BRAIN2SPEECH - A Deep Learning Homework

Katica Bozsó

January 14, 2024

Abstract

This document serves as the essential documentation for the BMEVITMMA19 2023/24/1 deep learning homework, centered on the BRAIN2SPEECH project. The project delves into the field of speech synthesis from intracranial Electroencephalography (iEEG) data. Relying on a publicly accessible dataset, the goal of the project is to refine and improve the original codebase and methodology. The primary enhancement involves the introduction of a Deep Learning-based, efficient, and compact approach using the Pytorch Lightning framework. The implementation is designed to facilitate further exploration in this field, offering a robust foundation for future research in speech synthesis from iEEG data.

1 Introduction

Synthesising speech from neural features is an intriguing field of research, since this could pave the path to Brain–computer interfaces (BCIs) [1] that may help people with speech disabilities in speech production. In the core of such project, has to stand a well-established dataset. The Dataset of Speech Production in intracranial Electroencephalography [2] aims to provide this base. It involves 10 participants, each implanted with stereotactic EEG (sEEG) electrodes as part of their clinical therapy for epilepsy. The dataset contains recordings from a total of 1103 electrodes, offering high temporal resolution and coverage of various cortical and sub-cortical brain regions. Participants were asked to read aloud words displayed on a screen in Dutch, while their brain activity and audio were recorded.

Feature extraction from this data is a multi-step process, but a complete pipeline was provided by the authors. Initially, the Hilbert envelope of high-frequency activity (70–170Hz) is extracted from each iEEG contact using bandpass filters. These filters isolate the desired frequency range and attenuate line noise without causing phase shifts. The envelope data is then averaged over 50ms windows with a 10ms frame shift and combined with adjacent windows for temporal context. Finally, this data is normalized to zero mean and unit variance, preparing it for further analysis and decoding.

Meanwhile, audio recording pairs are first downsampled to 16 kHz. Features are extracted using the Short-Term-Fourier-Transform[1], followed by compression into a log-mel spectrogram. This ensures correspondence between audio and neural feature vectors, but the log-mel spectrogram loses the phase information this way, therefore an audio waveform can not be reconstructed directly. Authors utilize the Griffin-Lim Algorithm [3] to mitigate this problem, therefore reconstructing waveforms.

Although the initial introduction may have been somewhat technical, it is crucial to have a comprehensive understanding of the underlying dataset. This deep knowledge is essential for contributing to the existing work.

2 Method

The project employs a comprehensive methodology that begins with the creation of a Torch-based virtual environment, including all the necessary packages. This ensures a stable and reproducible computational platform for the experiments. Following the environment setup, data acquisition is the next step, which can be done manually or through the provided scipts. The SingleWordProductionDutch repository's authors' feature extraction script is used as an initial preprocessing step.

A key step in the preprocessing pipeline is the application of Principal Component Analysis (PCA) [4] for dimensionality reduction of the extracted features. PCA, a statistical technique, helps simplify

[1] https://en.wikipedia.org/wiki/Short-time_Fourier_transform (access date: 2024.01.14.)

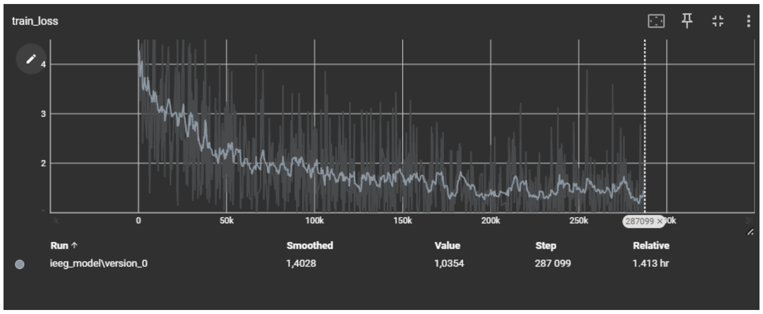
the complexity of high-dimensional data while retaining crucial patterns and relationships. Once the features are refined through PCA, a dataset is prepared - specifically formatted to yield pairs of features and corresponding spectrograms when loaded. This dataset is crucial for the effective training and evaluation of the model.

The core analytical technique employed in this project is a simple Convolutional Neural Network (CNN), which is used to predict spectrograms from the input features. The architecture is really minimal - it only consists of 1D convolutional, linear, pooling and activation layers. This neural network plays a pivotal role in establishing the relationship between the extracted features and the spectrograms.

