ITA–LIP

LIPREADING ITALIAN WORDS THROUGH TRANSFER LEARNING AND SEQ2SEQ

Jacopo P. Gargano
Computer Science and Engineering
Politecnico di Milano
Milano, Italy
jacopopio.gargano@mail.polimi.it

Loris Giulivi
Computer Science and Engineering
Politecnico di Milano
Milano, Italy
loris.giulivi@mail.polimi.it

September 19, 2019

ABSTRACT

In this work we accomplish the task of lipreading 100 words selected from the 1000 most common words in the Italian language through deep learning. We create our own dataset by recording a person pronouncing words facing a camera. The chosen set of words consists of nouns, adjectives and verbs and can possibly be extended to the whole dictionary.
We use an adapted version of *GCNet* [3] for our purpose, which we explain in detail in section 5. The model we present should be taken as a follow up, simplified version of *LipNet* [2], *Deep Speech 2* [1] and *Lip Reading Sentences in the Wild* [5], where more complex networks are capable of recognizing entire sentences coming from different speakers, possibly with different orientations and in various settings.

***K*eywords** Lipreading · Visual Speech Recognition · Transfer Learning · seq2seq

1. Introduction

Lipreading consists in decoding text through the observation of a person speaking, without hearing what they are saying.
This ability is specifically developed by those who are completely or partially deaf. In fact, the child born deaf or with a severe hearing impairment cannot acquire speech by the normal process but must attend special classes to be taught speech and lipreading [11].
However, lipreading is used even by persons with normal hearing who, in the presence of background noise, need these visual clues to supplement hearing [11].

The value of having a machine able to lipread lies in the possibility of transcribing speech, either working side by side with speech-to-text or on its own when audio is not available. In fact, machine lipreading could be a great support to deaf people in their everyday life. For example, a camera could be installed on glasses or clipped to clothes and it could process speech, which would later be displayed to a smartphone.
Another practical application consists in decoding what a person is saying in the presence of loud background noise. For instance, one could be driving with loud music on and want to ask a virtual assistant to place a call.

The purpose of this work is to develop a model, as an adapted version of *GCNet* [3], that is able to perform lipreading in a controlled environment receiving as input the sequence of preprocessed images making up the word a speaker is pronouncing and outputting the word as text.

There are some existing lipreading models, a few examples of which is cited in the abstract. However, to the best of our knowledge, so far none can lip-read Italian.

Differently from English there are no public Italian datasets that suit our purpose, so we created our own - see section 2. As an example, the dataset with English audiovisual sentences used by *LipNet* is the GRID corpus [6]. It was impossible for us to obtain a dataset large enough to support lipreading from multiple users, so we focused on having a good amount of recordings from one speaker and on the realization of the model. However, we recorded also other speakers in order to test the performance of our model on speakers never seen. Our results are presented in section 6.

1.1 Baseline

The baseline for the results will be the same used in *LipNet* [2] - word error rate (WER) of 47.7% - where hearing impaired people were asked to label 300 random videos.

It must be noted that this result is skewed due to the small dataset used in *LipNet*; in the same paper it is in fact specified that "according to the literature, the accuracy of human lip readers is around 20% [7], [9]".
LipNet exhibits an error of 11.4% when performing on unseen speakers [2]. Note that *LipNet* lipreads sentences, not words, which makes the process easier thanks to the correlation between words in a sentence, that is the context.

2. Dataset Creation

Words Choice
The size of the dictionary of chosen words only depends on data availability, that is the reason why it was chosen to be small: 100 words. Moreover, increasing dictionary size would result in a rise in the model complexity.
We chose 100 words among the "1000 Most Used Words in Italian" [13]. Words are selected so that they vary between 6 and 8 characters in length. Some words, like *signore* and *signora*, are the male and female version of a word, which results in some words being different for only one letter. This is done to test the performance of the model also on very similar words. The complete list of words may be found in appendix A.

Data Collection
First, we collected recordings of single words through a custom made desktop application. Each word is recorded in its entirety through a camera capturing at a target of 10 frames per second. The same camera is used throughout the dataset populating process so as to reduce the need of domain adaptation. The process has the user start with their mouth closed, press a key to start the recording, pronounce the word, close their mouth and finally release the key. The frames are saved together with the needed metadata. The recording application hints to the user via a progress bar how fast the word should be pronounced. The average number of frames for a word happens to be 7, as the software estimates 0.1s for each letter in a word. This value has been selected empirically based on a relatively fast articulation of the word, but is quite variable as syllable pronunciation does not match linearly with letter count. The maximum number of frames is hence set to 16 to allow for slower words but at the same time avoid high complexity when training the model.
The size of the collected dataset is of about 5000 recordings.

