applied sciences

MDPI

Article

A Simplified Query-Only Attention for Encoder-Based Transformer Models

Hong-gi Yeom [1,2] and Kyung-min An [3,*]

[1] Department of Electronics Engineering, Chosun University, 309 Pilmundae-ro, Dong-gu, Gwangju 61452, Republic of Korea; honggi@chosun.ac.kr
[2] Interdisciplinary Program in IT-Bio Convergence System, Chosun University, Gwangju 61452, Republic of Korea
[3] Centre for Human Brain Health, School of Psychology, University of Birmingham, Birmingham B15 2TT, UK
* Correspondence: k.m.an@bham.ac.uk

Abstract: Transformer models have revolutionized fields like Natural Language Processing (NLP) by enabling machines to accurately understand and generate human language. However, these models' inherent complexity and limited interpretability pose barriers to their broader adoption. To address these challenges, we propose a simplified query-only attention mechanism specifically for encoder-based transformer models to reduce complexity and improve interpretability. Unlike conventional attention mechanisms, which rely on query (Q), key (K), and value (V) vectors, our method uses only the Q vector for attention calculation. This approach reduces computational complexity while maintaining the model's ability to capture essential relationships, enhancing interpretability. We evaluated the proposed query-only attention on an EEG conformer model, a state-of-the-art architecture for EEG signal classification. We demonstrated that it performs comparably to the original QKV attention mechanism, while simplifying the model's architecture. Our findings suggest that query-only attention offers a promising direction for the development of more efficient and interpretable transformer-based models, with potential applications across various domains beyond NLP.

Keywords: transformer; EEG conformer; self-attention; deep learning; brain–computer interface

check for updates

Citation: Yeom, H.-g.; An, K.-m. A Simplified Query-Only Attention for Encoder-Based Transformer Models. *Appl. Sci.* **2024**, *14*, 8646. https://doi.org/10.3390/app14198646

Academic Editors: Pedro Couto and Douglas O'Shaughnessy

Received: 23 July 2024
Revised: 28 August 2024
Accepted: 20 September 2024
Published: 25 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Transformers have revolutionized Natural Language Processing (NLP), enabling machines to accurately understand and generate human language [1]. The transformer-based models, like the generative pre-trained transformer (GPT) [2–5] or Gemini [6,7], have empowered users worldwide to interact with artificial intelligence (AI) and conveniently access information. A recent multi-modal language model can recognize not only text but also images or videos [8]. Furthermore, the state-of-the-art model allows us to communicate in real-time with voice instead of text [9]. These transformer-based models leverage self-attention mechanisms to effectively capture the interactions between words in a sentence, moving beyond traditional sequential methods. The attention mechanism allows the transformer to understand a context better and perform well in NLP tasks, such as machine translation [10–14], summarization [13–15], and question answering [4,6,13,14]. As a result, the transformer has become the most dominant architecture in NLP [14,16–19].

The transformer is widely used not only in NLP but also in various fields, such as image processing [20–23], speech recognition [18,24,25], and neural signal processing [26–29]. In this regard, modified versions of transformer algorithms have been proposed so that the model can adequately extract features from different data types [18–21,26–29]. The vision transformer (VIT) outperformed recent convolutional neural network (CNN) algorithms by splitting an image into patches and feeding the patches to an encoder of a standard transformer [20]. The conformer could capture an audio sequence's local and

global dependencies by combining CNNs and transformers [18]. Similarly, the EEG conformer performed excellently by combining CNNs and transformers for neural signal processing [26]. The D-FaST applied attention to frequency, space, and time domains [29].

The transformer is composed of encoders and decoders [1]. The encoder consists of multi-head attention and a feed-forward network, and the decoder comprises masked multi-head attention, multi-head attention, and a feed-forward network. The feed-forward network is a conventional neural network, and multi-head attention is the core of the transformer. Attention is calculated from query, key, and value inputs by computing the dot product, scaling, softmax, and another dot product. Multi-head attention performs the process several times to take different attention views. Masked multi-head attention computes attention by masking a part of the input data to prevent the use of future data as input.

Numerous modified versions of the transformer algorithms have been proposed, using only the encoder part for classification [20,26–29]. Most of them use query (Q), key (K), and value (V) to calculate the attention as initially proposed [20,26–29]. The K and V can be redundant in the encoder. However, many modified algorithms for the encoder-based transformer still use all the Q, K, and V components.

