Homework 2

小组名称: convguai
小组成员: 仇嘉星 熊光智

1 任务介绍

本次实验的目标是将图像尽量精准地分为 100 个类别。使用的训练集、验证集选自 ImageNet，训练集共 100K 个样本，验证集共 10K 个样本。测试集共 10K 个样本，其中 5K 选自常规 Imageenet 的验证集，另外 5K 选自 ImageNet-A 数据集。

2 实验方法

2.1 模型结构

因为 50% 测试数据来源于 ImageNet-A 数据集，所以本次实验的重点在于 ImageNet-A 样本的预测。我们阅读了相关文献，发现自注意力机制能够显著提高模型在 ImageNet-A 上的准确率 [2]。另外，较大的模型也能提高模型的准确率。

因此，我们在现有的 ImageNet 模型中寻找满足上述条件的模型作为出发点，最终选中了 Vision Transformer(ViT) 系列的 ViT-base 在 ImageNet-21k 上预训练的模型。

ViT 没有采用计算机视觉领域过去常用的 CNN 模型，而是借鉴了如今 NLP 领域常用的 Transformer 框架 [3]，使用纯注意力的机制来实现图像识别的任务。为了适应 Transformer 的数据处理格式，保证模型性能，ViT 首先对二维的图片进行了切块，再对每个子图的信息做 embedding，将其转化为一维向量再传入 transformer 中。输入向量中还加入了 position embedding 的信息，避免序列化后的子图丢失了它们在图片中的位置信息。随后的模型结构和标准 transformer 相同，由 Multiheaded self-attention(MSA) 和 MLP 构成，并在每个模块前使用了 LayerNorm，最后引入残差连接。[1]

为了与 TinyImageNet 数据集兼容，我们调整了该模型，在前加入了图像缩放层，在其后加入了全连接层。

2.2 数据预处理

我们对数据做了常见的预处理。考虑到图片每个像素点的取值范围为 0∼255，为了让模型能训得更加精准，我们将每个图片的像素点数值除以了 255 以保证其取值在 [0,1] 范围内。此外，我们对处理后的数据进行了标准化，使得其分布更加集中，进一步提升模型的性能。除此以外，考虑到 ImageNet-A 样本会有背景颜色干扰判断等问题，我们还对训练数据进行了随机灰阶、旋转、水平翻转处理。

3 实验过程及结果

损失函数：CrossEntropyLoss

优化器：SGD with learning rate = 0.003 and momentum = 0.9

batch size: 40

我们使用以上超参训练了 10 个 Epoch，模型在验证集上的准确率为 0.9316，在 Kaggle 公开测试集上的预测结果为 0.54625。

我们也使用了 Adam、Rmsprop、Adagrad 等优化器，按模型基本收敛时的准确率排名为 SGD > Adagrad > Rmsprop > Adam。

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] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.