首都师范大学
Capital Normal University

首都师范大学本科生课程作业

基于双向融合注意力的多模态多路特征融合方法

作　　者：李珂
院　　系：信息工程学院
专　　业：人工智能
学　　号：1200604012

课程名称：信息科学前沿 1
完成日期：2024 年 6 月 6 日

摘　　要

本文基于 CMU-MOSI 多模态（语音、文本、视频）情感分析数据集对情感识别技术进行研究，并基于现有的 CubeMLP 模型提出改进方案。CMU-MOSI 数据集包含 YouTube 独白视频，涵盖文本、音频和视频三种模态，标注为[-3, 3]之间的情感分数。CubeMLP 是一种基于多层感知器（MLP）的多模态特征融合框架，通过 Sequential-mixing、Channel-mixing 和 Modality-mixing 三种方向进行特征融合。

在实验中，本文将数据集任务分为回归、二分类和七分类三种，分别评估模型在这些任务上的表现。本文基于 CubeMLP 模型，提出了两种改进方案：方案 A 为**简单多路模态融合**，方案 B 为**多路模态融合+双向融合注意力机制**。实验结果显示，改进方案 A、B 在回归和二分类任务上取得了更好的效果，**MAE(0.018↓)**、**Corr(0.021↑)**，尤其是多路模态融合在二分类任务上表现出更高的准确率和 F1 分数，**ACC-2(0.030↑)**、**F1-2(0.031↑)**。

此外，本文可视化了三种方法的训练的收敛过程和验证集上的效果变化。本研究展示了多模态特征融合在情感分析中的潜力，为未来的研究提供了有益的参考。但因个人算力有限，未进行更多的调参优化，未来的工作将继续优化模型参数和融合策略，以进一步提升模型性能。

关键词：多模态融合、情感识别、深度学习

目 录

1 背景介绍

多模态特征融合是当前心理状态预测和情感分析等领域的关键技术之一。随着社交媒体和数字交流的普及，个人和公众产生的数据不再局限于单一形式，而是呈现出多模态的特征，包括文本、声音和视觉等信息[1] ~ [4] 。因此，如何有效地整合和利用这些多模态数据成为了研究的重要课题。

过去的研究主要集中在开发有效的融合策略，以整合来自不同模态的与心理状态相关的信息。其中一些技术基于机器学习模型，特别是多层感知器（MLP）[5] [6] [7] ，在计算机视觉等任务中取得了显著的成功。受到这些成果的启发，近期的研究开始从特征融合的角度探索多模态方法。

在多模态特征融合的研究中，主要涉及以下几个方面的工作：

1、特征表示学习：不同模态的数据具有不同的特点和表示方式，研究者致力于通过深度学习等技术，学习到适合多模态数据的特征表示。例如，可以使用卷积神经网络（CNN）[8] [9] [10] 来学习图像特征，使用循环神经网络（RNN）来学习序列数据特征。

2、特征融合方法：多模态特征的融合是多模态特征融合的核心任务。研究者提出了多种融合方法，包括早期融合[12] 、中间融合[13] 和后期融合[14] 等。早期融合将不同模态的数据在特征层面进行融合，中间融合将不同模态的数据在网络层面进行融合，后期融合将不同模态的数据在决策层面进行融合。

3、融合策略的选择：在多模态特征融合中，选择适合当前任务的融合策略非常重要[15] [16] [17] 。不同的融合策略会导致不同的性能和效果。研究者通过实验和比较，选择最佳的融合策略来提高任务的性能。

多模态特征融合旨在充分利用多源数据中的信息，以提高心理状态预测和情感分析等任务的性能。通过有效地整合不同模态的特征，这些方法能够更准确地捕捉个体的心理状态变化，为相关领域的研究和应用提供了重要的支持和参考。

2 相关工作

2.1 CubeMlp

CubeMLP[7] 是一个多模态特征融合的框架，它完全基于 MLP 构建，并采用了特定的多模态特征处理策略。CubeMLP 由三个独立的 MLP 单元组成，每个单元包含两个仿射变换，这使得它能

