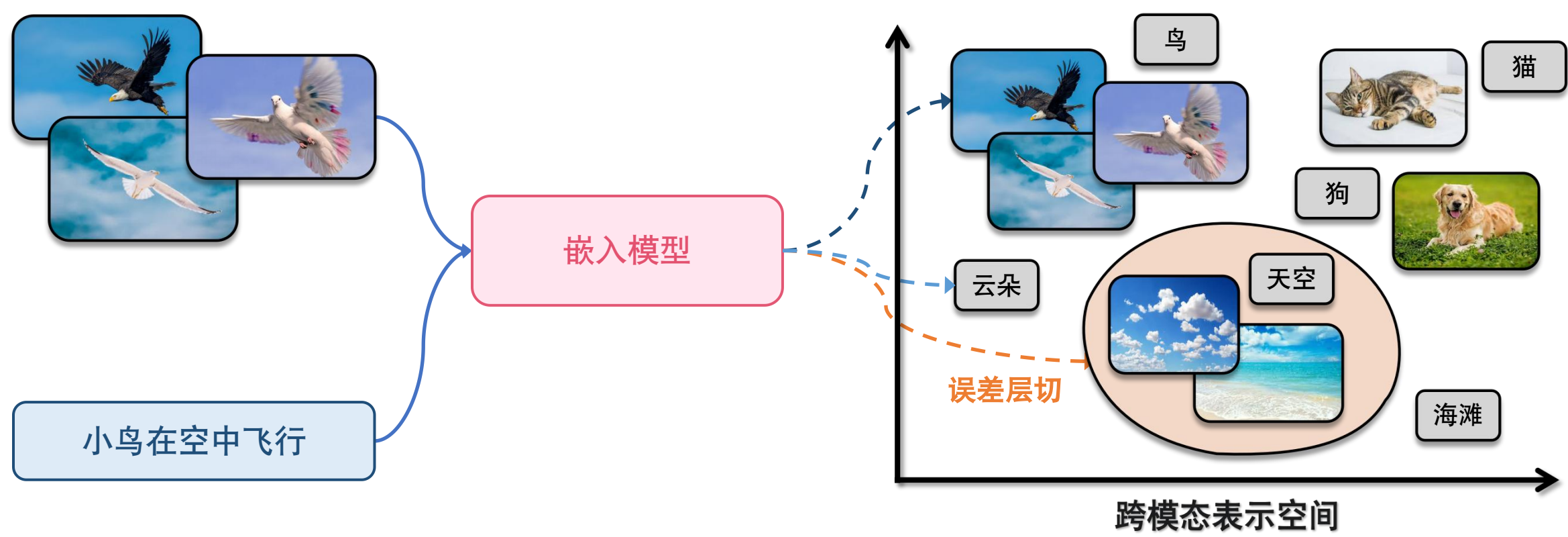
机器学习与数据挖掘

# 多模态感知与理解



# 跨模态表示学习的目标

学习共享的表示空间，对齐不同模态的语义信息



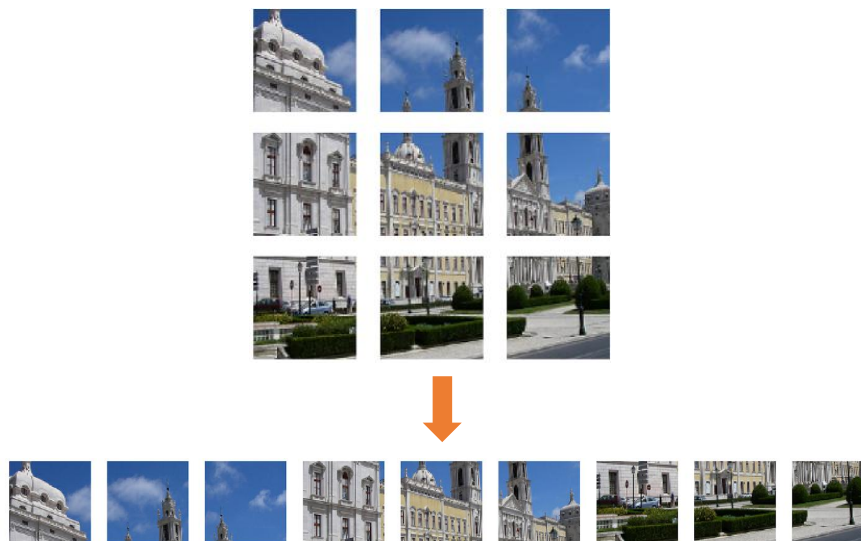
# 大纲

---

- 基于Transformer的视觉处理架构ViT
- CLIP
- BLIP

# 视觉与文本统一网络架构

- 思路：利用BERT类似的模型来处理视觉数据
  - 通过将图像分割成  $N$  个图像块的序列来创建“词”的概念

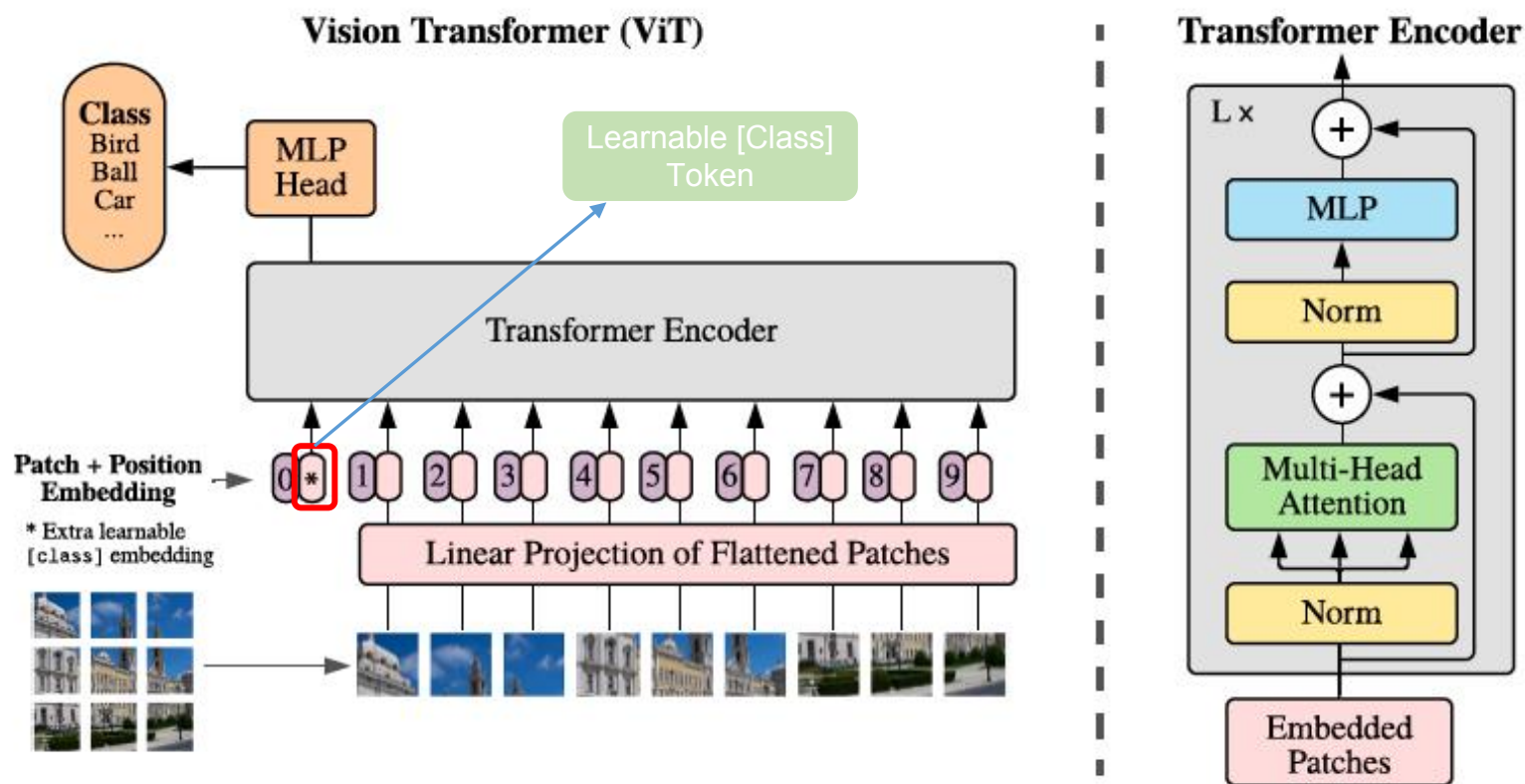


- 将压平的图像块进行线性投影  $\mathbf{x}_p^i \in \mathbb{R}^{16^2 \cdot 3}$  for  $i = 1, 2, \dots, N$

$$\mathbf{z}_0^i = \mathbf{x}_p^i \mathbf{E},$$

其中  $\mathbf{E} \in \mathbb{R}^{16^2 \cdot 3 \times D}$  是线性嵌入矩阵

- 引入一个可学习的 [class] 标记，表示为  $z_0^0$ ，类似于 BERT 的 [CLS] 标记
- 为位置 0, 1, 2, ... N 引入可学习的一维位置嵌入，可以表示为  $E_{pos} \in \mathbb{R}^{(N+1) \times D}$



