Cross-modal Information Retrieval using Latent Representations of Brain Activities

Albert Cai   Omar Abul-Hassan

Department of Mechanical Engineering   Department of Mathematics, Computer Science

Introduction

We study the problem of returning optimal matches of content from visually-envoked electroencephalography (EEG) signals. Formally, the central challenge is one of classification: given 17 channel EEG time series data at 100 time points, classify which of the 1,654 images classes was shown (aardvark, abacus, accordion, ...).

- EEG data is very noisy - completely impossible for a human to map from EEG → image class
- Scarcity of EEG-image tuple data

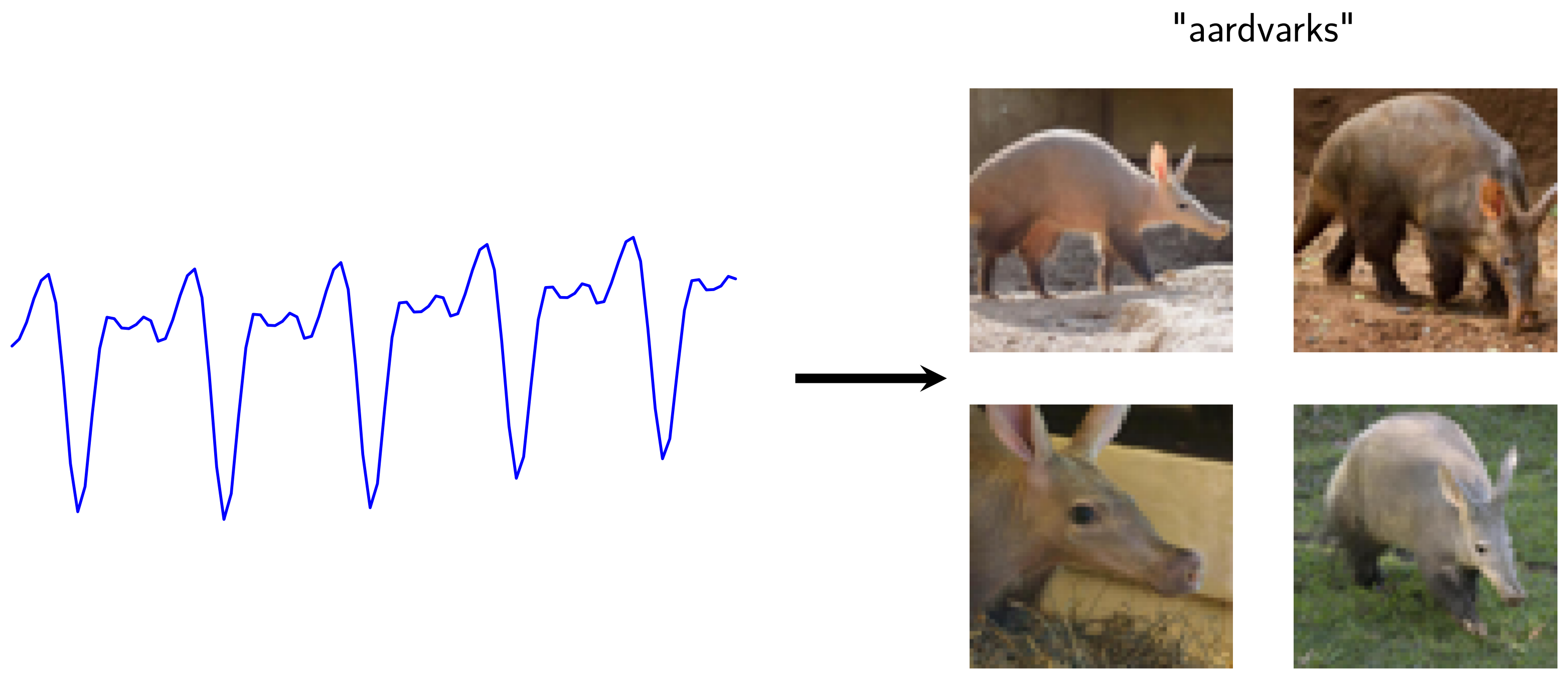
From these challenges, it is desirable to produce lower-dimensional latent representations of EEG data that are similar within classes and different between classes. Then, much simpler classification algorithms can be run on the latent representations.

