Automated Lip Reading using Deep Reinforcement Learning

Dung Le
Bennington College
Bennington, VT 05201

1. Introduction

Lip reading, also known as audio-visual recognition, has been considered as a solution for speech recognition tasks, especially when the audio is corrupted or when the conversation happened in noisy environments. It can also be an extremely helpful tool for people who are hearing-impaired to communicate through video calls. This task, however, is challenging, due to factors such as the variances in the inputs (facial features, skin colors, speaking speeds, etc.) and the one-to-many relationships between viseme and phoneme (Chung et al., 2016; Garg et al., 2017). This project aims to tackle lip reading by modeling an agent that is capable of learning the features by interacting with the environment using reinforcement learning methodology.

Task Description: Given a video of a speaker with no audio file and no hidden facial features (especially the lip region), the system transcribes lip movements into text.

2. Related Works

Most of the works done in lip reading focused on using either variations of Hidden Markov Model (Rekik et al., 2015; Gergen et al., 2016) or deep neural network models (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Garg et al., 2017). Chung et al. developed a *Watch, Listen, Attend and Spell* network which utilises a novel dual attention mechanism in addition to LSTM networks. Deep reinforcement learning, on the other hands, is becoming more popular, especially for tasks concerning language generation (Young et al., 2018). Razanto et al. applied reinforcement learning to train RNN-based models for several sequence generation tasks, i.e. text summarization, machine translation and image captioning. The ultimate goal of an automated lip reading system is to generate text from lips movement; thus, it fits into the tasks where deep reinforcement learning can be applied.

3. Methods

3.1. Dataset

For this project, I proposed using two datasets, the GRID dataset (available at http://spandh.dcs.shef.ac.uk/gridcorpus/) and the BBC-Oxford LWR dataset (available at http://www.robots.ox.ac.uk/~vgg/data/lip_reading_sentences/). The two datasets are similar in the way that both are designed for learning lip reading at sentence level, instead of just words and utterances. Each dataset consists of videos with visible facial features and annotated text transcriptions of the content.

3.2. Video Processing and Image Encoder

The video is first separated into still frames (images). For the LWR dataset, each image only covers the lip region of the face, therefore, nothing else needs to be done. For the GRID dataset, however, I will use a face-detector module in OpenCV to detect and extract facial features that focuses on the lip region. After this step, a video is split into a sequence of images, with each containing the mouth region of the speaker. Each image is then passed into the ConvNet architecture (a five-layer convolutional network with a fully-connected layer) trained on the ImageNet dataset, which outputs the vector representation encoding the features of mouth region.

3.3. Reinforcement Learning and Deep Q-network

In order to use a reinforcement learning method to solve this task, it is important to cast the problem in the reinforcement learning framework (Sutton and Barto, 1998). The components of a reinforcement learning task are an agent, an environment, a set of actions and policies, and a reward. In this framework, the lip reading model acts as an agent, which interacts with the external environment (the input vectors encoding lips region). By interacting with the environment, the agent learns the best policy (model's parameters) by picking an optimal action (refers to predicting the next word in the sequence at each time step) that maximizes the reward. In order to promote lip reading at sentence level, a reward is only given when the whole sentence is output.

Deep Q-network (Mnih et al., 2015), is a method that combines reinforcement learning with deep neural network. In *Human-level control through deep reinforcement learning,* Mnih et al. used the deep convolutional network to exploit the local spatial correlations presented in the classical Atari 2600 game. For this project, instead of using convolutional network, I will replace it with the recurrent neural network with LSTMs to perform sequence level training in text. In addition, I will also implemented Experience Replay, which allows the network to train itself using stored memories from its experiences, and the second 'target' network (to compute the the target Q-values during each update.)

Word Error Rate, a common evaluational metric in tasks concerning language generation, will be used to measure the performance of the network.

3.4. API for pre-trained network

Since I want my lip reading system to be integrated into video call and/or lip reading applications, I will build a streaming API for the network pre-trained using the BBC-Oxford LWR dataset.

REFERENCES

Amit Garg, Jonathan Noyola, and Sameep Bagadia. 2017. *Lip Reading using CNN and LSTM*.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2016. *Lip Reading Sentences in the Wild.*

Sebastian Gergen, Stenffen Zeiler, Ahmed Hussen Abdelaziz, Robert Nickel, and Dorothea Kolossa. 2016. *Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR.*

Ahmed Rerik, Achraf Ben-Hamadou, and Walid Mahdi. 2015. *An adaptive approach for lip-reading using image and depth data.*

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. *ImageNet classification with deep convolutional neural networks.*

Karen Simonyan and Andrew Zisserman. 2015. *Very deep convolutional networks for large-scale image recognition.*

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. *Recent Trends in Deep Learning Based Natural Language Processing.*

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. *Sequence Level Training with Recurrent Neural Networks.*

Volodymyr Mnih, Koray Kavukcuoglu, and David Silver. 2015. *Human-level control through deep reinforcement learning.*

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*.