《人工智能中的编程》大作业报告

——细粒度图像分类

1 模型

我们使用 [2] 所提出的基于 Vision Transformer[1] 的 TransFG 模型。

1.1 Vision Transformer

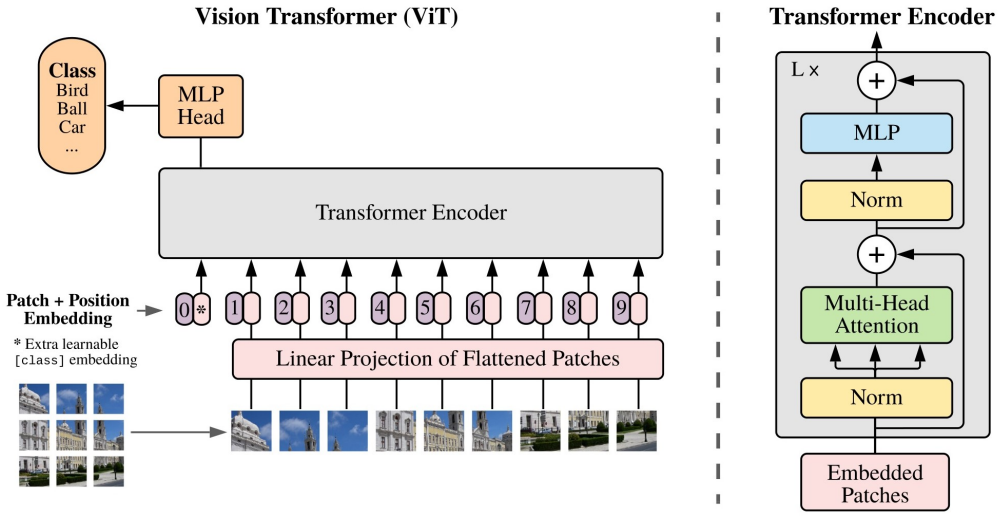

图 1: ViT 模型框架 (from [1])。

1.1.1 Image Sequentialization

对于高、宽、通道数分别为 H、W、C 的图像 $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$，将其分割为 $P\times P$ 的 patches，经 flatten 得到 $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$，其中 $N = HW/P^2$。

在 ViT 中，各 patch 之间是互不相交的，这一点在 TransFG 得到改进。记 sliding window (of size (P, P)) 滑动的步长为 S，则相邻两个 patch 的重叠区域大小为 $(P - S) \times P$，相应的，patch 的个数为

$$N = \lfloor \frac{H - P + S}{S} \rfloor \times \lfloor \frac{W - P + S}{S} \rfloor \tag{1}$$

增加重叠区域，也即减小 S，可以保留图像局部信息，从而得到更好的训练效果，但同时也会增加计算花销。因此在实际训练过程中，需要调整合适的 S 以取得计算性能与效果的平衡。

