黄品溢个人学习笔记

Title 个人思考

Tips 关于未来研究的思考

学生姓名 黄品溢

指导教师 王磊

学　　院 计算机学院

目　录

第一章 个人思考内容

1.1 个人的奇思妙想

1.1.1 关于动作单元 AU

AU 是一个在 FACS（*Facial Action Coding System*，面部行为编码系统）提出的先验概念，在所有使用到 AUs 的方法中，均默认 AU 是正确并可行的，如使用 AU 作为 GCN 中的邻接矩阵进行归一化。是否存在一种可能，可以采用不同的特征作为邻接矩阵或提取特征的单位，由此可以得到更好的空间特征提取的效果？

理由：

1. 往大了说：AU 可以有很多种不同的划分方式，例如可以划分为 11、16 或 36 个 ROI 人脸兴趣区域[1]，是否有可能采用这些划分方式，可以达到更好的特征提取或分类效果？
2. 往小了说：是否将一整个 AU 区域进行特征提取，会带有一定程度的图像噪声，或将更多不相关的特征同时提取进来？

缺点：

1. 若采用 ROI 的方式，最后还是只能使用光流进行特征提取，最后又回到原点；
2. 若要将 AUs 进一步细分，则可能需要人工对视频样本进行标定，提高成本；

1.1.2 关于样本集

这里存在着一个矛盾点：越复杂的神经网络就需要越庞大的训练集和样本数量，由此才能具备更好的分类效果和防止严重的过拟合，而微表情数据库一直很稀缺；若是采用结构稍简单的神经网络，就无法达到更好的分类性能。

由此想到了，是否可以通过预处理，对视频样本的数量进行一定程度的增广，由此扩大样本数量[2]（即老师于 2020 年发表的论文），便可以更好地对神经网络进行训练。

1.1.3 关于神经网络模型的想法

根据论文[1]中的描述，其认为 Encoder of Transformer 具有很高的研究价值，并提出希望他们的研究能够有助于推进未来对于 Transformer 的深入研究。

论文原话：*These findings strongly motivate further research on the use of transformer based architectures rather than convolutional neural networks in micro-expression analysis, and we hope that our theoretical contributions will help direct such future efforts*

这些发现强烈地激发了对在微表达分析中使用基于变压器的架构而不是卷积神经网络的进一步研究，我们希望我们的理论贡献将有助于指导这些未来的努力。

这篇论文中完全摒弃了 CNN 而使用了 Encoder of Transformer 来进行短期和长期空间关系的提取，其创新点在于：虽然同为特征提取，但两者有一个很重要的区别。

