Hypercomplex Transformer: Novel Attention Mechanism

Maxim Gordeev[1,2], **Alexander Zuev**[1], **Mikhail Bakulin**[1], **Andrey Latyshev**, **Dmitry Kozlov**, **Yiwu Yao**[1], **Anastasia Voronova**

[1]Huawei, [2]Nizhny Novgorod State Technical University n.a. R.E. Alekseev,

Correspondence: gordeev.maxim@huawei.com

Abstract

Self-attention mechanisms have become foundational across modern deep learning architectures. Recent efforts focus on improving their efficiency, particularly for signal processing tasks. The existing approaches employ complex-valued representations for inputs and weights and achieve higher accuracy at the cost of increased model size and inference latency. Dual-numbered algebra offers a promising alternative that allows efficient multiplication and faster inference with the same representational capacity. Inspired by previous studies in the field of hypercomplex neural networks, we introduce a generalized hypercomplex attention block and integrate it into Transformer-based models for EEG classification. Our experiments include adaptation of the hypercomplex models, so that the number of parameters is equal to that of their real-valued counterparts. Across all scenarios, the dual- and complex-numbered models consistently outperform the real ones, demonstrating superior accuracy. This work presents hypercomplex attention as a competitive and computationally efficient strategy with potential value to solve multiple NLP tasks.

1 Introduction

Complex numbers inevitably appear in signal processing and time series tasks, since Fourier Transform (FFT) is the most common way of operating with signal sequences. This type of numbers allows to describe the full picture of the signal, which is comprised of both time and frequency representation. The nature of input values encourages researchers to use complex-valued neural networks in speech enhancement [Zhao et al., 2021b; Hu et al., 2020], seismic interpretation [Dramsch et al., 2021], radar velocity estimation [Cho et al., 2021] and MusicNet [Yang et al., 2020] tasks.

Attention is an ultimate mechanism that allows neural networks to focus on any parts of the input, emphasizing its specific ones that are the most relevant to the task at hand. Nowadays, multiple works concentrate their efforts on improving the efficiency of attention. There are several works dedicated to extension of transformers to the complex domain that obtain promising results in the field of sequence processing [Yang et al., 2020; Tay et al., 2020; Zhang et al., 2021; Li et al., 2023]. Developing of complex-valued transformers is crucial because usually speech, signal and audio data are naturally complex-valued through Fourier Transform. The main reason for performance improvement provided by transition to complex numbers is that we can capture more information from the input data, including the phase component.

Despite all the advantages, complex-valued neural networks are more than 4 times slower compared to the corresponding real-valued models with the same architecture, due to the increased computational complexity required by complex-valued arithmetic operations. We define a complex linear operator as a matrix multiplication of a complex multi-dimensional tensor $c = x + iy$ and a complex weight matrix $W = A+iB$ with an additional complex-valued bias $z = a + ib$, which can be expressed as the following:

$$cW + z = (x + iy)(A + iB) + (a + ib) = \\ = (xA - yB + a) + i(yA + xB + b)$$

To reduce the computational overhead and save the two-component nature of complex numbers, we turn to dual numbers. They are a special kind of numbers, which can be written as $z = x + \varepsilon y$, where $x, y \in \mathbb{R}$, and ε is a nilpotent element, which satisfies the conditions $\varepsilon^2 = 0$ and $\varepsilon \neq 0$. A dual linear layer computes a product of a dual vector $d = x+\varepsilon y$ and a dual weight matrix $W = A+\varepsilon B$ with an additional bias $k = a + \varepsilon b$ as follows:

$$dW + k = (xA + a) + \varepsilon(yA + xB + b)$$

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1845–1851

The dual-valued multiplication requires fewer real-valued multiplications than the complex one. So, replacement of complex operators with the corresponding dual ones theoretically provides a 25% inference time improvement. The definition and applications of dual convolutional neural networks were presented in [Kozlov et al., 2022; Pavlov et al., 2023]. In this research, we focus on second-order hypercomplex numbers and introduce a novel self-attention block, referred to as hypercomplex attention. Our proposed block processes inputs with dual or complex values.

In the original attention mechanism [Vaswani et al., 2023] the SoftMax function is used to normalize the output probabilities and to emphasize the highest attention scores, thus stimulating the model to focus on the most relevant elements of the input sequence. Here, we study some extensions of SoftMax to the dual and complex domains with similar properties and find out that Component-wise SoftMax show good results.

2 Related work

2.1 Complex-valued neural networks

Multiple researches show promising results of using complex-valued models, such as faster learning [Arjovsky et al., 2015] and larger representational capacity [Nitta, 2003], comparing with the real-valued networks. Most of the existing models and real-valued layers were extended to the scope of complex-valued numbers. Basic operators for convolutional, fully-connected, and LSTM complex-valued architectures are defined in [Trabelsi et al., 2018]. The implementation of complex-valued transformers is mentioned in [Wang et al., 2020], but the article is focused on comparing ways of encoding text into complex-valued tokens. In [Yang et al., 2020], the authors offer a complex-valued model for the automatic music transcription task.

