Entropy Based Feature Pooling in Speech Command Classification

Christoforos Nalmpantis[1](✉), Lazaros Vrysis[1], Danai Vlachava[2], Lefteris Papageorgiou[3], and Dimitris Vrakas[1]

[1] Aristotle University of Thessaloniki, Thessaloniki, Greece
{christofn,dvrakas}@csd.auth.gr, lvrysis@auth.gr
[2] International Hellenic University, Thessaloniki, Greece
[3] Entranet Ltd, Thessaloniki, Greece
papageorgiou@entranet.gr

Abstract. In this research a novel deep learning architecture is proposed for the problem of speech commands recognition. The problem is examined in the context of internet-of-things where most devices have limited resources in terms of computation and memory. The uniqueness of the architecture is that it uses a new feature pooling mechanism, named entropy pooling. In contrast to other pooling operations, which use arbitrary criteria for feature selection, it is based on the principle of maximum entropy. The designated deep neural network shows comparable performance with other state-of-the-art models, while it has less than half the size of them.

Keywords: Entropy pooling · Speech classification · Convolutional neural networks · Deep learning

1 Introduction

Internet-of-Things emerged from the amalgamation of the physical and digital world via the Internet. Billions of devices that are used in our daily life, are connected to the internet. Our environment is surrounded by mobile phones, smart appliances, sensors, Radio Frequency Identification (RFID) tags and other pervasive computing machines, which communicate with each other and most importantly with humans. From the humans perspective the most natural way to communicate is by speaking. Speech recognition has been one of the most difficult tasks in artificial intelligence and machine-to-human user interfaces have been restricted so far to other options such as touch screens. Yet, two technological advancements paved the way for more friendly user interfaces based on sound.

The first technological advancement is the rise of multimedia devices like smart phones. Especially the development of digital assistants and their incorporation not only in mobile phones but also in smart home or smart car kits, has established the need for audio based interactions with humans. The second advancement is the Deep Learning revolution in many applications of artificial intelligence.

K. Arai (Ed.): *Intelligent Computing*, LNNS 285, pp. 1083–1091, 2021.
https://doi.org/10.1007/978-3-030-80129-8_71

Deep neural networks have shown a tremendous success in many domains including, but not limited to, computer vision, natural language processing, speech recognition, energy informatics, health informatics etc. Such models are already applied to real world applications such as medical imaging [15], autonomous vehicles [4], activity recognition [8], energy disaggregation [11] and others. Speech recognition is not an exception and there is an increasing interest in audio based applications that can run on embedded or mobile devices [13].

Some examples of sound recognition tasks are automatic speech recognition (ASR), speech-to-text (STT), speech emotion classification, voice commands recognition, urban audio recognition and others. For several years, researchers were trying to manually extract features from sound that are relevant to the task. Thus, the traditional pipeline of such systems includes a preprocessing step, feature extraction and a learning model [14,20]. The first two steps mainly include unsupervised signal processing techniques, extracting information in the frequency domain, exploiting frame-based structural information and others [1]. Recently, deep neural networks have demonstrated unprecedented performance in several audio recognition tasks, outperforming traditional approaches [5,16,17,20].

This research focuses on the challenge of voice commands classification task. Despite the fact that ASR has reached human performance, such models are gigantic and would not fit on a device with limited resources. Moreover, ASR is not so robust in a real world environment where noise is present in many unexpected ways. Thus, a more direct, computationally efficient and resilient to noise system is required. A solution to the voice command classification task seems promising in both achieving an acceptable performance and meeting the aforementioned requirements.

In this manuscript a novel 2D convolutional neural network is developed, utilizing a recent pooling operation named entropy pool [10] and applied on the Speech Commands dataset [18]. The paper is organized as follows. Firstly, previous work on this task is presented. Next, there is a detailed description of the proposed system and all aspects of the experimental arrangement. Afterwards the experimental results are demonstrated and analysed. Finally, conclusions and future research directions are presented.

2 Related Work

Recently, deep learning approaches have demonstrated superior performance than classic machine learning systems in various audio classification tasks. Until now there has been a race on achieving state-of-the-art performance in terms of accuracy on specific tasks. This lead to the development of huge neural networks with millions or billions of parameters that are prohibitive for resource-constrained and real-time systems. In this context, researchers now put their effort on improving the efficiency of deep neural networks.