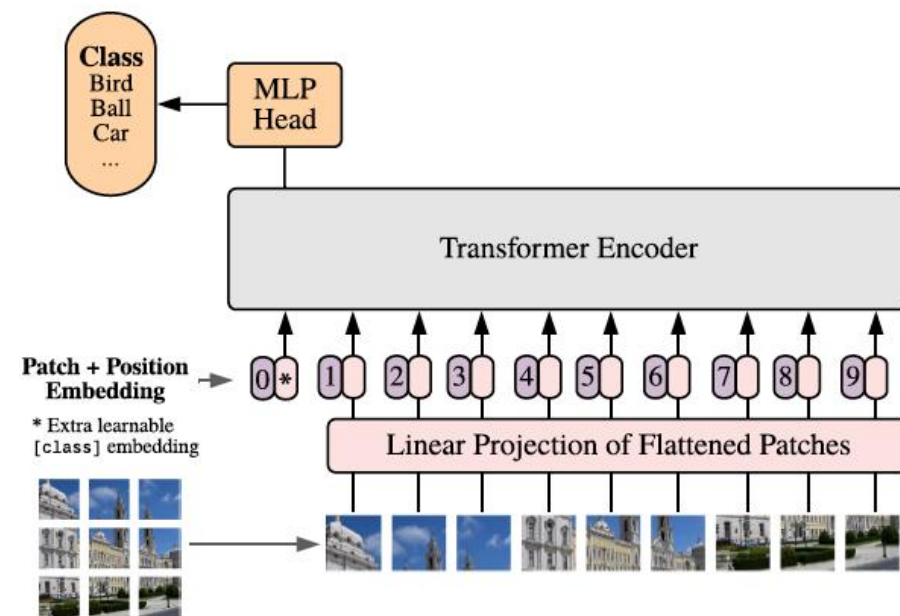
## ➤ 更新规则

$$\mathbf{z}_0 = [\mathbf{x}_{class}; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \cdots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{pos}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell$$

其中  $\text{MSA}(\cdot)$  表示多头注意力



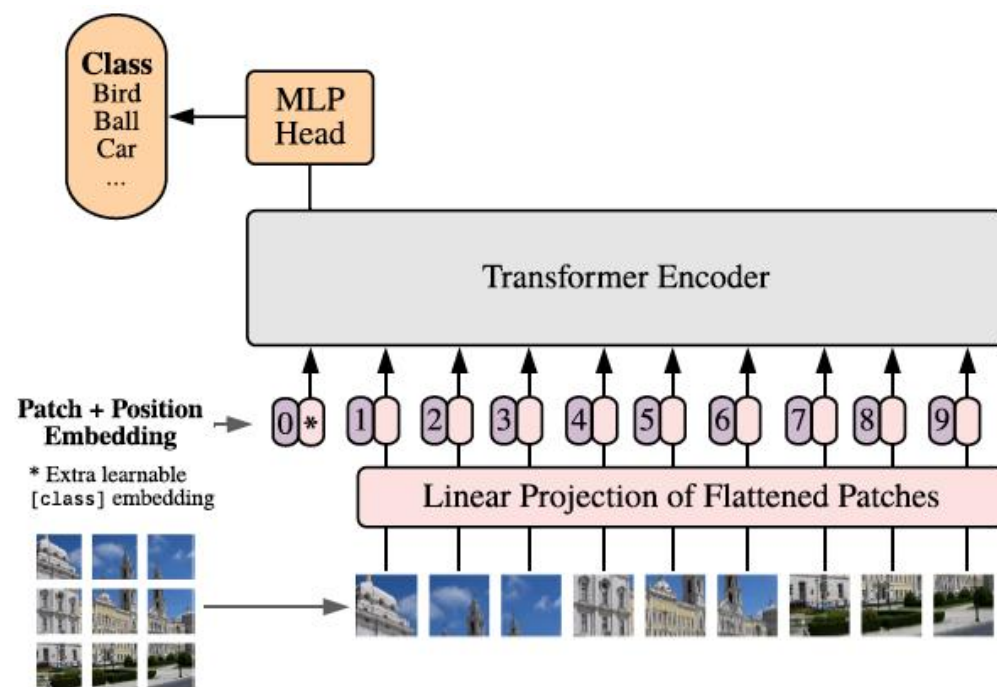
- ViT模型不同版本（Base, Large, Huge）的模型参数信息

Model	Layers	Hidden size $D$	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

# ViT的预训练——监督方式

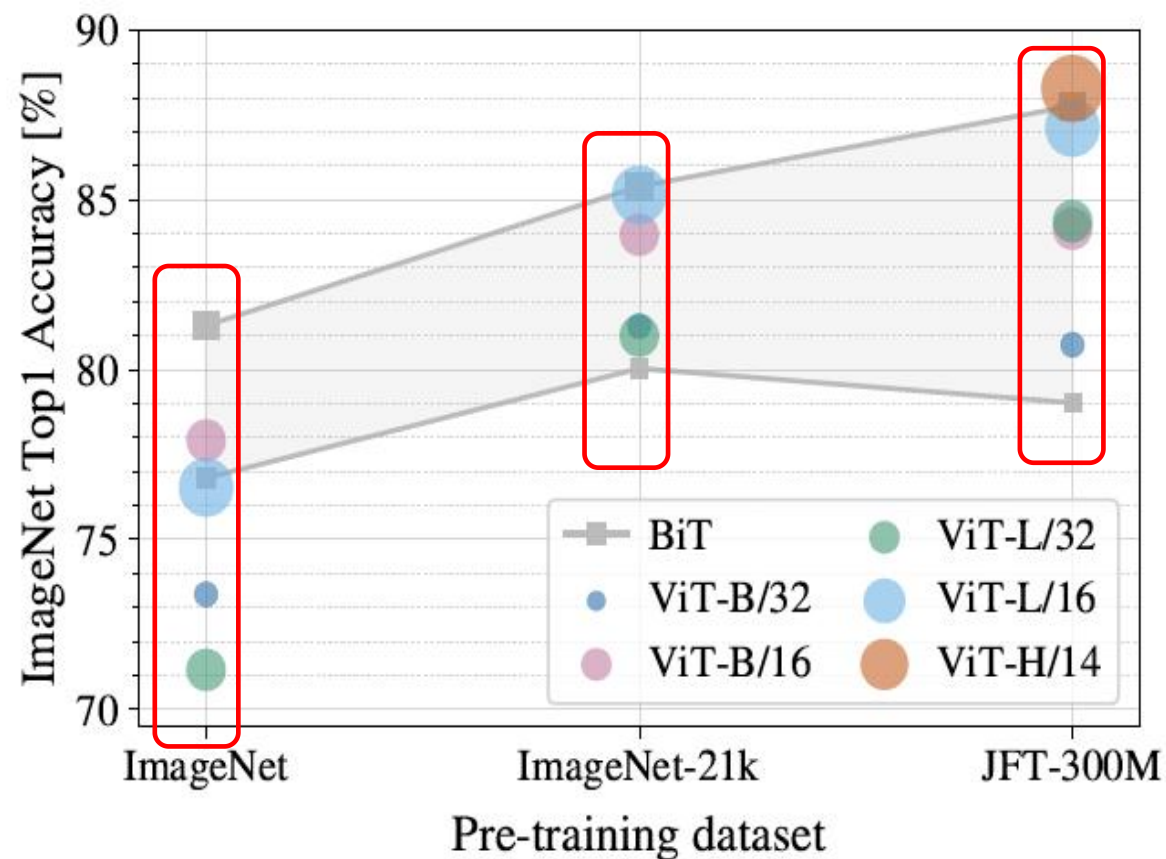
JFT300M包含 3 亿张带标签的图像，是谷歌为内部使用而开发的专有数据集

JFT300M比著名的 ImageNet 数据集大得多



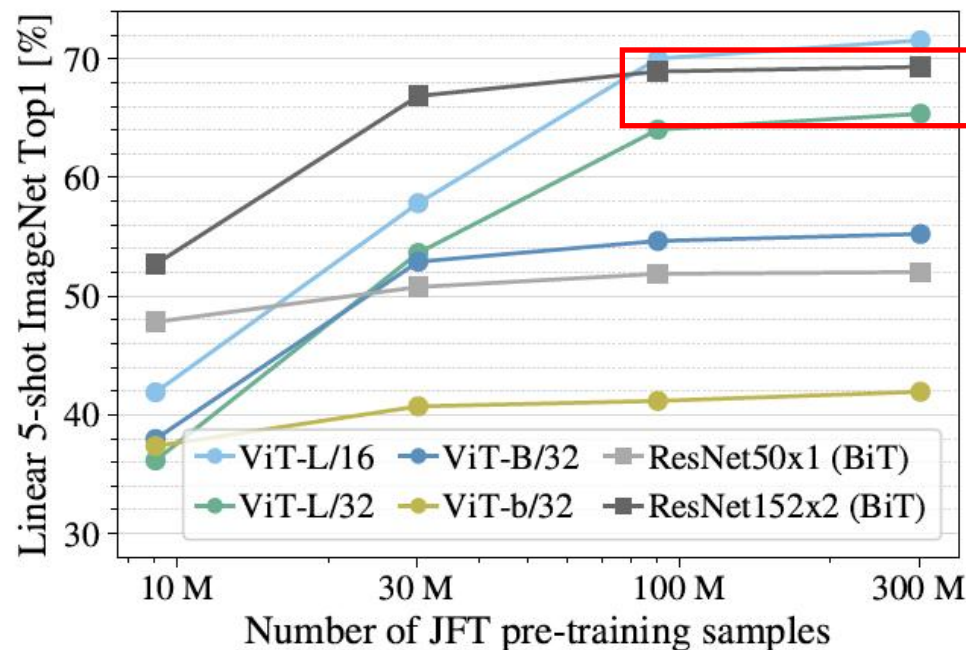
- 在不同的数据集上进行监督预训练，然后在 ImageNet 上进行微调以进行评估

