Pretrained ViT as Vision Encoder

Paper Reading by Zhiying Lu

2023.12.12

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ViT
ICLR 2020

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu[†*] Yutong Lin[†*] Yue Cao[*] Han Hu[*‡] Yixuan Wei[†]
Zheng Zhang Stephen Lin Baining Guo
Microsoft Research Asia
{v-zeliu1,v-yutlin,yuecao,hanhu,v-yixwe,zhez,stevelin,bainguo}@microsoft.com

Swin Transformer
ICCV 2021 (best paper)

Learning Transferable Visual Models From Natural Language Supervision

---

Alec Radford[*1] **Jong Wook Kim**[*1] **Chris Hallacy**[1] **Aditya Ramesh**[1] **Gabriel Goh**[1] **Sandhini Agarwal**[1] **Girish Sastry**[1] **Amanda Askell**[1] **Pamela Mishkin**[1] **Jack Clark**[1] **Gretchen Krueger**[1] **Ilya Sutskever**[1]

CLIP
ICML 2021

Sigmoid Loss for Language Image Pre-Training

Xiaohua Zhai[*] Basil Mustafa Alexander Kolesnikov Lucas Beyer[*]
Google DeepMind, Zürich, Switzerland
{xzhai, basilm, akolesnikov, lbeyer}@google.com

SigLIP
ICCV 2023

Masked Autoencoders Are Scalable Vision Learners

Kaiming He[*,†] Xinlei Chen[*] Saining Xie Yanghao Li Piotr Dollár Ross Girshick

[*]equal technical contribution [†]project lead

Facebook AI Research (FAIR)

MAE
CVPR 2022

DINOv2: Learning Robust Visual Features without Supervision

Maxime Oquab[**], Timothée Darcet[**], Théo Moutakanni[**],
Huy V. Vo[*], Marc Szafraniec[*], Vasil Khalidov[*], Pierre Fernandez, Daniel Haziza,
Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba,
Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat,
Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal[1],
Patrick Labatut[*], Armand Joulin[*], Piotr Bojanowski[*]

Meta AI Research [1]*Inria*

[*]core team [**]equal contribution
Reviewed on OpenReview: https://openreview.net/forum?id=a68SUt6zFt

DINOv2
TMLR 2024

- 研究背景
- Fully-Supervised
- Weakly-Supervised
- Self-Supervised
- 总结

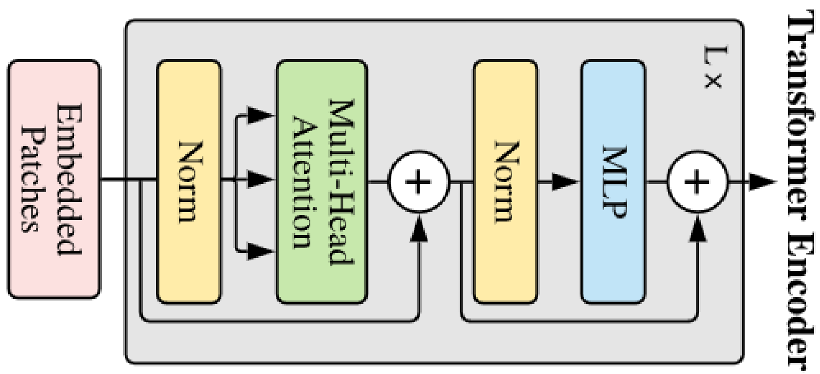
研究背景

- 当今研究中，如何利用预训练模型进行迁移成为主流
- 开源社区中存在众多预训练模型，由于各自预训练方法和预训练数据集不同，使得模型具有不同的表征模式
- 根据预训练的不同，大致可以分为全监督、弱监督、自监督三种

Joint Tuning	Supervised	Visual Tokenizer	# Pretraining Images	VQA Acc	Captioning CIDEr	Captioning SPICE	OC Acc	MCI Acc	Avg
×	Fully	DeiT [16]	1.28 M	48.3	65.8	15.9	37.5	83.6	58.8
	Self	DINO [19]	1.28 M	50.1	45.0	13.5	46.5	80.8	55.6
		MAE [18]	1.28 M	48.4	37.3	11.8	**47.5**	82.7	53.4
		DINOv2 [20]	142 M	51.3	67.9	16.1	47.0	86.0	**63.1**
	Weakly	CLIP [17]	400 M	**52.2**	**69.3**	**16.6**	42.5	86.0	62.5
✓	Fully	DeiT [16]	1.28 M	50.7	38.4	10.0	41.0	86.9	54.3
	Self	DINO [19]	1.28 M	47.3	54.1	14.5	44.5	86.6	58.1
		MAE [18]	1.28 M	48.9	48.0	14.2	**47.5**	**88.7**	58.2
		DINOv2 [20]	142 M	50.5	49.6	13.0	43.5	84.1	56.9
	Weakly	CLIP [17]	400 M	47.7	64.2	15.4	45.5	88.0	61.4

研究背景

- 主要讨论基于ViT架构的预训练模型，因为ViT模型具有多项良好性质
- **Variable in length**：可计算任意尺度的特征，不受特征图形状影响
- **Scalable**：传统卷积网络需要设计金字塔架构，当网络扩大参数时调参较为困难，而ViT系列网络可直接堆叠层数并任意改变每层的维度
- **Global Field**：对整个序列具有全局感受野，不受限
- **Unified Architecture**：与NLP实现统一架构



研究背景

- 当今研究中，如何利用预训练模型进行迁移成为主流
- 开源社区中存在众多预训练模型，由于各自预训练方法和预训练数据集不同，使得模型具有不同的表征模式
- 根据预训练的不同，大致可以分为全监督、弱监督、自监督三种

Joint Tuning	Supervised	Visual Tokenizer	# Pretraining Images	VQA Acc	Captioning CIDEr	Captioning SPICE	OC Acc	MCI Acc	Avg
×	Fully	DeiT [16]	1.28 M	48.3	65.8	15.9	37.5	83.6	58.8
	Self	DINO [19]	1.28 M	50.1	45.0	13.5	46.5	80.8	55.6
		MAE [18]	1.28 M	48.4	37.3	11.8	**47.5**	82.7	53.4
		DINOv2 [20]	142 M	51.3	67.9	16.1	47.0	86.0	**63.1**
	Weakly	CLIP [17]	400 M	**52.2**	**69.3**	**16.6**	42.5	86.0	62.5
✓	Fully	DeiT [16]	1.28 M	50.7	38.4	10.0	41.0	86.9	54.3
	Self	DINO [19]	1.28 M	47.3	54.1	14.5	44.5	86.6	58.1
		MAE [18]	1.28 M	48.9	48.0	14.2	**47.5**	**88.7**	58.2
		DINOv2 [20]	142 M	50.5	49.6	13.0	43.5	84.1	56.9
	Weakly	CLIP [17]	400 M	47.7	64.2	15.4	45.5	88.0	61.4

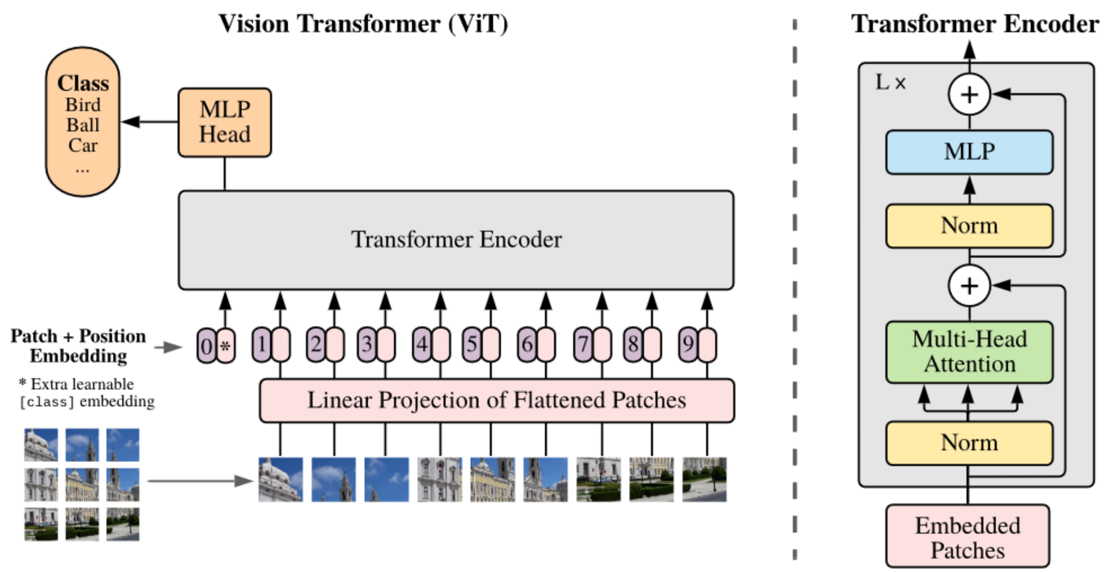
- 研究背景
- Fully-Supervised
- Weakly-Supervised
- Self-Supervised
- 总结

