ViT：Vision Transformer

An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale #论文

> **Note**
>
> - 卷积神经网络不适用：
> - [ ] 遮挡
> - [ ] 分布偏移
> - [ ] 对抗性patch添加
> - [ ] 图片打散排列组合
> - ViT 打破cv与nlp的模型上的壁垒。
> - nlp ： Bert GTP3 T5 模型
> - cv ： 对卷积神经网络的依赖不必要
> - > sota模型计算资源开销大。（2500天 TPUv3）

图片看作16x16个patch，每个patch对应一个单词。

主流方式：

- 大规模数据集上与训练，特定领域小数据集上微调：Bert（512序列长度）

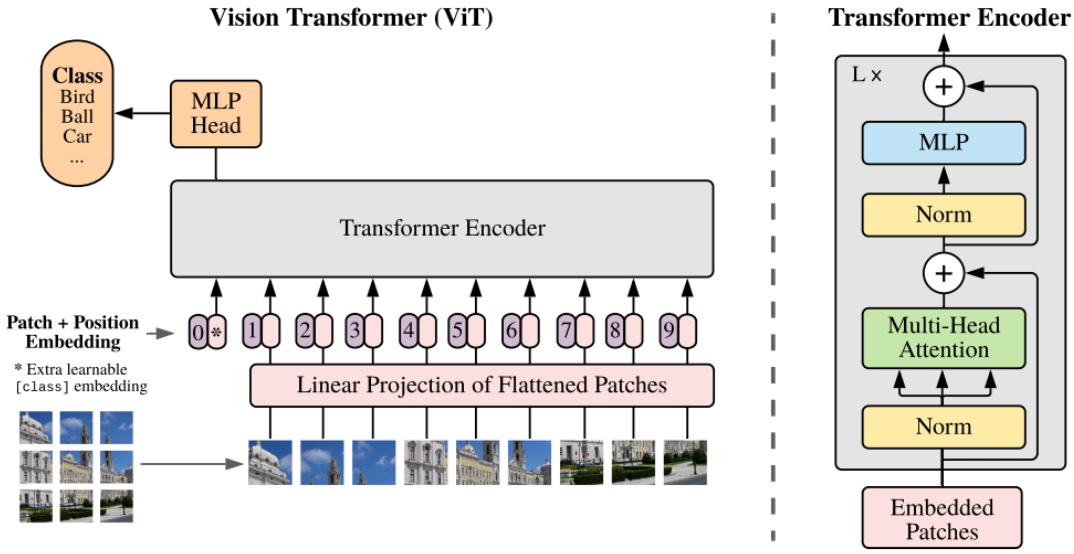
有趣的现象，增加数据没有出现性能饱和的现象。

自注意力模型，序列元素两两相互作用，求解权重计算图。

- 核心问题：
  - 2d图片变成1d
  - 用patch办法
  - 用local neighborhoods （局部注意力，相当于全局注意力的一个近似）
  - 自注意力用到不同大小的block，极端情况直接用轴注意力。
  - [ ] 问题：需要设计复杂的工程模型，还有模型参数和训练规则。

- 自注意力应用到CV中：
  - [ ] 用特征图作为自注意力的输入
  - 孤立自注意力（用局部小窗口降低复杂度）
  - 轴自注意力，分别对高度和宽度两个dim的方向做自注意力机制模型。
- ViT: vision transformer
  - 直接用transformer架构。（采用图片分patch，16×16）
  - 图片块相当于nlp里面的每一个的单词。
  - 有监督。
  - 可以扩展至 224x224
    - 大量数据加持
    - 充分预训练
    - 缺少CNN某些特定的归纳偏置（即一些先验的知识）
      - 例如提前做好的假设
      - inductive biases：
        - locality 相邻区域有相邻的特征，靠得越近，相关性越强
        - translation equivariance （平移不变性）写成公式
          - $f(g(x) = g(f(x)))$
          - f:卷积， g:平移
    - 在大规模的数据的预训练之后，就会别CNN的先验偏置效果更好。在下游任务获得很好的迁移学习效果。
- Transformer 怎么学习相同的偏置信息？
  - 迁移学习
- image GPT 同样运用了 Transformer
- 生成式网络一般比判别式网络差。这也是MAE备受关注的原因。

内容部分

- 

> We split an image into fixed-size patches, linearly embed each of them,
> add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable “classification token” to the sequence.

- Patch Embedding
  - patches 序列经过线性投射层得到的特征
  - 自注意力特点：元素之间两两存在交互，不存在一个顺序的问题。但对于图片来说这是一个整体。
  - #想法
    - Transformer 模型用于提取全局特征的本质可以将图片块配合位置信息，作为一个nlp模型处理cv问题，实现跨领域结合。
  - 在patch embedding 加上了一个position embedding 加上位置编码信息，整体图像的token包括了原本有的图像信息，包含了图像块所在位置。（图片中的0-9代表位置信息position）
  - cls 分类字符 *
    - cls token （position 0）只有一个token，同样维度D。（占用1个token）
    - 借鉴BERT class token 与图像的特征有一样的维度，将他的特征作为整体的特征，作为全局的
  - MLP HEAD 通用接头。
  - 训练使用交叉熵。

    - linearly embed:实际上是一个全连接层（E），他的维度DxD，其中D为前面patch算出来的。例如这里：D = 16×16×3=768，
    - 位置信息，sum（直接加）
- 多头自注意力 K Q V
- norm
- tanh非线性激活
- MLP 一般会放大四倍维度（768变3012）

The MLP contains two layers with a GELU non-linearity.

