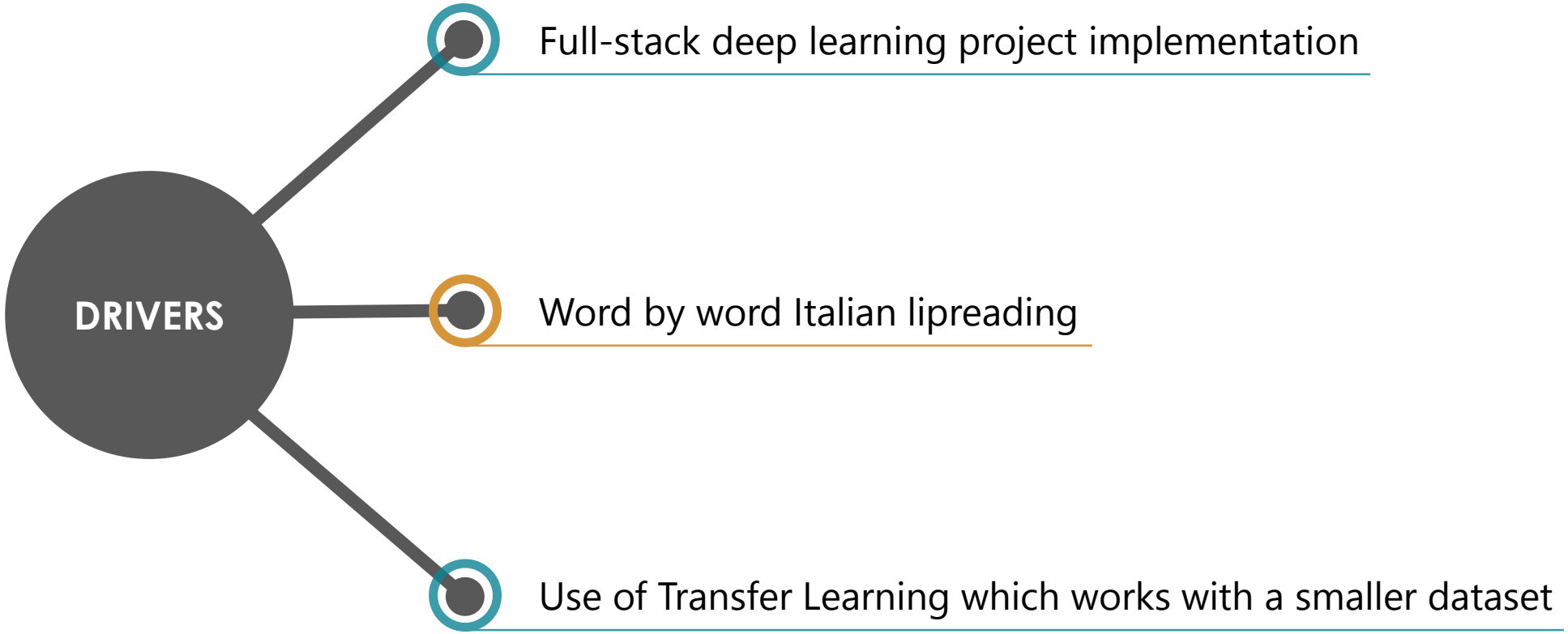
Ita–Lip

Lipreading Italian Words Through Transfer Learning and seq2seq

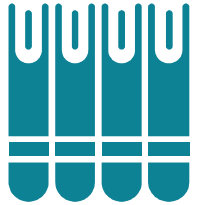
Project Drivers



Baseline

Hearing impaired people

When labelling 300 random videos, word error rate (WER) of 47.7%.

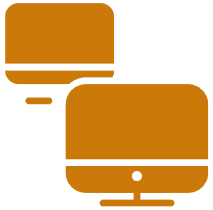


Literature

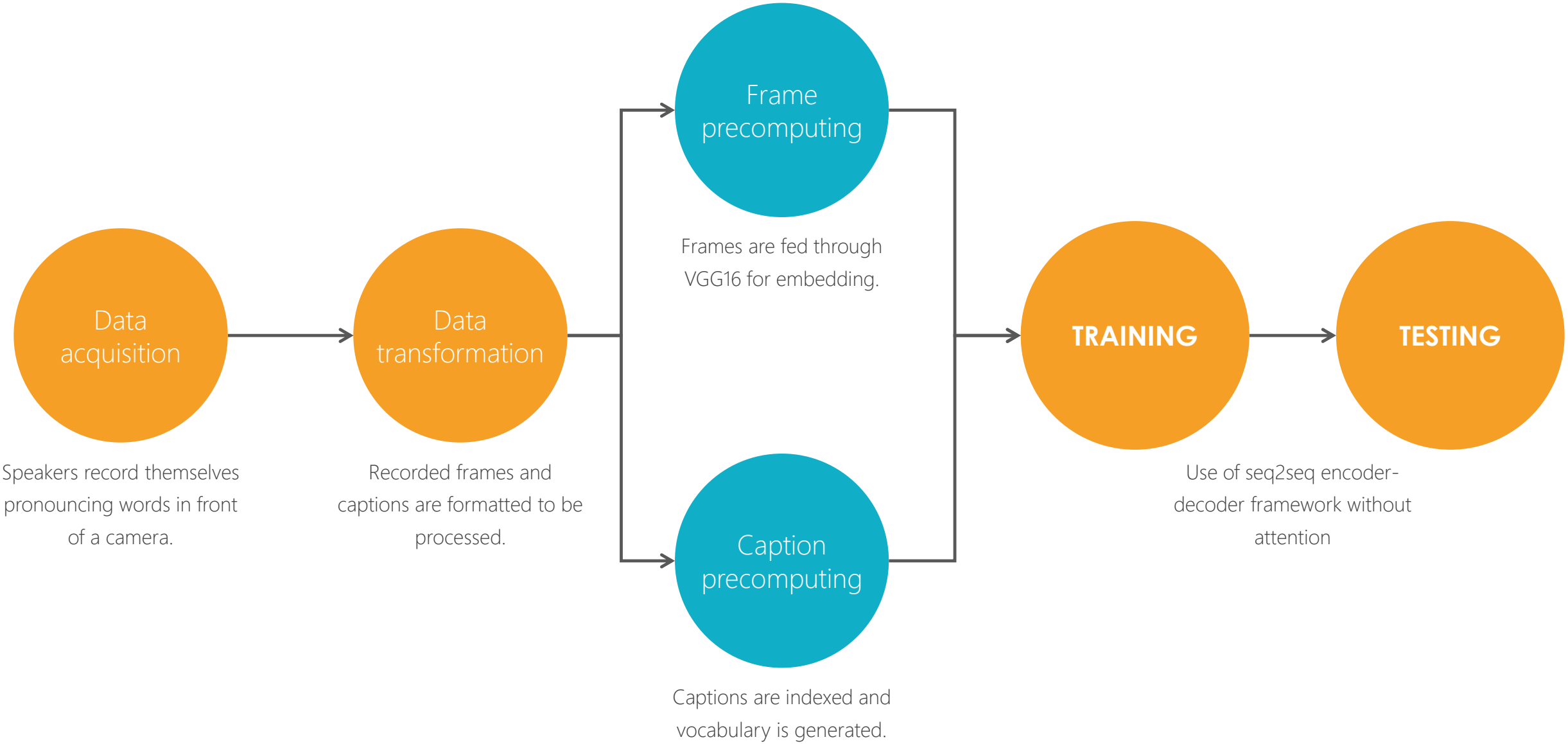
According to the literature, the accuracy of human lip readers is around 20% - WER of 80%.

