ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447463

CEMOAE: A DYNAMIC AUTOENCODER WITH MASKED CHANNEL MODELING FOR ROBUST EEG-BASED EMOTION RECOGNITION

*Yu-Ting Lan, Wei-Bang Jiang, Wei-Long Zheng and Bao-Liang Lu**

Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Emotion recognition through electroencephalography (EEG) has been an area of active research, but the inherent sensitivity of EEG signals to noise and artifacts poses significant challenges, especially in real-world settings. These complications often necessitate the removal of corrupted channels, making it crucial to develop robust models capable of maintaining performance even when few channels are available. To address this, we propose the Corrupted EMOtion AutoEncoder (CEMOAE), an innovative approach that leverages masked channel modeling to maintain robust performance, achieved through three components: masked autoencoder pretraining for robust representation learning, random masked auxiliary task for implicit modeling of channel corruption, and masked auto-repair to explicitly narrow the data distribution gap between high-quality and corrupted EEG signals. Specifically, we first pretrain a masked autoencoder with the dynamic masking strategy for feature extractor initialization and channel recovery. During the finetuning stage, we mask EEG data using the auxiliary task to mimic real-world EEG corruption. We then employ the pretrained autoencoder to repair these signals and finetune the feature extractor for emotion recognition. Experiments on the SEED dataset demonstrate that CEMOAE achieves SOTA performance for emotion recognition under the random channel corruption simulation, validating the effectiveness of the proposed techniques.

***Index Terms*—** Masked channel modeling, robust EEG-based emotion recognition.

1. INTRODUCTION

Electroencephalography (EEG) enables investigations into the temporal dynamics of the brain and its cognitive processes for a wide range of purposes, including fatigue detection, mental disease diagnosis, and affective computing [1, 2, 3]. As a physiological signal that directly measures brain activities, EEG has been demonstrated to be a simple, reliable, and easy-to-use solution for recognizing human emotions [4]. However, it is still unfeasible to translate EEG from lab and clinic to real-world settings such as at-home and ambulatory environments. One of the most critical bottlenecks is the noise and artifacts during EEG recordings [5]. The weak EEG signals are extremely sensitive to the body signal interference and external environments, including the electrical activities of the eyes, heart, and muscles, electrical artifacts due to cable movements, and electromagnetic interferences from the surroundings [6]. In real-world settings, the quality of EEG signals is hard to control, and we have to discard many corrupted channels, which imposes significant limitations on the practical applications of EEG signals.

To build a robust model in real-world scenarios, researchers have employed various methods to reduce the influence of excessive noise or channel corruption on the performance. Those methods can be roughly divided into three categories: directly ignoring, implicit denoising, and explicit denoising [7]. Directly ignoring is the simplest way to deal with noise to assume it is negligible or to simply discard bad segments [8]. Implicit denoising approaches can be used to design noise-robust processing pipelines. For example, Pierre *et al.* proposed a robust feature by capturing spectral, temporal, and spatial patterns of EEG signals [9]; Hubert *et al.* proposed dynamic spatial filter (DSF) to conduct interpretable modeling [7]. Explicit denoising deals explicitly with noise by correcting corrupted signals or predicting missing or additional channels from those available, *e.g.*, autoreject [10] and gated-layer autoencoders [11].

However, these approaches have some limitations. Simply discarding bad channels will cause the loss of usable information and requires algorithms to be robust to the removal of any number of EEG channels, *e.g.*, Neural Processes family [12, 13, 14]. The implicit denoising methods and noise-robust processing pipelines might not work with limited channels available, *e.g.*, only 10 channels out of 62 are retained. Traditional explicit denoising methods often treat channel corruption as a distinct time-series imputation

*Corresponding author.

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, National Natural Science Foundation of China (Grant No. 62376158), Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

