HW3 Report

11775-ISR: Large Scale Multimedia Analysis ISR Section

10/25/2022
Gyuseok Lee

1. Introduction

In this homework, I have to implement the fusion model, which connects each other modality. Specifically, through HW1 and HW2, I can get audio and 3D image feature, respectively. Therefore, by using these features together, I try to improve the performance of model. Specially, I can learn the fusion methods through this homework like early, late, double fusion. To compare these fusion methods, I conduct the experiments based on each method. As a result, early fusion shows better performance than any other fusion method. In the rest of parts, I would like to explain dataset, fusion model, experiment, and conclusion.

2. Dataset

Through HW1 and HW2, I can get audio and 3D image feature per each video. So, I use these features again. Concretely, the audio feature is extracted based on PaSST and the 3D image feature can be extracted through *R2 Plus1D18* 3D CNN model. The dimension of audio and image feature are 1024 and 512, respectively.

3. Model

In this homework, I have to develop at least three fusion schemes, so I implement the methods: early, late, and double. The detailed structure is shown in the figure below. Note that blue and orange rectangle means feature and fully connected layer, respectively.

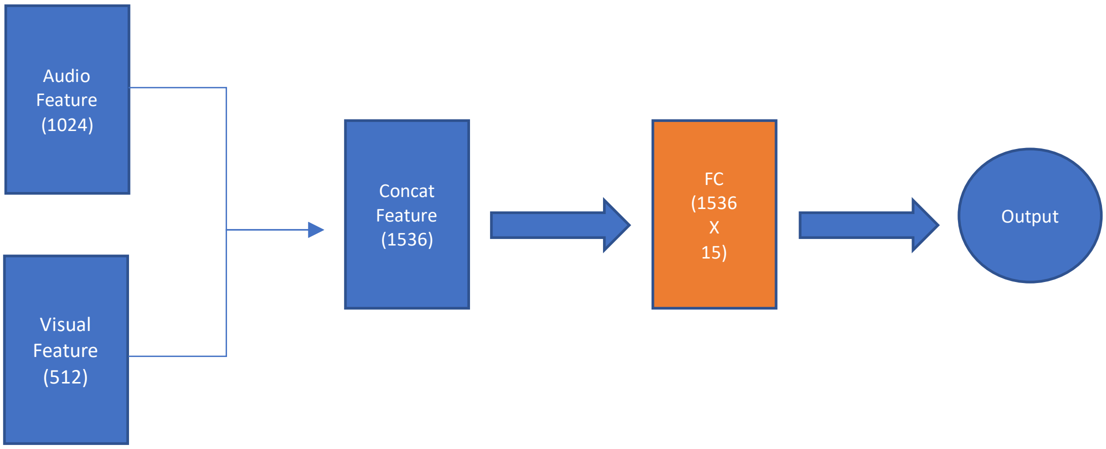


Figure1. Early fusion

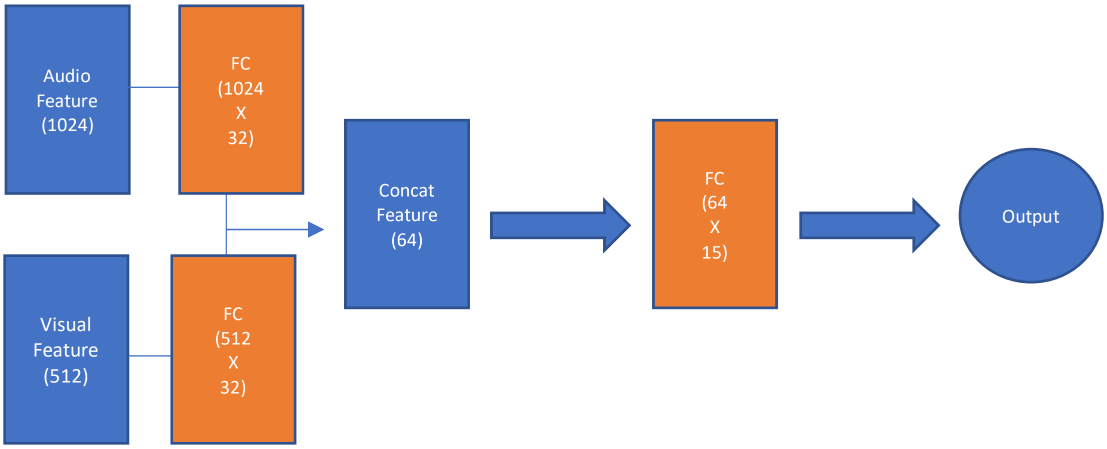


Figure2. Late fusion

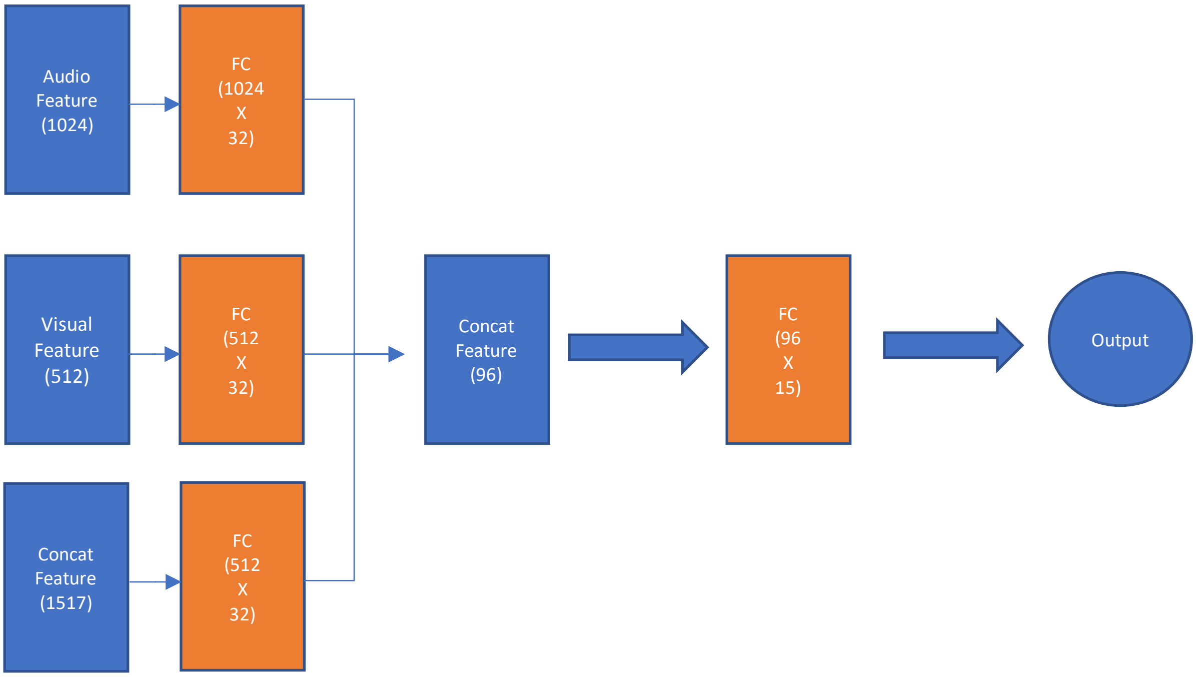
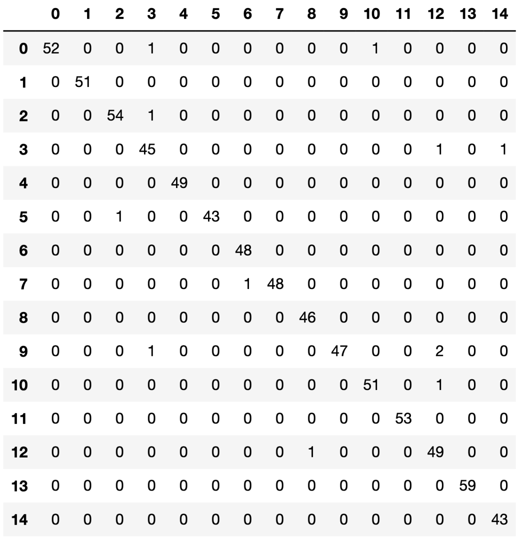