LipNet

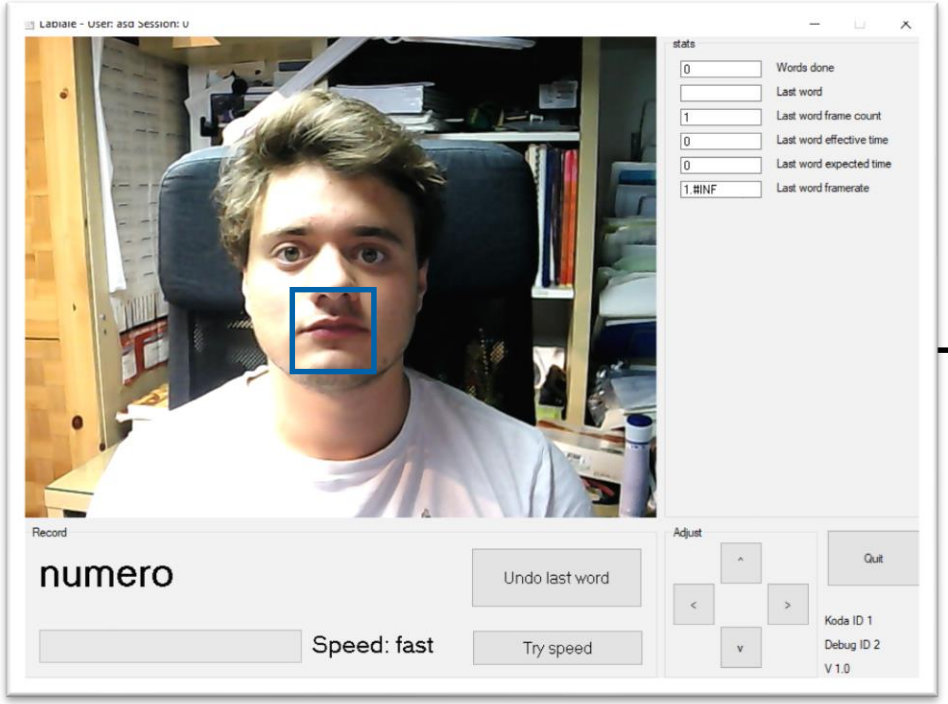
LipNet exhibits a WER of 11.4% when performing on unseen speakers.

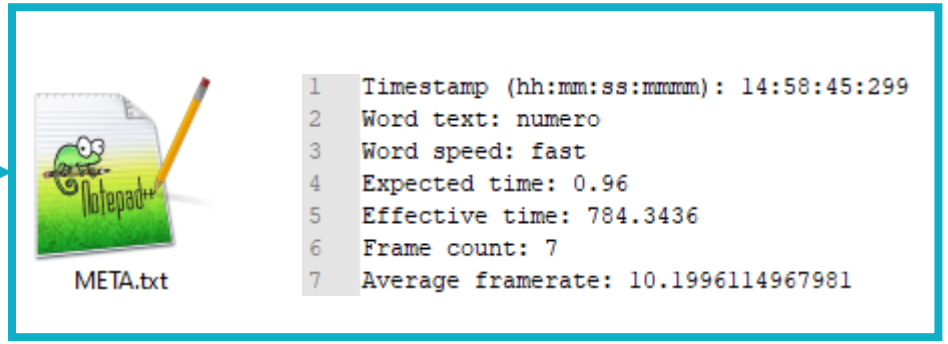


Process Overview



Data Collection

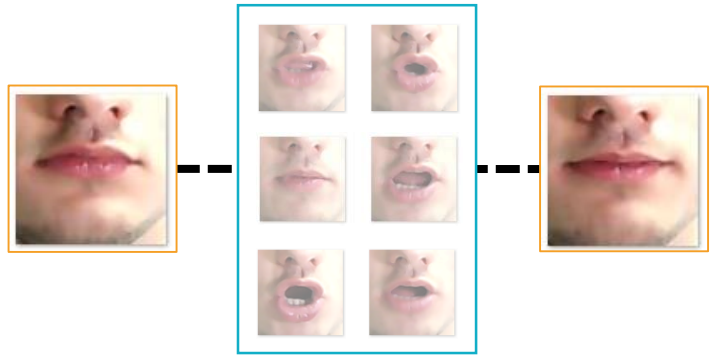




Speaker

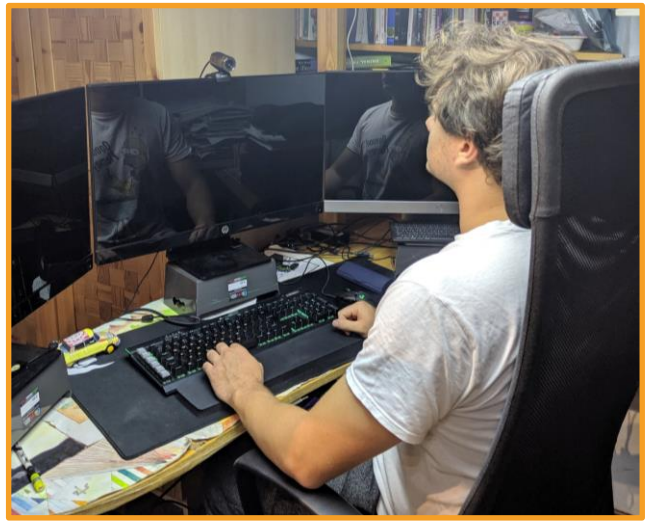
Speech recording

Recording starts right before the speaker opens his/her mouth and stops immediately after their lips reach the rest position after the word.