够接受所有相关的模态特征作为输入，并在三个轴上进行混合。在使用 CubeMLP 提取特征后，混合的多模态特征被平坦化，以便用于情绪分析和抑郁估计等任务的预测。

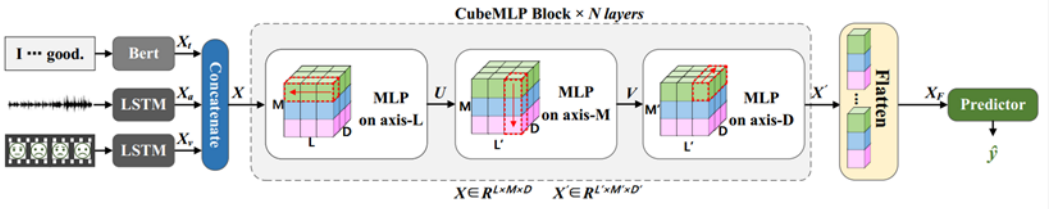
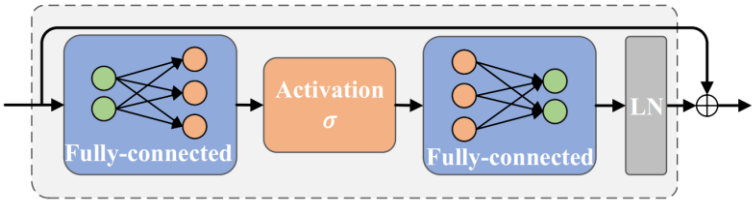


图 1 CubeMlp 模型架构

2. 1. 1 多模态特征提取

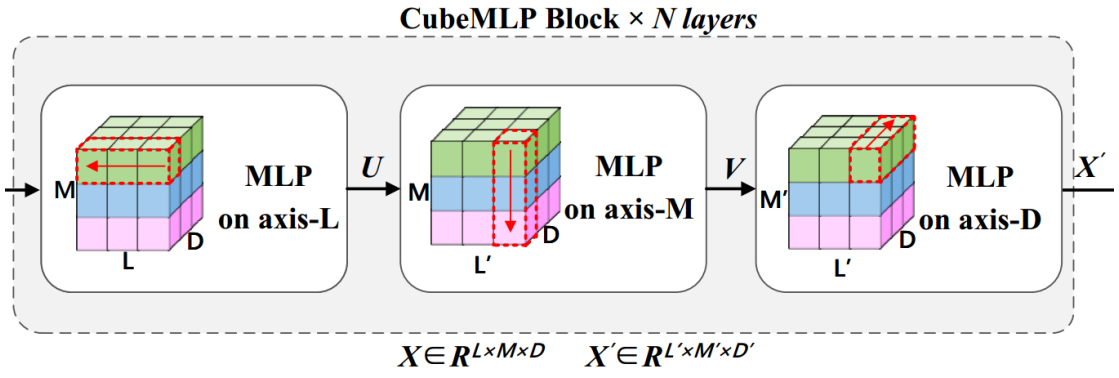
数据集提供原始文本与语音和视频模态的提取特征向量，其中的原始文本经过预训练 Bert[1]进行特征提取。语音和视频模态的特征向量$V \in R^{L\times d}$会经过 LSTM 进行进一步捕捉时序特征。最终得到输入模态张量为$X_i \in R^{L_i\times d}$，其中$i \in \{t,a,v\}$分别为文本模态、语音模态、视频模态特征。为了统一模态长度 L，该模型采用 0 填充的方式来补齐张量形状。最终获得长度相同的特征模态$X_i \in R^{L\times d}$。随后将三个模态特征，在新维度M堆叠起来，最终得到组合特征$X \in R^{L\times M\times D}$。

2. 1. 2 多模态特征融合



图表 1 CubeMLP 的 MLP 单元

多模态特征融合使用一个基本的非线性映射作为算子。MLP 单元设计为两个全连接层中间添加一个非线性激活，然后使用 layer normalization 对输出进行归一化。



图表 2 多维度特征融合

CubeMLP 定义了三个方向的特征融合方法。

[1] Huggingface 提供的预训练模型：“bert-base-uncased”

1. Sequential-mixing，L 轴方向特征融合：L 轴向量代表每个模态不同时刻特征向量，对其使用 MLP 单元进行非线性映射，并对输出在 L 轴进行 layer normalization。作用在于对单模态内特征进行混合。
2. Channel-mixing，M 轴方向特征融合：M 轴方向为多个模态特征的堆叠。在 M 轴上使用 MLP 单元，作用在于多个模态的特征融合。
3. Modality-mixing，D 轴方向特征融合：D 轴方向为单模态的单特征的向量嵌入特征，是相对独立的特征，但经过上述两个方向的融合已经变得不在独立。

