Google DeepMind Lip Reading Review

Michael Wu
UID: 404751542
CS35L Winter 2017 Lab 2

The topic I am reviewing is a research paper by Google DeepMind and members of the University of Oxford entitled *Lip Reading Sentences in the Wild*[1]. DeepMind is a company that Google acquired who does research with neural networks and AI. *Lip Reading Sentences in the Wild* describes a new way to use neural networks to lip read sentences from video data, something that is hard even for humans to do. Although there had been previous attempts at creating lip reading software, none had been as successful as this attempt. Previously, researchers had tried to do lip reading of limited words or phrases. The goal of *Lip Reading Sentences in the Wild* was to create a versatile lip reading software model that could be used in everyday life. It would have to recognize complete sentences and work on a variety of different people. This presents a challenge due to differences in different people's appearances and the similarity of mouth shapes for certain sounds such as when pronouncing 'p' or 'b'.

In the paper, the lip reading solution that DeepMind came up with was called *Watch, Listen, Attend, and Spell* (WLAS). WLAS is a new technique that uses different types of neural networks to generate text output from video and audio data. The technique allows the software to predict what is being said based on the previous input data it has processed. In order to do this it uses four different modules. The name of the WLAS model comes from the four modules "Watch", "Listen", "Attend", and "Spell" that it uses. The "Watch" module uses a convolutional neural network to process the video data sent to the lip reader. The "Listen" module processes the audio data sent to the lip reader. Both the "Watch" and "Listen" modules generate attention vectors that make up the "Attend" module, which describes how far back the neural network should look in order to generate the next letter of a word. Finally, the output of "Watch", "Listen", and "Attend" are sent to "Spell", which uses a long short-term memory neural network to generate text output[1].

The convolutional neural network in the "Watch" module works by splitting each input video frame into different areas and processing them to generate an output distribution. It processes the data using different filters, or neurons, that are receptive to different characteristics of the image. The output distribution is a set of probabilities for each characteristic that represents the likelihood that the frame in the video contains that characteristic[3]. The "Watch" module uses a deep neural network, which means it uses multiple layers to filter out features of the video frames as shown in figure 1. The "Listen" module processes the audio input to the lip reader using a protocol called MFCC that does not use a neural network. It transforms the waves of the audio data and sends that as output to the "Spell" module. The "Spell" module receives the outputs of the three other modules and generates the text output by using a long short-term memory neural network, as shown in figure 2. A long short-term memory neural network allows for the outputs of the network to be reused as inputs. This means that it can look back in its history in order to use its previous inputs to predict the next output[2]. The "Spell" module's neural network generates a probability distribution for each alphabetical character that allows it to select the most likely character to output.

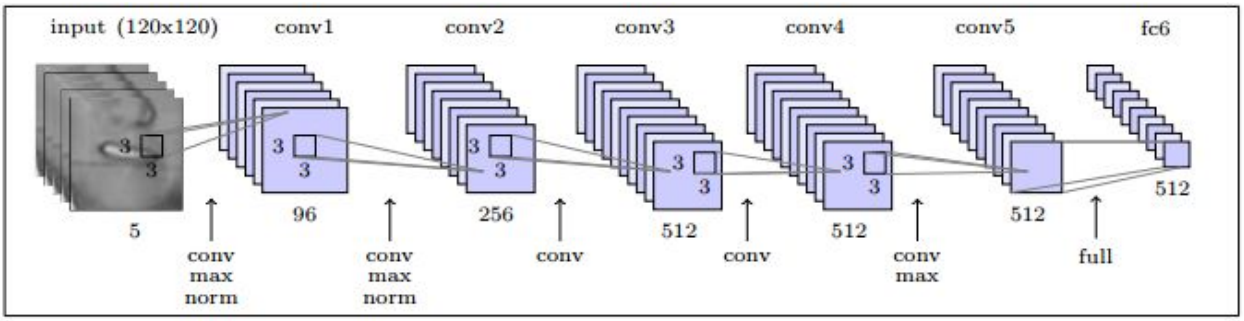

Figure 1. The convolutional neural network used in the “Watch” module. This image shows how the neural network processes different areas of the input frames, reduces them repeatedly, and generates an output vector for different characteristics of the data.

