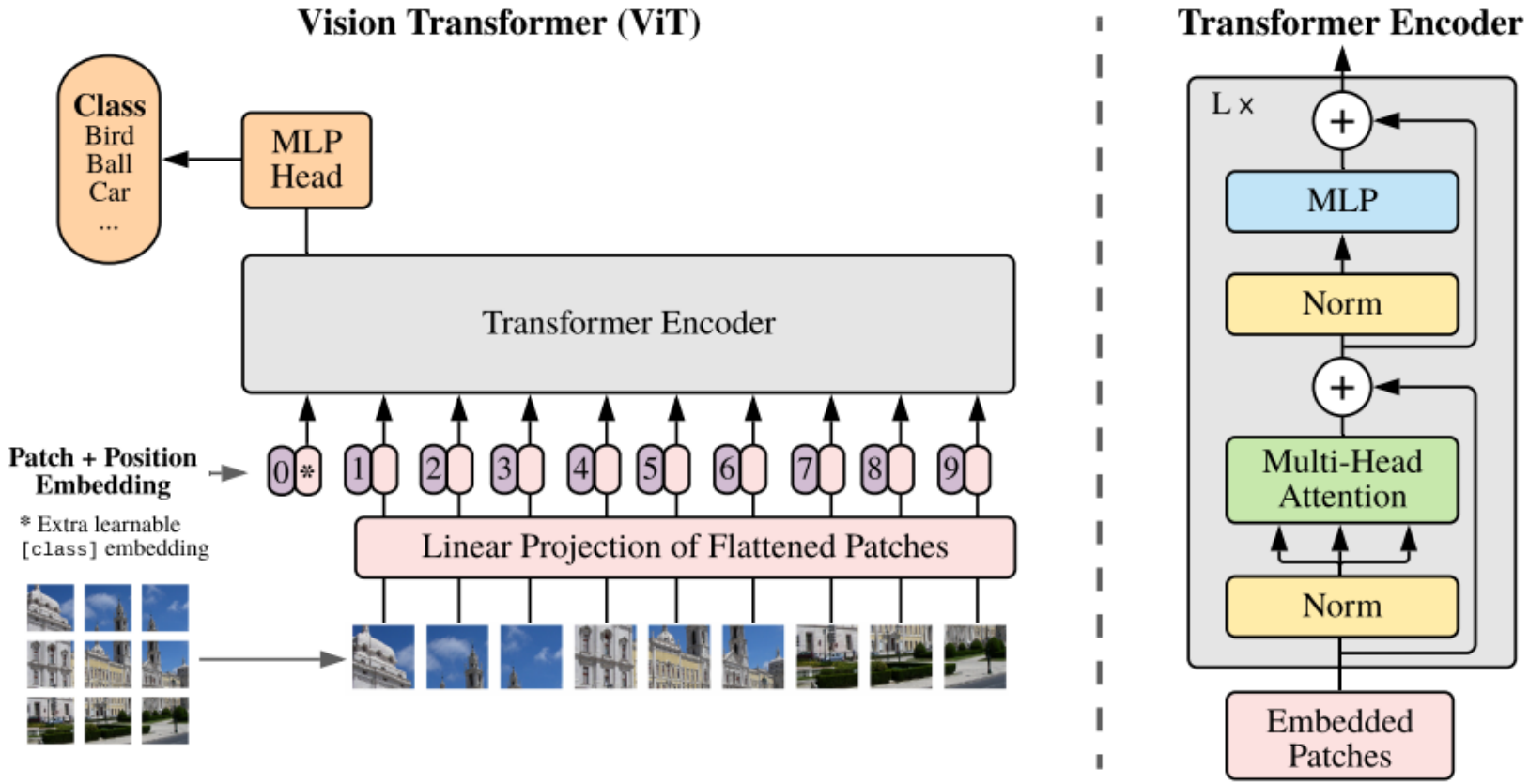
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