Recent interest in deploying speech recognition models on the edge has lead to new work on ASR model compression [9] and other sound recognition tasks.

Coucke et al. [3] developed a model utilizing dilated convolution layers, allowing to train deeper neural networks that fit in embedded devices. It is worth noting that the dataset that they created, named "Hey Snips" is public with utterances recorded by over 2.2K speakers. Kusupati et al. [7] proposed a novel recurrent neural network (RNN) architecture named FastGRNN, which includes low-rank, sparse and quantized matrices. This architecture results in accurate models that can be up to 35x smaller than state-of-the-art RNNs. FastGRNN was tested on a variety of datasets and tasks including speech, images and text. The scope of this research was to build models that can be deployed to IoT devices efficiently. Zeng et al. [19] proposed a neural network architecture called DenseNet-BiLSTM for the task of keyword spotting (KWS) using the Google Speech Command dataset. Their main contribution was the combination of a new version of DenseNet named DenseNet-Speech and BiLSTM. The former component of the architecture captures local features whereas maintaining speech time series information. The latter one learns time series features. Solovyev et al. [13] used different representations of sound such as Wave frames, Spectrograms, Mel-Spectograms and MFCCs and designed several neural network architectures based on convolutional layers. Two of their best performing networks had very similar architecture with VGG [12] and ResNet [6]. The models were evaluated on the Google Speech Command dataset, showing very strong results with accuracy over 90%.

In this research a novel neural network architecture has been developed. The proposed model is based on convolutional layers and pooling operations, has six convolutional layers and is more efficient than other deep architectures like VGG and ResNet, which usually have more than 12 layers. As a strong baseline Solovyev's et al. models are used. The experiments show that the proposed model performs on par with the deep models but using much less computation power.

3 Materials and Methods

3.1 Dataset

The Speech Commands dataset [18] has become a standard data source for training and evaluating speech command classification models that are targeted for devices with constraint resources. The primary real world application of such models are scenarios where a few target works have to be recognised in an unpredictable and noisy environment. The challenge in this type of problems is to achieve as very few false positives while restricting the energy consumption as much as possible.

One common scenario is the recognition of keywords that trigger an interaction like the keywords "Hey Google", "Hey Siri", etc. It is obvious that such devices are usually placed in houses or offices where the presence of human conversations is strong and there are many other noises that the devices should ignore.

The training data consists of 60K audio clips with size around 1 s. In total there are 32 different labels, from which only 10 are the target ones. The rest of the labels are considered as silence or unknown. The target labels are left, right, up, down, yes, no, go, stop, on, off. Figure 1 illustrates a pie chart which shows the proportion of each target command in the training set. The audio files are 16-bit PCM-encoded WAVE files with sampling rate 16K.

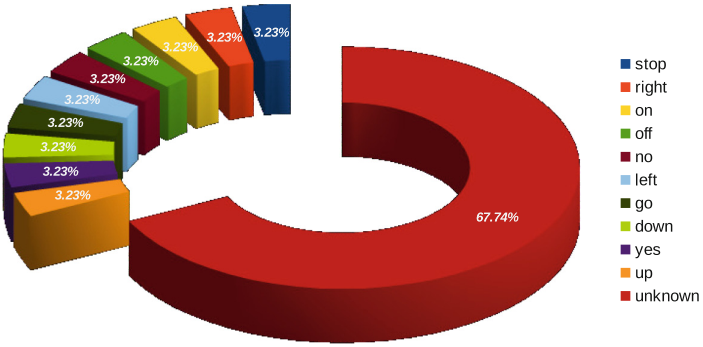


Fig. 1. Proportions of target commands in speech commands training dataset.

3.2 Preprocessing and Audio Features

Audio is in the form of a time series, but it is very common to convert it to a representation in the frequency domain. The most popular sound representations are spectogram, log-mel spectrogram and MFCC. Spectogram is used as the main representation in this research. It is computed using the algorithm of Short Time Fourier Transform (STFT). STFT has the advantage of Fourier transformation converting small segments of a time series to the frequency domain, whereas at the same time preserves temporal information. STFT has three non-default inputs: the signal that will be transformed, the frame length and the stride. The latter one determines how much consecutive windows will overlap each other. The output is a matrix of complex numbers from which we get an energy spectrogram. The spectrogram is extracted using the magnitude of the complex numbers and then taking the logarithm of these values. In the final feature set the angle of the complex numbers is also considered, which improves the accuracy of the final model.

