Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models

T. OZCAN and A. BASTURK

Abstract—**Lip reading has become a popular topic recently. There are widespread literature studies on lip reading in human action recognition. Deep learning methods are frequently used in this area. In this paper, lip reading from video data is performed using self designed convolutional neural networks (CNNs). For this purpose, standard and also augmented AvLetters dataset is used in train and test stages. To optimize network performance, minibatchsize parameter is also tuned and its effect is investigated. Additionally, experimental studies are performed using AlexNet and GoogleNet pre-trained CNNs. Detailed experimental results are presented.**

Index Terms—**Convolutional neural networks, data augmentation, deep learning, human action recognition, human computer interaction, lip reading, transfer learning**

I. Introduction

HUMAN ACTION RECOGNITION is an important phase for human computer interaction [1]. Lip reading, a subcategory of human action recognition, has begun to be used in various applications [2-6].

Sound and image assisted features can be used for lip reading. In particular, data containing image-assisted features seem to have higher success rate in applications where they are used. Success rate of lip reading is also directly related to the classification techniques which are used extensively with feature selection. Hidden Markov models [7,8], support vector machines [7,9], k-nearest neighbor algorithm [7] are most basic conventional classification algorithms. After deep learning methods have been used frequently in classification problems [10-13], researchers have started to use them on lip reading which is another classification problem.

Lip reading process is applied at alphabet, word and sentence levels [14]. In alphabet level lip reading process, still image and time series classification methods can be used. On the other hand, classification methods such as long short-term memory (LSTM), recurrent neural network (RNN), and so on are preferred in word and sentence level operations.

In this paper, we aim to classify an alphabet level lip reading dataset, AvLetters [15], by using a designed convolutional neural network (CNN) model and pre-trained model supported CNNs. A CNN model with 3 convolution layers, 3 max pooling layers, and 3 regularization layers is designed. On the other side, AlexNet [16] and GoogleNet [17] supported CNN structures are employed to compare performance with designed CNN model. Dataset size also affects performance as much as the classification model. In this study, data augmentation is also addressed by using data augmentation techniques. Experimental studies, which are based on CNN models with and without augmented dataset depends on different “minibatchsize” values, are performed. The contributions of this study are as follows:

- To the best of our knowledge, transfer learning supported CNN using ALexNet was used for the first time on the AvLetters dataset.
- By using some techniques, the problem is converted to a still image based lip reading from time series data.
- A CNN architecture is proposed and compared with transfer learning supported CNNs.
- The proposed methods in this study, not only have better performance than some methods used in [18] in the literature but also are easy to develop.

The presentation of this paper will be as follows: Section 2 provides a literature review on the subject. Section 3 describes the information of CNN and pre-trained models used through the study. In section 4, experimental studies are presented and in conclusion section, which is the last section of this paper, success rate results obtained from experiments and future studies are reviewed.

II. Related Work

Garg et al. [2], combined CNN and LSTM deep learning methods and applied it on lip reading problem. In this combination, CNN was used for feature extraction and LSTM was used for classification. MIRACL-VC1 [7] dataset which consists words and phrases was used for testing the model. Li et al. [3], who performed classification with CNN using the dynamic feature image instead of original image, tested the model with ATR Japanese speech dataset. Petridis et al. developed an LSTM supported work on visual speech recognition [4]. The model consisted of two streams. The first was feature extraction from mouth and the second was the change between images. The temporal dynamics of each stream were performed with LSTM. Proposed method was

TAYYIP OZCAN, is with Department of Computer Engineering University of Erciyes, Kayseri, Turkey, (e-mail: tozcan@erciyes.edu.tr).
https://orcid.org/0000-0002-3111-5260

ALPER BASTURK, is with Department of Computer Engineering University of Erciyes, Kayseri, Turkey, (e-mail: ab@erciyes.edu.tr).
https://orcid.org/0000-0001-5810-0643

Manuscript received December 7, 2018; accepted April 3, 2019.
DOI: 10.17694/bajece.479891

tested with OuluVS2 [19] and CUAVE [20] datasets. A model using LSTM with 5 convolution layers and 256 hidden units was developed by Dong and his team [21]. The method's performance was tested by combining 2 datasets. Word-level visual speech recognition models using deep learning was proposed by Stafylakis and Tzimiropoulos [22]. The proposed method was combined with CNN, ResNet and bi-directional LSTM. The model tested on LRW [23] dataset and experimental results were presented. Takashima et al. [5] developed a deep learning-supported speech recognition system for people with severe hearing loss. Both voice and visual data were used in the method and extracted features were included in system for classification. In another work, Takashima et al. [24] proposed a new approach for lip reading using the combination of lip image and sound features by using deep learning. The proposed method was tested on ATR Japanese speech dataset. A method which classifies the dataset consisting of Turkish color names by taking image and angle values with Kinect device was proposed by Yargic and Dogan [6]. Authors, who took the knowledge of lip co-ordinates from Kinect camera, used the angles between the points and classify them with k-nearest neighbor search algorithm.

