Task-level特征融合策略1：

语音：用Wav2Vec计算每个片段的特征（具体长度要看论文），然后平均值。

文本：用Bert计算句子特征。

人脸：随机选取一张图像，用DeiT计算特征。

Task-level特征融合策略2：

语音：用Wav2Vec计算每个片段的特征+attention

文本：用Bert计算句子特征。

人脸：用DeiT计算特征+attention融合，但是有些样本人脸数量不够

Fusion策略：

Early fusion

Late fusion

TensorNet

EmbraceNet

Attention? (no enough data)

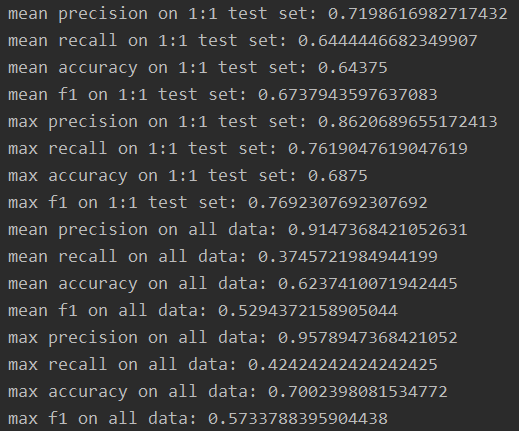
Hypergraph

实验结果

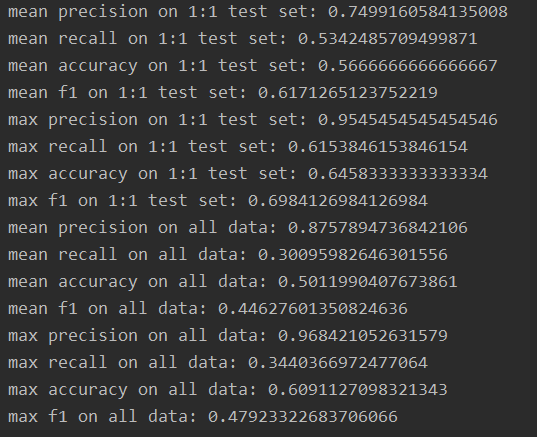
单模态：

100次训练，记录acc

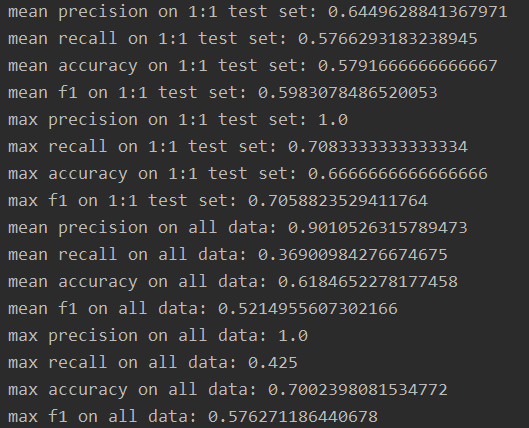
Text



Wav



Video

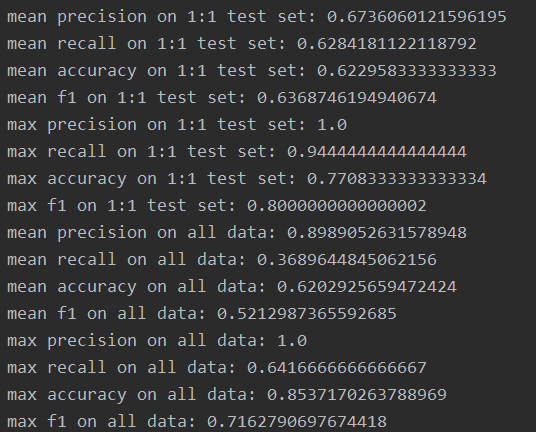


多模态融合

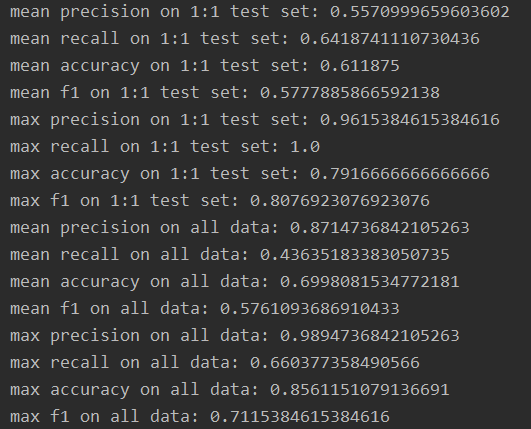
100次训练

记录模型的Precision/ recall/ Acc/ F1最好值和平均值

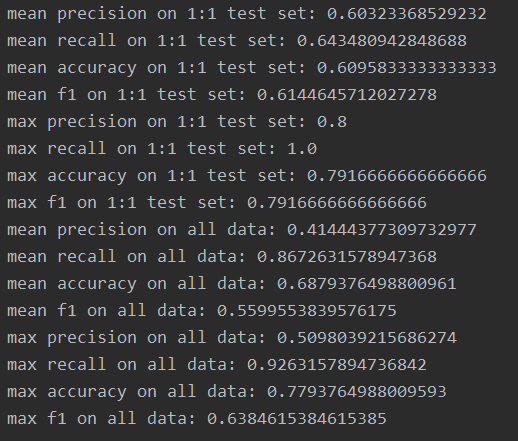
Early fusion：



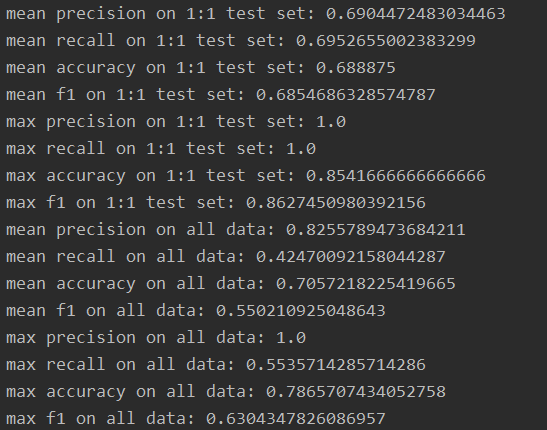
Late fusion:



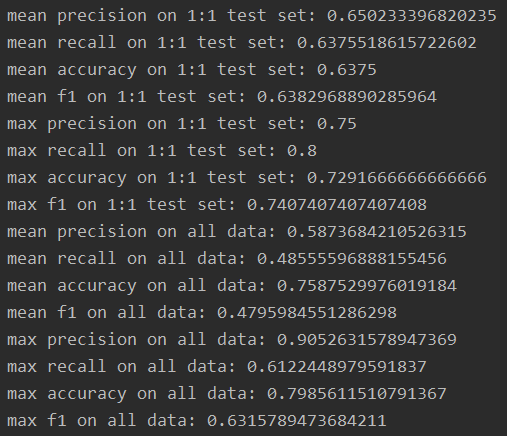
MKL:



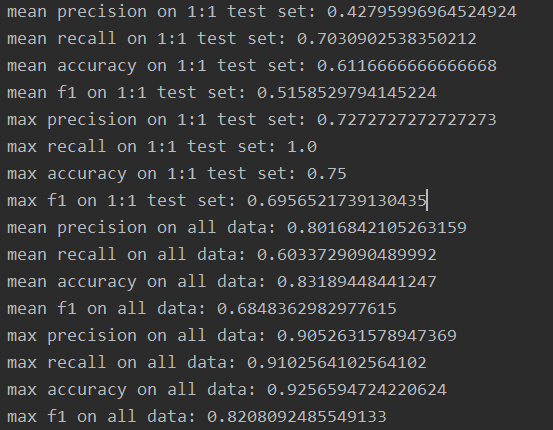
EmbraceNet:



TEN:



LMF:



Hypergraph:

1:1 test set

Best (mean)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method Name | Accuracy | Recall | Precision | F1 score |
| Early fusion+SVM | 77.08%(62.30%) | 94.44%(62.84%) | 100.00%(67.36%) | 80.00%(63.69%) |
