Lips reading to visual speech recognition.

Universidad Industrial de Santander

Juan Felipe Chacón Lopéz
Mario Hernan Vallejo Huertas

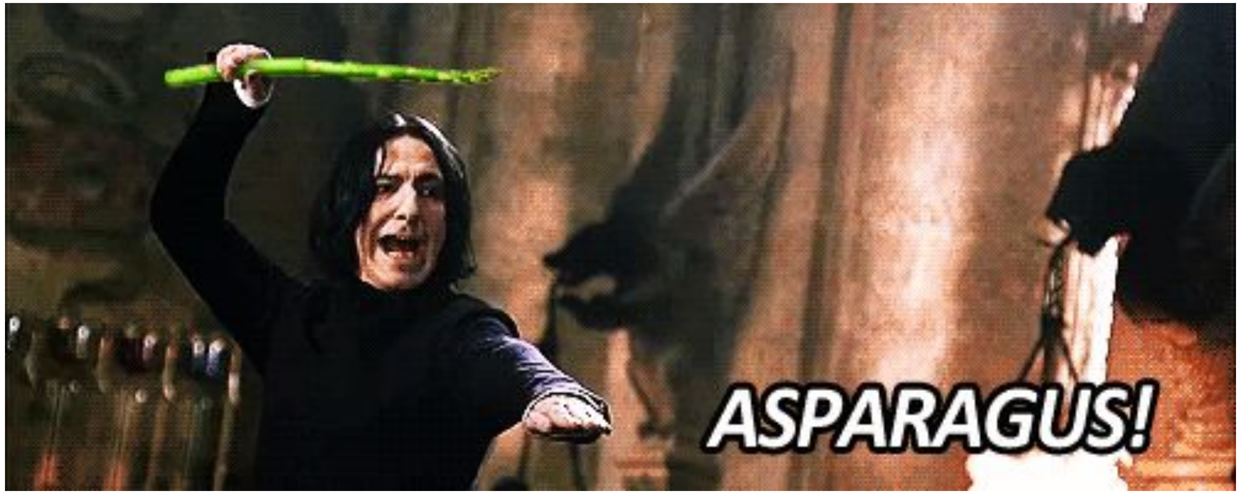
Inteligencia Artificial II
First delivery

#LaUISqueQueremos

72 años Universidad Industrial de Santander Patrimonio educativo y cultural

What’s lip reading?

Lip reading is a speech understanding technique by visually interpreting the movements of the lips, face and tongue when **normal sound is not available.**



Why lip reading?

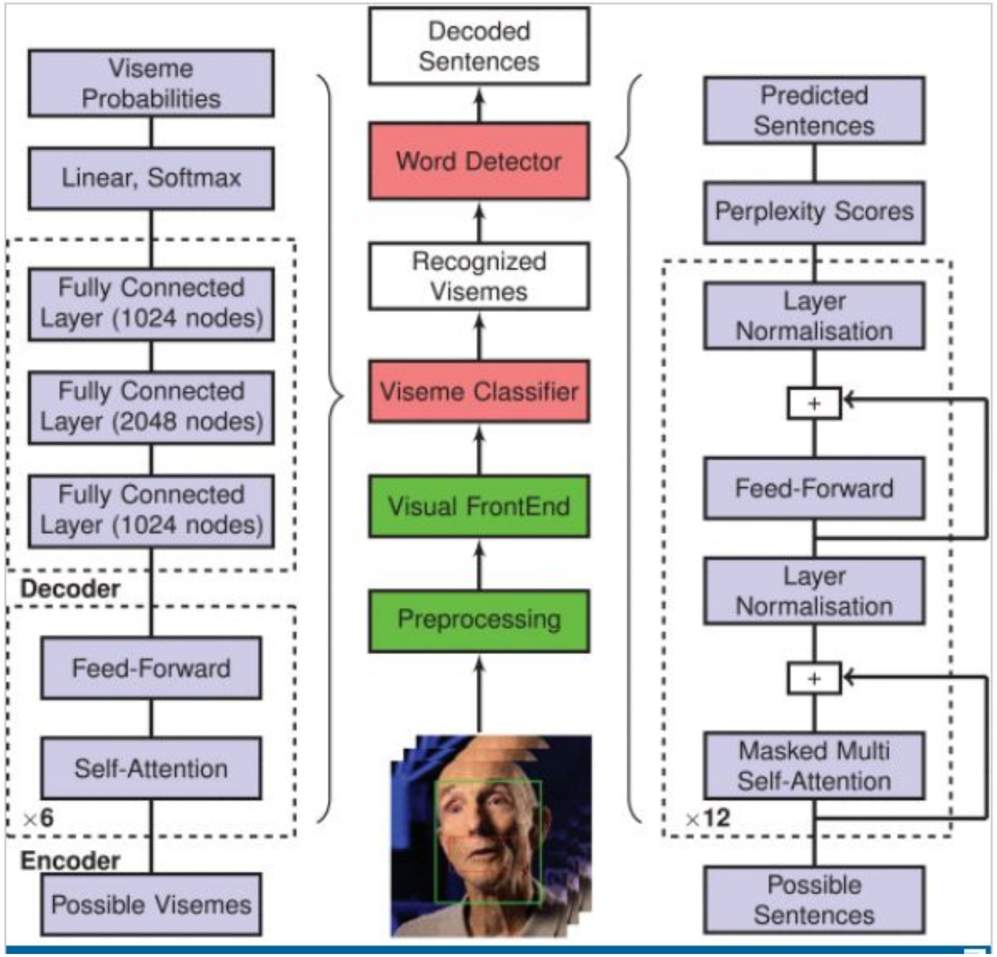
Lip reading is mainly used by deaf and hard of hearing people, people with normal hearing generally process visual information from the mouth that moves at a subconscious level.

Related work

Universidad Industrial de Santander

- **Survey on automatic lip-reading in the era of deep learning**
- **Lip-Reading Driven Deep Learning Approach for Speech Enhancement**
- Lip Reading Sentences Using Deep Learning With Only Visual Cues

Lip Reading Sentences Using Deep Learning With Only Visual Cues



Universidad Industrial de Santander

Somos **el mejor** escenario de creación e innovación.

www.uis.edu.co

Related works approaches.

