Toward Generating Plausible Artificial Electroencephalography Data: Evaluating the Effects of Convolution-based Upsampling Methods on Data Quality

Clemens Kaiser (c.kaiser@rug.nl)
Bernoulli Institute of Mathematics, Computer Science & Artificial Intelligence, University of Groningen
Groningen, 9747 AG Groningen, The Netherlands

Natasha Maurits (n.m.maurits@umcg.nl)
Department of Neurology, University of Groningen, University Medical Center Groningen
Groningen, 9713 GZ Groningen, The Netherlands

Joukje van der Naalt (j.van.der.naalt@umcg.nl)
Department of Neurology, University of Groningen, University Medical Center Groningen
Groningen, 9713 GZ Groningen, The Netherlands

Marieke van Vugt (m.k.van.vugt@rug.nl)
Bernoulli Institute of Mathematics, Computer Science & Artificial Intelligence, University of Groningen
Groningen, 9747 AG Groningen, The Netherlands

Abstract

Electroencephalography is key for clinical and cognitive research, yet limited data availability restricts deep learning (DL) applications. This study compares thee upsampling techniques in convolution-based Generative Adversarial Networks – transposed convolutions (TC), interpolation with convolutions (IC), and a mixed approach – that are used to transform noise vectors into artificial EEG data. We evaluate artificial signal quality with EEG-specific metrics across time, frequency, and spatial domains. Kolmogorov–Smirnov tests indicate that the mixed approach mitigates the high-frequency noise commonly introduced by TC, while better preserving lower-frequency components and inter-channel dependencies than IC. Moreover, the findings underline the importance of EEG-specific evaluation metrics for guiding the development of more explainable and efficacious generative models, advancing DL applications in neuroscience.

Keywords: Generative Adversarial Network; electroencephalography; data augmentation; convolutional neural network; evaluation

Introduction

In principle, electroencephalography (EEG) data is promising for addressing clinical and cognitive questions. However, difficulties collecting sufficient within- and between-participant EEG data hinder the use of powerful deep learning methods. One way to tackle this issue is by augmenting the data with artificial samples. Generative Adversarial Networks (GAN) have become the gold standard for EEG data augmentation (Habashi et al., 2023). However, the convolution-based generation process introduces systematic high-frequency noise into the generated EEG data (see Figure 1b).

Convolutional neural networks (CNN) are the most common architecture for GANs in EEG research (Habashi et al., 2023). A CNN can generate complex, high-dimensional data by iteratively upsampling and transforming an unstructured, low-dimensional noise vector. This upsampling process allows the model to refine the intermediate feature maps step-by-step, learning a hierarchical representation that mirrors the target data structure. The most common upsampling method in CNN-based GANs uses transposed convolutions. Transposed convolutions upsample and transform an intermediate feature map simultaneously. Unfortunately, this upsampling method introduces high-frequency artifacts into the generated EEG data – an issue rarely addressed in the literature (Habashi et al., 2023).

Odena, Dumoulin, & Olah (2016) designed a convolution-based upsampling method to avoid the high-frequency noise introduced by transposed convolutions. In this approach, the upsampling is separated from the computation of features. An intermediate feature map is upsampled by a non-parametric interpolation layer and then transformed by one or more convolutional layers. However, interpolations are averaging operations, introducing a "smoothness bias" into the generated

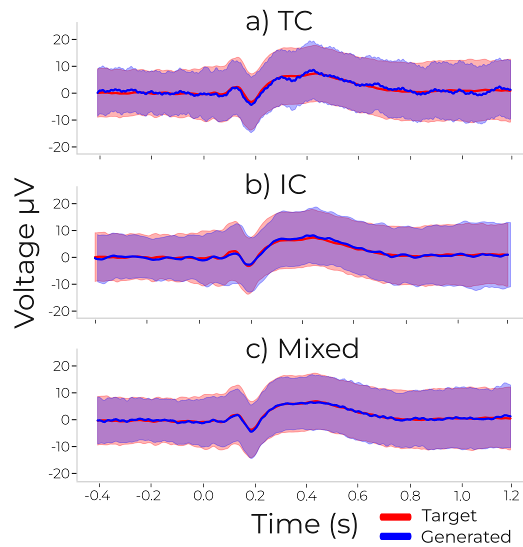

Figure 1: Comparison of the average trial (channel P08) for the (a) Transposed-Convolutions (TC), (b) Interpolation-and-Convolutions (IC), and (c) Mixed models. The model data (blue) is plotted on top of the target data (red) for comparison.

data. Although the subsequent convolutions counteract this bias, the generated data commonly has degraded amplitudes (Hartmann, Schirrmeister, & Ball, 2018).