Speaker position

The speaker faces the camera directly straight and aligns their mouth with the recording area.

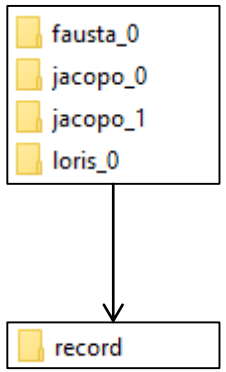


Data Transformation

Integration

Recordings from more speakers are integrated in a single dataset containing all spoken words.

Captions are also integrated into one single caption file.



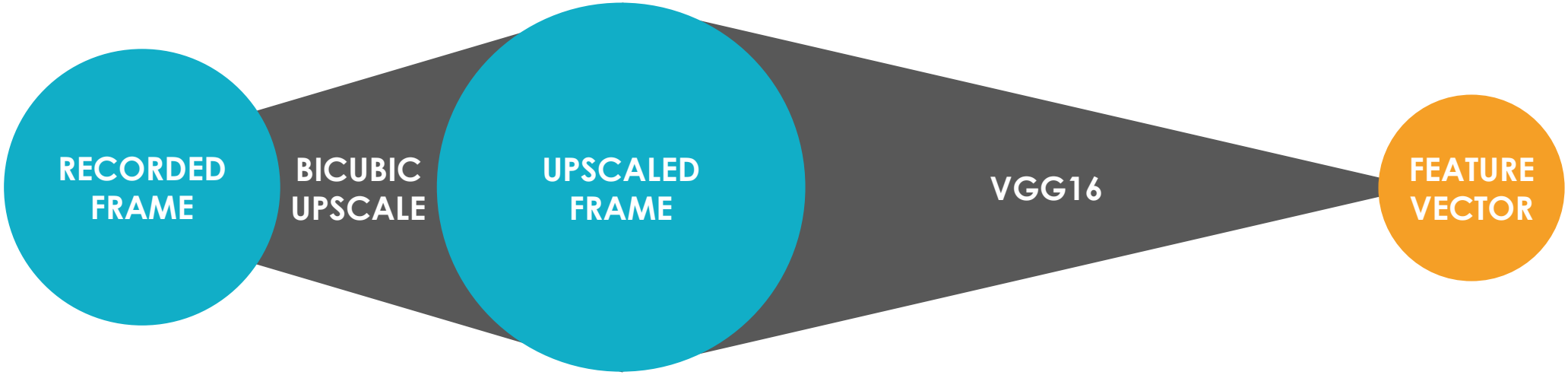
Caption cleaning

Captions are formatted as sentences of single letter words.

Captions are indexed to be fed to the pipeline.

1	0	n u m e r o
2	1	o r d i n e
3	2	d e n a r o
4	3	t i t o l o
5	4	f a t i c a
6	5	l i m i t e
7	6	d o l o r e
8	7	e r r o r e
9	8	a t t i m o
10	9	c u c i n a

Frame Precomputing



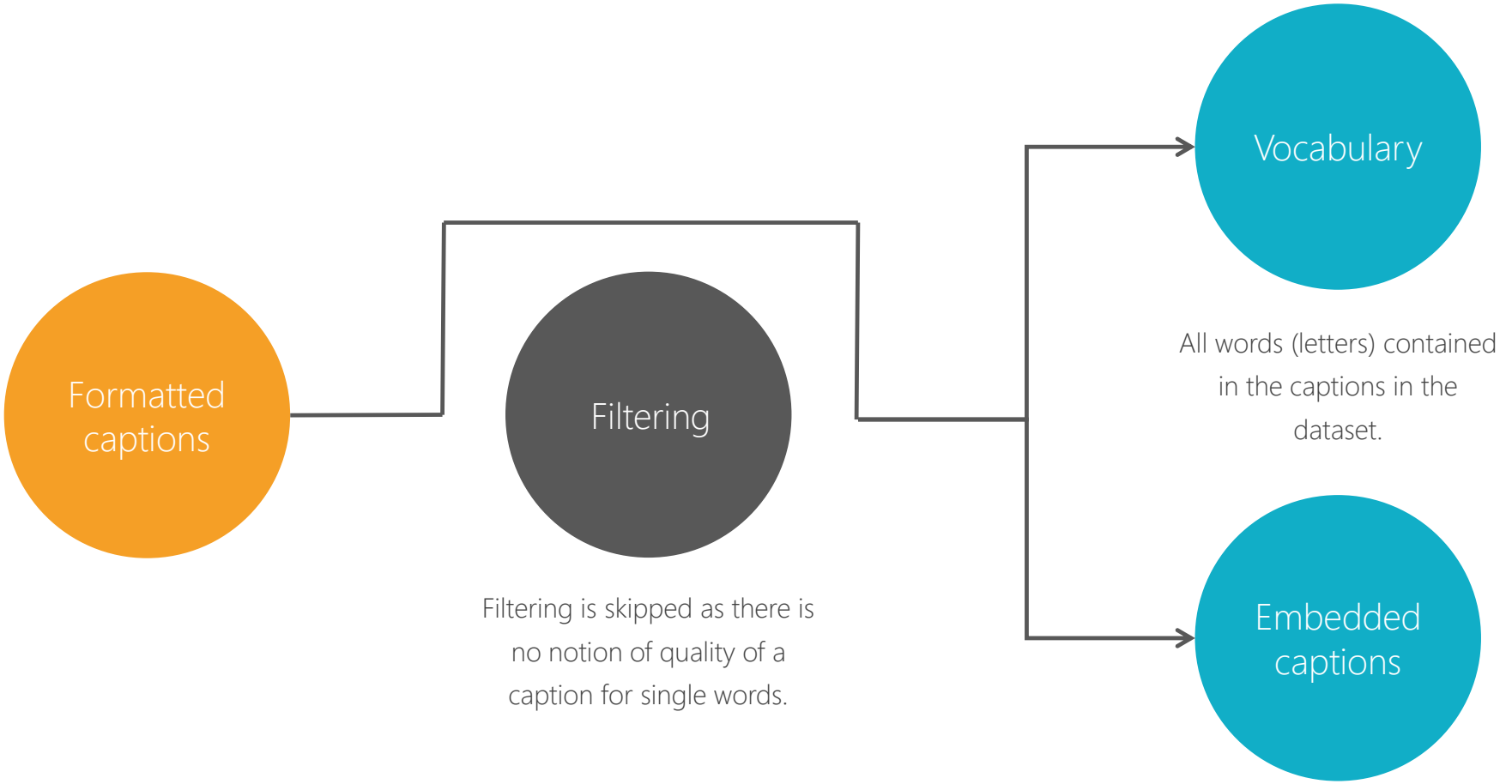
Original recorded frame (100x100)