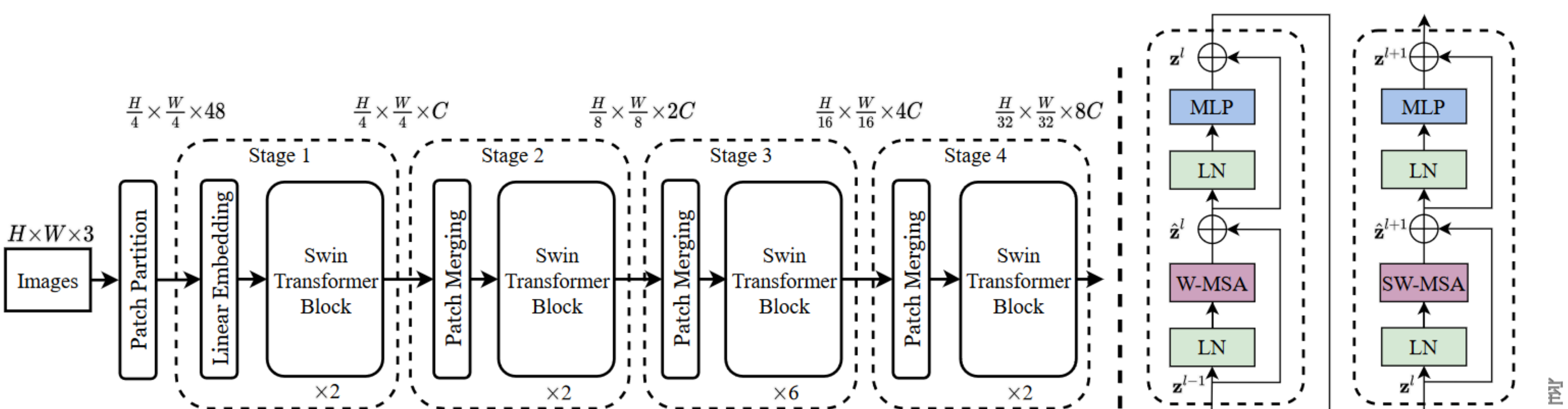
Fully-Supervised

- 全监督即每张图像与对应类别作为成对数据进行训练，使用交叉熵函数
- 绝大部分工作考虑如何在原版ViT的基础上改进网络结构，包括层级化设计、注意力、MLP的设计等
(详见之前的backbone主题报告)
- 全监督的主要预训练数据集为：ImageNet21k，ImageNet1k，JFT300M，iNaturalist等

small model ~ 4.5G	DeiT-S [43]	22	4.6	79.9
	Swin-T [32]	29	4.5	81.3
	ConvNeXt-T [33]	29	4.5	82.1
	Focal-T [56]	29	4.9	82.2
	InceptionNeXt-T [60]	28	4.2	82.3
	FocalNet-T [57]	29	4.5	82.3
	RegionViT-S [2]	31	5.3	82.6
	CSWin-T [9]	23	4.3	82.7
	MPViT-S [26]	23	4.7	83.0
	ScalableViT-S [58]	32	4.2	83.1
	MOAT-0 [55]	28	5.7	83.3
	Ortho-S [22]	24	4.5	83.4
	InternImage-T [49]	30	5.0	83.5
	CMT-S [15]	25	4.0	83.5
	FAT-B3 [13]	29	4.4	83.6
	MaxViT-T [44]	31	5.6	83.6
	SMT-S [31]	20	4.8	83.7
	BiFormer-S [66]	26	4.5	83.8
	RMT-S	27	4.5	84.1

