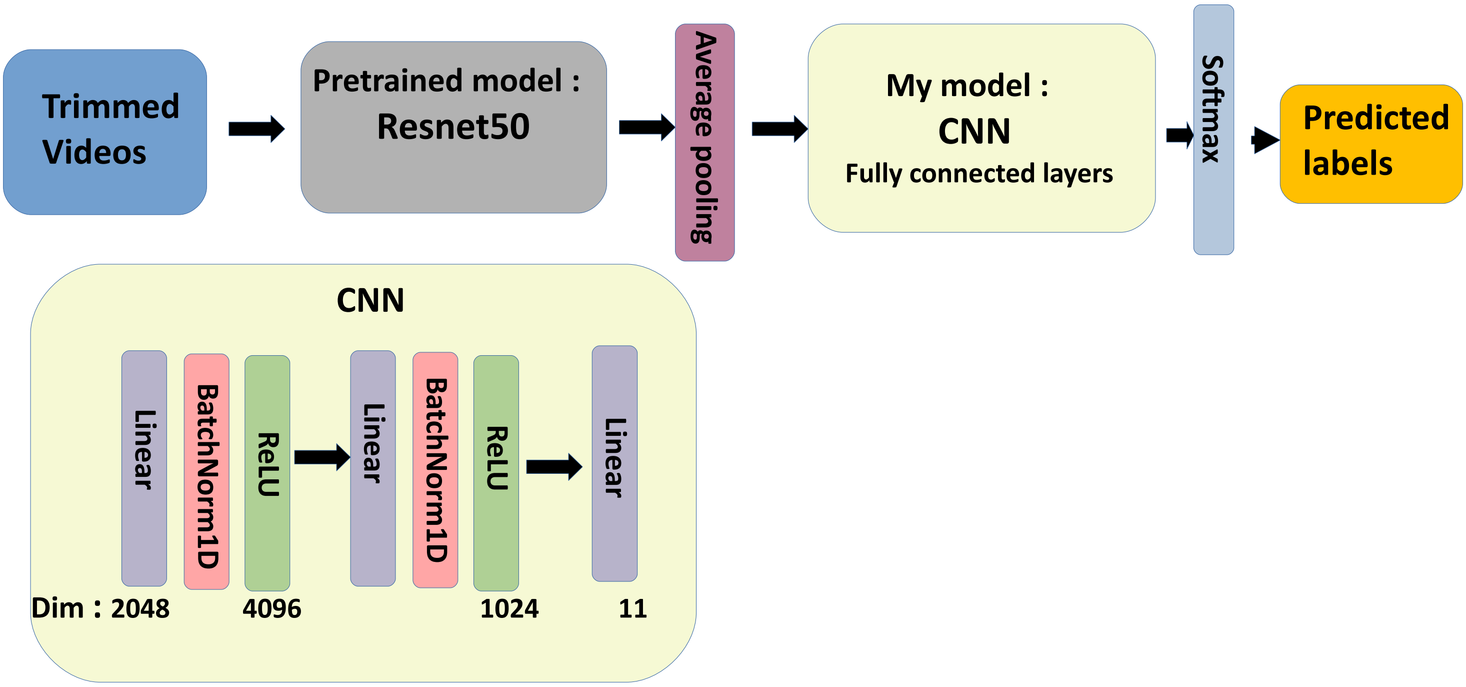
Clautour Amélie A08527201 No collaborator

