Increasing Classes in Classification of EEG Waves from Visual Features

Albert Cai
albertc4@stanford.edu

Omar Abul-Hassan
omarah@stanford.edu

1 Introduction

The focus of our research is to find the best matches for content using signals from visually-evoked electroencephalography (EEG). When people view different images, their brain activity captured by EEG varies. The goal is to simply first classify the image that they saw.

Despite sounding simple, there are two major challenges:

1. EEG data is very noisy - completely impossible for a human to map from EEG $\rightarrow$ image class
2. Scarcity of EEG-image tuple data

These two challenges make it so the task is not as simple as training a multi-class classifier. We propose, like others (in 1.2) to learn low dimensional latent space representation of the EEG signals and train the multi-class classifier on the latent space instead. We explore novel and capable approaches for Variational Autoencoders (VAEs) to compress EEG signals into a promising latent space. The bigger picture goal we are working towards, out of scope of this project, is a generalized understanding to classes it has never even seen.

1.1 Motivation

Traditional search engines operate based on textual queries. While powerful, textual-search approaches relies on how well a user can accurately articulate their informational needs in words. There is a gap between a user's thoughts and cognitive abilities and their ability to express this in a concise, clear, searchable textual query.

Additionally, the internal process of converting our thoughts into words inherently loses information. Textual queries cannot capture all of the nuanced mental visualizations we have of our desired query. Hence, our project is mainly motivated by the problem of translating our rich space of human thoughts into words. Rather than relying on textual search, our project aims to bypass textual queries, and match content directly from EEG scans.

1.2 Related Works

Over the past 6 years there have been 4 cutting edge models for EEG classification.

- *SyncNet (2017) [Li+17]* CNNs with structured 1D convolutions. This model uses unique convolutional filters, which are essentially scaled and rotated Morlet wavelets. These filters are tailored to each EEG channel, allowing the network to effectively target spectral properties across different channels collectively. The network structure includes a Gaussian Process adapter to accommodate various electrode layouts and mitigate overfitting. Key to SyncNet's functionality is the focus on cross-spectral densities, providing insights into the synchronous nature of EEG signals across frequency bands.

Preprint. Under review.

- *EEGNet (2018) [Law+18]* EEGNet is a three-stage CNN architecture. The first stage employs 2D convolutions across time series to extract band-pass frequency features, followed by depthwise convolutions to learn frequency-specific spatial filters. This approach is inspired by the filter-bank common spatial pattern, allowing the model to efficiently capture spatial and temporal dynamics in EEG signals. In the second stage, EEGNet utilizes separable convolutions to decouple and summarize the temporal and spatial features. The final stage is a classification block that directly employs a softmax layer.
- *EEG-ChannelNet (2020) [Spa+17]*: This method introduces a novel approach to EEG signal processing by employing 1D convolutions across the time series and then across the channels. The architecture of EEGChannelNet is designed to efficiently extract relevant features from EEG data. After the initial convolutional layers, the network utilizes four residual layers. The use of residual layers aids in mitigating the vanishing gradient problem, allowing for deeper network architectures without compromising training efficiency.
- *GRUGate Transformer (2021) [Tao+21]* This approach is the current SOTA: it adapts the transformer architecture with a gating mechanism to enhance stability, particularly for EEG signal classification. Initially, raw EEG data undergoes input embedding, enhanced with positional encoding vectors based on sine and cosine functions, to preserve the temporal sequence of the EEG signals. The model employs several encoder blocks, each consisting of a multi-head attention layer for capturing non-local correlations across long EEG sequences, followed by a feed-forward neural network for deeper embedding. There is a gating layer post each sub-layer, based on the gated recurrent unit (GRU) mechanism, providing enhanced stability over traditional residual connections.