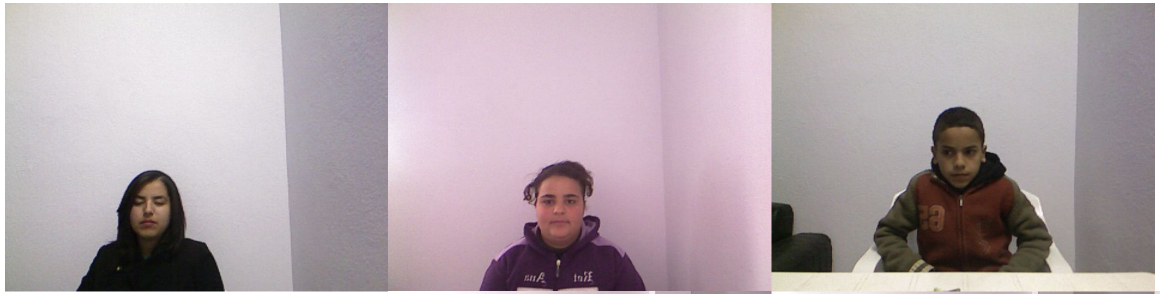
The most recent approaches to automated lip reading are deep learning-based and they largely focus on decoding long speech segments in the form of:

- Words and sentences using either words.
- ASCII characters.

Both of them used as the classes to recognize.

About dataset

MIRACL-VC1 is a lip-reading dataset including both depth and color images (in this work we only use color images), it was obtained from kaggle.



- 10 Women
- 5 Men

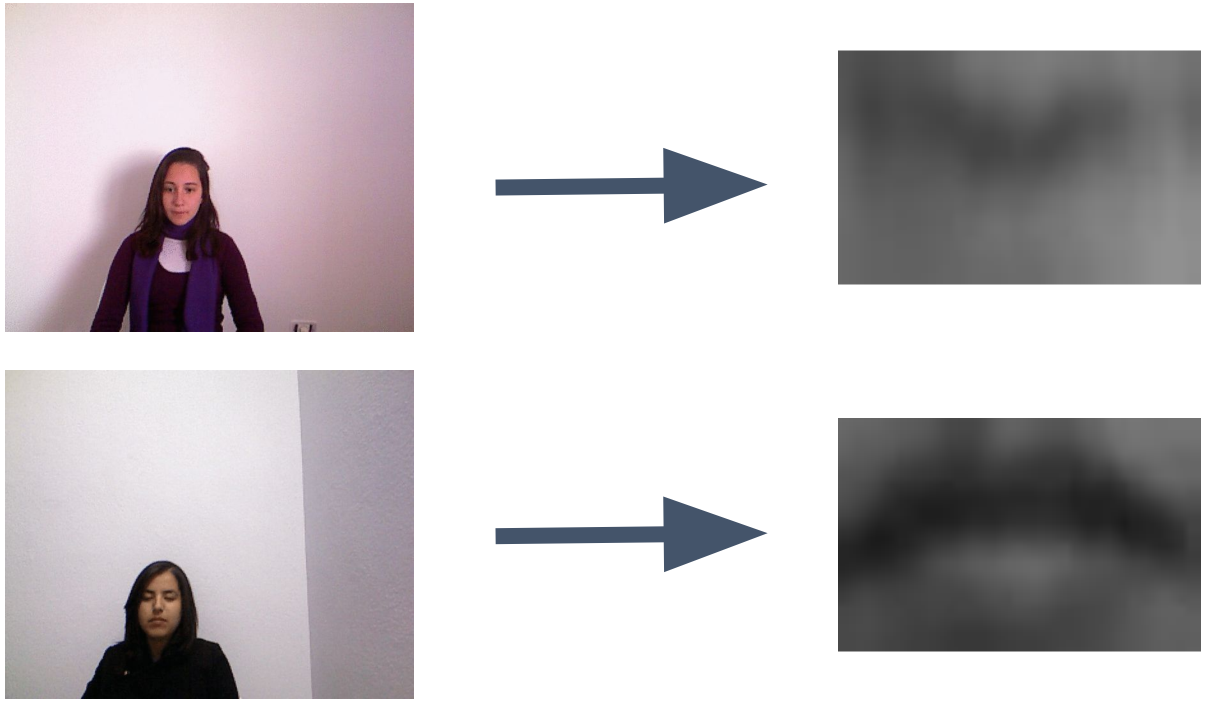
- 10 Sentences
- 10 Instances

['F01','F02','F04','F05','F06','F07','F08','F09', 'F10','F11','M01','M02','M04','M07','M08']

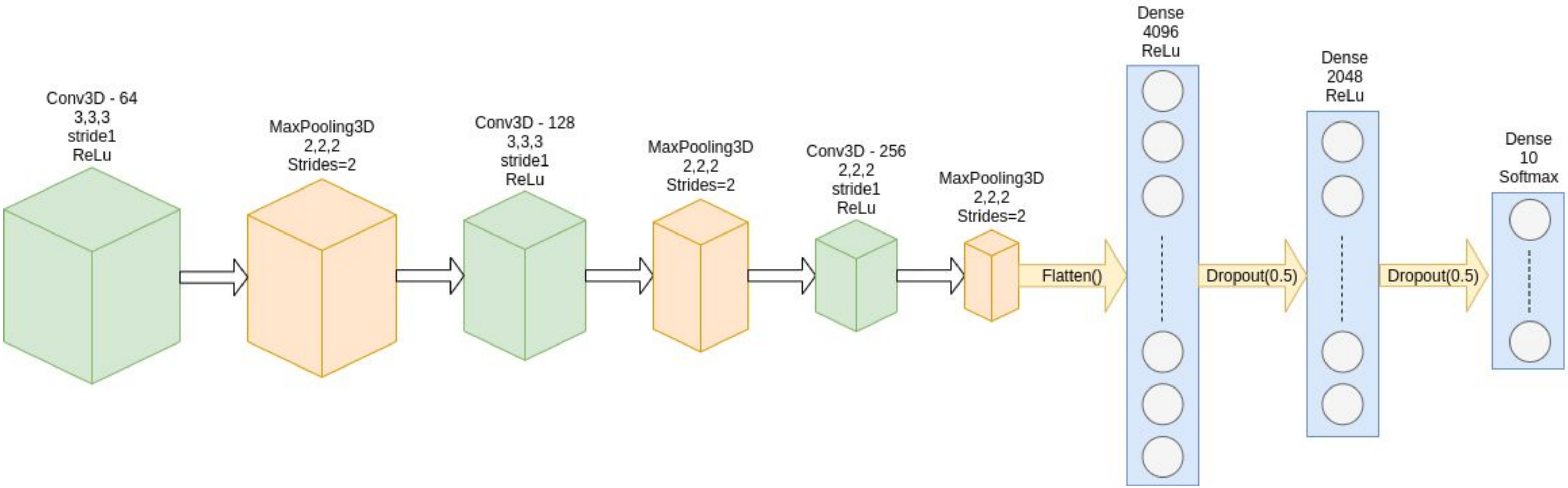
['Begin', 'Choose', 'Connection', 'Navigation', 'Next', 'Previous', 'Start', 'Stop', 'Hello', 'Web']