CubeMLP 模型经过连续的三个方向的特征融合，显式地增强了神经网络的特征融合能力。

3 本文方法

3.1 方案 A：多路模态融合

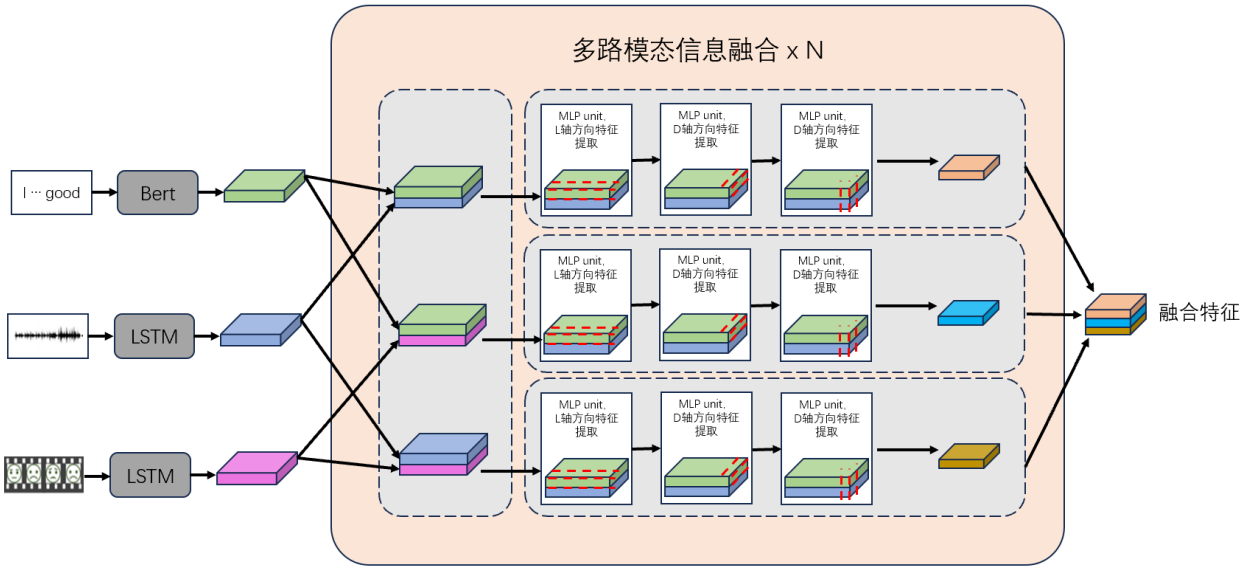


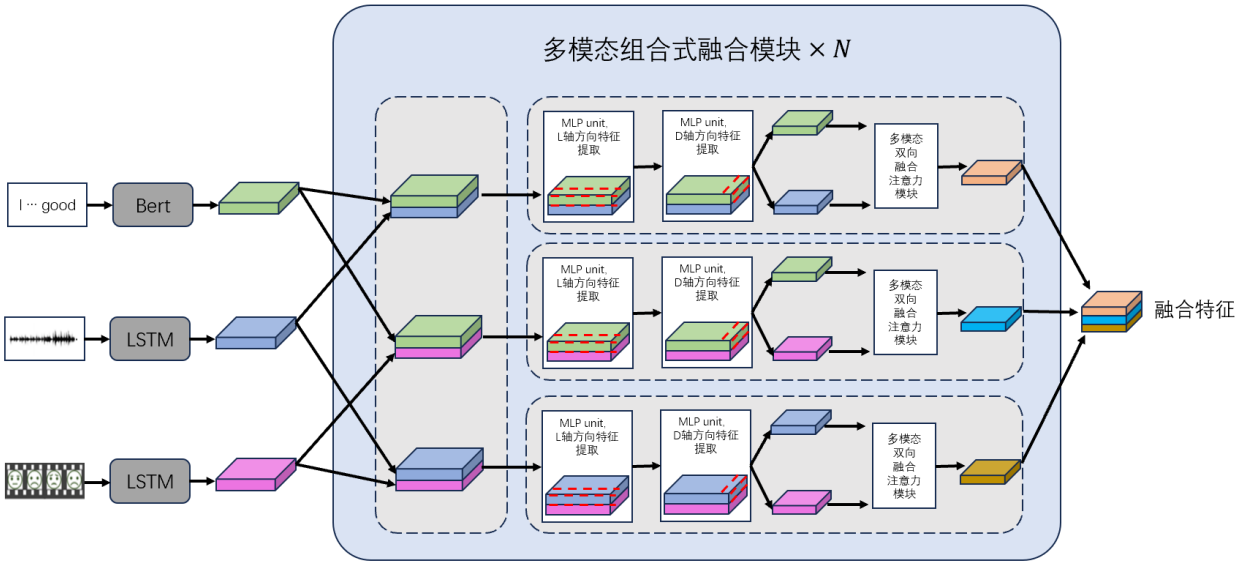
图 2 方案 A: 多路模态信息融合

3.1.1 组合式模态信息增强

CubeMLP 的方法仅单路信息流通，其限制在于造成方法特征数量有限，因此本文将文本、语音、视频三个模态数据进行全组合来进行数据特征增强[22] [23]，并希望得到更加有用的任意两模态的混合特征。

在模态数量维度将双模态$X \in R^{L\times 2\times D}$通过全连接层变换成单个特征图$X \in R^{L\times 1\times D}$。

