BIGLIP: A PIPELINE FOR BUILDING DATA SETS FOR LIP-READING

Umar Jamil

University of Leeds, Leeds, UK

ABSTRACT

Lip-reading, the process of deciphering text from visual mouth movements, has garnered significant research attention. While numerous data sets exist for training lip-reading models, their coverage of diverse languages remains limited. In this paper, we introduce an innovative pipeline for constructing data sets tailored to lipreading models, leveraging web-based videos. Notably, this pipeline is the first of its kind to be made publicly available. By employing this pipeline, we successfully compiled a data set comprising Italian videos—a previously unexplored language for lipreading research. Subsequently, we utilized this data set to train two lip-reading models, thereby highlighting the strengths and weaknesses of employing wild-sourced videos (e.g., from YouTube) for lip-reading model training. The proposed pipeline encompasses modules for audio-video synchronization, audio transcription, alignment, cleaning, and facilitates the creation of extensive training data with minimal supervision. By presenting this pipeline, we aim to encourage further advancements in lip-reading research, specifically in the domain of multilingual data sets, thus fostering more comprehensive and inclusive lip-reading models.

KEYWORDS

Deep Learning, Lip Reading, Visual Speech Recognition, Datasets

1. INTRODUCTION

Lip reading, also known as speech-reading, is the process of interpreting spoken language by observing the movements of a speaker's lips, tongue, and jaw. Lip reading can be an important tool for individuals who are deaf or hard of hearing, as well as in noisy environments where hearing speech may be difficult. However, lip reading is a challenging task that requires a high level of visual acuity and linguistic knowledge.

While in ASR (automated speech recognition) the smallest unit of sound in a language that serves to distinguish words from another is called phoneme (with the English language having 44 of them), the number of visemes, that is, the visually distinctive units, is much smaller. This is because many phonemes are produced within the mouth and the throat and, for example, the letters P and B are invisible to the observer. This makes Lip-reading a hard task even for a human specialist.

Lip-reading, apart from being used by people with hearing difficulties, can be applied to other scenarios:

- Healthcare: lip-reading can be used to read the patient’s lip when the patient is unable to produce sounds.

David C. Wyld et al. (Eds): NET, DNLP, MLIOB -2024
pp. 17-30, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140302

- Forensics: lip-reading can be used to reconstruct the dialogues in a footage where the audio has been lost or it is noisy.
- Automated Speech Recognition [1]: automakers can integrate lip-reading systems to complement their ASR model in order to understand commands (for example “turn on the A/C”) from the driver or the passengers in smart cars when the music’s volume is too high. Lip-reading is also necessary in this case to recognize the active speaker in the scene.
- Film and television [2]: lip-reading is used in the film and television industry to synchronize the dialogue with the actors' lip movements, especially when dubbing foreign language films.
- Sports broadcasting: in sports broadcasting, lip-reading can be used to analyze the tactics and strategies of players and coaches, as well as to commentate on the game.
- Security and surveillance: lip-reading is used to reconstruct the speech of individuals that are under surveillance, especially in environments that make it impossible to listen (a train station or an airport).

The main contributions of this research project are the following:

- We developed BigLip, an open source (https://github.com/hkproj/ai-project) pipeline for building data sets for lip-reading models.
- We used BigLip to build ItaLip, a data set made of Italian videos for training lip-reading models. The videos are taken from the web. To our knowledge, it is the first data set for the Italian language.
- We developed data augmentation techniques, explored below in the *transcription cleaning* section, specifically made for the lip reading task that allow to better utilize small data sets.

2. BACKGROUND

Automatic Lip Reading, sometimes also called Machine Lip Reading, has undergone extensive research and development in the recent two decades, driven also by the progress in the field of deep learning.

In the 1950s, Sumby, W.H. [3] was among the first to suggest that lip movements during human speech could be utilized to gather information, marking the inception of lip reading as a research field. This introduction of the concept of lip reading initiated a new phase of study in this domain. Petajan [4] proposes an automatic lip-reading system to enhance the performance of speech recognition. Traditional systems first extract features from a video and then classify them, using models such as SVM (Support Vector Machines) [5] and HMM [6] [7] (Hidden Markov Models). In recent years, deep learning has made significant progress in the field of computer vision. Several approaches to lip reading with deep learning have been proposed, including both 2D and 3D convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models that combine CNNs and RNNs [8] [9] [10]. The combination of convolutions and recurrent networks have proven to be successful [11] [12] [13].