III. Methods

A. Convolutional Neural Networks

CNN is a type of artificial neural network (ANN) specialized to handle multi-dimensional, large data. Convolutional networks are neural networks that use at least one layer of convolution processing instead of general matrix multiplication [25-27].

Basic components of CNNs are; convolution layer, pooling layer, activation functions, fully connected layers, loss layer, regularization, and optimization.

The convolution layer includes a learnable filter set. This layer is the structure of convolutional network that has the ability to learn along like as fully connected layer. The parameters that are important for this layer are spatial extent, number of filters and stride [25].

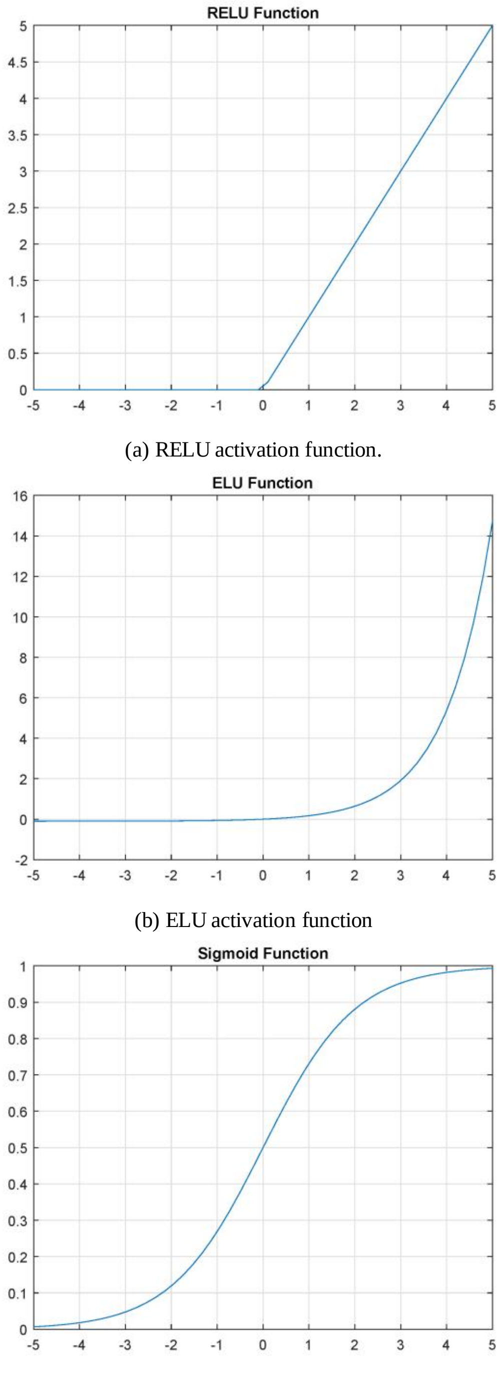
Another important structure that reduces the network model's cost is the pooling layer. In this layer, which makes the system resistant to small position changes, pooling process usually uses operations such as sum, maximum, average, etc.

Choice of activation functions such as ReLU, ELU, sigmoid etc., significantly affects the performance of CNN. ReLU, which is a piecewise linear function, is an activation function that returns negative inputs to zero, and positive inputs to output without changing them. The ReLU function graph is shown in Figure 1a. ELU is an activation function that allows neural networks to learn faster and achieve higher classification accuracy [28]. The ELU function graph is shown in Figure 1b. The sigmoid function is a continuous and derivable function and is frequently used. The Sigmoid function graph is shown in Figure 1c.

Fully connected layer comes after convolution and pooling layers in CNN. In this layer, neurons have full contact with the previous layer.

Loss layer, which is the last layer of CNN, determines how the difference is to be evaluated between predicted and actual labels during training. Softmax is the most commonly used loss function.