DLCV Homework #4

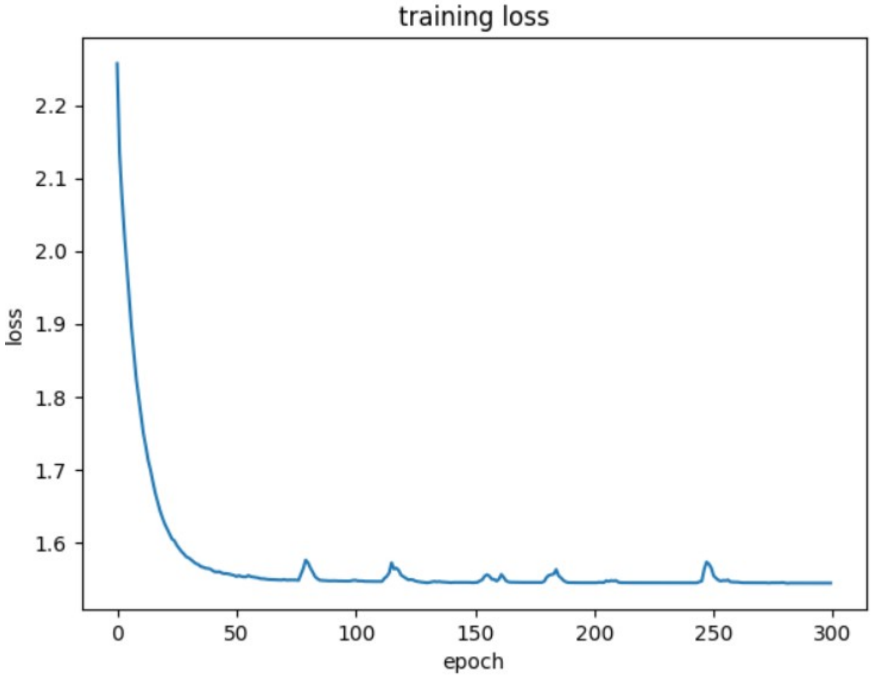
Problem 1 : Trimmed action recognition w/o RNN

Describe your strategies of extracting CNN-based video features, training the model and other implementation details (which pretrained model) and plot your learning curve.



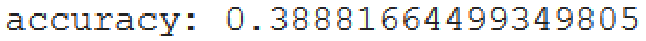
Implementation details :

batch size = 64 epochs = 300
learning rate = 0.000025 feature size = 2048
loss function = CrossEntropyLoss normalized images with (0.485, 0.456, 0.406),(0.229, 0.224, 0.225)



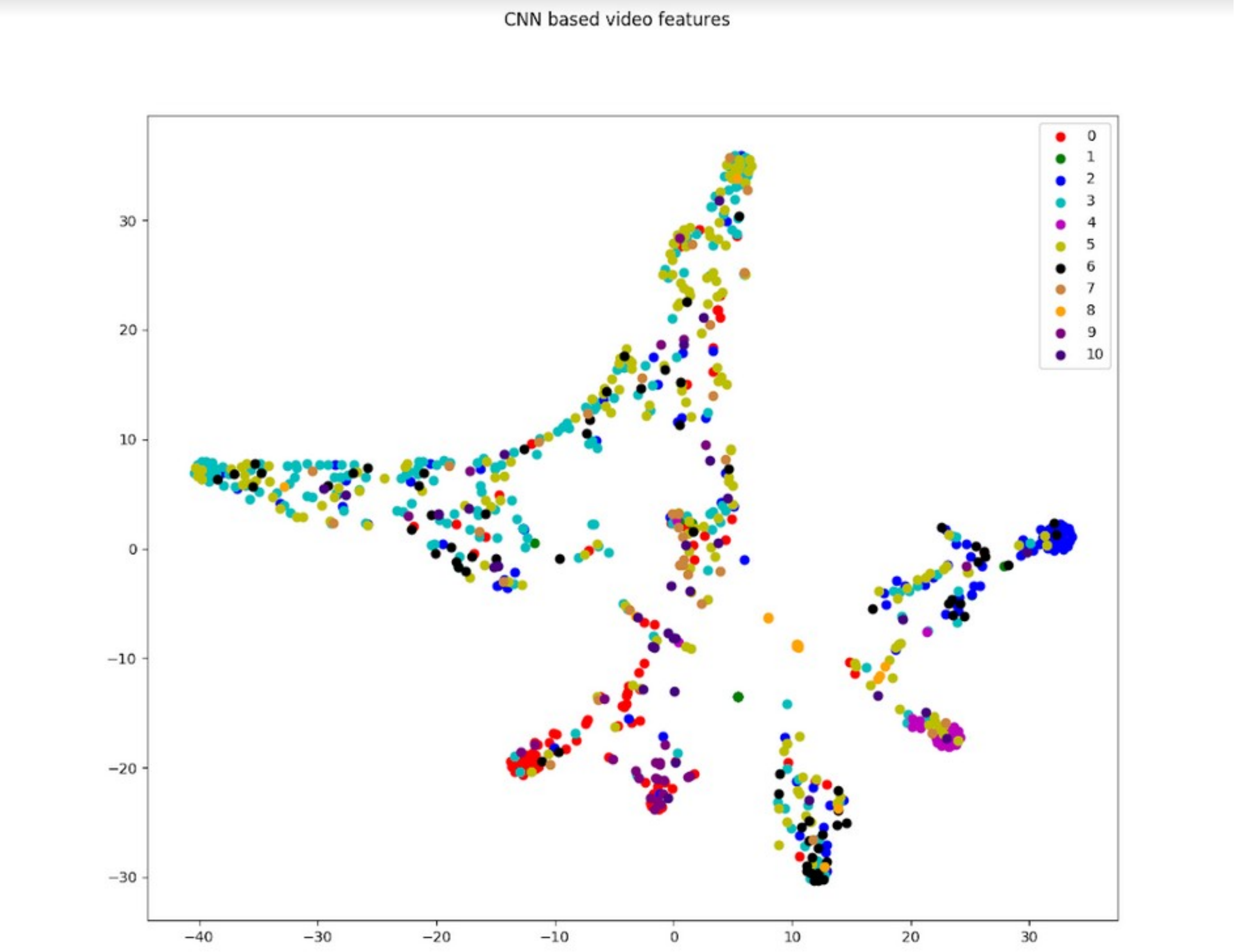
Report your video recognition performance (valid) using CNN-based video features and make your code reproduce this result.