Panwar et al. (2020) proposed a model alternating between the two upsampling methods. They showed that the mixed upsampling method outperformed the individual upsampling methods. However, they relied on a global metric for comparing the models, which provides no information on how the data generated from different models differs.

This study has two goals. First, we compare three upsampling methods for convolution-based GANs for generating EEG data: (1) transposed convolution and (2) interpolation and convolutions. We demonstrate that combining both upsampling methods leads to a more accurate reproduction of the target data in the frequency domain. Second, while most research in this field relies on global metrics, we propose EEG-specific metrics in the time, frequency, and spatial domains to evaluate artificially generated EEG data. EEG-specific metrics yield more fine-grained, and thus, actionable insight. This constitutes an important step towards building more explainable, generalizable, and efficacious generative models for EEG data in the future.

Methods

We used EEG data from Jin, Borst, & van Vugt (2019), restricting ourselves to channels Fz, Pz, P07, and P08. Participants performed a Go/No-Go task, responding to frequent lowercase words (non-targets) and withholding responses to rare uppercase words (targets). We used only target trials to keep the data relatively simple, yielding 5,568 trials.

Table 1: Comparison of features between (a) target data and data generated by (b) Interpolation-and-Convolutions (IC), (c) Transposed-Convolutions (TC), and (d) Mixed models. Values are reported as means with standard deviations ($M \pm$ Std). Reported p-values are from KS tests comparing each generated distribution to the target. Correlations with significant deviations in at least one model are shown. Statistically significant deviations are indicated by * ($p < 0.05$) and ** ($p < 0.01$).

		Target	IC		TC		Mixed	
		$M \pm$ Std	$M \pm$ Std	p	$M \pm$ Std	p	$M \pm$ Std	p
Time	$P3_L$	443 ± 96	453 ± 98	0.59	437 ± 91	0.58	449 ± 95	0.87
	$P3_A$	24.5 ± 11.7	26.2 ± 11.5	0.53	23.9 ± 11.3	0.89	25.0 ± 11.4	0.83
Frequency	SEn	3.33 ± 0.40	3.41 ± 0.34	$< 0.01^{**}$	3.41 ± 0.37	0.01^{*}	3.38 ± 0.39	0.14
	SEn_T	1.69 ± 0.12	1.70 ± 0.12	0.20	1.69 ± 0.12	0.88	1.69 ± 0.11	0.96
	SEn_A	1.76 ± 0.15	1.82 ± 0.12	$< 0.01^{**}$	1.77 ± 0.13	0.24	1.77 ± 0.13	0.17
	SEn_B	3.04 ± 0.16	3.03 ± 0.17	0.35	3.07 ± 0.13	0.01^{*}	3.05 ± 0.13	0.34
Space	Fz - Pz	0.61 ± 0.23	0.59 ± 0.16	0.04^{*}	0.58 ± 0.21	0.32	0.63 ± 0.19	0.45
	Fz - P08	0.27 ± 0.27	0.34 ± 0.20	0.03^{*}	0.24 ± 0.25	0.63	0.36 ± 0.22	0.03^{*}
	Pz - P07	0.70 ± 0.17	0.63 ± 0.14	$< 0.01^{**}$	0.68 ± 0.16	0.57	0.72 ± 0.14	0.38

We built three models to compare the upsampling methods used in GAN generators for EEG data. These models differed only in the composition of the four upsampling blocks they employed in their generators: (1) transposed convolution (TC), (2) interpolation and convolutions (IC), and (3) alternating between them (Mixed). We adapted these models from Panwar et al. (2020).