Furthermore, a testing step was implemented using Pytorch-Lightning [2], which also includes voice reconstruction from the predicted spectrograms. One of the key aspects of the approach is its modularity. The code provided is designed to be easily adaptable, allowing for future developments such as replacing or modifying the model or dataset. This flexibility is intended to accommodate continuous improvements and adapt to varying requirements in the field.

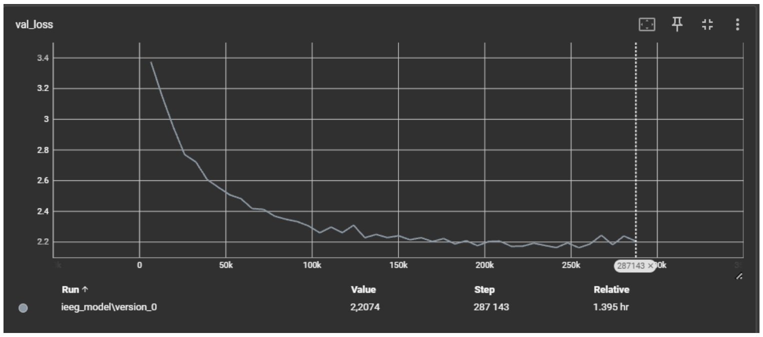
3 Results

Throughout the training process, the Mean Squared Error (MSE) loss was meticulously logged for both the training and validation datasets. These metrics served as key indicators of the model's learning progression and performance. The logged losses (see Figure 1 and Figure 2) demonstrated that the model was effectively learning to predict the correct spectrograms, which is a crucial step in the pipeline.



Run ↑	Smoothed	Value	Step	Relative
ieeg_model\version_0	1,4028	1,0354	287 099	1.413 hr

Figure 1: Training Loss



Run ↑	Value	Step	Relative
ieeg_model\version_0	2,2074	287 143	1.395 hr

Figure 2: Validation Loss

However, it was observed that the spectogram2voice component of the pipeline requires further refinement. While the model learned spectrogram prediction, the resynthesis of a voice from the test-set spectrograms did not yield successful results. This aspect of the project poses a significant challenge and is a potential area for further research and development.

Additionally, upon testing with the authors' 'reconstruction_minimal.py' script, it was confirmed that their approach to voice reconstruction from spectrograms also encountered similar difficulties.

[2] https://lightning.ai/ (access date: 2024.01.14.)

This outcome underscores the complexity of accurate voice reconstruction from spectrograms and highlights the need for continued exploration and improvement in this area.

4 Summary and Future Work

This project entailed a comprehensive examination of the Dataset of Speech Production in intracranial Electroencephalography and its associated code-base. An enhanced PyTorch-Lightning repository was established, complete with virtualization and a detailed description. Data acquisition and feature extraction steps were successfully reproduced, and further extended these with custom preprocessing and modeling techniques.

In summary, the project achieved success in the implementation of a model capable of learning to predict spectrograms, as evidenced by a significant reduction in Mean Squared Error (MSE) loss. However, chalenges were encountered in the voice reconstruction phase from these spectrograms. This aspect of the pipeline, while critical, remains a complex and unresolved issue, necessitating further research and development.

Looking ahead, there are several promising avenues for future work. One potential direction involves integrating the available text labels into the model. Developing a multi-task model that includes text-embeddings could offer a more comprehensive approach to understanding and predicting speech patterns. Finally, the current spectrogram-to-audio reconstruction script, provided by the authors of SingleWordProductionDutch, has room for improvement. Even with accurate spectrogram predictions, the reconstructed audio remains distorted, highlighting a crucial area for refinement.

References

[1] Wolpaw, J., Birbaumer, N., McFarland, D., Pfurtscheller, G. Vaughan, T. Brain–computer interfaces for communication and control. Clinical neurophysiology 113, 767–791 (2002).

[2] Verwoert, M., Ottenhoff, M.C., Goulis, S. et al. Dataset of Speech Production in intracranial Electroencephalography. Sci Data 9, 434 (2022). https://doi.org/10.1038/s41597-022-01542-9

[3] Griffin, D. Lim, J. Signal estimation from modifed short-time fourier transform. IEEE Transactions on acoustics, speech, and signal processing 32, 236–243 (1984).

[4] Wold, Svante, Kim Esbensen, and Paul Geladi. ”Principal component analysis.” Chemometrics and intelligent laboratory systems 2, no. 1-3 (1987): 37-52.