The reason behind the success of CNNs is due to multiple factors:

- CNNs are designed to automatically extract spatially localized features from images. This is achieved through the use of convolution layers, which employ small filters to convolve across the input image, capturing patterns and features in a local neighbourhood. This ability allows CNNs to capture both low-level features like edges and corners, as well as higher-level features such as textures and shapes, which are crucial for image understanding and classification.

- CNNs leverage parameter sharing and sparse connectivity to efficiently process images. In convolution layers, the same filter weights are shared across different spatial locations, enabling the network to learn and recognize similar patterns regardless of their position in the image.
- Deep architectures: CNNs are typically deep networks [14] [15], consisting of multiple stacked convolution and fully connected layers. The deep architecture allows them to learn complex and abstract representations of images, gradually capturing high-level semantics.
- Transfer learning: CNNs trained on large-scale data sets, such as ImageNet [14] have demonstrated excellent generalization capabilities. This property enables the use of transfer learning, where pre-trained CNN models can be fine-tuned on smaller, specialized data sets. By leveraging the learned features from the large data set, the model can quickly adapt to the new image classification task, even with limited labeled data. For example [16] employs a pre-trained VGG [15] network.

Lip-reading has also been used to complement ASR systems [12] [17].
Recent models [16] also make use of the Transformer [18] [19] [20] model to align the video input with the output sentence. This allows the model to be trained on non-aligned video-text pairs and lets the Transformer model automatically learn the alignment through its attention mechanism [18]. Alignment can also be avoided using loss functions like the CTC (Connectionist Temporal Classification) [21] that, however, need to explore all possible alignment paths between the input and the output sequence and employ a dynamic programming algorithm to explore them.

2.1. Feature extraction methods

Pu et al [22] classifies the lip feature extraction methods into three categories.
Pixel-based method: the pixel-based approach extracts features by utilizing all the pixels in the image, including those in the lip region, as the original feature. This method involves using a series of pre-processing techniques to reduce the dimension of the original feature, resulting in a specific expressive feature.
Shape-based method: instead of using the raw pixels of the image, a preprocessing phase is added to extract geometrical features from the image, for example the height, opening, angle of the lip and the area surrounding it.
Methods based on a hybrid model: The hybrid model is a method of lip feature extraction that combines both pixel-based and shape-based techniques. This approach takes advantage of the strengths of both the pixel-based method, which uses all the pixels in the lip region, and the shape-based method, which establishes a lip contour model and extracts visual features based on geometric and contour features.

2.2. Data sets

Multiple data sets have been developed to train lip-reading and audio-visual speech recognition models, among them we have:
OuluVS [23]: It consists of 10 daily-use English phrases uttered by 20 speakers (17 male and 3 female). Each utterance was repeated by a speaker up to nine times. Videos were recorded at 25 fps with a resolution of 720 × 576 pixels.
OuluVS2 [24]: OuluVS2 is a multi-view audiovisual database for non-rigid mouth motion analysis, composed of more than 50 speakers, with European, Chinese, Indian/Pakistani, Arabian and African ethnicities, uttering three types of utterances and thousands of videos recorded by six cameras from five different views spanned between the frontal and profile views.

AVICAR [25]: The authors present a description of a large audio-visual speech corpus that was recorded in a car environment. The corpus was created using a multi-sensory array of eight microphones on the sun visor and four video cameras on the dashboard. It includes speakers of various language backgrounds, with 50 male and 50 female speakers, and consists of four categories of speech: isolated digits, isolated letters, phone numbers, and sentences, all in English. To provide a range of signal-to-noise ratios, each script was recorded in five different noise conditions, including idling, driving at 35 mph with windows open and closed, and driving at 55 mph with windows open and closed.

GRID [26]: GRID is a collection of audiovisual recordings of 34 talkers (18 male, 16 female) speaking 1000 sentences each, in order to aid computational-behavioral studies in speech perception. The corpus is composed of high-quality audio and video recordings, featuring the speakers' facial expressions. The sentences follow a specific structure, such as "put red at G9 now." The corpus and its transcriptions are available free of charge for research purposes. Each talker's audio, video, and word transcriptions are available separately. The video files come in two formats: normal quality (360x288; ~1kbit/s) and high quality (720x576; ~6kbit/s).

LRW [27]: Lip Reading in the Wild, one of the most popular data set for lip-reading, proposed by Chung and Zisserman in 2016. The data set is built by extracting thousands of hours of BBC TV recordings, covering an extensive vocabulary of thousands of words with more than 1 million instances, and over 1000 different speakers. The data set is composed of single words, which may be a limitation for sentence level lip-reading.

