Use Pretrained ResNet for Picture Embedding in Transformer

林子清
2212279@mail.nankai.edu.cn

1 Dataset Construction

1.1 Base Dataset

基础的数据集包含 80 个数据点，成功与失败各 40 个，我们将其中的 75% 作为训练集，其余作为验证集。

1.2 Data Augment

我们使用了数据增强的方法将 80 个数据增广成 2560 个数据，这样可以扩大数据的规模以便进行迁移学习。输入的数据为一个八帧的视频，我们进行数据增强时对同一个视频采用相同的增广参数以保证提取特征的对应，同时也因为即使在测试数据中，八张图像的位置依然可以保持一致性。

增广的流程如下：

1. 以 50% 的概率进行 $\frac{\pi}{2}$ 的旋转
2. 以 50% 的概率进行 $\frac{3\pi}{2}$ 的旋转
3. 以 50% 的概率进行 $[-\frac{\pi}{4}, \frac{\pi}{4}]$ 的随机旋转
4. 以 50% 的概率进行 $[-0.5, 0.5]$ 的随机倾斜
5. 以 50% 的概率进行 $[-1, 1]$ 的随机扭曲
6. 以 50% 的概率进行 $[-\frac{\pi}{5}, \frac{\pi}{5}]$ 的随机剪切
7. 以 50% 的概率进行左右翻转

2 Framework

2.1 Overview

对于序列数据我们往往使用 RNN、LSTM 等模型处理，但已经有实验证明 transformer 在序列数据和分类任务上的优越性，同时其并行性也可以改善推理速度，因此我们这里采用 transformer 架构。值得说明的是，我们依然采用 CNN 作为提取 feature map 的手段而并不使用 ViT，这里只使用 transformer 作为处理序列数据的方法。

2.2 Picture Embedding

在 transformer 中存在 Word Embedding 层，这里我们设计了类似的 Picture Embedding 层，这层中我们将一张图片嵌入到 256 维的向量空间中。
正如 Word Embedding 层需要在大量本文中进行预训练一样，我们的 Picture Embedding 层也需要在大量图片中进行预训练，这里我们选择采用 ImageNet 数据集，鉴于图片特征的提取相较于文本更难，并且往往需要使用卷积层，我们选择 ResNet-18 作为嵌入层的架构。

2.3 Positional Encoding

在一般情况下，对于图片，其位置信息并不像文本那样重要，因为文本的相同词在不同位置有不同含义，但图片的内容中天然地包含了时序信息，但在这里我们为了保证模型的鲁棒性，我们依然采用了 RoPE 位置编码。

2.4 Classify Layer

最终 transformer 输出结果为 8 个向量，其中每个向量都包含了全局的信息，因此我们只需要利用其中一个，并在其上架设分类器就可以进行分类。但我们为了保证分类器能够充分的利用全局的信息，我们将 8 个向量进行连接，将连接后的向量输入进分类器中。分类器架构：
1. 256 输出全连接层
2. ReLU 层
3. 20% 概率丢弃的 Dropout 层
4. 2 输出的全连接层
5. LogSoftmax 层

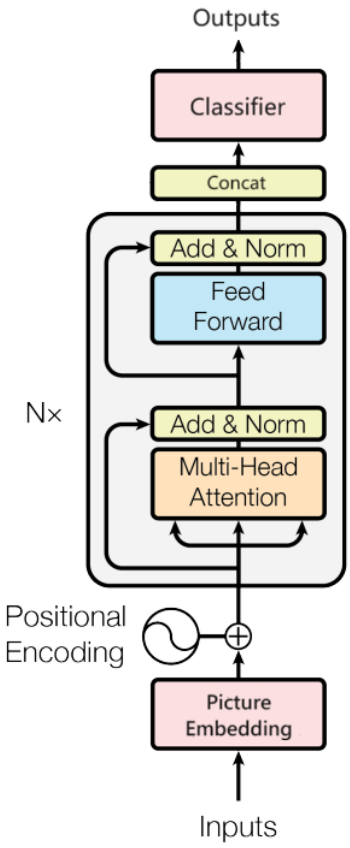

图 1: 网络结构

3 Training

这里我们采用 $NNLoss$ 作为损失函数：

$$\text{NLLLoss} = -\frac{1}{N}\sum_{i=1}^{N}\log(\text{output}[i][\text{target}[i]]) \tag{1}$$

选取 Encoder 层数为 2，注意力头数为 8，注意力维度为 64，前向传播维度为 256，同时采用 AdamW 优化器和 $1e-5$ 的学习率，以 32 为 `batch_size`，共训练 100 个 epoch。在训练过程中我们使用了单张 RTX4090 显卡，共训练 8h，在验证集上得到了超过 99% 的准确率。

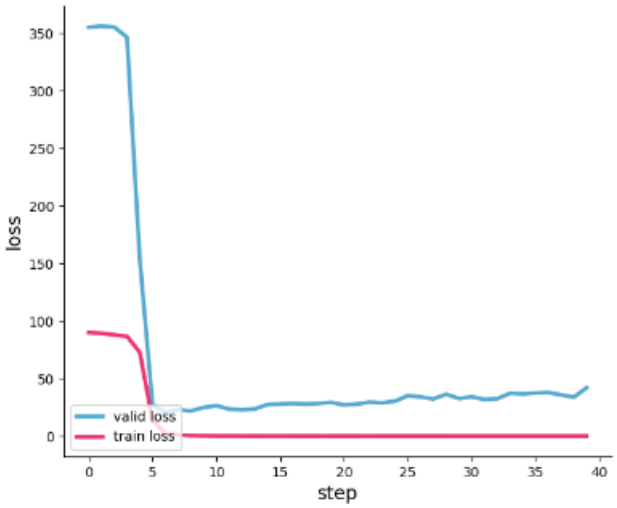

图 2: loss(前 40epoch)

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). Cambridge, MA: IEEE Press.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). Long Beach, CA: Curran Associates, Inc.