Data Transformation
We adapted our dataset to *GCNet* in the following way. *GCNet* generates captions for GIFs, which are made up of a sequence of frames as in our problem; the output is a sentence made of several words. We adapted our dataset by splitting each of our word labels into separate, single-letter words. In this way, a n letters long word becomes a sentence of n words.

3. Embedding Matrix for Labelling

The first step of our work consists in adapting our dictionary to the one used in *GCNet*. In *GCNet*, *GloVe* [10] is used to obtain vector representations for words used in captions. In our case, we use letters instead of words (section 2 - Data Processing).
The word embedding matrix is, in general, a 2-dimensional matrix that assigns a feature vector to each word in the dictionary. Differently than in *GCNet*, in our case words are made of single letters, and being letters independent one from another, we do not need any particular learning algorithm to obtain an embedding matrix. This is because words carry a meaning which is emergent from the letters they are made of, for instance *useful* and *useless*; letters instead do not, and can be treated as orthogonal vectors in the feature space.
Even if word embedding is not needed for our purposes, an embedding matrix is kept to maintain the same pipeline used in *GCNet*, however, the matrix used in our project differs as it is predetermined and not computed. The first row represents the stop symbol and is the zero vector. The rest of it is a diagonal matrix in a one-hot encoding fashion. Finally, all samples are joined in a matrix where each row represents a word as an array of numbers, each representing a letter (token index). The size of this matrix is $n \times 16$, where n is the size of the dataset.

4. Embedding Frames Through Transfer Learning on VGG16

The model used by *GCNet* is *VGG16* [12] which is pre-trained on *ImageNet* [8]. Frames are transformed into 224x224 pixels images through bicubic interpolation, without loss of information, so as to be fed to *VGG16*.
We use transfer learning on *VGG16*: we keep the knowledge of *VGG16* in recognizing the features of the images from *ImageNet* to learn the features of the frames making up our words, thus obtaining a $16 \times n \times 1000$ matrix. Here 16 is the maximum number of frames that make up each word, n is the number of samples in the dataset and 1000 is the number of features set in *VGG16*.
The result of this step is therefore a 3-dimensional matrix composed of all embedded word frames, padded to match the maximum sequence length of 16.

5. Adapted GCNet Model

Training Phase
Once the frames are precomputed through *VGG16* and the embedding matrix is generated, these are fed as inputs to Ita-Lip, our model - which is an adapted version of *GCNet* - and is shown in Figure 2, appendix B.

In the model scheme, in the top left, the model receives the precomputed *VGG16* output, that is a sequence of 16 frames representing a word. Each frame is represented by a combination of 1000 features. These go through an

LSTM, referred to as the GIF Encoder in *GCNet*. For each frame that goes through it the LSTM updates a vector representing the word, finally outputting an encoding of the sequence of frames.
In the bottom left, the model is fed with the label of each word, which is a vector of length 16 containing the letters that make up the word. The embedding matrix previously created is then used to obtain a word embedding.
This process can be observed in the encoding part - colored in blue - of the *seq2seq* training phase in Figure 1.

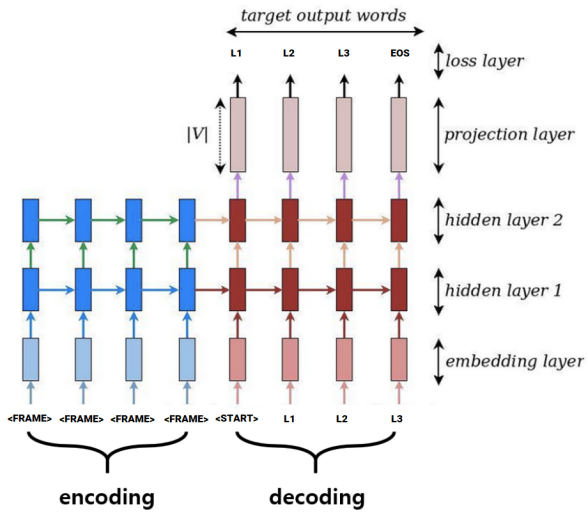

Figure 1: *seq2seq* Training Phase [4]

The encoded sequence of frames is combined with the caption embedding generating a vector of data. Each one of the 16 components is made up by all the embedded frames and the corresponding embedded letter. The vector is fed into another LSTM that decodes it, component by component, in an online supervised learning fashion through letter classification: the LSTM trains its weights to output the letters of the word.
This process can be observed in the decoding part - colored in red - of the general *seq2seq* scheme in Figure 1.

Testing Phase
The testing phase is performed similarly to the training phase with the exception that the captions are not provided to the model. For each step of the decoding process, the input of the decoder is the output of the previous decoding step, instead of the next target letter. The output of the model could be a sequence of letters making up a word outside our dictionary. The benefits of this consequence are explained in section 7.