3.2 方案 B：多路模态融合+双向融合注意力机制



图表 3 方案 B: 多路模态融合+双向融合注意力机制

据集提供原始文本与语音和视频模态的提取特征向量，其中的原始文本经过预训练 Bert[2]进行特征提取。语音和视频模态的特征向量$V \in R^{L\times d}$会经过 LSTM 进行进一步捕捉时序特征。

最终得到输入模态张量为$X_i \in R^{L_i\times d}$，其中$i \in \{t,a,v\}$分别为文本模态、语音模态、视频模态特征。

为了统一模态长度 L，该模型采用 0 填充的方式来补齐张量形状，最终获得长度相同的特征模态$X_i \in R^{L\times d}$。

3. 2. 1 单模态特征提取

为了便利之后的多模态特征融合，本文在融合之前对单个模态特征进行特征提取，借鉴[7] [18] [19] 基于张量网络的方法的特征提取方法。

1、Sequential-mixing ， L 轴方向特征融合：

L 轴向量代表每个模态不同时刻特征向量，对其使用 MLP 单元进行非线性映射，并对输出在 L 轴进行 layer normalization。<u>作用在于对单模态内特征进行混合</u>。

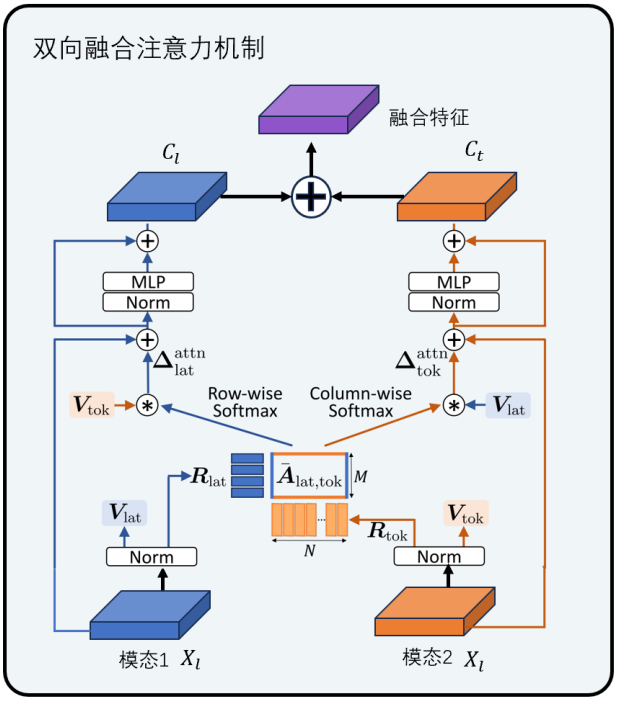
2、Modality-mixing ， D 轴方向特征融合：

获得单模态特征不同 token 之间信息交互。

通过上述两步操作，在低计算代价下完成了特征提取。

同 CubeMlp。

3. 2. 2 跨模态特征融合——双向融合注意力机制



图表 4 双向融合注意力机制

将两个组合模态完成特征提取后，本文使用一个改进自[21] 的 双向融合注意力模块进行特征融合，该模块可以通过向量内积将两个模态特征进行对齐。

假设输入模态特征为$X_l \in R^{L\times D}$和$X_t \in R^{L\times D}$，其将计算注意力矩阵：

$$sim = \frac{W_l * LN(X_l) * [W_t * LN(X_t)]^T}{\sqrt{D}}$$

其中 LN 为归一化层，W为全连接层参数。

同时，使用$V_{lat} = W_{l1} * LN(X_l)$,$V_{tok} = W_{t1} * LN(X_t)$将两个特征进行特征变换。

Sim 矩阵两个维度分别代表两个模态的相似度值分布，因此

$$sim = V_{lat} * V_{tok}^T$$

$$sim_l = softmax(sim, dim = 0)$$

$$sim_t = softmax(sim, dim = 1)$$

sim_l中每一行元素代表V_{lat}的第一行向量与V_{tok}每一列向量的相关分数。

然后通过点积计算保留的模态特征，即

$$O_{lat} = sim_l * V_{tok}$$

$$O_{tok} = sim_t * V_{lat}$$

为了保证训练的稳定性，最终结果表述为：

$$C_l = W_2 * LN(O_{lat} + X_l) + O_{lat} + X_l$$

$$C_t = W_2 * LN(O_{tok} + X_t) + O_{tok} + X_t$$

3.3 预测头

经过连续 N 个多模态组合式特征融合模块，将得到 3 个形状为$X \in R^{L\times 1\times D}$的**融合特征**，可用于下游任务。本文将三个张量在 dim=1 维度拼接，再展平成向量，通过多层感知机进行分类与回归任务。

4 实验设置

4.1 数据集

CMU-MOSI 数据集是一个 YouTube 独白的集合，说话者表达他们对电影等主题的意见。MOSI 总共有 93 个视频，横跨 89 个距离说话者，包含 2198 个主观语词-视频片段。这些话语被人工标注为[-3, 3]之间的连续意见分数，其中-3/+3 代表强烈的负面/正面情感。