EEG-ChannelNet also built on previous neuroscience work which disputed that human brains process things in a parallel manner. In primate visual systems, the information comes down the optic nerve on the order of 10^8 bits per second. This is bottlenecked by the capability of brain, so instead, the brain selects "certain portions of the input to be processed preferentially, shifting the processing focus from one location to another in a serial, [not parallel], fashion" [IK00]. EEG-ChannelNet employed saliency maps, suggesting that upon seeing an image, the first quarter second in the brain is spent on low-level feature extraction, while the next quarter second is spent on "cognitive aggregation into different abstraction levels" [Spa+17]. They also tested their classification among different time interval, finding that the first 20 ms to 240 ms and the last 240 to 460 ms yielded equal classification results, indicating a balanced importance between low level and high level features.

1.3 Differentiation

The models mentioned in 1.2, are trained on the EEG-ImageNet dataset, with 40 classes (see 2.3 for more information). However, to reach the end goal of search engine via EEG, potentially many more than 40 classes are needed for generalized understanding. It is expected for classification accuracy to decrease significantly as the number of classes increases.

So, we instead create our models and train on the THINGS-EEG Dataset, which has 1,654 classes, more than 40x that of the EEG-ImageNet dataset. The models demonstrated in this paper are trained on 50 classes and 100 classes, which is 250% the classes of ImageNet. Not only is the classification task harder because of the increased classes, but the THINGS-EEG dataset has 17 channels and 100 time points, which, compared to the 128 channels and 440 time points of the ImageNet dataset, is 33 times less in input dimensionality. With an end goal of search by EEG, a 17 channel measurement system is far more commercial. See 2.2 for more information.

Our models are designed for better performance as classes increase (by 2.5x) and data decrease (by 1/33x).

2 Problem Statement

2.1 EEG Brain Waves

The occipital and parietal lobes are located at the back and middle of the brain respectively. The occipital lobe is the primary visual processing center of the brain. The parietal lobe processes other senses, and is responsible for touch, pressure, heat, cold, and pain. The parietal lobe has been thought as crucial for semantic analysis: turning "sensorimotor features into coherent and generalizable concepts" [CS18].

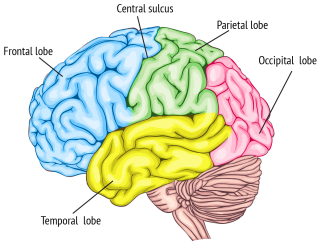

Figure 1: Labeled Lobes of Brain

Together, they process visual stimuli and turn them into generalizable concepts: object recognition. It is expected that when shown images, the occipital and parietal lobes activate in predictable patterns that correspond to the features and concepts represented in those images. This activation can be captured through EEG, which records electrical activity along the scalp produced by the firing of neurons within the brain.

2.2 THINGS-EEG Dataset

We use the **THINGS-EEG** dataset. The THINGS dataset consists of 1654 classes (called object concepts in original dataset paper), which can be any high-level human visual concept: e.g aardvark, abacus, airplane, zebra, basketball. For each of these classes, 10 images are collected, resulting in a dataset size of 16540 images. For each of these images, 4 image conditions are imposed (can be a rotation, or any sort of non-modifying change to the same image). Human subjects are then shown these 16540 images for every image condition, and the resulting EEG scan is recorded over 100 time data points.

For a brief explanation, each EEG scan is recorded over 00 time data points, each with 17 channels: electrodes placed at different locations on the occipital and parietal surface on the scalp. The 17 locations follow the internationally recognized 10-20 system by Herbert Jasper [58]. According to the initial dataset paper, the 17 channels correspond to locations: *(O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2, P4, P6, P8)*. For more information, see dataset paper at [Gro+22])

EEG plots can be visualized with a contour map. The white dots are the electrode positions, and red corresponds to greater readings.