To assess the quality of the generated data, we picked measures in the time, frequency, and spatial domain and determined them for each EEG trial. We compared the generated and target distributions on each metric using Kolmogorov-Smirnov (KS) tests. The KS test determines if two samples follow the same distribution function (Berger & Zhou, 2014).

For the time domain, we determined the P3 peak amplitude ($P3_A$) and latency ($P3_L$) as the maximum voltage between 250 and 600 ms after stimulus onset at Pz to measure a model's ability to capture global features of an EEG signal. For the frequency domain, we computed the spectral entropy for the full frequency range (SEn) as well as for the frequency bands theta (4-8 Hz; SEn_T), alpha (8-13 Hz; SEn_A), and beta (13-30 Hz; SEn_B). SEn quantifies the degree of randomness in a signal's frequency content by measuring the spread of power across a frequency range (Inouye et al., 1991). For the spatial domain, we computed by-trial correlations for every channel pair.

Results

Figure 1 displays the average signal for channel P08 per model. While models captured the overall trajectory of the target EEG data, the IC model exhibited the closest resemblance, with minimal residual noise. In contrast, the TC and Mixed models introduced systematic high-frequency noise.

Table 1 presents the quantitative results, contrasting each feature's distribution between the generated and the target EEG data. In the time domain, none of the models yielded distributions that differed significantly from the target data, with high p-values ($p > 0.50$). By contrast, in the frequency domain, the TC and the IC model significantly deviate from the target SEn for the full frequency range ($p = 0.01$ and $p < 0.01$, respectively). Furthermore, whereas the IC model failed to capture the SEn_A ($p < 0.01$), the TC model failed to capture the SEn_B ($p = 0.01$). Spatially, only the TC model captured all correlations, while the IC model significantly deviated from the target data in three channel pairs.

Discussion

This study contributes to the existing literature in two main ways. First, it proposes a set of EEG-specific metrics to holistically evaluate the quality of generated EEG data in the time, frequency, and spatial domain. These measures provide actionable insights, allowing targeted steps to improve a generative model. Second, it contrasted two upsampling methods commonly employed in convolution-based Generative Adversarial Networks (GAN): transposed convolution and interpolation followed by convolutions. Our results indicate that a mixed method – alternating between the two upsampling methods – is generally advantageous, preserving the target data's frequency content and mitigating high-frequency noise. However, this comes with a trade-off in spatial accuracy, which is best captured by transposed convolution alone

References

Berger, V. W., & Zhou, Y. (2014, September). Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online.* John Wiley & Sons, Ltd. Retrieved 2025-03-11, from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06558

Habashi, A. G., Azab, A. M., Eldawlatly, S., & Aly, G. M. (2023). Generative adversarial networks in EEG analysis: an overview. *Journal of NeuroEngineering and Rehabilitation*, *20*(1), 40.

Hartmann, K. G., Schirrmeister, R. T., & Ball, T. (2018). EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv preprint arXiv:1806.01875*.

Inouye, T., Shinosaki, K., Sakamoto, H., Toi, S., Ukai, S., Iyama, A., ... Hirano, M. (1991, September). Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalography and Clinical Neurophysiology*, *79*(3), 204–210. Retrieved 2025-01-28, from https://www.sciencedirect.com/science/article/pii/001346949190138T doi: 10.1016/0013-4694(91)90138-T

Jin, C. Y., Borst, J. P., & van Vugt, M. K. (2019, August). Predicting task-general mind-wandering with EEG. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(4), 1059–1073. Retrieved 2024-03-01, from https://doi.org/10.3758/s13415-019-00707-1 doi: 10.3758/s13415-019-00707-1

Odena, A., Dumoulin, V., & Olah, C. (2016, October). *Deconvolution and Checkerboard Artifacts.* Retrieved 2025-03-11, from https://staging.distill.pub/2016/deconv-checkerboard/

Panwar, S., Rad, P., Jung, T.-P., & Huang, Y. (2020, August). Modeling EEG Data Distribution With a Wasserstein Generative Adversarial Network to Predict RSVP Events. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(8), 1720–1730. (Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering) doi: 10.1109/TNSRE.2020.3006180