Lip Reading with CNN, LSTM and Transformer

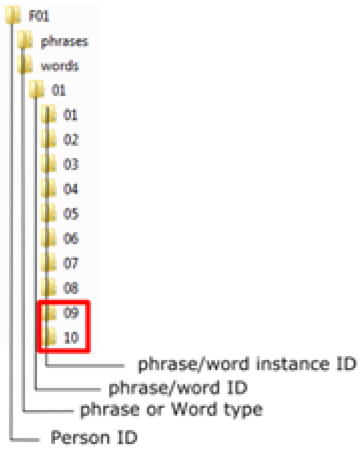
Task Introduction

- Given a sequence of images which show a person speaking a word
- Classify which word the person speaks
- Compare the performances of different neural network architectures
  - CNN + FC
  - CNN + LSTM
  - Transformer

Words
Begin
Choose
Connection
Navigation
Next
Previous
Start
Stop
Hello
Web

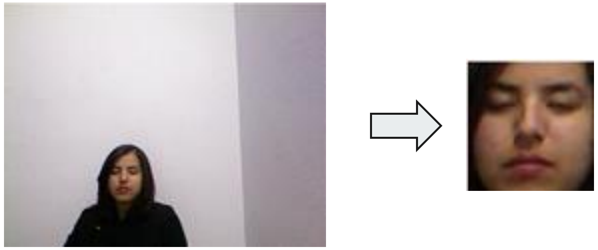
Dataset

- 10 females and 5 males
- Each person speaks 10 times for each of the 10 words
- Total: 15 x 10 x 10 = 1500 instances
- Train and test set:
  - Seen test set: take two instances from each word spoken by each person (i.e. 300 instances)
  - Unseen test set: take the instances from two person (i.e. 300 instances)
  - Validation set: take 20% from train set

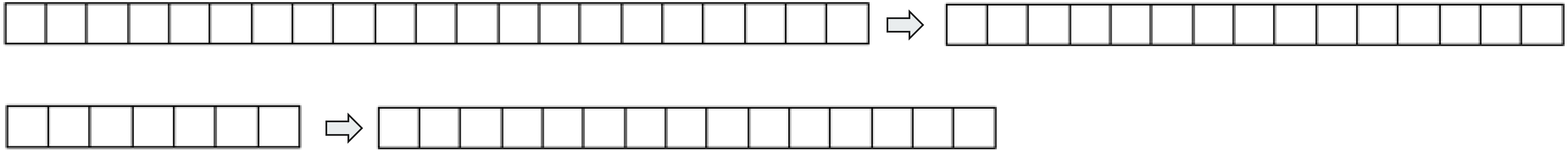


Data Preprocessing

- Original image are too large (640 x 480)
  - Redundant information from the background
  - Extract the face from the whole image using existing tools (e.g. dlib)
  - Resize to 64 x 64 pixels

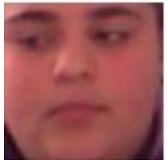
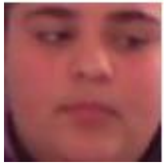


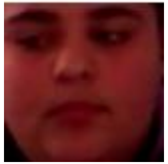
- Different lengths for the image sequences
  - Trim down to the 15th frame
  - Pad up to the 15th frame



Data Augmentation

- Horizontal Flip (1200 instances)
- Adjust Brightness (1200 instances)
- Channel Shift (1200 instances)
- Results in total 4800 instances for training set

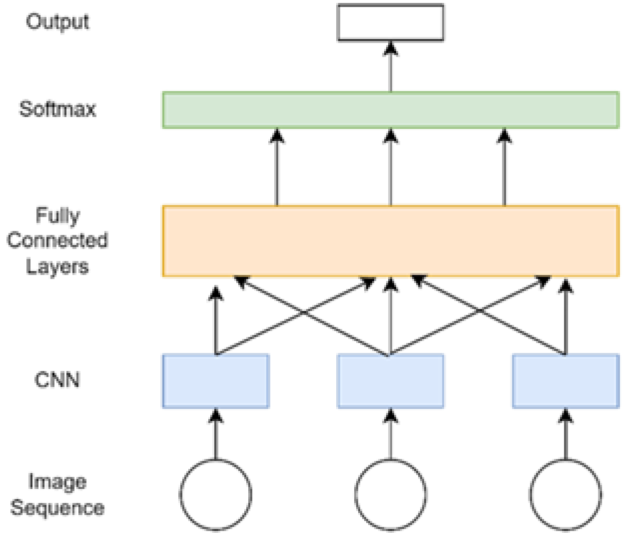


Models Introduction

- Compare CNN + FC (baseline) , CNN + LSTM, Transformer
- We expect CNN + FC performs worse and Transformer performs the best

