Towards Mechanistic Interpretability for Autoencoder compression of EEG signals

Leon Hegedić[1], Luka Hobor[1], Nikola Marić[1], Martin Ante Rogošić[1] and Mario Brcic[1]

[1]*University of Zagreb Faculty of Electrical Engineering and Computing, Unska ulica 3, Zagreb, Republic of Croatia*

Abstract

Convolutional Variational Autoencoders (VAEs) have found extensive application in dimensionality reduction, data compressibility, assessment, and signal analysis. However, a comprehensive understanding of their internal mechanisms remains elusive. This study aims to use mechanistic interpretability to elucidate the inner workings of VAEs. By training VAEs on images generated by interpolating EEG signals from the human brain and analyzing the resulting latent space, as well as the signal propagation through the network layers, we aim to construct an explanation of how these specific models internally analyze the generated images. Since we work under hardware constraints, we devised an iterative approach that breaks big task into easier, more manageable steps.

Keywords

Mechanistic Interpretability, EEG, Convolutional Variational Autoencoder, Iteratively Shaped Search

1. Introduction

Electroencephalogram (EEG) is a non-invasive tool for measuring brain activity by placing electrodes on the human scalp, which detect neuronal discharge voltage. While EEG technology possesses limitations such as a poor signal-to-noise ratio and capturing only surface brain activity, it remains a reliable method for diagnosing conditions like epilepsy and sleep disorders [1]. Autoencoders [2] are a specialized class of neural networks functioning as encoder-decoder pairs. The encoder compresses input data into a condensed representation, known as the latent space, by progressively reducing neuron count across layers, culminating in a bottleneck layer. Conversely, the decoder reconstructs input data from this compressed form by gradually increasing neuron count in subsequent layers. This compression and reconstruction process enables the network to capture salient features of the input data effectively. Convolutional Variational Autoencoders (CVAEs) [3, 4] extend this framework by incorporating convolutional layers, making them particularly adept at processing image data. Unlike standard autoencoders, CVAEs generate a probabilistic latent space. This probabilistic approach facilitates learning robust features and enhances the model's capability to generate new data instances resembling the training data. Leveraging convolutional layers, CVAEs exploit spatial hierarchies within data, enhancing their ability to analyze and reconstruct complex patterns and textures inherent in image data. Consequently, CVAEs find extensive application in tasks demanding detailed

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

Leon.Hegedic@fer.hr (L. Hegedić); Luka.Hobor@fer.hr (L. Hobor); Nikola.Maric@fer.hr (N. Marić); Martin-Ante.Rogosic@fer.hr (M. A. Rogošić); Mario.Brcic@fer.hr (M. Brcic)

analysis and synthesis of image content, offering significant improvements in both data reconstruction quality and interpretability of learned representations. Mechanistic interpretability [5] involves eliciting a simple algorithm from a learned ML model, under the assumption that there exists a human-understandable algorithm with low complexity that closely approximates the ML model [6]. This technique enhances the interpretability of ML models and aids in understanding the underlying mechanisms driving their predictions [7, 8]. It originates from the field of AI safety [9], but is increasingly finding application across various domains.

The motivation of this paper is to build upon the work in [10], where the authors used a Conditional Variational Autoencoder (CVAE) on EEG data. Our goal is to use the technique of mechanistic interpretability to uncover the basic algorithmic approach learned within the neural network. Given that this technique requires extensive manual work, intuition, hardware, and domain knowledge in neuroscience (the latter two of which we lack), we introduce a *stepwise guiding procedure through gradually relaxing constraints*. We begin with a highly constrained CVAE designed to plausibly approximate only one known algorithm. Our objective is to understand how the neural network implements this algorithm. Subsequently, we relax these constraints incrementally and monitor how the learned underlying algorithm adapts. We reckon that this procedure of tracking evolution from a known origin is easier than immediately interpreting mechanistically the unconstrained system.

This paper is organized as follows: Section 2 reviews related and relevant work. Section 3 presents our hypothesis, describes the stepwise guided approach to mechanistic interpretability, and introduces the tools we developed for this process. In Section 4, we detail our experimental results. Finally, Section 5 offers our conclusions and ideas for future work.

