Single-Word Speech Recognition using Convolutional CNN Neural Networks

Nurbapa Mekebayev[1], Orken Mamyrbayev[2], Dina Oralbekova[3], Madiyar Tasbolatov[4]

[1,2,4] al-Farabi Kazakh National University, Almaty, Kazakhstan
[2] Kazakh national Women's Teacher Training University, Almaty, Kazakhstan
[3] Satbayev University, Almaty, Kazakhstan
[1,2] Institute of Information and Computational Technologies, Almaty, Kazakhstan
Email: [1]nurbapa@mail.ru, [2]morkenj@mail.ru, [3]dinaoral@mail.ru, [4]aizhir.mt@gmail.com

ABSTRACT

This work focuses on monosyllabic speech recognition, where the ultimate goal is to accurately recognize a set of predefined words from short audio clips. It uses a data set of speech commands that consist of 64,000 one-second utterances of 30 short words, from which we learn to classify 10 words, as well as classes for "unknown" words, and also "Silence". We use a convolutional neural network (CNN) with one-dimensional convolusions on the raw audio signal to classify the samples. The results show that the model can predict samples of words it saw during training with high accuracy, but it somewhat struggles with generalizing to words that are beyond the training data, and extremely noisy samples.

Keywords

Speech Recognition, Neural Network, CNN, Audio Signal, MFCC
Article Received: 20 October 2020, Revised: 21 November 2020, Accepted: 13 December 2020

Introduction

Currently, one of the most pressing problems in the field of information technology is the problem of speech recognition. The efficiency of using computer systems directly depends on this, since speech is considered the most common and natural phenomenon of human communication and significantly speeds up the process of entering information and managing mobile systems. Information technologies are rapidly developing and are widely used in information exchange. In this regard, the development of speech recognition plays an important role.

In everyday life, language is a natural means of human communication. Everyone knows that in the course of the development of science and technology, scientists and engineers have been studying the problem of verbal communication between the user and the machine for many years.

It should be recognized that many companies and individual developers have made significant progress in the development of speech recognition technologies, but it should also be recognized that they are still not widely used in Kazakhstan. This is due to the presence of sound interference and the peculiarities of the speaker's speech style.

In this regard, the most urgent task in solving this problem is to identify the speaker in the automatic speech recognition system.

Machine learning is a branch of artificial intelligence, the characteristic feature of which is not the direct solution of a problem, but learning in the process of applying solutions to many similar problems. Machine learning is located at the intersection of mathematical statistics, optimization methods, and classical mathematical disciplines, but it also has its own specifics related to information extraction, data mining, and computational efficiency problems [1]. The scope of application of machine learning technologies is wide and constantly increasing, including pattern recognition, medical and technical diagnostics, statistical analysis, forecasting, management and decision-making tasks, text array processing, and speech synthesis. For example, in linguistics, machine learning helps determine the grammatical characteristics of words; in phonetics, machine learning is used to predict the pitch frequency and length of phonemes. Prosodic characteristics, such as determining the length and location of pauses, and predicting intonations, can also be taught.

Theory And Methods

This section presents the basic theory of using audio data for speech recognition models, the models used in the final architecture, and the basic principles of learning neural networks.

A. Sound as Data

The data set consists of a set of wave files, each of which is approximately one second long. To use this data, each file is sampled into a vector with a sampling rate of 16000. A common speech recognition strategy is to first extract features from the raw waveform. Commonly used speech functions, such as spectrograms, log-Mel filter banks, and Mel-frequency cepstral coefficients (MFCC), convert the raw waveform to the time-frequency domain [2]. These functions are then used as input to the model. [3] show how log-Mel filter banks can be used as input features for a neural network.

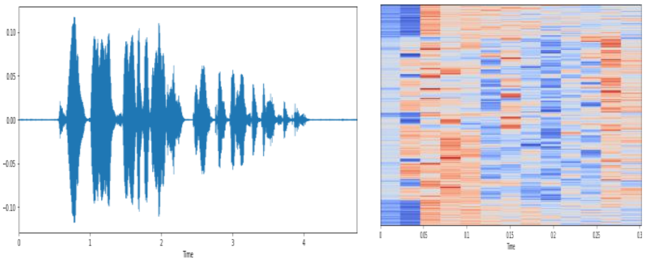

Fig. 1. Sample showing the word as a raw waveform (left) and MFCC (right)

B. Convolutional cnn neural networks for speech recognition

The most common machine learning models are artificial neural networks, which are used to process input data from a combination of distributed simple operations that depend on the parameters being trained. Modern neural networks are used in forecasting tasks, image and speech recognition, text generation, and many others. The most optimal types are convolutional networks and recurrent neural networks.

Convolutional neural networks consist of four basic operations: convolution, nonlinearity, union and classification. Optionally, models can include batch normalization as well as dropout. These operations are usually added together so that convolutions are followed by non-linearity such as ReLU, then this operation is repeated several times, after which the join operation is used. When the network is deep enough and the original input signal is counted by pooling to a size that can be controlled, it is passed to the classification part of the network. If batch normalization is used, it is usually used after convolution, but before non-linearity.