- ✓ 对于小型预训练数据集，ResNet 表现更好，ViT-B 表现优于 ViT-L，这可能是由于 CNN 对图像的归纳偏置以及大型 ViT 容易过拟合
- ✓ 对于大型数据集，ViT-L 和 ViT-H 的优势比 ResNet 和 ViT-B 更为突出
- ✓ 小图像块 (16 x 16) 比大图像块 (32 x 32) 更受欢迎



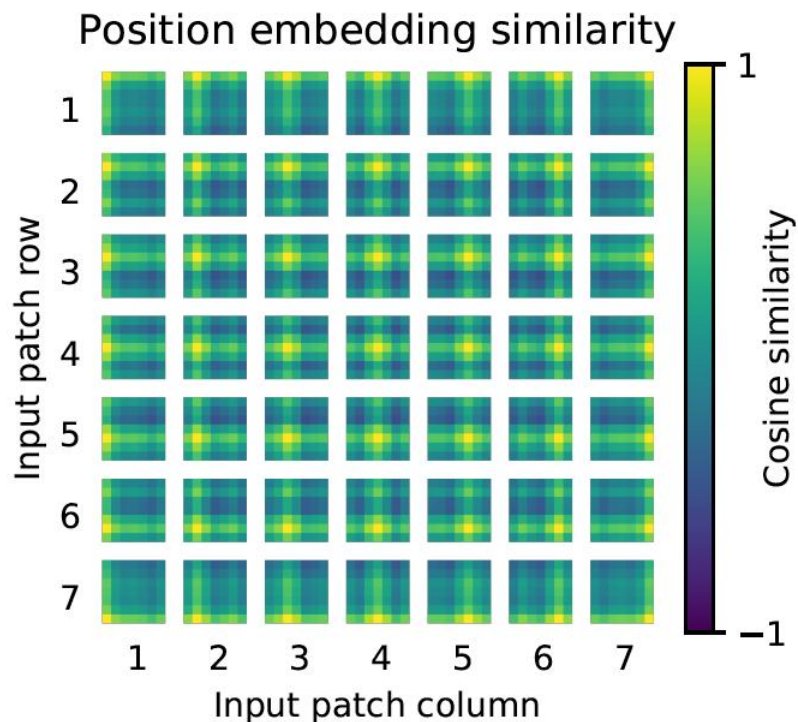
- ViT-L/16 意味着图像块大小设置为 16 x 16
- BiT 指在 ImageNet 和辅助数据上训练的 ResNet

- 在不同比例的 JFT-300 数据集上进行监督预训练，然后在 ImageNet 上对输出表示进行线性探测



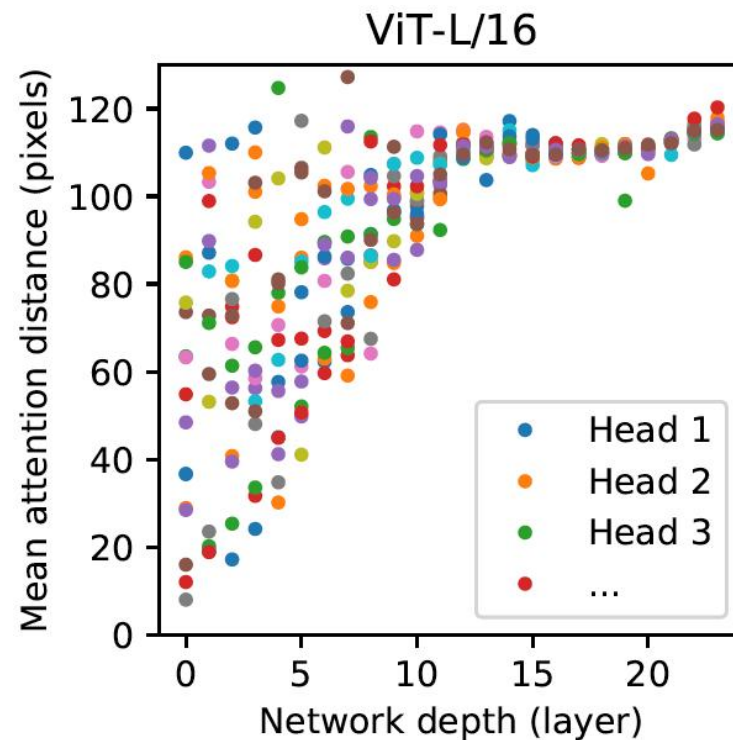
- ✓ 结果表明，当预训练数据集足够大时，ViT-L 学习到的表示比 ResNet 更好
  - 从海量数据中学到的知识压倒了 CNN 的归纳偏置