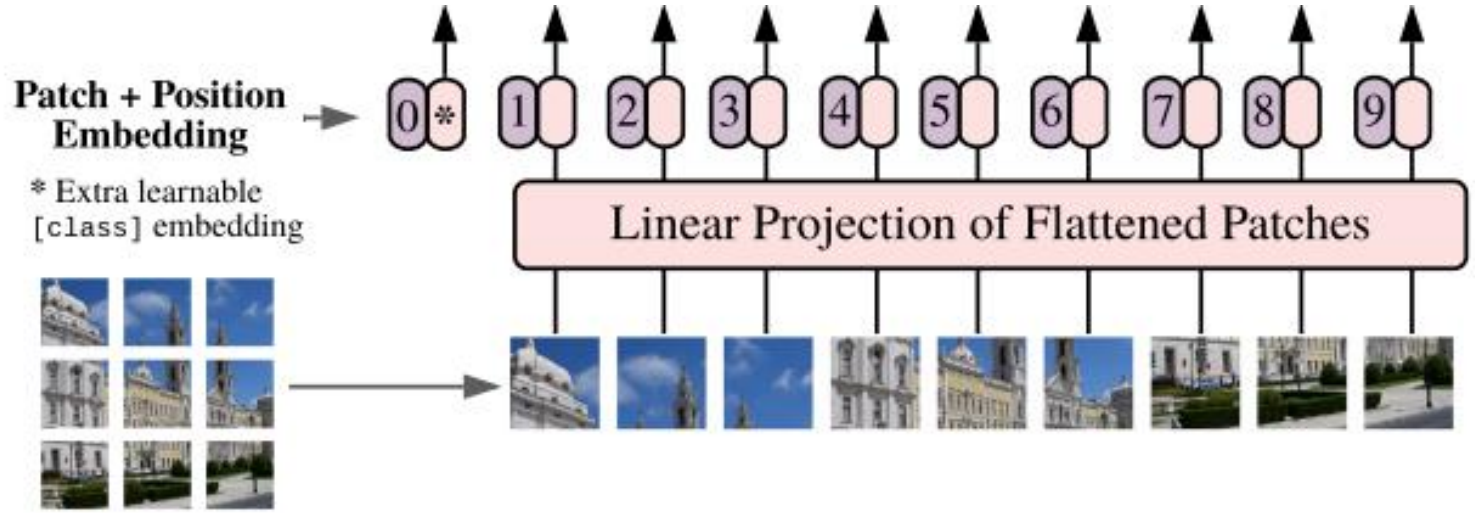
2025.11.08

➢ 模型架构概览



- 遵循原始Transformer架构
- Embedding层
- Transformer Encoder
- MLP Head

➢ Embedding层



➢ 可训练的用于分类的class token

- Cat([1, 768], [196, 768]) --> [197, 768]