Recorded frame upscaled to 224x224 to be fed to VGG16

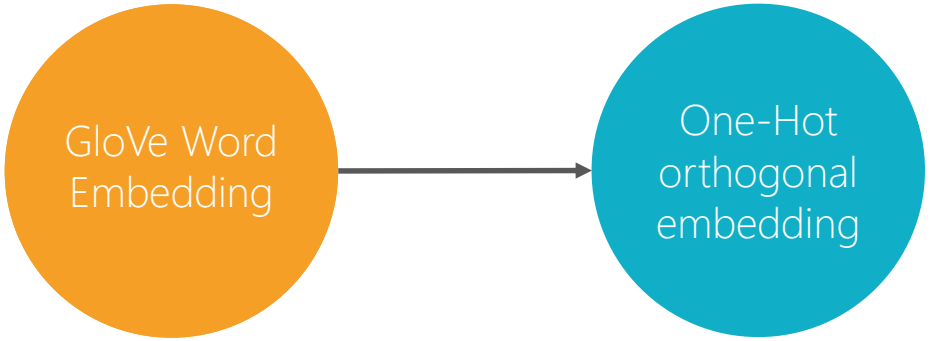
VGG16 model pretrained on ImageNet.

Encoded frame in feature space.

Caption Precomputing



Caption Precomputing



Learned embedding for English words.

Hand coded feature matrix.

Orthogonal embedding

Letters do not carry meaning; hence they are all orthogonal to each other.

Keeping an embedding matrix allows to leave the pipeline unchanged

Training

Hardware

GPU Training for main model on 1 GTX 1080

CPU for dataset acquisition and processing Ryzen 7 2700x

Parameters

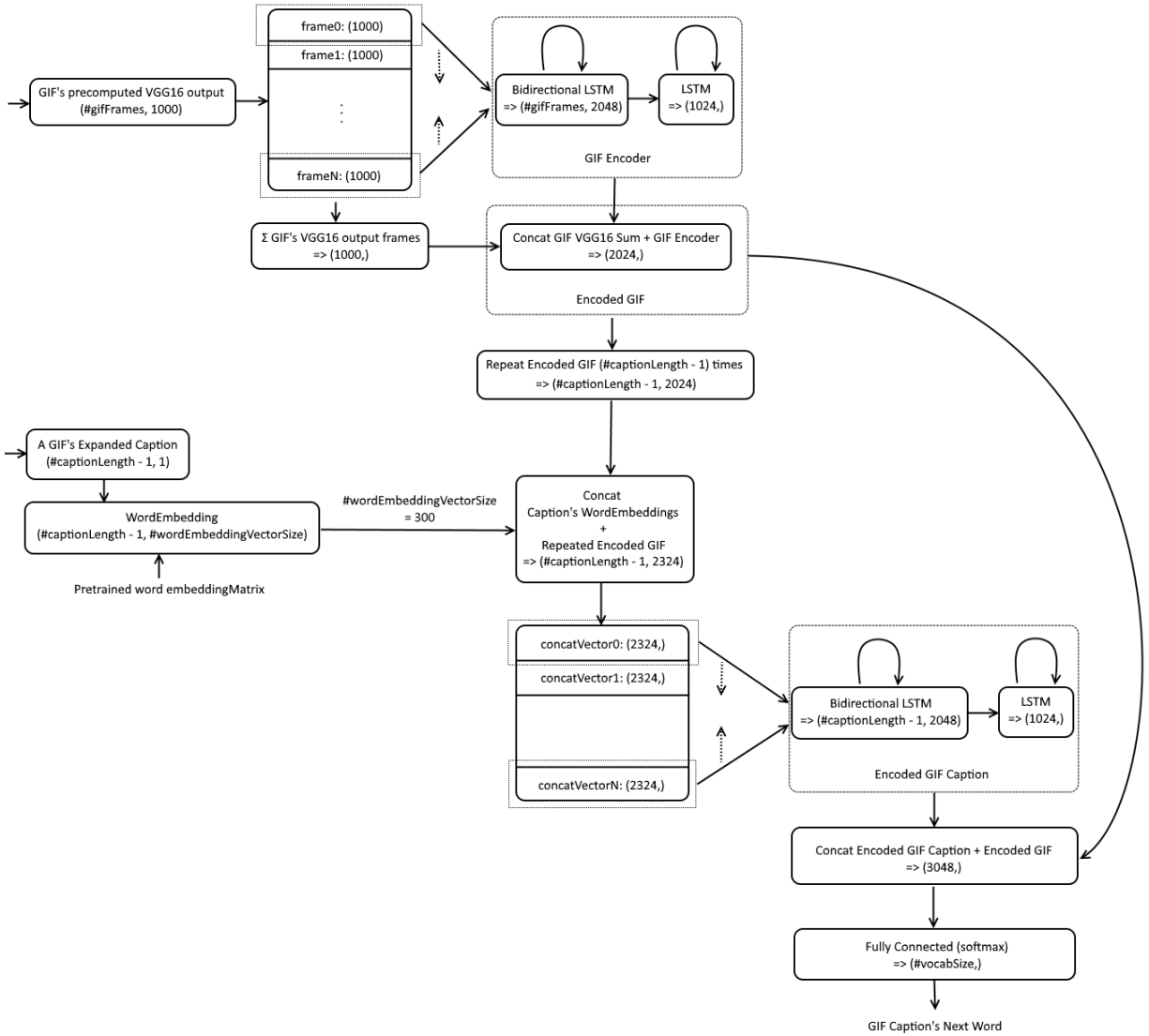
100 epochs using sparse categorical cross entropy as loss

