Keyword Spotting

Stefano Ivancich
Department of
Information Engineering
University of Padova, Italy
Email: stefano.ivancich@studenti.unipd.it

Luca Masiero
Department of
Information Engineering
University of Padova, Italy
Email: luca.masiero.8@studenti.unipd.it

***Abstract*—Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla est purus, ultrices in porttitor in, accumsan non quam. Nam consectetur porttitor rhoncus. Curabitur eu est et leo feugiat auctor vel quis lorem. Ut et ligula dolor, sit amet consequat lorem. Aliquam porta eros sed velit imperdiet egestas. Maecenas tempus eros ut diam ullamcorper id dictum libero tempor. Donec quis augue quis magna condimentum lobortis. Quisque imperdiet ipsum vel magna viverra rutrum. Cras viverra molestie urna, vitae vestibulum turpis varius id. Vestibulum mollis, arcu iaculis bibendum varius, velit sapien blandit metus, ac posuere lorem nulla ac dolor. Maecenas urna elit, tincidunt in dapibus nec, vehicula eu dui. Duis lacinia fringilla massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut consequat ultricies est, non rhoncus mauris congue porta. Vivamus viverra suscipit felis eget condimentum. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer bibendum sagittis ligula, non faucibus nulla volutpat vitae. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In aliquet quam et velit bibendum accumsan. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Vestibulum vitae ipsum nec arcu semper adipiscing at ac lacus. Praesent id pellentesque orci.**

I. Introduction

The goal of *keyword spotting* is to detect a relatively small set of predefined keywords in a stream of user utterances, often in the context of an intelligent agent on a mobile phone.

With the rapid development of mobile devices, speech-related technologies are nowadays becoming very popular. For example, Google gives us the chance to search by voice on Android phones, while personal assistants (Google Now, Apple's Siri, Amazon's Alexa or Microsoft's Cortana) use speech recognition to interact with these systems. Google has enabled a fully hands-free speech recognition experience, known as "Ok Google", which continuously listens to what we say for specific keywords to initiate voice input.

Keyword spotting can be used to detect "command triggers" ("Hey Siri" for example) that provide explicit cues for interactions between the device and the human that gave such cue. It is also desirable that such models have a small footprint (for example, measured in the number of model parameters) so they can be deployed on low power and performance-limited devices.

In recent years, neural networks have been shown to provide effective solutions to the small-footprint keyword spotting problem. Research typically focuses on a tradeoff between achieving high accuracy and having a small footprint.

In this work we focus on recurrent neural networks (RNNs), a ground-breaking advance in deep learning, a class of models that has been successfully applied to small-footprint keyword spotting in recent years.

SPIEGARE PERCHè LE RNN VANNO BENE PER QUESTO TASK

For this project we use the recently-released Google Speech Command Dataset as our benchmark.

(AGGIUNGERE QUALCOSA DA INTRO DI Attention-Based Models ... SOLO SE SERVE)

DA METTERE ALLA FINE DELL'INTRODUZIONE: This paper is structured as follows. In Section 2 we give an overview of the related work of the Keyword Spotting Problem, Section 3 presents the model implentation. The experimental setup is described in Section 4, while results comparing... are presented in Section 5. Finally, Section 6 concludes the paper and discusses future works.

June 29, 2020

A. Subsection Heading Here

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

II. Related work

The first system similar to a modern ASR was built in the 1952 by researchers at the Bell laboratories and was able to recognize numerical digits from speech using formats of the input audio. These are a concentration of the acoustic energy around a particular frequency in the input file wave. For the next thirty years, various researchers developed devices capable of recognizing vowels and consonants using different types of features like *phonemes*, until the introduction, in the mid 1980s of the Hidden Markov Models (HMM).

This approach represented a significant shift from the simple pattern recognition methods that were based on templates (and a spectral distance measure), to a statistical method for speech processing and was possible thank to the incredible advances in the computer computational power obtained(?) in those years. In recent years, however, the HMMs faced the challenge of the introduction of Deep Learning and several architechtures that work well with these type of problems like Convolutional Neural Networks (CNN) shift-invariant in the data representation domain, and Recurrent Neural Networks (RNN) and their ability to store information.

III. Model implementation

This section describes our base model and its variants. All code necessary to replicate our experiment has been made oper source in our Github repository[1].

IV. Evaluation

DA MODIFICARE We evaluated our models using Google's Speech Commands Dataset, which was released in August 2017 under a Creative Commons license[2]. The dataset contains 65,000 one-second long utterances of 30 short words by thousands of different people, as well as background noise samples such as pink noise, white noise, and human-made sounds.

V. Experimental Setup

VI. Results

VII. Conclusion

PRENDI SPUNTO DA QUESTO: This paper describes the application of deep residual learning and dilated convolutions to the keyword spotting problem. Our work is enabled by the recent release of Google's Speech Commands Dataset, which provides a common benchmark for this task. Previously, related work was mostly incomparable because papers relied on private datasets. Our work establishes new, state-of-the-art, open-source reference models on this dataset that we encourage others to build on. For future work, we plan to compare our CNN-based approaches with an emerging family of models based on recurrent architectures. We have not undertaken such a study because there do not appear to be publicly-available reference implementations of such models, and the lack of a common benchmark makes comparisons difficult. The latter problem has been addressed, and it would be interesting to see how recurrent neural networks stack up against our approach.

Acknowledgment

All the experiments were conducted using Keras and blabla libraries.

The authors would like to acknowledge the support of...

References

[1] Douglas Coimbra de Andrade, Sabato Leo, Martin Loesener Da Silva Viana, Christoph Bernkopf. *A neural attention model for speech command recognition*. arXiv:1808.08929

[2] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio. *Attention-Based Models for Speech Recognition*. arXiv:1506.07503

[3] Sainath, Tara N. / Parada, Carolina (2015). *Convolutional neural networks for small-footprint keyword spotting*, In INTERSPEECH-2015, 1478-1482.

[4] Raphael Tang, Jimmy Lin. *Deep Residual Learning for Small-Footprint Keyword Spotting*. arXiv:1710.10361

[5] Alon, G.. *Key-Word Spotting-The Base Technology for Speech Analytics*.

[6] Pete Warden. *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*.

[1] https://github.com/ivaste/KeyWordSpotting

[2] https://research.googleblog.com/2017/08/launching-speech-commands-dataset.html