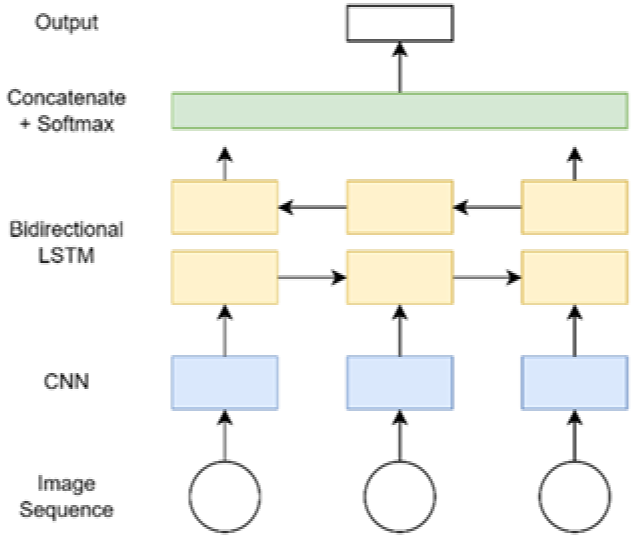
CNN + FC (baseline)

- Each image in the sequence is fed into the CNN
- The output features from CNN are fed into the fully-connected layers



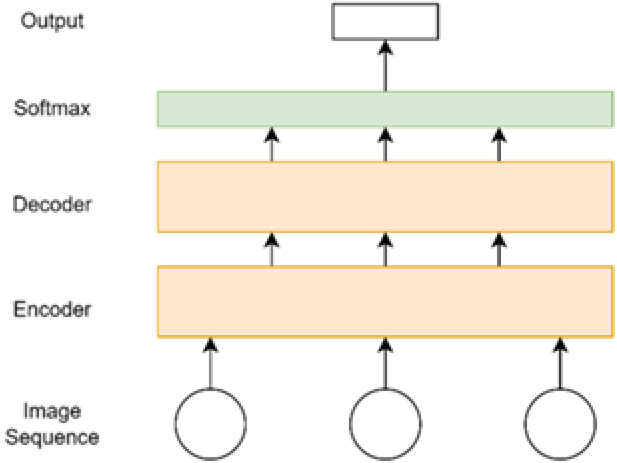
CNN + Bidirectional LSTM

- Each image in the sequence is fed into the CNN
- The output features from CNN are fed into the bi-LSTM layers
- Expect to perform better since the order of the images matters

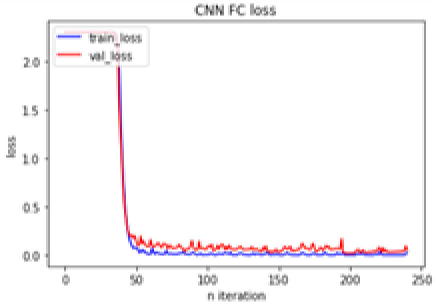


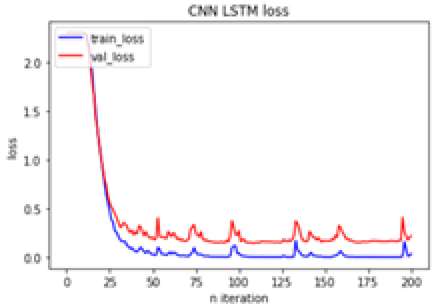
Transformer

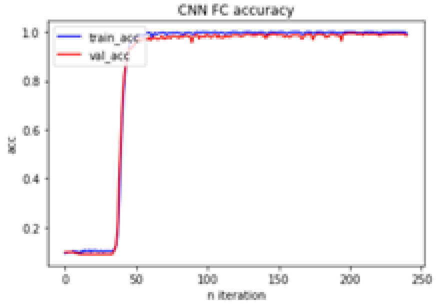
- The images in the sequence are fed to the encoder and decoder architecture
- Expect to perform better due to the attention mechanism

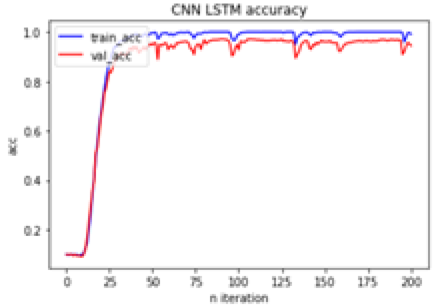


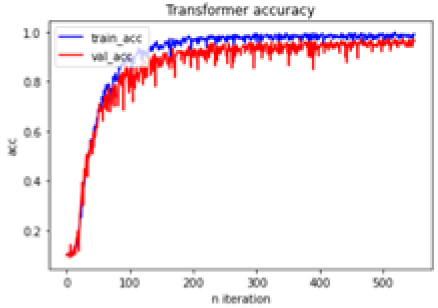
Models Training











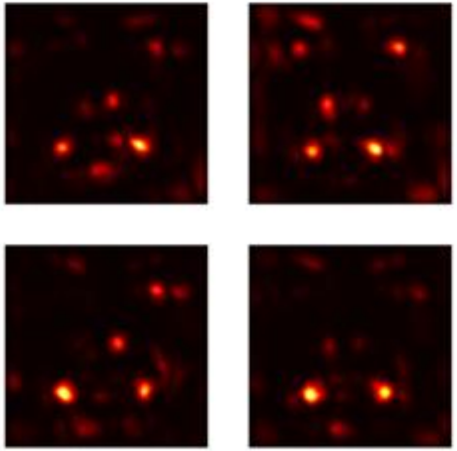
Results: Seen Dataset

Model	Train Accuracy	Validation Accuracy	Test Accuracy
CNN + FC (baseline)	100.0%	99.6%	62.7%
CNN + LSTM	99.7%	95.2%	86.3%
Transformer	97.2%	90.1%	85.0%