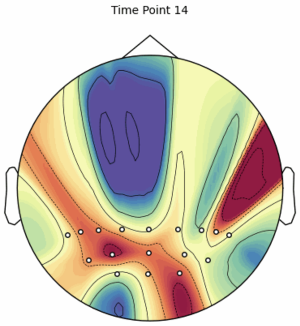

Figure 2: EEG of Patient 1. Shown: Aardvark, Time point: 14

2.3 EEG-ImageNet Dataset [Spa+17]

As a comparison of our models against state-of-the-art (SOTA) models, we train our models on the **EEG-ImageNet** dataset. The use of this dataset is soley for comparison.

We used a subset of the ImageNet dataset, comprising 40 different object classes, for the visual stimuli. In the experiment, participants were exposed to 2,000 images (50 per class), each displayed for 0.5 seconds in 25-second bursts, interspersed with 10-second breaks displaying a black screen, totaling approximately 23 minutes and 20 seconds. This data was acquired over 128 channels, where the THINGS-EEG dataset only used 17 channels.

The initial 40 ms (40 samples) of EEG data following each image were discarded to avoid interference from the previous image. This ensured that the data represented the cognitive processing of the current image. We analyzed the subsequent 440 ms (440 samples) for each image. Data values were centered around zero and underwent non-linear quantization.

2.4 Goals and Expected Results

To start, we wish to classify unseen images based on EEG scans, on classes the model has been trained on. To formalize with an example, aardvarks are a class the model will train on. After training, a patient is shown a new picture of an aardvark and his EEG scan is recorded. We wish to be able to identify that the patient was seeing an aardvark without seeing the image he was shown.

To evaluate our model's performance, we will employ several metrics. Accuracy on a test-train split will be a key metric, assessing how well our model generalizes to unseen data. Additionally, the Area Under the Receiver Operating Characteristic curve (AUROC) will be crucial. AUROC will help us distinguish between EEG scans corresponding to a specific class (e.g., aardvark) and scans of any other class, providing a measure of the model's discriminative ability.

To elaborate with an example on the THINGS-EEG dataset: We have 10 images (and 10 corresponding EEG recordings) corresponding to the image class of "cat". We train on nine EEG-image pairs, and hold out one for testing. We report the accuracy of this one for testing across the n classes we test on.

This is a building block to what we would like to further explore after this project. We hope to generalize to classes not in the 1654 classes in the THINGS dataset. This will be done over a separate testing dataset where patients have been shown different images and their EEGs have been recorded in a similar manner.

3 Technical Approach

We have $N = 1654 \times 10$ EEG-image tuples. We first separate the EEG scans from corresponding images to independently encode both modalities (EEG scans and images) using domain-specific variational autoencoders. We then hope to (not in progress report) use some sort of contrastive learning approach or method of matching these two latent spaces to return our initial matching of EEG-image tuples.

3.1 Mathematical Description

Variational Autoencoders (VAEs) consist of an encoder and a decoder. The encoder maps input data x to a latent representation z through a probabilistic mapping $q_\phi(z|x)$, where ϕ are the learned parameters. The decoder reconstructs the input data from this latent representation through another probabilistic mapping $p_\theta(x|z)$, where θ are its learned parameters.

The loss function for a VAE comprises two terms: the reconstruction loss and the KL divergence. The reconstruction loss ensures the output closely matches the input, and the KL divergence enforces a regularized latent space.

$$\mathcal{L}(\theta, \phi; x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \text{KL}(q_\phi(z|x) \| p(z))$$

We use F.mse for our reconstruction loss:

$$\text{MSE}(x, \hat{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

where x is the original input data, $\hat{x}$ is the reconstructed data from the decoder, and n is the number of data points.

The KL divergence term is given by:

$$\mathrm{KL}(q_\phi(z|x)\|p(z)) = -\frac{1}{2}\sum_{j=1}^{J}(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2)$$

where μ and σ are the mean and standard deviation of the latent variables, and J is the dimensionality of the latent space.