CMU-MOSI 数据集是一个用于情感分析和情感识别研究的公开数据集，由卡内基梅隆大学（Carnegie Mellon University）的研究人员创建。该数据集主要基于 YouTube 上的影片片段，涵盖了多种情感表达和情感体验，包括愉快、悲伤、愤怒等。这些影片片段来自于各种主题，如电影、新闻报道、独白等，以确保数据的多样性和代表性。

CMU-MOSI 数据集提供了多模态的数据，包括视频、音频和文本。其中，视频部分包括了演讲者的面部表情和身体语言，音频部分则包括了演讲者的语音信号，而文本部分则是演讲者的言论内容。每个样本都配有相应的情感标签，标记了影片片段中表达的情感类别，如高兴、悲伤、愤怒等。

CMU-MOSEI 中的数据每种模态的提取特征如下：

文本：所有视频都有手动转录。使用 P2FA 强制对齐模型在音素级别对齐单词和音频。 在此之后，视觉和听觉模式通过插值与单词对齐。 由于英语单词的发音持续时间通常较短，因此这种插值不会导致大量信息丢失。

视觉：从 30Hz 的完整视频中提取帧。使用 MTCNN 人脸检测算法提取人脸的边界框。我们通过面部动作编码系统提取面部动作单元。 提取这些动作单元可以准确跟踪和理解面部表情。 还使用 Emotient FACET 从静态人脸中提取了一组六种基本情绪。 MultiComp OpenFace 用于提取 68 个面部标志、20 个面部形状参数、面部 HoG 特征、头部姿势、头部方向和眼睛注视的集合。最后，我们从常用的面部识别模型中提取人脸嵌入。

声学：我们使用 COVAREP 软件提取声学特征，包括 12 个 Mel 频率倒谱系数、音高、浊音/清音分段特征、声门源参数、峰值斜率参数和最大色散商数。 所有提取的特征都与情绪和语气有关。

即，<u>涉及隐私的视频与语音数据，均提供的是特征提取完的向量嵌入</u>。

4.2 评估指标

该数据集有文本、语音、视频三个模态，对应标签为情感分数标量，范围[-3,3]，因此可将数据集转化为三个任务：

1、回归任务：模型回归分数值。

2、二分类任务：情感分数>=0，为正面情绪。情感分数<0，为负面情绪。

3、七分类任务：-3~+3，包括-3,-2,-1,0,1,2,3 共 7 个分数值，可设置为 7 分类任务。

本文将对三种任务上效果进行指标评价，回归任务计算 MSE，分类任务计算准确率与 F1 分数。

4.3 超参数设置

4.3.1 基线模型

部分关键超参

!python /kaggle/input/cubemlp-main/Train.py --dataset mosi_SDK --batch_size 128 --features_compose_t mean --features_compose_k cat --d_hiddens 50-2-128=10-2-32 --d_outs 50-2-128=10-2-32 --res_project 1-1 --bias --ln_first --dropout_mlp 0.1-0.1-0.1 --dropout 0.1-0.1-0.1-0.1 --bert_freeze part --bert_lr_rate 0.01 --learning_rate 4e-3

4.3.2 方案 A：简单多路融合

部分关键超参

!python /kaggle/input/cubemlp-main/Train.py --dataset mosi_SDK --batch_size 128 --features_compose_t mean --features_compose_k cat --d_hiddens 50-2-128=10-2-32 --d_outs 50-2-128=10-2-32 --res_project 1-1 --bias --ln_first --dropout_mlp 0.1-0.1-0.1 --dropout 0.1-0.1-0.1-0.1 --bert_freeze part --bert_lr_rate 0.01 --learning_rate 4e-3

4.3.3 方案 B：多路模态融合+双向融合注意力机制

部分关键超参

!python /kaggle/input/cubemlp-main/Train.py --dataset mosi_SDK --batch_size 128 --features_compose_t mean --features_compose_k cat --d_hiddens 50-2-128=10-2-32 --d_outs 50-2-128=10-2-32 --res_project 1-1 --bias --ln_first --dropout_mlp 0.1-0.1-0.1 --dropout 0.1-0.1-0.1-0.1 --bert_freeze part --bert_lr_rate 0.01 --learning_rate 4e-3

详细超参数见补充材料代码。

5 实验结果与可视化

5.1 效果评估

将 CMU-MOSI 数据集建模为三个任务：

1、回归任务：直接回归情绪分数。计算 MAE 和皮尔斯相关系数(Corr)。
2、二分类任务：情绪分为正面情绪、负面情绪。计算准确率与 F1 分数。
3、七分类任务：将-3~+3 的情绪得分，每份作为一类。计算准确率与 F1 分数。
基线模型复现 CubeMlp 的官方实现提供的复现指令[28]，看起来比论文中公布结果略低。