Regularization techniques prevent the overfitting problem, which is an important problem for deep neural networks. Dropout and Dropconnect are two most important regularization techniques [25].



(a) RELU activation function.

(b) ELU activation function

(c) Sigmoid activation function

Fig.1. Activation functions

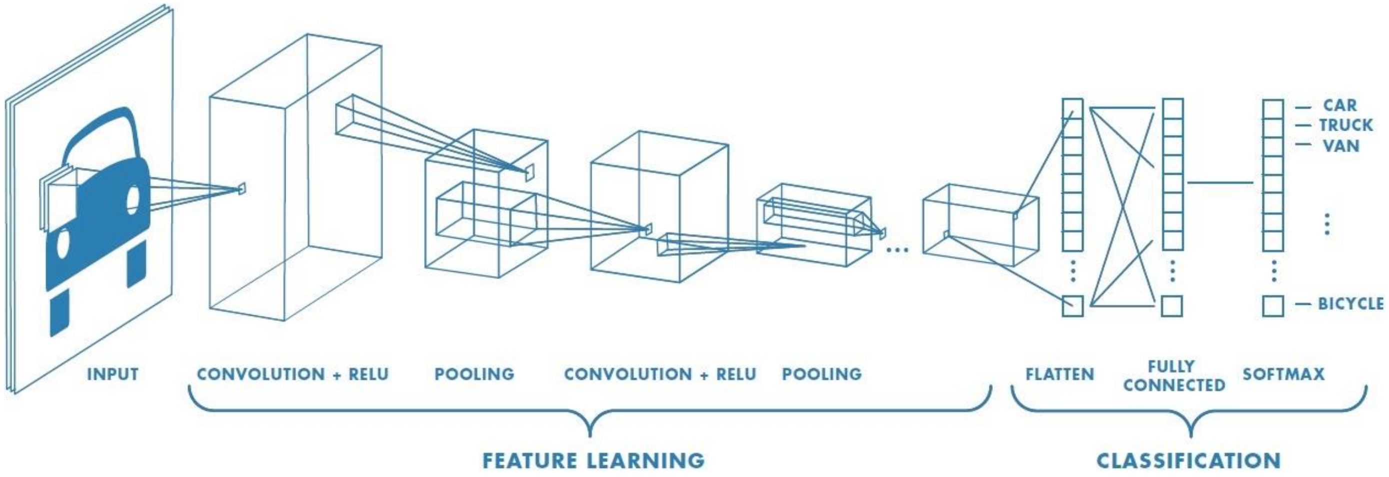


Fig.2. Convolutional Neural Network architecture [25]

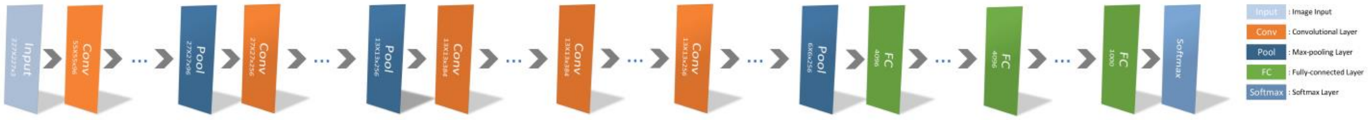


Fig.3. AlexNet architecture

B. *Pre-Trained Models*

The easier and faster option than developing a new model is using pre-trained networks. Pre-trained networks are used for purposes such as classification, transfer learning, and feature extraction.

Figure 2 presents the architecture of a typical neural network with basic components. The most commonly used pre-trained CNN architects are AlexNet [16], VGG16 [29], VGG19 [29], GoogleNet [17] and ResNet 50 [30]. In this paper, AlexNet and GoogleNet will be used for experiments.

1) *AlexNet*

Transfer learning or feature extraction can be applied to different problems with AlexNet, which is trained with a large library of images and performs powerful feature extraction. A subset of ImageNet dataset was used to train this network model. AlexNet, which has 8 learnable layers, 5 convolution layers, and 3 fully connected layers, won first place in 2012 ImageNet Large Scale Visual Recognition Competition (ILSVRC). The architecture of AlexNet model is shown in Figure 3.

2) *GoogleNet*

GoogleNet model, which won the 2014 ILSVRC competition, has a smaller and faster network structure than the VGG models. GoogleNet has a higher performance than AlexNet on ILSVRC dataset. It has a more complex structure than AlexNet and VGG models. The architecture of the GoogleNet model is shown in Figure 4.