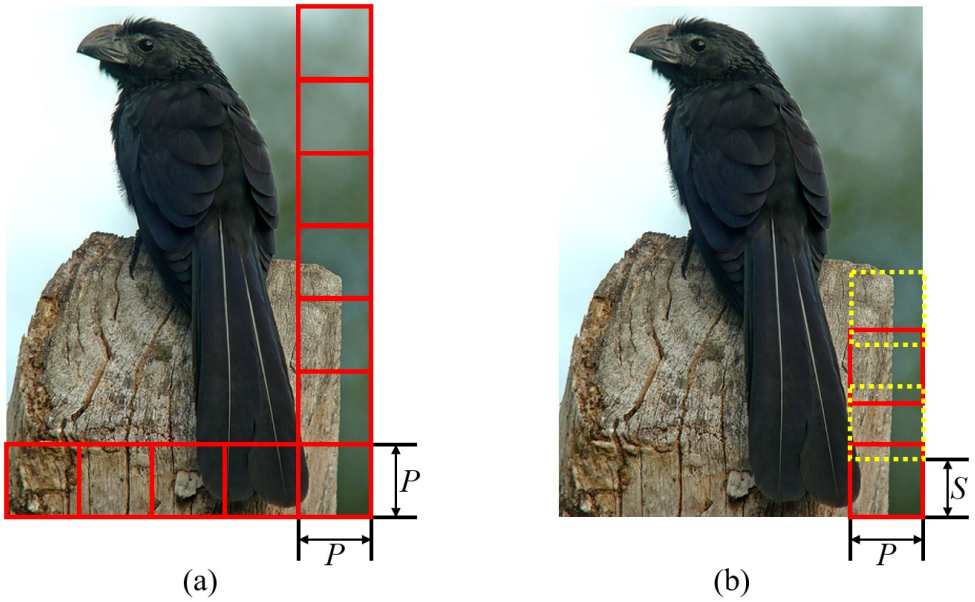

图 2: patch 的分割方法。(a) ViT; (b) TransFG。

1.1.2 Patch Embedding

将序列化后的图像 $\mathbf{x}_p$ 的每个 patch 映射成一个 D 维向量

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

其中 $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$，$\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$，均为可训练的参数，$\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$ 是可训练的 token。

1.1.3 Transformer Encoder

Transformer Encoder[3] 由 L 层 block 组成，每个 block 包含一个 multi-head self-attention (MSA) 层和一个 multi-layer perceptron (MLP) 层，通过每一层前进行一次 LayerNorm (LN)。(结构见 fig. 1右侧)

eq. (2) 中得到的 $\mathbf{z}_0$ 将作为 Transformer Encoder 的输入，则在第 l 层所得的结果可记为

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, l = 1, 2, \cdots, L \tag{3}$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, l = 1, 2, \cdots, L \tag{4}$$

在 ViT 中，直接将 $\mathbf{z}_L^0$ 作为特征传入分类器。TransFG 将在这方面做出改进。

1.2 TransFG

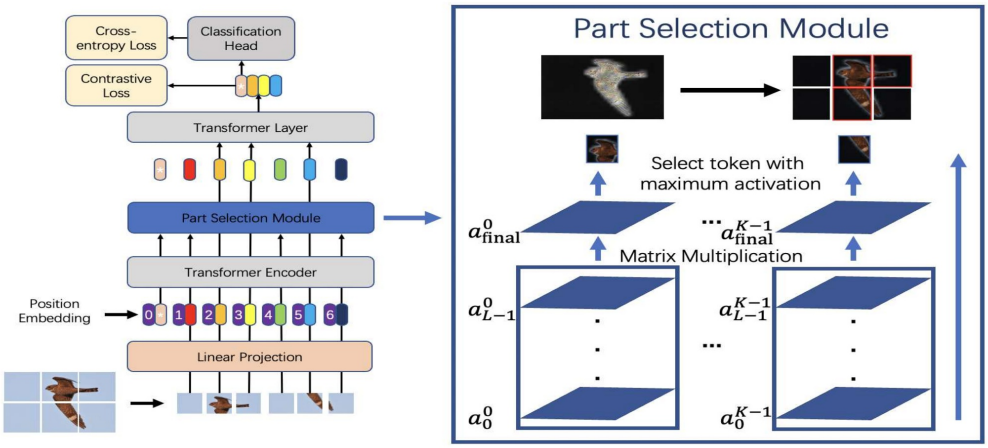

图 3: TransFG 模型框架 (from [2])。

TransFG 在 ViT 的架构上主要做出了两个改进。

1.2.1 Part Selection Module

在通过最后一层 Transformer Layer 前，先利用 Part Selection Module 选取具有代表性的 patches。

若模型具有 K 个 self-attention head，则各层的注意力权重 $\mathbf{a}_l$ 具有如下形式:

$$\mathbf{a}_l = [a_l^0; a_l^1; a_l^2; \cdots; a_l^K], l = 1, 2, \cdots, L-1 \tag{5}$$

$$a_l^i = [(a_l^i)_0; (a_l^i)_1; (a_l^i)_2; \cdots; (a_l^i)_K], i = 1, 2, \cdots, K \tag{6}$$