LRS [28]: Lip Reading Sentences is another data set by Chung et al which is built by processing recordings of BBC TV programs. In particular, this data set is composed of over 100,000 sentences and nearly 5,000 hours of videos. The sentences are separated by full stops, commas and question marks; and are clipped to 100 characters or 10 seconds.

LRW-1000 [29]: Introduced in 2019, it is one of the few large Mandarin language corpora. It is composed of 1,000 words with 718,018 instances from over 2,000 individual speakers. The high-definition videos are 1920 pixels × 1080 pixels, and the standard-definition videos are 1024 pixels × 576 pixels. The collection process considered various conditions such as speaker posture, age, gender, and lighting. This corpus focuses on lip movements rather than an audiovisual cross-corpus and is valuable for studying Chinese lip pronunciation on a large scale.

Lip2Wav [30]: While most of the other models convert lip movements into text, Lip2Wav converts lip movements into audio, which can be useful for compensating audio loss during video calls or audio loss in a footage. To create the Lip2Wav data set, the authors collected a total of about 120 hours of talking face videos across 5 speakers. The speakers are from various online lecture series and chess analysis videos. The data set consists of only English. It has about 20 hours of natural speech per speaker and vocabulary sizes over 5000 words.

LRWR [31]: The paper introduces a new benchmark data set for lip-reading called LRWR, which focuses on the Russian language. The data set includes 235 classes and 135 speakers, making it a large-scale benchmark. The authors compare the performance of popular lip reading methods on the LRWR data set and analyze the results to highlight the differences between benchmarked languages. Their findings demonstrate promising directions for fine-tuning lip reading models and lead to new state-of-the-art results on the LRW benchmark. The LRWR data set is expected to be useful for advancing research in lip-reading beyond the English and Chinese languages.

3. DESIGN AND METHODOLOGY

3.1. The BigLip pipeline

The main objective of this project is to build an open source pipeline that allows to build data sets of videos taken from the web which can then be used to train lip-reading models. We have used

the pipeline to build a data set called It aLip, that consists of videos in Italian taken from the web. To our knowledge, there is no other data set for the Italian language. The primary source for the videos is the popular video sharing platform YouTube. Since it is not possible to share videos downloaded from YouTube without the company's or the video owner's permissions, the videos will not be shared, but rather, links to the videos and the timestamps to crop will be provided. Researchers can use the pipeline to download their own copy of the videos for further processing.

3.2. Video selection

The videos have been selected to bring as much variety as possible into the data. This includes videos of males and females, videos of people with different skin tones and backgrounds (the landscape surrounding the main speaker). Videos in which there is primarily one speaker whose face is fully visible have been preferred, as our goal is to catch the lip area in the video and map it to the corresponding text. Educational videos, self-help videos and videos in which there is one active speaker telling a story are the most suitable for this task. Even with this rigid selection there are some challenges:

1. The face of the active speaker is not always facing the camera, but just like in natural conversations, the face may be facing left, right, down or up from time to time.
2. It is not uncommon for Italian men to grow a beard or mustache. This can sometimes partially cover the lip.
3. Italian people are known for integrating hand gestures in their speech, and more often than not, hands come in between the speaker's face and the camera, covering the lip area partially or fully.
4. All the videos have different lighting conditions and the quality of the video (pixel density) is also not homogeneous, as authors may have used different cameras and settings.

All of the above-mentioned challenges augment the data variety, and for this reason, no action has been taken to avoid or remove these details from the videos. The idea is to include as many videos as possible to provide the data set with a variety taken from real videos instead of employing data augmentation techniques.
Moreover, there are also challenges in the spoken language:

1. Most speakers employ terms borrowed from the English language (so-called *loan words*), for example the word *online*, *training*, *web*, etc but also proper names like *YouTube*, *TikTok*, *Wikipedia*, etc. Since a lip-reading model is concerned with the sequence of visemes each word is mapped to, non-Italian words that are used in daily life will be kept as is.
2. Words with accents, whose pronunciation depends on the type of accent, will also be kept without any further post-processing. For example, in the Italian language the words cantò and *canto* are two conjugations of the same verb *cantare* (*to sing*); they are written similarly but pronounced differently. However, there is not much difference in the visemes, that is, the configuration of the lip area when pronouncing these two variants. The model should learn to select the right word based on the context of the frame sequence. This is a clear indication that the model must learn to handle long dependencies.