6. Performance Evaluation

Performance has been evaluated on an unseen validation set made up of 100 samples, one for each word in the dictionary. The trained model correctly predicted 61 of these, resulting in a 39% WER. Of the wrongly predicted examples, the majority - 33 of them - were recognized as a different word belonging to the dictionary, 5 were similar to the correct one, and 1 was a meaningless sequence of characters.
Out of the 5 similar words, 4 only differed for one letter; for example, the feminine and masculine form of *signore* was wrongly predicted both ways: *signore* was predicted as *signora* and vice versa. One of the almost correctly identified words was *mattino*, which the network interpreted as *mottino*, which is not a word in our dictionary. The word with a meaningless sequence of characters as output was *ricordo*, which was notably predicted as *errordo*. This is not a word in our dictionary, but it shows how some of the structure of the pronunciation is kept even in a wrong prediction.

Comparison with Baseline
The accuracy of Ita-Lip is 61%. Compared to the baseline accuracy of 47.7%, Ita-Lip has a better performance. The accuracy of human lipreaders is around 20% [7], [9], which means our model has a 3x better performance compared to humans. Taking *LipNet* as the machine baseline, the performance of Ita-Lip is worse - 89% against 61% respectively. However, considering the limitations due to the small dataset and the absence of a context, the performances of Ita-Lip are considerable.

7. Conclusions

This work brings contributions to the fields of lipreading, transfer learning and *seq2seq*. Our model shows a good performance on unseen data, with a word error rate (WER) of 39%.

We are confident that with a bigger dataset and with a larger dictionary our model would be able to correctly predict words outside of the dictionary. This would allow our model to use a small subset of the whole Italian dictionary for the training phase and still be able to predict almost all, if not all, the words that make it up.
Despite increasing the variance of the data and consequently the dataset complexity, different settings and various speakers would make the model less biased and more flexible to changes.
It would be interesting to estimate or empirically obtain the size of both the dataset and the dictionary to collect more data and allow for these improvements.

Lipreading is usually performed on sentences, not on single words. Differently from letters, words carry a meaning and differently from words by themselves, those in a sentence are contextualized. Therefore words are not independent and the embedding matrix will be more complex but more useful to the model. Using the *seq2seq* framework with attention would surely increase the model predictive performances.

Moreover, lipreading entails attentive observation of the entire facial expression rather than the movements of the lips alone [11]. In our work we only focused on the speaker's mouth, so it would be interesting to expand the captured area to the entire face and evaluate the model performance in different settings, including the simple one we used: one speaker, constant setting, small dictionary.

Finally, Ita-Lip learned to lipread Italian words in a controlled environment, with performances noticeably better than hearing impaired humans and worse but comparable to the artificial baseline, considering the reduced size of the dataset we used.

References

[1] D. AMODEI, R. ANUBHAI, E. BATTENBERG, C. CASE, J. CASPER, ET AL., *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*, 2015. arXiv:1512.02595.

[2] Y. M. ASSAEL, B. SHILLINGFORD, S. WHITESON, AND N. DE FREITAS, *LipNet: End-to-End Sentence-level Lipreading*, 2016. arXiv:1611.01599.

[3] C. CARUSO, *GCNet (GIF Caption Network) | Neural Network Generated GIF Captions*, 2016. `https://github.com/chcaru/gcnet`, appropriately modified for our purposes.

[4] P. CHANSUNG, *Seq2Seq Model in TensorFlow*, 2018. `https://towardsdatascience.com/seq2seq-model-in-tensorflow-ec0c557e560f`, appropriately modified for our purposes.

[5] J. S. CHUNG, A. SENIOR, O. VINYALS, AND A. ZISSERMAN, *Lip Reading Sentences in the Wild*, 2016. arXiv:1611.05358.

[6] 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/`.

[7] R. D. EASTON AND M. BASALA, *Perceptual Dominance During Lipreading*, 1982. Perception & Psychophysics, 32(6):562–570.

[8] L. FEI-FEI AND K. LI, *ImageNet*. Stanford University and Princeton University, `http://www.image-net.org/`.

[9] S. HILDER, R. HARVEY, AND B.-J. THEOBALD, *Comparison of Human and Machine-based Lip-reading*, 2009. In AVSP, pp.86–89.

[10] 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.

[11] G. E. SHAMBAUGH, *Inner Ear Disease - Rehabilitation*. Encyclopedia Britannica, `https://www.britannica.com/science/ear-disease/Inner-ear#ref65106`.

[12] K. SIMONYAN AND A. ZISSERMAN, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2014. arXiv:1409.1556.

[13] TELELINEA, *Le 1000 Parole Più Usate in Italiano*. `http://telelinea.free.fr/italien/1000_parole.html`.

