Visual Speech Recognition System with Deep Neural Networks

Jin Sol Choi[1], Daeyeol Kim[1], Sooyoung Cho[1], Sinwoo Yoo[1], and Chae-Bong Sohn[1]

[1]Department of Electronics and Communications Engineering, Kwangwoon University
20, Gwangun-ro, Nowon-gu, Seoul, 01897, Republic of Korea.

Abstract

Recent artificial intelligence manufactures based on voice recognition cannot be used by the deaf. In order to solve this problem, we present 'Visual Speech Recognition System' using deep learning with lip movement. This system analyzes mouth shape and process time series data through the 3-dimensional convolutional neural network and gated recurrent unit. Our visual speech recognition system deals with Korean vocabulary, and creates subtitles based on oral movements of the subjects in the video. This system recognizes individual words rather than the whole sentences. We achieved 91.8% accuracy. This system could be applicable for someone who being deaf, having the difficulty of hearing, or anyone who requires communication without the voice.

Keywords: Deep neural networks, Visual speech recognition system, Gated recurrent unit, Hidden markov model

INTRODUCTION

Most the deaf people communicate through sign language. However, the counterpart cannot understand what they meant, if the counterpart hasn't learned the sign language. Then it is impossible to continue the conversation. Some deaf people are capable of reading others' lips and interpreting the words from others, but this requires plenty time for training. When deaf people read lips as a form of communication, they also rely on facial expressions and body language in order to interpret the meanings. In this paper, we will discuss how we have considered lip shapes and lip reading to develop our deep learning system based on these factors [1] [2].

RELATED WORKS

Before constructing the system, we separated it into 2 parts as 'feature extraction' and 'time series data analysis'. We applied a 3-dimensional convolutional neural network in order to extract the features from input data which consist of moving pictures based on a certain timeline. Also we borrowed a gated-recurrent-unit model for the following process of given information of timeline that would be obtained after. HE initialization was used regarding overall process without obstacles.

A. Convolutional neural network

CNN is a model that can handle any 3-dimensional type of input-data without losing information, not like such as other types of neural network that's built by only fully connected layer as a single dimension.

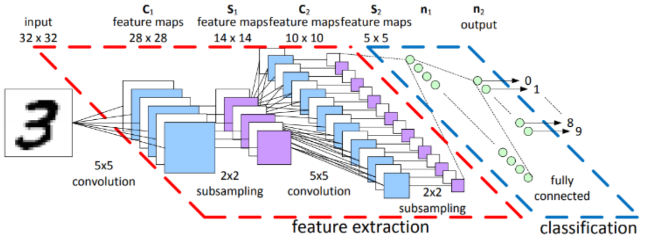

Figure. 1: Convolutional neural network architecture

Fig. 1 shows the overall hierarchy of the CNN model [3]. We can see the convolutional layer, the part of feature-extraction of accumulated Max-Pooling-layer, and the part of classification applied by softmax function that consists of rest of fully-connected-layers. This CNN performs required convolutions by circulating filtered input-data in order to extract the specific image characteristic and generates a feature-map as the result of this process. The designed system itself uses CNN also to obtain the key element of mouth as a specific feature.

B. 3-dimensional Convolutional neural network

Normally, Convolutional Neural Network (CNN) builds on 2-dimensional images, but this system treats videos which are spatiotemporal data. Fig. 2 indicate architecture of 3-dimensional CNN [4] [5].

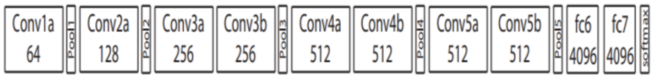

Figure 2: 3-dimensional CNN architecture

When input data type is video volume, then output maintain volume form. This network preserves temporal information of the input video. Hence, our system extracts lip shape features through the 3-dimensional CNN.

