CMU 11-775 Fall 2023 HW2
Jooeon Kang
Report on Video Classification Pipeline

Introduction

The task at hand is to perform a video classification using visual features. This report elaborates on the designed pipeline, findings, results, and analysis of the experiment conducted as a part of CMU 11-775 Fall 2023 Homework 2.

Pipeline Design

The pipeline designed for this video classification task encompasses various stages, including feature extraction, feature transformation, model training, and evaluation. Below is a detailed description of each stage:

1. Feature Extraction

- **SIFT Features**: Scale-Invariant Feature Transform (SIFT) features are extracted from the video frames.
- **CNN Features**: Convolutional Neural Network (CNN) features are extracted for a more advanced representation.
- **3D CNN Features**: 3D CNN features are extracted to capture the temporal relationships between frames.

2. Feature Transformation

- **K-Means Clustering on SIFT Features**: SIFT features are used to train a K-Means model with 128 clusters, resulting in a Bag-of-Words (BoW) representation.
- **Bag-of-Words Representation**: The trained K-Means model is used to transform the SIFT features into a BoW representation.

3. Model Training

- **MLP Classifier**: A Multi-Layer Perceptron (MLP) classifier is trained using the different feature representations (SIFT BoW, CNN, and 3D CNN features).

4. Evaluation

- The MLP classifier's performance is evaluated to understand the effectiveness of each feature representation.
- Submitted the CSV file to Kaggle and see the scores.

Findings

- The README provides a comprehensive guide on how to execute each part of the pipeline, including the installation of dependencies, dataset preparation, and execution of scripts for feature extraction, transformation, and model training.
- Performance : 3D_CNN(Video) > 2D_CNN(Image) > SIFT

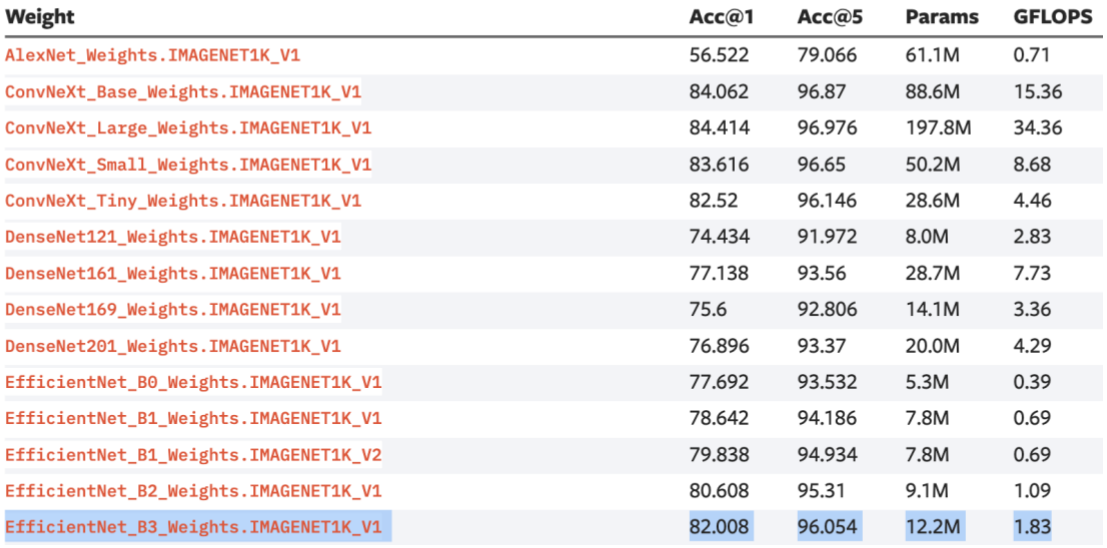
Results and Analysis

1. SIFT Features

- SIFT features are robust to scale and rotation, but they might not capture high-level semantic information in the videos.
- Final Kaggle Score : 0.28571

2. CNN Features

- CNN features are expected to provide a more comprehensive representation of the video content, capturing complex patterns and structures.
- I used EfficientNet_B3_Weights.IMAGENET1K_V1, which has relatively small parameters and good accuracy