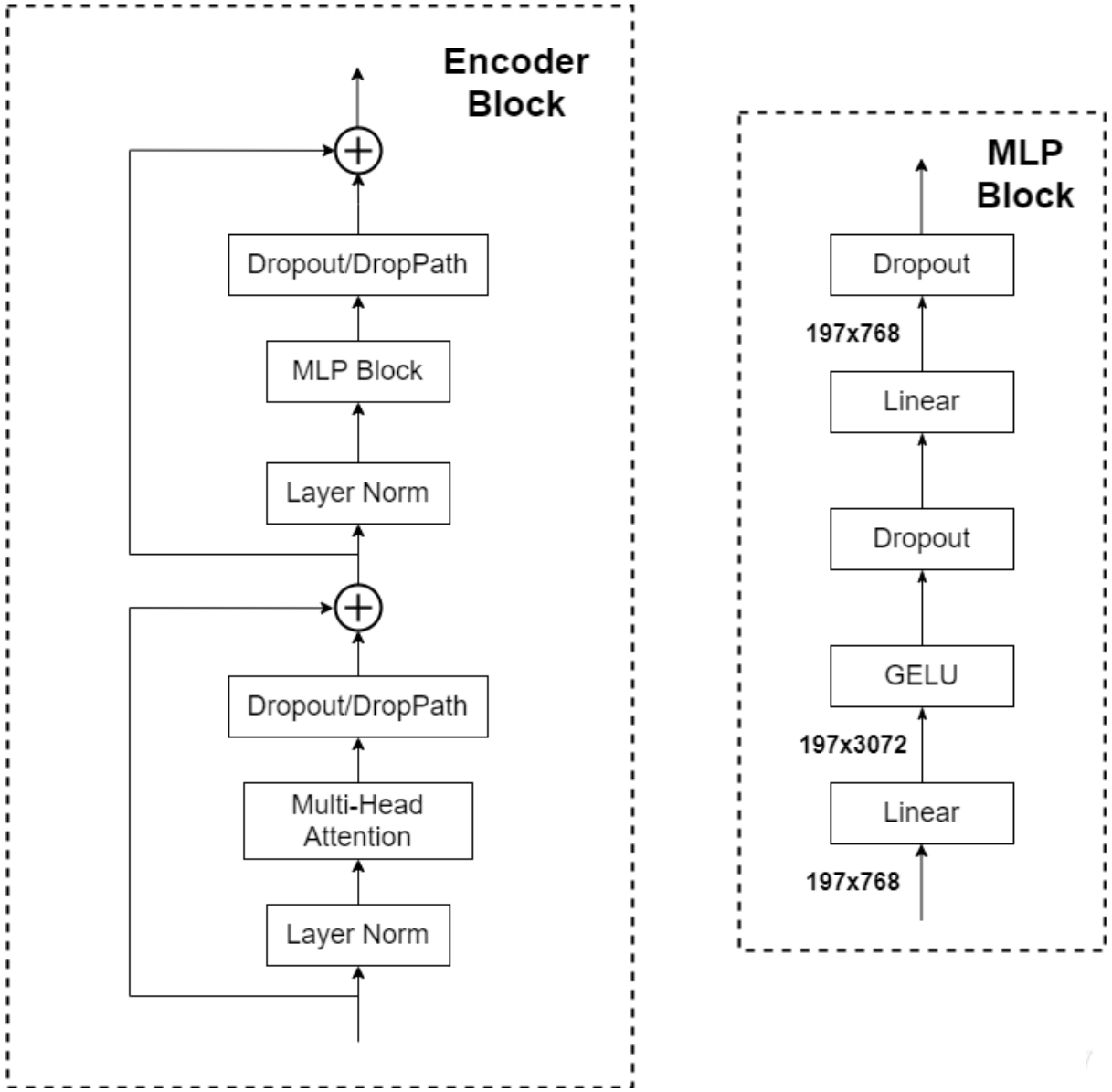
➢ 可训练的Position Embedding

- 同shape直接叠加[197, 768]+[197, 768]=[197, 768]

➢ 图像分割

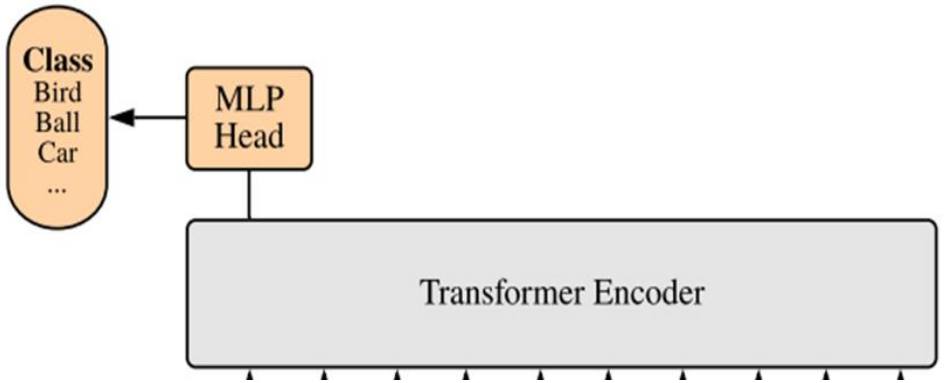
- 224*224-->196个16*16的patches
- 每个patch的shape为[16,16,3]-->16*16*3=768长度的向量（token）
- 实现方式：卷积（卷积核大小16*16，步距16，个数768）将[224,224,3]-->[14,14,768]，再展平得到[196,768]，以二维矩阵形式输入Transformer