2. Related work

Neuroscience is a multidisciplinary field which aims to explain the functioning of the brain and the nervous system in general. The place where neuroscience meets machine learning is in trying to explain the computations which the brain preforms. The authors of [10] have used CVAEs in order to analyze EEG images of the brain. They took EEG signals from 32 electrodes placed on the human scalp, this way each data point represents a 32 dimensional vector, afterwards they used geometric transformations to project the positions of the electrodes onto a 40 by 40 grid. Finally, they performed cubic interpolation upon the grid which resulted in the images which comprised the dataset. Then they trained the models, one for each person. After training, the models showed significant capabilities in reproducing the images, showing that the data has an underlying structure.[1] They did this in an attempt to filter out spikes of neural activity from blinks, head movement etc. which in this context amount to noise.

Mechanistic interpretability is an emerging field in machine learning aimed at understanding the algorithms discovered by neural networks to solve specific problems. The papers [6, 11] were the first to explore the concept of "grokking" and explaining it, linking it to uncovering the underlying algorithm. Another paper ([12]) subsequently demonstrated that models do not consistently uncover the same algorithm. This finding highlights that the algorithm's nature is highly reliant not only on the model architecture and learning process but also on the

[1] It is important to note that in [10] the dimensionality of the latent space was set to 27.

inductive bias added by the initial chosen weights. The most significant results of mechanistic interpretability have been achieved on simple logical tasks using unimodal language models [5], e.g. modular arithmetic [6, 12] and othello gameplaying [13, 14]. There are attempts at automation of mechanistic interpretability such as automatic circuit detection [15] and attribution patching [16], but this work is only at the beginning.

3. Hypotheses, Approach, and Tools

We build upon the work in [10] by examining the performance of the Convolutional Variational Autoencoder (CVAE) in reconstructing images. Unlike the previous study, we also aggregate data from multiple individuals. To ensure mechanistic interpretability, we first conduct experiments to get grokking on each model, a process that may require extensive training. The task at hand is significantly more complex than previous problems addressed by mechanistic interpretability. While earlier works focused on discrete tasks such as arithmetic and board-game playing, our challenge involves the continuous approximation of EEG readouts. This necessitates an iterative approach to reduce and constrain the complexity of the analysis. To tackle this, we propose bootstrapping from a simple target algorithm, the *pick&interpolate* method. This method replicates how the 40x40 EEG images are generated from electrode values. Specifically, we constrain the neural network to select electrode positions from the input image and interpolate these values, mimicking the known method used to produce the input images. This approach allows us to incrementally increase complexity while tracking the underlying algorithm at each step, starting from the initial, known algorithm implemented in a neural manner.

3.1. Hypotheses

After analyzing the underlying mechanisms, we aim to gradually relax constraints and track changes in mechanisms. This iterative process enables us to break down the complex task into simpler steps conducive to analysis with more limited hardware resources. While the original EEG signal comprises 32 dimensions, the interpolation process expands it to 40x40 (1600 dimensions), indicating redundancy beyond the original 32 dimensions. We identify two factors complicating the learned algorithms:

1. Bottleneck layer size below 32: we simplify by setting the layer size to 32.

2. Biases of convolutional neural networks: Specifically, the bias away from locating objects. To mitigate this bias, we introduce an injection layer at the input to inject helpful artifacts into the image. Additionally, the network architecture may not be powerful or expressive enough to learn the direct interpolation algorithm - we will check this with an experiment.

3.2. Tools

The **injection layer** introduces helpful artifacts into the input image, including reference watermarks and occlusion of non-informative points. We have created 20 reference 2x2-pixel watermarks that uniquely locate 20 out of 32 electrodes. These watermarks are positioned 1px below its pertinent point. The remaining 12 electrodes are at the edge of head, so can be located

by the local curvature. **One-hot signals** are synthetic signals with one electrode set to 1 and others to 0, to investigate signal propagation through the VAE. This is a further simplification from the complex mixtures on which VAE is trained.

3.3. Approach

Our iterative approach involves the following steps:

E1 Test the decoder with pure electrode signals to assess its capacity to reproduce interpolated images faithfully.