For my CNN model the accuracy is:



```
accuracy: 0.38881664499349805
```

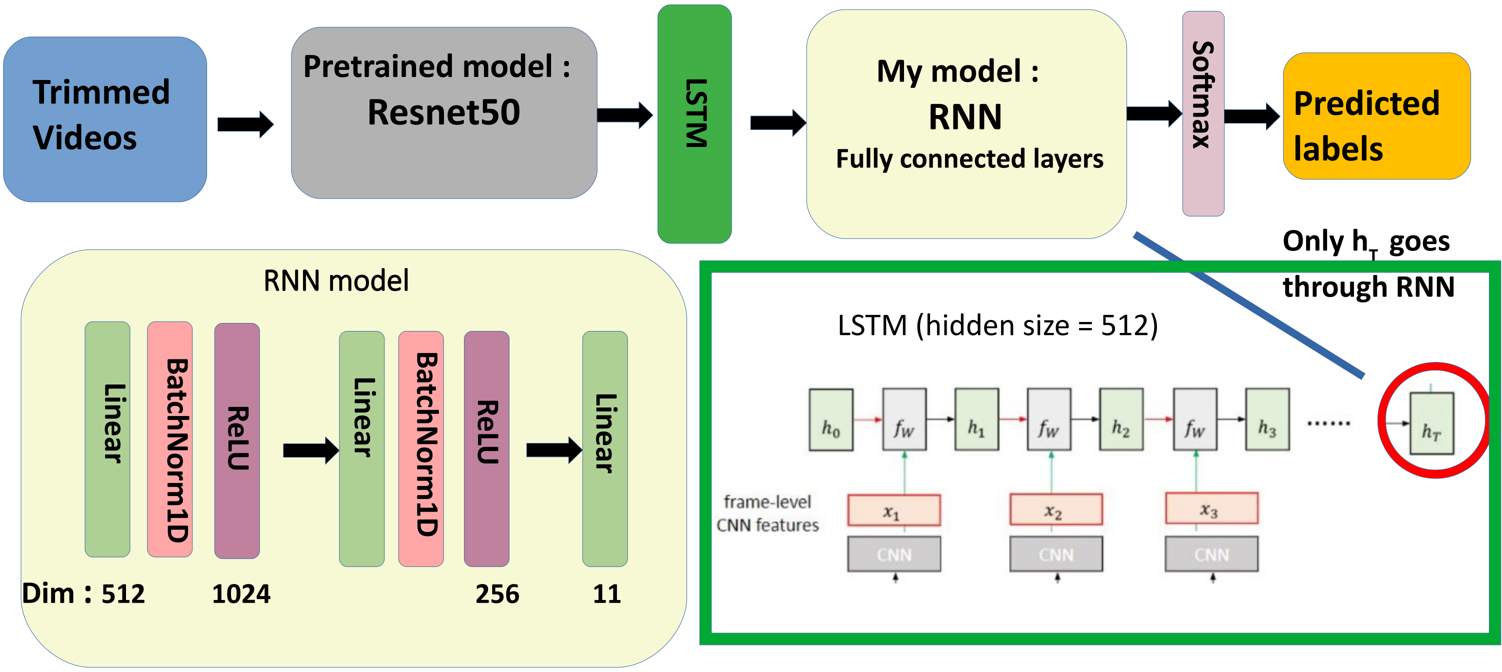
Visualize CNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels



