Lip Reading Number Recognition

W251 Final Project

Marcus Chen, Marcelo Scatolin Queiroz,

Sylvia Yang, Wei Wang

Github Repo

Background

- Lip Reading:
  - Definition
  - Challenges
  - Practical Applications:
    - Hearing Aids
    - Speech Recognition
    - Multi-talker
- Model Approach:
  - Word Level Recognition / Sentence Level Recognition (contextual understanding)
- Project Objectives:
  - **LipNet for Digits 0-9!**
  - Improve speech recognition in noisy environments
  - Combine with facial recognition techniques for password identification

LipNet

95.2%

Accuracy for word level prediction

Rich, Well Documented Code

GRID, STCNNs, RNNs, Bi-GRUs, CTC

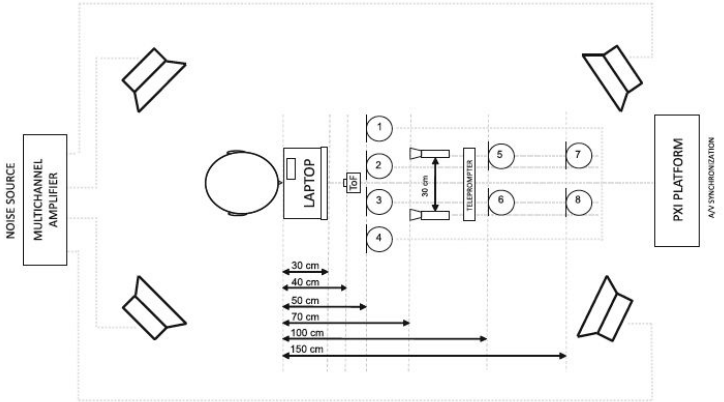
GRID Corpus (2007)

- Original dataset for LipNet
- 34 speakers
- Each speaker produce 1000 sentences (64,000 possible comb.)
- Contains Digits and Other Words
- Hoping to improve the LipNet's performance on digit inference we decided to go with a new option.
- Training LipNet model on MODALITY Corpus using IBM Cloud Virtual Machines

Modality Corpus (2016)

- Vast vocabulary including digits
- Multi camera/mic array
- high-resolution (1080x1920 pixels at 100 fps) video streams
- 35 native and non-native english speakers
- Approx. 2.5 TB clipped to 790 MB

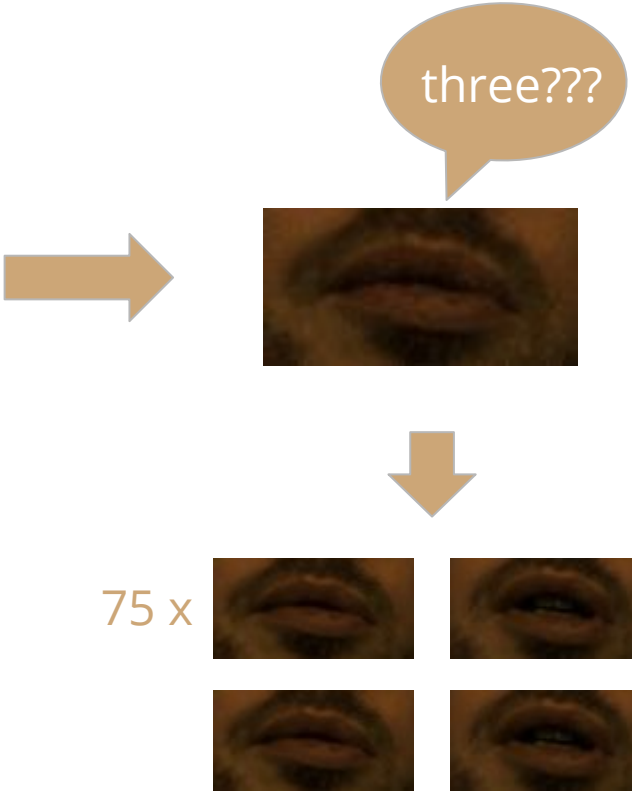
```
141650000 147330000 PAUSE
157580000 162460000 CALL
173570000 176870000 TAB
190040000 195850000 PICTURES
205970000 211590000 MOVE
223670000 229010000 ALARM
239600000 245220000 MUTE
256470000 259700000 FORTY
259770000 264010000 FIVE
274140000 278970000 JUNE
289520000 295370000 FILE
306360000 314450000 MINIMIZE
324240000 329760000 VIDEOS
324240000 329760000 VIDEOS
340350000 345320000 RUN
355370000 361540000 SEND
374200000 380730000 MILLION
391580000 398950000 DOCUMENTS
408960000 414980000 BROTHER
426450000 432070000 VOLUME
```