C. Gated recurrent unit

In order to deal with the spatiotemporal data, the result from '3-dimensional CNN', we applied a gated recurrent unit (GRU). The GRU is a type of recurrent neural network (RNN). This is the same as the Long Short-Term Memory(LSTM) which of one another transformed RNN as the perspective that can resolve the Gradient Vanishing or Explosion issue. As shown in Fig 3., GRU have no separate memory cells [6]. For this reason, the computational complexity of GRU is going to be simple, and also it has a benefit as to be possible to train with fewer datasets than usually required. Therefore, our spatiotemporal data were processed with GRU.

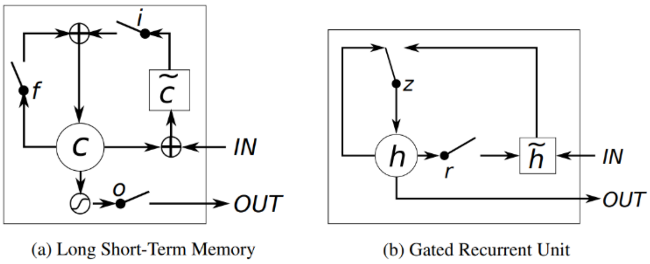

Figure 3: LSTM and GRU

D. HE initialization

Initialization is one of the key elements of a deep learning process. The overall performance of training results relies on how these weight values can be initialized. There are some methodologies being used for weight value initialization, we applied HE initialization for that purpose. In general, these initialization values of weights are usually taken by very small numbers. but this may cause of converging zero if the training process performs with hidden layers. In order to avoid such case, we found and applied the optimal initialization values by using HE initialization, which derives by half values of the standard normal distribution of the number of input nodes that could be obtained from Xavier initialization.

E. Datasets

As we have previously mentioned, our system organized datasets consisting of individual words. Datasets are composed of 4 categories; animal, food, number and fruit. In order to confirm the more obvious result, we placed more efforts to exclude word samples that make very similar orally shaped. The GRID audiovisual sentence corpus dataset was referenced for this training process [7] [8].

Table 1: Example of Korean word training dataset

Korean word dataset category			
Animals (Meaning)	Foods (Meaning)	Numbers (Meaning)	Fruits (Meaning)
하마 /ha̠ma̠/ (Hippopotamus)	만두 /ma̠ndu/ (Dumpling)	이 /i/ (two)	배 /pɛ̝/ (Pear)
노루 /no̞ru/ (Roe deer)	녹차 /no̞k̚t͡ɕʰ/ (Green tea)	삼 /sʰa̠m/ (three)	사과 /sʰa̠gwa̠/ (Apple)
오리 /o̞ri/ (Duck)	초밥 /t͡ɕʰopa̠p̚/ (Sushi)	오 /o̞/ (five)	자두 /t͡ɕa̠du/ (Plum)
호랑이/ɸʷo̞ra̠ŋi/ (Tiger)		팔 /pʰa̠ɭ/ (eight)	
새우 /sʰɛ̝u/ (Shrimp)		십 /sip̚/ (ten)	

As shown in Table. 1, Korean phonetics are presented with IPA phonetic notation[9]. Each experimenter recorded 225 videos and every video's running time standardized 3 second. In total, there were 6 experimenters and they recorded 1530 videos. Regarding any of interference that may affect the training result, we secured datasets those are from all same condition, including the space and length of assets. Additionally, we generated the assets for the dataset based on the assumption that each of the pronunciations makes clear and obvious oral shapes. The Align file that is composed is based on the Hidden Markov Model Toolkit(HTK) form. Among this align file information, sil indicates a silent syllable, also the starting and ending of each word. To produce the align file, we extracted .wav file from the videos and used a program, SFSWin(Speech Filing System). SFSWin is a voice analysis software which supports features of overall functions for voice signal analysis, subroutine libraries, and integrated scripts for other automation processes, so it's widely utilized on the various purpose of sound and voice analysis.

