The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2023

Volume 26, Pages 327-331

IConTES 2023: International Conference on Technology, Engineering and Science

Inception Model for Automatic Arabic Speech Recognition

Zoubir Talai
Badji Mokhtar University of Annaba

Nada Kherici
Badji Mokhtar University of Annaba

Abstract: Reproducing basic human abilities has always been the main purpose for Artificial Intelligence (AI) systems. Since speech is essential to people's communication, AI was applied to this major field to achieve Automatic Speech Recognition (ASR). In this paper, we focus on the inception model as a solution for Arabic speech recognition, due to its remarkable results on image classification tasks. We adapted this model for ASR problems and tried it on a dataset of spoken Arabic digits collected from social media apps and published corpora which resulted in more than 54000 utterances. A comparison between the proposed model and a traditional Convolutional Neural Network (CNN) shows the superiority of the inception model in ASR tasks. The inception model achieved 99.70% accuracy on the training dataset which is far better than the traditional CNN that achieved 87.46% on the same set, it did also great performance on the test subset with 88.96% accuracy compared to the traditional model with 84.78% recognition rate.

Keywords: Inception model, Convolution neural network, Artificial intelligence, GoogLeNet

Introduction

Automatic speech recognition (ASR) has grown to be a major field of research these last few years. A great number of commercial products were created based on the use of ASR as a user interface that proved to be useful and provides helpful assistance to humans in all kinds of everyday tasks. Consumer-targeted applications require ASR to be robust facing language dialects and noisy environments which can be challenging for such systems.

For this purpose, scholars issued several propositions using different approaches. Han et al. (2020) proposed ContextNet, a Convolutional Neural Network (CNN) combined with Recurrent Neural Network (RNN) to incorporate global context information into convolution layers with a custom scaling technique for the proposed model to achieve good accuracy in acceptable computation time. Another end-to-end model was proposed by Kriman et al. (2020), where they combined 1D time-channel separable convolution layers with batch normalization and ReLU layers to create blocks that were connected using the residual connection. They trained their model on the LibriSpeech dataset and claimed near state-of-the-art accuracy while needing fewer parameters. CNN was also used by Passricha & Aggarwal (2020) combined with Bidirectional Long Short Term Memory (LSTM) to achieve continuous speech recognition tasks. The authors experimented with different weight-sharing methods and pooling strategies to decrease Word Rate Error (WER). CNN proved also to be very efficient in the work proposed by Shahamiri (2021) where authors trained a CNN model and used it for transfer learning to create an ASR system to help people affected with Dysarthia. This medical condition is a result of paralysis of muscles involved in the articulation process hence making pronunciation for the affected individual very difficult. It was also used by Palaz et al. (2019) to estimate Hidden Markov Model (HMM) states class conditional probabilities. To achieve this goal, the authors trained a CNN model with raw speech signal to do feature extraction and initiate an HMM to finally recognize the spoken word. Houri et al. in (2021)

proposed a new approach to extract speaker characteristics by constructing CNN filters linked to the speaker. They also proposed new vectors to identify speakers, which they called convVectors. Experiments have been performed with a gender-dependent corpus (THUYG-20 SRE) under three noise conditions: clean, 9db, and 0db. Results showed that the convVectors method was the most robust. CNN was also used by Mustaqeem & Kwon (2020) where authors used it to learn discriminative features from whole utterances and then fed them to a Long Short-Term Memory (LSTM) network for sequence learning to find emotions of the speaker. Finally, Kiranyaz et al. (2021) wrote a good paper on 1D convolutional neural networks and their applications showing their advantages in ASR systems.

In this paper, we adapt the Inception Model (also called GoogLeNet) to ASR by modifying its architecture and changing its layers making them suitable for audio signal processing. These changes include one-dimensional (1D) convolution filters and 1D max pooling layers. This paper is organized into four sections. Starting with an introduction that presents the context of the study and some related works. The second section outlines the used CNN's architecture. The next section summarizes the results along with discussions. Finally, we conclude the paper with what we think of the obtained results and perspectives.

1D Convolution Neural Network

Like a conventional 2D CNN that is principally used for image classification, one-dimensional convolution networks are majorly used for signal processing problems such as Automatic speech recognition, Electrocardiogram monitoring, structural damage detection in civil infrastructure based on vibration, predictive maintenance for industrial machines…etc.

1D CNN operates the same way as 2D CNN, the first layer is dedicated to data input which generally consists of raw signal. The next set of layers is a combination of convolution/pooling layers that extract features from that signal. Finally, a fully connected network receives those features and classifies them into several classes. This last one must have in its final layer a number of neurons equal to class's number.