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

- 消融实验：
    - global average pooling处理 （GAP）可以代替cls
        - 注意两个learning rate 不一样。

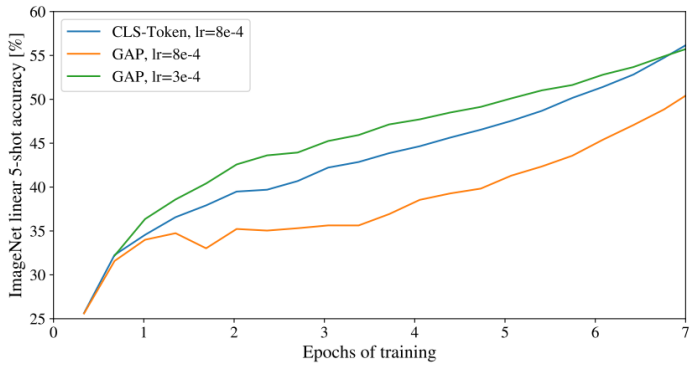


Figure 9: Comparison of class-token and global average pooling classifiers. Both work similarly well, but require different learning-rates.

    - 位置编码：（这几个编码的等价）
        - 从1D
        - 到2D 例如11,12,22,...
        - 相对位置编码 用两个编码的相对位置信息

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

-
- 三种位置编码performance都是64，解释排列组合比较少，编码不影响位置理解。

实验

模型变体：

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.

模型除了本身和Transformer有关，同时也要考虑模型输入patch-size对于位置变量的影响。每个图像块越小，那么序列长度越高，位置position越多。

	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).

- Noisy Student（TPUv3 10000天）
  - Imagenet表现最佳的方法 伪标签（pseudo label）进行 self-training
- ViT-H/14 训练比较贵。
- 训练数据需求：

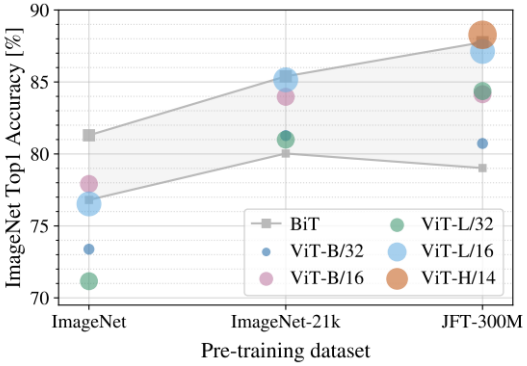


Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

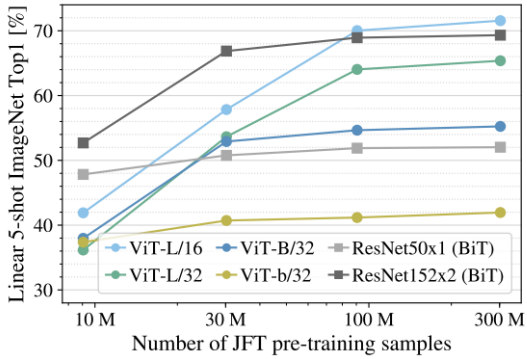


Figure 4: Linear few-shot evaluation on ImageNet versus pre-training size. ResNets perform better with smaller pre-training datasets but plateau sooner than ViT, which performs better with larger pre-training. ViT-b is ViT-B with all hidden dimensions halved.

- Figure 3: 最重要，不同大小的数据集，灰色区域 ResNet效果。
  - 小数据集ViT全面低于ResNet
  - 中数据集差不多
  - 大数据集上的ViT要比ResNet更好 （ResNet152*2和ResNet50*1），拓展性更好一些。
  - 预训练，比一般卷积神经网络要便宜。
  - [ ] 训练的 Tricks，提升性能以可以和ResNet比肩：
    - drop out
    - weight decay
    - label smoothing
- Figure 4:
  - 通过类似的 少样本（这里是每一类采取了5个样本）
    - 作者也**采取了这种方式做了大量的消融实验。**
  - 致力于分析ViT的本身的特性
    - 采用linear few-shot evaluation
  - 实验方法：
    - 拿到预训练模型之后，直接当做一个特征提取器。
    - 不做fine tune
    - 直接得出的特征做一些logistic regression
    - 采用统一数据集的子集（JFT）这样模型的数据集之间不存在很大的gap（不同数量样本10M,30M,100M,300M）
      - 更能体现模型的本质
    - 特点，小样本训练上 ViT没有上述的Tricks引入，容易过拟合，导致训练出来的样本没办法拓展到其他任务中去。 （缺少归纳偏置和其他约束方法）

