Fine-tuning SpeechT5 Text-To-Speech Model for Italian Language

Model Overview

This report provides a detailed description of the fine-tuning process of the microsoft/speecht5_tts model on an Italian language dataset. The training was conducted using the freds0/cml_tts_dataset_italian dataset available on Hugging Face. The goal was to adapt the SpeechT5 Text-to-Speech (TTS) model for Italian text-to-speech generation.

The report covers the dataset, training configuration, evaluation results, and detailed training logs.

Dataset Description

Dataset: freds0/cml_tts_dataset_italian
Link: CML TTS Dataset Italian

This dataset consists of paired Italian text and corresponding speech waveforms. Each data sample in the dataset includes:

- **Text**: Italian sentences and phrases meant to be converted into speech.
- **Audio**: Speech recordings in WAV format, corresponding to each text input, recorded by native Italian speakers.

The dataset is designed for Text-to-Speech tasks, where the objective is to generate high-quality Italian speech from text. It provides a solid foundation for training and fine-tuning models like SpeechT5, ensuring that they can accurately reproduce the prosody and pronunciation typical of native Italian.

Key details of the Dataset:

- Total number of samples: 10,000+
- Sampling rate: 16 kHz
- Audio format: WAV
- Text format: UTF-8 encoded Italian sentences

Model Description

We used the microsoft/speecht5_tts model as a base model for fine-tuning. This is a pre-trained model developed by Microsoft for Text-to-Speech tasks. It leverages transformer-based architectures to perform TTS with high-quality speech synthesis across multiple languages.

The fine-tuning task was aimed at adapting this model to generate high-quality Italian speech.

More information regarding the architecture and technical details of the SpeechT5 model can be found here.

Intended Uses & Limitations

The fine-tuned SpeechT5 model can be used for generating Italian speech from text, making it suitable for applications like:

- Text-to-speech systems for Italian.
- Conversational AI systems in Italian.
- Italian-language audiobook generation.

Limitations:

- The model may not perform as well on out-of-distribution data that differs significantly from the fine-tuning dataset.
- Handling of colloquial, regional, or dialectal variations of Italian may require additional fine-tuning.
- Some inaccuracies in pronunciation may arise with complex or rarely used Italian words.

Training and Evaluation Data

Training Set:

The training set consists of Italian text-to-speech pairs, with a variety of sentence structures, phrases, and dialogues that cover diverse linguistic features of the Italian language. Each text entry has a corresponding audio file spoken by native Italian speakers.

Evaluation Set:

The evaluation set contains a separate, unseen portion of the dataset for assessing the model's performance. This set was used to compute validation loss and ensure that the model generalizes well beyond the training data.

Training Procedure

The training process fine-tuned the SpeechT5 model using the Italian dataset. The model was trained with the following configurations:

Training Hyperparameters:

- **Learning rate**: 0.0001
- **Train batch size**: 4
- **Evaluation batch size**: 2
- **Seed**: 42
- **Gradient accumulation steps**: 8
- **Total train batch size (per update)**: 32
- **Optimizer**: Adam with betas (0.9, 0.999) and epsilon 1e-08
- **Learning rate scheduler**: Linear
- **Warmup steps**: 100
- **Total training steps**: 500
- **Mixed precision training**: Native Automatic Mixed Precision (AMP)

Optimizer:

The Adam optimizer was used with default parameters (betas=(0.9, 0.999) and epsilon=1e-08). This optimizer was chosen for its effectiveness in handling sparse gradients, which are common in transformer models.

Learning Rate Scheduler:

A linear learning rate scheduler was applied, with a warmup of 100 steps. This helps the model gradually adjust to the learning process and prevents instability in the early stages of training.

Mixed Precision Training:

To improve training speed and reduce memory usage, we used Native Automatic Mixed Precision (AMP). This allows the model to benefit from the faster training speeds provided by GPU acceleration while maintaining performance.

Training Results

The following logs summarize the model's performance across different epochs:

Training Loss	Epoch	Step	Validation Loss
0.5185	26.6667	100	0.4566
0.4471	53.3333	200	0.4438
0.4255	80.0	300	0.4479
0.4026	106.6667	400	0.4517

Training Loss	Epoch	Step	Validation Loss
0.3995	133.3333	500	0.4555

Observations:

- The training loss decreased consistently throughout the training process, which indicates that the model was learning effectively.
- The validation loss showed some fluctuations, suggesting that the model's generalization was improving but might need additional fine-tuning to stabilize.
- The final loss on the evaluation set was **0.4555**, which is reasonable but can be further reduced with additional epochs or a larger dataset.

Performance Evaluation

Metrics:

The primary metric used for evaluation was loss on the validation set. Loss functions such as Mean Squared Error (MSE) were applied to measure the difference between the generated and target audio features. Lower loss values indicate better performance.

Final Performance:

- **Training Loss**: 0.3995
- **Validation Loss**: 0.4555

While the model performed well on the training data, its performance on the validation set suggests there is room for further optimization. Fine-tuning over more steps, increasing the dataset size, or adjusting hyperparameters may help further reduce the validation loss.

Audio Quality:

Subjective evaluations of the generated Italian speech samples suggest that the fine-tuned model produces intelligible and natural-sounding Italian speech, although some minor errors in prosody and pronunciation were observed in complex sentences.

Framework Versions

The following versions of libraries were used during training:

- **Transformers**: 4.44.2
- **PyTorch**: 2.4.1+cu121
- **Datasets**: 3.0.2
- **Tokenizers**: 0.19.1

Conclusion

This fine-tuning task successfully adapted the SpeechT5 model to generate Italian speech. The model's performance was evaluated based on loss metrics, and it achieved satisfactory results. For further improvement, adjustments to the dataset, hyperparameters, and training steps can be explored.