Multi-Task Multi-Modal Self-Supervised Learning for Facial Expression Recognition

Supplementary Material

6. Implementation Details

We extract features from the input modalities using pretrained fixed feature extractors. The frames for the 2D ResNet-152 [24] (ImageNet [13]) are subsampled to 1fps and for the 3D ResNet-101 [3] (Kinetics [27]) to 16fps. The audio is transformed into Mel spectrograms before processing with a DAVENet [23] pretrained on affective audio. For text, we take the last hidden layer of a DistillBERT [48] that was trained for sentiment analysis. To obtain a fixed-size feature vector per modality, we process the inputs sequentially and average the resulting vectors over time. The 2D and 3D frame features are concatenated after averaging. The result is stored on disk, *i.e.* no finetuning of the backbone feature extractors happens. We use AdamW optimizer [36] for both pertaining and downstream training, in combination with cosine annealing with warm restarts [37] as learning rate scheduling. Our batch size is 4096 and the image size is 180 by 180 pixels (cropped or resized). The size of the representations is 4096. More details are given in the supplementary material. We used the following learning rates in our experiments:

Pretraining:

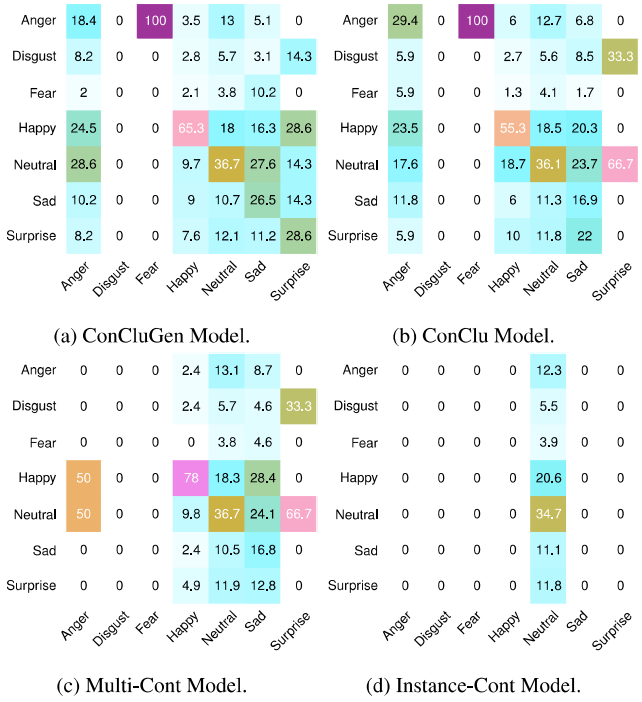
- ConCluGen: lr=0.00009, weight decay=0.00032
- ConClu: lr=0.00009, weight decay=0.00032
- ConGen: lr=0.00036, weight decay=0.00032
- Multi-Cont: lr=0.00036, weight decay=
- Instance-Cont: lr=0.00036, weight decay=00032
- Generative: lr=0.00036, weight decay=0.00032

Downstream: (We chose common hyperparameters that worked well for all methods)

- CAER: lr=0.0061, weight decay=0.08216
- MELD: lr=0.00967, weight decay=0.00004
- MOSEI: lr=0.00996, weight decay=0.00007

For the K-means clustering, we choose a queue size of 4 (*i.e.* 4 batches were considered in the clustering), 8 clusters and started the clustering in epoch 12.

7. Confusion Matrices for CAER



(a) ConCluGen Model. (b) ConClu Model.

(c) Multi-Cont Model. (d) Instance-Cont Model.

Figure 3. Confusion metrics for CAER dataset over different models.