E2 Utilize a latent layer size of 32 and bias VAE in a way to target approximating pick&interpolate algorithm. We initially mark the inner electrodes and the edges of the skull, then occlude all positions that do not correspond to the marked areas to guide the VAE's focus. Special constant watermarks were added to the inner electrodes to provide the model with relative information about the position of the electrodes. We train this occlusion&mark version of the VAE using original images from DEAP and a loss function that ignores pixels outside the circle. We investigate the signal propagation first with one-hot signals, and if necessary expand to original signals.

E3 Remove occlusion to allow the VAE more freedom in selecting where to focus, which may steer from picking the electrode pixels. Again, we investigate the signal propagation first with one-hot signals, and if necessary expand to original signals. The focus is on changes with respect from the mechanisms in the previous step.

E4 Further relax constraints by reducing the latent layer size to 27, as in [10], and track differences with respect to previous step.

E5 Analyze the original setting in [10] without marking.

This iterative approach allows us to systematically explore variations and uncover the mechanisms underlying the CVAE's image reconstruction capabilities.

4. Experiment

We build upon the basic experimental setup from [10]. That means we are using publicly available **DEAP dataset** where EEG data were collected from 32 persons who watched 40 one-minute music video clips [17]. Therein standard 10–20 systems were applied with the following 32 electrode positions: 'Fp1', 'AF3', 'F7', 'F3', 'FC1', 'FC5', 'T7', 'C3', 'CP1', 'CP5', 'P7', 'P3', 'Pz', 'PO3', 'O1', 'Oz', 'O2', 'PO4', 'P4', 'P8', 'CP6', 'CP2', 'C4', 'T8', 'FC6', 'FC2', 'F4', 'F8', 'AF4', 'Fp2', 'Fz', 'Cz'. We conducted our experiments in the Google Colab environment with V100 GPU. Prior to training, we preprocessed the dataset by clipping values to the lower and upper 5 quantiles and normalizing them to address issues related to blinks. Our implementation was based on the Keras framework, employing the AdamW optimizer with weight decay set to 1, as suggested in [6]. We initialized the learning rate (LR) to 10^{-3} and applied LR reduction on plateau with a patience of 30 epochs and factor of 0.3.

The VAE encoder architecture consists of three convolutional layers (kernel=4, stride=2), followed by three LeakyReLU layers, and concludes with two fully connected layers, one for the mean and one for the variance of the distribution. The number of filters is set to 32 for the first layer, 64 for the second layer, and 128 for the third layer. Conversely, the decoder comprises one fully connected layer, one reshape layer, and three Conv2DTranspose layers (kernel=4, stride=2). These layers collaborate to reconstruct the images, with a sigmoid function applied at the end to ensure proper output scaling. We have conducted the three initial experimental iterations, with the rest left as continuation. The preliminary results are given below. **E1** After the initial epoch, the decoder displayed an MSE of 1.803e-04 for the training set and 2.6914e-04 for the validation set, with a slight yet consistent decline thereafter. Upon analyzing the outcomes from 1, it becomes evident that the reconstruction shows promise, exhibiting minor discrepancies around the brighter areas of the image. The metrics for the test set were SSIM = 0.980, MSE = 2.807e-4 and MAE = 0.007. As indicated by both the metrics and 1, the decoder demonstrated high accuracy in reproducing the images, indicating that the architecture possesses ample capability to learn the interpolation algorithm. This enables our study based on initially targeting biasing our model to pick&interpolate algorithm.

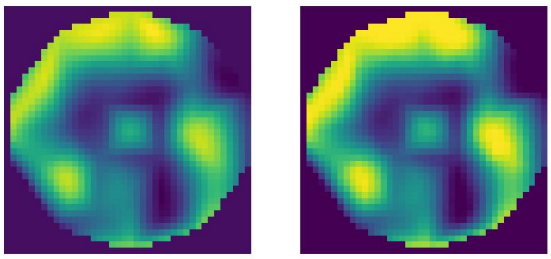

Figure 1: The image on the left is the original interpolated image and the image on the right is the output of the decoder.