Model Overview

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 16, 1000)	0	
dropout_1 (Dropout)	(None, 16, 1000)	0	input_1[0][0]
bidirectional_1 (Bidirectional)	(None, 16, 2048)	16588800	dropout_1[0][0]
dropout_2 (Dropout)	(None, 16, 2048)	0	bidirectional_1[0][0]
lstm_2 (LSTM)	(None, 1024)	12587008	dropout_2[0][0]
lambda_1 (Lambda)	(None, 1000)	0	input_1[0][0]
input_2 (InputLayer)	(None, 15)	0	
concatenate_1 (Concatenate)	(None, 2024)	0	lstm_2[0][0] lambda_1[0][0]
embedding_1 (Embedding)	(None, 15, 20)	420	input_2[0][0]
repeat_vector_1 (RepeatVector)	(None, 15, 2024)	0	concatenate_1[0][0]
concatenate_2 (Concatenate)	(None, 15, 2044)	0	embedding_1[0][0] repeat_vector_1[0][0]
dropout_3 (Dropout)	(None, 15, 2044)	0	concatenate_2[0][0]
bidirectional_2 (Bidirectional)	(None, 15, 2048)	25141248	dropout_3[0][0]
dropout_4 (Dropout)	(None, 15, 2048)	0	bidirectional_2[0][0]
lstm_4 (LSTM)	(None, 1024)	12587008	dropout_4[0][0]
concatenate_3 (Concatenate)	(None, 3048)	0	lstm_4[0][0] concatenate_1[0][0]
dropout_5 (Dropout)	(None, 3048)	0	concatenate_3[0][0]
dense_1 (Dense)	(None, 21)	64029	dropout_5[0][0]

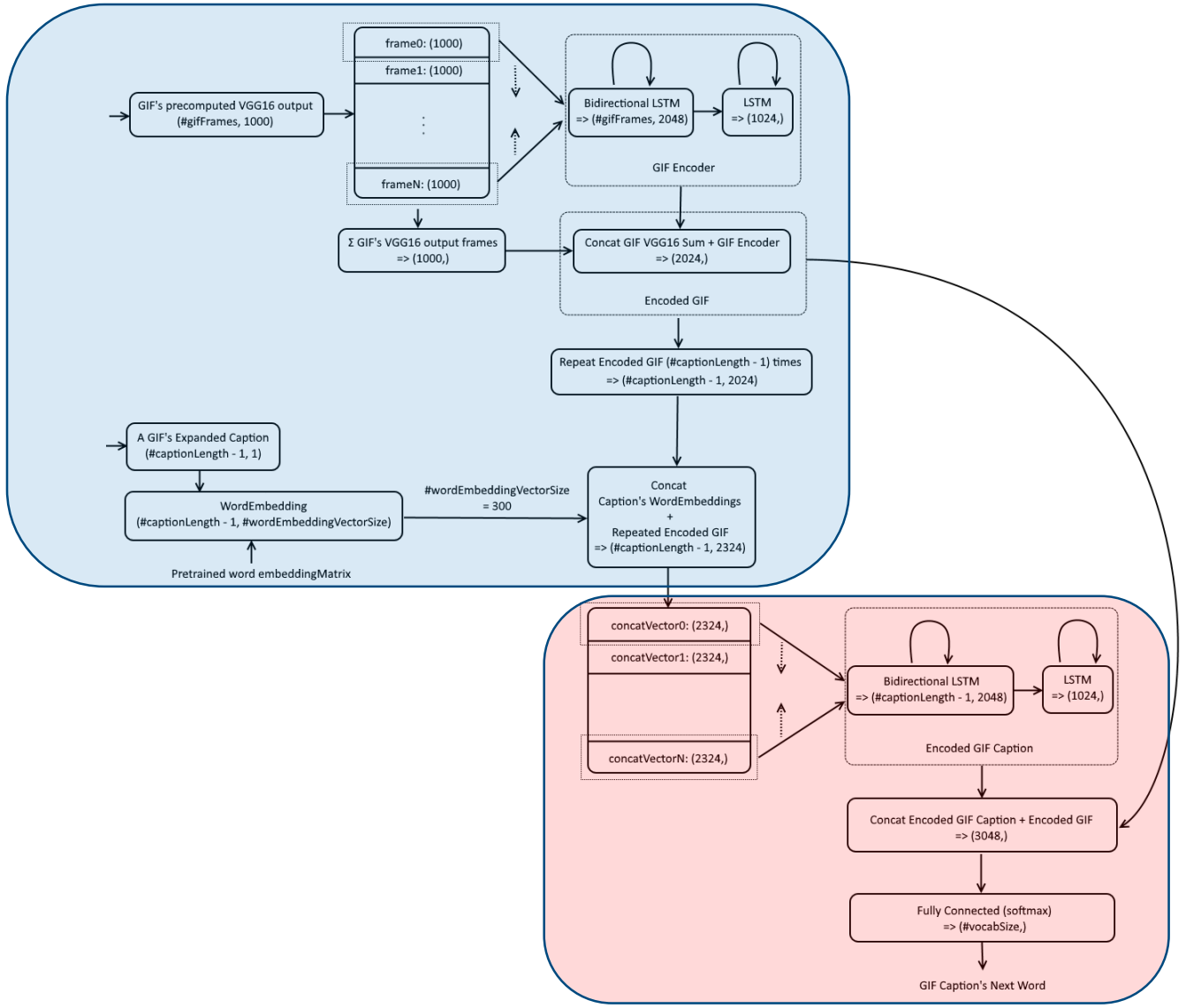
Total params: 66,968,513
Trainable params: 66,968,093
Non-trainable params: 420