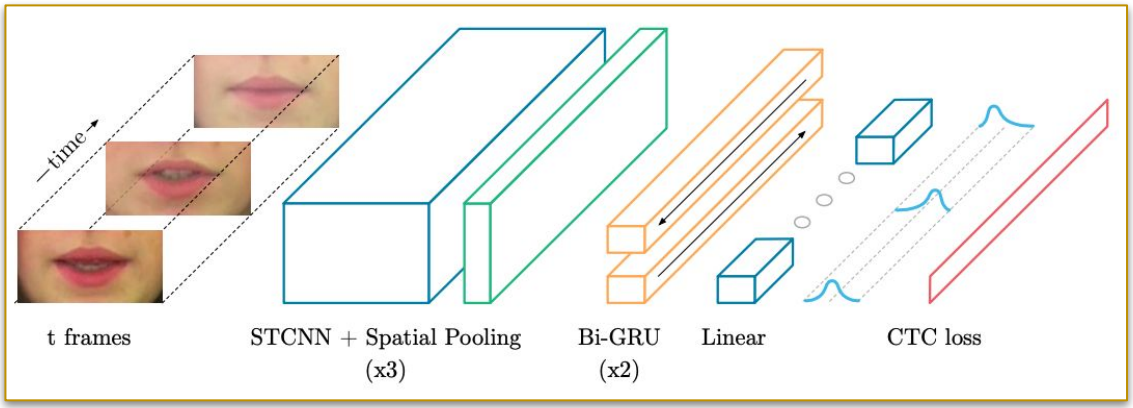
Modality Corpus

Codes	Goals	Effect
raw_downloader.py	Improve time of downloading videos	Eliminate corrupted transcriptions
file_clipper.py	Clip videos to process only where digits are spoken	790 MB of 5538 files saved to IBM Cloud
modality_to_GRID_Converter.py	Dataset Adaptation to LipNet	75 Frozen Frames each file
extract_mouth_batch.py	Clip for mouth movements only	100x50 pixels crop of mouth Train: 13 speakers (4170 files) Valid.: 4 speakers (1159 files)

Data Processing



LipNet Architecture



Layer (type)	Output Shape	Param #
the_input (InputLayer)	(None, 75, 100, 50, 3)	0
zero1 (ZeroPadding3D)	(None, 77, 104, 54, 3)	0
conv1 (Conv3D)	(None, 75, 50, 25, 32)	7232
batc1 (BatchNormalization)	(None, 75, 50, 25, 32)	128
actv1 (Activation)	(None, 75, 50, 25, 32)	0
spatial_dropout3d_1 (Spatial	(None, 75, 50, 25, 32)	0
max1 (MaxPooling3D)	(None, 75, 25, 12, 32)	0
zero2 (ZeroPadding3D)	(None, 77, 29, 16, 32)	0
conv2 (Conv3D)	(None, 75, 25, 12, 64)	153664
batc2 (BatchNormalization)	(None, 75, 25, 12, 64)	256
actv2 (Activation)	(None, 75, 25, 12, 64)	0
spatial_dropout3d_2 (Spatial	(None, 75, 25, 12, 64)	0
max2 (MaxPooling3D)	(None, 75, 12, 6, 64)	0
zero3 (ZeroPadding3D)	(None, 77, 14, 8, 64)	0
conv3 (Conv3D)	(None, 75, 12, 6, 96)	165984
batc3 (BatchNormalization)	(None, 75, 12, 6, 96)	384
actv3 (Activation)	(None, 75, 12, 6, 96)	0
spatial_dropout3d_3 (Spatial	(None, 75, 12, 6, 96)	0
max3 (MaxPooling3D)	(None, 75, 6, 3, 96)	0
time_distributed_1 (TimeDist	(None, 75, 1728)	0
bidirectional_1 (Bidirection	(None, 75, 512)	3048960
bidirectional_2 (Bidirection	(None, 75, 512)	1181184
dense1 (Dense)	(None, 75, 28)	14364
softmax (Activation)	(None, 75, 28)	0

Total params: 4,572,156.0
Trainable params: 4,571,772.0
Non-trainable params: 384.0

Implementation

LipNet

https://github.com/rizkiarm/LipNet

x86
Python 2.7
TF-GPU 1.0.1
Keras 2.0.2

TX2 (nvidia-docker)