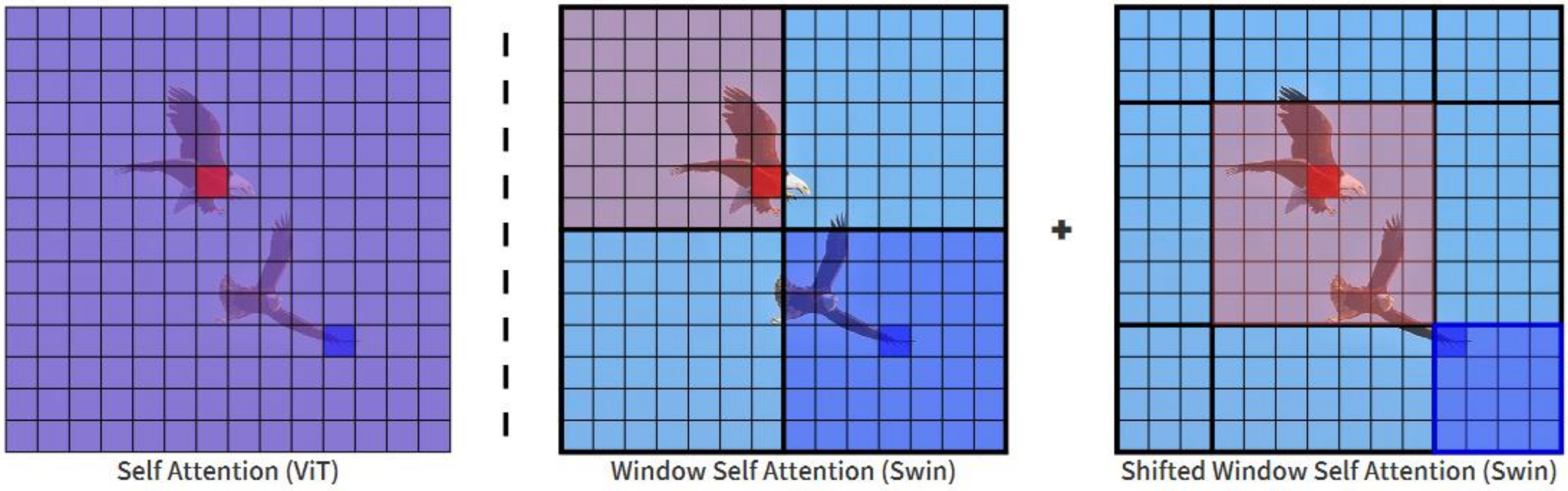
ViT与Swin



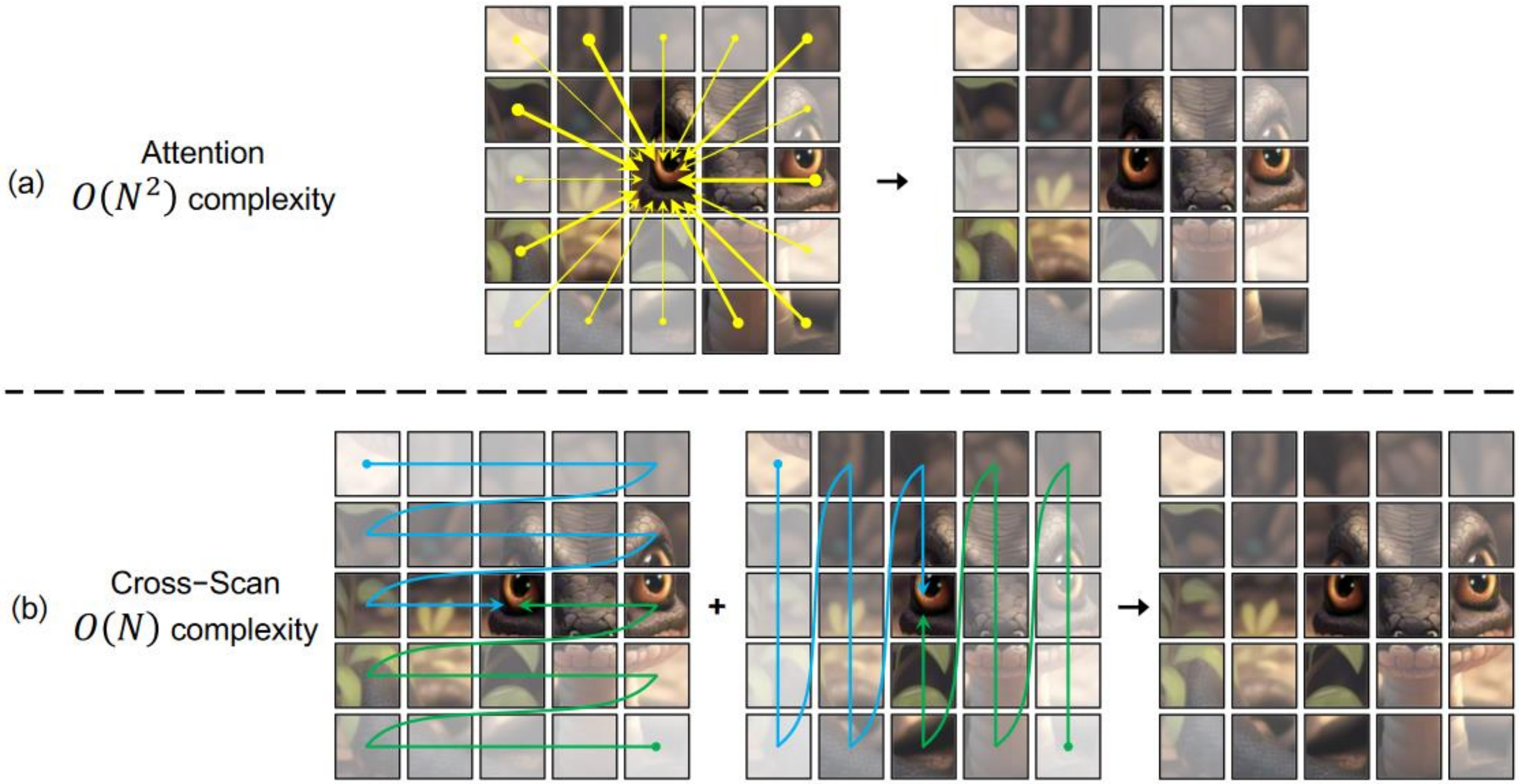


(a) Architecture (b) Two Successive Swin Transformer Blocks

ViT与Swin


Self Attention (ViT)
Window Self Attention (Swin)
Shifted Window Self Attention (Swin)

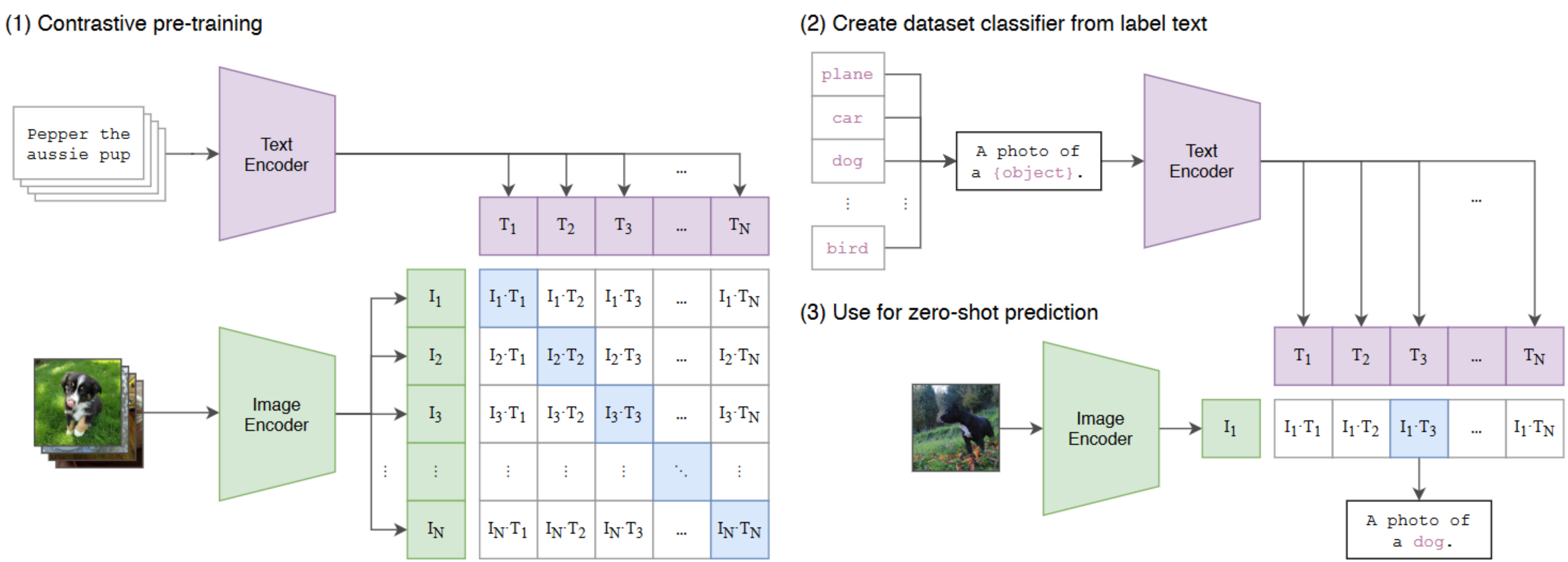
全监督的新发展 Mamba?



- 研究背景
- Fully-Supervised
- Weakly-Supervised
- Self-Supervised
- 总结

Weakly-Supervised

- 主要讨论来自于进行视觉语言对比学习预训练的模型CLIP及其变体
- 弱监督即为使用图像与文字匹配预训练的方式，并不是传统全监督直接对应图像和label，而是在一个batch中匹配图像和文本，增强正例，排除负例