3.3. Pipeline

3.3.1. General overview

A lip-reading model takes as input a sequence of images representing the lip area of a speaker and produces a text. No audio is ever sent to the model, which must learn to read the lip area only. To produce the data set given a set of videos taken from the web, the first operation is to extract clips from the video in which the face of the active speaker is visible. Secondly, the audio needs to be transcribed and the text aligned to the audio/video sequence: for each word, there should be two timestamps indicating when that word starts being pronounced and when it ends.
Given a video in which the active speaker is fully visible and its aligned transcript, it is possible to generate shorter videos of a few words each as data samples for training a deep learning model.

Listed below are the steps that compose the pipeline.

3.3.2. Video download

There are many tools available online to download videos from video sharing platform YouTube. We used *youtube-dl*, which allows to download videos in the chosen format and with the chosen quality. We used the following argument for the *format* parameter:

```
youtube-dl [...] -f "best[ext=mp4]" [...]
```

This tells the tool to download the best available format that uses the mp4 extension. Different videos may have a different pixel density.

3.3.3. Video Synchronization

Each video is evaluated using a custom SyncNet pipeline (https://github.com/hkproj/syncnet_pipeline), based on the work by Chung et al.[32], to determine if the audio is synchronized with the active speaker. This pipeline utilizes a contrastive loss to accurately assess the temporal alignment between the audio and visual components of the video. SyncNet has been trained on multiple languages and can thus scale easily to unseen languages.
Videos that exhibit discrepancies in synchronization are identified and subsequently excluded from the data set.

3.3.4. Facial recognition

Facial recognition is important for two reasons:

1. Find all the faces that appear in the video so as to select which one to consider to then cut the video into shorter clips, in which only one face is visible.
2. Make sure the face is actually visible throughout the video, insomuch that a facial recognition software detects it.

Facial recognition is performed on each frame for each video using the open source *face_recognition* (https://github.com/ageitgey/face_recognition) library: which is based on d-lib and claims to have an accuracy of over 99% on the *Labeled Faces in the Wild* [33] benchmark.
While performing this job, the following challenges have been faced:

1. The same face in different frames of the video may not be recognized as the same by the facial recognition library. This can be due to many reasons and the easiest solution is to use models with higher accuracy. However, every model will come with a certain degree of imperfection.
2. The face may not be fully visible, even if it is recognized. For example, the face of a person that is initially facing the camera and after some time is facing left can be recognized by a facial recognition library as still belonging to the same person, because facial recognition libraries can reconstruct the 3D mesh of the face even if it is not fully visible. The lip area, however, may not be fully visible. It is possible to discard such intervals of time in which the lip area is not fully visible, but these intervals are not the majority of the intervals, they are kept in the final videos and the model is trained upon them.

The output of this phase is a list of all the faces that are present in the video along with the timestamps corresponding to the frames in which a particular face is detected.

3.3.5. Face selection and intervals mapping

A video may contain multiple faces, so the facial recognition software exports one picture for every face it has recognized. For each face, the software does export not only a picture of the frame in which the face was first seen, but also a list of timestamps (in a text file), one for each of the frames in which the same face is seen again.

With this approach, there are still some challenges to overcome:

1. The facial recognition software may not recognize the same face but treat two different frames with the same person as different people. Of course this is a limitation of the facial recognition model used, as every model has its own non-zero error rate. This can be recognized by the user because the software exports two pictures of the same person with their own list of timestamps.
2. A face may not be recognized in a given frame. This usually happens when the frame is blurry (the camera is out of focus).

For example, a video in which a single person speaks at frames 0, 1, 2, 3, ..., 100, may produce the following output timestamps:
Face0, with timestamps (frame numbers):
0, 1, 2, [...], 55, [missing 2 frames], 58, [...], 100
Face1, with timestamps (frame numbers):
56, 57
The timestamps in Face1 are the ones missing from Face0
The facial recognition software may treat the same person as two different people, one of them appearing only for two frames. This is obviously an error and the user can merge the two intervals into a single interval by telling the pipeline that these two faces actually belong to the same person. This operation is manual and it is the only manual operation that needs to be done by the user due to a limitation of the facial recognition library.

Gaps in the intervals shorter than a threshold, which is a hyper-parameter set to 200ms are also merged as a contiguous interval, because it means that the facial recognition library just didn't recognize the person for a short duration of time because of a blurry frame or a malfunction of the facial recognition library itself. The choice of 200ms is realistic as in a typical video setting a person doesn't disappear and reappear after 200ms, as the user may not even notice such an effect. Usually if a slide is displayed in an educational video, the slide is displayed for a much longer period of time.