将所有层的原始注意力权重作矩阵乘法得到

$$\mathbf{a}_{\text{final}} = \prod_{l=0}^{L-1} \mathbf{a}_l \tag{7}$$

由于 $\mathbf{a}_{\text{final}}$ 蕴含了从输入层到最后一层的所有注意力相关信息，故利用 $\mathbf{a}_{\text{final}}$ 选择具有代表性的 patches 是比直接用 a_{L-1} 更佳的选择。

设 $\mathbf{a}_{\text{final}}$ 中 K 个 attention-head 所对应的最大值的索引为 $A_1, A_2, \cdots, A_K$，则将

$$\mathbf{z}_{\text{local}} = [z_{L-1}^0; z_{L-1}^{A_1}; z_{L-1}^{A_2}; \cdots; z_{L-1}^{A_K}] \tag{8}$$

作为最后一个 Transformer Layer 的输入。这使得最后一个 Transformer Layer 专注于不同子类别之间的细微差异，而放弃背景之间的共同特征等区分度较小的区域。

1.2.2 Contrastive Loss

由于在细粒度分类的条件下，不同种类之间的差距可能十分细微，所以交叉熵损失函数 $\mathcal{L}_{\text{cross}}$ 无法起到完美的监督学习作用。

TransFG 在交叉熵之外，引入了新的 contrastive loss

$$\begin{aligned} \mathcal{L}_{\text{con}} = \frac{1}{B^2} \sum_{i=1}^{B} [& \sum_{j:y_i=y_j}^{B} (1 - \text{sim}(z_i, z_j)) + \\ & \sum_{j:y_i \neq y_j}^{B} \max(\text{sim}(z_i, z_j) - \alpha, 0)] \end{aligned} \tag{9}$$

其中 B 是 batch size，$\text{sim}(\cdot, \cdot)$ 表示向量的点乘。

最终将 $\mathcal{L} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{con}}$ 作为模型的损失函数。

2 训练

2.1 数据集

使用 CUB_200_2011[4] 的类别标注数据。

2.2 细节

eq. (1) 中的 H, W, P, S 分别设置为 448, 448, 16, 12。eq. (9) 中的 α 设为 0.4。

使用了 ViT 的预训练模型`ViT-B_16`作为 Base Model，batch size 为 16，learning rate 初始值为 0.03，训练轮数为 10000。

使用了 SGD 优化器，momentum 设置为 0.9。

使用了余弦退火 (cosine annealing) 作为 scheduler。

2.3 结果

每训练 1000 次进行一次 validation。各次 validation 的 Loss 和 Accuracy 曲线如 fig. 4所示。最大的 Accuracy 为 90.991%。

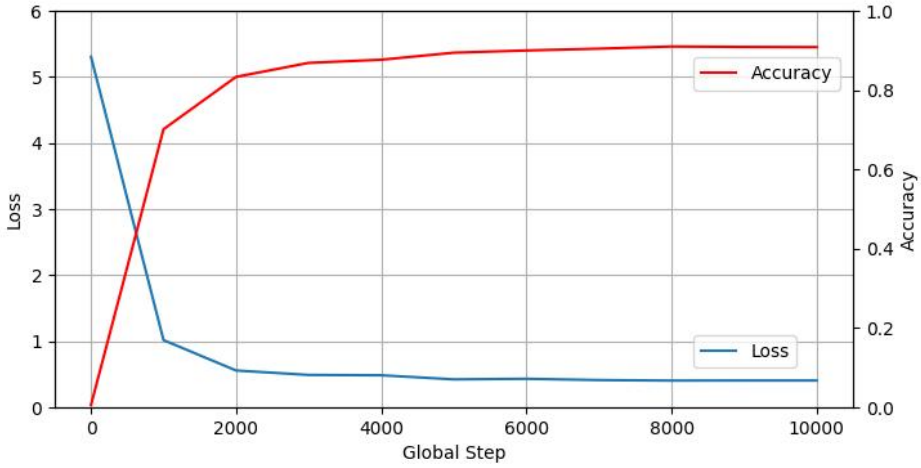

图 4: 训练结果。

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. *arXiv preprint arXiv:2103.07976*, 2021.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[4] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.