AlexNet

AlexNet is a deep convolutional neural network architecture that was introduced by Krizhevsky et al. (2012). It consists of 8 layers, including 5 convolutional layers and 3 fully connected layers. AlexNet uses traditional stacked convolutional layers with max-pooling in between as shown in figure 1. Its deep network structure allows for the extraction of complex features from sound files. The architecture employs overlapping pooling layers to reduce spatial dimensions while retaining the spatial relationships among neighboring features. It uses the ReLU activation function and dropout regularization, which enhance the model's ability to capture non-linear relationships within the data.

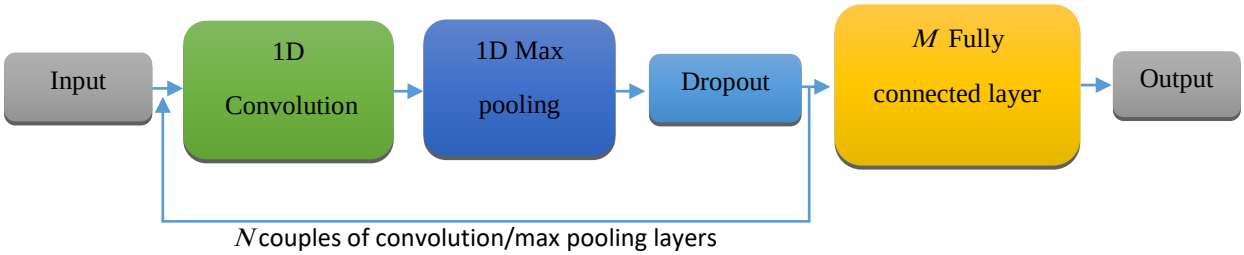


Figure 1. AlexNet architecture for ASR

Inception Model v1 (GoogLeNet)

GoogleNet, also known as Inception v1, was introduced by Szegedy et al. (2015). It was built on the success of AlexNet by introducing a number of new innovations. In GoogleNet architecture, there is a method called global average pooling used at the end of the network. This method reduces the number of parameters in the model and helps prevent overfitting. GoogleNet also introduced inception modules shown in Figure 2, which are designed to capture features at different scales and resolutions. The architecture is deeper than AlexNet and has 22 layers.

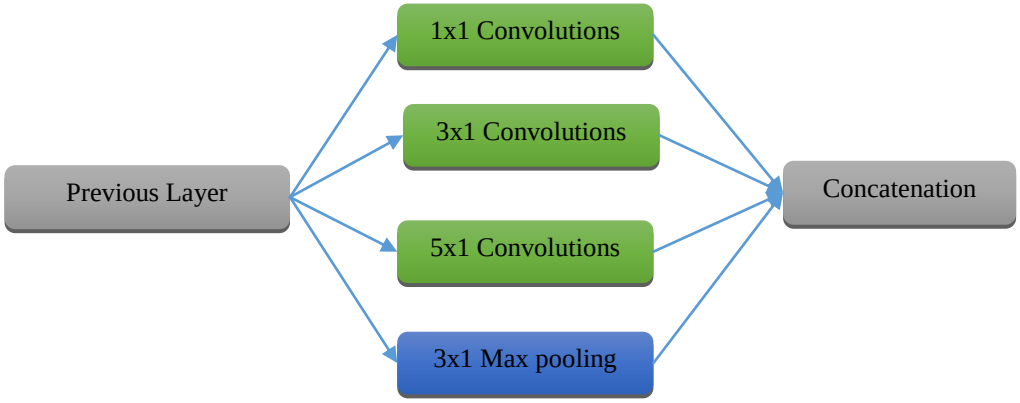


Figure 2. 1D inception block

Dataset

For experimentations, we used a dataset collected for previous work by Talai et al. (2022). We collected utterances of the first ten Arabic digits, from 107 speakers; the recordings were received via Internet and social media apps. A mixed group of people participated in the recording including women, men, and children aged from 4 to 64 years. The sound files were pretreated and converted to have segments of two seconds duration. We also resorted to data augmentation to increase the initial dataset which proved to be too small for the training of the proposed model. This operation resulted in 14 538 samples. The corpus was divided in two parts (80% + 20%), 80% was dedicated to the training/validation of the model, and 20% was kept for testing the models' performance.

Results and Discussion

To achieve good recognition using AlexNet and after several experiments, we chose to progressively increase kernel size and decrease extracted features combined with a dropout layer to avoid overfitting. As for the used GoogLeNet, we used the Inception block present in Table 2, we also chose to add one auxiliary classifier noticeable in the sequence: Average pooling – Conv1D – Dense. Note that the number of features extracted and kernel size are to be chosen meticulously so that the output size of each layer matches the input needed for the next one. Table 1 shows the used architectures.