Background

- **EEG-visual recognition:** Spampinato et. al [Spa+17] randomly selected 50 images of 40 classes from ImageNet to show to subjects, whose EEG recordings were measured when viewing these images. Every EEG clip is encoded with recurrent neural networks, then an image class is predicted for each clip with a CNN-based classifier.
- **EEG-visual reconstruction:** There have been numerous works that aim to reconstruct the originally viewed image given the corresponding EEG recording [Sin+23], [Kav+17]. Combinations of VAEs and GANs are employed here to produce an image for a given class from an EEG-class tuple. [Sin+23] uses a constrastive learning approach to extract features from EEG signals, then use a conditional GAN to synthesize the input images, conditioning on EEG signals. [Kav+17] uses an RNN for feature extraction and VAE for image generation.
- **EEG-visual retrieval:** Focuses on predicting the exact input shown given an EEG recording out of a given image library. For this task, it is necessary to rank the image library by which is most likely to be the one seen by the EEG recording. [Ye+22] uses a graph convolutional network as an EEG encoder, then uses contrastive learning for correlating EEG encodings with image encodings.

Dataset

The **THINGS-EEG** dataset consists of **1654** image classes, with **10** images each, and the corresponding EEG time-series recorded for each of these **16540** images for ten patients, 4 times per image, over 17 channel/electrodes, over 100 ms. The data is preprocessed, but very limited for generative modeling tasks (only 10 images per class).

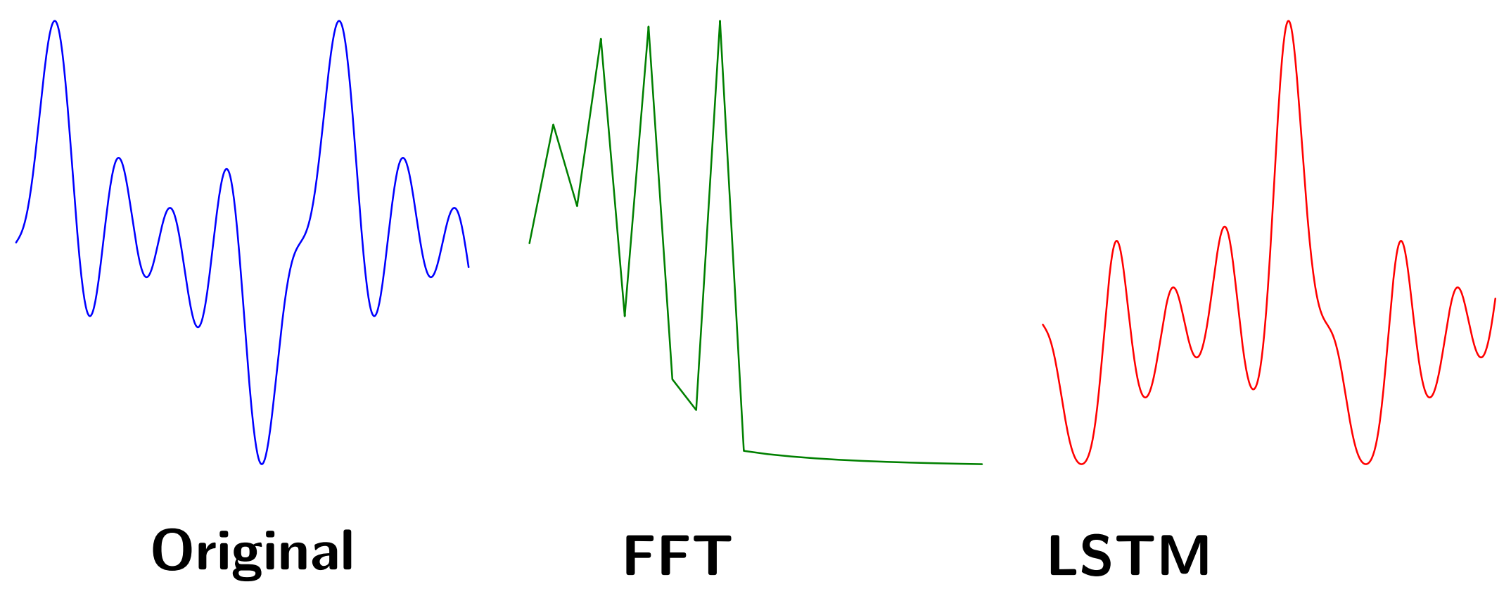


Models

Two stage process:

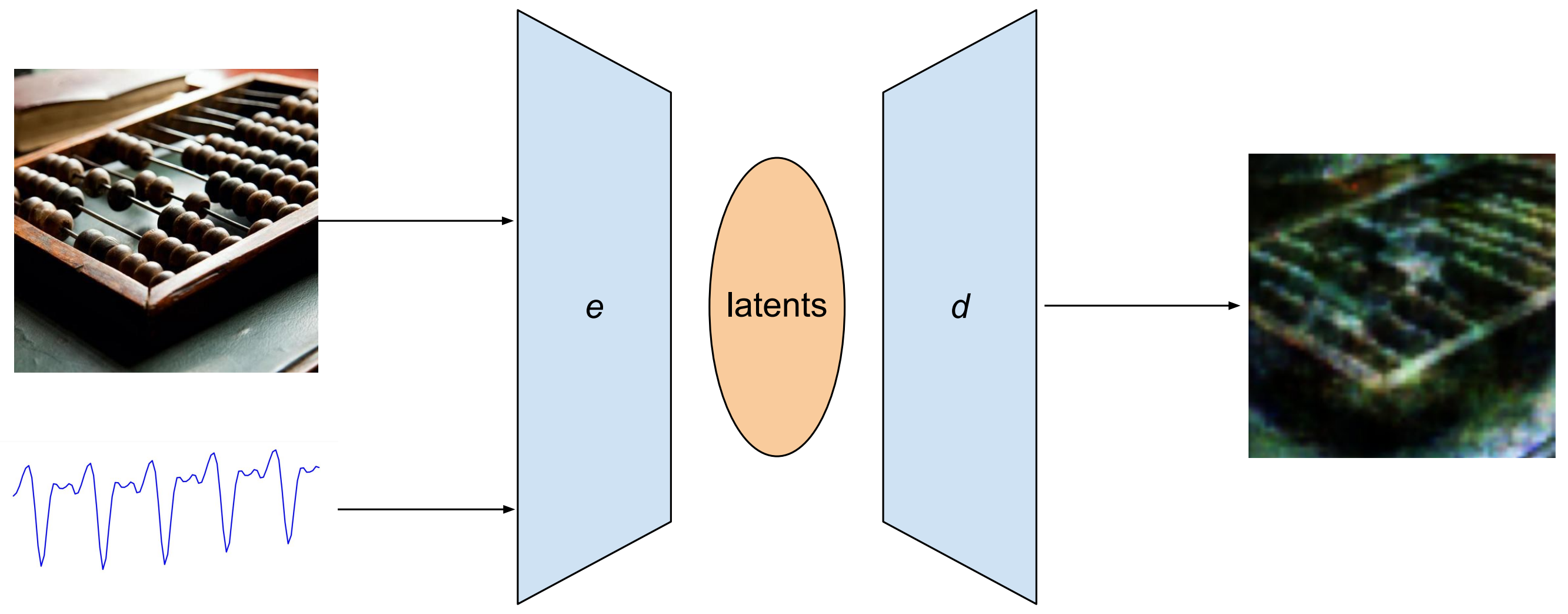
- Unsupervised latent representations of the EEG recordings using a variant of a VAE
- Train a multi-class classifier on these latents to predict image class

Vanilla models, along with VAEs with convolution and LSTM encoders, failed to uncover meaningful latent spaces, and the classifier performed no better than random.