IV. Experimental Studies

Two different methods have been applied to study on with and without augmented AvLetters dataset. The first of these methods is developing a user-designed CNN model. Another method is using AlexNet and GoogleNet pre-trained models instead of a user-designed one.

A. *AvLetters Dataset*

Lip reading is studied under the heading of human action recognition. AvLetters dataset, which consists of both video and audio recordings, was created by repeating the 10 alphabet letters in 3 different trials [15]. Visual data will be used in the study. Therefore, the Matlab mat data file containing image information of the dataset is handled. Numerical information of image frames formed during each alphabetical pronunciation is kept in mat files. For example, 'R3_Kate-lips.mat' file holds the frame information of 3rd pronounce of the letter R of Kate.

In AvLetters dataset, 10 subjects repeated 26 letters 3 times. “TrainData” and “TestData” folders have been created by the way first two pronunciations are recorded for training and the last pronunciation is recorded for testing.

1) *Data Pre-processing*

Duration of each letter for each trial may vary. This situation causes the number of different frames to be formed between samples. As the first step in dataset preparation, average number of frames is set to 20, and adding image process for below 20 and deleting image process is performed for above 20. After each letter is formed in 20 frames, lip images are rendered as still images in 5x4 format. By this way, the dataset consisting of the time series is transformed into a structure composed of still images. A sample image of the pre-processed dataset is shown in Figure 5.

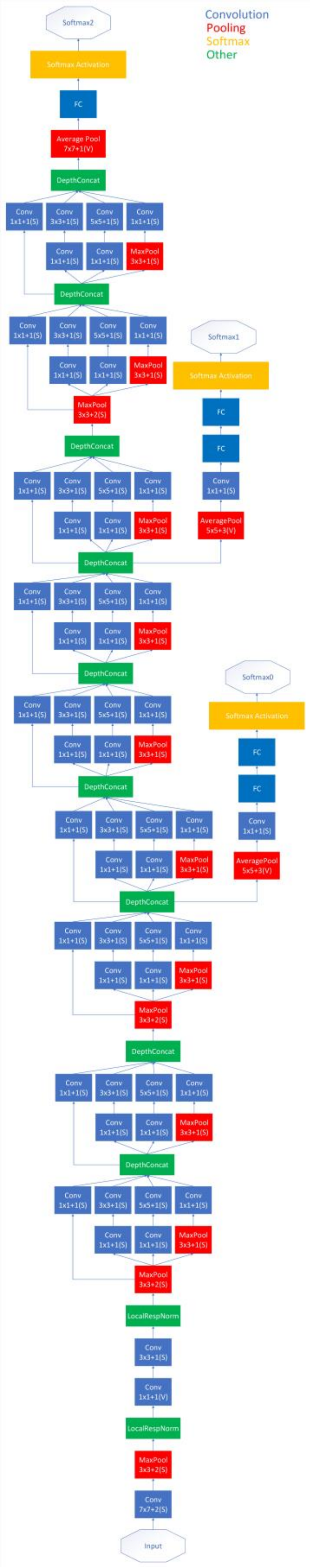


Fig.4. GoogleNet architecture [31]

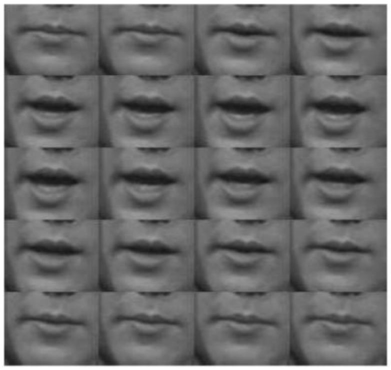
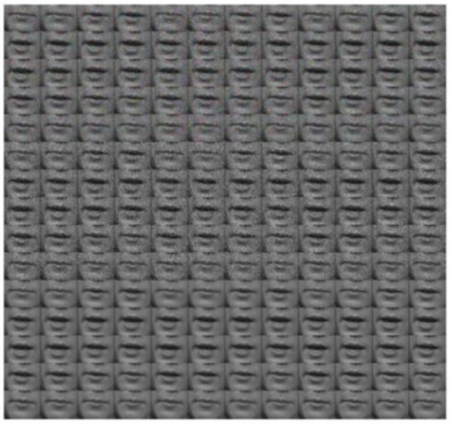


Fig.5. Avletters, pre-processed data - letter A

2) *Data Augmentation*

520 samples can be regarded as a small number for training in deep learning. Data augmentation (DA) can be used to increase the size of samples. The training dataset is increased using various DA methods. Adding noise with “gaussian”, “salt and pepper” and “speckle”, sharpening with “unsharp” and softening with “median” filtering are the operations of DA used in this paper. Also, RGB and grayscale format of noised images are used additionally. Therefore, eight different types of original images are replicated. An example for DA is presented in Figure 6a and augmented data matrix is depicted in Figure 6b.