Mouth segmentation

Due to the size of the images it was necessary to reduce their size, openCV was used to segment the region corresponding to the mouth.

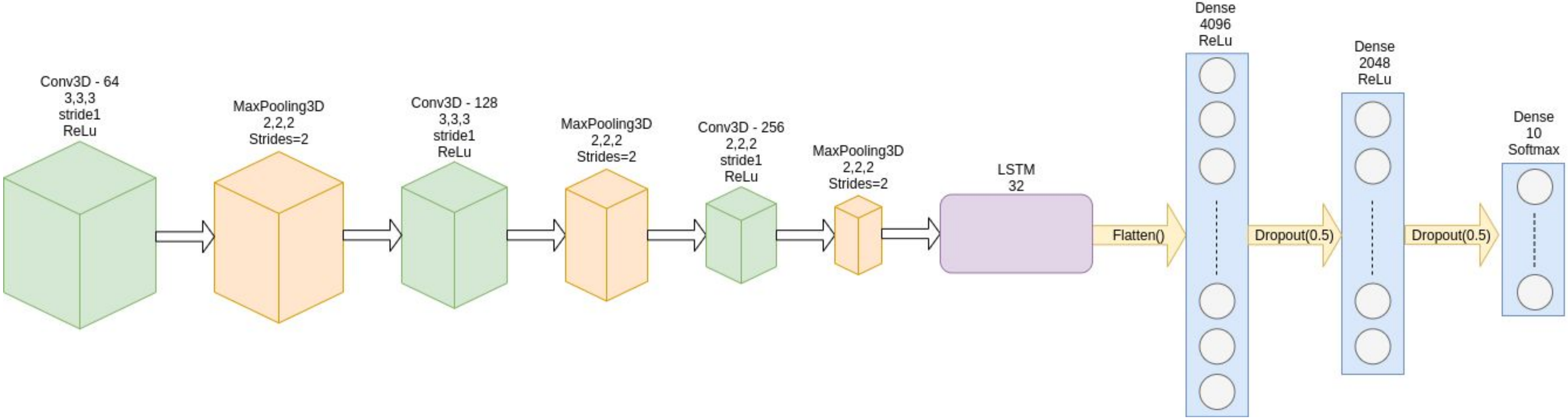


Model1_3D-CNNs



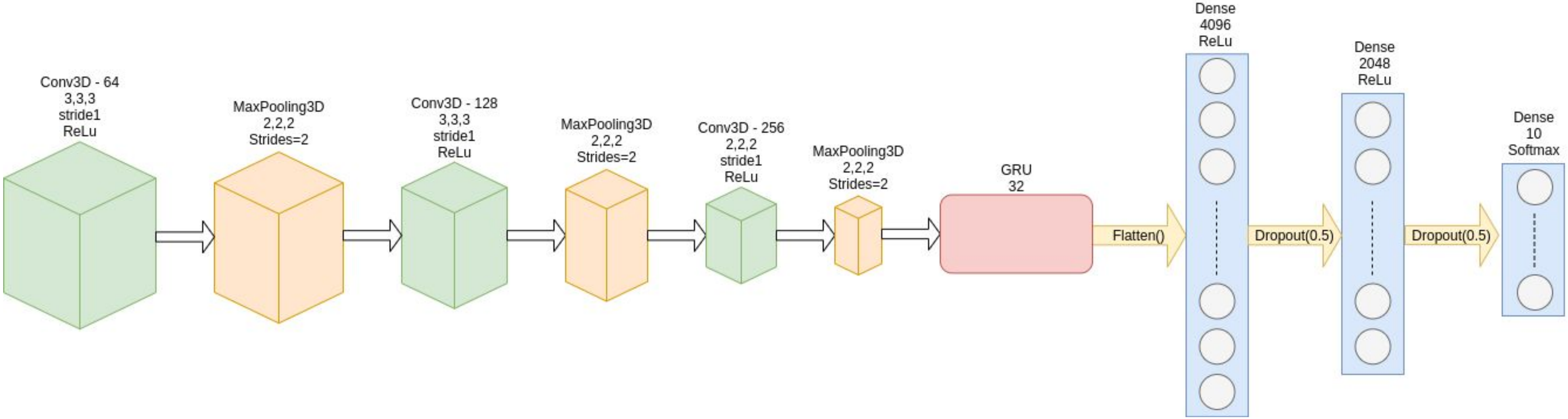

```
loss='categorical_crossentropy',
optimizer='Adagrad',
metrics=['accuracy']
```

Model2_3D-CNNs-LSTM



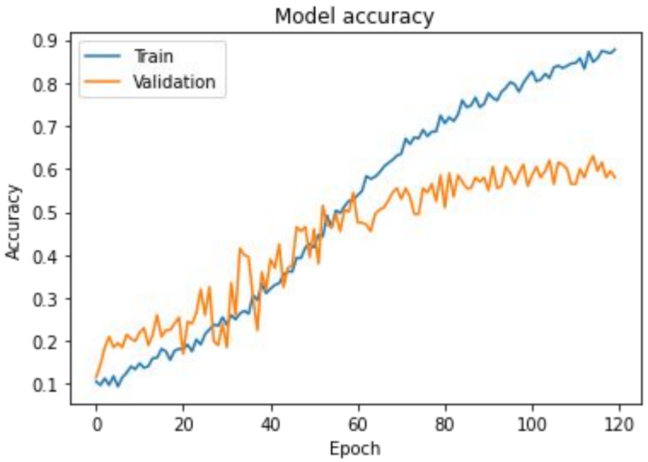

```
loss='categorical_crossentropy',
optimizer='Adagrad',
metrics=['accuracy']
```