In this paper, we propose a simplified query-only attention mechanism to enhance the efficiency and interpretability of transformer models by removing redundant elements. The transformer, using only a Q, performed similarly to the transformer using Q, K, and V, even in a simple structure. To evaluate the proposed method, we applied the simple attention and conventional attention mechanisms to the equal transformer model and compared their accuracy. To show that the proposed method plays the same role as the previous method, even in simple structures, we applied the simple attention method and the previous attention method to the equal structure of the transformer. We compared the performance of the two cases. An EEG conformer was used as the transformer structure because the state-of-the-art model performs well by extracting local features with a convolutional layer and recognizing global relations with attention. The open datasets (BCI competition IV dataset 2a and b) were used, which are commonly used for evaluation [30].

2. Related Works

Transformers were proposed for sequence-to-sequence tasks, such as machine translation, in which a sentence in one language is converted into another [1]. The transformer architecture consists of an encoder and a decoder. The encoder is responsible for encoding the input sequence into a compact representation, while the decoder generates the output sequence (e.g., a translated sentence) from this representation. The attention mechanism plays a crucial role by allowing the model to focus on different parts of the input sequence at each step of the output generation, modeling the relationships within and between the input and output sequences.

Owing to their exceptional performance in NLP, transformers have been widely adopted across domains beyond text processing [20,26–29,31–34]. In computer vision, where convolutional neural networks (CNNs) were traditionally dominant, the vision transformer (ViT) introduced a novel approach by dividing an image into patches and processing them using a transformer without any convolutional layers [20]. The ViT could achieve superior performance compared to state-of-the-art (SOTA) CNNs. This breakthrough led to the widespread adoption of transformers in image recognition tasks [31–34].

Transformers have also been increasingly applied in neural signal processing [26–29]. For instance, the EEG conformer model combined convolutional layers and a transformer [26]. The EEG conformer extracts features from EEG signals through convolutional layers and processes them within a transformer to predict the user's intent. The EEG conformer has shown superior performance compared to previous state-of-the-art models in this domain. Many BCI studies have employed transformers for various classification tasks [27–29].

In many classification tasks, such as image recognition and neural signal processing, the encoder is often used independently, without the decoder. This is because the

encoder compresses the high-dimension input sequences into the low-dimension outputs, whereas the decoder expands the low-dimension inputs into the high-dimension output sequences [20,26–29]. As a result, the encoder-based transformer does not require attention between input and output sequences, as used in sequence-to-sequence tasks like translation. In the encoder, the Q, K, and V vectors are all derived from the same input sequence through linear transformations, while in the decoder, Q, K, and V are generated from the input and output sequences [1]. Consequently, the Q, K, and V vectors in the encoder can be similar to each other, leading to redundancy. Despite this, many applications use distinct Q, K, and V components, even when only the encoder is utilized.

In this paper, we propose a simplified query-only attention mechanism for encoder-only transformer models. Our experiments demonstrated that the query-only attention mechanism achieves good performance comparable to traditional attention, which uses Q, K, and V, even in a simple structure. The proposed mechanism saves computational time and memory by eliminating the need to generate and learn the K and V components. This approach can be applied to most transformer models that rely solely on the encoder.

3. Materials and Methods

Unlike previous attention mechanisms that utilize all three Q, K, and V vectors, our proposed method employs only the Q vector, reducing computational complexity and model complexity. To evaluate the proposed method, we applied simple attention and previous attention to the EEG conformer structure [26]. We utilized the open-source EEG conformer codes to implement simple attention. The EEG conformer codes are available at https://github.com/eeyhsong/EEG-Conformer (accessed on 20 September 2024). The data are available at http://www.bbci.de/competition/iv (accessed on 20 September 2024).

3.1. Dataset

The proposed method was evaluated with two widely used EEG datasets, BCI competition IV dataset 2a and 2b. Details of the data can be found in paper [30]. The sampling rate was 250 Hz for both datasets. The signals were bandpass filtered between 0.5 and 100 Hz and notch filtered at 50 Hz in both cases.