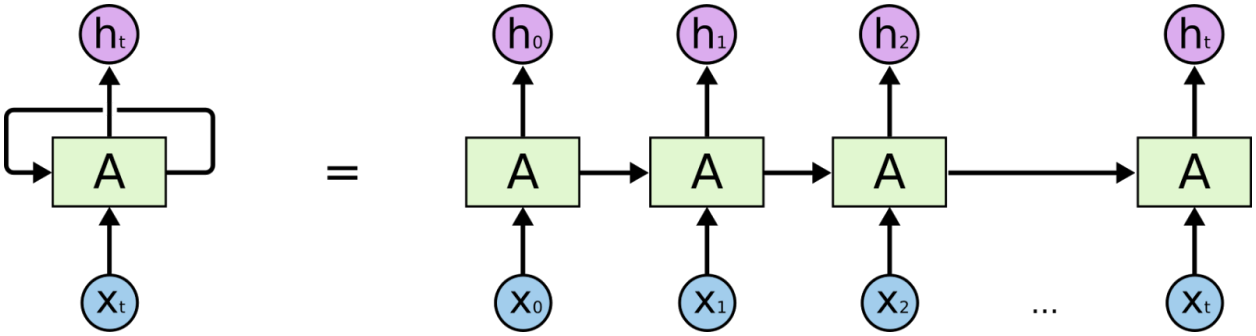

Figure 2. A long short-term memory neural network. This shows how the outputs of the neural network A are reused, allowing it to effectively “look back” at previous inputs.

In order to train this model for lip reading, the researchers had to use supervised machine learning. Supervised machine learning involves sending training data to the WLAS system and checking its outputs against the desired correct output. The system repeatedly changes what the neurons in its neural networks are receptive to until it generates correct output for most cases.

To generate the data set, the researchers took around five thousand hours of video data from BBC news channels and split it into segments of around one sentence or phrase each. They used subtitle data to generate the desired text outputs to train the lip reader with for each video clip. They also trained the WLAS model with only audio data, so that it could be receptive to both the video and audio data[1]. This data set represents a significant improvement over previous data sets, which only consisted of a limited set of words or phrases.

Once the WLAS model was trained, the researchers turned off the “Listen” module so that it would work with only video data and not audio data. This simulates the software doing lip reading. They found that on their test data set from the BBC video, the WLAS model had a word error rate of 53.2% when lip reading, which is lower than the 73.8% word error rate of a professional human lip reader[1].

In summary, the researchers at DeepMind and University of Oxford have created an impressive new lip reading model that outperforms humans and previous attempts to create lip reading software. The researchers have also generated a large lip reading training data set,

which is helpful for future research in lip reading. The lip reader that DeepMind has created relies on two different neural networks, convolutional neural networks and long short-term memory neural networks, to predict what people are saying. This software works on people speaking in normal situations, such as on the news. It recognizes full sentences and phrases. Eventually, possible applications of this technology include assisting in speech input or captioning videos. Many smartphones today have speech recognition software, and using this lip reading technology in addition to speech recognition could improve accuracy and usability on these devices. It could even allow speech input to work in noisy environments. With regards to captioning, current automated captioning technology such as what is found on YouTube frequently generates nonsensical output that does not assist understanding at all. Using lip reading in addition to audio processing on these videos could generate more accurate captions. Overall, lip reading is an exciting new way of using neural networks that can soon be used in a variety of applications. It can greatly impact everyday life through its inclusion in consumer products.

Bibliography

1. Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2016) Lip Reading Sentences in the Wild. *Computing Research Repository.* Retrieved from https://arxiv.org/list/cs.CV/1611.
2. Sak, H., Senior, A., Beaufays, F. (2014) Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *Computing Research Repository.* Retrieved from https://arxiv.org/list/cs.NE/1402.
3. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems, 25,* 1097-1105. Retrieved from https://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012.