- **预训练**数据集的增大，ViT的**稳健性**上升。

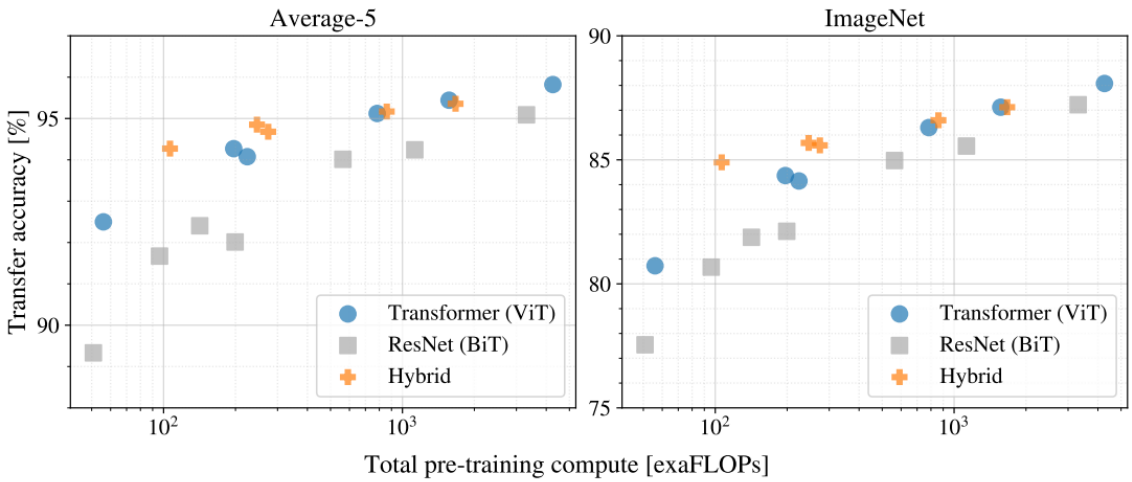


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger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

- 证明ViT经济实惠的实验
  - Hybrid ： 混合模型CNN+transformer
  - 大大小小的点是相同颜色类型的模型的变体（例如ResNet50*1）
  - #想法 是否可以利用下特点
    - 混合模型在小样本数据集上，不需要很多样本预训练，同时可以达到ViT同样的效果
    - 随着模型增大，混合模型和ViT趋同，但是甚至有稍微低于ViT

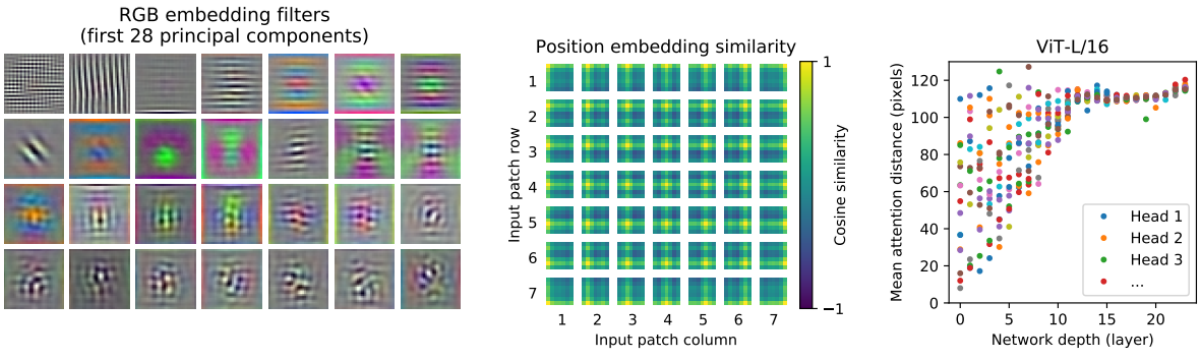


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

> 网络学到了什么
> - ViT

> - First layer
> - Linear projection (E)
> - 类似gobar filter
> - input patch column
> - 从1D的位置编码学到了2D的位置编码
> - 自注意力操作
> - 模拟长距离的关系 （ViT是否有效？）
> - ViT-Large （24 layers）
> - head 多头注意力中的头
> - mean attention distance = $d(\hat{a}, \hat{b}) \times attention$

- [ ] mask patch prediciton
  -

结论

- 抽图像块、位置编码：用了图像特有的归纳偏置，其他地方没有引入了。
- 简单、扩展性很好，与大规模与训练结合。不需要很多领域外了解。
- （训练起来相对便宜）
- 已在分类表现良好。
- 应用到分割和检测怎么样。
- 检测 : ViT FR-CNN， 分割:SETR
- Swin Transformer: 多尺度融合进Transformer。（更适合做视觉）

- [ ] 探索自监督的学习方式
  - NLP大网络都是靠自监督训练方式，能否迁移应用？Scaling Vision Transformer
  - ViT-G

- [ ] 多模态工作能否用Transformer来做

- 大规模语料做与训练，具体目标任务fine tune
- GPT language modeling

- BERT 挖词填空式

展望

- [ ] 如何用Transformer去做小样本的学习，是一个相当有前途的方向。