3.3.6. Clips extraction from the videos

The output of the previous step is a list of contiguous intervals in which the same face is detected. This is used to cut the video into clips, rejecting intervals that are shorter than a threshold set to 5s (hyper-parameter).
For example, a video of many minutes in which the speaker alternates with a screen cast of his computer, should automatically be cut and resulting in shorter clips in which only the speaker is visible.
The videos are cut using the popular tool *ffmpeg* with the following parameters:

```
ffmpeg -i in_path -ss from_ts -to to_ts -c:a copy -crf 10 out_path
```

3.3.7. Audio extraction

The audio is extracted for each of the clips produced by the previous step using *ffmpeg*.
The following parameters are used to extract the audio:

```
ffmpeg -i in_path -vn -acodec copy out_path
```

That is, the audio codec is kept as is.

3.3.8. Audio transcription and alignment

The audio extracted by the previous step is transcribed using the Whisper [34] framework from OpenAI. Since Whisper does not automatically generate aligned subtitle files, we have used WhisperX [35], a popular fork of Whisper that automatically generates SRT files with one line for each word spoken and the corresponding time bounds.

3.3.9. Transcription cleaning

The transcriptions generated by WhisperX need further processing. This is due to some incorrect alignments. For example, it may happen that the transcriptions generated give a very short time period for a long list of words (i.e. a person pronouncing 50 words in a one second, which is humanly *impossible*), even if it this doesn't happen in the original video. To find these damaged alignments, we have developed a sliding window algorithm that calculates the average *words/sec* and *characters/sec* for every time window in the transcriptions and just remove the ones that are too fast to be human. Since the parameters we have chosen for the algorithm are quite restrictive, some transcriptions have been cut by 30%.

Due to the data set being relatively small (approximately 5,000 samples of short videos of up to 3s), we have also converted all numbers into text (for example the number *10* converted into its equivalent Italian word *dieci*). To our knowledge, this has not been done before in the literature about lip reading models and it is a useful method to teach the model the mapping between visemes and a sequence of characters rather than the correct written form of the words. The same has been applied to common symbols like *%*, which has been replaced by its Italian word *percentuale*.

WhisperX and other ASR (Automatic Speech Recognition) systems may write thousands as a comma between digits, for example transcribing *five thousands* as *5,000*. However, since we are interested in learning the visemes/letters mapping, we have chosen to remove all thousands separators from the transcriptions.

That is, if in an unseen video the speaker says "five thousands" we want our model to predict "five thousands" as output and not "5,000".

The whole transcription cleaning process is automatic and the algorithms' efficacy have been manually verified on the ItaLip data set.

3.3.10. Mini clip creation

Having generated and cleaned the transcriptions, the pipeline then proceeds to create short videos, called mini clips, that will be then used to extract the lips and train the model.
We have extracted videos between 1 and 3 seconds using a sliding window algorithm that then considers only parts of videos where at least two words are spoken. Because we have the aligned transcripts, we can cut the corresponding video and turn it into a mini clip.

3.3.11. Lip area extraction

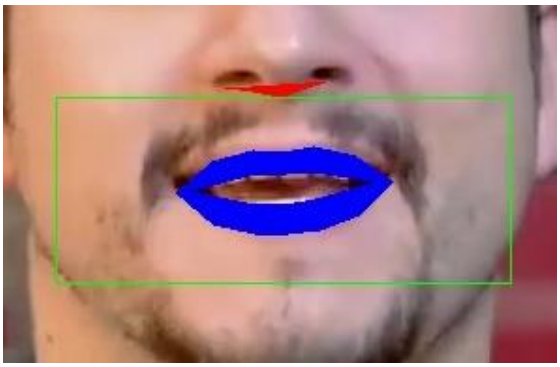

Figure 1: Lip area extracted (in green) and the landmarks used to define it.

Given the miniclips (short videos of maximum duration of 3 seconds), the pipeline proceeds to extract the lip area from the video's frames and resize the resulting image to a size of 160 (width) by 80 (height). To extract the lip area we used the same face recognition library and used it to output the face landmark points. To add some margin to the lip area, we used the distance between the upper lip and the nose and then resized the lip area's rectangle proportionally to touch the nose. Figure 1 shows the rectangle that is cropped and the landmarks used.