➢ Transformer Encode

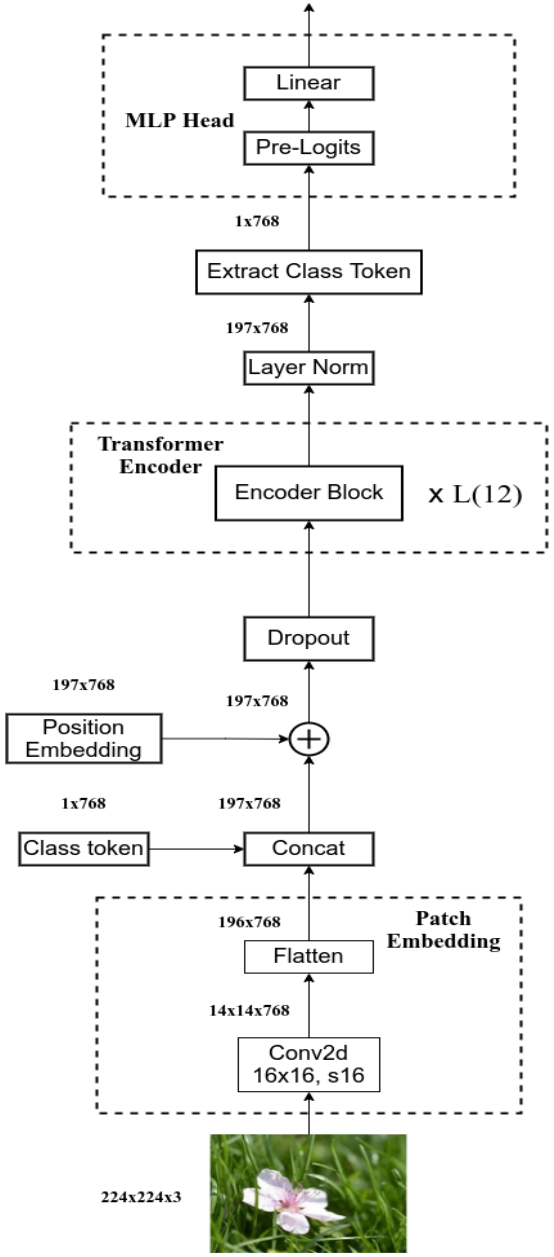


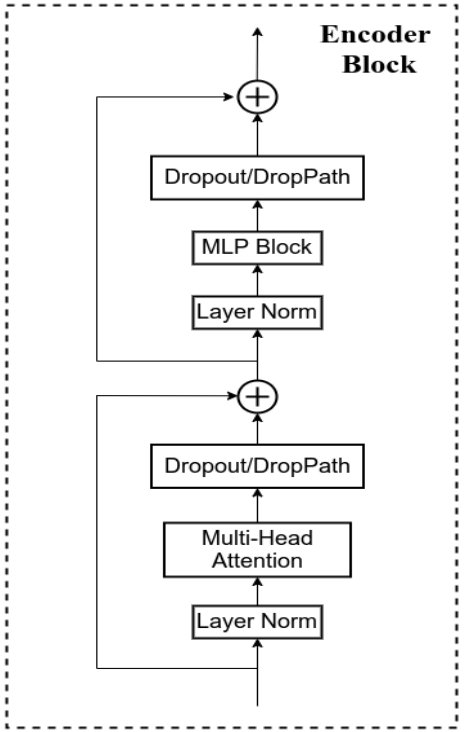
➢ MLP Head

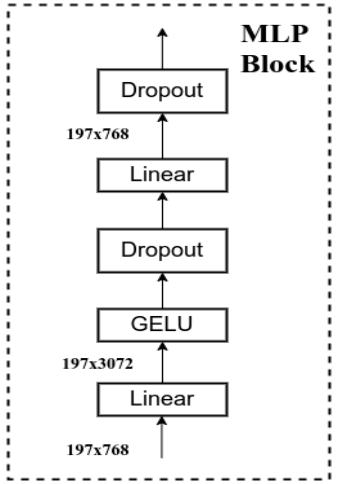
从Transformer Encode 输出的[197, 768]中抽取出class token对应的[1, 768]，输入MLP Head得到最终分类结果



➢ 模型架构详解（ViT-B/16）







➢ 模型参数设置

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

Table 1: Details of Vision Transformer model variants.

- Layers是Transformer Encoder中重复堆叠Encoder Block的次数
- Hidden Size是对应通过Embedding层后每个token的长度
- MLP size是Transformer Encoder中MLP Block第一个全连接的节点个数
- Heads代表Transformer中Multi-Head Attention的heads数。

