Homework 3: Multimodal Fusion for MED

11-775 Large-Scale Multimedia Analysis (Fall 2022)

Due on: Wednesday October 26, 2022 11:59 PM

1 Overview

Homework 3 will be the final homework, in which you complete your own MED system that utilizes both video and audio inputs. In homework 1 and 2, you have learned how to process/extract audio and video features. In this homework, we ask you to fuse them together with the knowledge about multimodal fusion you learned in the class. Please **START EARLY**! It takes time to integrate different modules and process the large-scale dataset. Homework 3 is due on Wednesday October 26, 2022 11:59 PM.

2 MED Pipeline Overview

To minimize your effort, please reuse the MED pipeline developed in homework 1 and 2. In this final assignment, the only thing you need to do is to fill in the last missing piece of MED: **Multi-modal fusion**. You've learned many fusion methods in class (e.g., early fusion, late fusion, double fusion, middle fusion ...) to combine different features and improve the model performance. After you finishing homework 1 and 2, your pipeline should be able to utilize audio features (MFCC, SoundNet ... etc) and video features (SURF, ResNet, AlexNet, ... etc) with various feature encoding techniques (Bag-of-Words, pooling, VLAD...) for classification. Please develop **at least THREE fusion schemes** to fuse your selected features. By definition of "fusion", for each fusion scheme, you should fuse **at least TWO features** in your MED pipeline and compare the fusion results with those of individual features. As features from different modalities are likely to be complementary to each other, it's preferable to use both audio and visual domains for MED. Another challenge in this homework is to deal with large-scale data with limited computation power. Please make sure your pipeline would fully utilize the computation power available and is efficient enough to process all videos. In the following we briefly review three feature fusion techniques (early/late/double fusion) from related research papers. Note that you are not limited to use the fusion methods listed above, and **we encourage you to explore other State-of-the-Art fusion methods** such as [1, 2]. You will use the same dataset in HW1 (HW2P1) with both audios and videos as inputs.

2.1 Early Fusion

As shown in Figure 1, for a practical MED system which relies on early fusion for decision, it firstly extracts individual features separately. The extracted features are then combined into a single vector representation for each video. A commonly used feature combination strategy is concatenating vectors from different feature extractors into a long vector. Please think about whether you need to normalize the feature before or after early fusion. After combination of individual feature vectors for a multimodal representation, the supervised classifiers (such as SVM or MLP you used in HW1 and HW2) are employed for classification.

2.2 Late Fusion

As illustrated in Figure 2, a MED system which uses late fusion for classification also starts with extracting different feature descriptors. In contrast to early fusion, where features are then combined into a multimodal representation, approaches for late fusion firstly learn separate supervised classifiers directly from unimodal features. In the testing phase, the prediction scores from different models are

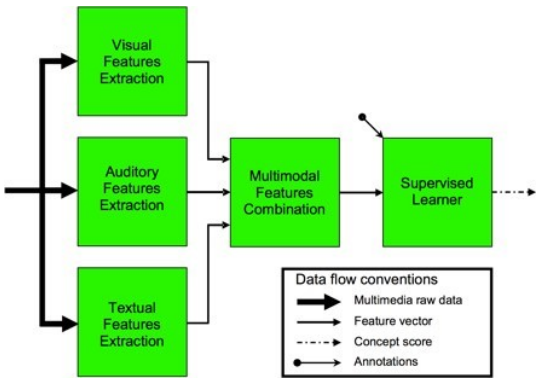

Figure 1: General scheme for early fusion. Output from different feature extractors are fused before classifier learning.

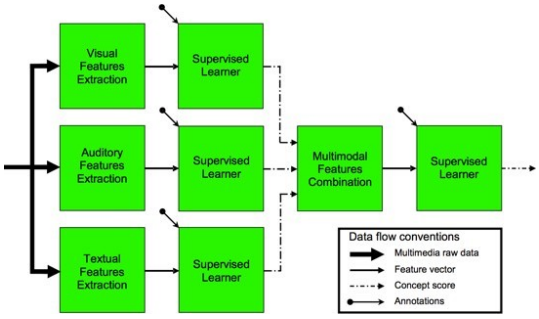

Figure 2: General scheme for late fusion. Outputs of different feature extractors are used to learn separate scores for a concept. After fusion, a final score is learned for the concept.

then combined to yield a final score. In general, late fusion schemes combine learned unimodal scores into a multimodal representation. Compared to early fusion, late fusion focuses on the individual strength of modalities.

2.3 Double Fusion

In double fusion shown in Figure 3, we first perform early fusion to generate different combinations of features from subsets on the single features pool. After that, we train classifiers on each feature or feature combination and carry out late fusion on the output of these classifiers. For example, we may want to extract three features (ResNet-34, ResNet-101 and SoundNet). After that, pairwise early fusion (ResNet-101+ResNet-34, ResNet-101+SoundNet) are carried out in these three features based on their kernel matrices. In the training step, five classifiers are trained based on five features and their combinations (ResNet-34, ResNet-101, SoundNet, ResNet-101+ResNet-34, ResNet-101+SoundNet). For each video, there are five output scores indicating how likely it is that this video belongs to the event. Lastly, late fusion is used to fuse five output score vectors into one score vector, on which the final interpretation can be executed.

3 Submission

3.1 Canvas

Please compress your submission into a zip file named as `andrewid_hw3.zip` and submit it through Canvas. The contents of your zip file should be organized as the following:

1. **report.pdf**: Your PDF report with your pipeline design, findings, results, and analysis.
2. **code.zip**: A .zip file with your code only. Please be sure to add a `README.md` with the instructions on how to run your code to reproduce the results.

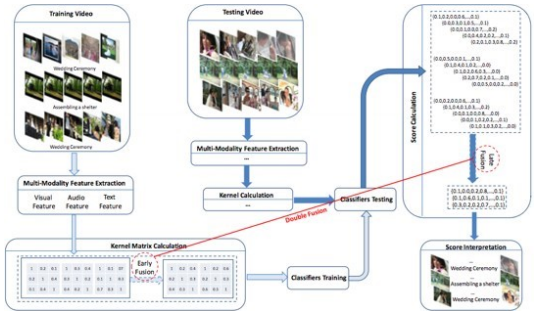

Figure 3: General scheme for double fusion.

3. **best.csv**, The best classification results for each testing video,

For the CSV files, the following format should be used:

```
Id,Category
LTExODM2MzcOODQyOTc1ODE4NDM=,7
LTUwNDU3NzgyNjE2MzkOOTU1NjQ=,1
ODU3OTEOMDU5NzM5NDI2MDQ2,0
...
```

We will grade this homework by the completeness of experiments and analysis. In the report, please:

1. Describe the fusion schemes you choose to implement and the features to fuse. Does fusion improve the result?
2. Report the confusion matrix for multi-class classification in your validation set.
3. Report the time your MED system takes for feature extraction and classification on the testing set.

3.2 Kaggle

Please submit your **best.csv** file to this Kaggle competition and report their Public test set accuracy in your report. You can submit up to 5 times/day. The final performance on the whole test set will be revealed on October 26.

References

[1] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33:4835–4845, 2020.