HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

Using Deep Learning to Understand and Model how a Virtual Assistant, like Siri, knows when to Act

Shravan Devraj[1] and Ross Greer[2#]

[1]Oak Park High School, Oak Park, CA, USA
[2]University of California, San Diego, CA, USA
[#]Advisor

ABSTRACT

In the era of technology, virtual assistants are all around us and have changed the way we interact with technology. To better understand the inner workings of virtual assistants, we visualized and demonstrated one way that mimics the audio classification techniques of virtual assistants by developing a deep convolutional neural network (DCNN) trained on mel spectrograms to classify audio. Our hypothesis is that mel spectrograms of the wake and non-wake words can be used to accurately classify audio. Out of the 85 files in our dataset, our classifier was trained and validated on 58 files of data and tested on 27 files of data. When evaluating our test performance, our model achieved a value of 1 for precision, recall and accuracy. Our classifier achieved a 100% accuracy in classifying wake words and non-wake words.

Introduction

Virtual assistants, like Siri, can be used to simplify tasks and make your life more efficient. As the presence of virtual assistants increases in our lives, it is important to understand how they work. Understanding how virtual assistants work can enhance our experiences, make us more aware of how to handle information regarding privacy and security, and, most importantly, by better understanding their inner workings, developers can create more advanced and intelligent virtual assistants, further contributing to the ongoing innovation in this field. Virtual assistants are passive listeners, meaning that they listen to everything around them. But, they do not act until they hear a wake word. How does a virtual assistant know when the wake word is spoken? To demonstrate this, let's take a look at Siri, a well known virtual assistant. Siri uses a deep neural network (DNN) to interpret human voices and commands [2]. To see this process in action, we developed and trained a deep convolutional neural network (DCNN), a branch of DNN that is a highly effective neural network when it comes to image/audio classifications [3, 4, 5], on a dataset containing recordings of wake and non-wake words.

Method

The DCNN model we created consisted of three convolutional layers with 16, 32, and 16 filters respectively, each followed by a max pooling layer. The output is then flattened and passed through two fully connected layers, the first with 256 neurons and the final layer with a single neuron, using ReLU and sigmoid activation functions respectively (model layers shown in Figure 1). All the steps of this DCNN model were coded in Python 3.10.9 and the model was constructed using TensorFlow-MacOS 2.9.0 (model structure shown in Figure 2).

Various recordings of a person saying the wake and non-wake words were used as data for training the DCNN. The 'Siri' dataset consists of the wake words, while the 'not_siri' dataset consists of non-wake words. Our dataset contains a total of 85 audio files, 42 for Siri and 43 for not_siri. Our wake word recordings in the Siri dataset

include the voice of AI and a person saying “Siri” in various different pitches while the non-wake word recordings in the not_siri dataset include the voices of AI and a person saying non-wake words as well as recordings of background noises. All the data in both datasets were converted to a wav file, padded to the same length and converted to mel spectrograms. The mel spectrograms, shown in Figure 3, are then labeled (0 for not_siri and 1 for siri) and scaled before being used in the training, validating and testing phases of the DCNN model.

```
# Adding the model layers
model.add(Conv2D(16, (3,3), 1, activation="relu", input_shape=(256,256,3)))
model.add(MaxPooling2D())

model.add(Conv2D(32, (3,3), 1, activation="relu"))
model.add(MaxPooling2D())

model.add(Conv2D(16, (3,3), 1, activation="relu"))
model.add(MaxPooling2D())

model.add(Flatten())

model.add(Dense(256, activation="relu"))
model.add(Dense(1, activation="sigmoid"))
```

Figure 1: This figure shows the DCNN model layers that were added while creating this model.