➤ ViT-L/32 的位置嵌入的相似性



- 学习到的位置嵌入可以自己发现相关的行和列

➤ 每个头和网络深度对关注区域大小的影响。图中每个点代表16个注意力头中的一个，并显示了其在所有图像上的平均注意力距离。



- 浅层：一些头关注局部图像块，而一些头关注远距离图像块
- 深层：大多数关注远距离图像块

# 大纲

---

- 基于Transformer的视觉处理架构ViT
- CLIP
- BLIP

- 使用标签作为监督的局限性
  - 1) 标注图像成本高昂
  - 2) 难以获取网络规模的数据集
  - 3) 仅使用一个标签来描述图像既不准确也不完整



如何只用一个标签来定义该图像表达的信息？

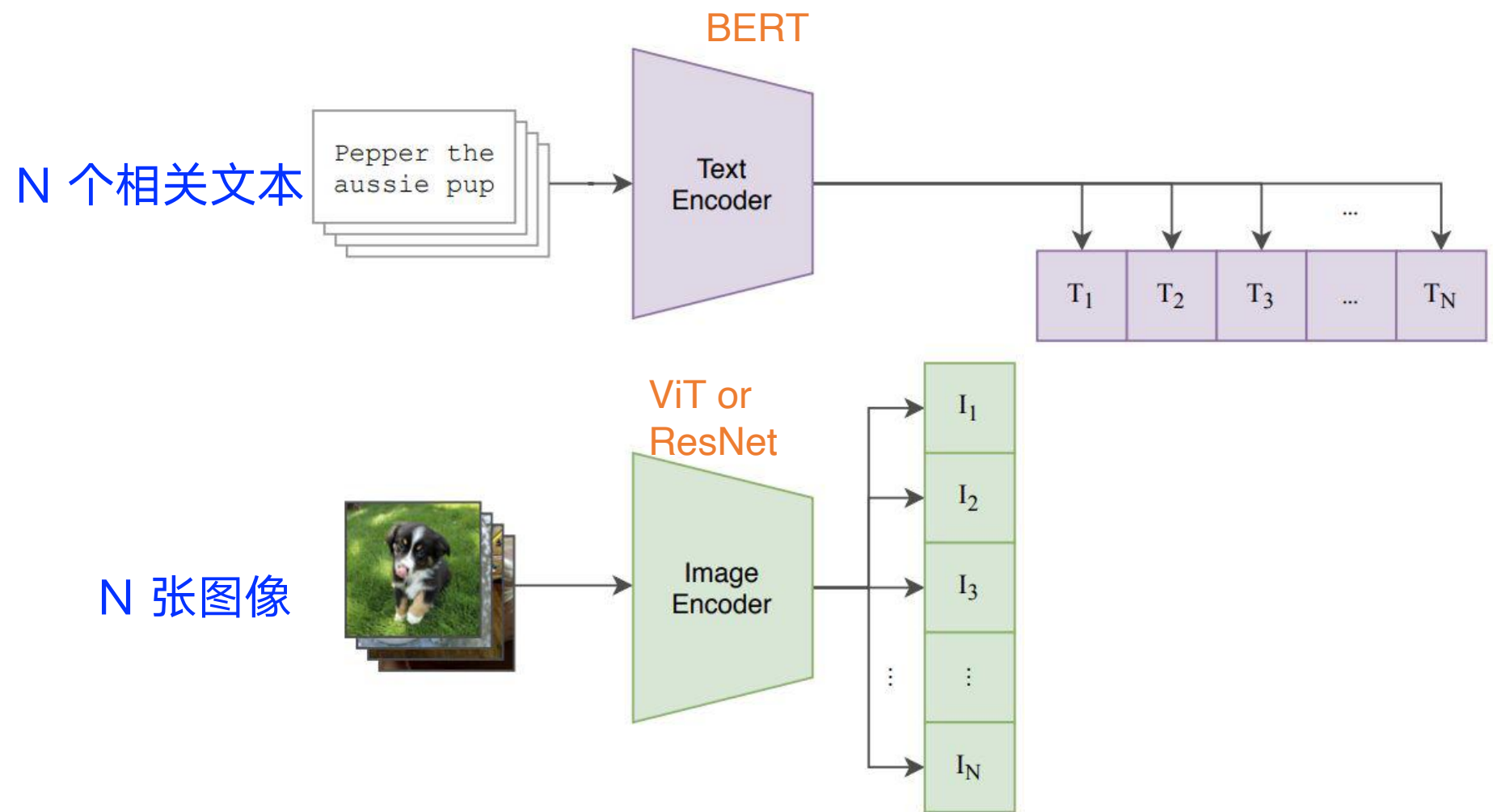
- 4) 从不利用或对齐自然语言中的标签名称

从自然语言监督中学习可转移的视觉模型

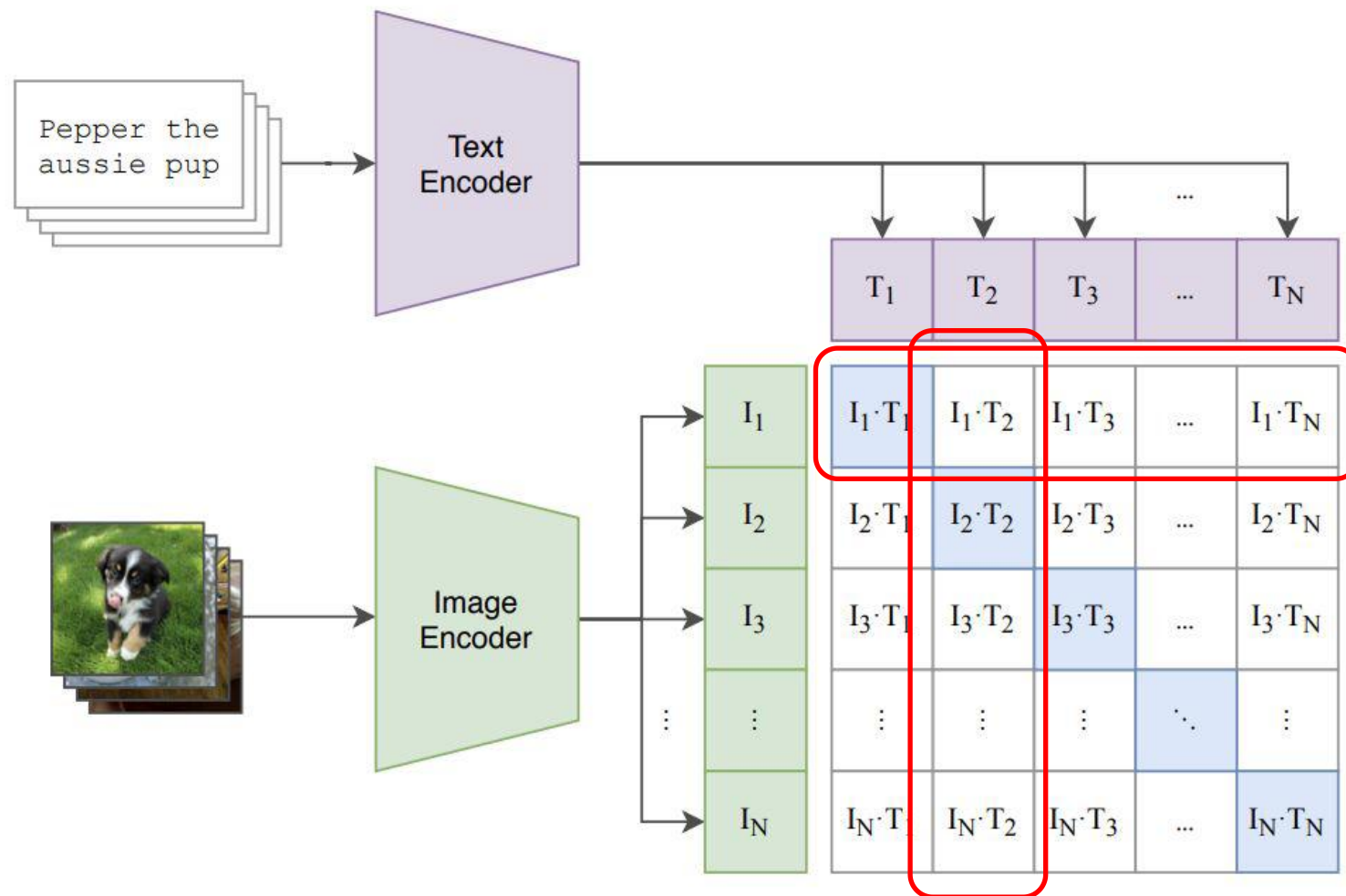
Radford et. al, Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

- 构建训练数据集——WebImageText (WIT)
  - 从互联网上各种公开可用的来源收集了 4 亿个 (图像, 文本) 对
  - 进行一些预处理以使数据集涵盖尽可能广泛的视觉概念, 并实现大致平衡的类别分布
- 基于收集到的<图像, 文本>数据集可能的预训练方式
  - 1) 方式一: 图像 -> 确切的文本 (Image captioning)
  - 2) 方式二: 图像 -> 相关文本的词袋 (预测某个单词是否出现, 图像二分类)
  - 3) 方式三: 图像和文本之间的对比学习

