HW2 - Multimodal Machine Learning for Emotion Recognition

CSCI 535 - University of Southern California

Due Date: March 23, 2021

1 Goal

In this exercise, we perform multimodal machine learning tasks on multiple emotion categories using features obtained from pre-trained models.

2 Dataset

For this assignment, you have access to the **IEMOCAP** (Interactive Emotional Dyadic Motion Capture) database [1]. IEMOCAP is a well-established and largely used database in Affective Computing.

This database is recorded from 10 actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expressions and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions.

The IEMOCAP dataset provides segments of data including multiple utterances from both subjects in the dyadic interaction. The dataset has been annotated for 11 different emotion classes by multiple annotators. The labels are: anger, sadness, happiness, neutral, other, unknown, frustration, surprise, fear, excitement, and disgust.

In this excecrise, you should only focus on 4 of these 11 classes, namely **anger(0)**, **sadness(1)** and **happiness(2)**, and **neutral(3)**.

You are provided with visual, speech and text features for the dataset.

For the visual features, you have face embeddings obtained from a **ResNet** model [4] pre-trained on ImageNet. You have $T \times 2048$ matrix for each utterance, where T denotes the number of frames.

For the acoustic features, you have **VGGish**, a deep convolutional neural network pre-trained on audio spectrograms extracted from a large database of videos to recognize a large variety of audio event categories [3]. The 128-dimensional embeddings were generated by VGGish after dimensionality reduction with Principal Component Analysis (PCA).

For the textual features, you have Bidirectional Encoder Representations from Transformers (**BERT**) [2]. BERT is a pre-trained language representation model that has substantially advanced the state-of-the-art in a number of natural language processing (NLP) tasks including sentiment analysis. The 768-dimensional vectors are obtained using the bert-base-uncased model.

Among the provided features, BERT encodes the whole text sequence into a fixed size vector, and unlike audiovisual modalities, the temporal dimension is latent in the text representation. Additionally all the features have been normalized per speaker.

3 Your tasks

In the file **dataset.csv**, you are provided with the relative address for the audio, visual and text feature files along with their corresponding emotion labels. There are 5 sessions and each session has one male and one female speaker.

1. You can use different pooling methods (e.g., max pooling, mean pooling) for reducing the temporal dimension of the audio and visual files, or use your preferred temporal modeling (e.g., RNN, GRU, LSTM) to obtain feature vectors per data point.

2. Perform a 4-class emotion classification using your preferred classifier with the obtained feature vectors. Select the parameters using Grid Search (search over a range for hyper-parameters). Perform any additional steps you see fit to obtain the best results.

3. Report your classification results on individual modalities (vision, speech, and text) using **F1-micro** metric on a 10-fold subject-independent cross validation.

4. How do you handle the problem of class imbalance? Plot the confusion matrix for the 4 classes.

5. Use both early fusion (concatenate features from different modalities) and late fusion (majority vote over the outputs of the unimodal models) to obtain multimodal classification results. Report and compare the results for both fusion techniques.

6. Provide an interpretation on your results from the performed unimodal and multimodal classification tasks. Which one is performing best and why?

Note: You are only allowed to use the features and labels provided by us with this assignment. Please refrain from using the original data; assignments submitted with any other labels or data will not be graded.

4 Submission deadline

Please submit your report and source codes named as **lastname_firstname** on Blackboard until **03.23.2021 at 23:59**. Late submissions will incur a 10% penalty per day. If you have any questions, please email it to Yufeng Yin (yufengy@usc.edu).

References

[1] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**(4), 335 (2008)

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[3] Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)

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