E2 After 200 epochs, the model converged with a training MSE of 3.6375e-4 and a validation MSE of 0.0866. Although the model exhibited some limitations in reproducing details, it effectively captured the global features of the one-hot signals. Since we ignored the pixels outside the circle in our loss function, the model did not eliminate them. Instead, it extended the patterns from inside the circle to the outside in a continuous and smooth manner, suggesting the use of some interpolation algorithm.

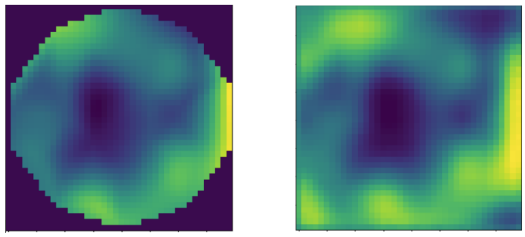

Figure 2: The image on the left is the original interpolated image and the image on the right is the output of the occluded model.

E3 The model exhibited solid performance after several hundred iterations and continued to improve gradually even at the final iteration of 2000. Throughout the iterations, the model consistently maintained a validation loss approximately twice the MSE of the training set. Upon analyzing the materials presented in 3, the image reconstruction appeared nearly flawless, with only minute differences observed around the bright and dark spots, which would not be noticeable without a closer examination. The obtained metrics, though smaller than expected, were as follows: SSIM = 0.899, MSE = 0.004 and MAE = 0.046. This indicates that the model was indeed capable of performing well and reconstructing the image accurately. Subsequently, we proceeded with the analysis using one-hot signals to hypothesize of the algorithm. Upon observing the activations transitioning from layer to layer, we made the following observations:

1. The encoder, spanning from the 1st to the 3rd convolutional layer, transforms the input image into feature maps. Before reaching the fully connected layer for the latent space, we obtain 128 maps with spatial dimensions of 5x5. Each map is activated to some extent, reflecting the intensity at a relative position of the input signal, akin to resizing the input image to 5x5 and introducing some noise. Despite slight variations, these activation maps serve to accurately encode and differentiate different input signals. Consequently, the information at the output of the encoder comprises two key components: the activation position, providing a rough estimate of the input signal's position, and subtle differences surrounding the activation, offering a finer estimate of which signal is active and to what extent.

2. When mapping into the latent space, two main factors are emphasized: the sigma, or standard deviation, is consistently very small (1e-7) and therefore insignificant, and the bias of the fully connected layer, which is also close to zero (around 1e-2). Consequently, the combination of convolutions, LeakyReLU activation and a linear layer with a negligible bias suggests an overall linear transformation.

3. Activation maps at this deconvolution layers are very similar, due to small differences in scale. This suggests that the latent space exhibits a high degree of similarity across all inputs, which is something we aimed to avoid in order to simplify the interpretation. Consequently, the decoder relies on minor differences in its initial layers to extract information. Already at the 1st deconvolutional layer, the identity of the signal being processed becomes evident.

The gap from E1 to E2 could in future be circumvented by stage-wise (instead of end-to-end) training where we would take decoder from E1 and teach the encoder to pick points. Other, easier alternative is initialization of the VAE close to the pick&interpolate algorithm and then tracking the evolution of the underlying algorithm.

5. Conclusion and future work

Our future work involves finalizing our experimental plan through all its steps, as only the initial three are partially completed. Meanwhile, based on our current experiments, we have gained an understanding and experience with mechanistic interpretability that may also benefit

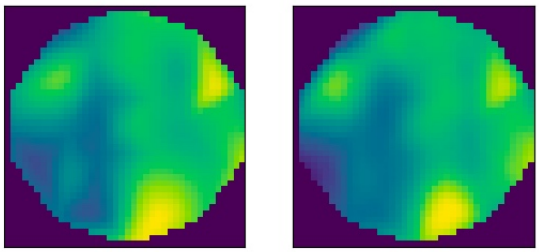

Figure 3: The image on the left is again the original EEG topographic map placed into the model and the image on the right is the model output.