(a) Avletters, augmented data - letter A

Gaussian RGB Image	Salt & pepper RGB Image	Speckle RGB Image
Gaussian Grayscale Image	Salt & pepper Grayscale Image	Speckle Grayscale Image
Original Image	Median Filtered Image	Unsharp Filtered Image

(b) Avletters, augmented data matrix

Fig.6. Data augmentation structure

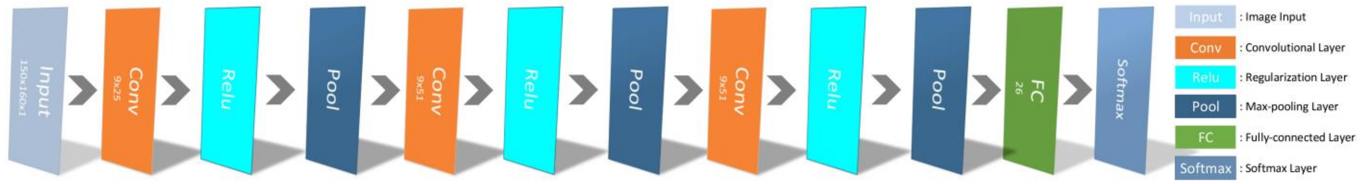


Fig.7. Designed CNN architecture

B. Proposed CNN Model

Deep learning methods have recently been used frequently in classification problems. CNN, which is a deep learning method, is accepted as a popular method in classification problems. A CNN model that separates from ANN models because of the number of layers is designed for lip reading problem. Designed CNN model is presented in Figure 7. This model, consisting of 3 convolution layers, 3 max pooling layers, 3 regularization layers, achieved 54.23% success rate on augmented AvLetters dataset with '8' value of minibatchsize parameter. A confusion matrix graph is presented in Figure 8.

C. Pre-Trained Model Supported CNN

Developing a new CNN model is a challenging task. Transfer learning from pre-trained models is an easier and faster way.

Designed CNN architecture is compared with AlexNet and GoogleNet supported CNN on with and without augmented AvLetters dataset. The presentation of results obtained with different “minibatchsize” values is done and presented in Table I. According to this table, when the “minibatchsize” value is 8, AlexNet supported CNN + DA is more successful than other models with 54.62\% success rate. Designed CNN model has the highest accuracy rate when “minibatchsize” value is equal to 16. Designed CNN + DA model achieves better success rate when “minibatchsize” value is equal to 32. One of the main reasons for the low achievement success rates is needing more data for training in deep learning. Another important factor that may increase success rate is the choice of the correct initial parameters. The success of deep learning based models used in this study and available in the literature is presented in Table II. The studies in this table use standardized split training and test data. The autoencoder based method gave the best result. On the other hand, the success of studies in literature has been achieved with our methods. Although the proposed methods are less successful than the most successful method, architectural installation is simpler and more flexible by using transfer learning supported CNN. Confusion matrix figures for AlexNet and GoogleNet supported CNN are shown as in Figure 9 and Figure 10.

According to results, AlexNet supported CNN model gave the best success rate. Minibatch size based accuracy and loss value of training are shown in Figure 11.