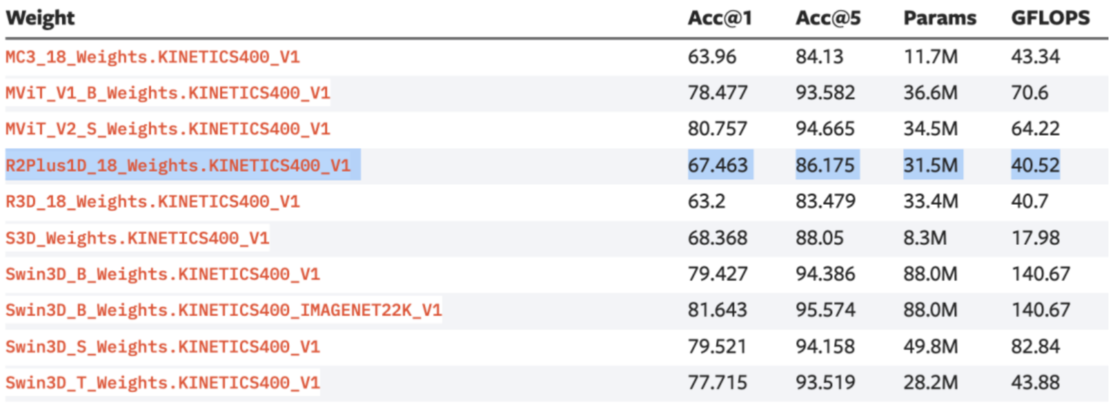


Weight	Acc@1	Acc@5	Params	GFLOPS
AlexNet_Weights.IMAGENET1K_V1	56.522	79.066	61.1M	0.71
ConvNeXt_Base_Weights.IMAGENET1K_V1	84.062	96.87	88.6M	15.36
ConvNeXt_Large_Weights.IMAGENET1K_V1	84.414	96.976	197.8M	34.36
ConvNeXt_Small_Weights.IMAGENET1K_V1	83.616	96.65	50.2M	8.68
ConvNeXt_Tiny_Weights.IMAGENET1K_V1	82.52	96.146	28.6M	4.46
DenseNet121_Weights.IMAGENET1K_V1	74.434	91.972	8.0M	2.83
DenseNet161_Weights.IMAGENET1K_V1	77.138	93.56	28.7M	7.73
DenseNet169_Weights.IMAGENET1K_V1	75.6	92.806	14.1M	3.36
DenseNet201_Weights.IMAGENET1K_V1	76.896	93.37	20.0M	4.29
EfficientNet_B0_Weights.IMAGENET1K_V1	77.692	93.532	5.3M	0.39
EfficientNet_B1_Weights.IMAGENET1K_V1	78.642	94.186	7.8M	0.69
EfficientNet_B1_Weights.IMAGENET1K_V2	79.838	94.934	7.8M	0.69
EfficientNet_B2_Weights.IMAGENET1K_V1	80.608	95.31	9.1M	1.09
EfficientNet_B3_Weights.IMAGENET1K_V1	82.008	96.054	12.2M	1.83