- 方式三：通过图像和相关文本之间的对比学习进行预训练

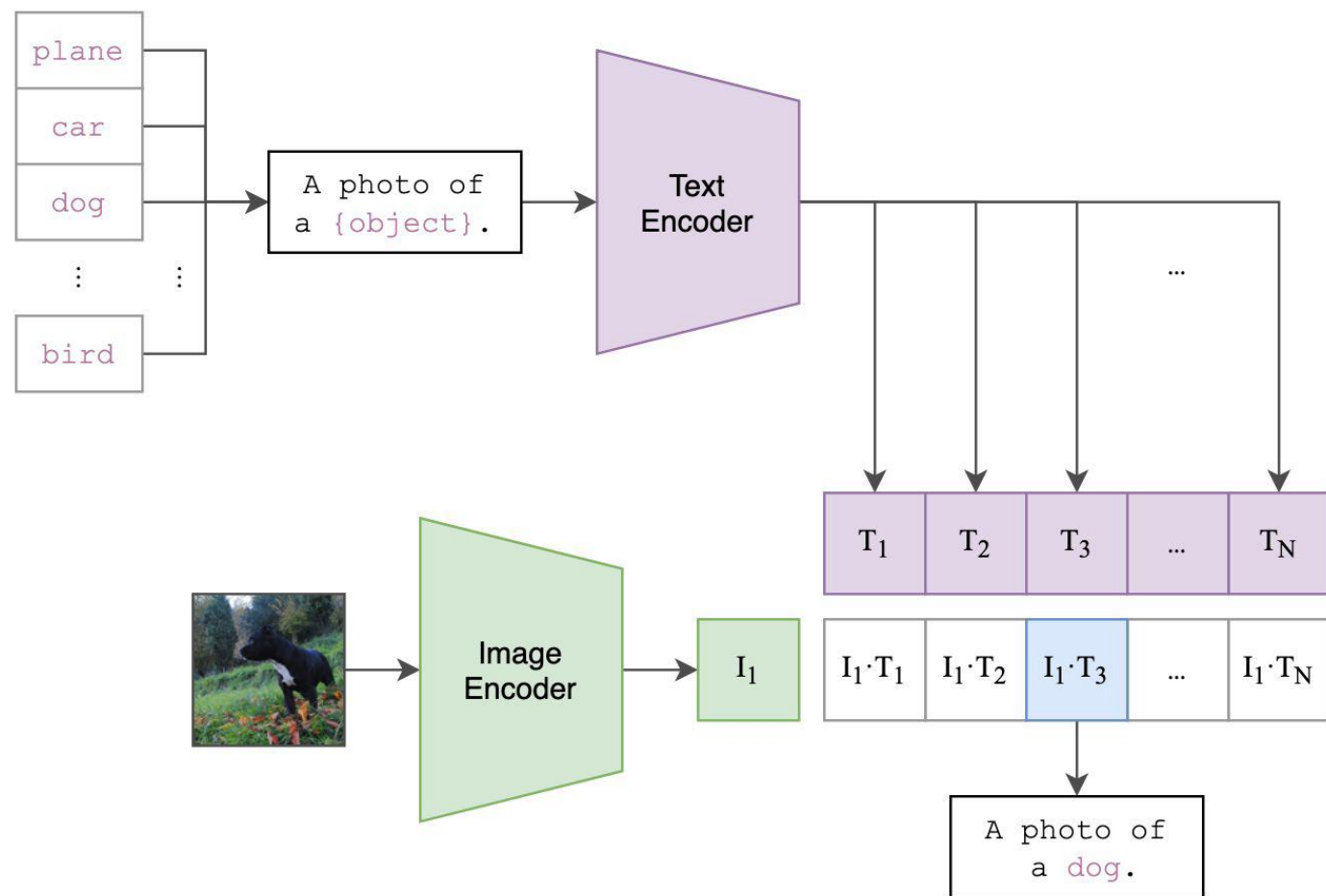


➤ 分别沿文本和图像维度进行对比训练

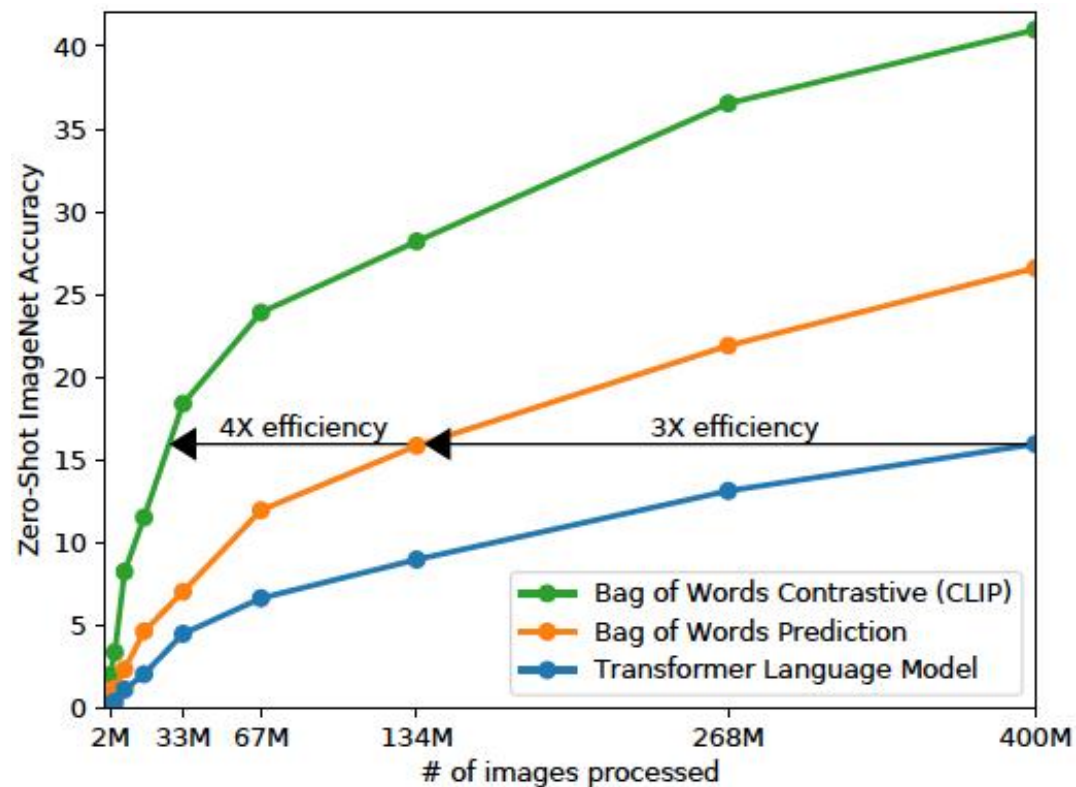


- 训练参数说明
  - 使用来自互联网的 4 亿个图像-文本对进行训练
  - Batch size 为 32768
  - 在数据集上进行 32 个 epoch
  - 余弦学习率衰减

- 如何使用 CLIP 进行零样本学习



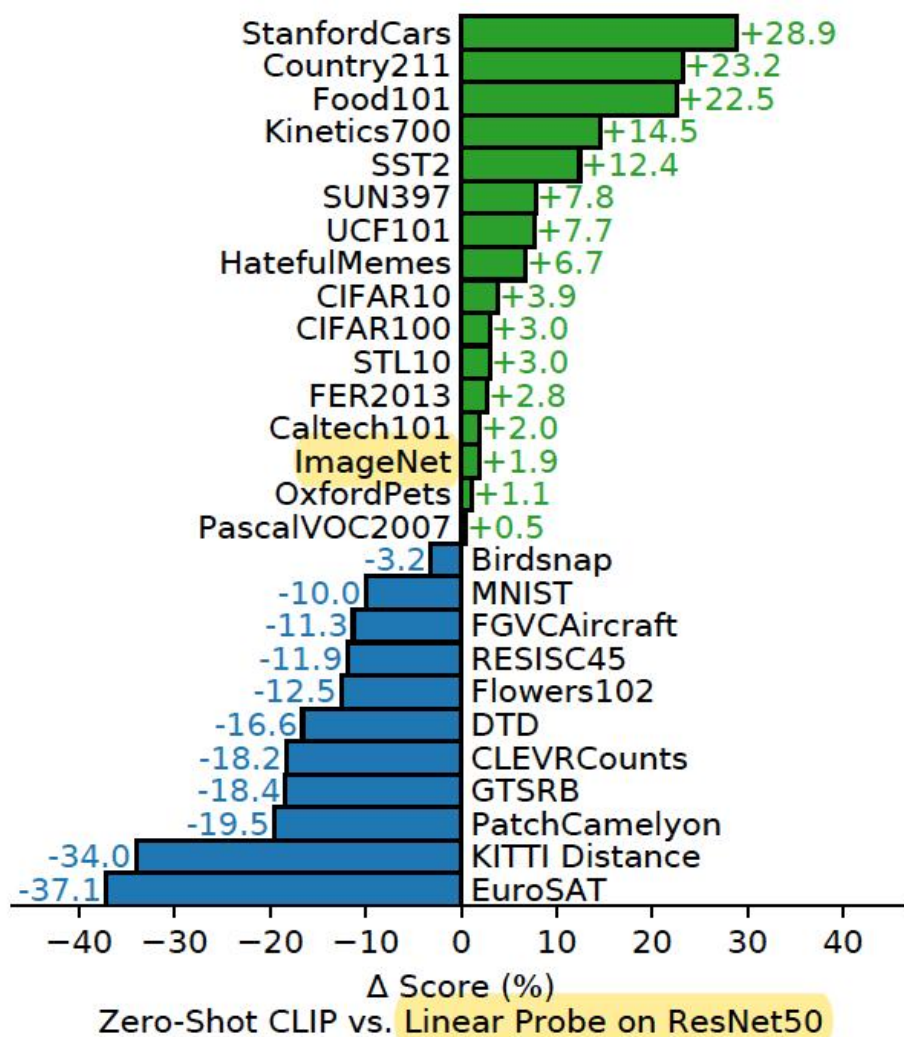
- 在 ImageNet 准确率的零样本性能上比较三种预训练方法



- Transformer Language Model: 预训练方式一
- Bag of Words Prediction: 预训练方式二
- Bag of Words Contrastive: 预训练方式三