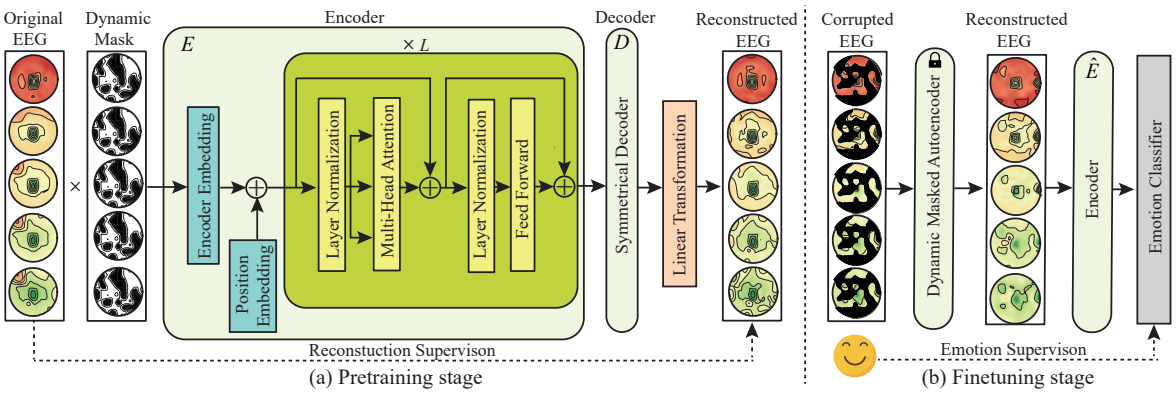

Fig. 1: The framework of CEMOAE. The black channels indicate the masked channels in the random sampling. All the modules with dotted lines, i.e. reconstruction supervision and emotion supervision are only used during the training phase and would be removed during the inference phase.

problem, leading to inconsistencies and degraded performance, especially when the corruption is relatively moderate. Additionally, most existing solutions are designed for raw spatio-temporal EEG data and may not suitably address the challenges associated with differential entropy (DE) [15] features in the spatial spectrum domain, which is crucial for EEG-based emotion recognition. Further, the question of effectively combining these diverse approaches to enhance their strengths and address their limitations remains unanswered.

To tackle with aforementioned challenges in this task, we propose a Corrupted EMOtion AutoEncoder, namely, CEMOAE, a comprehensive pipeline using the masked channel modeling through masked autoencoder pretraining, random masked auxiliary task, and masked auto-repair, to handle the corrupted emotion recognition. The masked autoencoder pretraining serves to construct robust representations by predicting the masked EEG channels. The random masked auxiliary task and masked auto-repair act in concert to provide both implicit and explicit denoising, thereby effectively modeling the complex channel corruption scenarios and narrowing the data distribution gap between high-quality and corrupted EEG signals. We systematically evaluate the performance of CEMOAE on a public dataset SEED [16] for emotion recognition with DE features under the simulation of random channel corruption. The experimental results demonstrate that CEMOAE achieves SOTA performance and further improves the feasibility and the performance of EEG-based emotion recognition in practical applications.

2. METHODOLODY

In this section, we design our method CEMOAE, which handles emotion recognition of corrupted EEG signals via masked channel modeling. We briefly introduce the intuition of our masked channel modeling in three corresponding components: masked autoencoder pretraining for robust representation learning, random masked auxiliary task for implicit modeling of channel corruption, and masked auto-repair to explicitly narrow the data distribution gap between high-quality and corrupted EEG signals. The three masked channel modeling methods work in unison to construct a robust emotion recognition model, effectively tackling the challenges associated with corrupted EEG signals.

In the following, we first formulate the problem and give an overview of our method. Then, we outline the masked channel pretraining with the dynamic masking strategy during the pretraining phase. Finally, we describe the masked auxiliary task and masked auto-repair in the finetuning stage.

2.1. Problem Statement and Overview of CEMOAE

In this section, we formalize the problem of dealing with corrupted EEG channels and give an overview of our method.
Problem Statement Let $X \in \mathcal{R}^{C\times F}$ denote the DE features of EEG signals. Here, C and F stand for the number of channels and the dimension of the DE features, respectively. During the training phase, we assume that EEG recordings are obtained in a controlled clinical environment, where all C channels are available. Conversely, in the inference stage, we assume that EEG data are gathered in ambulatory settings denoted by $X_{\text{real}} \in \mathcal{R}^{C_{\text{real}}\times F}$ and only limited C_{real} channels are available. The primary objective is to ensure that the performance of the emotion recognition model remains robust when its inputs transform from X to X_{real}.
Overview As delineated in Figure 1, the initial phase involves the pretraining of the masked autoencoder. This is accomplished by predicting the masked EEG data $X_m \in R^{C_m\times F}$ using the visible data $X_v \in R^{C_v\times F}$. Here $C_v + C_m = C$, which is sampled based on the dynamic masking strategy. Such a dynamic masking autoencoder serves dual roles: initializing the parameters of the feature extractor and recovering the corrupted channels. During the finetuning stage, we mask EEG data using the random auxiliary task A to mimic

real-world EEG corruption. We then employ the pretrained dynamic autoencoder to repair these signals and finetuning the pretrained feature extractor $\hat{E}$ for emotion recognition.