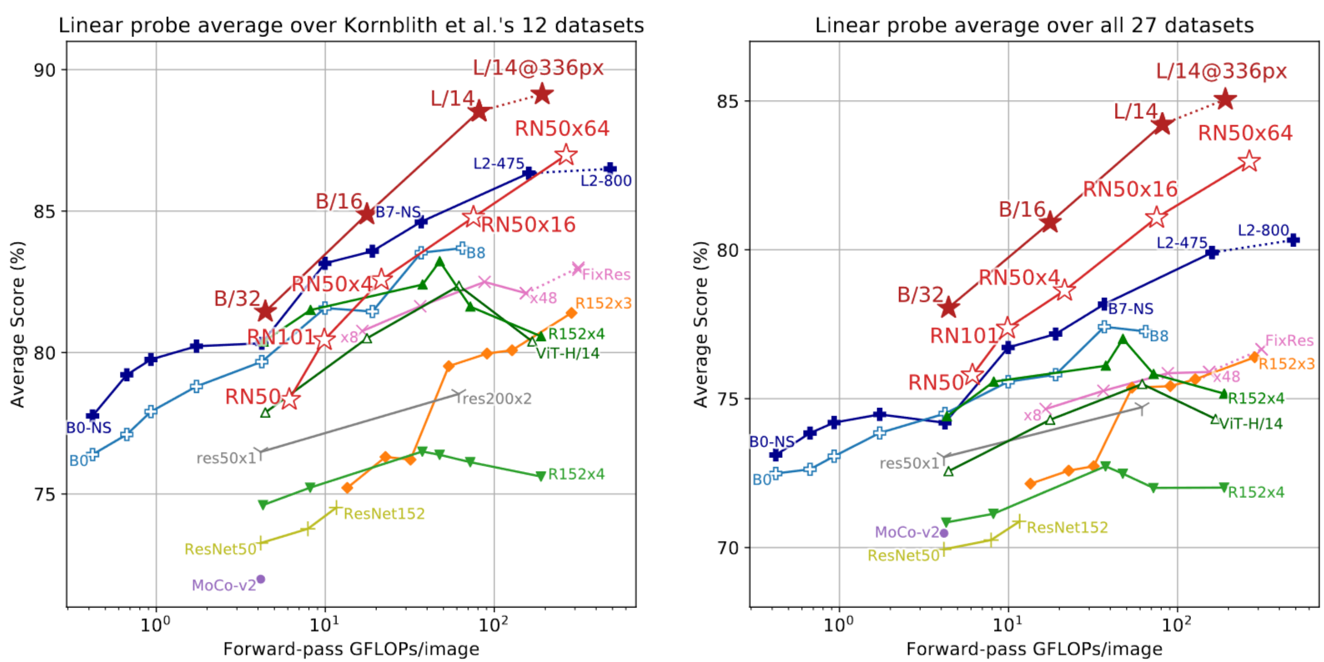
Weakly-Supervised

- CLIP的训练主要分为三部分：**数据集构造，训练目标，缩放模型**
- **数据集构造：**构造充分大数据集（之前的图文对数据集MSCOCO，Visual Genome都只有0.1M，YFCC100M的高质量数据只有15M），CLIP用500k个查询，每个查询20k的 image-text pair来构造了一个**400M大**小的WIT400M数据集
- **训练目标：**在一个N batch中利用cos距离匹配image-text，构造双向的对称的loss，**图像匹配文本，文本匹配图像**

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

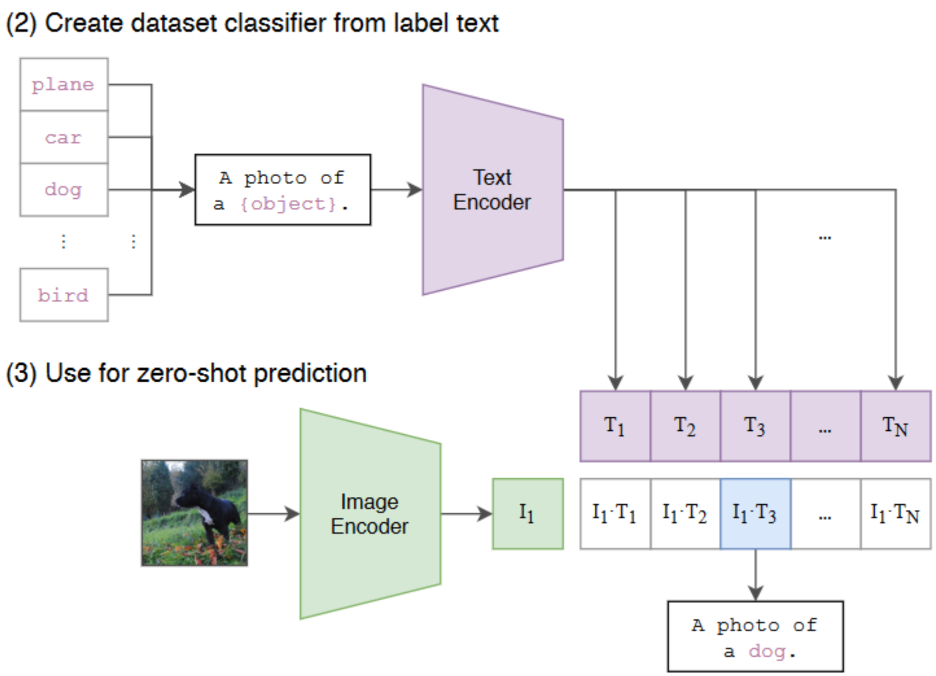
Weakly-Supervised

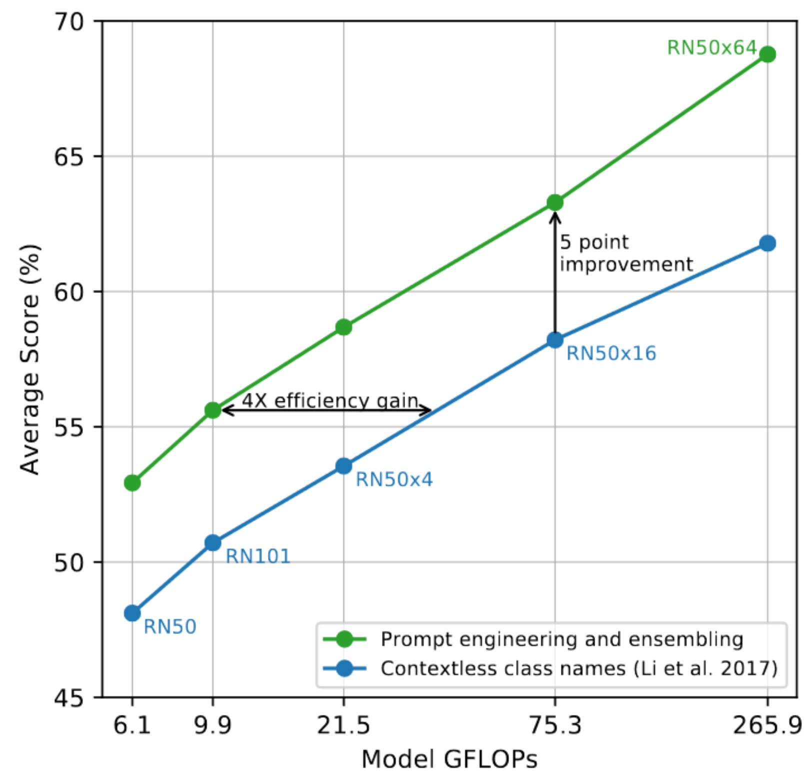
- **缩放模型**：CLIP采用双塔结构，图像编码器负责处理图像，文本编码器负责处理文本
- **文本编码器**：标准Transformer，每层宽度随着图像编码器设置，模型架构影响较小
- **图像编码器**：采用ResNet50或ViT架构，由于在匹配时文本段采用class token，因此在ResNet50的输出端还要加上一个attention层，并取输出class token作为最终图像特征。结果为，ViT优于ResNet，且ViT越大效果越好，最终规模为ViT-L/14-336px



Weakly-Supervised

- **Inference阶段：**图像端直接输入图像，文本段根据人工设置的文本，与图像进行匹配
- 使用合适的或集成的prompt可以达到较高的零样本效果，即对于同一个类别使用多个text template，于是诞生了prompt-engineering



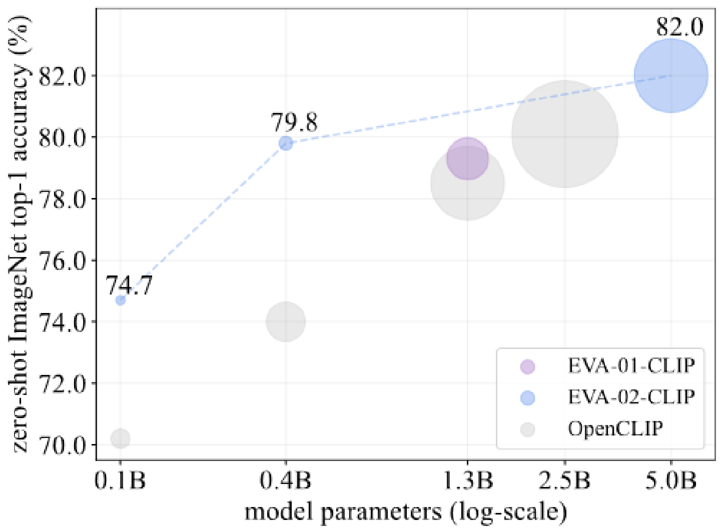


Weakly-Supervised

- CLIP的一大特点在于使用**双塔匹配训练**，但可以单塔使用，两塔的输出在一个相近的映射空间，即它可以**把图像映射到文本空间中，因此可以作为MLLM的visual encoder.**
- **由于CLIP数据未开源，**因此诞生了后续很多开源CLIP系的工作
- **OpenCLIP (CVPR2023)**：采用LAION2B数据集预训练，相对于以往的ViT-L (0.3B), 训练了更大的ViT-H (0.6B) 和ViT-G (1B)