- Final Kaggle Score : 0.90977

3. 3D CNN Features

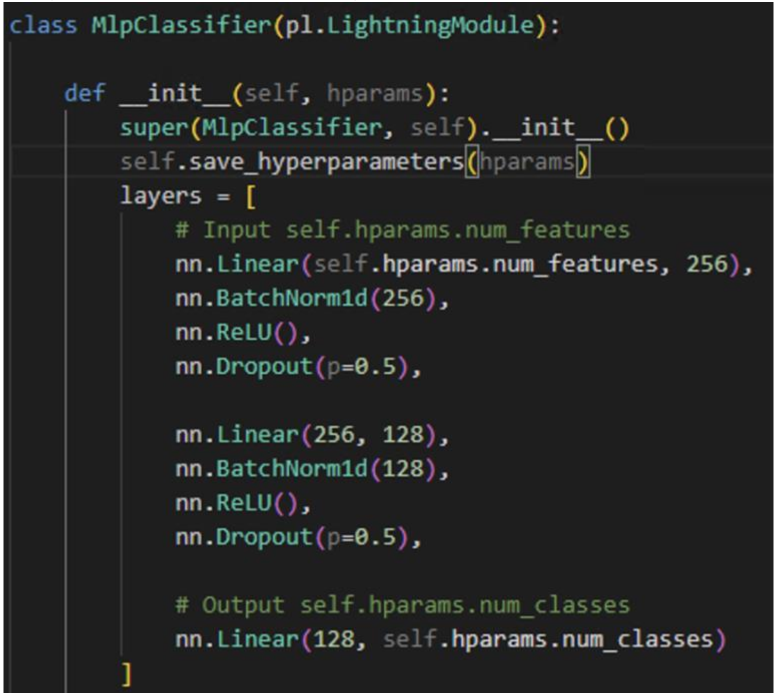
- 3D CNN features should capture both spatial and temporal relationships, potentially leading to better performance in video classification.
- I tried to use Swin3D or MViT, but an error occurred and I could not use it.
- I used R2Plus1D_18_Weights.KINETICS400_V1.
- Final Kaggle Score : 0.95989s



Weight	Acc@1	Acc@5	Params	GFLOPS
MC3_18_Weights.KINETICS400_V1	63.96	84.13	11.7M	43.34
MViT_V1_B_Weights.KINETICS400_V1	78.477	93.582	36.6M	70.6
MViT_V2_S_Weights.KINETICS400_V1	80.757	94.665	34.5M	64.22
R2Plus1D_18_Weights.KINETICS400_V1	67.463	86.175	31.5M	40.52
R3D_18_Weights.KINETICS400_V1	63.2	83.479	33.4M	40.7
S3D_Weights.KINETICS400_V1	68.368	88.05	8.3M	17.98
Swin3D_B_Weights.KINETICS400_V1	79.427	94.386	88.0M	140.67
Swin3D_B_Weights.KINETICS400_IMAGENET22K_V1	81.643	95.574	88.0M	140.67
Swin3D_S_Weights.KINETICS400_V1	79.521	94.158	49.8M	82.84
Swin3D_T_Weights.KINETICS400_V1	77.715	93.519	28.2M	43.88

4. MLP Classifier

- The performance of the MLP classifier would depend on the feature representation used. CNN and 3D CNN features are expected to outperform SIFT BoW features due to their ability to capture more complex and high-level information.
- Simple Classifier that I implemented:



```
class MlpClassifier(pl.LightningModule):

    def __init__(self, hparams):
        super(MlpClassifier, self).__init__()
        self.save_hyperparameters(hparams)
        layers = [
            # Input self.hparams.num_features
            nn.Linear(self.hparams.num_features, 256),
            nn.BatchNorm1d(256),
            nn.ReLU(),
            nn.Dropout(p=0.5),

            nn.Linear(256, 128),
            nn.BatchNorm1d(128),
            nn.ReLU(),
            nn.Dropout(p=0.5),

            # Output self.hparams.num_classes
            nn.Linear(128, self.hparams.num_classes)
        ]
```

Recommendations for Future Work

- **Evaluation Metrics**: Implement and report various evaluation metrics such as accuracy, precision, recall, and F1-score to have a holistic view of the model’s performance.
- **Hyperparameter Tuning**: Perform hyperparameter tuning for the MLP classifier to optimize its performance.
- **Feature Fusion**: Explore the combination of different feature representations to enhance the model's predictive power.
- **Model Complexity**: Experiment with different MLP architectures and increase complexity if necessary to capture more intricate patterns in the data.

Conclusion

The pipeline designed for video classification in this task is comprehensive, covering feature extraction, transformation, and model training. As it was expected, the CNN-based features would yield superior performance compared to SIFT features due to their ability to capture more complex and high-level information, especially 3D CNN outperformed. Future work should focus on handling ***Swin3D*** & ***MViT*** model . Also focus on evaluating the models comprehensively, tuning hyperparameters, and exploring feature fusion to further enhance performance.