对比学习预训练方式在收敛速度和最终性能方面都展现出突出优势

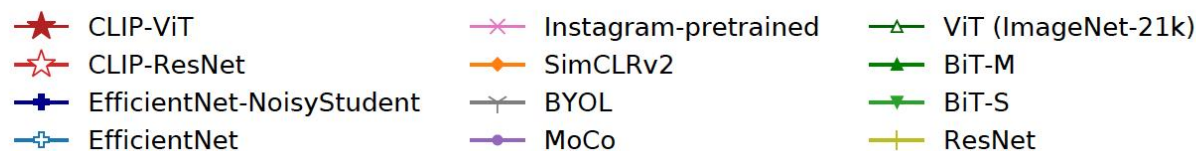
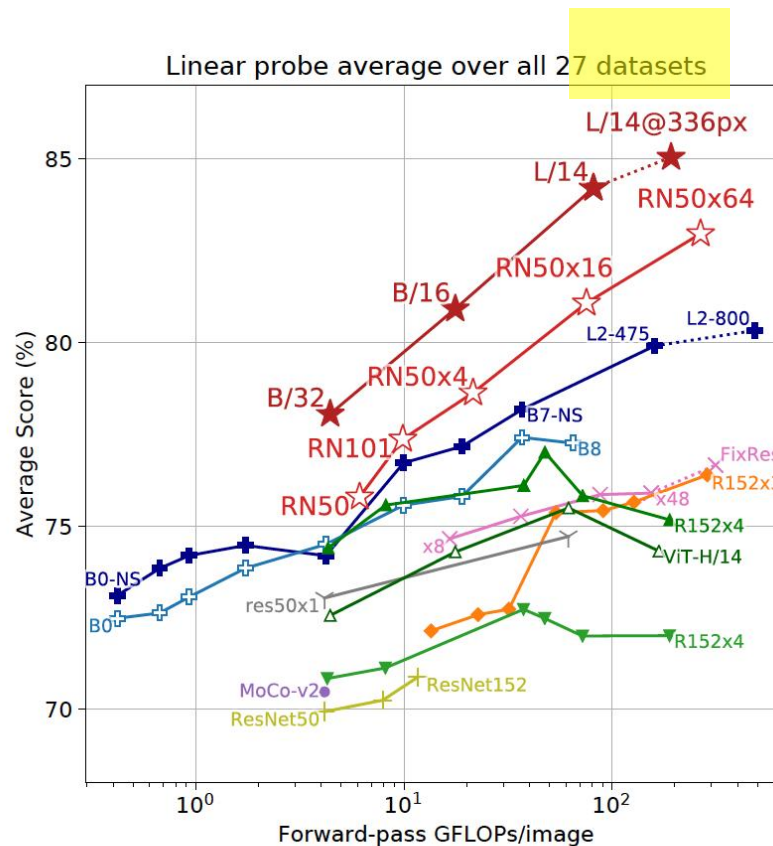
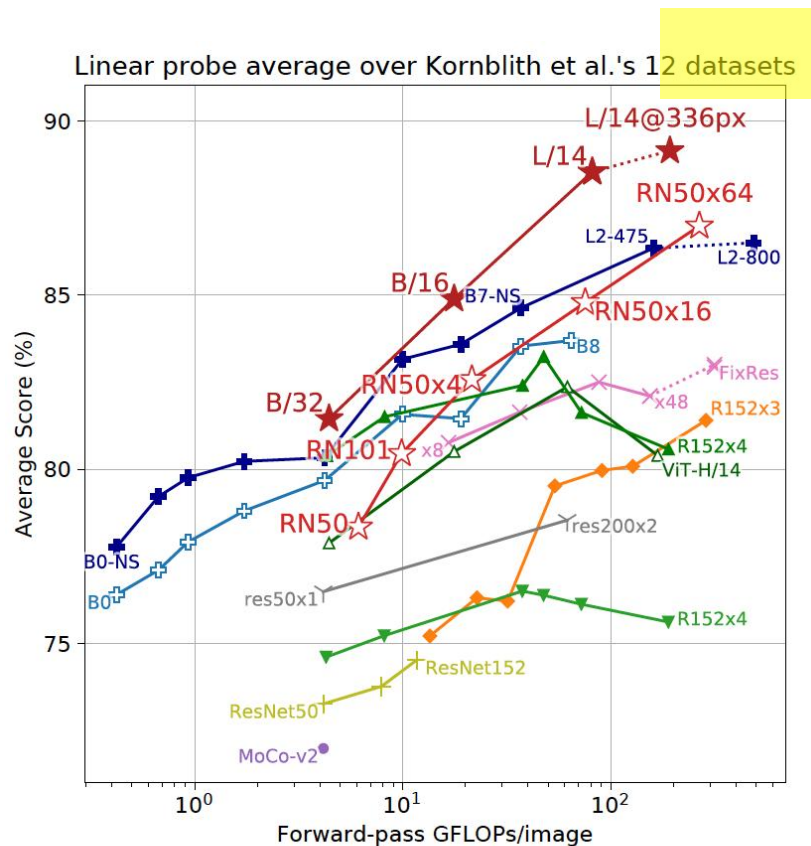
- 零样本 CLIP vs ResNet 50 特征上完全监督的线性探针









➤ 在大多数自然图像数据集上，零样本训练的 CLIP性能超过ResNet 50特征的完全监督线性探测性能，包括ImageNet数据集

➤ CLIP 只在几个特殊数据集上表现较弱，例如卫星图像分类、淋巴结肿瘤检测、合成场景中物体的计数、德国交通标志识别

- CLIP 学习到的视觉表示的线性探测



- 对数据分布偏移的强鲁棒性

	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	$\Delta$ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