	MAE	Corr	Acc-2	F1-Score-2	Acc-7	F1-Score-7
基线复现	0.808	0.751	0.803	0.803	**0.420**	**0.398**
方案 A	**0.790**	0.770	**0.832**	**0.831**	0.374	0.348
方案 B	0.797	**0.771**	0.818	0.820	0.341	0.316

表 1 CMU-MOSI 的测试集效果。基线模型为 CubeMlp 开源代码复现超参。方案 A 为简单多路融合模型。方案 B 为多模态融合+双向融合注意力机制。

实验结果显示，多路特征融合有比 CubeMlp 更有竞争力的结果。在二分类任务中，简单多路融合方案有着比基线模型更明显的提升。但在七分类任务中，基线模型效果较好。

5.2 可视化效果

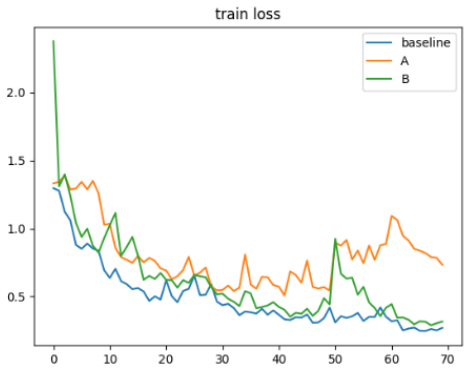


图 3 训练损失曲线

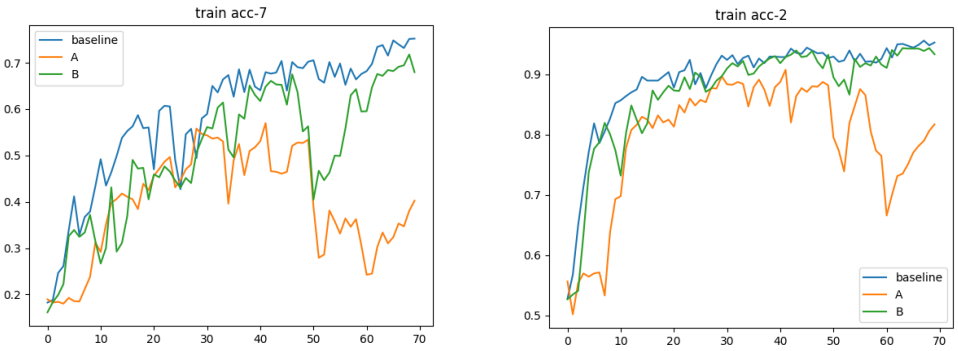


图 4 训练集准确率。左图为七分类，右图为二分类。

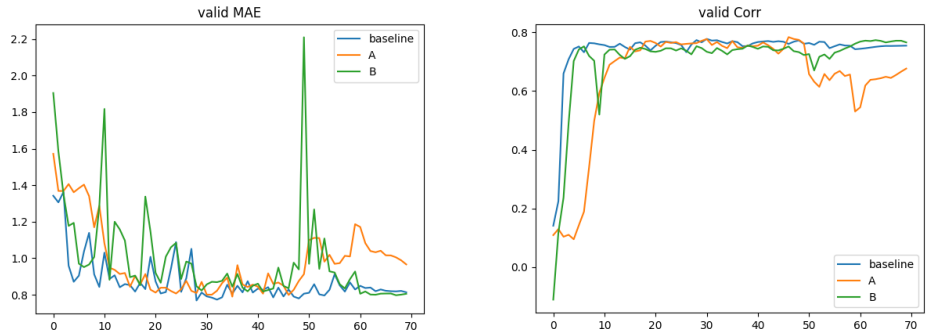


图 5 验证集指标变化。左图为验证集 MAE，右图为验证集皮尔森相关系数。

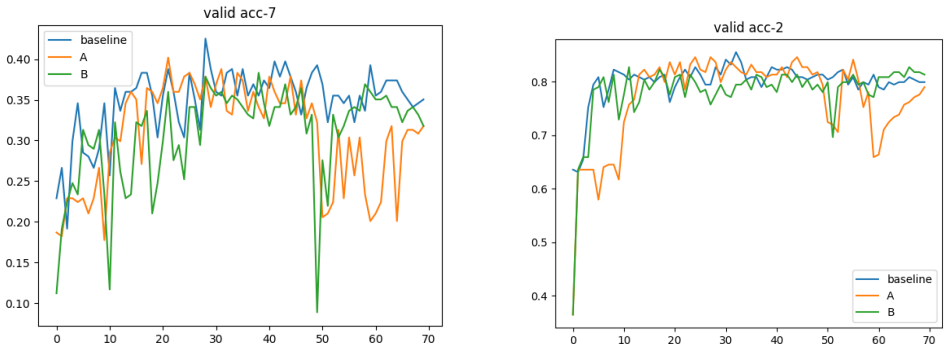


图 6 验证集准确率。左图为七分类，右图为二分类。

1、由图 2 可以看出，多路特征融合的方式收敛速度变慢。

2、由图 4 可以看出，多路特征融合的方式稳定性较差。

3、图 5 可以看出双向融合注意力机制在 2 分类任务下效果好。

由 1、2 推测多路融合的方式优化过程不稳定可能源于特征两两组合这一步未设计**残差连接**，导致深层神经网络优化问题，后续可从这一步进行改进。

6 总结与回顾

回顾在《信息科学前沿 1》这门课程，让我了解了一些语音信号处理的基本概念、技术和算法，用于处理和分析语音信号。

首先，我们学习了语音信号的基本特性和表示方法。语音信号通过采样和量化将其数字化表示。我们还学习了语音信号的频域和时域表示方法，如傅里叶变换和时域波形。

接下来，我们学习了语音信号的预处理技术。预处理是在语音信号进入后续处理步骤之前对其进行的一系列操作。这包括去噪、降噪、均衡化、滤波等技术，目的是消除噪声和增强语音信号的质量。

我们研究了语音信号的特征提取技术。这些特征提取方法用于从语音信号中提取有用的信息，如音频能量、频谱等。这些特征可以用于语音识别、语音合成、语音转换等各种应用中。

最后，我们还学习了一些基于神经网络的语音特征提取算法。这包括语音识别、语音分离、情感识别等。

这门课最后的课程设计，我选择深入探索多模态特征融合方向，这也是我个人希望未来发展的方向，并尝试进行了一些改进，初步取得了一些可观效果。我相信经过适当调参，本文引入的改进很有潜力超越 baseline 模型，但因个人算力有限，没有进一步优化实验。

7 参 考 文 献

[1] Gao, Jing, et al. "A survey on deep learning for multimodal data fusion." Neural Computation 32.5 (2020): 829-864.

[2] Stahlschmidt, Sören Richard, Benjamin Ulfenborg, and Jane Synnergren. "Multimodal deep learning for biomedical data fusion: a review." *Briefings in Bioinformatics* 23.2 (2022): bbab569.

[3] Jabeen, Summaira, et al. "A review on methods and applications in multimodal deep learning." *ACM Transactions on Multimedia Computing, Communications and Applications* 19.2s (2023): 1-41.

[4] Sun, Zhaoyi, et al. "A scoping review on multimodal deep learning in biomedical images and texts." *Journal of Biomedical Informatics* (2023): 104482.

[5] Touvron, Hugo, et al. "Resmlp: Feedforward networks for image classification with data-efficient training." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022): 5314-5321.

[6] Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." *Advances in neural information processing systems* 34 (2021): 24261-24272.

[7] Sun, Hao, et al. "CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation." Proceedings of the 30th ACM international conference on multimedia. 2022.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*, 25, 1097-1105.

[9] Simonyan, K., & Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations (ICLR)*.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

[11] Gregory, Philip A., et al. "The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1." Nature cell biology 10.5 (2008): 593-601.