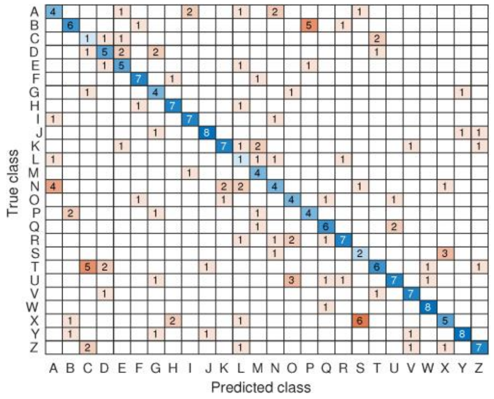


Fig.8. Confusion matrix for best result of designed CNN

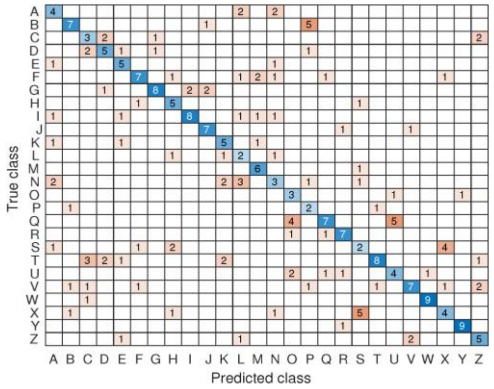


Fig.9. Confusion matrix for best result of AlexNet supported CNN

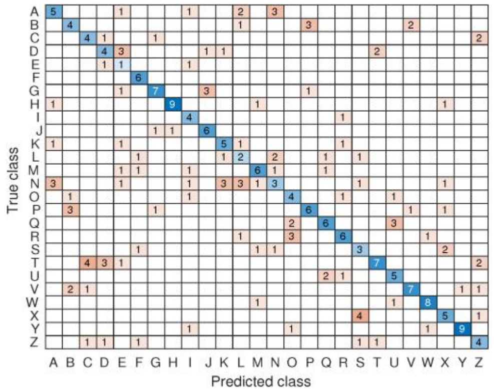


Fig.10. Confusion matrix for best result of GoogleNet supported CNN

TABLE I

ACCURACY COMPARISON OF DESIGNED, ALEXNET AND GOOGLENET SUPPORTED CNN MODELS

	Initial Parameters			
CNN Models	**MiniBatchSize**	**InitialLearnRate**	**MaxEpochs**	**Accuracy**
Designed CNN	8	0.0001	100	0.0346
Designed CNN + DA	8	0.0001	100	0.5423
AlexNet Supported CNN	8	0.0001	100	0.4769
AlexNet Supported CNN+ DA	8	0.0001	100	**0.5462**
GoogleNet Supported CNN	8	0.0001	100	0.3692
GoogleNet Supported CNN+ DA	8	0.0001	100	0.5231
Designed CNN	16	0.0001	100	**0.5231**
Designed CNN + DA	16	0.0001	100	0.4808
AlexNet Supported CNN	16	0.0001	100	0.4423
AlexNet Supported CNN+ DA	16	0.0001	100	0.4808
GoogleNet Supported CNN	16	0.0001	100	0.5192
GoogleNet Supported CNN+ DA	16	0.0001	100	0.5192
Designed CNN	32	0.0001	100	0.4885
Designed CNN + DA	32	0.0001	100	**0.5385**
AlexNet Supported CNN	32	0.0001	100	0.4577
AlexNet Supported CNN+ DA	32	0.0001	100	0.4500
GoogleNet Supported CNN	32	0.0001	100	0.3962
GoogleNet Supported CNN+ DA	32	0.0001	100	0.5000

