**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.

1. **Dataset**
2. **Model**

Input

channel

x

64

64

X

15

(# of class)

Figure1. Model Architecture

1. **Experiment**

|  |
| --- |
| 테이블이(가) 표시된 사진  자동 생성된 설명  Table1. Confusion Matrix for SIFT  테이블이(가) 표시된 사진  자동 생성된 설명  Table2. Descriptive Statics Table for SIFT |

1. CNN

As I already mentioned, I used the *convnext(base)* which was pretrained with ImageNet dataset. So, the embedding size is 1024. At the bottom of the line is that the top-1 accuracy is 0.94 for validation set and the training time for convergence is 144 seconds. The hardest class is 12 and the easiest class is 11. I attached the confusion matrix and descriptive statistics table below.

|  |
| --- |
| Table3. Confusion Matrix for CNN    Table4. Descriptive Statics Table for CNN |

1. CNN3D

I used the pretrained *R2Plus1D18* for CNN3d which has the 512 dimensions for embedding. The top-1 accuracy is 0.982 and the training time for convergence is 174.6. The hardest class is 8 and the easiest class is 0,10,11,14.

I attached the confusion matrix and descriptive statistics table below.

|  |
| --- |
| Table5. Confusion Matrix for CNN3D    **Table6. Descriptive Statics Table for CNN3D** |

1. **Conclusion**