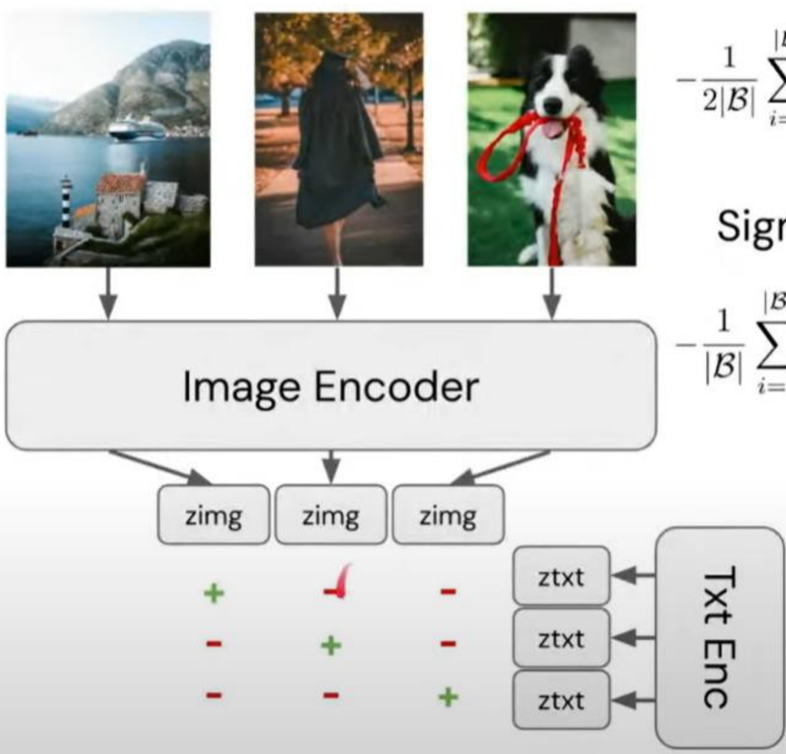
	Data	Arch.	ImageNet	VTAB+	COCO
CLIP [55]	WIT-400M	L/14	75.5	55.8	61.1
Ours	LAION-2B	L/14	75.2	54.6	71.1
Ours	LAION-2B	H/14	78.0	56.4	73.4

- EVA-CLIP: 采用Merge-2B (LAION+COCO) 进行训练，结合FLIP的对图像mask的操作，EVA (BeiT与iBOT结合) 权重初始化视觉编码器，CLIP权重初始化文本编码器。最大版本为EVA-CLIP-18B



SigLIP

From CLIP to SigLIP



Boat on a mountain-lake with lighthouse
Woman in dress standing on pathway
Cute dog sitting on grass with leash

Softmax-based (CLIP):

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\rightarrow\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\rightarrow\text{image softmax}}\right)$$

- 👎 Bi-directional
- 👎 Multiple global sums
- 👎 Weird learning task(?)

Sigmoid-based (SigLIP):

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

- 😍 Simpler
- 😍 Each entry individual
- 😍 works & scales better

Google DeepMind

SigLIP

- 采用Sigmoid Loss而不是对比学习loss进行匹配，使得每个样本对的loss计算无需计算整个batch里的值，容易扩大batchsize，同时Sigmoid Loss在小batch时也更好
- 使用谷歌PaLM同款的WebLI数据集（文本包含100+种类语言）进行预训练，选取其中的英语部分

Algorithm 1 Sigmoid loss pseudo-implementation.

```
# img_emb      : image model embedding [n, dim]
# txt_emb      : text model embedding [n, dim]
# t_prime, b   : learnable temperature and bias
# n            : mini-batch size

t = exp(t_prime)
zimg = l2_normalize(img_emb)
ztxt = l2_normalize(txt_emb)
logits = dot(zimg, ztxt.T) * t + b
labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
l = -sum(log_sigmoid(labels * logits)) / n
```

	Image	Text	BS	#TPUv4	Days	INet-0
SigLiT	B/8	L*	32 k	4	1	79.8
SigLiT	g/14	L	20 k	4	2	84.5
SigLIP	B/16	B	16 k	16	3	71.0
SigLIP	B/16	B	32 k	32	2	72.1
SigLIP	B/16	B	32 k	32	5	73.4

* *We use a variant of the L model with 12 layers.*

Batch Size	3 B		9 B	
	sigmoid	softmax	sigmoid	softmax
512	**51.5**	47.7	-	-
1 k	**57.3**	53.2	-	-
2 k	**62.1**	59.3	-	-
4 k	**65.3**	63.8	**68.4**	66.6
8 k	**68.6**	66.6	**70.6**	69.4
16 k	-	-	**72.3**	71.7
32 k	**69.9**	**69.9**	**73.4**	72.9
98 k	69.5	**69.7**	73.0	**73.2**
307 k	-	-	71.6	**72.6**