Dataset I: BCI competition IV 2a was acquired from nine subjects performing a cue-based motor imagery task with four classes (left hand, right hand, both feet, and tongue). The EEG signals were recorded using 22 Ag/AgCl electrodes. The left mastoid was a reference, and the right mastoid was a ground. Two sessions were conducted on separate days for each subject. Each session included six runs, with 48 trials per run (12 trials per class), resulting in 288 trials per session. Each trial began with a fixation cross on a black screen and a brief auditory cue (t = 0 s), as shown in Figure 1a. After two seconds (t = 2 s), a directional cue (arrow) appeared for 1.25 s, indicating the desired motor imagery. The subjects performed the imagery task until the fixation cross disappeared at t = 6 s. No feedback was provided during the trials.

Dataset II: BCI competition IV 2b consists of EEG data from nine subjects performing a cue-based motor imagery task with two classes (left and right hand). Three bipolar electrodes (C3, Cz, and C4) were recorded. The left mastoid was a reference, and the Fz electrode was a ground. For each subject, five sessions were recorded. The first two sessions were acquired without feedback, and the last three sessions were measured with online feedback.

The two sessions without feedback measured six runs, with 20 trials per run (60 trials per class) for each person. A fixation cross was presented on a screen at the beginning of each trial (t = 0 s), as shown in Figure 1b. After two seconds, a beep sound was given (t = 2 s). After three seconds from the beginning (t = 3 s), a directional cue (arrow) appeared for 1.25 s, indicating the desired motor imagery. The subjects were instructed to imagine the corresponding hand movement for 4 s.

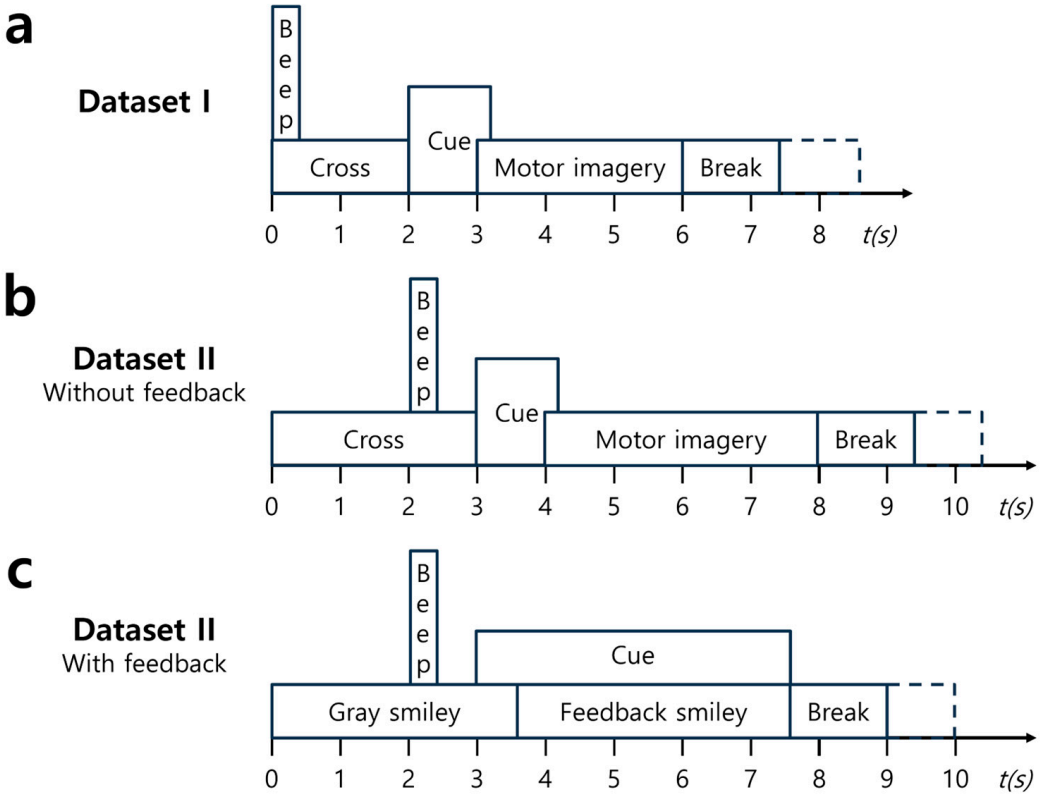


Figure 1. Experimental paradigm. (**a**) Paradigm of dataset I. (**b**) Paradigm without feedback of dataset II. (**c**) Paradigm with feedback of dataset II.