3.3.12. The ItaLip data set

The ItaLip data set, constructed through the utilization of the BigLip pipeline, comprises approximately 8 hours of video content distributed across 37 distinct videos. Certain videos that were initially included in the selection process were subsequently excluded due to inadequate audio-video synchronization as assessed by the SyncNet pipeline. Running the complete pipeline on the remaining videos yielded a corpus of 5,666 brief video segments ranging in duration from 1 to 3 seconds. Each segment (mini clip) encompasses a minimum of 2 pronounced words. In relation to the combined duration of the input videos, the resultant data set has a total duration of 2 hours and 55 minutes, an average duration of 1.85 seconds and an average content of 7 words per mini clip.

3.4. Models

3.4.1. LipNet

The LipNet [8] model is a deep learning architecture designed for lip-reading. It was introduced in 2016 by researchers at the University of Oxford, DeepMind and CIFAR (Canada) and has since gained significant attention due to its remarkable performance.

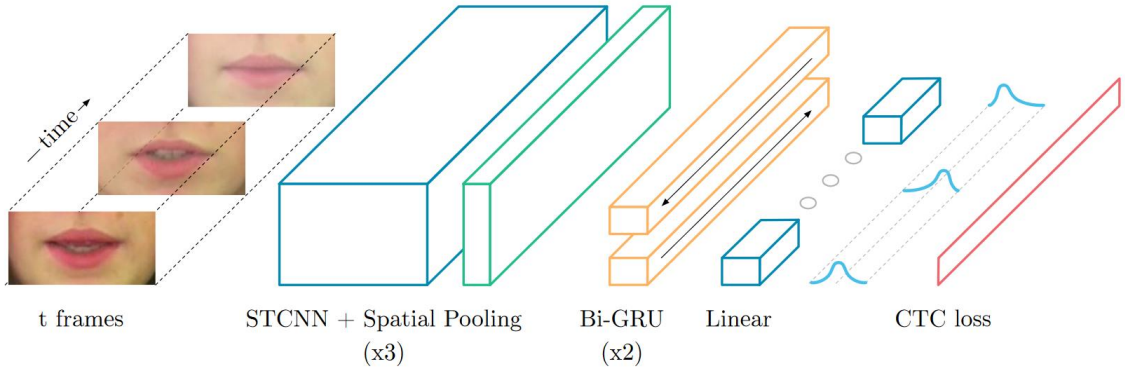

Figure 2: Architecture of the LipNet model. Source: [8]

LipNet employs a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to effectively extract spatio-temporal information from video sequences of lip movements. The architecture is described in the Figure 2. The model's architecture enables it to automatically learn meaningful features from the visual input and capture the temporal dependencies between frames, crucial for accurate lip-reading. The model outputs a sentence one character at a time using a CTC [21] loss.

The training process of LipNet involves a large data set of videos [26] containing spoken sentences along with their corresponding transcriptions. By leveraging this annotated data, the model learns to map visual input sequences to the corresponding textual representation.

The LipNet model uses the CTC (Connectionist Temporal Classification) [21] as loss, so as to not depend on but rather learn the alignment between the video frames and the sentence.
Training the LipNet model on the ItaLip data set resulted in the model being unable to be trained, as the CTC loss requires the input sequence length to be at least as long as the required output sentence. In the case of ItaLip, with sequences of 3 seconds at an FPS of 30, sometimes the number of frames was not enough to reach the number of characters of the target sentence. This problem did not happen in the original data set used for training LipNet as the sentences in the GRID data set are quite short [26] and definite in their structure. The full source code of the training pipeline used is publicly available (https://github.com/hkproj/LipNet-PyTorch).