➢ 实验对比

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

➢ 实验对比

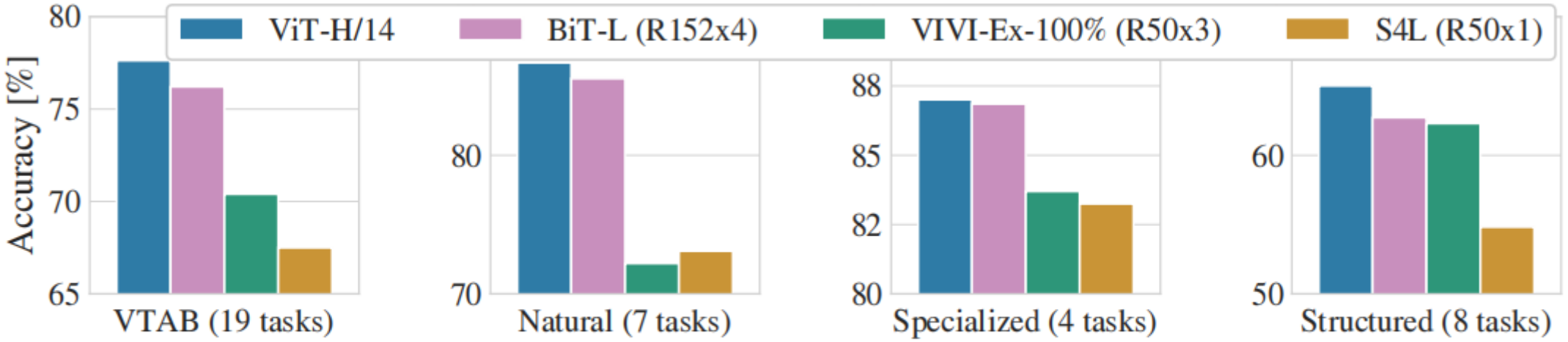


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.