The three sessions with online feedback recorded four runs, with 40 trials per run (80 trials per class) session for each person. A gray smiley was presented on a screen at the beginning (t = 0 s), as shown in Figure 1c. A beep sound was given after two seconds (t = 2 s). The directional cue was presented from 3 s to 7.5 s, instructing the motor imagery.

3.2. Transformer Architecture

We applied the proposed simple attention and the previous attention to the structure of the EEG conformer and compared the accuracies. The EEG conformer is a state-of-the-art deep learning model for EEG signal classification [26]. It extracts local and global features using convolution and attention modules. The EEG conformer consists of three main modules: a convolutional module, a self-attention module, and a classifier module, as illustrated in Figure 2.

(1) Convolutional module: This module consists of two convolutional layers and one pooling layer. The first convolutional layer extracts temporal features using 40 kernels (1×25) with a stride of (1, 1). The second convolutional layer plays a role as a spatial filter, using 40 kernels (22×1) with a stride of (1, 1). The number of electrode channels was 22. The average pooling layer reduces the amount of feature dimension using a kernel (1 * 75) with a stride of (1, 15). Consequently, the convolutional module effectively captures the local spatio-temporal features of EEG signals.

(2) Self-attention module: This module utilizes a self-attention mechanism to capture global interactions. Self-attention learns relationships between distant EEG signals, enabling more accurate classification.

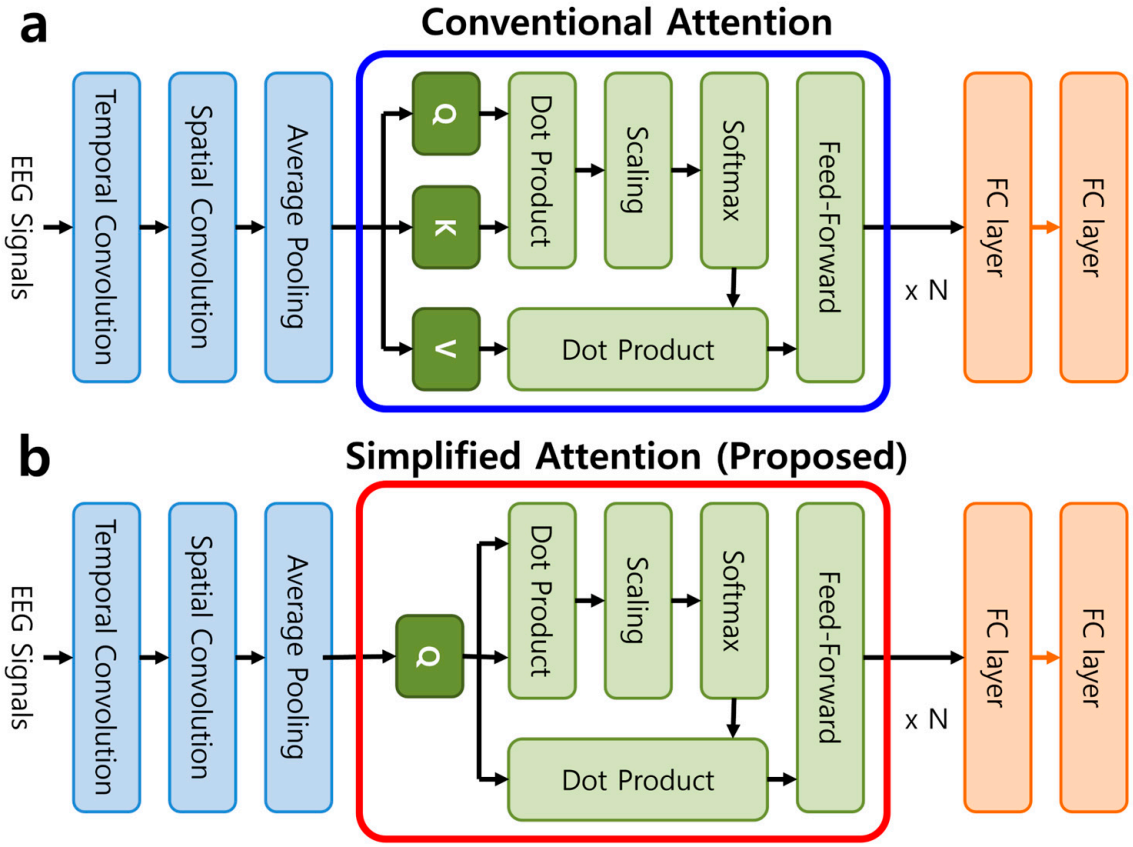


Figure 2. EEG conformer structure. The convolutional module is illustrated in sky blue, the self-attention module in light green, and the fully connected layer module in orange. (**a**) EEG conformer with the conventional attention mechanism using query (Q), key (K), and value (V) vectors. (**b**) EEG conformer with the proposed simplified attention mechanism using only a query (Q).