3.4.2. Transformer-based Model

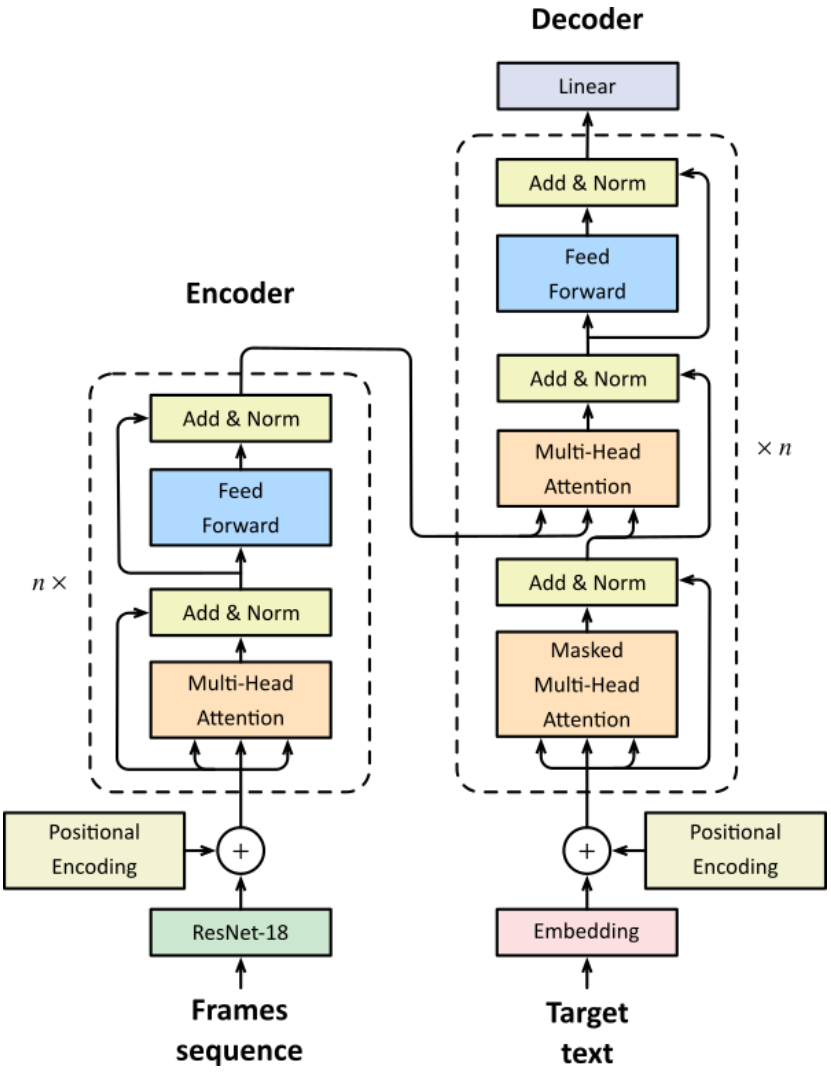

Figure 3: Transformer-based model with a ResNet-18 backbone

The second model upon which we have trained ItaLip is inspired by a Transformer-based model [16] in which the encoder of the Transformer is fed with features taken from the sequence of images and the decoder is fed with the target sentence. Unlike [16], which is only predicting single words, we have trained the model to predict a full sentence and replaced the VGG backbone with a pretrained ResNet-18 model. The architecture is shown in the Figure 3.
This architecture resulted in a model that gets trained very slowly, because predicting a full sentence adds more complexity compared to predicting a single word as done by Huang et al. [16] in which an accuracy of just 45.81% is achieved on single words. Overall, we believe a Transformer-like model can be used for this task, but it needs massive amount of data for sentence-level prediction.

4. ETHICAL CONSIDERATIONS

The development and application of deep learning models for lip reading raise a number of important ethical considerations that should be carefully addressed. While these models have the

potential to revolutionize various fields, including accessibility for individuals with hearing impairments, there are several key ethical concerns that need to be taken into account.

- Privacy: deep learning models for lip reading involve processing and analyzing visual information, often obtained from video recordings. It is crucial to ensure that individuals' privacy rights are respected, and their informed consent is obtained when using their video data.
- Bias and fairness: deep learning models are highly dependent on the quality and diversity of the training data. It is essential to address potential biases in the data, such as imbalances in demographics or cultural representation, to avoid perpetuating unfair outcomes or discriminatory practices.
- Accuracy: lip reading models, like any machine learning system, can make errors. These errors can have significant consequences, particularly in applications such as legal proceedings or security contexts.
- Ethical use cases: The technology of lip reading models can be used for both beneficial and potentially harmful purposes. It is important to consider the ethical implications of deploying such models in different domains. Applications that enhance accessibility, education, and communication for individuals with hearing impairments should be prioritized, while potential misuses that violate privacy, enable surveillance, or infringe on personal freedoms should be carefully regulated and monitored.

In conclusion, the development and deployment of deep learning models for lip reading come with ethical considerations that require careful attention. While the debate on accountability is still open, we believe researchers play an important role in preventing harmful use of AI systems.