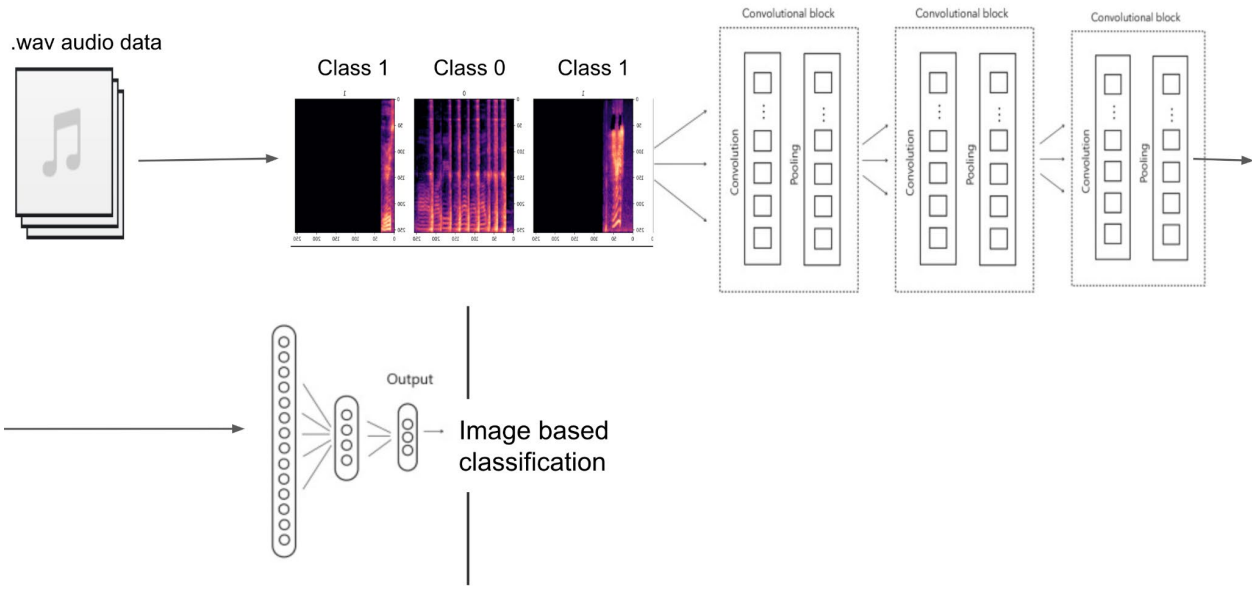

Figure 2: This figure is a visual representation of the DCNN model structure. Images adapted from [11,12]

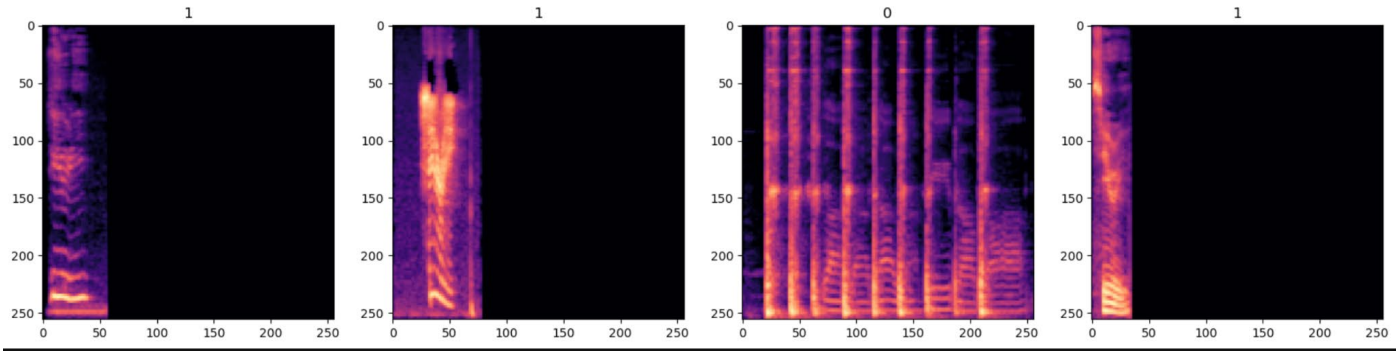

Figure 3: This figure shows a labeled batch of 'Siri' and 'not_siri' data as mel spectrograms.