C. Architecture

A convolutional neural network receives a sequence of raw input signals, splits them into frames, and outputs a score for each class, for each frame. The network architecture consists of several filtering stages, followed by a classification stage. The filter stage includes a convolutional layer, followed by a time layer of maximum Union and non-linearity (tanh ()). Our optimal architecture included three filtering stages. The processed signals coming out of these stages go to the classification stage, which in our case is a multi-layer perceptron with one hidden layer. It outputs conditional probabilities p (i/x) for each class i, for each frame x, using the Soft Max layer [19]. The network is trained by the cross-entropy criterion maximized by the stochastic gradient ascent algorithm [20].

D. Convolutional layer

While «classical» linear layers in standard MLP accept a fixed-size input vector, it is assumed that the convolutional layer is fed by a sequence of t vectors frames: $X = \{x_1, x_2, \dots, x_T\}$.. The convolutional layer applies the same linear transformation over each successive (or overlapping dR frames) window of kR frames. For example, the transformation in frame t is formally written as:

$$N \begin{pmatrix} x^{t-(kR-1)/2} \\ \vdots \\ x^{t+(kR-1)/2} \end{pmatrix} \qquad (1)$$

Where, N is a $d_{out} \times d_{in}$ matrix of parameters. In other words d_{out} filters (rows of the matrix N) are applied to the input sequence.

E. Max-pooling layer

These kinds of layers perform local time maximum operations on the input sequence. More formally, the transformation in frame t is written as:

$$\max_{t-(kR-1)/2 \leq 6 \leq t+(kR-1)/2} x_6^d \qquad (2)$$

with x being the input, kR the kernel width and d the dimension

Data

Set consists of 64,000 one-second utterances of 25 short words such as “кел”,“бар”,“аяқ” and "қол". This data set is designed to help build voice interfaces for applications with the definition of keywords that can be useful on mobile devices and microcontrollers.

The goal is to classify the following words: "кел", "бар", "жүгір", "сақтан", "аяқ", " қол", "орман"," тоғай"," алыс "and " жақын ". All other words are marked as "unknown" and are used to help the model learn a representation for all words that are not included in the 10 words to be classified. The last class is "silence", which is samples without a word.

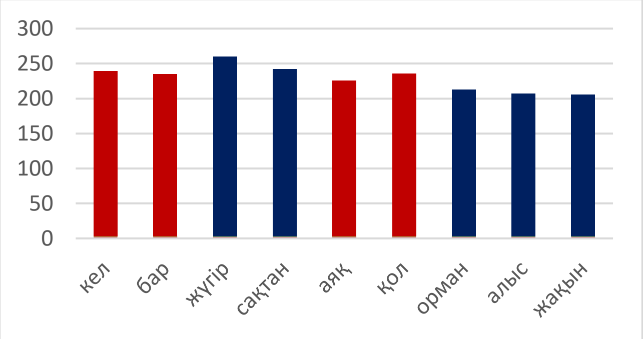

Fig. 2. Calculations for each class in the data set of voice commands. The blue examples represent words that need to be classified, the red examples represent an unknown class

The test set provided by the Kazakh speech corpus consists of approximately 150,000 additional utterances, including words and speakers that do not appear in the speech command dataset.

The files in the dataset are arranged in folders by label, with file names starting with a hash representing the speaker, followed by a number representing the number of times a statement by the same speaker appears in the dataset. The data is split into training and validation sets using a hash, so that the same speaker does not appear in both sets.

The Kazakh speech corps also provided a set of audio files that are designated as background noise, using these files, we create data for the silence class, as well as background noise that is used during training. Some files that were recognized as silent but were incorrectly marked were corrected and used as additional silence during training.

A. Data augmentation

Is a way to increase the amount of training data by modifying the available data so that it can still be identified with the same label. Increasing the amount of data showed that this is a simple and effective way to reduce overvoltage and thereby improve the performance of the model. Increasing the data can also help the model learn a wider range of features, since the extended samples can be very different from the original training samples, but they can still be identified as the same.

Three methods of data augmentation were used: time-shifting the sound, scaling the amplitude, and adding noise. The first step, time offset, is applied with a probability of 49.8 % to each sample. Enlarged samples are shifted forward or backward in time up to 19.6 % of the original length of the samples. Sometimes this results in a partial excision of the utterance in the sample, but the probability of this is negligible. This augmentation method should help the model learn a more time-invariant representation of statements, since they can appear anywhere in the sample.