Complex numbers and quaternions are also used for high-level compression of deep learning models ($\times 2$ and $\times 4$) by exploiting the matrix representation of these hypercomplex spaces, achieving comparable performance for NLP [Zhang et al., 2021; Tay et al., 2020].

2.2 Dual-valued neural networks

The idea to use dual numbers in deep learning was introduced in [Okawa and Nitta, 2021]. The building blocks needed to construct dual-valued convolutional neural networks were presented in [Kozlov et al., 2022]. It includes the definitions of Linear, Convolution, Average Pooling, ReLU layers in the dual domain. In addition, [Kozlov et al., 2022] contains an algorithm for dual-valued Batch Normalization. The authors of [Kozlov et al., 2023] selected a subclass of dual-valued operators, which satisfy the equivalent of the Cauchy-Riemann equations for the dual domain.

To the best of our knowledge, we are the first researchers who propose to use dual numbers in the attention layer.

3 Definition of hypercomplex attention

3.1 Hypercomplex input

In all of our experiments we use FFT before the hypercomplex attention layers. FFT converts the input into a complex spectrogram, which has real and imaginary parts. Dual numbers, just like complex ones, are two-component. This property makes it possible to treat the spectrogram as dual-valued data and feed them into the dual-valued attention.

We use an additional dimension to represent the hypercomplex input, which is stored as real-valued tensors:

$$Z = X + \tau Y \in \mathbb{H}^{b \times s \times d} => \\ Z = [X, Y] \in \mathbb{R}^{2 \times b \times s \times d},$$

where the first dimension is used to denote the real or the imaginary part, $\tau = \varepsilon$ or $\tau = i$ correspondingly, b is the batch size, s is the length of a token sequence, d is the dimension of a token vector, $\mathbb{H}$ is the notation of Hypercomplex numbers.

3.2 Hypercomplex attention block

Creating the hypercomplex attention mechanism requires two steps: developing a method for representing matrix multiplication of weights and to design a hypercomplex polarization function. The SoftMax is used as a polarization function in the real-valued attention, but we show below in this section that it is not suitable for the hypercomplex algebra.

Previous studies have identified the approach to forming the complex-valued attention layer with complex weights [Zhao and Ma, 2023].

We define the Hypercomplex attention for a model with complex-valued or dual-valued weights

$W_{\{Q,K,V\}} = W_{\{Q,K,V\}_r} + \tau W_{\{Q,K,V\}_d} \in \mathbb{H}$ as:

$$f[QK^T]V = f[((X + \tau Y)(W_{Q_r} + \tau W_{Q_d}))$$
$$((X + \tau Y)(W_{Kr} + \tau W_{Kd}))^T]$$
$$((X + \tau Y)(W_{Vr} + \tau W_{Vd}))$$

Figure 1 presents the hypercomplex attention mechanism, implemented with hypercomplex weight matrix.

One of the essential elements of Transformer architecture is SoftMax operator. It serves the purpose of emphasizing the closest matches between the rows of Query and Key matrices. In the domain of real numbers the SoftMax function is defined as:

$$(SoftMax(x))_k = \frac{e^{x_k}}{\sum_{j=1}^{n} e^{x_j}},\ k \in \{1, 2, ..., n\} \tag{1}$$

However, this definition is hardly scalable beyond the field of real numbers. In the case of complex numbers it is not well-defined because, unlike for the real-valued domain, the sum of exponents in the denominator can be too close to zero. For example, $e^{i\pi} + e^0 = -1 + 1 = 0$. For the domain of dual numbers this problem does not exist. However, there is a different issue with definition (1). The exponent of a given number $x + \varepsilon y$ can be derived using the Taylor series formula:

$$e^{x+\varepsilon y} = e^x \cdot e^{\varepsilon y} = e^x \cdot \sum_{j=0}^{\infty} \frac{(\varepsilon y)^j}{j!} = \tag{2}$$
$$= e^x \cdot (1 + \varepsilon y) = e^x + \varepsilon y e^x$$

One can see that the real part of the sum of exponents can never be equal to zero, and this is enough to be able to divide over it:

$$\frac{x_1 + \varepsilon y_1}{x_2 + \varepsilon y_2} = \frac{(x_1 + \varepsilon y_1)(x_2 - \varepsilon y_2)}{(x_2 + \varepsilon y_2)(x_2 - \varepsilon y_2)} =$$
$$= \frac{x_1}{x_2} + \varepsilon \frac{x_2 y_1 - x_1 y_2}{x_2^2},\ \ x_2 \neq 0$$