Therefore, the total loss function for the VAE is:

$$\mathcal{L}(\theta, \phi; x) = \mathrm{MSE}(x, \hat{x}) + \mathrm{KL}(q_\phi(z|x)\|p(z))$$

This loss function is crucial for training the VAE, as it balances the fidelity of the reconstruction with the regularization of the latent space.

3.2 Model Architecture

Vanilla VAE Our initial approach involved using a standard VAE with a Convolutional Neural Network (CNN) for the encoder and a mirrored architecture for the decoder. The encoder comprises convolutional layers followed by fully connected layers to produce the mean and log-variance of the latent distribution. The decoder utilizes transposed convolutional layers to reconstruct the input from the latent representation.

LSTM VAE To better capture the temporal dynamics in EEG data, we also experimented with a LSTM-based VAE. The LSTM encoder consists of LSTM layers to process the time-series EEG data. The final hidden state is then passed to fully connected layers to generate the latent distribution. The decoder reconstructs the EEG data from the latent space. It first maps the latent vectors back to the hidden state dimension using a fully connected layer, then employs LSTM layers for the reconstruction.

Bi-GAN An additional encoder component is integrated alongside the traditional generator and discriminator components of a GAN. The encoder maps the input data (only EEG) to a latent space, while the generator learns to map these latent representations back to the data space. This bidirectional mapping encourages the model to learn more meaningful and robust representations of the EEG data, which was hoped to be reflected in the latent space.

The encoder in our BiGAN is designed with convolutional layers, followed by fully connected layers, to effectively capture the spatial features of EEG signals. The generator has transposed convolutional layers to generate EEG-like data from latent space representations.

Conditional VAE We also explored the efficiency of a conditional VAE, where we learn a latent space that is conditioned on the EEG data. This consisted of two encoders: one for EEG data (two Linear layers) and one for image data (convolutional and linear layers). Because this approach seemed to be more initially promising, we explored various feature extractors and found that using a simple LSTM feature extractor before on EEG data worked best for the conditional VAE.

Siamese VAE For our best-performing model, we explored the efficiency of Siamese networks with VAEs. We understood that Siamese networks could help with learning discriminative features, particularly in tasks involving similar inputs. For the first part of this approach, we incorporated a Fast Fourier Transform (FFT) for our feature extractor when preprocessing EEG data, allowing for a more compact and computationally efficient feature representation.

We also found an increase in performance after incorporating the suggestions in [Spa+17], where they noted that the first quarter second in the brain is spent on low-level feature extraction, while the next quarter second is spent on “cognitive aggregation into different abstraction levels.” Hence, we used FFT on each quarter separately, and then appended results.

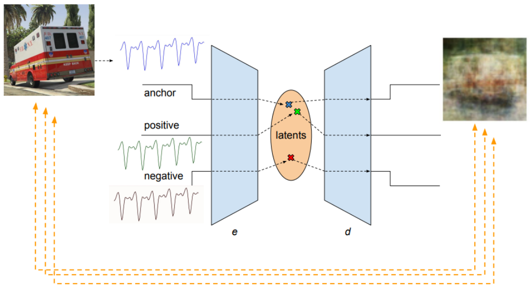

Figure 3: Siamese VAE

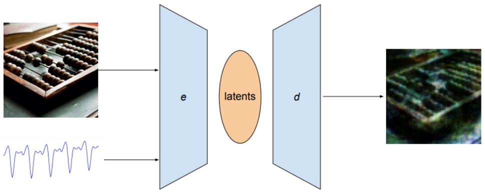

Figure 4: Conditional VAE

The Siamese architecture consisted of two encoders, which process similar and dissimilar EEG input pairs, and maps these to a shared latent space. The encoder and decoder are both two simple residual layers. We train this network via a joint training approach with a low learning rate, combining the traditional reconstruction and ELBO loss from the VAE architecture, as well as the triplet contrastive loss from the Siamese network:

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

where $\mathrm{d}(x_i, y_i) = \|x_i - y_i\|_p$, a_i is the anchor, p_i is a positive example, and n_i is a negative example for each sample in the mini-batch. The p-norm distance between the anchor and the positive example is minimized, and the distance between the anchor and the negative example is maximized, subject to the constraint imposed by the margin.