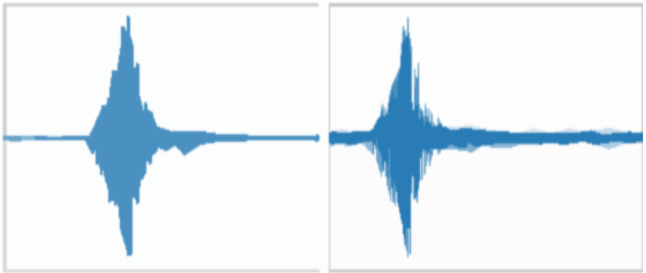

Fig. 3. Sample of the word " Кел” as the original sample (left) and its augmented version (right)

The second amplification method is mixing noise with a sample, this is also applied with a probability of 49.8 %. Noise is generated in the same way as silence is generated during training, when up to two samples from the background sound of the data set are scaled and added together. This noise is then added to the original input, which scales between 74.9 % and 124.6% of its original amplitude. Adding noise should help the model better distinguish relevant information from data.

Results

The evaluation of results is divided into three sets, a validation Set, and two sets of tests. The validation set contains approximately 10% of the samples in the dataset. These two test suites are based on two test leaderboards, public and private, which contain 30% and 70% of the test data, respectively, for clarity, they will be designated as a public test Suite and a private test Suite. Accuracy, accuracy, recall, and F1-scores are reported for validation data, but since we don't have access to the correct labels for test suites, only accuracy scores will be reported for them. Some analysis of manually validated predictions on the test set will be discussed, as the test sets contain words and speakers that are missing from the training data, and analyzing the predictions will provide a deeper understanding of the strengths and weaknesses of the model.

Before using pseudo-tags in the training process, the model achieves 96.5% accuracy on the validation set, 86.6% accuracy on the public test set, and 87.8% accuracy on the private test set. After including pseudo-labels in the training process, the model achieves 96.7% accuracy on the validation set, 87.8% accuracy on the public test set, and 88.6% accuracy on the private test set.

TABLE 1. FOR EACH CLASS THE ACCURACY, RECALL AND F1-SCORES FOR THE VALIDATION SET ARE SHOWN IN

Lebel	Percision	Recall	F1-score	Support
кел	0,98	0,98	0,98	236
бар	0,95	0,97	0,96	232
жүгір	0,93	0,96	0,95	257
сақтан	0,96	0,97	0,98	239
аяқ	0,96	0,97	0,97	223
қол	0,94	0,96	0,95	233
орман	0,87	0,97	0,94	210
алыс	0,93	0,96	0,94	204
жақын	0,94	0,95	0,93	203

Conclusion

This work showed that a model that uses one-dimensional convolutions along with reasonable learning methods can be used effectively to recognize monosyllabic speech, but a lot of work needs to be done to generalize to invisible samples and work with extremely noisy samples.

Acknowledgements

AP08855743 «Development of an end-to-end automatic speech recognition system for agglutinative languages»

References

[1] Abdrahmanova, Je. R. (2019) ‘Mashinnoe obuchenie: Obzor i primenenie’.Tehnika i tehnologii: Puti innovacionnogo razvitija: sbornik nauchnyh trudov 8-j Mezhdunarodnoj nauchnoprakticheskoj konferencii. Jugo-Zapadnyj gosudarstvennyj universitet [Machine Learning: Overview and Application.Technique and Technology: Ways of Innovative Development: a collection of scientific papers of the 8th International Scientific and Practical Conference. Southwestern State University]. Kursk. Pp. 9-12.

[2] Zhang, Y., Suda, N., Lai, L. & Chandra, V. 2017, Hello Edge: Keyword Spotting on Microcontrollers, CoRR, vol. abs/1711.07128.

[3] Sainath, T.N. and Parada, C. 2015, Convolutional neural networks for small-footprint keyword spotting. In Sixteenth Annual Conference of the International Speech Communication Association.

[4] J. Bridle, “Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition,” in Neuro-computing: Algorithms, Architectures and Applications, 1990, pp. 227–236.

[5] L. Bottou, “Stochastic gradient learning in neural networks,” in Proceedings of Neuro-Nmes 91. Nimes, France: EC2, 1991.

[6] Orken Mamyrbayev, Turdalyuly, Nurzhamal Oshanova,Tolga Ihsan Medeni, Aigerim Yessentay. Voice Identification Using Classification Algorithms // We are IntechOpen, the world’s leading publisher ofOpen Access books Built by scientists, for scientists. June 25, 2019. London.

[7] Kalimoldayev, M., Mamyrbayev, O., Mekebayev, N., Kydyrbekova, A. Algorithms for detection gender using neural networks // International Journal of Circuits, Systems and Signal Processing. 2020

[8] Mamyrbayev, O., Alimhan, K., Zhumazhanov, B., Turdalykyzy, T., Gusmanova, F. End-to-End Speech Recognition in Agglutinative Languages Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioin formatics), 2020

[9] Aizat, K., Mohamed, O., Orken, M., Ainur, A., Zhumazhanov, B. Identification and authentication of user voice using DNN features and i-vector // Cogent Engineering. 2020

[10] Mamyrbayev O., Toleu A., Tolegen G., Mekebayev N. Neural Architectures for Gender Detection and Speaker Identification // Cogent Engineering, ISSN: 2331-1916. – 2020. Volume 7, - Issue 1