The previous self-attention used query (Q), key (K), and value (V) to calculate the attention, as in Figure 2a. The previous attention is computed using Equation (1).

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k denotes the length of Q, K, and V. A single-layer, feed-forward network generates the Q, K, and V from processed EEG signals. These values can be considered latent vectors of the input (EEG). Therefore, the K and V are redundant with the Q, although the Q, K, and V might differ slightly. The Q can substitute the K and V.

The proposed query-only attention mechanism calculates attention scores using only the Q vector, as in Figure 2b. It can be formulated using Equation (2).

$$Attention(Q) = softmax\left(\frac{QQ^T}{\sqrt{d_k}}\right)Q \tag{2}$$

In the encoder, the Q, K, and V are linear combinations of the same input X. Therefore, the Q, K, and V can be represented as AX, BX, and CX, respectively. The expression QK^TV in Equation (1) can be rewritten as AXX^TB^TCX. If we represent Q in Equation (2) as DX, where D is a matrix trained from a different linear network, then QQ^TQ becomes DXX^TD^TDX. Thus, when A equals D and B^TC equals D^TD, Equations (1) and (2) become equivalent. Therefore, in this case, Equation (1) can be entirely replaced by Equation (2). However, a matrix D that mathematically satisfies the conditions $A = D$ and $B^TC = D^TD$ may not always exist, and A, B, and C are not fixed matrices with specific val-

ues. The matrices can be changed based on the initial conditions and the training iterations. Therefore, a D matrix may be found that produces outputs similar to those generated by the flexible combination AXX^TB^TCX.

The replacement makes the model efficient and allows for a more intuitive understanding. Figure 2 illustrates the procedures of the previous and the proposed attention.

Multi-head attention performs a self-attention h times to capture the diverse interactions. In this study, the h was 10. The calculation of the multi-head attention was repeated N times. In this case, the N was 6.

(3) Classifier module: This module classifies the category of the EEG signals, based on the features extracted from the previous modules. Two fully connected layers were applied. A cross-entropy loss function was used for model training.

3.3. Evaluation

We evaluated our proposed method on an EEG conformer model and compared its performance to the original attention mechanism. We applied the same training and testing methods with the EEG conformer study for consistency [26]. For dataset I, the first session of the experiment was used for model training, and the second was used for testing. The EEG signals were segmented from 2 to 6 s based on the visual cue for each trial that corresponded to movement imagination. For dataset II, the first three sessions were used for training, and the last two sessions were used for testing. The signals were epoched from 3 to 7 s based on the visual cue for each trial. The epoched signals were used as input for the EEG conformer model for both datasets I and II. The data augmentation method was used because a large dataset was required to train the model. Segmentation and reconstruction (S&R) were performed in the time domain for the data augmentation. The number of segments, Ns, in S&R was 8. An Adam optimizer was used to train the EEG conformer. The parameters were 0.0002, 0.5, and 0.999 for the learning rate, $\beta 1$, and $\beta 2$, respectively. The number of self-attention computations (N) and multi-heads were 6 and 10, respectively. The number of epochs and the batch size were 2000 and 72, respectively. We evaluated the prediction accuracy for each epoch of training and testing. The average accuracy of all the epochs and a kappa value were shown for evaluation. A paired-sample t-test was performed between the prediction accuracies of the previous and the proposed methods. The kappa value can be calculated using Equation (3).

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o denotes the average accuracy of all the epochs and p_e means the accuracy of the random guesses.

4. Results

The results indicate that our method achieves a performance that is comparable to the original, while also simplifying the model's architecture. The accuracy transitions on the dataset I were similar between the previous attention and the simple attention, as shown in Figure 3. The prediction accuracies rapidly increased, according to the training in both methods. Then, the accuracy was saturated. The blue and red lines show the average accuracy among the subjects. The sky blue and light red bounded lines illustrate the standard deviation among the subjects. The X-axis shows the number of training epochs. The Y-axis represents the prediction accuracy. The accuracy of the previous method is depicted in Figure 3a. The results of the proposed methods are shown in Figure 3b.

