Word2Brain2Image : Visual Reconstruction from Spoken Word Representations

Ajay Subramanian, Rajaswa Patil, Veeky Baths
Birla Institute of Technology and Science Pilani, K. K. Birla Goa Campus

INTRODUCTION

- The brain is divided into several Brodmann areas of which areas 17, 18 and 19 form the Visual Cortex. BA 17 is known as the primary or striate visual cortex and is responsible for much of the brain's visual processing and pattern recognition tasks. **BA 18 and 19** form the extrastriate visual cortex and their primary function is believed to be **visual association**. Both the structure and function of this cortex are relatively unexplored.
- Recent work on understanding the role of visual association in handling non-visual stimuli has shown that the listening to spoken words can activate **low-level visual representations** that aid in the processing of visual stimuli. While this work confirms the involvement of spoken word encodings in visual processing, the nature of their representations are yet to be studied.
- Recently, Deep artificial neural networks have been extremely useful in feature extraction and operating on time series data. **Variational Autoencoders (VAEs)** and **Generative Adversarial Networks (GANs)** can help us identify compact latent representations in data and to produce samples of the source distribution from these representations.

OBJECTIVES

- Analyse the nature of low-level visual representations by attempting to **generate images of numbers from 32 channel EEG data** obtained while a subject listens to spoken versions of numbers
- Understand the **temporal onset** of this effect by using attention layers in the EEG data encoder.

METHODOLOGY

- Subjects are asked to listen to audio recordings of spoken digits from 0-9 (Speech MNIST dataset) and visualize them on a screen in front of them. During the experiment, we collect 32 channel EEG data that we later use to train our EEG feature extractor/encoder.
- We independently train a VAE and GAN on the MNIST dataset to generate MNIST-like images from a learned latent representation. We use convolutional layers for feature extraction in the encoder/discriminator and inverse convolution to produce feature maps in the decoder/generator.
- Using the EEG data collected, a LSTM-based classifier is trained which will later serve as the encoder half of the model. This classifier consists of LSTM layers followed by FC layers (to convert to a Gaussian distribution) and a Softmax layer to classify into one of 10 digits (this will be removed later).
- Finally we combine the EEG encoder and MNIST image generator and finetune weights to form the complete EEG to image generator model.

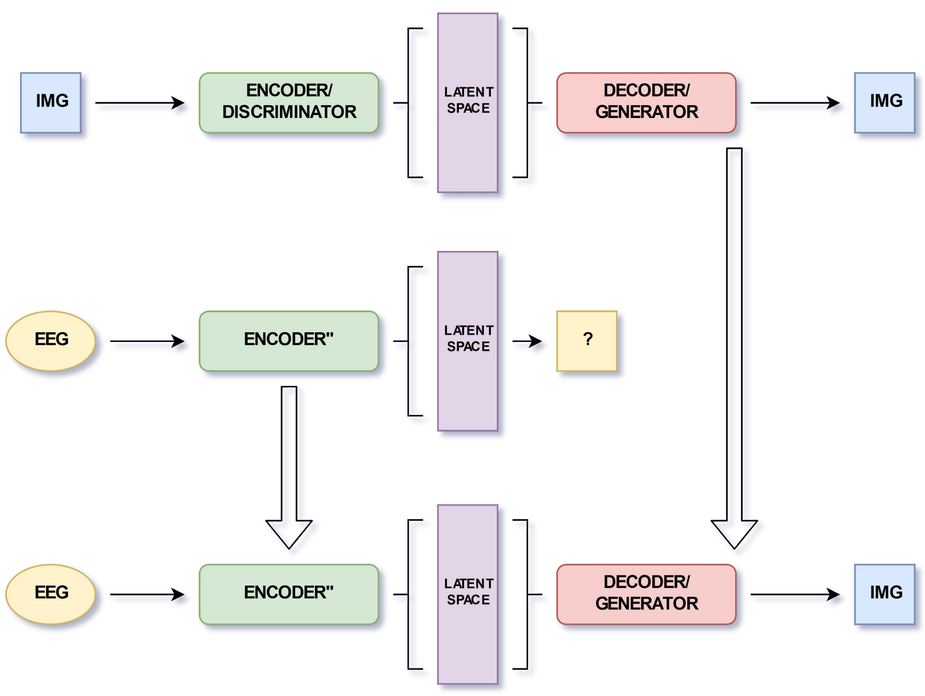

Figure 1: *Neural Networks based model architecture for the EEG to Image Generation task*

- Using the EEG data collected, a LSTM-based classifier is trained which will later serve as the encoder half of the model. This classifier consists of LSTM layers followed by FC layers (to convert to a Gaussian distribution) and a Softmax layer to classify into one of 10 digits (this will be removed later).
- Finally we combine the EEG encoder and MNIST image generator and finetune weights to form the complete EEG to image generator model.