对于 CNN 来说：*CNN* 仅在固定窗口大小内提取短期特征，无法提取全局的长期特征。

对于 Encoder 来说：而 *Encoder* 的 *Self-Attention* 可以很好地解决这个问题。

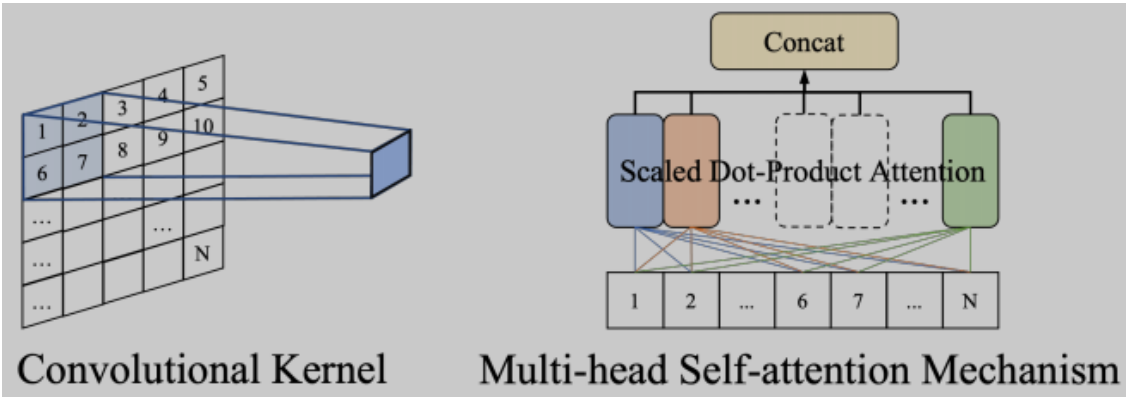

图 1.1 两种不同的空间特征提取方法的对比

而其完整的神经网络 SLSTT 模型结构图如下：

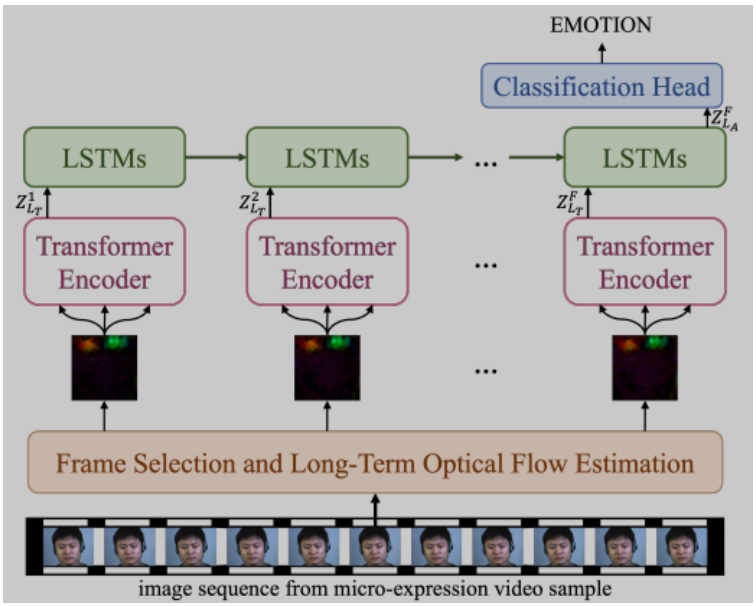

图 1.2 SLSTT 模型总体结构

那么是否存在一种可能，我们可以对 Transformer 的结构进行更进一步的研究，从而提出一些更好的基于 Encoder of Transformer 的模型进行 MER 任务？关于论文[1] 的大致总结如第二章所示。

1.1.4 关于梯度下降方法的想法

由于微表情数据的稀缺，即使是最大的数据库，其包含的样本也很少。由此必须采用一些方法，来防止严重的过拟合现象发生，如：随着神经网络参数逐渐接近最佳时，学习率需要适当降低。

受[1]启发，可以采用 *cosine annealing*[3]方法（暂时还没具体去了解是什么）。

但是在 Pytorch 的优化器（*torch.optim*）中，有一类优化器 SGD（*stochastic Gradient Decent*）随机梯度下降可以采用冲量进行梯度下降。这样的梯度下降算法，在每一步时都会考虑到上一步的更新量，由此达到更好的收敛效果。是否 *cosine annealing* 是比 SGD 等优化器更好的方法？

1.1.5 关于人种问题

有一个很细节的问题：当考虑人种问题时，所有方法的准确率都出现了明显的下滑。所有 MER 方法在 CASME II 上的识别率都显然高于其他数据库[1]，而 CASME II 内只有中国人的样本。

是否有可能改进这一现象，由此能够写一篇论文？

根据论文[1]的描述，这一问题：*The performance of all methods on CASME II is consistently higher than when applied on other data sets, which suggests that the challenge of MER is increased with ethnic diversity of participants – this should be born in mind in future research and any comparative analysis.*

CASME II 中所有方法的表现始终高于应用于其他数据集的方法，这表明 MER 的挑战随着受试者的种族多样性而增加——这应该在未来的研究和任何比较分析中牢记。

第二章 关于已读论文

2.1 *Short and Long Range Relation Based Spatio-Temporal Transformer for Micro Expression Recognition(2021.12)* [1]

个人的一些总结：

1. 将视频中每一帧与微表情起始帧计算光流场，而非任意两连续帧。
   计算相邻帧之间的光流场时，在前半部分有相似的光流，后半部分则强度相似，方向相反。
   而用这里提出的方法，发现光流场总是沿同样的方向，只是前半部分的强度逐渐递增，后半部分的强度逐渐递减。这导致了与每个微表情相关的更稳定和可区别的特征。
2. 首个完全摒弃 CNN 进行特征提取的 MER 神经网络模型。其认为，CNN 只能在固定窗口大小中进行短期空间特则提取，由此无法学习到同样十分重要的长期空间特征。而 Encoder of Transformer 中使用到的 Multi-head Self-attention Mechanism（MSM）自注意力机制，可以很好的学习到短期和长期空间特征，由此其完全摒弃了 CNN 的使用。
3. 时间聚合：使用了 LSTM Aggregator 进行聚合，并与 Mean Aggregator 进行了对比。聚合函数确保了 Transformer 模型可以被训练并应用于每一帧的空间特征集，然后处理每个样本中各帧之间的时间关系。

其具体细节的总结和个人理解见 2022.07.31.pdf

参考文献

[1] L. Zhang, X. Hong, O. Arandjelovic, and G. Zhao, "Short and long range relation based spatio-temporal transformer for micro-expression recognition," *arXiv preprint arXiv:2112.05851*, 2021.

[2] L. Wang, J. Hou, X. Guo, Z. Ma, X. Liu, and H. Fang, "Micro-expression video clip synthesis method based on spatial-temporal statistical model and motion intensity evaluation function," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 211–217, IEEE, 2020.

[3] M. Inc., "Face++ research toolkit," *[Online]*, 2013. Available: www.faceplusplus.com.