2.2. Pretraining Stage

In this section, we outline the methodology for pretraining with the mask channel transformer-based autoencoder. Additionally, we detail our dynamic masking strategy, for both representation learning and auto-repair of corrupted channels.

Masked Channel Autoencoder During the pretraining stage, we randomly partition the EEG feature signals X of each batch into a visible set X_v and a masked set X_m. For each batch, the masked set X_m is dropped, and the visible DE features X_v are inputted into an encoder E. The encoder's output is concatenated with the learnable masking tokens and then inputted into a decoder D to reconstruct the EEG data $\hat{X}$. The quality of the reconstruction is supervised using the Mean Squared Error (MSE) loss as defined in Equation (1):

$$\mathcal{L}_{\text{reconstruction}} = \|\hat{X} - X\|_2^2 = \|D(E(X_v)) - X\|_2^2. \quad (1)$$

Masking Strategy We talk about the design of the masking strategy here. We first consider the straightforward fixed masking strategy: random sampling on channels with a fixed masking ratio. However, unlike the fixed masking ratio, *e.g.*, 90% and 15%, which perform well in vision tasks [17] and NLP tasks [18], respectively. Our task requires the autoencoder to understand the complex channel corruption scenarios. The fixed masking ratio may limit the potential of our model, especially when the number of simulated corrupted channels differs significantly from the number of masked channels during pretraining. Therefore, we finally choose the random sampling with a dynamic masking ratio, which means the visible channel number of EEG signals is randomly sampled in a set S_v with equal probability for each batch.

2.3. Finetuning Stage

In this section, we describe the finetuning stage, which employs the random masked auxiliary task and the masked auto-repair. These modules serve both implicit and explicit denoising functions to implicitly model the channel corruption scenarios and explicitly bridge the data distribution gap between high-quality and corrupted EEG signals.

Masked Auxiliary Task To model the channel corruption, we introduce the random masked auxiliary task A. Specifically, for each batch, we randomly mask some EEG channels, denoted as $A(X) \in R^{C_a \times F}$ like the dynamic masking strategy. This approach enables the model to understand channel corruption scenarios, thereby enhancing its robustness.

Masked Auto-Repair After simulating channel corruption, we employ the pretrained autoencoder to reconstruct the EEG features as $D(E(A(X))) \in \mathcal{R}^{C \times F}$. Given that some channels are dropped, this method serves to bridge the data distribution gap, particularly in complex channel corruption situations, thereby enhancing performance.

Finetuning Pipeline In this stage, each subject's model is finetuned individually. After masking EEG signals with A and making auto-repair with the dynamic autoencoder E and D, we instantiate a new encoder $\hat{E}$ initialized with pre-trained parameters from E. This encoder $\hat{E}$ is then finetuned with the loss as follows:

$$\mathcal{L}_{\text{emotion}} = \mathcal{L}_{\text{CE}}(Y, \hat{Y}) = \mathcal{L}_{\text{CE}}(Y, \hat{E}(D(E(A(X))))), \quad (2)$$

where Y and $\hat{Y}$ is the ground-truth and predicted emotion label, respectively.

3. EXPERIMENT

In this section, we present the experimental settings and the results with extended analysis. We list the following research questions (RQs) to lead the experimental discussion. **RQ1**: Does the CEMOAE achieves the best performance by using the masked channel modeling among all the compared methods? **RQ2**: Is the combination of masked autoencoder pretraining, random masked auxiliary task, and masked auto-repair better than utilizing a single strategy only? **RQ3**: Is the dynamic masking strategy better than the fixed masking strategy for the robust EEG-based emotion recognition?

3.1. Experimental Settings

Dataset The proposed model is evaluated on a public affective EEG dataset SEED [16] with DE features [15]. SEED contains 15 participants, three sessions each, with video stimuli inducing negative, neutral, and positive emotions. For each session, the first 9 trials are considered for training and the remaining 6 for testing. Following [19], the model is pretrained on the training data from all subjects and individually finetuned using the emotion supervision.

Evaluation Metrics Our corruption simulations strictly follow the previous baseline NPA [13]. Specifically, we conduct simulations with varying channel availabilities, including Full, 50, 40, 30, 20, and 10 channels, repeating each configuration 50 times, and consider the mean and standard deviation of emotion classification accuracy as evaluation metrics.