CNN based video features

Problem 2: Trimmed action recognition w/ RNN

Describe your RNN models and implementation details for action recognition and plot the learning curve of your model (The loss curve of training set is needed, others are optional).



Trimmed Videos

Pretrained model : Resnet50

LSTM

My model : RNN Fully connected layers

Softmax

Predicted labels

Only h_T goes through RNN

RNN model

Linear BatchNorm1D ReLU → Linear BatchNorm1D ReLU → Linear

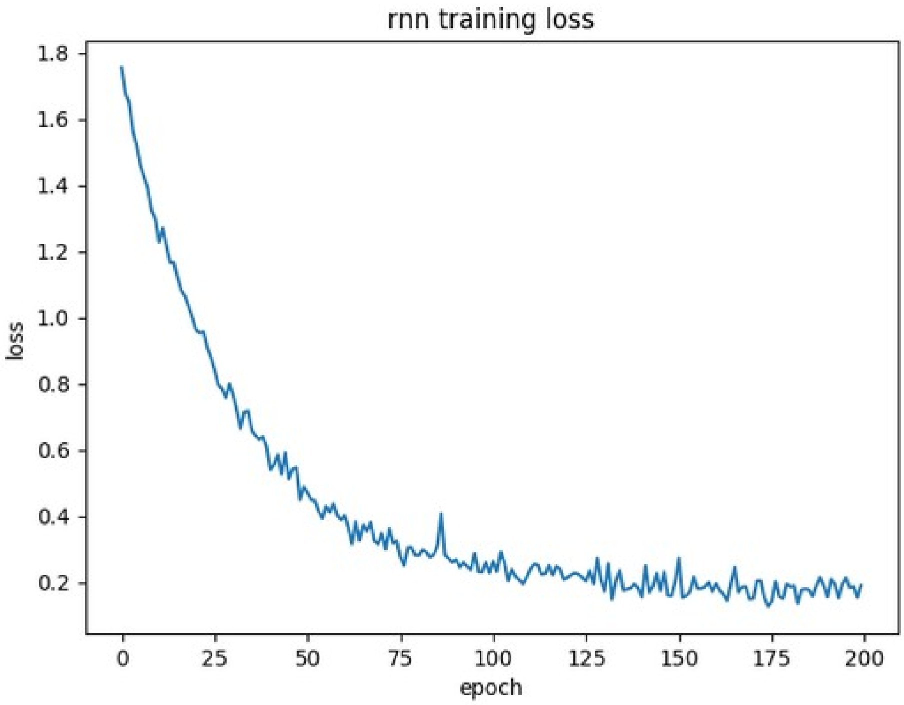
Dim : 512 1024 256 11

LSTM (hidden size = 512)