3.3 Entropy Pooling

Feature pooling has been an established layer that helps in sub sampling features with high cardinality. The two most popular pooling operations in deep learning are max and average. However, choosing the right pooling operation is mainly done through a series of experiments and based on the final performance of the model.

To the best of the author's knowledge, in the literature there are two main efforts that shed light on the properties of these two mechanisms. Firstly, Boureau et al. [2] presented a theoretical analysis and described the statistical properties of max and average feature pooling for a two-class categorization problem. The analysis evaluated the two methods in terms of which properties affected the models performance in separating two different classes. The most important outcome was that among other unknown factors, experiments showed that the sparsity and the cardinality of the features affect the model's performance. More recently, Nalmpantis et al. [10] investigated feature pooling operations from the information theory point of view. The authors showed theoretically and empirically that max pooling is not always compatible with the maximum entropy principle. In practice a model's performance can vary a lot with different weight initialization. On the contrary average pooling gives more consistent results because it will always give a more uniform feature distribution. In this context, a novel pooling operation, named entropy pooling, was presented with guarantees to select features with high entropy.

Entropy pooling calculates the probabilities p of the features with cardinality N. Then, the values of probabilities are spatially separated using a kernel and a stride size. For each group the most rare feature is selected. Given a group with size r the mathematical formula is:

$$f_{entr}(X_r) = X_r[g(P_r)], \tag{1}$$

$$g(P_r) = \operatorname*{arg\,min}_{1 \leq i \leq r} p_i, \tag{2}$$

where X_r is the input feature map and P_r the constructed map of probabilities.

3.4 Neural Network Topology

For the problem of speech command recognition recent research has borrowed popular architectures from computer vision such as VGG16, VGG19, ResNet50, InceptionV3, Xception, InceptionResnetV2 and others [13]. These neural networks all have in common the utilization of convolutional neural networks. In this research convolutional layers are included but with the scope to meet the following requirements. Firstly the model has to be smaller than the original models that were introduced for cloud based applications. In the context of speech command recognition, the developed model has to be deployed in devices with limited storage. Moreover, a reduced size also means less computational complexity which is essential for embedded devices or devices with constraint resources. Furthermore, these devices also depend on a battery, magnifying the need for more energy-efficient models. Finally, regarding the input data of the model, they will often be very short and most of the time irrelevant sounds. This means that false positives have to be eliminated.

It is not a surprise that a larger architecture usually achieves better results in terms of accuracy at the expense of computation. After trying many different configurations, a neural network architecture with a satisfying trade off between best performance and efficiency has been developed. It consists of six 2D convolutional layers with activation function ReLU. Each of the first four convolutional layers is followed by a batch normalization layer and an entropy pooling operation. Figure 2 shows the details of the architecture.

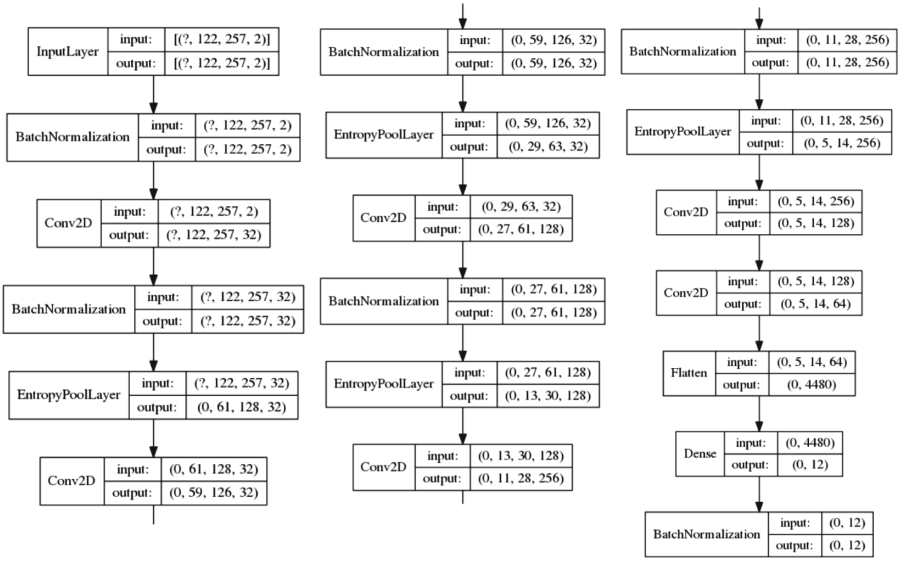


Fig. 2. The proposed neural network with 2D convolutions and entropy pooling operations.