| Late fusion+SVM | 79.17%(61.19%) | 100.00%(64.19%) | 96.15%(55.71%) | 80.77%(57.78%) |
| MKL | 79.17%(60.96%) | 100.00%(64.35%) | 80.00%(60.32%) | 79.17%(61.45%) |
| TEN | 72.92%(63.75%) | 80.00%(63.76%) | 75.00%(65.02%) | 74.07%(63.83%) |
| LMF | 75.00%(61.17%) | 100.00%(70.31%) | 72.73%(42.80%) | 69.57%(51.59%) |
| EmbraceNet | **85.42%(68.89%)** | 100.00%(69.53%) | 100.00%(69.04%) | 86.27%(68.55%) |
| Hypergraph |  |  |  |  |

On all data

Best (mean)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method Name | Accuracy | Recall | Precision | F1 score |
| Early fusion+SVM | 85.37%(62.03%) | 64.17%(36.90%) | 100.00%(89.89%) | 71.63%(52.13%) |
| Late fusion+SVM | 85.61%(69.98%) | 66.03%(43.63%) | 98.95%(87.15%) | 71.15%(57.61%) |
| MKL | 77.93%(68.79%) | 92.63%(86.73%) | 50.98%(41.44%) | 63.85%(56.00%) |
| TEN | 79.86%(75.88%) | 61.22%(48.56%) | 90.53%(58.74%) | 63.16%(47.96%) |
| LMF | **92.57%(83.19%)** | 91.03%(60.34%) | 90.52%(80.17%) | 82.08%(68.48%) |
| EmbraceNet | 78.66%(70.57%) | 55.36%(42.47%) | 100.00%(82.56%) | 63.04%(55.02%) |
| Hypergraph |  |  |  |  |

实验结果分析