Implementation details :

batch size = 64
learning rate = 0.0001
loss function = CrossEntropyLoss

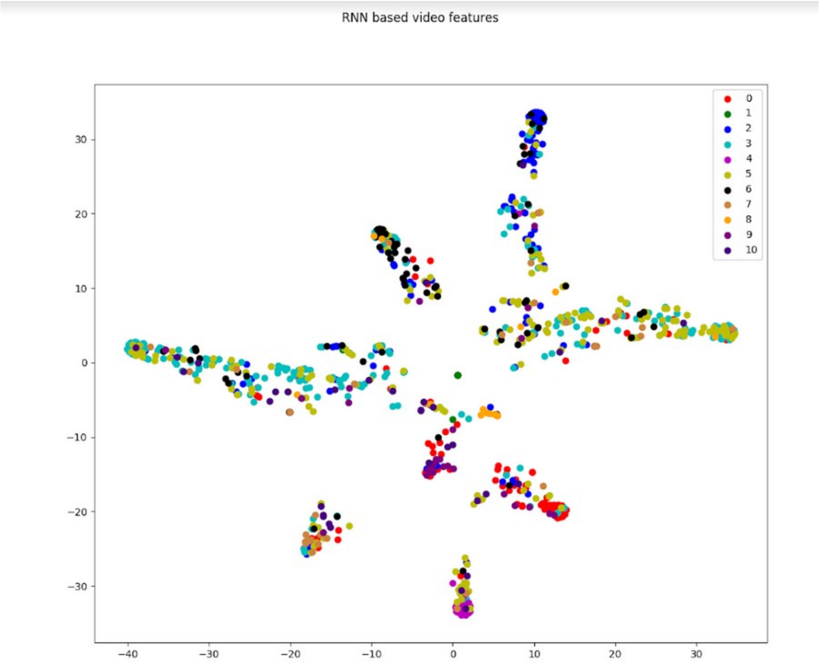
epochs = 200
feature size = 2048
normalized images with (0.485, 0.456, 0.406),(0.229, 0.224, 0.225)



accuracy on validation set :

```
('accuracy:', 0.46293888166449937)
```

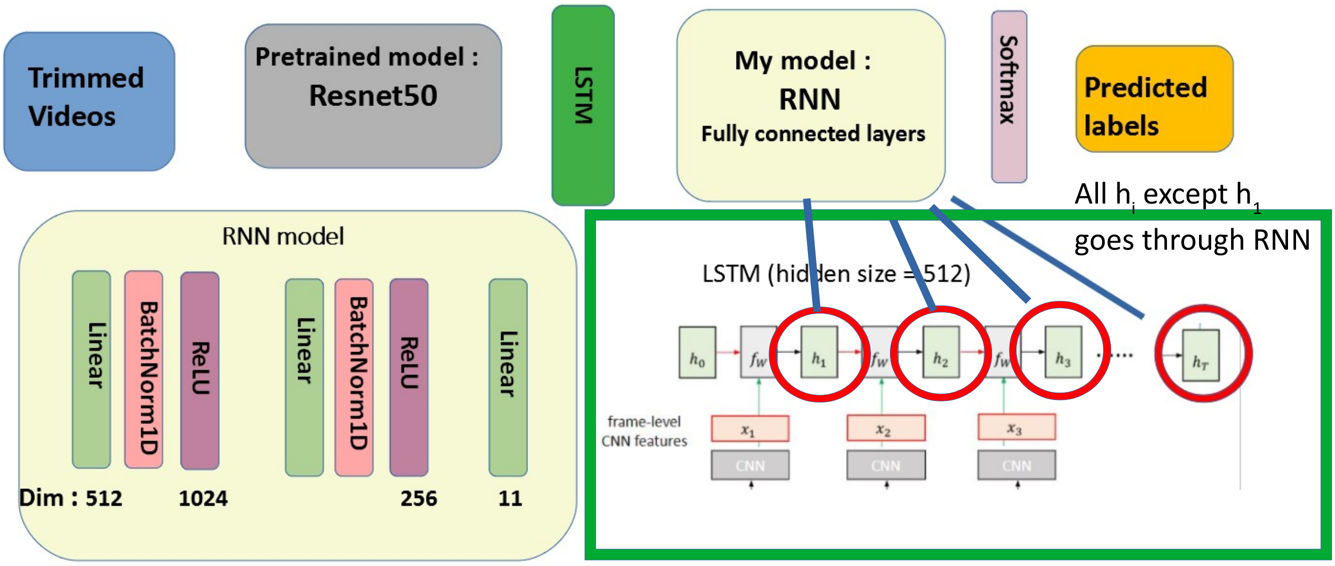
Visualize RNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels. Do you see any improvement for action recognition compared to CNN-based video features? Why? Please explain your observation