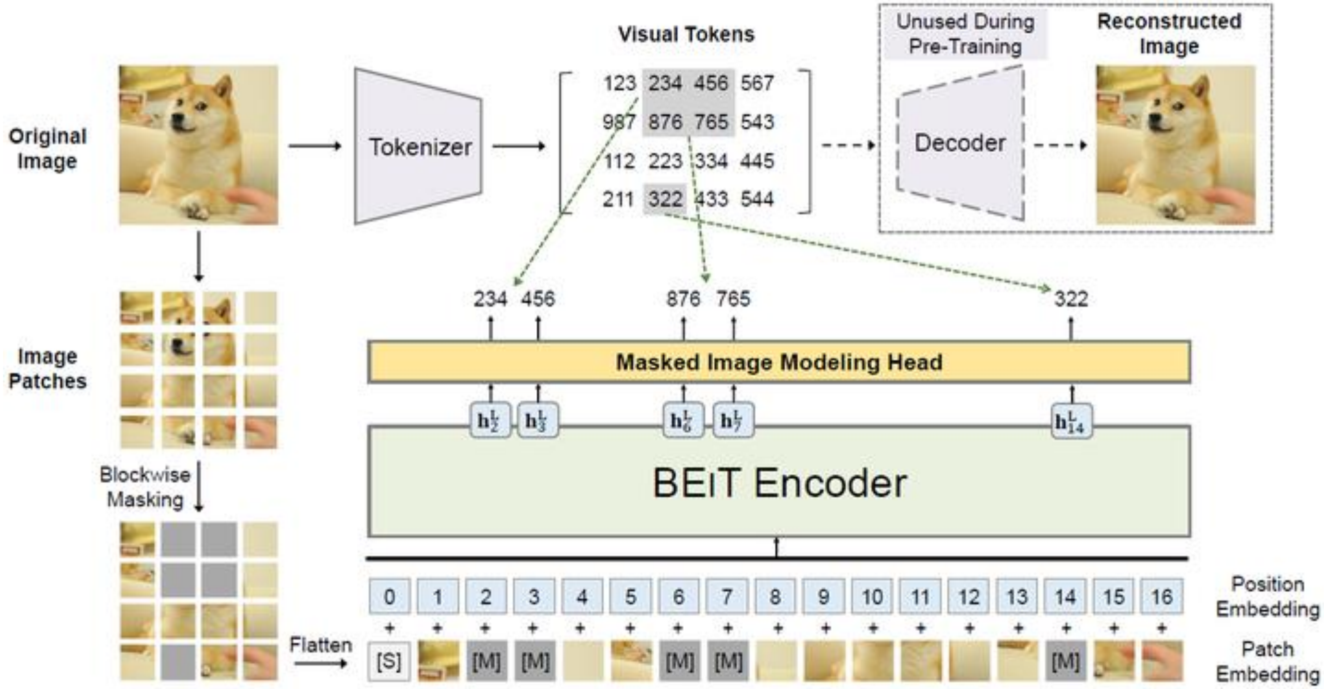
SigLIP

Method	Image Encoder		ImageNet-1k				COCO R@1	
	ViT size	# Patches	Validation	v2	ReaL	ObjectNet	I → T	T → I
CLIP	B	196	68.3	61.9	-	55.3	52.4	33.1
OpenCLIP	B	196	70.2	62.3	-	56.0	59.4	42.3
EVA-CLIP	B	196	74.7	67.0	-	62.3	58.7	42.2
SigLIP	B	196	**76.2**	**69.6**	82.8	**70.7**	**64.4**	**47.2**
SigLIP	B	256	76.7	70.0	83.1	71.3	65.1	47.4
SigLIP	B	576	78.6	72.1	84.5	73.8	67.5	49.7
SigLIP	B	1024	**79.2**	**73.0**	**84.9**	**74.7**	**67.6**	**50.4**
CLIP	L	256	75.5	69.0	-	69.9	56.3	36.5
OpenCLIP	L	256	74.0	61.1	-	66.4	62.1	46.1
CLIPA-v2	L	256	79.7	72.8	-	71.1	64.1	46.3
EVA-CLIP	L	256	79.8	72.9	-	75.3	63.7	47.5
SigLIP	L	256	**80.5**	**74.2**	**85.9**	**77.9**	**69.5**	**51.1**
CLIP	L	576	76.6	72.0	-	70.9	57.9	37.1
CLIPA-v2	L	576	80.3	73.5	-	73.1	65.5	47.2
EVA-CLIP	L	576	80.4	73.8	-	78.4	64.1	47.9
SigLIP	L	576	**82.1**	**75.9**	**87.0**	**81.0**	**70.6**	**52.7**
OpenCLIP	G (2B)	256	80.1	73.6	-	73.0	67.3	51.4
CLIPA-v2	H (630M)	576	81.8	75.6	-	77.4	67.2	49.2
EVA-CLIP	E (5B)	256	82.0	75.7	-	79.6	68.8	51.1
SigLIP	SO (400M)	729	**83.2**	**77.2**	**87.5**	**82.9**	**70.2**	**52.0**

- 研究背景
- Fully-Supervised
- Weakly-Supervised
- Self-Supervised
- 总结

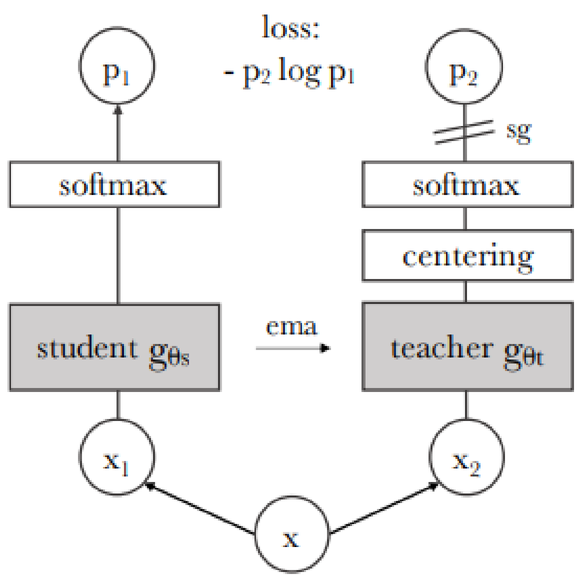
Self-Supervised

- 自监督学习，仅使用单模态进行训练，**充分挖掘模态内表征**，没有人为设定的label监督
- 自监督学习主要有mask reconstruction和self-distillation两种模式
- BEiT (ICLR2022 Oral) 设计了dVAE方法，构造了视觉词汇表，离散量化了像素，利用BERT方式实现掩码重建式预训练



Self-Supervised

- DINO (ICCV2021)将同一张图像经过两种不同数据增强，输入到学生模型和教师模型，并且对齐二者的输出，使得图像无论如何变换都可以对齐到同一表征

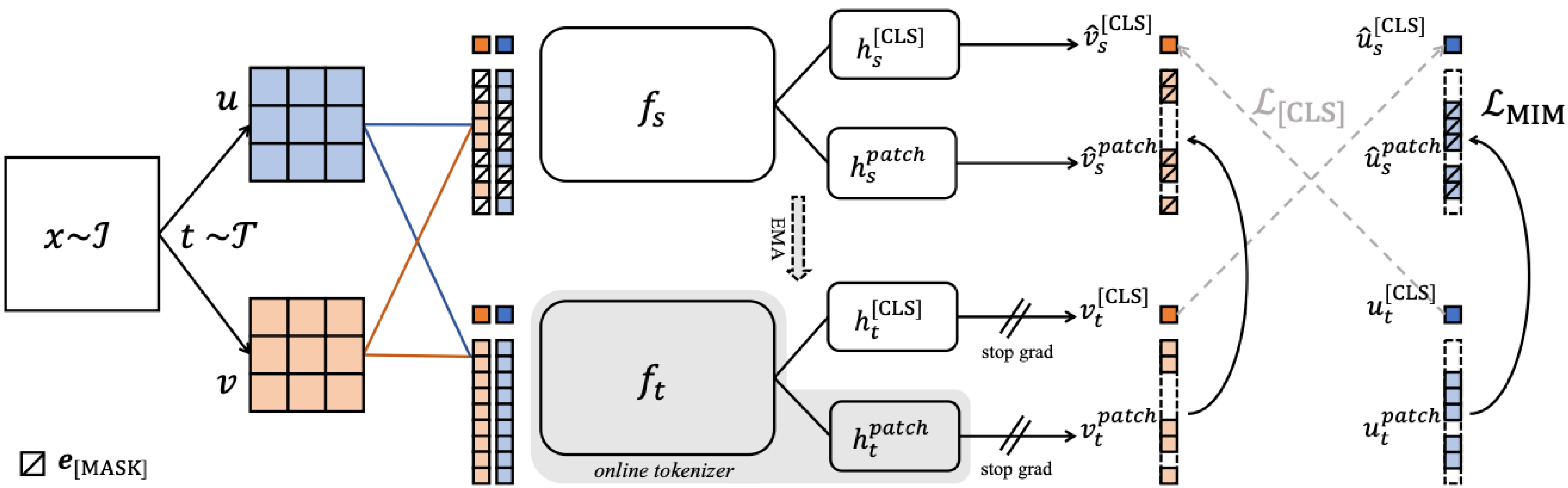


Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

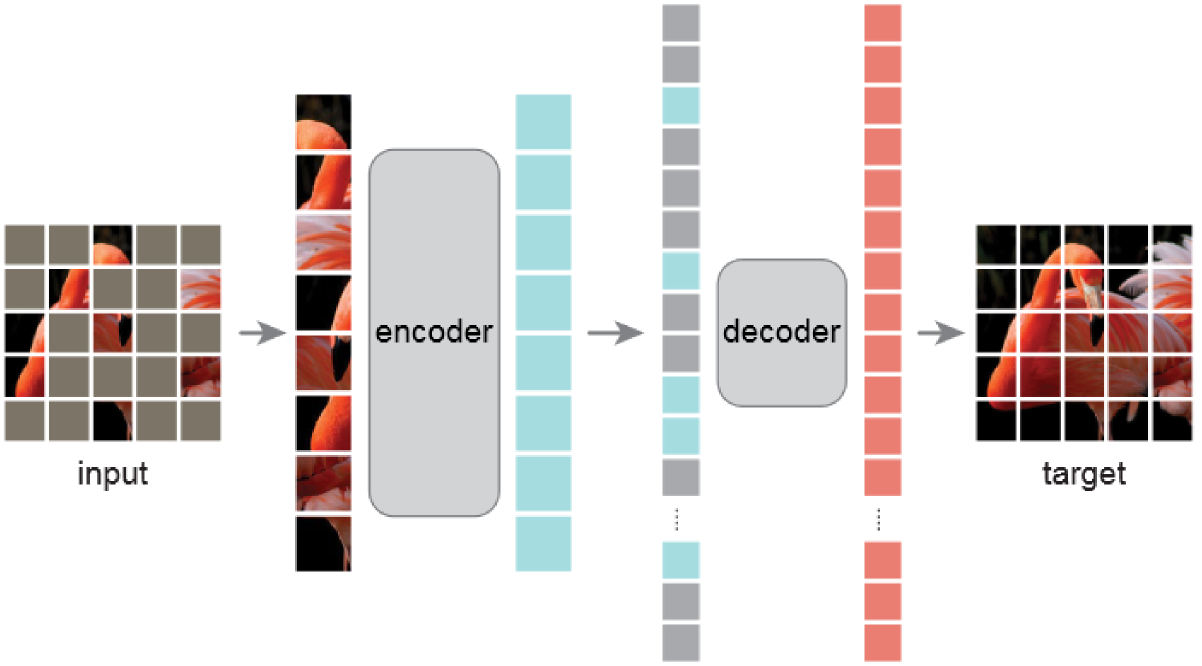
Self-Supervised

- iBOT (ICLR2022)结合了二者，利用学生和教师网络，在学生部分mask掉了一部分patch，教师部分保留，在输出部分，利用DINO相关方法交叉对齐class token，同时利用MIM损失重建masked patch