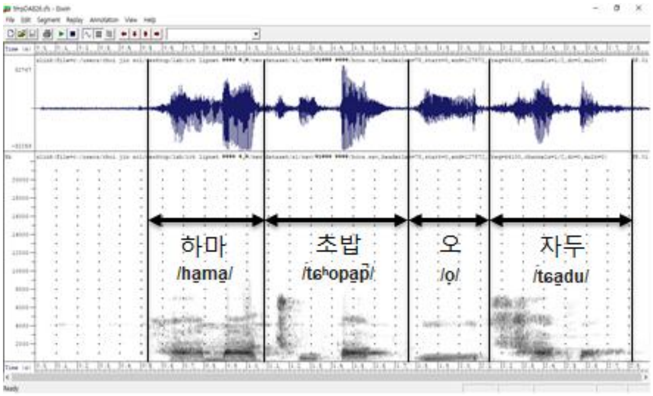

Figure 4: Speech filing system

Fig. 4 is the example of the SFS windows version usage. We extracted wave signals from voice assets of the dataset, such as pronunciation ″hippopotamus″ (하마 /ha̠ma̠/), ″sushi″ (초밥 /t͡ɕʰopa̠p̚/), ″five″ (오 /o̞/), ″plum″ (자두 /t͡ɕa̠du/) and utilized the duration of each pronunciation in order to compose the ″Align file″ [10].

EXPERIMENT

Fig. 5 present our overall architecture of lip analysis system. Specific features were obtained as the result of the CNN process by each oral shape from cropped images [11]. GRU will process these features with among given timestamps for each of feature. we initialized our weight values by applying HE initialization method, and CTC Loss function was used for the training process.

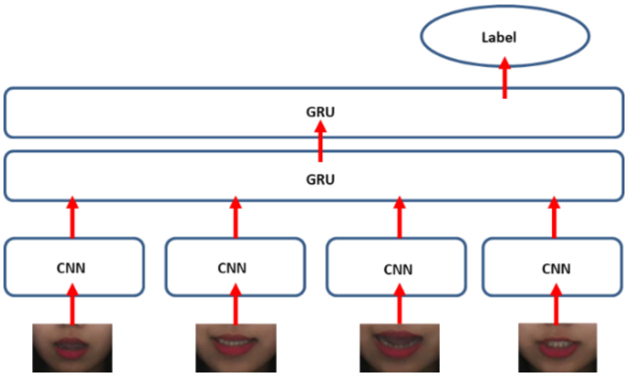

Figure. 5: Lip analysis system architecture

We secured needful computing power with GeForce GTX 1080 Ti graphic card, entire training was done by 729 times of iterations, and overall duration of training process was 9 hours, 37 minutes and 48 seconds. Loss value was indicated as 4.497, and calculated accuracy was 91.8%.

CONCLUSIONS

In this paper, we proposed lip analysis system based on deep learning. Fig, 6 top shows pronunciations of "Shrimp" (새우 /sʰɛ̞u/), "dumpling" (만두 /ma̠ndu/), "two" (이 /i/), and "pear" (배 /pɛ̞/), and botton shows pronunciations of "roe deer" (노루 /no̞ru/), "green tea" (녹차 /no̞k̚t͡ɕʰ/), "ten" (십 /sip̚/) and "apple" (사과 /sʰa̠gwa̠/).