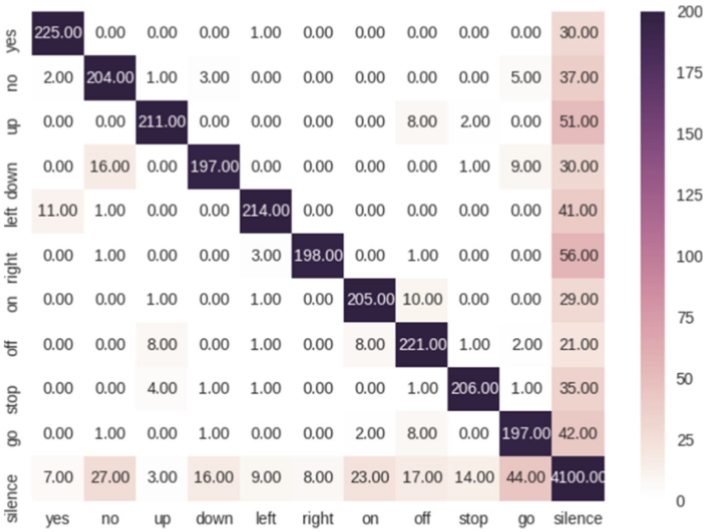
4 Experimental Results and Discussion

The evaluation of the proposed neural network was done using accuracy and cross entropy error. The latter one is the result of the cross entropy of the model's output and the target. Accuracy is more intuitive and estimates the total number of correctly classified instances to the total number of samples. The model achieves test error 0.398 and accuracy 90.4%. Other state-of-the-art models perform slightly better with accuracy around 94%, but the size of these neural networks is multiple times larger [13]. To give an example a modification of the popular deep neural network VGG [12] named VGG16 includes 16 convolutional layers in contrast to the current solution which involves only six and its performance is around 2% better.

For a more in depth analysis of the performance of the model, recall, precision and f1 score are also employed. Recall is defined as the number of correctly predicted positive observations divided by the false and true positive observations. It gives a percentage of total commands that should be recognized. Precision is defined as the true positives divided by the total number of true and false positives. It gives insight of how many of the commands are actually true. This metric is quite important because it is directly affected by the false positives, which is very critical in the real world, as explained previously. Finally f1 score is the harmonic mean of precision and recall. Table 1 presents a detailed classification report for each of the commands. According to the table the command "yes" is the easiest to detect whereas the worst performance is shown for the command "go". The latter one, as shown in the confusion matrix in Fig. 3, is mixed with silence.

Table 1. Analytical classification report.

	Yes	No	Up	Down	Left	Right	On	Off	Stop	Go	Silence	Macro	Micro
Precision	0.92	0.82	0.93	0.90	0.93	0.96	0.86	0.83	0.92	0.76	0.92	0.89	0.90
Recall	0.88	0.81	0.78	0.78	0.80	0.76	0.83	0.84	0.83	0.78	0.96	0.82	0.90
F1-score	0.90	0.81	0.84	0.84	0.86	0.85	0.85	0.84	0.87	0.77	0.94	0.85	0.90



	yes	no	up	down	left	right	on	off	stop	go	silence
yes	225.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	30.00
no	2.00	204.00	1.00	3.00	0.00	0.00	0.00	0.00	0.00	5.00	37.00
up	0.00	0.00	211.00	0.00	0.00	0.00	0.00	8.00	2.00	0.00	51.00
down	0.00	16.00	0.00	197.00	0.00	0.00	0.00	0.00	1.00	9.00	30.00
left	11.00	1.00	0.00	0.00	214.00	0.00	0.00	0.00	0.00	0.00	41.00
right	0.00	1.00	0.00	0.00	3.00	198.00	0.00	1.00	0.00	0.00	56.00
on	0.00	0.00	1.00	0.00	1.00	0.00	205.00	10.00	0.00	0.00	29.00
off	0.00	0.00	8.00	0.00	1.00	0.00	8.00	221.00	1.00	2.00	21.00
stop	0.00	0.00	4.00	1.00	1.00	0.00	0.00	1.00	206.00	1.00	35.00
go	0.00	1.00	0.00	1.00	0.00	0.00	2.00	8.00	0.00	197.00	42.00
silence	7.00	27.00	3.00	16.00	9.00	8.00	23.00	17.00	14.00	44.00	4100.00

Fig. 3. Confusion matrix with results of the proposed neural network recognizing 10 different speech commands.