The triplet dataset is created by selecting an anchor (eeg1, img1), a positive example (eeg2, img2), and a negative example (eeg3, img3) for each triplet. The positive example is chosen from the same class as the anchor, while the negative example is chosen from a different class. This selection process is crucial for learning discriminative features in the embedded space. The final loss used to train the SiamVAE:

$$\text{MSE}(x, \hat{x}) + \beta \text{KL}(q_\phi(z|x) \| p(z)) + \alpha \text{TripletLoss}(a, p, n)$$

With this, we ensure that our approach doesn't just reconstruct data efficiently, but also develops an understanding of the difference in EEG inputs. This characteristic is particularly useful as class size increases, as we can produce a general understanding of differences between image classes. However, we find that the joint training approach worked well on the 40-class ImageNet dataset, discussed further in 4, but not on the original training dataset, THINGS-EEG. We suspect this to be because of the low sample size per class in the THINGS-EEG dataset, and the much-higher dimensionality of the EEG data in the ImageNet dataset (33x more than THINGS-EEG).

For our classifier, we used a simple MLP classifier.

4 Results

In Table 1, we compare the classification accuracies from our Siamese VAE architecture, named SiamVAE, and the four state of the art models in 1.2, on the 40-class EEG-ImageNet Dataset.

Method	**Accuracy**
SyncNet (2017) [Li+17]	31.7%
EEGNet (2018) [Law+18]	31.9%
SiamVAE	40.94%
EEG-ChannelNet (2020) [Spa+17]	48.1%
GRUGate Transformer (2021) [Tao+21]	**61.11%**

Table 1: Comparison to SOTA on 40-class EEG-ImageNet Dataset, using the High Gamma (55-95 Hz) filter over the time interval 20 ms to 460 ms,

In Table 2, we compare our models after opening up more classes (50 and 100) in the THINGS Dataset. We also report the multiclass-AUROC metric for this task, since we wanted to closely evaluate our model's ability to discriminate between classes.

Method	Accuracy	AUROC
CVAE + RF Classifier	16%	76%
SiamVAE + NN Classifier	**25.30%**	**96%**
VAE + NN Classifier	2%	50%
Random	2%	50%

Table 2: Results of 50-class classification

Method	Accuracy	AUROC
SiamVAE + NN Classifier	**19.57%**	**97%**
VAE + NN Classifier	1%	50%
Random	1%	50%

Table 3: Results of 100-class classification

The results after opening up to more classes are very promising, since the classification task becomes much more challenging when the number of classes increase. This is discussed further in 5.

We also would like to note that other models, like [Pal+18] uses an InceptionV3-based architecture for their EEG encoder and similarly-deep network for their Siamese network, each with millions of parameters. We use a simple variational autoencoder architecture with parameters of many magnitudes lower, indicating that our model may learn better or have much more room for improvement given deeper networks and more compute power to train these networks.

5 Discussion

The results presented in this research signify a substantial advancement in the field of EEG-based image classification. Our models, especially the Siamese VAE, show promising results in classifying EEG signals into image classes, even with an increased number of classes and reduced input data dimensionality. This achievement aligns well with the overarching goal of developing a more intuitive and direct method of content retrieval based on brain signals, potentially transforming how we interact with digital information.