Model Overview



Model Overview

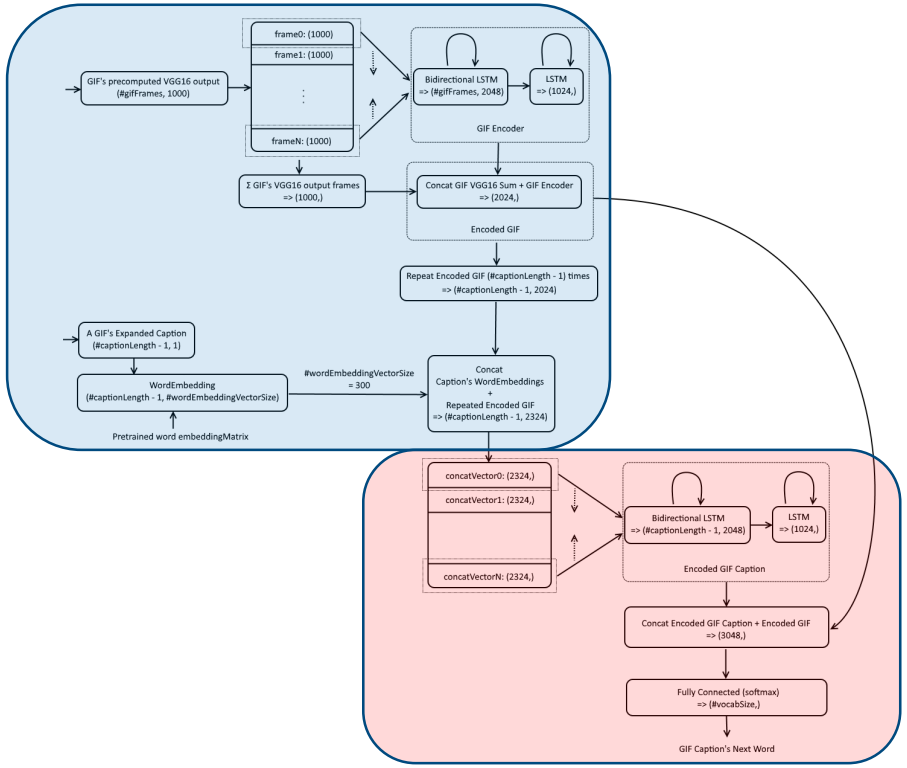
Encoding



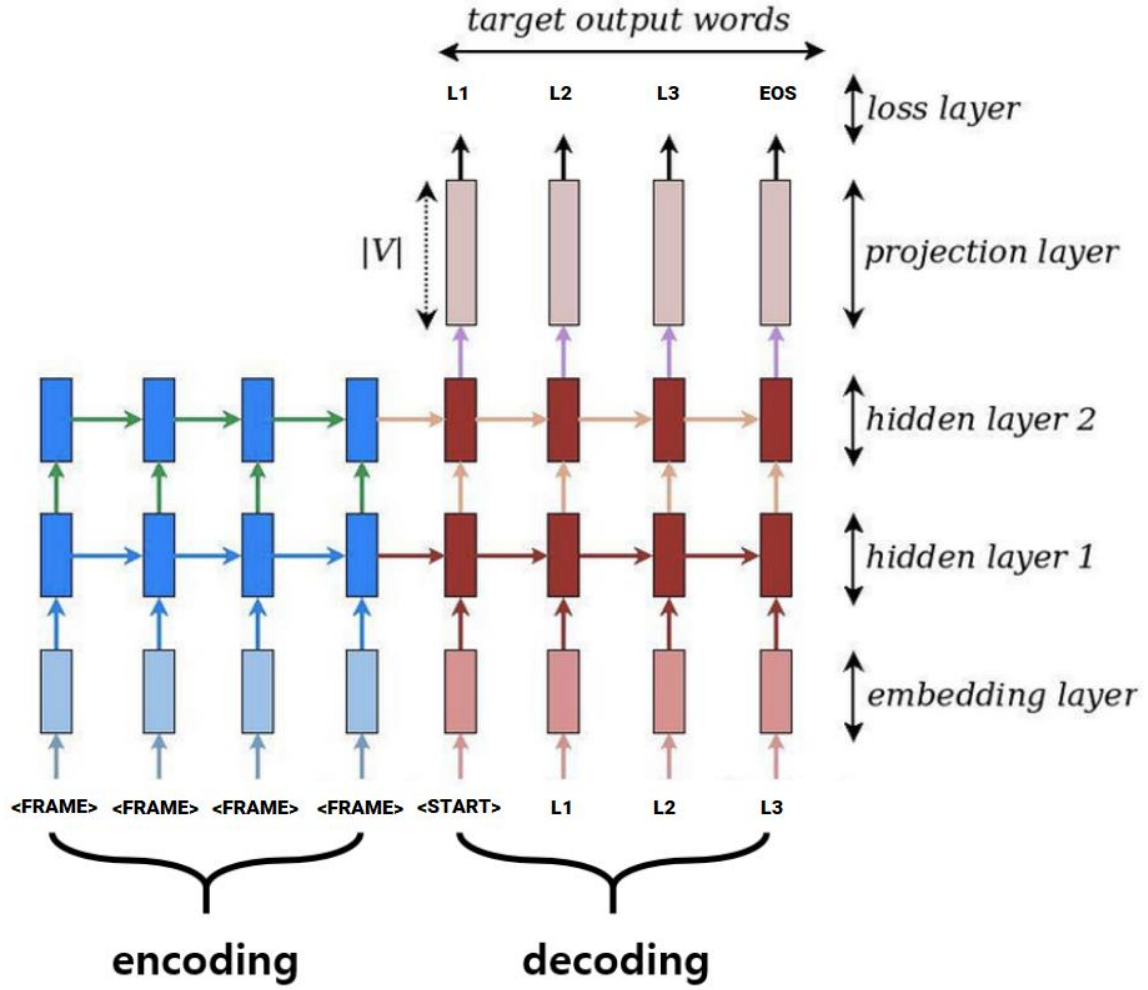
Decoding

Model Overview

Encoding



Decoding



Dataset Considerations

DICTIONARY SIZE

The dictionary is relatively small, this can impede generalization on unseen words.

SPEAKER COUNT

A dataset with multiple speakers largely increases the variance, hence we chose to discard the data and keep only one speaker.