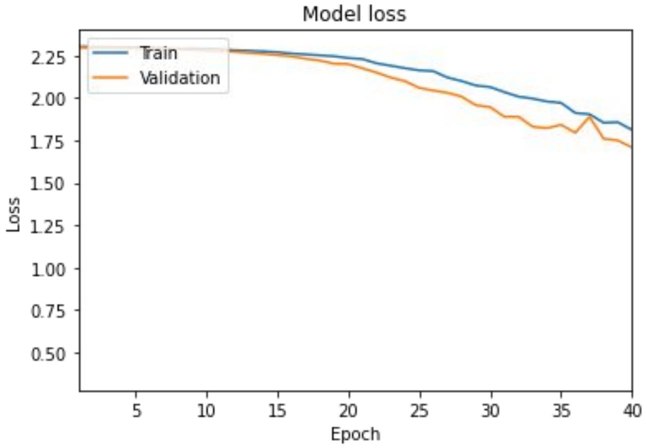
Model3_3D-CNNs-GRU




```
loss='categorical_crossentropy',
optimizer='Adagrad',
metrics=['accuracy']
```

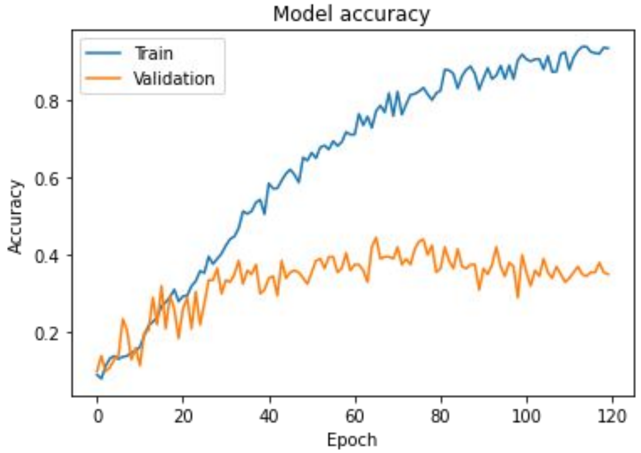
Model1_3D-CNNs

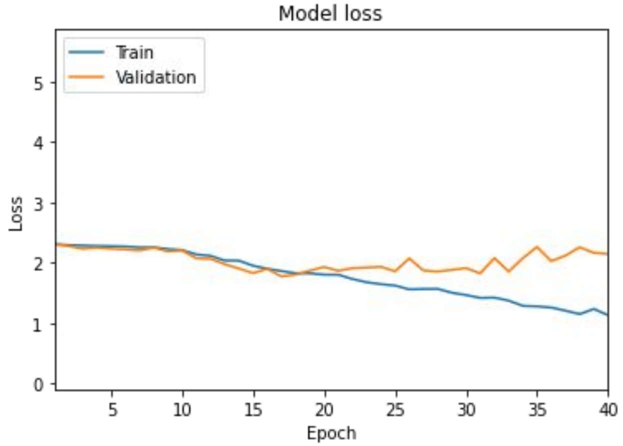




Accuracy = 0.26 on completely unseen data

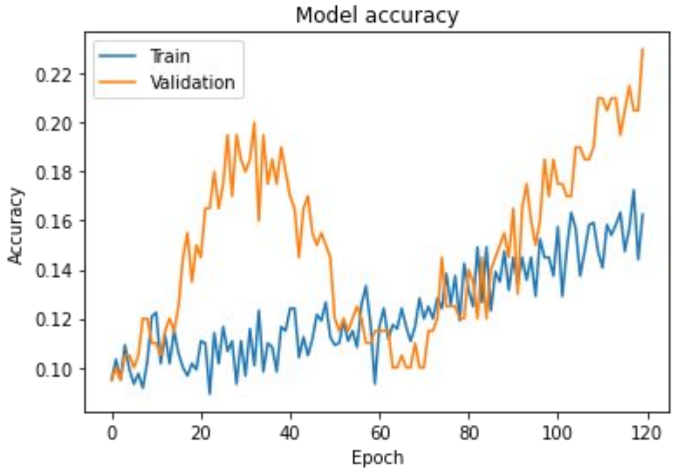
Model2_3D-CNNs-LSTM

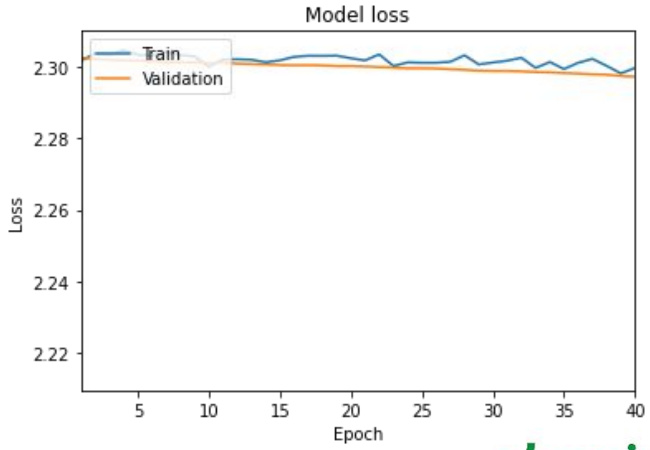




Accuracy = 0.14 on completely unseen data

