Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Marielen Marins Ferreira

23 de abril de 2018

Proposta

Histórico do assunto

Existe um *podcast* chamado Mamilos que está disponível nas plataformas *iTunes*, *Spotify*, entre outros. Os episódios sempre trazem assuntos polêmicos com o diferencial de colocar à mesa, pessoas especialistas no tema com diferentes pontos de vista, elevando o nível da discussão.

No episódio #58 sobre acessibilidade foi lido um comentário de uma pessoa que colocou o seguinte comentário num post do *podcast*:

> "Sou surdo, não sei como posso ouvir esse programa."

Essa frase abalou a equipe e começaram a pensar em como tornar o *podcast* mais acessível para surdos. Assim, a partir desse programa decidiram criar uma rede de colaboradores para transcrever o áudio e não restringir mais a audiência que eles poderiam atingir.

Descrição do problema

No episódio #128, que celebrava os 3 anos do *podcast* Mamilos, foi revelado diversas curiosidades sobre como era produzido cada episódio. Disseram que a cada 5 minutos de áudio demorava-se 30 minutos para transcrever o conteúdo. Considerando um episódio do Mamilos de 90 minutos, levaria 540 minutos para ser transcrito e mais 150 minutos para ser revisado antes da publicação, segundo o episódio.

A equipe de transcrição, carinhosamente chamada de Mamilândia, já teve mais de 70 pessoas e entregavam uma transcrição por semana nessa época. Entretanto, nos dias atuais, conta com um grupo de 22 voluntários que tentam encaixar essa tarefa em suas vidas atribuladas.

Sendo assim, o problema que será atacado por este projeto é a redução do tempo de transcrição dos episódios do Mamilos.

Conjuntos de dados e entradas

O conjunto de dados não foi retirado dos episódios do *podcast* Mamilos, pois para uma pesquisa inicial foi mais viável começar com uma base de dados já estruturada. Assim, a base foi retirada da competição *TensorFlow Speech Recognition Challenge* na plataforma *Kaggle*. O conjunto de dados

contém diversas pastas cujo nome era a palavra dita nos arquivos de áudio dentro dela.

Escolhi trabalhar com todas, exceto *background noise* que não tem relações com as palavras ditas no áudio, apenas simular sons que não são produzidos em estúdio.

Descrição da solução

A solução será feita a partir do pré-processamento do áudio para digitalizar as ondas sonoras e do desenvolvimento de *machine learning* para associar o áudio às palavras ditas.

Modelo de referência (benchmark)

Por ser um desafio do *Kaggle*, diversas pessoas desenvolveram algoritmos para resolver a competição com métodos distintos. Durante a minha busca sobre soluções referente ao tema, percebi que a solução mais executada foi utilizando *Mel-Frequency Cepstrum* para o pré-processamento do áudio e *Convolutional Neural Networks* para o algoritmo de *machine learning*. Chegando em alguns casos, a 86% de acurácia para configurações semelhantes à descrita.

Métricas de avaliação

Como a maioria das referências deste projeto utilizou a acurácia como métrica de avaliação, também a usarei para meios de comparação. Ela é calculada através da taxa de acerto do modelo de *machine learning* em uma base de dados que não foi utilizada para o treinamento do mesmo.

Design do projeto

O projeto foi estruturado da seguinte maneira:

1. Análise de dados - estudo para descobrir característica da base de dados utilizada;
2. Pré-processamento dos arquivos de áudio - conversão dos arquivos de áudio em vetores numéricos;
3. Construção do modelo - desenvolvimento da arquitetura de redes neurais;
4. Treinamento do modelo - treinamento do modelo construído com base numa amostra da base de dados.