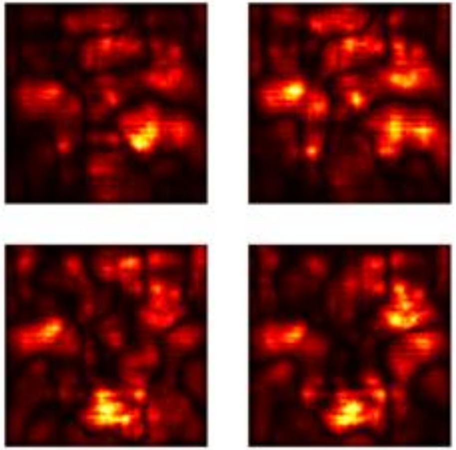
- Results are closed from the expectation
- We may need more data for the transformer to learn

Results: Seen Dataset (Saliency)

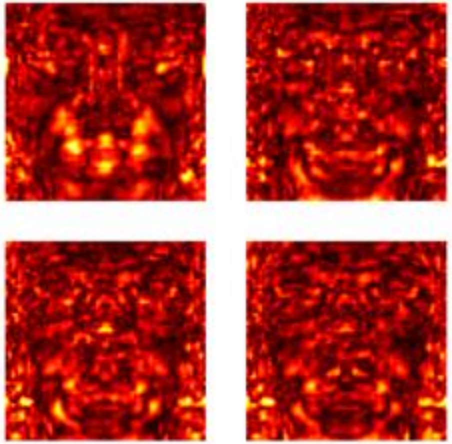
CNN + FC



CNN + LSTM



Transformer



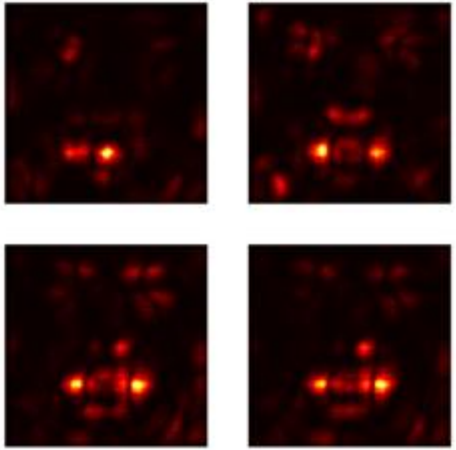
Results: Unseen Dataset (Unseen People)

Model (Unseen)	Train Accuracy (Unseen)	Validation Accuracy (Unseen)	Test Accuracy (Unseen)
CNN + FC (baseline)	100.0%	99.2%	42.5%
CNN + LSTM	100.0%	98.1%	17.0%
Transformer	97.8%	92.9%	16.0%

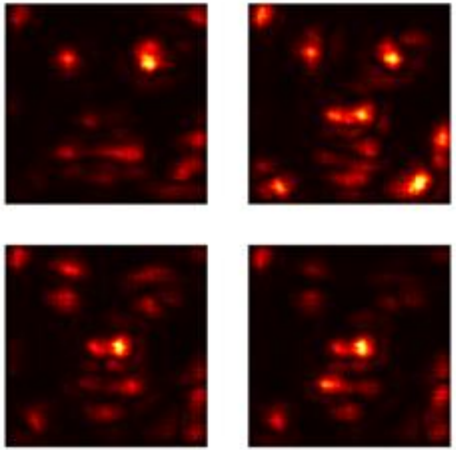
- There may be overfitting
- We may need more data

Results: Unseen Dataset (Saliency)

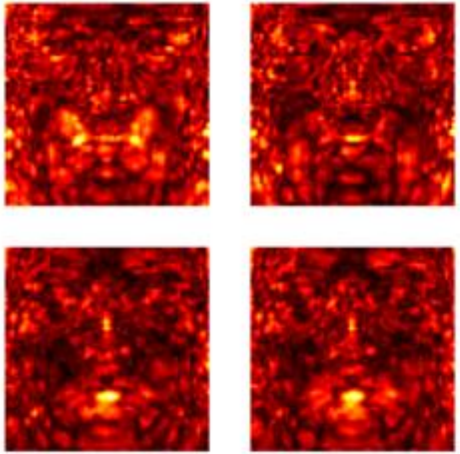
CNN + FC



CNN + LSTM



Transformer



Conclusion

Trained Model	Test Accuracy	Trained Model (Unseen)	Test Accuracy (Unseen)
CNN + FC (baseline)	62.7%	CNN + FC (baseline)	42.5%
CNN + LSTM	86.3%	CNN + LSTM	17.0%
Transformer	85.0%	Transformer	16.0%

- Results are quite acceptable for the models trained on the first dataset
- But the models trained using the unseen dataset perform bad
  - It may due to overfitting
  - More data are needed since there are only in total 13 people (13 different faces) used in training

Thank You