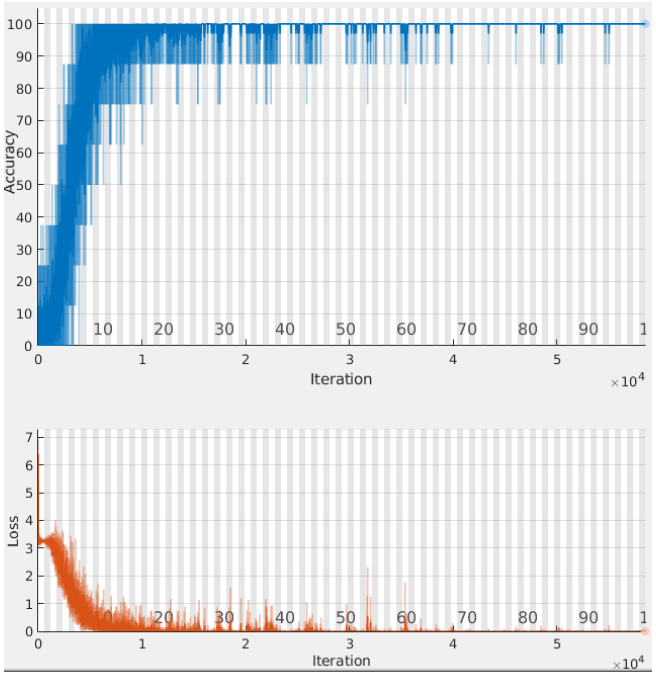


Fig.11. Training progress of AlexNet supported CNN

TABLE II
ACCURACY COMPARISON BETWEEN USED IN THIS PAPER AND APPLIED DEEP LEARNING MODELS

Methods	**Accuracy**
Deep auto-encoder [32]	**64.40%**
CNN [18]	49.90%
CNN & LSTM [18]	57.70%
CNN & bidirectional LSTM [18]	49.40%
Designed CNN & DA	54.23%
AlexNet supported CNN & DA	**54.62%**
GoogleNet supported CNN & DA	52.31%

V. Conclusion

Lip reading, which is a human action recognition subcategory, may be used on various applications such as interaction with deaf people, intelligence services, detection of swearing people in football stadiums and so on. Alphabet level can be acceptable for the first step of lip reading. In this paper, the success of the CNN models supported by the user-designed and pre-trained model on AvLetters dataset has been investigated. Selecting different initial parameters gives different results and there is not an absolute winning model. Using data augmentation increases the success rate generally. In experiments, 8 simple methods are also used for data augmentation on Avletters dataset. Experimental results show that data augmentation is a good way of getting more successful results.

For future studies, other pre-trained models supported CNN can be run on both AvLetters and different datasets. Also, other data augmentation methods can be used for getting a larger dataset.

References

[1] S. Agrawal, V. R. Omprakash, Ranvijay, “Lip reading techniques: A survey,” in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 753–757, July 2016.

[2] A. Garg, J. Noyola, S. Bagadia, “Lip reading using CNN and LSTM,” in Technical Report, 2016.

[3] Y. Li, Y. Takashima, T. Takiguchi, Y. Ariki, “Lip reading using a dynamic feature of lip images and convolutional neural networks,” in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–6, June 2016.

[4] S. Petridis, Z. Li, M. Pantic, “End-to-end visual speech recognition with LSTMs,” CoRR, vol. abs/1701.05847, 2017.

[5] Y. Takashima, Y. Kakihara, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, K. Nakazono, “Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss,” IPSJ Transactions on Computer Vision and Applications, vol. 7, pp. 64–68, 2015.

[6] A. Yargic, M. Dogan, “A lip reading application on MS Kinect camera,” in 2013 IEEE INISTA, pp. 1–5, June 2013.

[7] A. Rekik, A. Ben-Hamadou, W. Mahdi, “A new visual speech recognition approach for RGB-D cameras,” in Image Analysis and Recognition (A. Campilho and M. Kamel, eds.), (Cham), pp. 21–28, Springer International Publishing, 2014.

[8] A. Rekik, A. Ben-Hamadou, W. Mahdi, “Human machine interaction via visual speech spotting,” in Advanced Concepts for Intelligent Vision Systems (S. Battiato, J. Blanc-Talon, G. Gallo, W. Philips, D. Popescu, and P. Scheunders, eds.), (Cham), pp. 566–574, Springer International Publishing, 2015.

[9] A. Rekik, A. Ben-Hamadou, W. Mahdi, “Unified system for visual speech recognition and speaker identification,” in Advanced Concepts for Intelligent Vision Systems (S. Battiato, J. Blanc-Talon, G. Gallo, W. Philips, D. Popescu, P. Scheunders, eds.), (Cham), pp. 381–390, Springer International Publishing, 2015.

[10] M. Emin Yuksel, N. Sarikaya Basturk, H. Badem, A. Caliskan, A. Basturk, “Classification of high resolution hyperspectral remote sensing data using deep neural networks,” Journal of Intelligent & Fuzzy Systems, vol. 34, pp. 2273–2285, 04 2018.

[11] A. Caliskan, M. Yuksel, H. Badem, A. Basturk, “Performance improvement of deep neural network classifiers by a simple training strategy,” Engineering Applications of Artificial Intelligence, vol. 67,pp. 14 – 23, 2018.

[12] H. Badem, A. Basturk, A. Caliskan, M. E. Yuksel, “A new efficient training strategy for deep neural networks by hybridization of artificial bee colony and limited–memory bfgs optimization algorithms,” Neurocomputing, vol. 266, pp. 506 – 526, 2017.