DATA COLLECTION

- **MNIST** Dataset (**Image & Speech**)
- Experiment setup on **PsychoPy v3.0**
- Approx. Experiment Time = **45 min**
- **32 Channel EEG** Headset (1KHz)
- **5 Subjects (4-M / 1-F)**
- **50 Trials** per **Loop**, **4 Loops** per **Session**
- **2 Sessions** per **Subject**
- **400 Trials** per **Subject**

Figure 2: *Flow diagram for the Experiment*

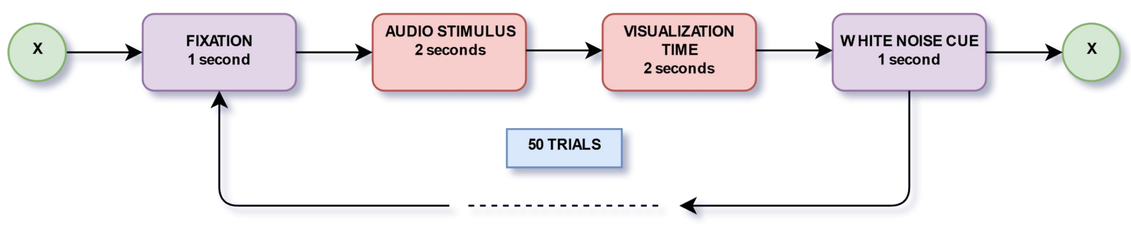

Figure 3: *Flow diagram for a single trials loop*

RESULTS & FUTURE WORK

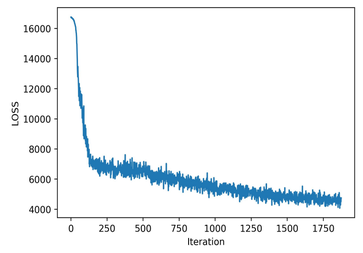

Figure 4: *Loss plot for the VAE model*

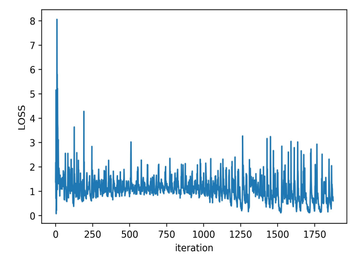

Figure 5: *Loss plot for the GAN Discriminator model*

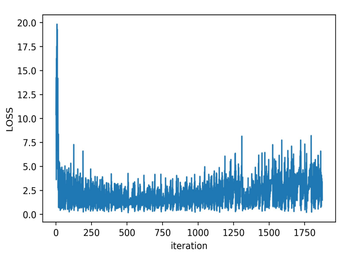

Figure 6: *Loss plot for the GAN Generator model*

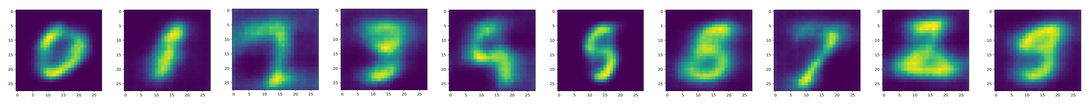

Figure 7: *Output Image Samples from the VAE model*

- **Dataset:** Thus, have collected a EEG dataset consisting of 2000 audio and visualisation events which can be used to analyse temporal visual and auditory responses to spoken word stimuli.
- **Model:** We have trained two image generator models, a VAE and GAN to generate MNIST-like images from latent vector representations. We are currently in the process of training the EEG classifier.
- **Future work:** Apart from performing the experiments we have proposed, we plan the following:
  - use attention weighted layers to analyse the temporal onset of the visual effects during spoken word processing.
  - Analyse the involvement of different areas of the brain in spoken object word processing.

ACKNOWLEDGEMENTS

This work was supported by:

- **Cognitive Neuroscience Lab, BITS Pilani Goa Campus**
- **Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands**

REFERENCES

1. **Ostarek M, Heuttig F**. *Spoken words can make the invisible visible – Testing the involvement of low-level visual representations in spoken word processing*
2. **I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano**. *Brain2Image: Converting Brain Signals into Images*
3. **LeCun Y, Cortes C**. *MNIST Handwritten Digits Dataset - 2010*
4. **Zohar J. et al**. *Free Spoken Digit Dataset (FSDD)*