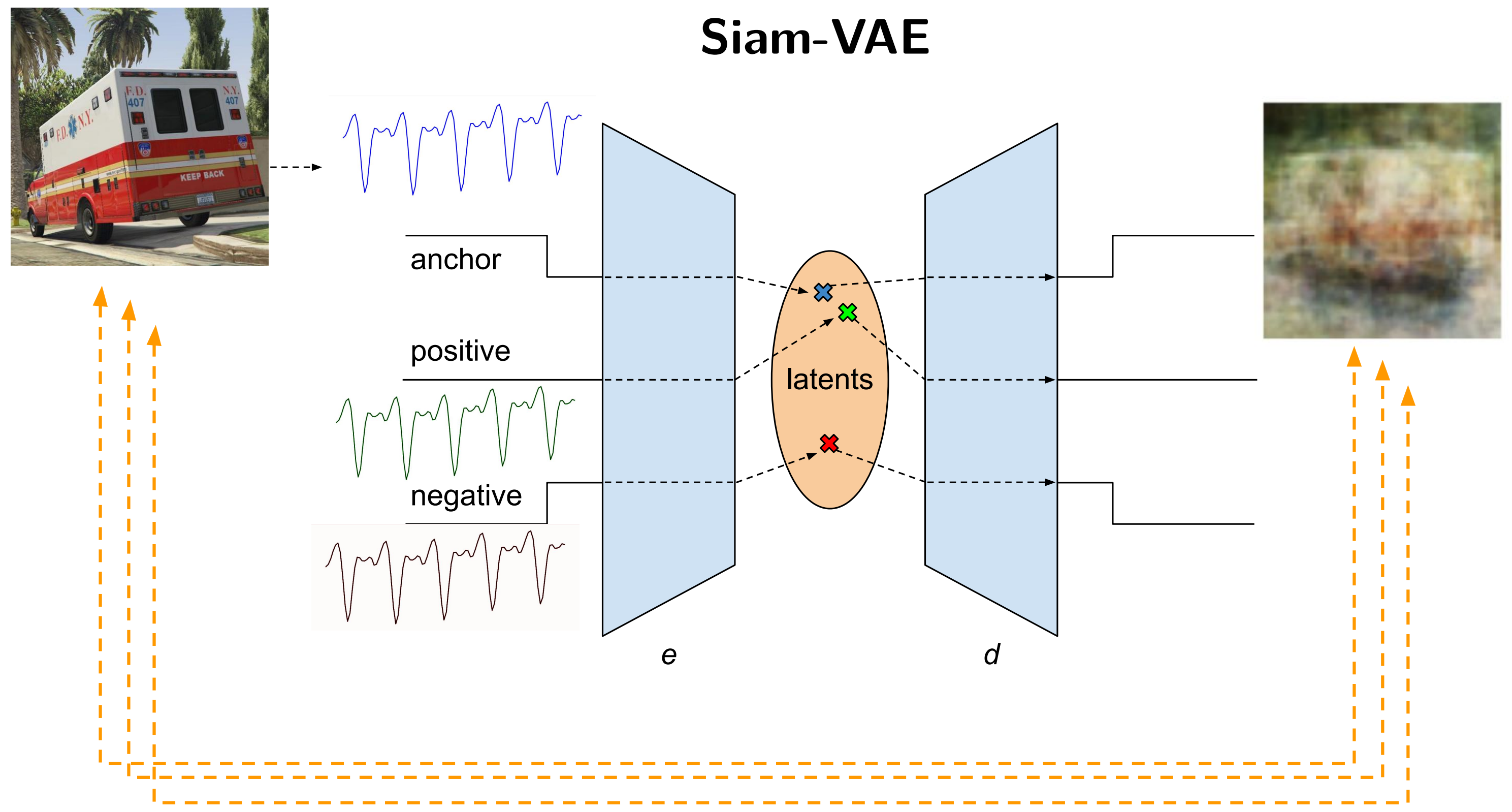


We used FFT and LSTM for feature extraction because we found that it reduces the amount of redundant data in the dataset the models are trained on.

Conditional β-VAE



Siam-VAE



$$\text{TRIPLET}(a, p, n) = \max\{d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0\}$$

$$\text{MSE}(a, p) = \frac{1}{N}\sum_{i=1}^{N}(a_i - p_i)^2$$

$$\text{Loss} = \text{TRIPLET} + \text{MSE} + \beta * \text{KL}$$

Intuition: good latents should be (1) quite different across different classes, (2) similar in the same class, (3) still follow the distribution of a prior, and (4) hold enough information for the decoder to reconstruct the image. This loss includes all four.

Results

We trained models specialized on classifying 50 image classes, 100 image classes, and all 16540 image classes. To evaluate, we used two key metrics: Percentage Accuracy, indicating the correct class identification rate from EEGs, and Multiclass AUROC, evaluating the model's ability to distinguish between various classes.

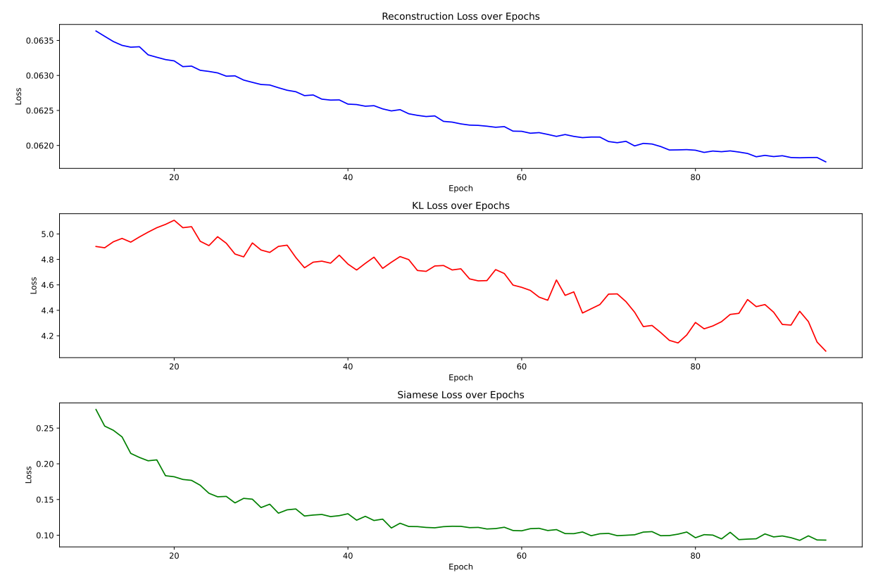
50 Class	% Accuracy	% AUROC
CVAE + RF Classifier	16%	76%
SIAMVAE + NN Classifier	**25.30%**	**96%**
VAE + NN Classifier	2%	50%
NN Classifier	2.17	1,392
Random	2%	50%