other researchers. Grokking [6, 11] may take a long time, if it happens at all, as it depends on hyperparameter values. Whether the model will grok under certain settings is not evident beforehand (maybe even not decidable [18]), and extensive experimentation is necessary to find suitable values. Training is often unstable, with many oscillations, and sensitivity to hyperparameters is high. For instance, the authors in [6] could not achieve grokking using L1 normalization. Considering the above, searching for good configurations for grokking is computationally expensive. Due to our modest hardware resources, we adopted an iterative approach. Initially, we shaped and constrained the initial step solution, which we could interpret anchored on the initial target algorithm pick & interpolate. Then, we gradually allowed more freedom to the model until it matched the architecture of interest. We also observed that domain expertise is necessary to facilitate easier interpretability. However, this was a hindrance to our team, as none of us is well-versed in neuroscience. Therefore, we resorted to more abstract and basic algorithmic features. Additionally, previous work in mechanistic interpretability focused on nicely structured domains in arithmetic and logic, while the problem addressed in this paper is qualitatively more challenging. Looking further ahead, automation of mechanistic interpretability is a valuable goal to pursue [5]. It aims to circumvent the manual work currently performed, which is subject to all human limitations. Such discovery algorithms would mine the inner workings of black-box systems, searching for patterns with low algorithmic complexity and reformulations similar to fragments in the current codebase available in repositories.

References

[1] B. CD, P. PF., Electroencephalography, Journal of Neurology, Neurosurgery and Psychiatry (1994).

[2] R. Marc'Aurelio, P. Christopher, C. Sumit, L. Yann, Efficient learning of sparse representations with an energy-based model, Conference on Neural Information Processing Systems (2006).

[3] K. Diederik P., W. Max, Auto-encoding variational bayes, International Conference on Learning Representations (2014).

[4] S. Aman, O. Tokunbo, An overview of variational autoencoders for source separation, finance, and bio-signal applications, MDPI (2022).

[5] L. Bereska, E. Gavves, Mechanistic interpretability for ai safety: A review, arXiv e-prints (2024). URL: https://arxiv.org/abs/2404.14082v1. arXiv:2404.14082.

[6] N. Neel, C. Lawrence, L. Tom, S. Jess, J. Steinhardt, Progress measures for grokking via mechanistic interpretability, arXiv (2023).

[7] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210–0215. doi:10.23919/MIPRO.2018.8400040.

[8] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000794. doi:https://doi.org/10.1016/j.inffus.2024.102301.

[9] M. Juric, A. Sandic, M. Brcic, Ai safety: state of the field through quantitative lens, in: 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 1254–1259. doi:10.23919/MIPRO48935.2020.9245153.

[10] A. Taufique, L. Luca, Interpreting disentangled representations of person-specific convolutional variational autoencoders of spatially preserving eeg topographic maps via clustering and visual plausibility, MDPI (2023).

[11] T. Vimal, L. Etai, Z. Shuangfei, S. Omid, P. Roni, S. Joshua, The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon, arXiv (2022).

[12] Z. Ziqian, L. Ziming, T. Max, A. Jacob, The clock and the pizza: Two stories in mechanistic explanation of neural networks, arXiv (2023).

[13] K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, M. Wattenberg, Emergent world representations: Exploring a sequence model trained on a synthetic task, arXiv preprint arXiv:2210.13382 (2022).

[14] Z. He, X. Ge, Q. Tang, T. Sun, Q. Cheng, X. Qiu, Dictionary learning improves patch-free circuit discovery in mechanistic interpretability: A case study on othello-gpt, arXiv e-prints (2024). URL: https://arxiv.org/abs/2402.12201. arXiv:2402.12201.

[15] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, A. Garriga-Alonso, Towards automated circuit discovery for mechanistic interpretability, arXiv e-prints (2023). URL: https://arxiv.org/abs/2304.14997. arXiv:2304.14997.

[16] A. Syed, C. Rager, A. Conmy, Attribution patching outperforms automated circuit discovery, arXiv e-prints (2023). URL: https://arxiv.org/abs/2310.10348. doi:10.48550/arXiv.2310.10348. arXiv:2310.10348.

[17] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis ;using physiological signals, IEEE Transactions on Affective Computing 3 (2012) 18–31. doi:10.1109/T-AFFC.2011.15.

[18] M. Brcic, R. V. Yampolskiy, Impossibility results in ai: A survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3603371. doi:10.1145/3603371.