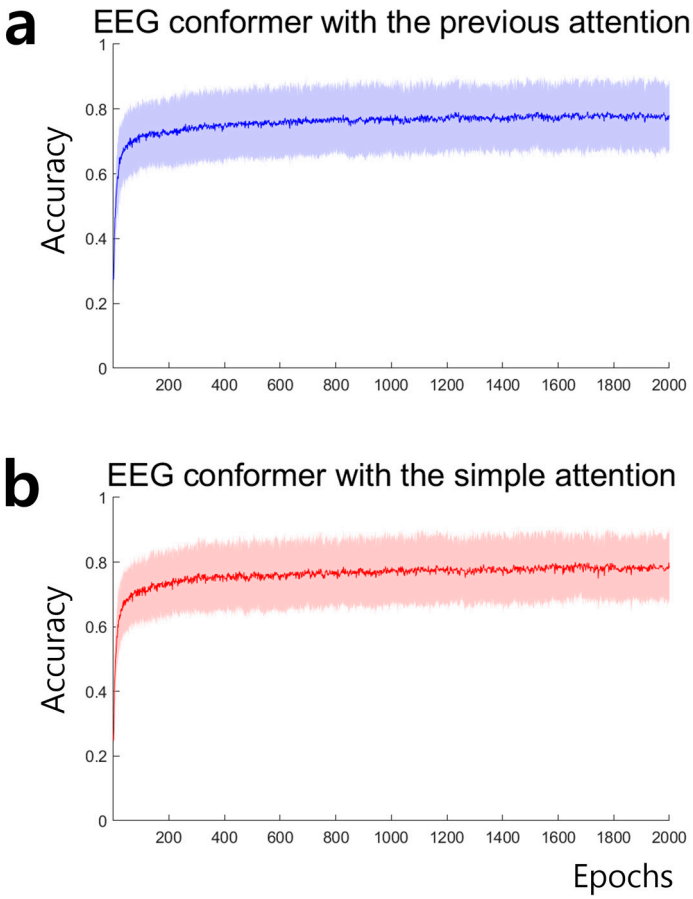


Figure 3. Accuracy transition on dataset I. The *X*-axis indicates the number of training epochs, and the *Y*-axis represents prediction accuracy. (**a**) The results from the EEG conformer with the conventional attention mechanism. The blue line represents the average accuracy across subjects, with the sky-blue shaded area showing the standard deviation. (**b**) The results from the EEG conformer with the simplified attention mechanism. The red line represents the average accuracy among subjects, with the light red shaded area indicating the standard deviation among subjects.

As shown in Tables 1 and 2, the accuracy across the subjects was similar between the previous and simple attention methods. Tables 1 and 2 depict the average accuracy of all the epochs for each subject on datasets 1 and 2, respectively. PrevConformer means the EEG conformer with previous attention. SimpleConformer implies the EEG conformer with simple attention. The SimpleConformer shows similar accuracy to the PrevConformer, despite its simple structure. Tables 1 and 2 also represent the accuracy of state-of-the-art models [26], such as FBCSP [35], ConvNet [36], EEGNet [37], and DRDA [38]. The results show that the transformer model that is based on attention mechanisms outperforms the other machine learning algorithms.

Table 1. Prediction accuracy of dataset I.

Model	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average	Kappa
ConvConformer	82.29	60.69	90.31	71.02	72.57	60.68	81.94	83.55	79.30	75.82	0.6776
SimpleConformer	81.11	60.96	90.12	73.87	73.06	61.05	82.58	85.03	77.41	76.13	0.6817
FBCSP	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75	0.5700
ConvNet	76.39	55.21	89.24	74.65	56.94	54.17	92.71	77.08	76.39	72.53	0.6337
EEGNet	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	86.81	74.50	0.6600
DRDA	83.19	55.14	87.43	75.28	62.29	57.15	86.18	83.61	82.00	74.75	0.6633

Table 2. Prediction accuracy of dataset II.