[12] Tziafas, Georgios, and Hamidreza Kasaei. "Early or late fusion matters: Efficient rgb-d fusion in vision transformers for 3d object recognition." *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.
[13] Mousa-Pasandi, Morteza, et al. "RGB-LiDAR fusion for accurate 2D and 3D object detection." *Machine Vision and Applications* 34.5 (2023): 86.
[14] Cai, Qi, et al. "Objectfusion: Multi-modal 3d object detection with object-centric fusion." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
[15] Zou, Shinan, et al. "A multi-stage adaptive feature fusion neural network for multimodal gait recognition." *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2024).
[16] Fu, Baole, et al. "A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals." *Frontiers in Neuroscience* 17 (2023): 1234162.
[17] Wang, Yuanyuan, et al. "Multimodal transformer augmented fusion for speech emotion recognition." *Frontiers in Neurorobotics* 17 (2023): 1181598.
[18] Zadeh, Amir, et al. "Tensor fusion network for multimodal sentiment analysis." arXiv preprint arXiv:1707.07250 (2017).
[19] Gu, Yue, et al. "Multimodal affective analysis using hierarchical attention strategy with word-level alignment." Proceedings of the conference. Association for Computational Linguistics. Meeting. Vol. 2018. NIH Public Access, 2018.
[20] Zadeh, Amir, et al. "Multi-attention recurrent network for human communication comprehension." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
[21] Hiller, Markus, Krista A. Ehinger, and Tom Drummond. "Perceiving Longer Sequences With Bi-Directional Cross-Attention Transformers." arXiv preprint arXiv:2402.12138 (2024).
[22] Fu, Yanping, et al. "Hybrid cross-modal interaction learning for multimodal sentiment analysis." *Neurocomputing* 571 (2024): 127201.
[23] Wang, Yue, et al. "Multimodal industrial anomaly detection via hybrid fusion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
[24] Wang, Yikai, et al. "Learning deep multimodal feature representation with asymmetric multi-layer fusion." Proceedings of the 28th ACM International Conference on Multimedia. 2020.
[25] Cui, Yuhang, Shengbin Liang, and YuYing Zhang. "Multimodal representation learning for tourism recommendation with two-tower architecture." *Plos one* 19.2 (2024): e0299370.
[26] Boulahia, Said Yacine, et al. "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition." *Machine Vision and Applications* 32.6 (2021): 121.
[27] Bodaghi, Morteza, Majid Hosseini, and Raju Gottumukkala. "A Multimodal Intermediate Fusion Network with Manifold Learning for Stress Detection." *arXiv preprint arXiv:2403.08077* (2024).
[28] https://github.com/kiva12138/CubeMLP