Figure3. Double fusion

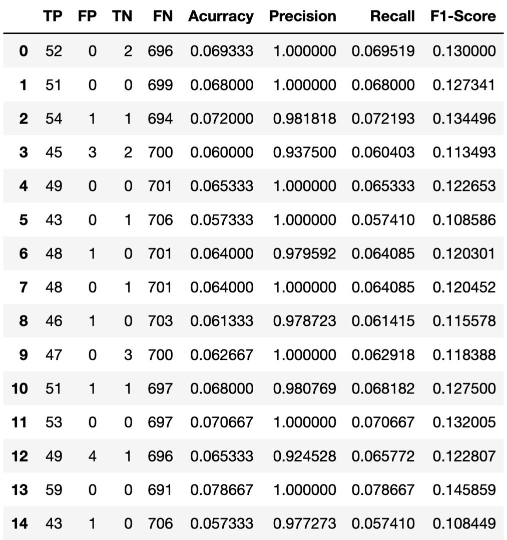
4. Experiment

I conduct the experiments based on each fusion method. The bottom of the line is that early fusion shows the best performance 0.992 (validation set), and other fusion models (late and double) show 0.674 and 0.686, respectively. I analyze these results and I concluded that its feature is already well trained, we don't have to train deeper and wider. Considering above figure, late and double fusion have deeper and wider network, but early fusion has only one fully connected layer. Moreover, I find that the model has the tendency of overfitting, so I have to deal with this problem by reducing the model size, adding the regularization term and so on. In this respective, I think that early fusion is the simplest method so it can overcome the problem of overfitting better than any other method. Furthermore, I find that early fusion model can improve the result (Kaggle submission), which gets 0.95 score, but using the single 3D image feature gets 0.92 score. Note that It took 58 minutes and 29 minutes to extract audio and 3D data, respectively, and less than 10 minutes to train the model based on early fusion. Due to the early stopping, training finished earlier than the total epochs (i.e., epochs = 100). Because model learning has already learned sufficiently for train dataset, it can no longer afford to learn.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	52	0	0	1	0	0	0	0	0	0	1	0	0	0	0
1	0	51	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	54	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	45	0	0	0	0	0	0	0	0	1	0	1
4	0	0	0	0	49	0	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	43	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	48	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	46	0	0	0	0	0	0
9	0	0	0	1	0	0	0	0	0	47	0	0	2	0	0
10	0	0	0	0	0	0	0	0	0	0	51	0	1	0	0
11	0	0	0	0	0	0	0	0	0	0	0	53	0	0	0
12	0	0	0	0	0	0	0	0	1	0	0	0	49	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	59	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43

Table1. Confusion Matrix for Early fusion



	TP	FP	TN	FN	Acurracy	Precision	Recall	F1-Score
0	52	0	2	696	0.069333	1.000000	0.069519	0.130000
1	51	0	0	699	0.068000	1.000000	0.068000	0.127341
2	54	1	1	694	0.072000	0.981818	0.072193	0.134496
3	45	3	2	700	0.060000	0.937500	0.060403	0.113493
4	49	0	0	701	0.065333	1.000000	0.065333	0.122653
5	43	0	1	706	0.057333	1.000000	0.057410	0.108586
6	48	1	0	701	0.064000	0.979592	0.064085	0.120301
7	48	0	1	701	0.064000	1.000000	0.064085	0.120452
8	46	1	0	703	0.061333	0.978723	0.061415	0.115578
9	47	0	3	700	0.062667	1.000000	0.062918	0.118388
10	51	1	1	697	0.068000	0.980769	0.068182	0.127500
11	53	0	0	697	0.070667	1.000000	0.070667	0.132005
12	49	4	1	696	0.065333	0.924528	0.065772	0.122807
13	59	0	0	691	0.078667	1.000000	0.078667	0.145859
14	43	1	0	706	0.057333	0.977273	0.057410	0.108449

Table2. Descriptive Statics Table for Early fusion

5. Conclusion

In this homework, I can learn what the fusion model is, connect the different modalities, and improve the classification result. The learning based on multi-modal can have the model get a fine-grained feature by using characteristics of each modality. Plus, I found that early fusion shows the better performance than any other method, so how to fuse the different modality would be very important.