A. Dictionary

numero, ordine, voglia, verità, marito, popolo, autore, parete, albero, dovere, denaro, titolo, fatica, limite, dolore, errore, attimo, cucina, teatro, secolo, camera, musica, potere, pagina, parola, misura, lavoro, natura, figura, motivo, nemico, diritto, regione, oggetto, lettera, capello, ragazza, termine, esempio, soldato, origine, polizia, ragazzo, ufficio, società, bisogno, bambino, potenza, momento, ritorno, sorriso, ragione, operaio, sistema, cortile, signore, cultura, signora, persona, governo, inverno, nazione, albergo, mattino, maniera, dottore, materia, passato, sorella, effetto, periodo, ricerca, animale, scienza, memoria, terreno, destino, ricordo, generale, famiglia, stazione, elemento, speranza, funzione, distanza, finestra, malattia, compagno, sviluppo, articolo, giudizio, processo, contatto, ambiente, proposta, corrente, opinione, politica, giardino, rispetto.

B. Ita-Lip Overview

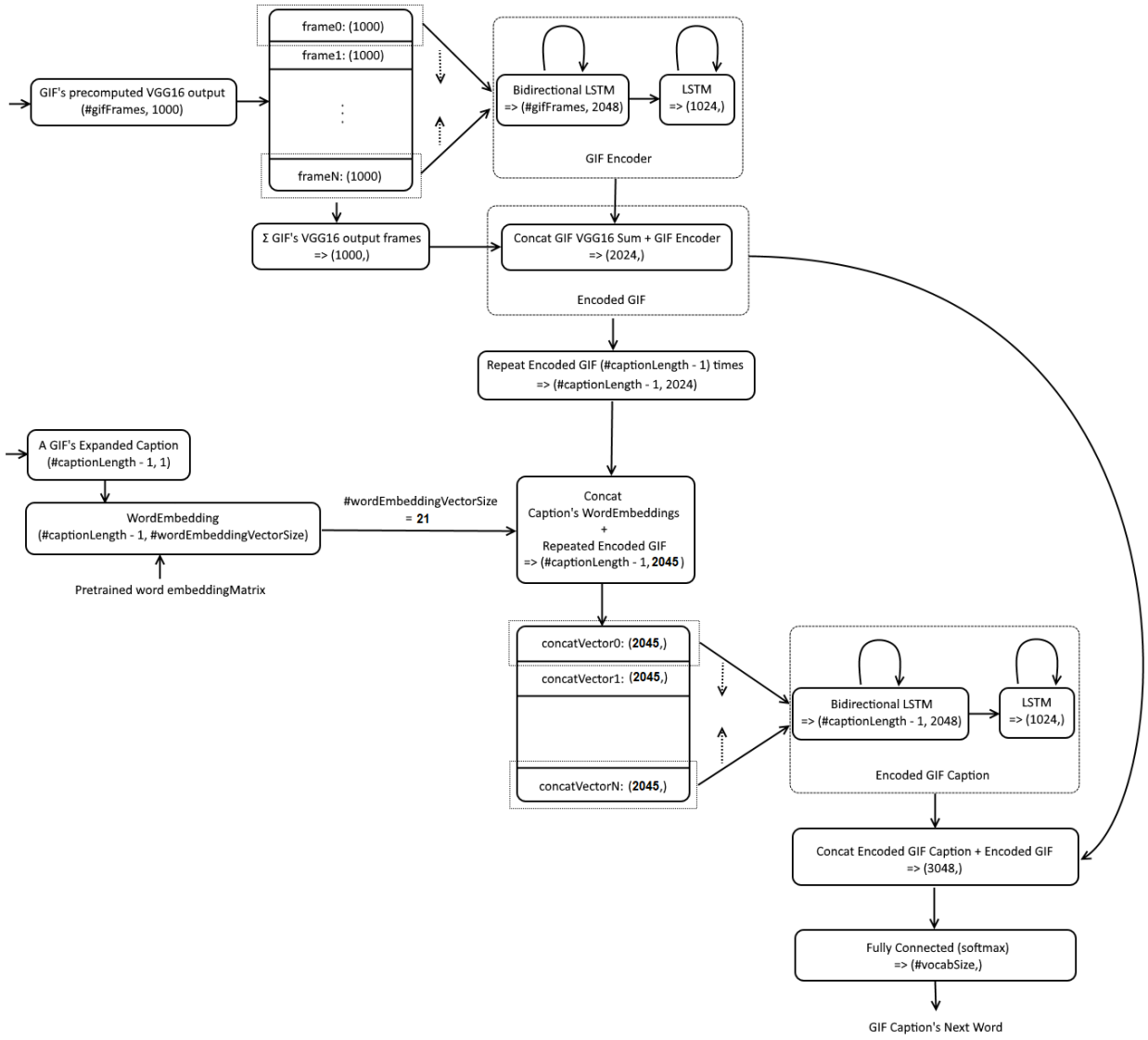

Figure 2: Ita-Lip Overview - Modified Overview of GCNet [3]