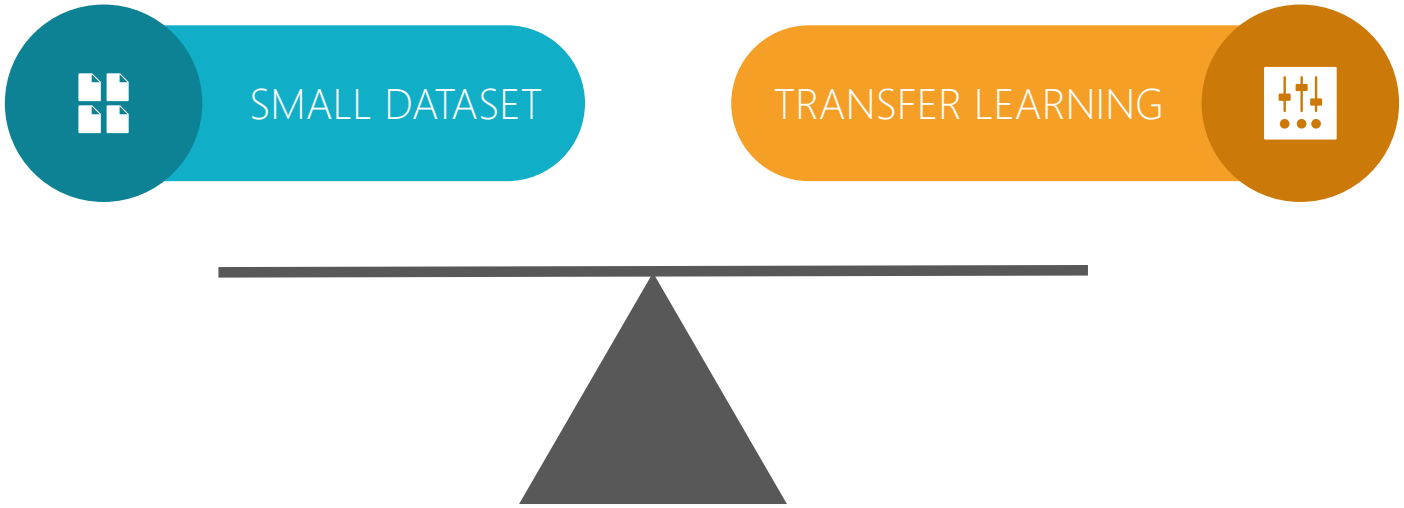
ANGLE AND LIGHT CONDITIONS

The dataset has been obtained in a controlled set, reducing the need for generalization on environment aspects.

DOMAIN ADAPTATION

The dataset was acquired using the same camera to have a consistent image condition on all the samples.

Dataset Considerations



Transfer Learning compensates for a Small Dataset

Advantages

Not as many samples are required to learn the same sequences.

Smaller dataset implies smaller training time.

Disadvantages

Pipeline complexity increases as frames must be precomputed.

More difficult to label unseen data.

Dictionary and Recordings

Words

Selection of 100 words from the 1000 most common words in the Italian language.
Words with roughly 3 syllables (6-8 letters long).

Pronunciation

Relatively fast speaking at 0.1s/letter
Average of 7 frames per word, with a target framerate of 10fps.

Recording device

All speakers use the same recording device to increase consistency throughout all the dataset and decrease the need for domain adaptation.

Issues

Generalization of recording conditions

The model works very well in standard conditions; but it is very sensitive to changes in settings and speaker.

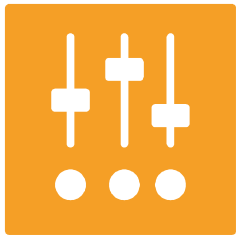
Generalization of words

The model works well only on already seen words. The model still generates unseen letter sequences, but these almost never make sense.

Performance and Results

Validation methods

Unseen dataset of 100 samples, one for each word in the dictionary, recorded in the same setting as the dataset.

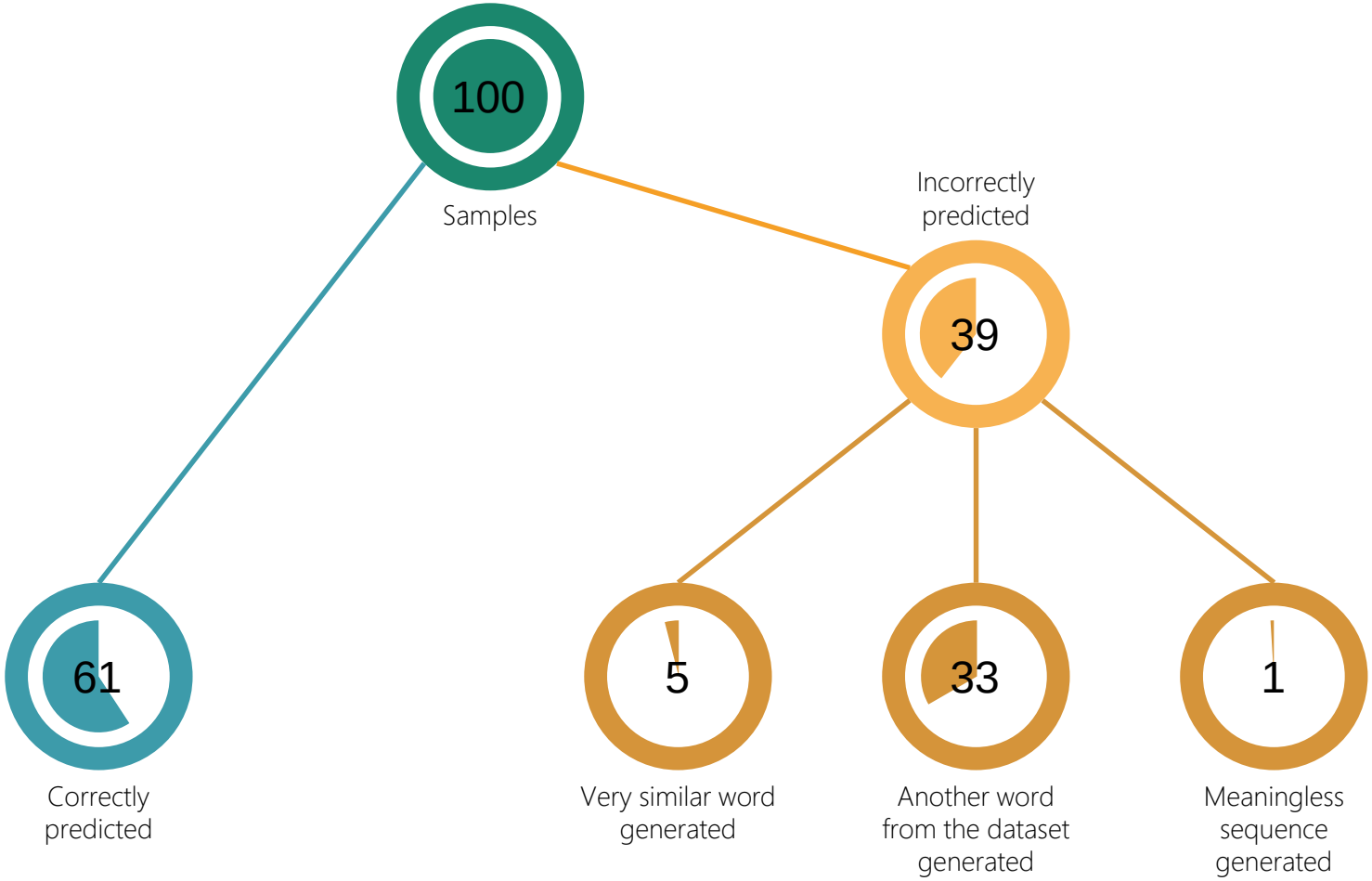


Validation results

Ita-Lip correctly predicted 61 of the 100 samples.

Some of the wrongly predicted samples only differ for one letter from their target.