Implementation Details The visible channel set S_v is [10, 20, 30, 40, 50] for the dynamic masking strategy and the masked auxiliary task. Further implementation specifics, such as the masked autoencoder (MAE) [20] structures, optimizers, learning rate, and more, adhere to [19].

3.2. Experimental Results

Comparing with Baselines We compare CEMOAE with **LSTMNet** [13], which models the channels as sequences; **NPA** [13], which combines the prior knowledge of Gaussian distribution to build robust pipelines; **MV-SSTMA**, which

Table 1: The mean/std (%) accuracy of various models on the SEED dataset. The LSTMNet [13], NPA [13], MV-SSTMA [19], 1 (Transformer), and 2 (MAE [20]) can be considered as baselines. The models from 1 to 6 are ablation studies of CEMOAE.

Model	P	A	R	M	62 (Full)	50	40	30	20	10
LSTMNet	-	-	-	-	83.79/10.04	80.58/10.17	77.63/10.42	73.47/10.93	67.66/12.26	58.73/14.00
NPA	-	-	-	-	79.66/12.49	78.66/12.63	77.64/12.70	76.70/13.03	75.26/13.29	71.96/13.03
MV-SSTMA	-	-	-	-	**95.32/03.05**	-	-	-	-	-
1	✗	✗	✗	-	90.03/07.12	78.47/08.72	71.28/09.26	64.34/09.31	56.25/08.44	46.82/07.42
2	✓	✗	✗	0.75	92.27/05.19	59.01/15.62	47.23/10.61	39.64/07.78	36.65/06.82	35.16/05.69
3	✗	✓	✗	-	88.46/08.60	86.43/08.87	84.48/09.16	82.02/09.56	78.29/09.88	71.09/10.43
4	✓	✓	✗	0.75	89.77/08.28	88.48/07.88	86.89/08.07	84.36/08.67	80.66/08.71	72.43/09.49
5	✓	✓	✗	Dynamic	90.09/07.44	88.38/07.68	86.81/08.08	84.30/08.40	80.36/09.28	72.48/09.49
6	✓	✓	DSF	Dynamic	87.81/09.92	84.48/09.95	82.25/10.13	77.21/09.80	72.09/09.70	65.45/09.19
CEMOAE	✓	✓	AR	Dynamic	92.03/07.29	**90.21/07.02**	**88.08/07.31**	**85.17/07.88**	**80.81/08.65**	**73.20/09.43**

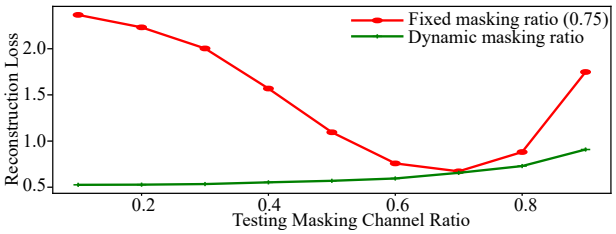

Fig. 2: The average MSE loss of reconstructing EEG signals with the different masking strategies for the test samples.

is the SOTA method for high-quality EEG signals in this dataset. In addition, models 1 and 2 can be considered as **Transformer** [21] and **MAE** [20]. As is shown in Table 1, by comparing the results of the CEMOAE and the baseline models, we see that the CEMOAE significantly outperforms the classical deep learning methods in corrupted channel simulations; In high-quality EEG signal scenarios (Full), CEMOAE still maintains a good performance compared to the SOTA MV-SSTMA. These quantitative results demonstrate the superiority of our model with masked channel modeling, addressing **RQ1**.

Abalation Study We also conduct an ablation study to analyze the effectiveness of each module of our CEMOAE to further answer **RQ2** and **RQ3** in Table 1. Here, P and A represent the masked channel pretraining and the masked auxiliary task, respectively. R denotes the approach to repair EEG signals using DSF [7] (a SOTA technique for recovering raw time-series EEG signals), either through masked auto-repair (AR) or not at all. Meanwhile, M designates the masking strategies, delineating between a fixed masking ratio and a dynamic masking ratio.