8 补充材料

8.1 CubeMlp 原文实验效果

图中分别展示了 CubeMLP 将数据集视为回归任务(MAE)、二分类任务(Acc-2)和七分类任务(Acc-7)时的效果。

Models	CMU-MOSI					CMU-MOSEI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
TFN[37]	0.970	0.633	73.9	73.4	32.1	0.593	0.700	82.5	82.1	50.2
MFN[38]	0.965	0.632	77.4	77.3	34.1	-	-	76.0	76.0	-
ICCN[30]	0.862	0.714	83.0	83.0	39.0	0.565	0.713	84.2	84.2	51.6
SWAFN[2]	0.880	0.697	80.2	80.1	40.1	-	-	-	-	-
MulT[33]	0.871	0.698	83.0	82.8	40.0	0.580	0.703	82.5	82.3	51.8
LMF-MulT[27]	0.957	0.681	78.5	78.5	34.0	0.620	0.668	80.8	81.3	49.3
MAT[3]	-	-	-	80.0	-	-	-	82.0	82.0	-
MNT[3]	-	-	-	80.0	-	-	-	80.5	80.5	-
MISA[12]	0.817	0.748	82.1	82.0	41.4	0.557	0.748	84.9	84.8	51.7
BBFN[11]	0.776	0.755	84.3	84.3	45.0	**0.529**	**0.767**	**86.2**	**86.1**	54.8
CubeMLP(Ours)	**0.770**	**0.767**	**85.6**	**85.5**	**45.5**	**0.529**	0.760	85.1	84.5	**54.9**

图表 5 CubeMLP 在 MOSI 和 MOSEI 上的效果

Models	AVEC2019	
	CCC(↑)	MAE(↓)
Baseline [25]	0.111	6.37
Adaptive Fusion Transformer [29]	0.331	6.22
EF [16]	0.344	-
Bert-CNN & Gated-CNN [26]	0.403	6.11
Multi-scale Temporal Dilated CNN [8]	0.430	4.39
Hierarchical BiLSTM [36]	0.442	5.50
CubeMLP(Ours)	**0.583**	**4.37**

图表 6 CubeMLP 在 AVEC2019 上的效果

8.1.1 消融实验

	MLP-L	MLP-M	MLP-D	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Model 1	✓			0.860	0.744	80.6	80.7	39.0
Model 2		✓		0.850	0.729	80.3	80.4	39.2
Model 3			✓	0.910	0.717	81.7	81.8	39.0
Model 4	✓	✓		0.806	0.753	81.5	81.6	42.4
Model 5	✓		✓	0.803	0.750	80.6	80.8	39.5
Model 6		✓	✓	0.874	0.718	82.4	82.4	41.6
Model 7(Ours)	✓	✓	✓	**0.770**	**0.767**	**85.6**	**85.5**	**45.5**

图表 7 消融实验效果

论文中作者对 Cube MLP block 的三中特征融合方式进行消融实验。其中 MLP-L 表示 L 轴方向融合，MLP-M 表示 M 轴方向融合，MLP-D 表示 D 轴方向融合。

从中可以得到以下结论：

1、MLP-M 的效果要好于 MLP-L 好于 MLP-D
2、增加融合方式效果会变好

可以看出 CubeMLP block 的有效性。

8. 1. 2 模型空间复杂度

Models	Space Consumption
TFN [37]	$O(L^M)$
Adaptive Fusion Transformer [29]	$O(L^2)$
ICCN[30]	$O(L \times D^2)$
MNT& MAT[3]	$O(L^2)$
BBFN[11]	$O(L^2)$
CubeMLP(Ours)	$O(max(L, M, D))$

图表 8 模型复杂度对比