# 大纲

---

- 基于Transformer的视觉处理架构ViT
- CLIP
- BLIP

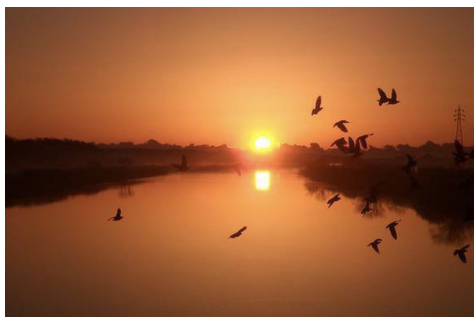
- CLIP 模型的局限

- 模型视角

- 仅采用基于编码器的结构，使其难以应用于涉及文本生成的任务，如Image captioning

- 数据视角

- 从互联网上抓取的 (图像，文本) 对是嘈杂的，对学习到的表示产生负面影响



from bridge  
near my  
house



blue sky bakery  
in sunset park

Junnan Li et. al, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022

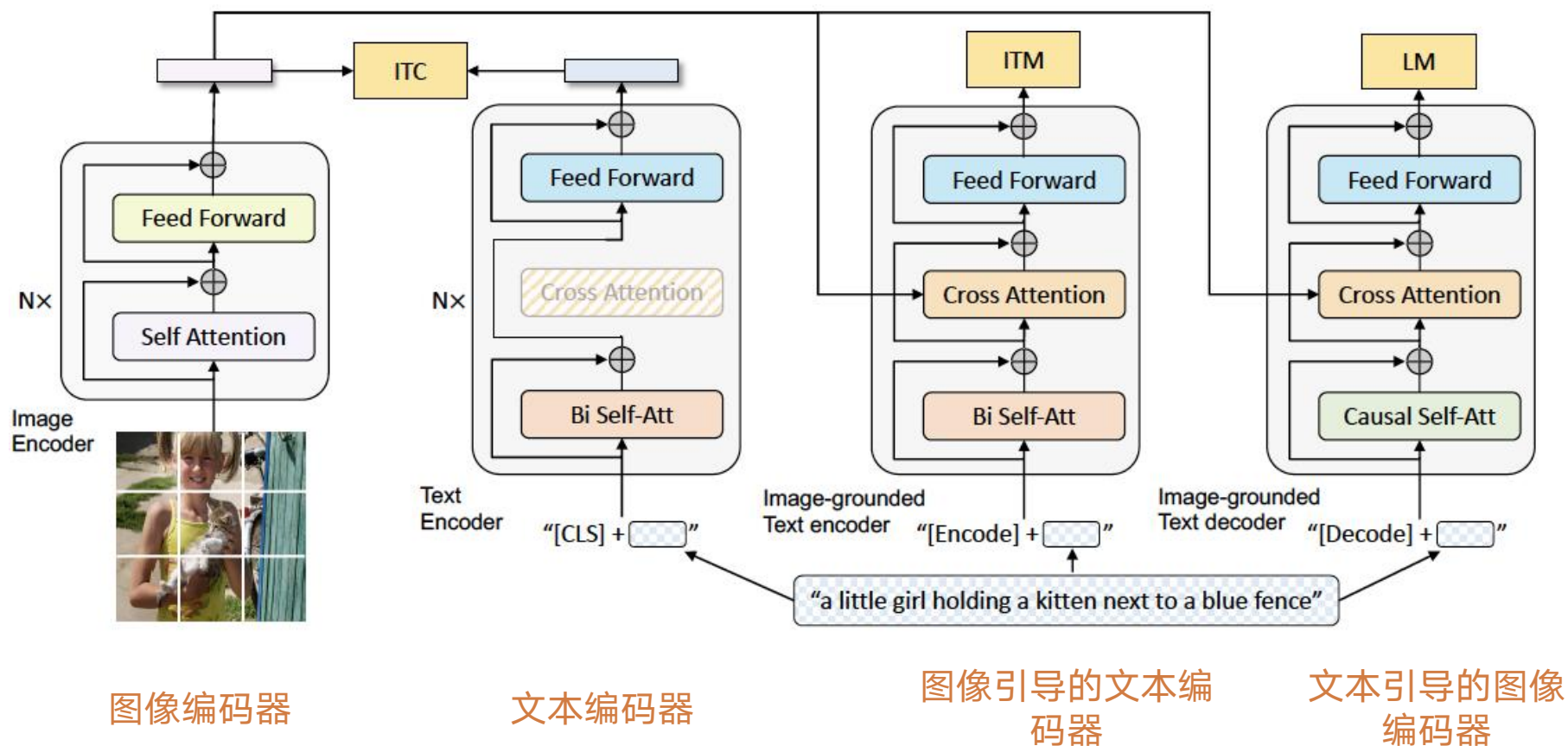
- 对于第一个问题，提出一个具有多个组件的结构

➤ 图像编码器

➤ 图像引导的文本编码器

➤ 文本编码器

➤ 文本引导的图像编码器



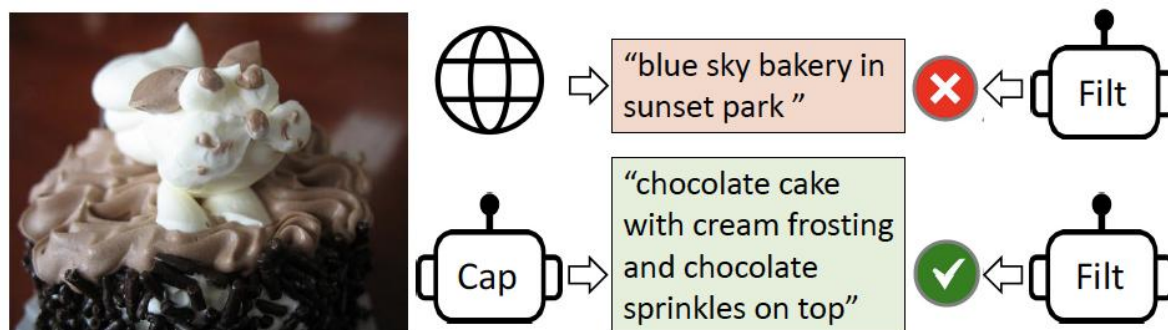
- 模型结构规格
  - 图像编码器：默认为 ViT-B
  - 文本编码器：默认为 BERT<sub>base</sub>
- 训练数据集
  - 总共 14M 张图像 (主要)
    - 两个人工注释的数据集：COCO 和 Visual Genome
    - 两个网络数据集：Conceptual 12M, SBU 字幕
  - 一个额外的具有 1.15 亿张图像的网络数据集 (可选)

BLIP训练数据集的大小、训练batch size、模型参数量都比 CLIP 小得多

- 对于第二个问题，提出一种字幕和过滤 (CapFilt) 机制

- 过滤不正确的字幕

- 用生成的字幕替换它们



$T_w$ : "from bridge near my house"

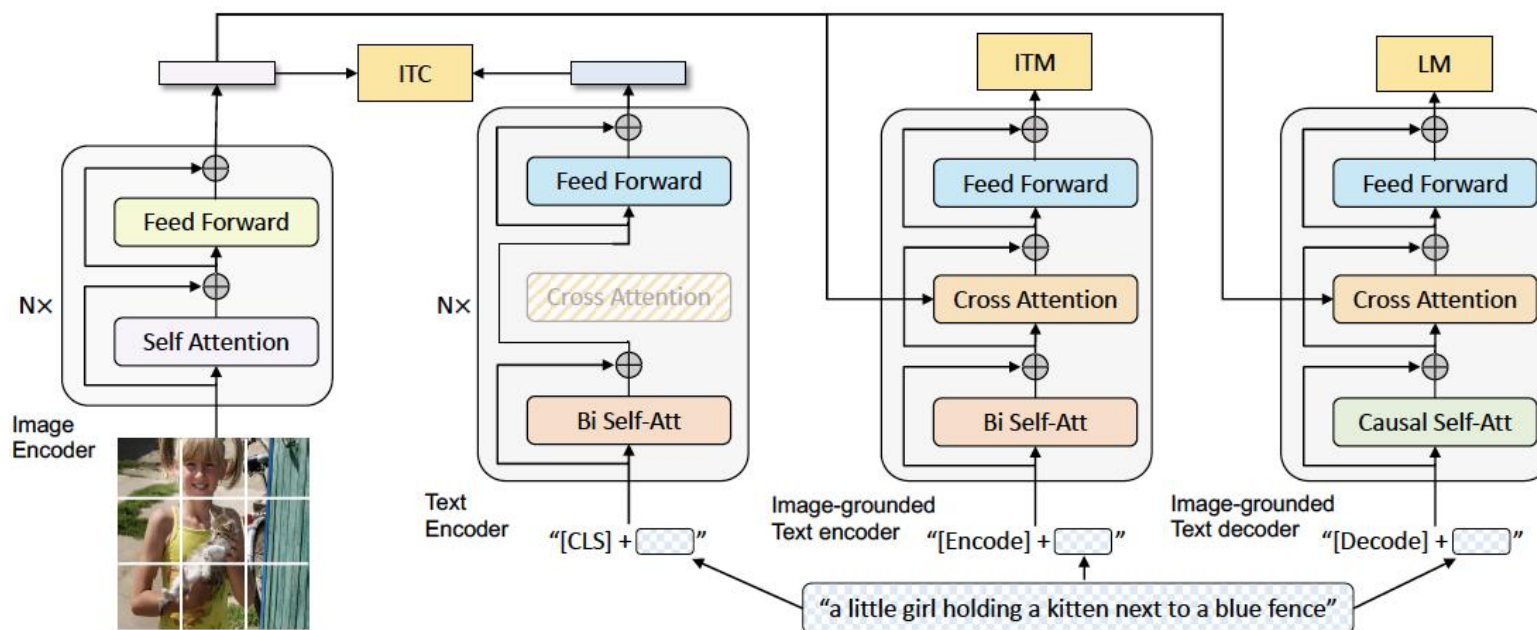
$T_s$ : "a flock of birds flying over a lake at sunset"



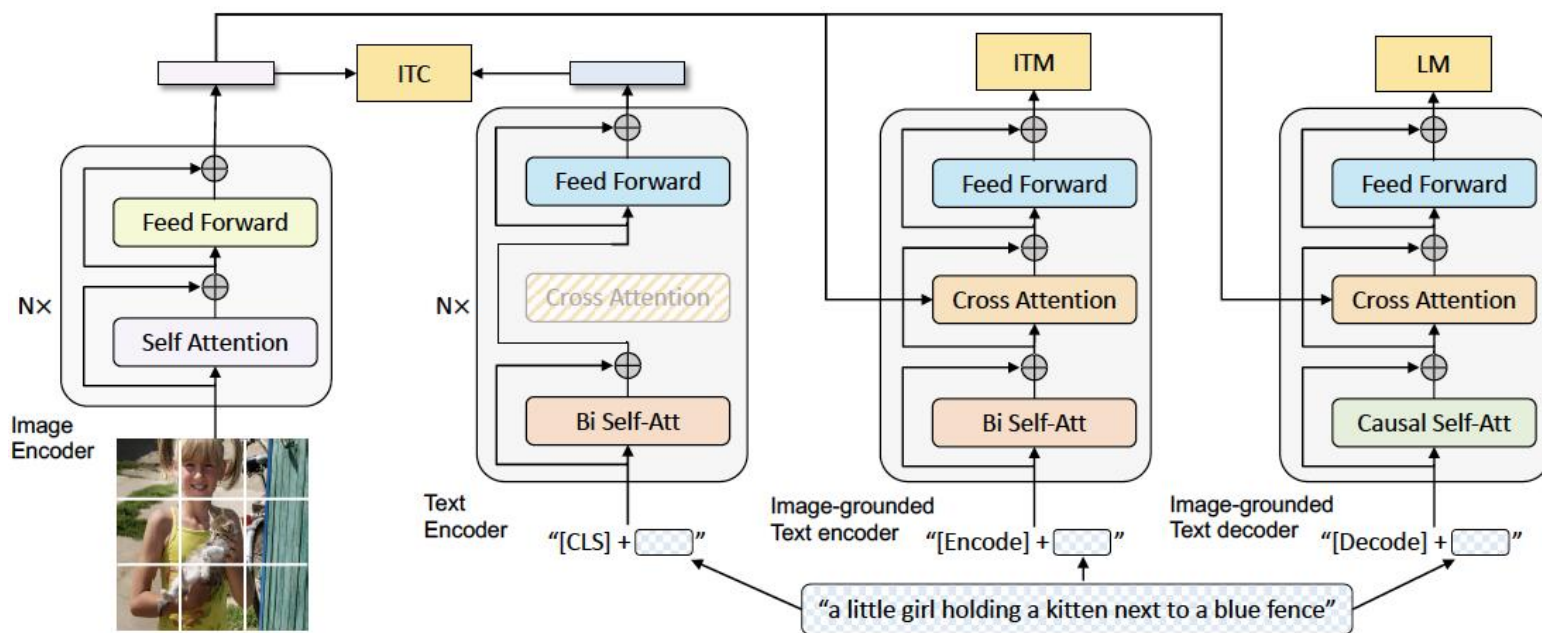
$T_w$ : "in front of a house door in Reichenfels, Austria"

$T_s$ : "a potted plant sitting on top of a pile of rocks"

- 预训练目标



- 图像-文本对比损失 (Image-Text Contrastive Loss, ITC): 通过类似 CLIP 的对比学习来对齐视觉 transformer 和文本 transformer 的特征空间
- 图像-文本匹配损失 (Image-Text Matching Loss, ITM): 用于预测图像-文本对是否匹配的 二元分类器。它学习图像相关的文本表示, 该表示通过交叉注意力机制 (cross attention) 涉及文本和图像。