➢ 预训练数据需求

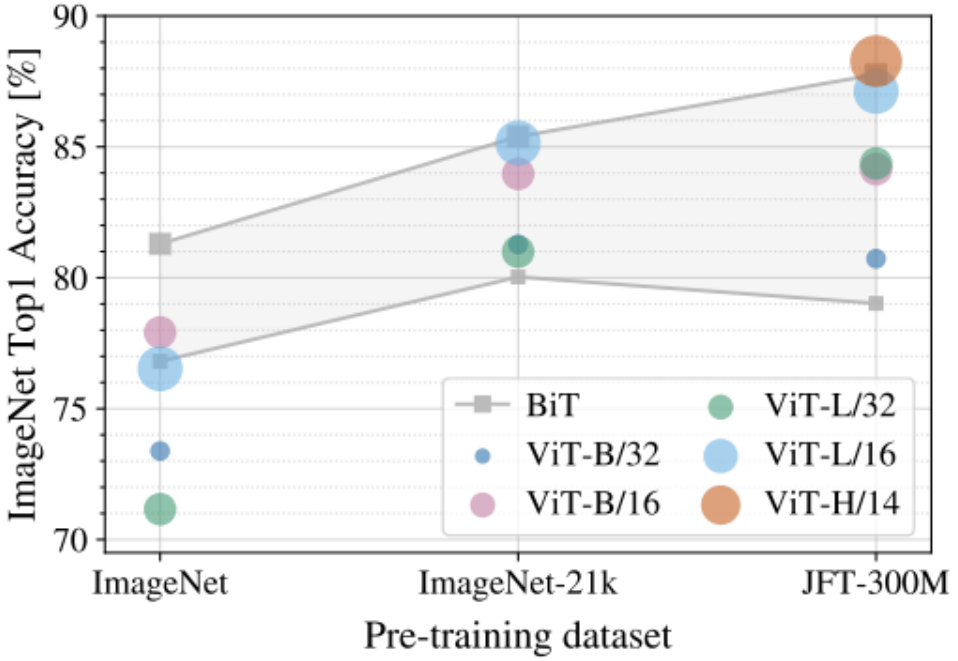


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

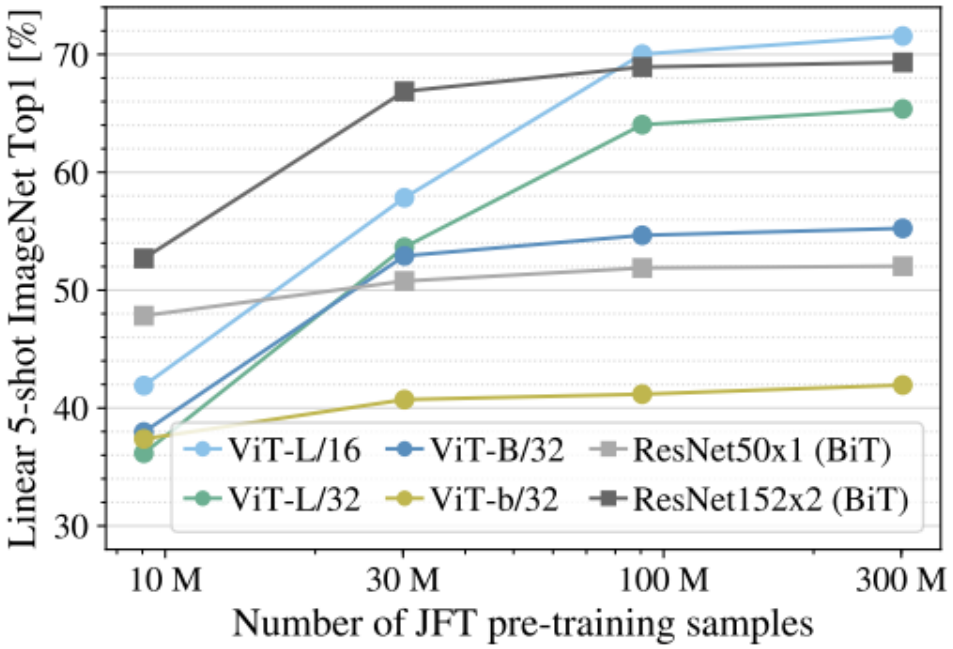


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

➢ 预训练成本分析

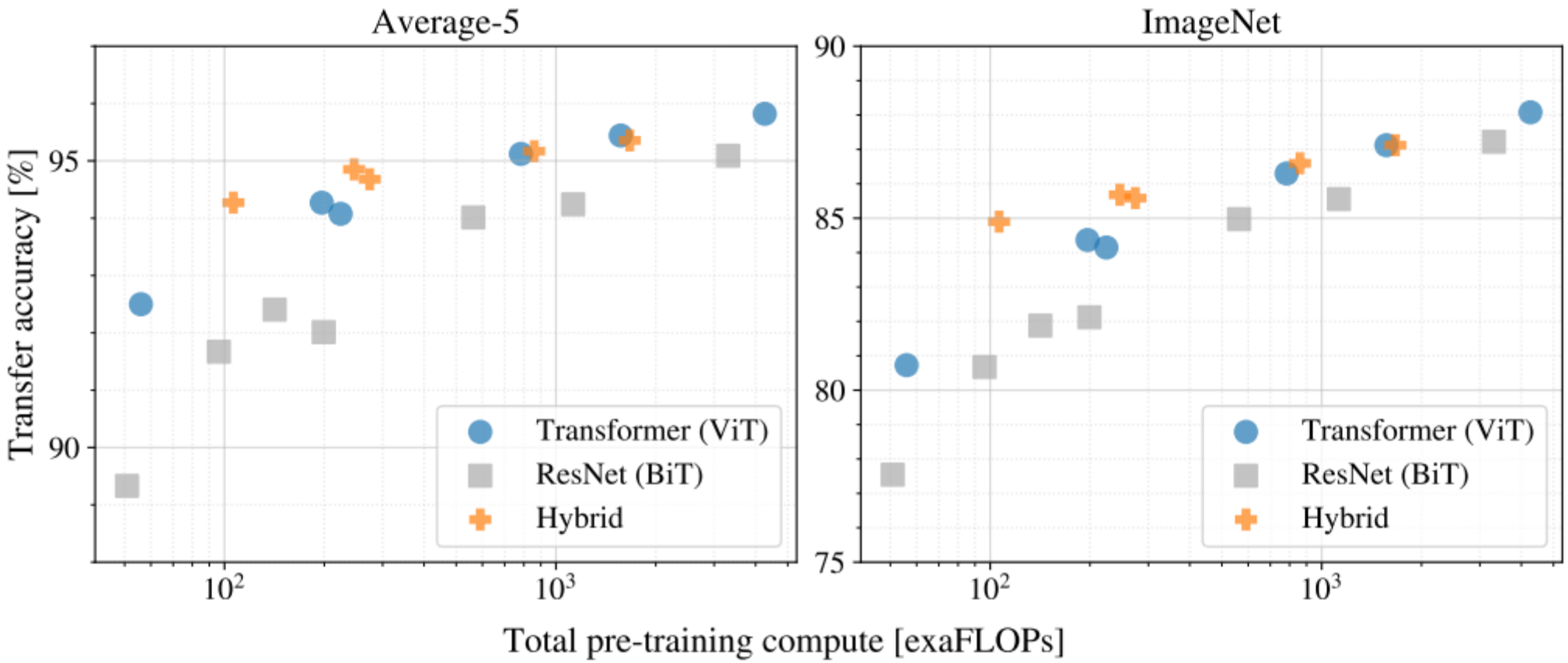


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.