Results

Our DCNN model has an accuracy of 100% after being trained and validated on 58 files of data and tested on 27 files of data. When we evaluated our test performance, our model achieved a value of 1 for precision, recall and accuracy. During the testing phase, our model classified wake and non-wake words correctly. This model proves the concept of our hypothesis, based on our limited dataset, that mel spectrograms can be used to accurately classify audio while mimicking the functionality of virtual assistants, like Siri.

Discussion

Our model having 100% accuracy shows how mel spectrograms are an effective tool that can be used to create accurate audio classification models. Even though our DCNN model achieved 100% accuracy on a small dataset, it is important to note that achieving 100% accuracy on a small dataset does not guarantee the same performance on a larger or more diverse dataset. The model might have overfitted the limited training data, meaning it learned to recognize specific samples extremely well but may struggle with unseen or different data. All the audio recordings in the dataset consisted of recordings of AI generated voices as well as voices of a person in different pitches and speeds. Even though this method added variety, it did not add diversity to the dataset, which potentially makes this model biased and may not correctly classify every single voice it hears in the real world. We have curated little clips of audio for our model to classify but pulling such clips from a continuous stream of real-world noise (multiple people talking at once, including background noises) is a complex and difficult task and this can impact the accuracy of our model. Despite these limitations, this model worked well for our research purposes and successfully proved the concept of our hypothesis.

Acknowledgement

I would like to express my sincere gratitude to my advisor, Ross Greer, PhD Candidate in Electrical and Computer Engineering, for his guidance throughout this research process.

References

1. Renotte, Nicholas, director. Build a Deep CNN Image Classifier with ANY Images. YouTube, YouTube, 25 Apr. 2022, https://youtube.com/watch?v=jztwpsIzEGc&t=0s.
2. "Hey Siri: An on-Device DNN-Powered Voice Trigger for Apple's Personal Assistant." Apple Machine Learning Research, https://machinelearning.apple.com/research/hey-siri.

3. *Unsupervised Feature Learning for Audio Classification Using ... - Neurips*, https://proceedings.neurips.cc/paper_files/paper/2009/file/a113c1ecd3cace2237256f4c712f61b5-Paper.pdf.
4. Nanni, Loris, et al. "An Ensemble of Convolutional Neural Networks for Audio Classification." *MDPI*, 22 June 2021, https://www.mdpi.com/2076-3417/11/13/5796.
5. Nanni, Loris, Yandre M. G. Costa, et al. "Ensemble of Convolutional Neural Networks to Improve Animal Audio Classification - EURASIP Journal on Audio, Speech, and Music Processing." *SpringerOpen*, 26 May 2020, https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-020-00175-3.
6. McLaughlin, Molly. "What Is a Virtual Assistant and How Does It Work?" Lifewire, 5 Aug. 2021, www.lifewire.com/virtual-assistants-4138533.
7. Doshi, Ketan. "Audio Deep Learning Made Simple: Automatic Speech Recognition (ASR), How It Works." Medium, 25 May 2021, https://towardsdatascience.com/audio-deep-learning-made-simple-automatic-speech-recognition-asr-how-it-works-716cfce4c706.
8. Doshi, Ketan. "Foundations of NLP Explained Visually: Beam Search, How It Works." Medium, 21 May 2021, https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24.
9. Doshi, Ketan. "Audio Deep Learning Made Simple: Sound Classification, Step-by-Step." Medium, 21 May 2021, https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5.
10. Doshi, Ketan. "Audio Deep Learning Made Simple (Part 1): State-of-the-Art Techniques." Medium, 21 May 2021, https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504.
11. https://www.mdpi.com/sensors/sensors-22-01521/article_deploy/html/images/sensors-22-01521-g001.png
12. https://ars.els-cdn.com/content/image/3-s2.0-B9780128188330000096-f09-03-9780128188330.jpg