Table 1. Used architectures

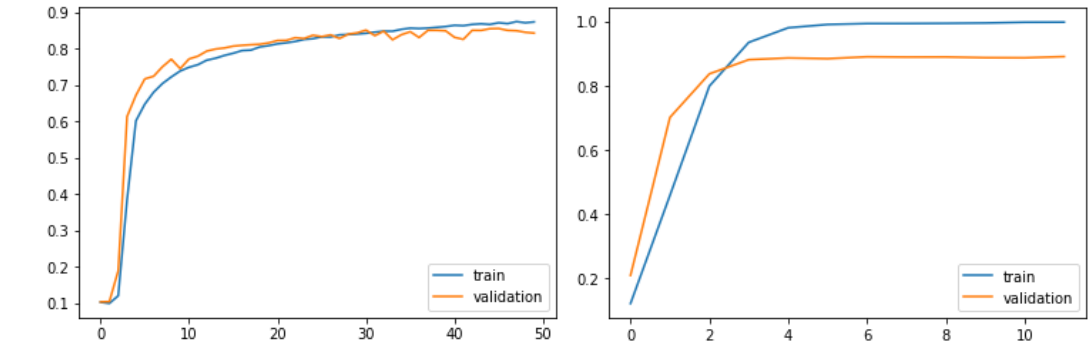
AlexNet	Inception Model
Conv1D (8, 13)	Inception(16,16,16,16,8,16)
Max-Pooling (3)	Inception(16,16,16,16,8,16)
Dropout (0.3)	Average-pooling(5)
Conv1D (16, 11)	Conv1D(32,1)
Max-Pooling (3)	Dense(10)
Dropout (0.3)	Inception(32,32,32,32,16,32)
Conv1D (32, 9)	Inception(32,32,32,32,16,32)
Max-Pooling (3)	Dense(10)
Dropout (0.3)	
Conv1D (64, 7)	
Max-Pooling (3)	
Dropout (0.3)	
Dense (256)	
Dense (128)	
Dense (10)	

After several experimentations, the best model's evolution can be seen in Figure 3. The results presented in Table 3 show the improvement gained using the Inception model. First, AlexNet's accuracy is acceptable considering the nature of input data, the simplicity of implementation, and the arguably small size of the network. Unfortunately, it stops evolving at a certain accuracy even with 50 epochs. On the other hand, the

Inception model outperforms AlexNet's results needing only 12 epochs. The combination of Inception blocks and auxiliary output layer proves that this model is efficient for automatic speech recognition tasks.

Table 2. Inception block architecture

Inception (a, b, c, d, e, f)			
Conv1D(a,1)	Conv1D(b,1) Conv1D(c,3)	Conv1D(d,1) Conv1D(e,5)	Max-Pooling(3) Conv1D(f,1)
Concatenation			



(a) (b)

Figure 3. Models' accuracy evolution: 3a AlexNet. 3b Inception Model

Table 3. Results summary

	AlexNet	**Inception Model**
Batch Size	16	32
Epochs	50	12
Accuracy achieved on training dataset	87.46%	99.70%
Accuracy achieved on the validation dataset	84.78%	88.96%

Conclusion

ASR is an exciting field that has seen significant progress in recent years. With major developments in computing power and the appearance of GPU parallel processing, even the most complex deep model can be trained in an acceptable time. This work introduces a different way of using the Inception model to achieve automatic speech recognition. We detailed the required modification to the model's architecture so it can work with sound data. The Inception model proved to be efficient for ASR tasks without needing long processing time. Inception model proved his superiority achieving 88.96% recognition rate compared to AlexNet which only had 84.78% accuracy.

References

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R., & Wu, Y. (2020). ContextNet: improving ocnvolutional neural networks for automatic speech recognition with global context. *arXiv.*

Hourri, S., Nikolov, N. S., & Kharroubi, J. (2021). Convolutional neural network vectors for speaker recognition. *International Journal of Speech Technology*, *24*(2), 389–400.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398.

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., & Zhang, Y. (2020). Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6124–6128.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Mustaqeem, & Kwon, S. (2020). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, *20*(1), 1.

Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. S*peech Communication, 108*, 15–32.

Passricha, V., & Aggarwal, R. K. (2020). A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. *Journal of Intelligent Systems, 29*(1), 1261–1274.

Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29*, 852–861.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. 2015 I*EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

Talai, Z., Bahi, H., & Kherici, N. (2022). Remote spoken Arabic digits recognition using CNN. In N. M. Mahyuddin, N. R. Mat Noor, & H. A. Mat Sakim (Eds.), Proceedings of the *11th International Conference on Robotics, Vision, Signal Processing and Power Applications* (pp. 639–645). Springer.

Author Information

Zoubir Talai
Badji Mokhtar University of Annaba
P.O Box.12, Annaba-23000 Algeria
Contact e-mail: talai_zoubir@yahoo.com

Nada Kherici
Badji Mokhtar University of Annaba
P.O Box.12, Annaba-23000 Algeria

To cite this article:

Talai, Z. & Kherici., N. (2023). Inception model for automatic arabic speech recognition. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 26*, 327-331.