Multi-Lingual Audio Classification

Cynthia Hua and Hiroshi Mendoza
[cynthiax; hmendoza]@stanford.edu

CS230- Fall 2018 - Stanford University

Objective

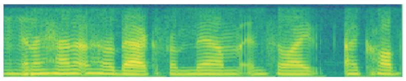
Language identification (LID) has the potential to greatly enhance multi-lingual ASR programs. Current ASR systems such as Siri or Google Assistant rely on a user to manually input their spoken language. However, this is less applicable to multi-lingual households, which is the case for 1 out of 5 residents in the U.S., or similar settings. **We attempt to develop an improved LID network using the visual properties of sound via convolutional networks.**

Setup

Scipy: Amplitude-frequency conversion from WAV file
Tensorflow GPU: CNN baseline, AlexNet, Wavenet

We approach the problem using CNN's because we expect LID on short audio sequences to benefit from analyzing the structure of individual sounds rather than simply time-based patterns. Subsequently, we develop a Wavenet-based architecture that takes both temporal and visual elements into account. The input to the trained neural network will be an audio sample, and the output will be the predicted language. We were able to achieve 80% with our AlexNet model.

Dataset



CALLHOME: Dataset of recorded phone conversations in six languages: English, German, Mandarin, Japanese, Spanish and Arabic. [1]The dataset contains 60 hours of data (in 120 files of 30-minute calls) per language. For the CNN's, we split the audio files into 3 second mono audio files which we convert into spectrogram arrays which are padded with zeros to make a [300, 300] training sample.

For the wavenet, we converted the audio directly to integer arrays using mu-law companding, which compresses the audio ranges to improve the signal to noise ratio:

$$f(x_t) = sign(x_t)\frac{ln(1+\mu|x_t|}{ln(1+\mu)}$$

The final wavenet dataset stores data as size [48000] integer arrays.

References

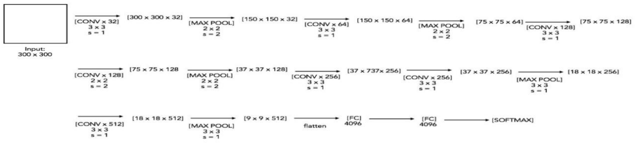
[1] https://catalog.ldc.upenn.edu/LDC97S42

[2] https://arxiv.org/abs/1609.03499

[3] https://deepmind.com/blog/wavenet-generative-model-raw-audio/

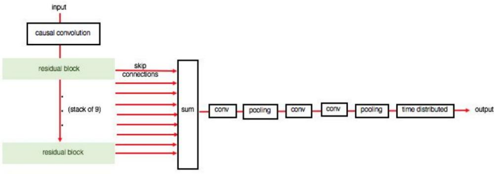
CNNs

We implemented 3 CNN architectures: a Basic 2 layer CNN, a 7 Layer CNN, and a Simplified AlexNet. The **7 layer CNN** is shown below.

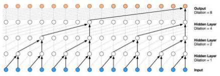


Wavenet

We implement a network loosely based on the Wavenet model [2], experimenting with parameters that were not specified in the wavenet paper and modifying the model to be classifier rather than a generator. Overall, the architecture consists of a causal convolutional layer followed by a series of residual blocks followed by a series of convolution and dimensionality reduction layers, ending in a softmax classification layer.



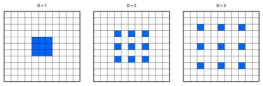
Causal Convolutions



Output Dilation = 8
Hidden Layer Dilation = 4
Hidden Layer Dilation = 2
Hidden Layer Dilation = 1
Input

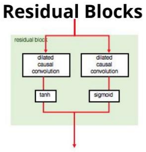
We employ casual convolutional layers, which ensure our convolutions consider the temporal nature of audio data. The masking ensures that each node only uses learnings on pixel inputs that are temporally prior, blocking connections to nodes in the previous layer that considered inputs which lie ahead in time.

Dilation



Dilation applies a convolution with gaps defined by their dilation rate, for example a dilation rate of k=2 indicates the layer looks at every other pixel in its input. Skipping inputs in this way allows nodes to consider overall a far larger area of the input, increasing its receptive field. In our model, we stack dilated layers with doubling dilation rates.

Residual Blocks

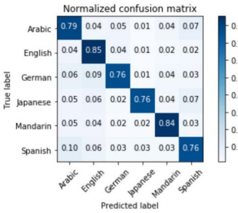


Inside each residual block are two dilated convolution layers. The convolution with the tanh activation acts a filter, which is responsible for feature learning. The convolution with the sigmoid activation acts as a gate, which decides how important various outputs from the tanh activation are.

Results

Model	Test Set Accuracy
2 Layer CNN	53%
7 Layer CNN	75%
Simplified Alex Net	80%
WaveNet	NA

Confusion Matrix (AlexNet)



	Arabic	English	German	Japanese	Mandarin	Spanish
Arabic	0.79	0.04	0.05	0.01	0.04	0.07
English	0.04	0.85	0.04	0.01	0.02	0.02
German	0.06	0.09	0.76	0.01	0.04	0.03
Japanese	0.05	0.06	0.02	0.76	0.04	0.07
Mandarin	0.05	0.04	0.02	0.02	0.84	0.03
Spanish	0.10	0.06	0.03	0.03	0.03	0.76

Conclusion

- We were able to achieve an accuracy of **80% with our simplified AlexNet CNN model.**
- We showed that image-based audio analysis can be an appropriate method for Language identification.
- We noticed that 1-2% of the dataset contained silent and incoherent sounds that if discarded would boost our accuracy.
- **Mandarin and English were the most identifiable**. The classifier was confused most with Spanish, Arabic, and Japanese which makes sense since there might be shared phonemes between the languages.
- A key challenge with image-based audio analysis is the high memory volume. For example, our wavenet input was 16GB. We were limited by time and GPU access, but believe these results would improve if multiple GPU's were used to train larger sample sets represented as higher memory data types (we used ints and float16's). .

Future Work

- For future work, we would try to incorporate other features such as pitch and tempo to better classify the languages.
- Train on larger and smaller sounds samples (~1sec).