5.1 Key Takeaways

1. **Performance with Increased Class Size:** Our models, particularly the Siamese VAE, demonstrate robustness in handling a larger number of classes. This is evident from the significant performance improvement over the baseline models when tested on the 50 and 100-class classifications from the THINGS-EEG dataset. Recall that other works have only reported metrics for the 40-class ImageNet dataset.
2. **Effectiveness of Siamese Architecture:** The Siamese VAE model's success suggests that leveraging the Siamese network's ability to discern between similar and dissimilar inputs is highly effective for EEG data. This approach could be vital for future research in brain-computer interfaces and EEG-based classification tasks.
3. **Adaptability to Reduced Input Data:** Despite the THINGS-EEG dataset having substantially fewer channels and time points than the EEG-ImageNet dataset, our models adapted well. This adaptability is crucial for practical applications, as it points towards the feasibility of using less complex and more user-friendly EEG recording setups.

5.2 Limitations and Future Directions

1. **Class Size vs. Performance:** While our models perform well with up to 100 classes, it remains to be seen how they would scale with even more classes. A critical future direction

would be to test these models on the entire THINGS-EEG dataset with 1,654 classes, which couldn't have been done with our current compute resources.

2. **Generalization to Unseen Classes:** A further extension of this work would be to test the models' ability to generalize to completely unseen classes, a step towards a more versatile and practical EEG-based search engine.
3. **Model Complexity and Training Resources:** We would like to greatly emphasize that our models were trained on local CPU, and hence we trained for very little epochs and purposefully made some of our network architecture very simple. The aspect of feature extraction was also inspired by our limited computational resources, and so future direction could look into deeper/more complex networks/feature extractors. Some of the SOTA related works used million-parameter networks, such as transfer learning used in [Spa+17].

6 Code

The Siamese VAE and visualization code can be found at 1 and the conditional VAE code can be found at 2.

References

[Li+17] Yitong Li et al. "Targeting EEG/LFP Synchrony with Neural Nets". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: `https://proceedings.neurips.cc/paper_files/paper/2017/file/7993e11204b215b27694b6f139e34ce8-Paper.pdf`.

[Law+18] Vernon J Lawhern et al. "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces". In: *Journal of Neural Engineering* 15.5 (July 2018), p. 056013. DOI: `10.1088/1741-2552/aace8c`. URL: `https://dx.doi.org/10.1088/1741-2552/aace8c`.

[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`.

[Tao+21] Yunzhe Tao et al. "Gated Transformer for Decoding Human Brain EEG Signals". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 2021, pp. 125–130. DOI: `10.1109/EMBC46164.2021.9630210`.

[IK00] Laurent Itti and Christof Koch. "A saliency-based search mechanism for overt and covert shifts of visual attention". In: *Vision Research* 40.10 (2000), pp. 1489–1506. ISSN: 0042-6989. DOI: `https://doi.org/10.1016/S0042-6989(99)00163-7`. URL: `https://www.sciencedirect.com/science/article/pii/S0042698999001637`.

[CS18] H. Branch Coslett and Myrna F. Schwartz. "Chapter 18 - The parietal lobe and language". In: *The Parietal Lobe*. Ed. by Giuseppe Vallar and H. Branch Coslett. Vol. 151. Handbook of Clinical Neurology. Elsevier, 2018, pp. 365–375. DOI: `https://doi.org/10.1016/B978-0-444-63622-5.00018-8`. URL: `https://www.sciencedirect.com/science/article/pii/B9780444636225000188`.

[58] "Report of the committee on methods of clinical examination in electroencephalography: 1957". In: *Electroencephalography and Clinical Neurophysiology* 10.2 (1958), pp. 370–375. ISSN: 0013-4694. DOI: `https://doi.org/10.1016/0013-4694(58)90053-1`. URL: `https://www.sciencedirect.com/science/article/pii/0013469458900531`.

[Gro+22] Tijl Grootswagers et al. *"Human electroencephalography recordings from 50 subjects for 22,248 images from 1,854 object concepts"*. OpenNeuro, 2022. DOI: `doi:10.18112/openneuro.ds003825.v1.2.0`.

[Pal+18] Simone Palazzo et al. "Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features". In: *CoRR* abs/1810.10974 (2018). arXiv: `1810.10974`. URL: `http://arxiv.org/abs/1810.10974`.