Model3_3D-CNNs-GRU

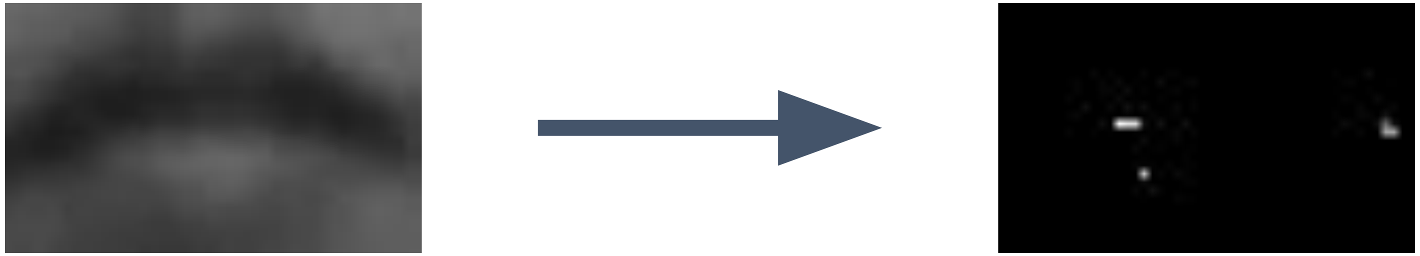




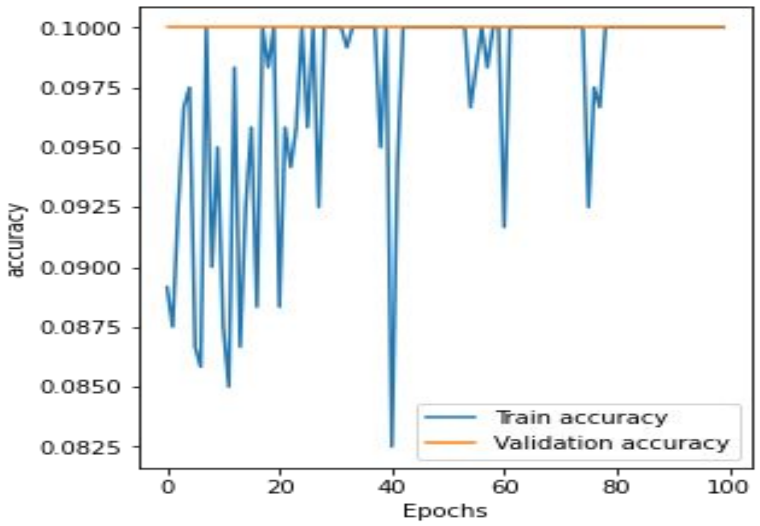
Accuracy = 0.18 on completely unseen data

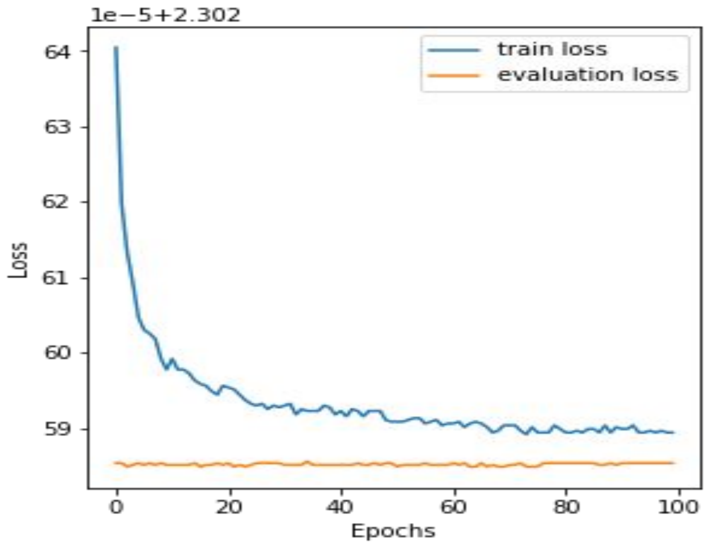
We also try

Optical flow (Lucas Kanade) describes a sparse or dense vector field, where a displacement vector is assigned to certain pixel position, that points to where that pixel can be found in another image.