Self-Supervised

- MAE (CVPR2022) 设计了deocder来重建模型，最终训练得到encoder
- Encoder的输入仅有unmask token，降低计算量
- 不需要BEiT的视觉codebook，直接重建最原始像素信息

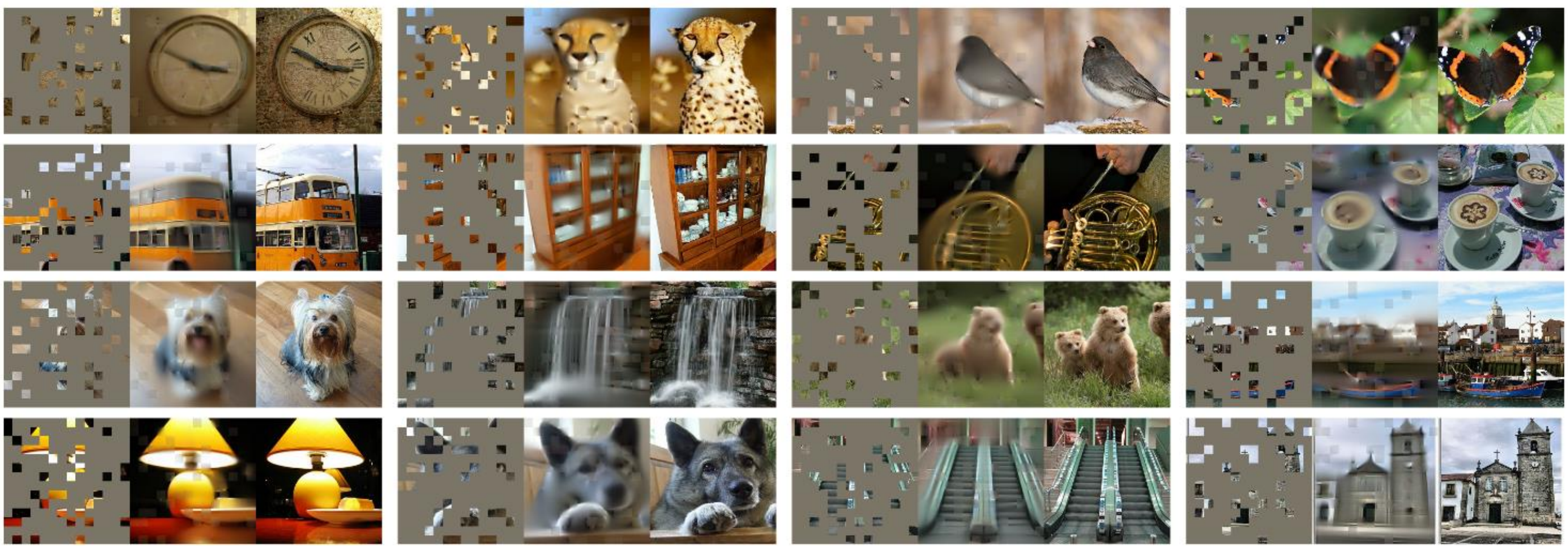


Self-Supervised

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

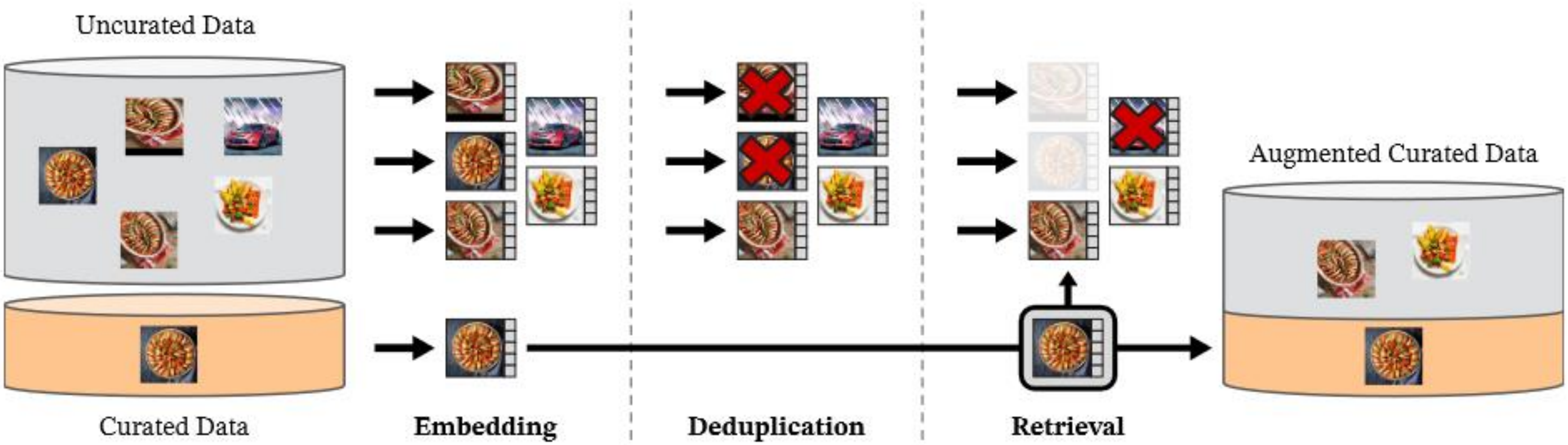
- MAE可以允许特别高的mask率
- 相比监督学习更能挖掘patch信息

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**



Self-Supervised

- DINOv2 (TMLR2024)，学习更强更通用的视觉表征，真正的视觉单模态大模型基础
- 贡献了LVD-142M数据集，利用curated数据集来检索大规模未标注数据集
- 主要采用的检索标准为Google Landmarks 2和ImageNet22k
- 自监督检索采用ImageNet22k上pretrain的ViT-H计算embedding