5 Conclusion

The problem of speech command recognition is a critical one for the success of personal assistants and other internet of things devices.In this research a novel neural network has been developed, utilizing 2D convolutional layers and a pooling operation which is based on entropy instead of randomly selecting audio features. The proposed model meets the real world requirements and can be used by a product with confidence. It achieves a descent performance when compared to larger models while at the same time it is energy efficient and computationally lightweight. For future work it is recommended to examine the performance of entropy pool in larger neural networks. Researchers are advised to conduct further experiments on more datasets and different sound recognition tasks.

Acknowledgement. This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code:T1EDK-00343(95699) - Energy Controlling Voice Enabled Intelligent Smart Home Ecosystem).

References

1. Bountourakis, V., Vrysis, L., Konstantoudakis, K., Vryzas, N.: An enhanced temporal feature integration method for environmental sound recognition. In: Acoustics, vol. 1, pp. 410–422. Multidisciplinary Digital Publishing Institute (2019)
2. Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 111–118 (2010)
3. Coucke, A., Chlieh, M., Gisselbrecht, T., Leroy, D., Poumeyrol, M., Lavril, T.: Efficient keyword spotting using dilated convolutions and gating. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6351–6355 (2019)
4. Fayyad, J., Jaradat, M.A., Gruyer, D., Najjaran, H.: Deep learning sensor fusion for autonomous vehicle perception and localization: a review. Sensors **20**(15), 4220 (2020)
5. Han, W., et al.: Contextnet: improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint* arXiv:2005.03191 (2020)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Kusupati, A., Singh, M., Bhatia, K., Kumar, A., Jain, P., Varma, M.: Fastgrnn: a fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. In: Advances in Neural Information Processing Systems, pp. 9017–9028 (2018)
8. Lentzas, A., Vrakas, D.: Non-intrusive human activity recognition and abnormal behavior detection on elderly people: a review. Artif. Intell. Rev. **53**, 1975–2021 (2020). https://doi.org/10.1007/s10462-019-09724-5
9. McGraw, I., et al.: Personalized speech recognition on mobile devices. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5955–5959. IEEE (2016)
10. Nalmpantis, C., Lentzas, A., Vrakas, D.: A theoretical analysis of pooling operation using information theory. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1729–1733. IEEE (2019)

11. Nalmpantis, C., Vrakas, D.: On time series representations for multi-label NILM. Neural Comput. Appl. **32**, 17275–17290 (2020). https://doi.org/10.1007/s00521-020-04916-5
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556 (2014)
13. Solovyev, R.A., et al.: Deep learning approaches for understanding simple speech commands. In: 2020 IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), pp. 688–693. IEEE (2020)
14. Tsipas, N., Vrysis, L., Dimoulas, C., Papanikolaou, G.: Mirex 2015: Methods for speech/music detection and classification. In Processing, Music information retrieval evaluation eXchange (MIREX) (2015)
15. Viswanathan, J., Saranya, N., Inbamani, A.: Deep learning applications in medical imaging: Introduction to deep learning-based intelligent systems for medical applications. In: Deep Learning Applications in Medical Imaging, pp. 156–177. IGI Global (2021)
16. Vrysis, L., Thoidis, I., Dimoulas, C., Papanikolaou, G.: Experimenting with 1d CNN architectures for generic audio classification. In: Audio Engineering Society Convention 148. Audio Engineering Society (2020)
17. Vrysis, L., Tsipas, N., Thoidis, I., Dimoulas, C.: 1d/2d deep cnns vs. temporal feature integration for general audio classification. J. Audio Eng. Soc. **68**(1/2), 66–77 (2020)
18. Warden, P.: Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint* arXiv:1804.03209 (2018)
19. Zeng, M., Xiao, N.: Effective combination of densenet and bilstm for keyword spotting. IEEE Access **7**, 10767–10775 (2019)
20. Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, Amr, E.D., Jin, W., Schuller, B.: Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans. Intell. Syst. Technol. **9**(5), 28 p. (2018). https://doi.org/10.1145/3178115. Article 49