- **语言建模损失 (Language Modeling Loss, LM):** 生成给定图像的文本描述。它通过最大化自回归方式的文本的对数似然来训练模型。

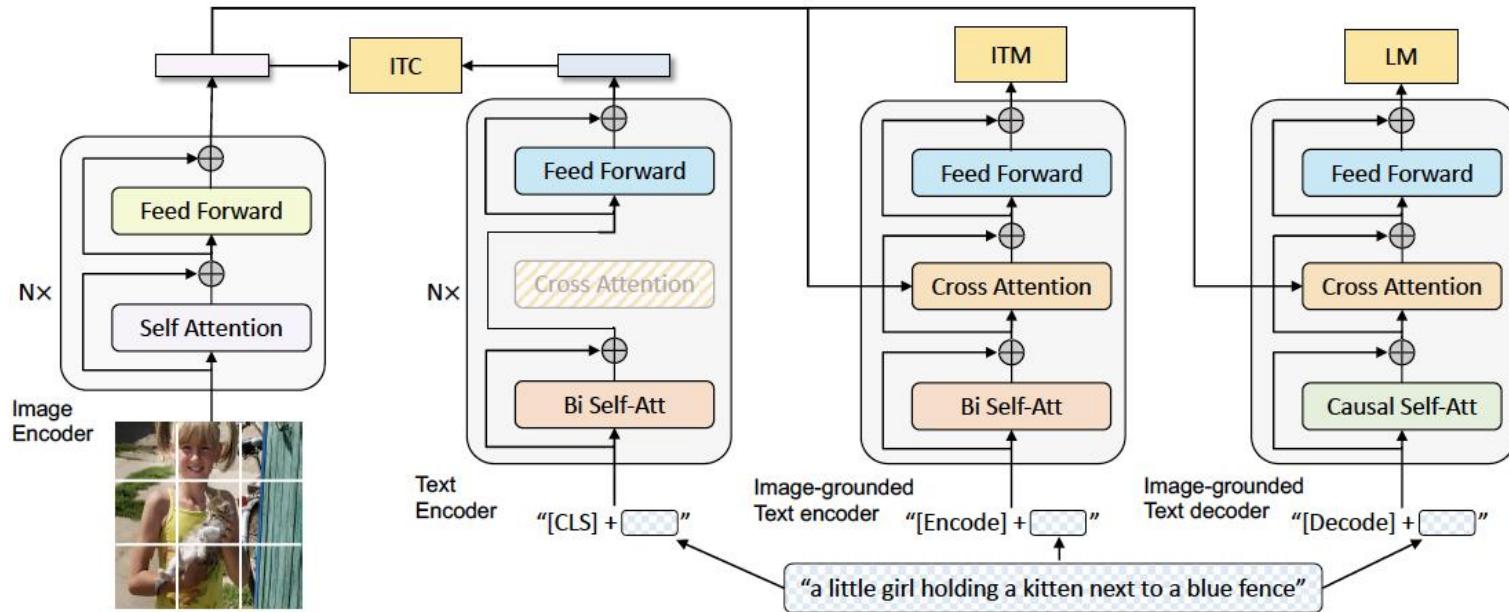
- 字幕 & 过滤 (CapFilt)

两个图像-文本对数据集

少量高质量的人工标注对  $\{(I_h, T_h)\}$

大量网络爬取的对  $\{(I_w, T_w)\}$

- 字幕生成器是以图像为依据的文本生成器，它会在高质量的  $\{(I_h, T_h)\}$  数据集上进行进一步的微调
- $\{(I_h, T_h)\}$  过滤器是以图像为依据的文本分类器，它会在高质量的  $\{(I_h, T_h)\}$  数据集上进行进一步的微调。



- CapFilt 的效果通过零样本检索任务进行评估

Pre-train dataset	Bootstrap		Vision backbone	Retrieval-FT (COCO)		Retrieval-ZS (Flickr)		Caption-FT (COCO)		Caption-ZS (NoCaps)	
	C	F		TR@1	IR@1	TR@1	IR@1	B@4	CIDEr	CIDEr	SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	✓ <sub>B</sub>		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓ <sub>B</sub>	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓ <sub>B</sub>	✓ <sub>B</sub>		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓ <sub>B</sub>	✓ <sub>B</sub>		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓ <sub>L</sub>	✓ <sub>L</sub>		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓ <sub>L</sub>	✓ <sub>L</sub>		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Table 1. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping. Downstream tasks include image-text retrieval and image captioning with finetuning (FT) and zero-shot (ZS) settings. TR / IR@1: recall@1 for text retrieval / image retrieval. ✓<sub>B/L</sub>: captioner or filter uses ViT-B / ViT-L as vision backbone.

- 图像-文本检索结果，其中文本和图像编码器在 COCO 和 Flickr30K 数据集上进行了微调

Method	Pre-train # Images	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER (Chen et al., 2020)	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA (Gan et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2021b)	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN (Jia et al., 2021)	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF (Li et al., 2021a)	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
BLIP	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	<b>100.0</b>	87.2	97.5	98.8
BLIP	129M	<b>81.9</b>	95.4	97.8	<b>64.3</b>	85.7	91.5	<b>97.3</b>	<b>99.9</b>	<b>100.0</b>	87.3	97.6	<b>98.9</b>
BLIP <sub>CapFilt-L</sub>	129M	81.2	<b>95.7</b>	<b>97.9</b>	64.1	<b>85.8</b>	<b>91.6</b>	97.2	<b>99.9</b>	<b>100.0</b>	<b>87.5</b>	<b>97.7</b>	<b>98.9</b>
BLIP <sub>ViT-L</sub>	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0

Table 5. Comparison with state-of-the-art image-text retrieval methods, finetuned on COCO and Flickr30K datasets. BLIP<sub>CapFilt-L</sub> pre-trains a model with ViT-B backbone using a dataset bootstrapped by captioner and filter with ViT-L.

- 图像字幕（Image Captioning）结果

Method	Pre-train #Images	NoCaps validation								COCO Caption Karpathy test	
		in-domain		near-domain		out-domain		overall		B@4	C
		C	S	C	S	C	S	C	S		
Enc-Dec (Changpinyo et al., 2021)	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL <sup>†</sup> (Zhang et al., 2021)	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
LEMON <sub>base</sub> <sup>†</sup> (Hu et al., 2021)	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
LEMON <sub>base</sub> <sup>†</sup> (Hu et al., 2021)	200M	107.7	14.7	106.2	14.3	107.9	13.1	106.8	14.1	<b>40.3</b>	<b>133.3</b>
BLIP	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
BLIP	129M	109.1	14.8	105.8	14.4	105.7	13.7	106.3	14.3	39.4	131.4
BLIP <sub>CapFilt-L</sub>	129M	<b>111.8</b>	<b>14.9</b>	<b>108.6</b>	<b>14.8</b>	<b>111.5</b>	<b>14.2</b>	<b>109.6</b>	<b>14.7</b>	39.7	<b>133.3</b>
LEMON <sub>large</sub> <sup>†</sup> (Hu et al., 2021)	200M	116.9	15.8	113.3	15.1	111.3	14.0	113.4	15.0	40.6	135.7
SimVLM <sub>huge</sub> (Wang et al., 2021)	1.8B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP <sub>ViT-L</sub>	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7

Table 7. Comparison with state-of-the-art image captioning methods on NoCaps and COCO Caption. All methods optimize the cross-entropy loss during finetuning. C: CIDEr, S: SPICE, B@4: BLEU@4. BLIP<sub>CapFilt-L</sub> is pre-trained on a dataset bootstrapped by captioner and filter with ViT-L. VinVL<sup>†</sup> and LEMON<sup>†</sup> require an object detector pre-trained on 2.5M images with human-annotated bounding boxes and high resolution (800×1333) input images. SimVLM<sub>huge</sub> uses 13× more training data and a larger vision backbone than ViT-L.

视觉问答与自然语言视觉推理 (Natural Language Visual Reasoning, NLVR<sup>2</sup>) 的结果

Method	Pre-train #Images	VQA		NLVR <sup>2</sup>	
		test-dev	test-std	dev	test-P
LXMERT	180K	72.42	72.54	74.90	74.50
UNITER	4M	72.70	72.91	77.18	77.85
VL-T5/BART	180K	-	71.3	-	73.6
OSCAR	4M	73.16	73.44	78.07	78.36
SOHO	219K	73.25	73.47	76.37	77.32
VILLA	4M	73.59	73.67	78.39	79.30
UNIMO	5.6M	75.06	75.27	-	-
ALBEF	14M	75.84	76.04	82.55	83.14
SimVLM <sub>base</sub> <sup>†</sup>	1.8B	77.87	78.14	81.72	81.77
BLIP	14M	77.54	77.62	<b>82.67</b>	82.30
BLIP	129M	78.24	78.17	82.48	<b>83.08</b>
BLIP <sub>CapFilt-L</sub>	129M	<b>78.25</b>	<b>78.32</b>	82.15	82.24

Table 8. Comparison with state-of-the-art methods on VQA and NLVR<sup>2</sup>. ALBEF performs an extra pre-training step for NLVR<sup>2</sup>. SimVLM<sup>†</sup> uses 13× more training data and a larger vision backbone (ResNet+ViT) than BLIP.

# 课堂小结

---

- 基于Transformer的视觉处理模型-ViT
- 视觉-文本多模态表示与理解模型CLIP
- 视觉-文本多模态表示与理解模型BLIP