Self-Supervised

Task	Dataset / Split	Images	Retrieval	Retrieved	Final
classification	ImageNet-22k / –	14,197,086	as is	–	14,197,086
classification	ImageNet-22k / –	14,197,086	sample	56,788,344	56,788,344
classification	ImageNet-1k / train	1,281,167	sample	40,997,344	40,997,344
fine-grained classif.	Caltech 101 / train	3,030	cluster	2,630,000	1,000,000
fine-grained classif.	CUB-200-2011 / train	5,994	cluster	1,300,000	1,000,000
fine-grained classif.	DTD / train1	1,880	cluster	1,580,000	1,000,000
fine-grained classif.	FGVC-Aircraft / train	3,334	cluster	1,170,000	1,000,000
fine-grained classif.	Flowers-102 / train	1,020	cluster	1,060,000	1,000,000
fine-grained classif.	Food-101 / train	75,750	cluster	21,670,000	1,000,000
fine-grained classif.	Oxford-IIIT Pet / trainval	3,680	cluster	2,750,000	1,000,000
fine-grained classif.	Stanford Cars / train	8,144	cluster	7,220,000	1,000,000
fine-grained classif.	SUN397 / train1	19,850	cluster	18,950,000	1,000,000
fine-grained classif.	Pascal VOC 2007 / train	2,501	cluster	1,010,000	1,000,000
segmentation	ADE20K / train	20,210	cluster	20,720,000	1,000,000
segmentation	Cityscapes / train	2,975	cluster	1,390,000	1,000,000
segmentation	Pascal VOC 2012 (seg.) / trainaug	1,464	cluster	10,140,000	1,000,000
depth estimation	Mapillary SLS / train	1,434,262	as is	–	1,434,262
depth estimation	KITTI / train (Eigen)	23,158	cluster	3,700,000	1,000,000
depth estimation	NYU Depth V2 / train	24,231	cluster	10,850,000	1,000,000
depth estimation	SUN RGB-D / train	4,829	cluster	4,870,000	1,000,000
retrieval	Google Landmarks v2 / train (clean)	1,580,470	as is	–	1,580,470
retrieval	Google Landmarks v2 / train (clean)	1,580,470	sample	6,321,880	6,321,880
retrieval	AmsterTime / new	1,231	cluster	960,000	960,000
retrieval	AmsterTime / old	1,231	cluster	830,000	830,000
retrieval	Met / train	397,121	cluster	62,860,000	1,000,000
retrieval	Revisiting Oxford / base	4,993	cluster	3,680,000	1,000,000
retrieval	Revisiting Paris / base	6,322	cluster	3,660,000	1,000,000
					142,109,386

Self-Supervised

- 训练目标包含class-level的DINO方法以及patch level的iBOT方法
- 共享了DINO和iBOT的MLP head （自监督学习模型中在ViT或ResNet后加MLP head，然后再到模型输出，效果更好）
- 其他trick：使用SwAV的中心化方法，KoLeo正则化方法
- 在预训练的最后阶段分辨率从224x224变成518x518
- 其他的技术细节包括使用Flash Attention以及多种数据并行策略，大版本蒸馏小版本等

$$\mathcal{L}_{DINO} = -\sum p_t \log p_s \qquad \mathcal{L}_{iBOT} = -\sum_i p_{ti} \log p_{si}$$

Self-Supervised

Training Data	INet-1k	Im-A	ADE-20k	Oxford-M	iNat2018	iNat2021	Places205
INet-22k	**85.9**	73.5	46.6	62.5	81.1	85.6	67.0
INet-22k \ INet-1k	85.3	70.3	46.2	58.7	80.1	85.1	66.5
Uncurated data	83.3	59.4	48.5	54.3	68.0	76.4	67.2
LVD-142M	85.8	**73.9**	**47.7**	**64.6**	**82.3**	**86.4**	**67.6**

Table 2: **Ablation of the source of pretraining data.** We compare the INet-22k dataset that was used in iBOT to our dataset, LVD-142M. Each model is trained for the same number of iterations, that is smaller than in our final run, without high-resolution adaptation. Pretraining on LVD-142M maintains the performance over INet-1k while leading to models that perform better in other domains.

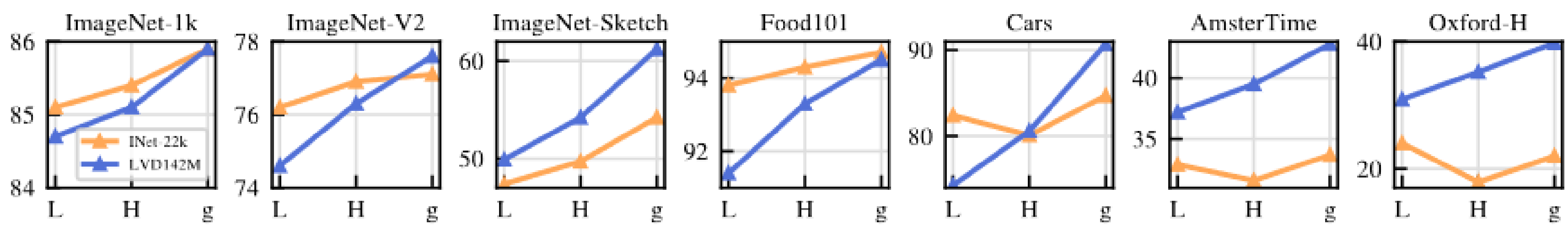


Figure 4: **Model scale versus data scale.** Evolution of performance as a function of model size for two different pretraining datasets: ImageNet-22k (14M images) and LVD-142M (142M images). The ViT-g trained on LVD-142M surpasses the ViT-g trained on ImageNet-22k on most benchmarks.

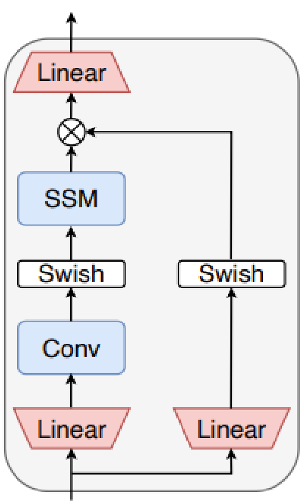
Self-Supervised

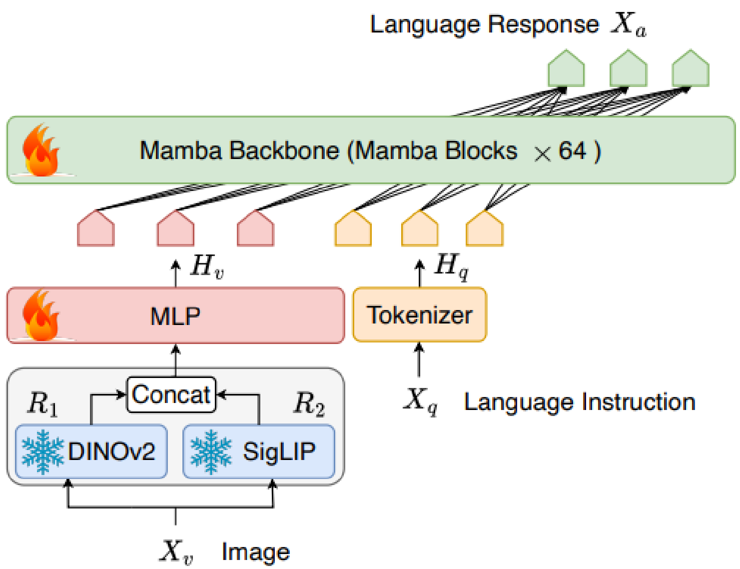
Method	Arch.	Data	Text sup.	kNN val	linear val	linear ReaL	linear V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14_{336}	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION-2B	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION-2B	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	✗	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	✗	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	✗	–	79.8	–	–
MSN	ViT-L/7	INet-1k	✗	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	✗	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	✗	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	✗	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	✗	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	✗	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	✗	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	✗	**83.5**	**86.5**	**89.6**	**78.4**

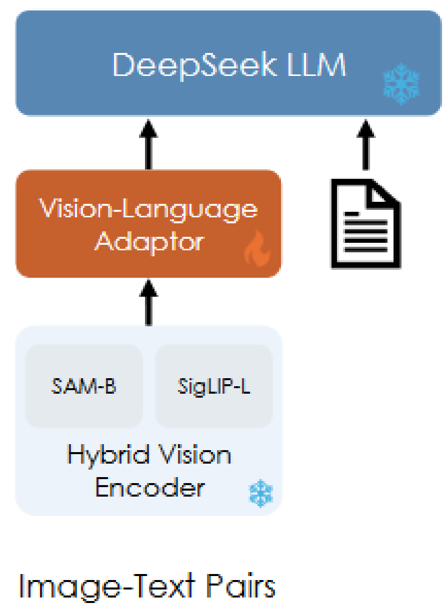
- 研究背景
- Fully-Supervised
- Weakly-Supervised
- Self-Supervised
- 总结

总结反思

- 视觉预训练模型可以作为视觉编码器，成为MLLM的视觉端
- 已有多篇文章评估了各个预训练模型在构建MLLM时的效果
- 单纯使用各种预训练模型本身来构建MLLM已经不再新鲜







谢谢！