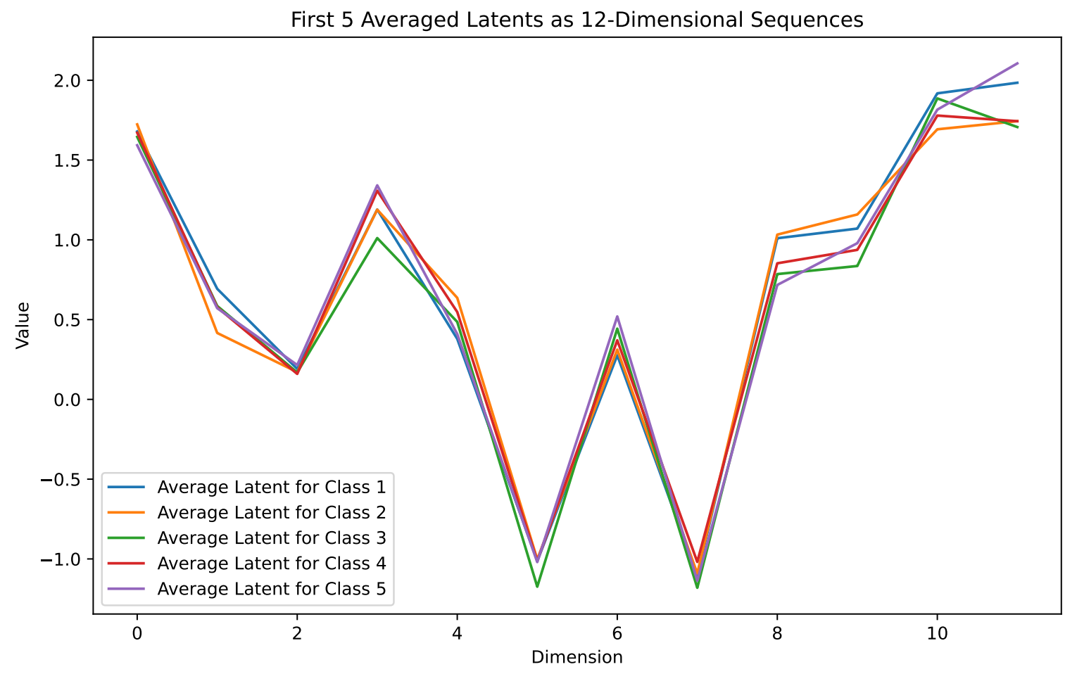
Table 1. results of 50-class classification.

100 Class	% Accuracy	% AUROC
SIAMVAE + NN Classifier	**19.57%**	**97%**
VAE + NN Classifier	1%	50%
NN Classifier	2.17	1,392
Random	1%	50%

Table 2. results of 100-class classification.

Note how the average EEG latents are very similar for different image classes





Further Steps

- Developing a basic prototype of a search engine operating on encoding unseen live EEG recordings
- Investigating more into using attention/transformers for the temporal aspect of EEG data
- Applying our two-stage approach to EEG-visual retrieval (instance-level)
- Conditioning on meta-categories (e.g, the image classes cat, dog, giraffe grouped under animals)

References

[Spa+17] C. Spampinato et al. "Deep Learning Human Mind for Automated Visual Classification". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4503–4511. DOI: 10.1109/CVPR.2017.479.

[Sin+23] Prajwal Singh et al. *EEG2IMAGE: Image Reconstruction from EEG Brain Signals*. 2023. arXiv: 2302.10121 [cs.HC].

[Kav+17] Isaak Kavasidis et al. "Brain2Image: Converting Brain Signals into Images". In: *Proceedings of the 25th ACM International Conference on Multimedia*. MM '17. Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 1809–1817. ISBN: 9781450349062. DOI: 10.1145/3123266.3127907. URL: https://doi.org/10.1145/3123266.3127907.

[Ye+22] Zesheng Ye et al. *See What You See: Self-supervised Cross-modal Retrieval of Visual Stimuli from Brain Activity*. 2022. arXiv: 2208.03666 [cs.MM].