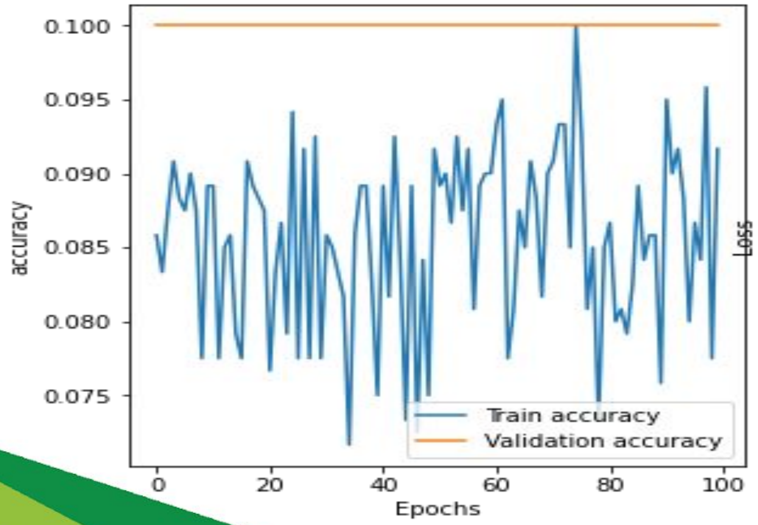


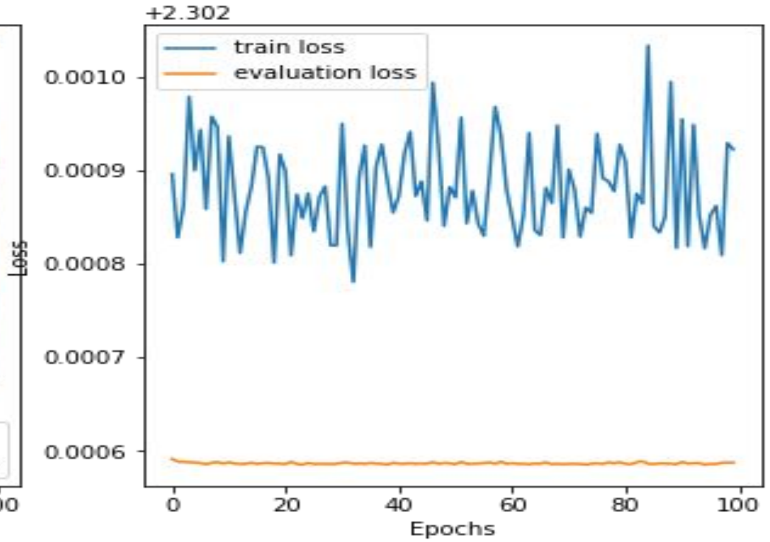
Model1_3D-CNNs_Lukas-Kanae

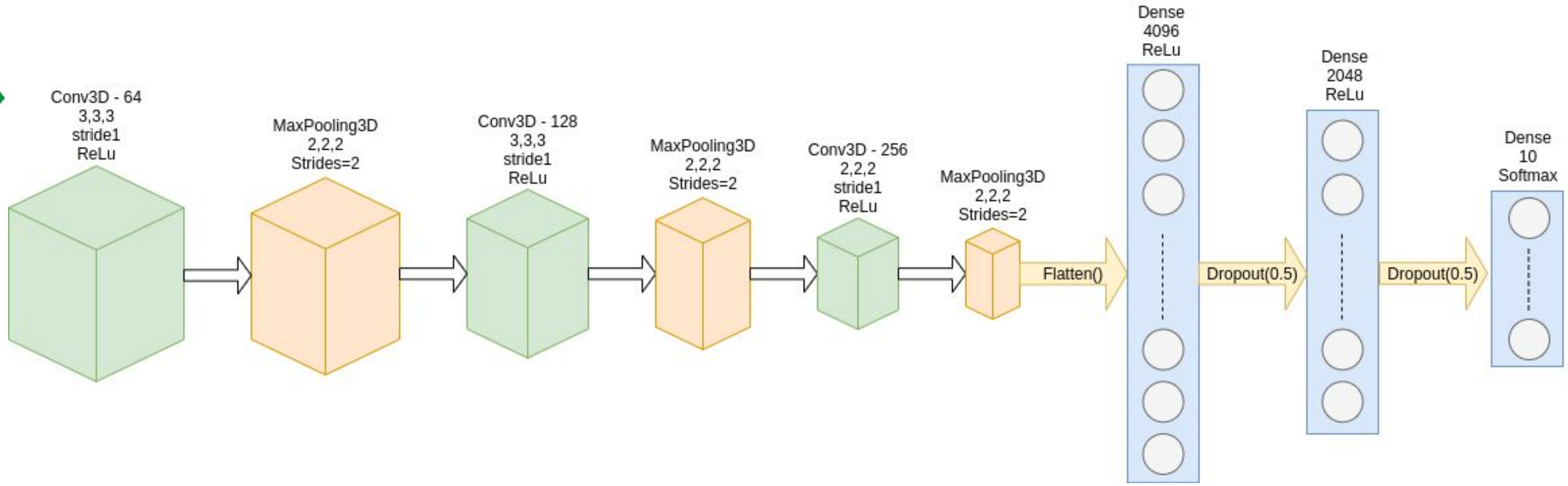




Model2_3D-CNNs-LSTM_Lukas-Kanae








```
loss='categorical_crossentropy',
optimizer='Adagrad',
metrics=['accuracy']
```

The best Model1_3D-CNNs

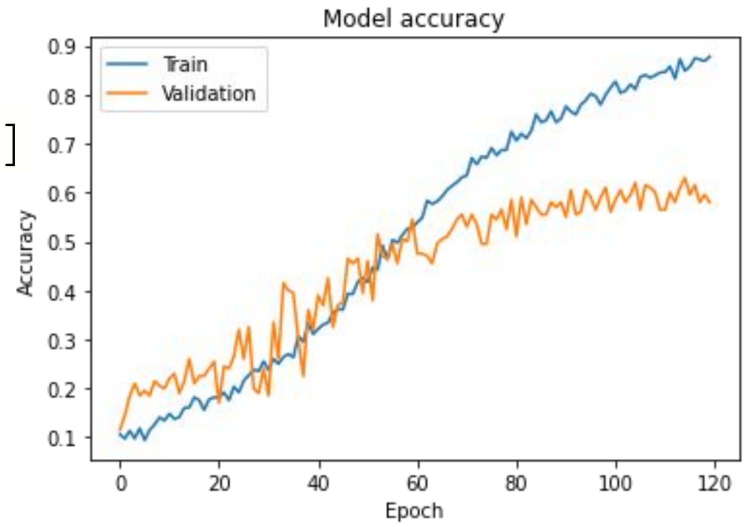
Star experimentation

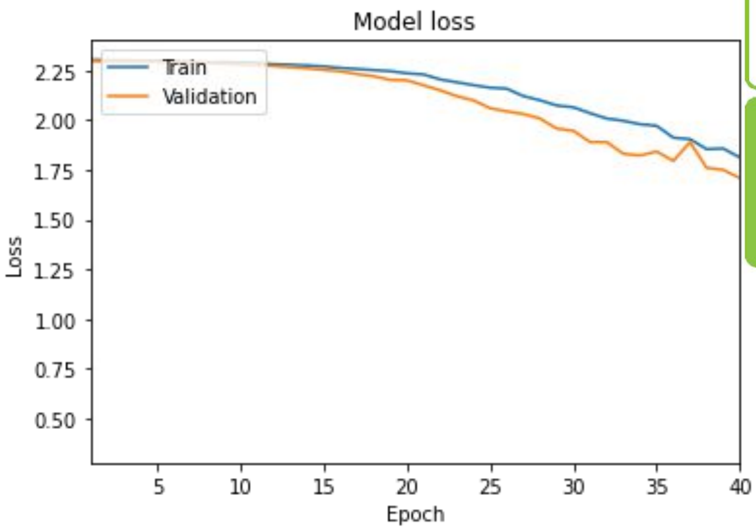
```
loss='categorical_crossentropy'
optimizer='Adagrad'
metrics=['accuracy']
```