We can see a slight improvement for action recognition compared to CNN based features. I think that's because the RNN model learns from sequenced information and not only from single information. For example the "inspect" action is often followed by the "open" action which can help the model predict an "open" action if there was before an "inspect" action. There is an effect of memory transfer between frames.

Problem 3: Temporal action segmentation

Describe any extension of your RNN models, training tricks, and postprocessing techniques you used for temporal action segmentation.



Implementation details:

batch size = 64

learning rate = 0.0001

loss function = CrossEntropyLoss

epochs = 100

feature size = 2048

normalized images with (0.485, 0.456, 0.406),(0.229, 0.224, 0.225)

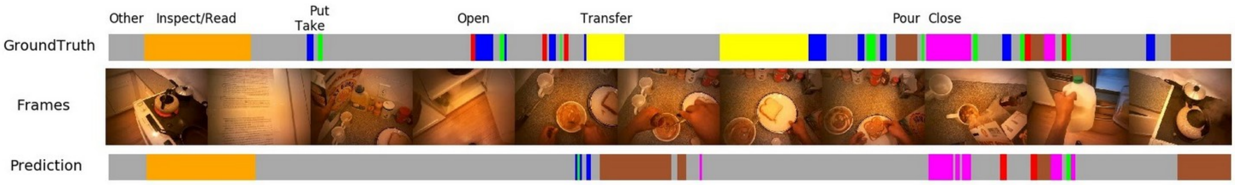
Report validation accuracy in your report and make your code reproduce this result

The validation accuracy is:

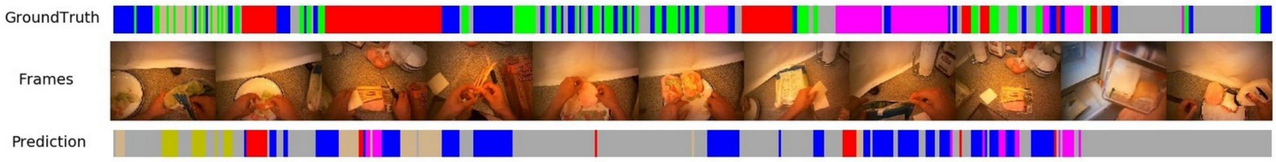
OP01-R07-Pizza: 0,539457709
OP01-R02-TurkeySandwich: 0,406126
OP06-R03-BaconAndEggs : 0.6402321083172147
OP05-R04-ContinentalBreakfast : 0,553797468
OP04-R04-ContinentalBreakfast: 0,65345622
OP03-R04-ContinentalBreakfast : 0,55568053
OP01-R04-ContinentalBreakfast: 0,590631

Choose one video from the 7 validation videos to visualize the best prediction result in comparison with the ground-truth scores in your report. Please make your figure clear and explain your visualization results (You need to plot at least 500 continuous frames).

For ***OP04-R04-ContinentalBreakfast (best case)*** :



For ***OP01-R02-TurkeySandwich (worst case)*** :



For every videos, when the actions change quickly we can see that it is where the predicted labels of the model are the less accurate which is not surprising since the model studies sequences and if those sequences change too fast the model has trouble identifying a pattern needed to predict accurate outputs. Moreover some information is lost during the sampling process. The best predicted label is "read/inspect" and the "transfer" action is not identified at all by the model, probable because the "read" action is associated with characteristic images/information like numbers and words while the "transfer" action is more abstract.