Universidade Estadual de Londrina

MARCELO FERREIRA DE ALMEIDA
RAFAEL PALHETA TOKAIRIN

IA GERADORA DE ÁUDIO:

LONDRINA - PR
2024

1 Objetivos

O objetivo deste projeto é desenvolver uma Inteligência Artificial (IA) capaz de gerar áudios de voz humana de alta qualidade com base em textos ou comandos específicos (como outro áudio). Esta IA será projetada para síntese da fala. O projeto explorará técnicas avançadas de processamento de linguagem natural (NLP) e aprendizado profundo para criar um sistema robusto e eficiente.

O projeto será feito utilizando a o modelo Tacotron-2 para a síntese de áudio, os áudios utilizados no projeto serão retirados principalmente de vídeos do youtube e a transcrição será feita pelo uso da Whisper Transcriber (IA que gera texto a partir de áudios). O critério de análise para os dados gerados será a qualidade do som e a semelhança da voz gerada com a voz humana natural.

2 Conjunto de dados

O conjunto de dados consiste em:

1. Áudios de vozes: os áudios foram retirados em sua maioria de vídeos do youtube e de sites como o Common Voice, os áudios utilizados foram escolhidos baseando se em sua uniformidade, buscamos áudios mais limpos e com menos ruídos.
2. Transcrição dos áudios usados: A maior parte das transcrições foi feita a partir da Whisper Transcriber, que é uma IA que converte áudios para textos, a ferramenta comete alguns erros então foi necessário tomar alguns cuidados para que o texto transcrito fosse fiel ao áudio.
3. Espectrogramas Mel: Esses espectrogramas eram o que a IA geradora de voz de fato utilizava para gerar o áudio em si.

3 Metodologia

3.1 Preparação dos dados

A preparação de dados consiste em 3 passos:

1. Coleta e análise de qualidade dos dados: Nesta etapa foi buscado várias fontes disponíveis de áudios que continham clareza no discurso e poucos ruídos de fundo, encontramos um banco de dados chamado Common Voice do Mozilla, que foi utilizado em testes iniciais, pois possuía uma grande variedade de dados, todos transcritos, porém aquele banco possuía muitas vozes diferentes o que causava inconsistência em testes grandes. Assim, a solução foi utilizar áudios extraídos de vídeos do youtube, pois continham mais qualidade e era possível pegar até horas de uma mesma pessoa falando.
2. Pré-processamento: O pré-processamento dos dados foi basicamente transcrever e validar as transcrições dos áudios feitos, pois como a IA usa textos como input, seria necessário usá-los no treinamento. Então foi feito o

uso da ferramenta Whisper Transcriber para transcrever os áudios adquirido em texto, após a transcrição foi necessário um trabalho manual para checar a qualidade da transcrição e corrigir eventuais erros de gramática cometidos pela IA.

3. Separar os dados**:** Os dados foram separados de forma que 80% foram destinados para treinamento e 20% para testes.

3.2 Arquitetura do modelo

O modelo escolhido para desenvolver a IA geradora de áudios foi o Tacotron-2, que é o modelo estado-da-arte para síntese da fala. Trata-se da segunda versão da Tacotron, um modelo desenvolvido pelo Google e que possui uma arquitetura seq2seq, o qual recebe como entrada o texto e produz como saída um espectrograma em escala mel.

O Tacotron-2 funciona, resumidamente, em três etapas:

1. Encoder: Tem a tarefa de transformar a sequência de caracteres de entrada (texto) em uma sequência de representações internas (embedding) que captura as características fonéticas e semânticas do texto. Essa representação é então usada pelo mecanismo de atenção para gerar espectrogramas mel

2. Attention Mechanism:O objetivo do mecanismo de atenção no Tacotron 2 é permitir que o modelo aprenda automaticamente quais partes do texto de entrada (representadas pelas saídas do Encoder) devem ser focadas ao gerar cada quadro do espectrograma mel. Isso é especialmente importante para a síntese de fala, onde a duração das partes do texto não é fixa e pode variar muito.

3. Decoder: O Decoder no Tacotron-2 é responsável por converter a representação intermediária do texto, fornecida pelo Encoder e ajustada pelo mecanismo de atenção, em um espectrograma mel, que pode ser usado para gerar a síntese de fala.

3.3 Treinamento do modelo

O treinamento do Tacotron 2 envolve ajustar os pesos do modelo para minimizar a diferença entre as previsões do modelo e os dados de treinamento (espectrogramas mel reais). O treinamento é feito de maneira supervisionada, onde o modelo é alimentado com pares de entrada (texto) e saída (espectrograma mel) conhecidos.

Nos treinamentos feitos, foram utilizados áudios de 30 min, que foram particionados em áudios menores de 5 a 10 segundos. Após o treinamento efetuado, foram feitos testes de geração de áudio para validar a eficácia do treinamento.

Nos testes iniciais em que usamos cerca de 5-10min de áudios (esses testes eram feitos apenas para checar a funcionalidade do algoritmo), os áudios gerados eram de

qualidade baixa, com voz bem robotizada e muitas falhas, porém isso se dava devido a baixa quantidade de áudios usadas para treinar.
No teste final, em que foi usado 30min de áudio, os áudios gerados foram satisfatórios, apenas com alguns problemas na pronúncia ou entonação de algumas palavras, mas podemos testar novamente posteriormente, com uma entrada de dados maior (cerca de 1 hora ou mais) e descobrir se esses erros eram também causados pela quantidade de arquivos usados no treinamento.

4 Resultados

A construção do projeto em si foi desafiadora, foram dias de pesquisa procurando o modelo que melhor atendia a nossas necessidades, além disso a geração de voz possui muitas etapas, como a transcrição e avaliação dos áudios transcritos, geração do espectrograma Mel e outras, que caso não estejam funcionando em perfeito estado, podem acarretar grandes problemas no resultado do treinamento, assim descobrir o que ocasionava o erro final era muito complexo e demandava otimizar todas as etapas.

Mesmo com isso aplicação do modelo Tacotron-2 para a síntese de voz humana mostrou resultados satisfatórios. O modelo foi treinado usando cerca 30 minutos de áudios o que foi uma quantidade suficiente para um resultado razoável contendo apenas alguns erros de fala, entonação e tonalidade das sílabas.

REFERÊNCIAS

PYTORCH. Tacotron 2: PyTorch Hub. Disponível em: https://pytorch.org/hub/nvidia_deeplearningexamples_tacotron2/ . Acesso em: 7 ago. 2024.

PAPERS WITH CODE. Tacotron 2. Disponível em: https://paperswithcode.com/method/tacotron-2 . Acesso em: 7 ago. 2024.

GOOGLE COLAB. Tacotron2_e_HiFi_GAN_PTBR.ipynb. Disponível em: https://colab.research.google.com/drive/1-XWvLVhD11ZFosHsEqLnhiF58y-LlQWh#scrollTo=GHIBEHtW-eHZ . Acesso em: 7 ago. 2024.