Universidad Industrial de Santander

Model1_3D-CNNs

Dense Activation ReLu

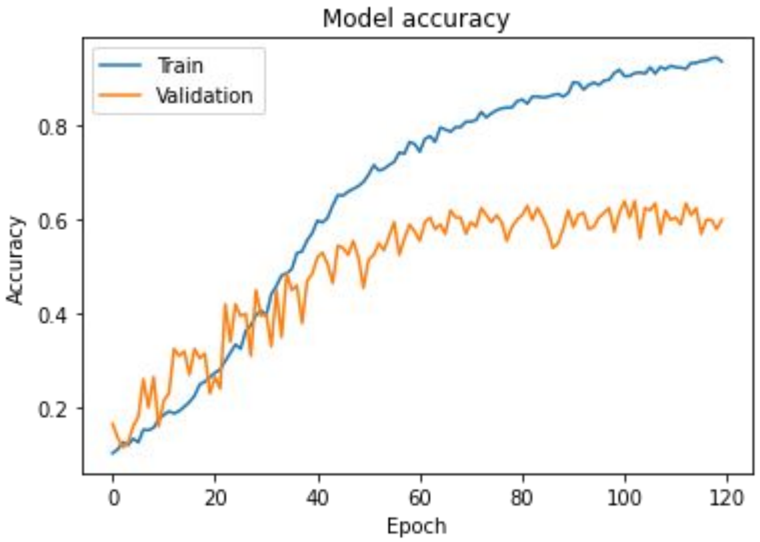


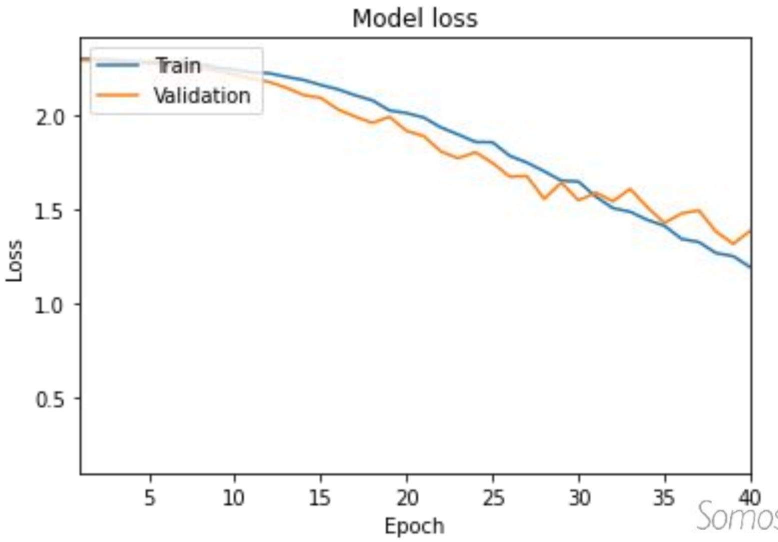


Accuracy = 0.26 on completely unseen data

Model1_3D-CNNs

Dense Activation tanh





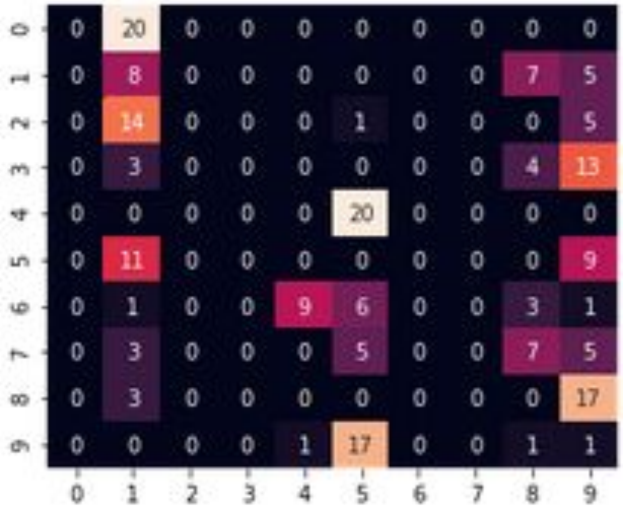
Accuracy = 0.31 on completely unseen data

```
loss='categorical_crossentropy'
optimizer='Adagrad'
metrics=['accuracy']
```

Model1_3D-CNNs

Dense Activation ReLu

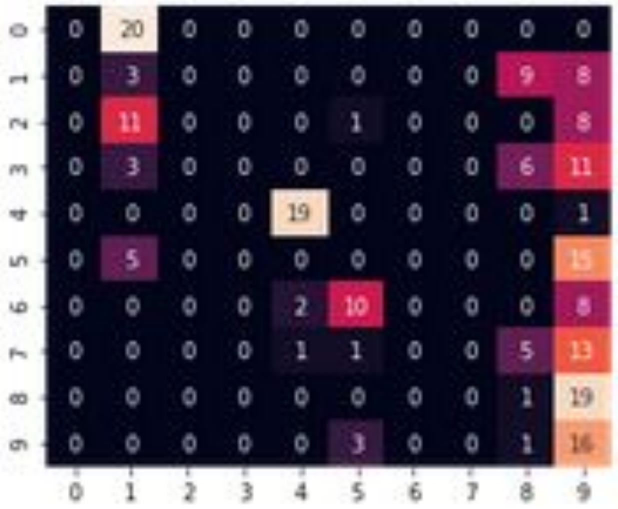
```
Accuracy: 0.045000
Precision: 0.014484
Recall: 0.045000
F1 score: 0.021909
Cohens kappa: -0.061111
```



Model1_3D-CNNs

Dense Activation tanh

```
Accuracy: 0.195000
Precision: 0.114214
Recall: 0.195000
F1 score: 0.131806
Cohens kappa: 0.105556
```