[13] A. Caliskan, M. Yuksel, H. Badem, A. Basturk, “A deep neural network classifier for decoding human brain activity based on magnetoencephalography,” Elektronika ir Elektrotechnika, vol. 23, no. 2, 2017.

[14] A. Fernandez-Lopez, F. M. Sukno, “Survey on automatic lip-reading in the era of deep learning,” Image and Vision Computing, vol. 78, pp. 53 – 72, 2018.

[15] I. Matthews, T. Cootes, J. A. Bangham, S. Cox, R. Harvey, “Extraction of visual features for lipreading,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, p. 2002, 2002.

[16] A. Krizhevsky, I. Sutskever, G. Hinton, “Imagenet classification with deep convolutional neural networks,” NIPS, vol. 25, pp. 1106–1114, 2012.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going deeper with convolutions,” CoRR, vol. abs/1409.4842, 2014.

[18] W. Feng, N. Guan, Y. Li, X. Zhang, Z. Luo, “Audio visual speech recognition with multimodal recurrent neural networks,” in 2017 International Joint Conference on Neural Networks (IJCNN), pp. 681–688, May 2017.

[19] I. Anina, Z. Zhou, G. Zhao, M. Pietikainen, “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, May 2015.

[20] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy, “Moving talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus,” EURASIP J. Appl. Signal Process., vol. 2002, pp. 1189–1201, Jan. 2002.

[21] W. Dong, R. He, S. Zhang, “Digital recognition from lip texture analysis,” in 2016 IEEE International Conference on Digital Signal Processing (DSP), pp. 477–481, Oct 2016.

[22] T. Stafylakis, G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” CoRR, vol. abs/1703.04105, 2017.

[23] J. S. Chung, A. Zisserman, “Lip reading in the wild,” in Asian Conference on Computer Vision, pp. 87–103, Springer, 2016.

[24] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, K. Nakazono, “Audio-visual speech recognition using bimodal trained bottleneck features for a person with severe hearing loss,” in INTERSPEECH, 2016.

[25] E. Kilic, Classification of Mitotic figures with convolutional neural networks. M.Sc. thesis, Erciyes University, Graduate School of Natural and Applied Sciences, 2016.

[26] H. S. Nogay, T. C. Akinci, “A convolutional neural network application for predicting the locating of squamous cell carcinoma in the lung,” Balkan Journal of Electrical and Computer Engineering, vol. 6, pp. 207 – 210, 2018.

[27] H. S. Nogay, “Classification of different cancer types by deep convolutional neural networks,” Balkan Journal of Electrical and Computer Engineering, vol. 6, pp. 56 – 59, 2018.

[28] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, “Recent advances in convolutional neural networks,” CoRR, vol. abs/1512.07108, 2015.

[29] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.

[30] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” CoRR, vol. abs/1512.03385, 2015.

[31] S. Das, “CNNs architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more . . . .” https://medium.com/@siddharthdas-32104, 2017.

[32] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, “Multimodal deep learning,” in Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696, 2011.

BIOGRAPHIES

TAYYIP OZCAN received his B.S. degree in Computer Engineering from Istanbul Kultur University, Istanbul, Turkey, in July 2010. He then joined as a research assistant to the Dept. of Computer Engineering of Erciyes University in 2013. He received his M.Sc. degree in 2016 from Computer Engineering, Erciyes University. He is currently Ph.D. candidate in Erciyes University, Department of Computer Engineering, Kayseri, Turkey. His research areas are intelligent optimization algorithms, image processing, machine learning, deep learning and Kalman filter. Lip reading, hand gesture recognition, human action recognition and human computer interaction are the main topics of his applications.

ALPER BASTURK received his B.S. degree in Electronics Engineering from Erciyes University, Kayseri, Turkey, in July 1998. He then joined as a research assistant to the Dept. of Electronics Eng. of Erciyes University. He received his M.S. and Ph.D. degrees in both of Electronics Engineering from Erciyes University in August-2001 and November 2006, respectively. In 2006, he joined the Computer Hardware Division, Department of Computer Engineering, Erciyes University, where he is currently an Associate Professor. Between 2010 and 2011, he was a visiting scholar to the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York, USA. He guest-edited several special issues for various journals and has published more than fifty articles in leading journals and conferences. His research areas are digital signal and image processing, machine learning, deep learning, neural networks, fuzzy and neuro-fuzzy systems, intelligent optimization and applications of these techniques.