- **Masked Channel Pretraining:** Comparing models 1 and 2, and then models 3, 4, and 5, we deduce that using the masked channel pretraining improves performance by better representation learning, especially when finetuning with the masked auxiliary task, which partially answers **RQ2**.
- **Masked Auxiliary Task:** Comparing models 1 with 3 and 2 with 4 or 5, we find that the auxiliary task notably increases performance by implicitly modeling the corruption scenarios, which partially answers **RQ2**.
- **Masked Auto-Repair:** From the evaluation of models 5, 6, and CEMOAE, we infer that the masked auto-repair significantly bolsters EEG emotion recognition performance. It achieves this by directly recovering the corrupted signals and reducing the distribution disparities in channel corruption. Additionally, using DSF [7] is not suitable for this task as DE feature modeling diverges from that of raw EEG signals, which partially answers **RQ2**.
- **Fixed Masking Strategy vs Dynamic Masking Strategy:** Comparing models 4 and 5, we find that the fixed masking ratio and dynamic masking ratio perform almost the same for the downstream tasks without auto-repair. However, as is shown in Figure 2, the reconstruction performance (loss) of the dynamic autoencoder (red) generalizes better with the different testing masking ratios, which can provide more robustness to repair the complex and diverse channel missing or corrupting situations in our experiments. This analysis further answers **RQ2** and **RQ3**.

4. CONCLUSIONS

In this paper, we have proposed a dynamic autoencoder with masked channel modeling for real-world EEG-based emotion recognition. Our innovative approach leverages masked channel modeling to maintain robust performance, achieved through three components: masked autoencoder pretraining for robust representation learning, random masked auxiliary task for implicit modeling of channel corruption, and random masked auto-repair to explicitly narrow the data distribution gap between high-quality and corrupted EEG signals. Extensive experiments on the SEED dataset have demonstrated the outstanding performance of our framework for emotion recognition compared with various advanced baseline models under complex random channel corruption. This work underscores the potential for devising EEG measurements, paving the way for feasible real-world applications.

5. REFERENCES

[1] Sadegh Arefnezhad, James Hamet, Arno Eichberger, Matthias Frühwirth, Anja Ischebeck, Ioana Victoria Koglbauer, Maximilian Moser, and Ali Yousefi, "Driver drowsiness estimation using EEG signals with a dynamical encoder–decoder modeling framework," *Scientific Reports*, vol. 12, no. 1, pp. 1–18, 2022.

[2] Wei-Long Zheng and Bao-Liang Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, pp. 026017, 2017.

[3] Dongrui Wu, Bao-Liang Lu, Bin Hu, and Zhigang Zeng, "Affective brain-computer interfaces (aBCIs): A tutorial," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1314–1332, 2023.

[4] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, 2019.

[5] Jesus Minguillon, M Angel Lopez-Gordo, and Francisco Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, 2017.

[6] Xiao Jiang, Gui-Bin Bian, and Zean Tian, "Removal of artifacts from EEG signals: a review," *Sensors*, vol. 19, no. 5, pp. 987, 2019.

[7] Hubert Banville, Sean UN Wood, Chris Aimone, Denis-Alexander Engemann, and Alexandre Gramfort, "Robust learning from corrupted EEG with dynamic spatial filtering," *NeuroImage*, vol. 251, pp. 118994, 2022.

[8] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, pp. 051001, 2019.

[9] Pierre Thodoroff, Joelle Pineau, and Andrew Lim, "Learning robust features using deep learning for automatic seizure detection," in *Machine Learning for Healthcare Conference*, 2016, pp. 178–190.

[10] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort, "Autoreject: Automated artifact rejection for meg and eeg data," *NeuroImage*, vol. 159, pp. 417–429, 2017.

[11] Heba El-Fiqi, Kathryn Kasmarik, Anastasios Bezerianos, Kay Chen Tan, and Hussein A Abbass, "Gate-layer autoencoders with application to incomplete eeg signal recovery," in *International Joint Conference on Neural Networks*. IEEE, 2019, pp. 1–8.

[12] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami, "Conditional neural processes," in *International Conference on Machine Learning*, 2018, vol. 80, pp. 1704–1713.

[13] Yan-Kai Liu, Wei-Bang Jiang, and Bao-Liang Lu, "Increasing the stability of EEG-based emotion recognition with a variant of neural processes," in *International Joint Conference on Neural Networks*, 2022, pp. 1–6.

[14] Chen-Li Yao and Bao-Liang Lu, "A robust approach to estimating vigilance from EEG with neural processes," in *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2020, pp. 1202–1205.

[15] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu, "Differential entropy feature for EEG-based emotion classification," in *International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2013, pp. 81–84.

[16] Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[17] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He, "Masked autoencoders as spatiotemporal learners," *arXiv preprint arXiv:2205.09113*, 2022.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu, "A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning," in *ACM International Conference on Multimedia*, 2022, pp. 6–14.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.