Model	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average	Kappa
ConvConformer	72.80	67.81	80.07	94.76	93.98	82.80	89.56	92.53	88.17	84.72	0.6944
SimpleConformer	71.79	66.18	79.49	96.63	93.91	82.40	89.52	92.66	88.98	84.62	0.6924
FBCSP	70.00	60.36	60.94	97.50	93.12	80.63	78.13	92.50	86.88	80.00	0.6000
ConvNet	76.56	50.00	51.56	96.88	93.13	85.31	83.75	91.56	85.62	79.37	0.5874
EEGNet	75.94	57.64	58.43	98.13	81.25	88.75	84.06	93.44	89.69	80.48	0.6096
DRDA	81.37	62.86	63.63	95.94	93.56	88.19	85.00	95.25	90.00	83.98	0.6796

The prediction accuracies of the conventional and proposed methods were not statistically different in either dataset I or II. For dataset I, the ConvConformer and the SimpleConformer accuracies were 75.82 $\pm$ 10.33 and 76.13 $\pm$ 10.09, respectively. For dataset II, the ConvConformer and the SimpleConformer accuracies were 84.72 $\pm$ 9.59 and 84.62 $\pm$ 10.45, respectively. The p-values of the paired-sample t-tests for datasets I and II were 0.513 and 0.770, respectively.

5. Discussion

This study proposed a simplified query-only attention mechanism for encoder-based transformer models. The primary goal was to improve the model's efficiency and interpretability by reducing the complexity of the attention mechanism, without compromising performance. Our experimental results indicate that the query-only attention mechanism can achieve a performance that is comparable to the traditional QKV attention mechanism, which uses separate query, key, and value vectors.

5.1. Model Simplification and Its Implications

The simplification of model architectures has often driven advancements in technology. Using only the attention mechanism, the original transformer achieved a superior performance by eliminating the recurrence and convolution [1]. Similarly, the performance of AlphaGo, the first algorithm to win over humans in the game of Go, was improved by simplifying the architecture. AlphaGo employed separate policy and value networks [39]. After AlphaGo's improvement, AlphaGo Zero, AlphaZero, and MuZero utilized a unified network for policy and value [40–42].

The proposed query-only attention mechanism could simplify the architecture by eliminating the redundant K and V vectors while achieving a performance that is comparable to the traditional model. As a result, the number of linear models to generate Q, K, and V vectors could be reduced by one-third. It implies that the linear model's training time and memory space also be decreased by one-third.

The relatively complex structure of conventional attention mechanisms can make it challenging to understand the underlying principles of the algorithm. The meanings and roles of the Q, K, and V vectors differ between the encoder and decoder, which can further complicate intuitive understanding. Particularly in the case of encoder-based transformers, the Q, K, and V vectors are generated through the same process from the same input, making it unclear what each of these vectors signifies and why different values are necessary. In contrast, the proposed method simplifies this process by replacing the ambiguous K and V vectors with Q, as illustrated in Figure 2, potentially making the algorithm more straightforward to comprehend. The Q vector can be considered as a transformed input. Therefore, the proposed method clarifies that attention values are calculated based on the relationships within the input itself. In conclusion, the proposed attention mechanism enhances interpretability by replacing the unnecessary and ambiguous K and V values with the interpretable Q value. This enhanced intuitiveness could facilitate the development of new algorithms.

5.2. Limitation

While the experimental results showed that the proposed attention was comparable to the traditional attention, it should be noted that the experiments were evaluated only with the EEG conformer and EEG datasets. The temporal relationships in EEG signals are relatively local. Therefore, additional experiments with other types of data and models are required in future work. Moreover, query-only attention cannot be applied to encoder–decoder- or decoder-based models; it can only be applied to encoder-based models.

6. Conclusions

In this paper, we proposed the simplified query-only attention mechanism to improve the efficiency and interpretability of encoder-based transformer models. Our method, which utilizes only the Q vector for attention calculation, significantly reduces the model's complexity while maintaining its performance. The experimental results on an EEG conformer model demonstrate that the proposed query-only attention mechanism performs similarly to the original QKV attention mechanism, while simplifying the model's architecture. These findings suggest that query-only attention offers a promising approach for developing more interpretable transformer-based models. Future work could involve applying our proposed method to a broader range of transformer-based models and comparing its performance to other attention mechanisms. The query-only attention mechanism could also be used to develop new transformer models.