Performance and Results

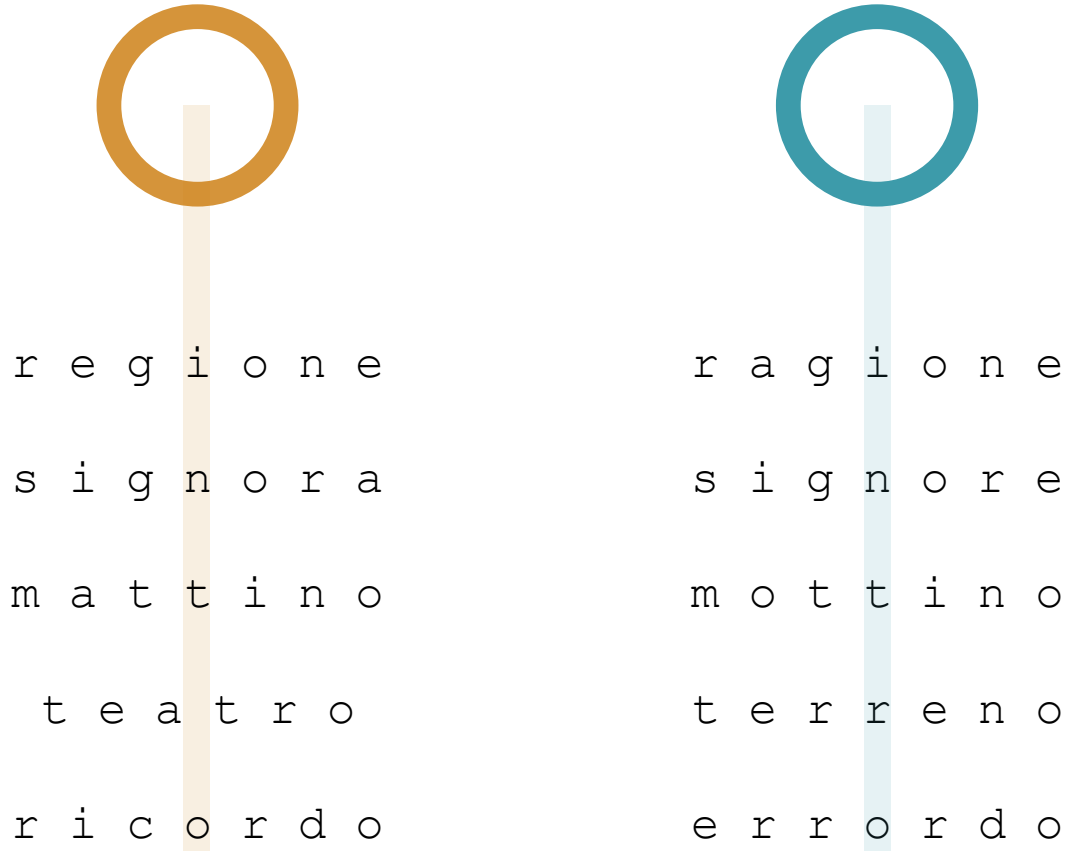


Performance and Results

Highlights

The model mistakenly produced unseen words.

The model classified some words as their similar counterpart.



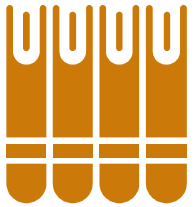
Conclusions

Performances

Accuracy of 61% on unseen words inside the dictionary.
Better performances compared to humans.
Worse compared to SOTA.

Methods

Use of Transfer Learning and seq2seq encoder-decoder framework without attention.



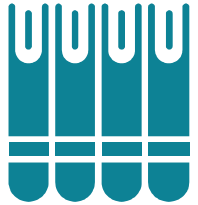
Contributions

First approach to lipreading Italian.
Detailed explanation of the adapted, slightly modified version of GCNet - C. Caruso

Future Work

Prediction of Words Outside the Dictionary

Need of a bigger dataset and larger dictionary.

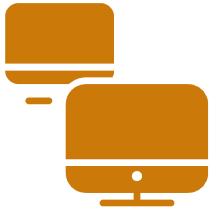


Sentences, not Words

Shifting the focus to sentences allows for the use of context, leading to the use of attention in seq2seq.

Expansion of the Captured Area

Lipreading entails attentive observation of the entire facial expression.



Thank You

References

- D. AMODEI, R. ANUBHAI, E. BATTENBERG, C. CASE, J. CASPER, ET AL., *Deep Speech 2: Endto-End Speech Recognition in English and Mandarin*, 2015. arXiv:1512.02595.
- Y. M. ASSAEL, B. SHILLINGFORD, S. WHITESON, AND N. DE FREITAS, *LipNet: End-to-End Sentencelevel Lipreading*, 2016. arXiv:1611.01599.
- C. CARUSO, *GCNet (GIF Caption Network) | Neural Network Generated GIF Captions*, 2016. https://github.com/chcaru/gcnet, appropriately modified for our purposes.
- P. CHANSUNG, *Seq2Seq Model in TensorFlow*, 2018. https://towardsdatascience.com/seq2seq-model-in-tensorflow-ec0c557e560f, appropriately modified for our purposes.
- J. S. CHUNG, A. SENIOR, O. VINYALS, AND A. ZISSERMAN, *Lip Reading Sentences in the Wild*, 2016. arXiv:1611.05358.
- M. COOKE, J. BARKER, S. CUNNINGHAM, AND X. SHAO, *An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition*, 2006. The Journal of the Acoustical Society of America, 120(5):2421–2424, http://spandh.dcs.shef.ac.uk/gridcorpus/.
- R. D. EASTON AND M. BASALA, *Perceptual Dominance During Lipreading*, 1982. Perception & Psychophysics, 32(6):562–570.
- L. FEI-FEI AND K. LI, *ImageNet*. Stanford University and Princeton University, http://www.image-net.org/.
- S. HILDER, R. HARVEY, AND B.-J. THEOBALD, *Comparison of Human and Machine-based Lipreading*, 2009. In AVSP, pp.86–89.
- J. PENNINGTON, R. SOCHER, AND C. D. MANNING, *Glove: Global vectors for word representation*, in Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
- G. E. SHAMBAUGH, *Inner Ear Disease - Rehabilitation*. Encyclopedia Britannica, https://www.britannica.com/science/ear-disease/Inner-ear#ref65106.
- K. SIMONYAN AND A. ZISSERMAN, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2014. arXiv:1409.1556.
- TELELINEA, *Le 1000 Parole Più Usate in Italiano*. http://telelinea.free.fr/italien/1000_parole.html.