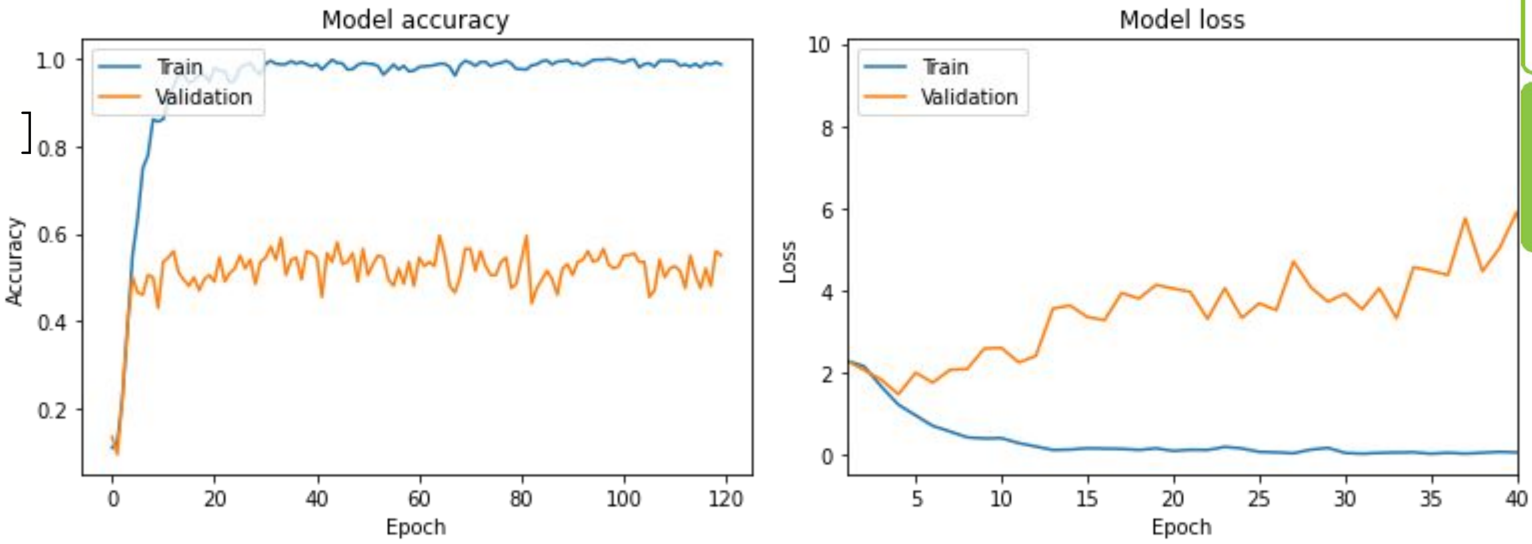



```
['Begin', 'Choose',
'Connection',
'Navigation', 'Next',
'Previous', 'Start',
'Stop', 'Hello', 'Web']
```

```
loss='categorical_crossentropy'
optimizer=Adam
metrics=['accuracy']
```

Model1_3D-CNNs

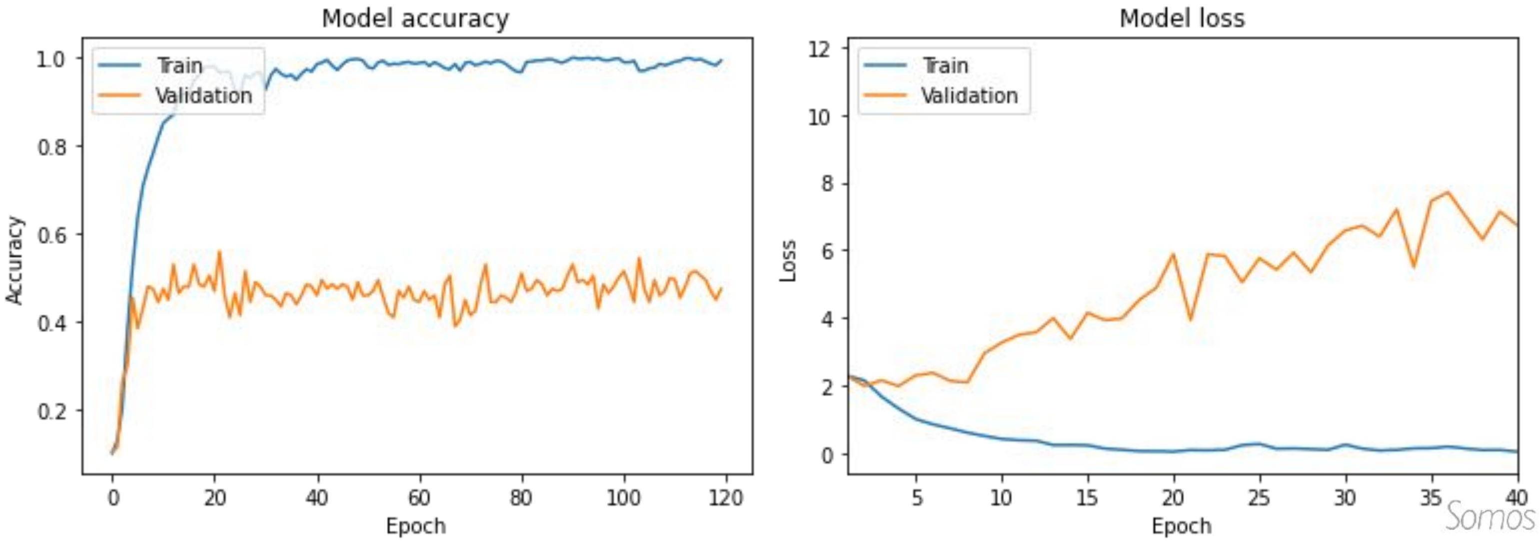
Dense Activation ReLu



Accuracy = 0.37 on completely unseen data

Model1_3D-CNNs

Dense Activation tanh



Accuracy = 0.41 on completely unseen data

```
loss='categorical_crossentropy'
optimizer=Adam
metrics=['accuracy']
```

Universidad Industrial de Santander

Model1_3D-CNNs

Dense Activation ReLu

```
Accuracy: 0.130000
Precision: 0.074604
Recall: 0.130000
F1 score: 0.087352
Cohens kappa: 0.033333
```


```
['Begin', 'Choose',
'Connection',
'Navigation', 'Next',
'Previous', 'Start',
'Stop', 'Hello', 'Web']
```

Model1_3D-CNNs

Dense Activation tanh

```
Accuracy: 0.230000
Precision: 0.171074
Recall: 0.230000
F1 score: 0.183518
Cohens kappa: 0.144444
```

Some conclusions

- It’s important to optimize use of resources in problems with a large size data to avoid OOM errors.
- RNN has shown improvements NLP problems like video to speech, but i this case our knowledge due to dataset structure has prevented us from obtaining good results.
- Real-life problems related with speech recognition are a hard problem to solve with classic CNN networks.
- Preprocessing data is not always a good idea and it’s necessary to be careful about.
- Dataset that we use is kind a real-life datasets.

Universidad Industrial de Santander

#LaUISqueQueremos

¡Gracias!