Author Contributions: Conceptualization, H.-g.Y. and K.-m.A.; methodology, H.-g.Y.; software, H.-g.Y.; validation, H.-g.Y.; writing—original draft preparation, H.-g.Y. and K.-m.A.; writing—review and editing, H.-g.Y. and K.-m.A.; supervision, H.-g.Y. and K.-m.A.; funding acquisition, H.-g.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2017R1A6A1A03015496).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used open EEG data for our evaluation. The data are shared at http://www.bbci.de/competition/iv (accessed on 20 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
2. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 20 September 2024).
3. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
5. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
6. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
7. Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* **2024**, arXiv:2403.05530.
8. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *arXiv* **2023**, arXiv:2306.13549.
9. OpenAI. Hello GPT-4o. Available online: https://openai.com/index/hello-gpt-4o (accessed on 20 September 2024).
10. Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning deep transformer models for machine translation. *arXiv* **2019**, arXiv:1906.01787.

11. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**, arXiv:1910.13461.
12. Conneau, A.; Lample, G. Cross-lingual language model pretraining. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
13. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
14. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Syst. Appl.* **2023**, *241*, 122666. [CrossRef]
15. Liu, Y.; Lapata, M. Text summarization with pretrained encoders. *arXiv* **2019**, arXiv:1908.08345.
16. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132. [CrossRef]
17. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
18. Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
19. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
21. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
22. Peng, L.; Zhu, C.; Bian, L. U-shape transformer for underwater image enhancement. *IEEE Trans. Image Process.* **2023**, *32*, 3066–3079. [CrossRef]
23. Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; Yu, Y. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Trans. Image Process.* **2023**, *32*, 4036–4045. [CrossRef]
24. Kim, S.; Gholami, A.; Shaw, A.; Lee, N.; Mangalam, K.; Malik, J.; Mahoney, M.W.; Keutzer, K. Squeezeformer: An efficient transformer for automatic speech recognition. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9361–9373.
25. Li, Y.; Lai, L.; Shangguan, Y.; Iandola, F.N.; Ni, Z.; Chang, E.; Shi, Y.; Chandra, V. Folding Attention: Memory and Power Optimization for On-Device Transformer-based Streaming Speech Recognition. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 11901–11905.
26. Song, Y.; Zheng, Q.; Liu, B.; Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *31*, 710–719. [CrossRef] [PubMed]
27. Abibullaev, B.; Keutayeva, A.; Zollanvari, A. Deep Learning in EEG-Based BCIs: A Comprehensive Review of Transformer Models, Advantages, Challenges, and Applications. *IEEE Access* **2023**, *11*, 127271–127301. [CrossRef]
28. Jiang, W.-B.; Zhao, L.-M.; Lu, B.-L. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.
29. Chen, W.; Wang, C.; Xu, K.; Yuan, Y.; Bai, Y.; Zhang, D. D-FaST: Cognitive Signal Decoding with Disentangled Frequency-Spatial-Temporal Attention. *IEEE Trans. Cogn. Dev. Syst.* **2024**, *16*, 1476–1493. [CrossRef]
30. Tangermann, M.; Müller, K.-R.; Aertsen, A.; Birbaumer, N.; Braun, C.; Brunner, C.; Leeb, R.; Mehring, C.; Miller, K.J.; Müller-Putz, G.R. Review of the BCI competition IV. *Front. Neurosci.* **2012**, *6*, 55. [CrossRef]
31. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.
32. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4195–4205.
33. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19175–19186.
34. Lu, M.Y.; Chen, B.; Williamson, D.F.; Chen, R.J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L.P.; Gerber, G. A visual-language foundation model for computational pathology. *Nat. Med.* **2024**, *30*, 863–874. [CrossRef]
35. Ang, K.K.; Chin, Z.Y.; Wang, C.; Guan, C.; Zhang, H. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Front. Neurosci.* **2012**, *6*, 21002. [CrossRef] [PubMed]
36. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef]
37. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef]
38. Zhao, H.; Zheng, Q.; Ma, K.; Li, H.; Zheng, Y. Deep representation-based domain adaptation for nonstationary EEG classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 535–545. [CrossRef]

39. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
40. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef] [PubMed]
41. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **2018**, *362*, 1140–1144. [CrossRef]
42. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T. Mastering atari, go, chess and shogi by planning with a learned model. *Nature* **2020**, *588*, 604–609. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.