5. CONCLUSION

We believe BigLip is a solid foundation for building massive data sets that can be used for lip reading models. We used it to build ItaLip, which to our knowledge is the first Italian data set for this task. Finally, we have trained ItaLip on two popular architectures and we saw that the results are far from good. We believe that this is an unfair comparison because existing data sets are mostly professionally-recorded videos with a very definite sentence structure, as is the case for the GRID [26] data set upon which LipNet [8] was initially trained, while ItaLip is a small data set of videos taken from the web, made to showcase the capabilities of BigLip. Building massive data sets downloading content from the web has proven successful in other areas of deep learning, for example Large Language Models, Image generation and ASR models like Whisper [34]. We built BigLip to pave the way for further research in this direction.

REFERENCES

[1] D. Li, Y. Gao, C. Zhu, Q. Wang, and R. Wang, “Improving speech recognition performance in noisy environments by enhancing lip reading accuracy,” Sensors, vol. 23, no. 4, p. 2053, 2023.
[2] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li, “Seeing what you said: Talking face generation guided by a lip reading expert,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14653–14662, 2023.
[3] W. Sumby and I. Pollack, “Erratum: visual contribution to speech intelligibility in noise [j. acoust. soc. am. 26, 212 (1954)],” The Journal of the Acoustical Society of America, vol. 26, no. 4, pp. 583–583, 1954.
[4] E. D. Petajan, Automatic lipreading to enhance speech recognition (speech reading). University of Illinois at Urbana-Champaign, 1984.

[5] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. C. Azemin, and J. Gubbi, "Lip reading using optical flow and support vector machines," in 2010 3Rd international congress on image and signal processing, vol. 1, pp. 327–330, IEEE, 2010.
[6] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.
[7] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using hmm," Expert Systems with Applications, vol. 38, no. 4, pp. 4477–4481, 2011.
[8] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," arXiv preprint arXiv:1611.01599, 2016.
[9] M. Miled, M. A. B. Messaoud, and A. Bouzid, "Lip reading of words with lip segmentation and deep learning," Multimedia Tools and Applications, vol. 82, no. 1, pp. 551–571, 2023.
[10] Y. Fu, Y. Lu, and R. Ni, "Chinese lip-reading research based on shufflenet and cbam," Applied Sciences, vol. 13, no. 2, p. 1106, 2023.
[11] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," arXiv preprint arXiv:1703.04105, 2017.
[12] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6548–6552, IEEE, 2018.
[13] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3d cnns for lipreading," arXiv preprint arXiv:1905.02540, 2019.
[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A largescale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[16] H. Huang, C. Song, J. Ting, T. Tian, C. Hong, Z. Di, and D. Gao, "A novel machine lip reading model," Procedia Computer Science, vol. 199, pp. 1432–1437, 2022.
[17] B. Shi, A. Mohamed, and W.-N. Hsu, "Learning lip-based audio-visual speaker embeddings with av-hubert," arXiv preprint arXiv:2205.07180, 2022.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[19] Z. Li, T. Lohrenz, M. Dunkelberg, and T. Fingscheidt, "Transformer-based lip-reading with regularized dropout and relaxed attention," in 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 723–730, IEEE, 2023.
[20] K. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 5162–5172, 2022.
[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proceedings of the 23rd international conference on Machine learning, pp. 369–376, 2006.
[22] G. Pu and H. Wang, "Review on research progress of machine lip reading," The Visual Computer, pp. 1–17, 2022.
[23] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1254–1265, 2009.
[24] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–5, IEEE, 2015.
[25] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "Avicar: Audio-visual speech corpus in a car environment," in Eighth International Conference on Spoken Language Processing, 2004.
[26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, 2006.
[27] S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III, vol. 10113. Springer, 2017.

[28] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp. 3444–3453, IEEE, 2017.
[29] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), pp. 1–8, IEEE, 2019.
[30] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "Learning individual speaking styles for accurate lip to speech synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13796–13805, 2020.
[31] E. Egorov, V. Kostyumov, M. Konyk, and S. Kolesnikov, "Lrwr: large-scale benchmark for lip reading in russian language," arXiv preprint arXiv:2109.06692, 2021.
[32] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, pp. 251–263, Springer, 2017.
[33] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.
[34] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022.
[35] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," INTERSPEECH 2023, 2023.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Noorhan Abbas from the University of Leeds. She provided not only technical guidance, but also mentorship and support throughout the project.

AUTHORS

Umar Jamil is a software engineer specialized in Machine Learning. He has founded two startups in his hometown of Milan, Italy and currently lives in Suzhou, Jiangsu, China. More information on his website https://umarjamil.org