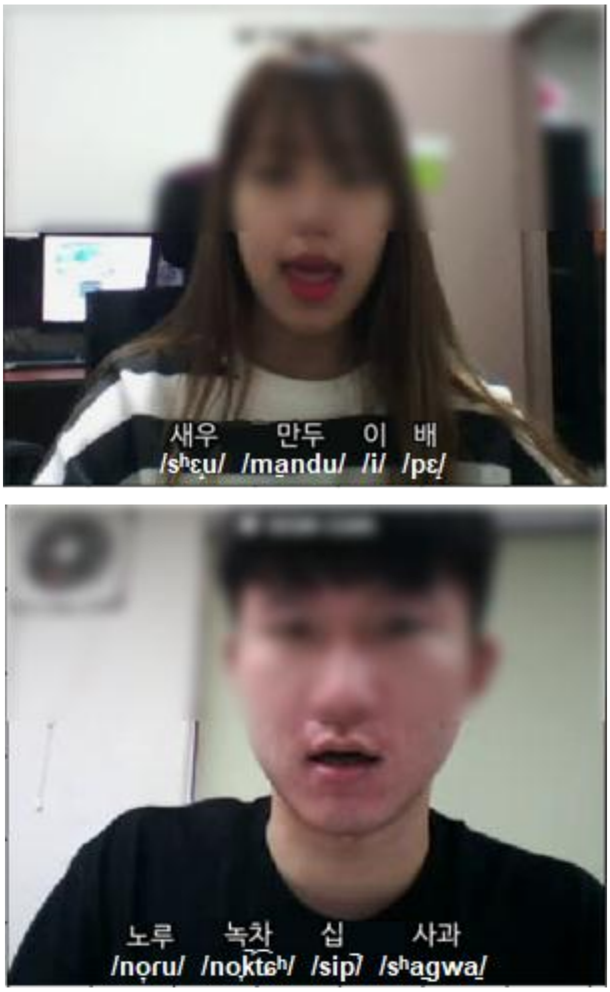
새우 만두 이 배
/sʰɛ̞u/ /ma̠ndu/ /i/ /pɛ̞/
노루 녹차 십 사과
/no̞ru/ /no̞k̚t͡ɕʰ/ /sip̚/ /sʰa̠gwa̠/

Figure 6: Result of our lip analysis system. **Top:** Pronunciations of "Shrimp" (새우 /sʰɛ̞u/), "dumpling" (만두 /ma̠ndu/), "two" (이 /i/), and "pear" (배 /pɛ̞/). **Bottom:** Pronunciations of "roe deer" (노루 /no̞ru/), "green tea" (녹차 /no̞k̚t͡ɕʰ/), "ten" (십 /sip̚/) and "apple" (사과 /sʰa̠gwa̠/).

As shown in Fig. 6, this system treats individual words. By using this system, deaf people can communicate conveniently with others

Moreover, if we consider diverse direction of face, it can be applied throughout the news, reports, Youtube and so on. And in CCTV environments, it can also prevent crime in advance. And this is expected to make possible to communicate each other under a severe situation, such as a middle of a war or very noisy circumstances.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-

2016-0-00288) supervised by the IITP (Institute for Information & communications Technology Promotion).

REFERENCES

[1] Chung, J. S., & Zisserman, A. Lip reading in the wild. In Asian Conference on Computer Vision, Springer, Cham, (2016), pp. 87-103

[2] Garg, A., Noyola, J., & Bagadia, S. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, (2016)

[3] PEEMEN, Maurice; MESMAN, Bart; CORPORAAL, Henk. Speed sign detection and recognition by convolutional neural networks. In: Proceedings of the 8th International Automotive Congress. 2011. p. 162-170.

[4] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, (2015), pp. 4489-4497

[5] JI, Shuiwang, et al. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2013, 35.1: 221-231.

[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, (2014)

[7] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. LipNet: End-to-End Sentence-level Lipreading, (2016)

[8] http://spandh.dcs.shef.ac.uk/gridcorpus/

[9] Handbook of the International Phonetic Association-A Guide to the Use of the International Phonetic Alphabet, International Phonetic Association, (1999)

[10] Janvale, G. B., & Deshmukh, R. R. Speech Feature Extraction Using Mel-Frequency Cepstral Coefficient (MFCC), The Emerging Trends in Computer Science, Communication and Information Technology, (2010)

[11] LAWRENCE, Steve, et al. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 1997, 8.1: 98-113.