ARM64
Python 3.7
TF-GPU 1.13.1
Keras 2.2.4

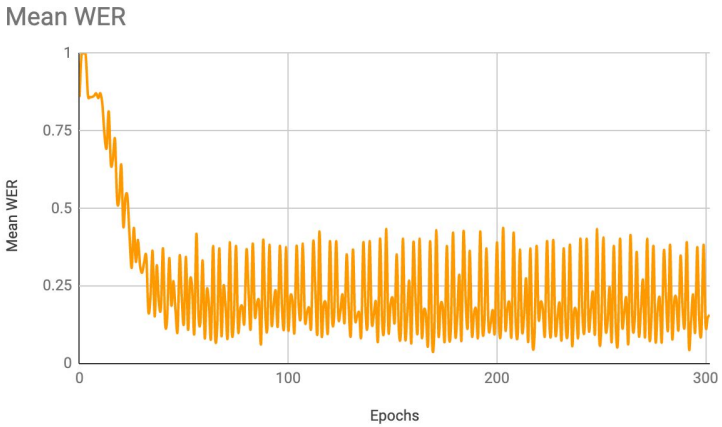
V100 (docker)

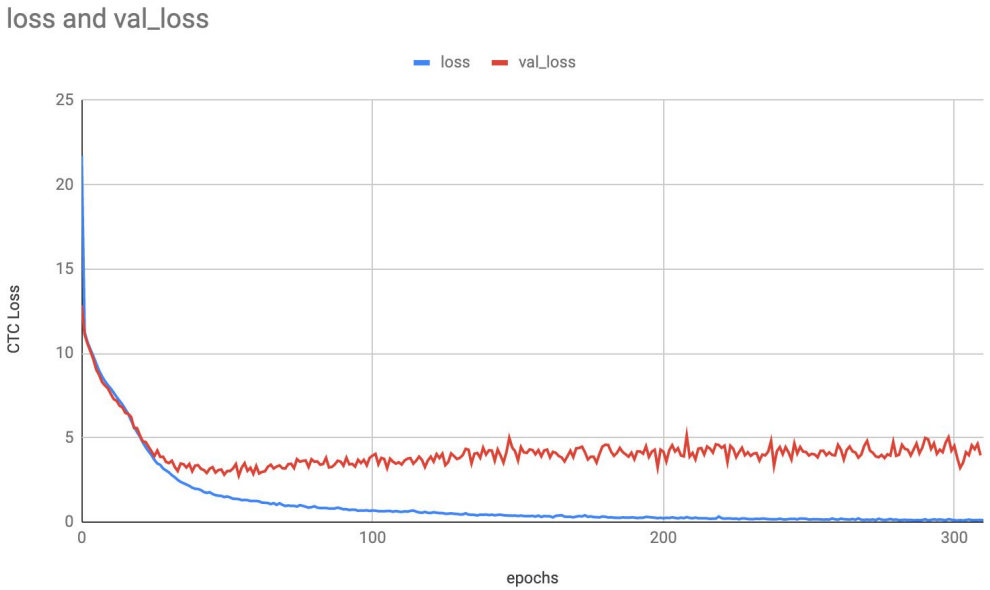
x86
Python 2.7
TF-GPU 1.0.1
Keras 2.0.2

Performance

Model Name	Epochs	Word Error Rate (WER)	Effect
Original LipNet	5000	11.4% for unseen speakers 4.8% for overlapped speakers	
Our W251 Model	300	12.5%	3 minutes per epoch
Pass et.al 2010	N/A	2%	
Stewart et. al 2014	N/A	30%	

Performance





Future Development

- Leverage Audio
  - Readily Available
  - Prepared script to handle audio tracks!
- Data Augmentation Tools
  - “Ground truth” generator script
- Adapting Other Datasets
- TX2
  - scikit-video to opencv
  - TX2 to detect, send to cloud and perform inference

References

- Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING. Retrieved August 4, 2019, from https://arxiv.org/pdf/1611.01599v2.pdf.

- Adriana Fernandez-Lopez, Federico M. Sukno. (2018). Survey on automatic lip-reading in the era of deep learning. Retrieved July 27, 2019, from https://www.sciencedirect.com/science/article/pii/S0262885618301276

- LipNet github

- GRID dataset: http://spandh.dcs.shef.ac.uk/gridcorpus/

- MODALITY dataset: http://www.modality-corpus.org/