Therefore, (1) is well-defined for dual numbers. However, from (2) it is clear that this definition only emphasizes the real component of the input vector, effectively making the imaginary part negligible. At the same time, that part is responsible for storing the information about signal frequency, which is very important. That is why in this paper we consider other approaches to the definition of the hypercomplex polarization function. We look for possible ways of highlighting important text parts that employ both components of a hypercomplex number. As a result, we suggest the following methods:

- $ScaleMax = \frac{Re(X) - min(Re(X))}{max(Re(X)) - min(Re(X))} + \tau \frac{Im(X) - min(Im(X))}{max(Im(X)) - min(Im(X))}$, where $\tau = i$ or $\tau = \varepsilon$.
- $Component\text{-}wise\ SoftMax = SoftMax(Re(X)) + \tau SoftMax(Im(X))$, where $\tau = i$ or $\tau = \varepsilon$.
- $NormMax = \frac{X^2}{\sum X^2}$, where X is complex or dual norm.

3.3 Hypercomplex Conformer architecture

Convolutional neural networks or transformers are typical solutions to a task of decoding electroencephalograph (EEG) data. However, as the authors of [Song et al., 2023] claim, with this approach the model is able to encapsulate either local or global features but not both, thereby missing some dependencies. To improve this, a Convolutional Transformer architecture called the EEG Conformer was proposed. This architecture is capable of performing efficient classification of EEG data. The model takes EEG signals as input. The result is the likelihood of belonging of the signal to each of the EEG categories. We plug the hypercomplex blocks (Figure 2) in place of their real-valued equivalents into the original Conformer architecture, because they are able to capture more information from the signal data and learn temporal-spatial features more efficiently. The most suitable and natural representation of signal data for complex and dual networks is its representation in the amplitude-frequency domain, which is easily obtained by FFT.

In order to evaluate the effectiveness of Conformer model with hypercomplex attention, two different public EEG datasets were used: BCI competition IV 2a [Brunner et al., 2008] and BCI competition IV 2b [Leeb et al., 2008].

We trained Conformer with hypercomplex attention using the same learning procedure as for the original solution [Song et al., 2023]. An important change is to halve the number of encoder blocks so our model has approximately the same number of parameters as the real-valued one (Table 1). It is worth noting that the inference time of these models also turned out to be almost the same. This similarity in inference time is explained by several factors, which do not allow to observe the maximum theoretical speedup in practice. The limit of 25% theoretical inference speedup is based on replacing complex-number operations with dual-number arithmetic. While both formats contain

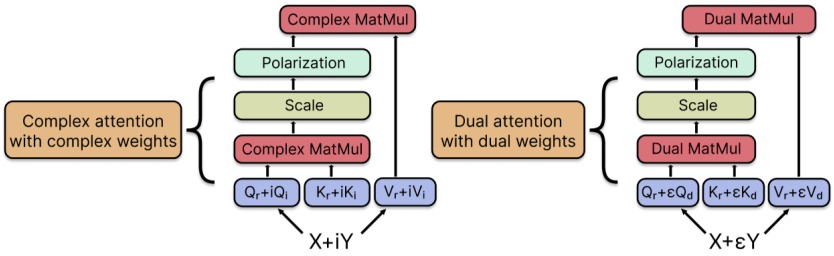

Figure 1: Hypercomplex attention

two components, dual numbers are algebraically simpler. For example, complex multiplication typically involves 4 real-valued multiplications (and 2 real-valued summations/subtractions), whereas dual multiplication is computationally equivalent to only 3 multiplications (and 1 summation). In attention and SoftMax mechanisms, this simplification translates to approximately 25% fewer arithmetic operations per linear layer. However, the actual gain depends on how much the linear operators take compared to the whole computation graph. In Conformer model the linear layers are not the only consumers of computational power, that is why the total gain is less than 25%. Besides, in PyTorch dual-valued tensors are treated as pairs of regular tensors, and their operations are executed using standard vectorized routines. To unlock the full speedup, a dedicated low-level kernel optimized for dual-numbered operations should be implemented. This kernel would directly encode dual arithmetic—such as component-wise summation, matrix multiplication, and SoftMax—without relying on high-level abstractions. By integrating this kernel it will be possible to achieve the full theoretical efficiency in practice.

Table 4 demonstrates the results of experiments with different polarization types for models with hypercomplex attention. The best accuracy value can be achieved using the Component-wise Softmax function on the BCI competition IV 2a and the BCI competition IV 2b datasets. In EEG signal processing, the real part typically encodes amplitude, while the imaginary part captures frequency characteristics, such as those derived via Fourier transform. Neglecting the imaginary part—as in naive extension of the traditional SoftMax to dual numbers — results in loss of frequency information, reducing the model's ability to distinguish patterns like alpha or beta rhythms. Component-wise SoftMax addresses this by applying SoftMax independently to the real and imaginary parts. The real part yields amplitude-based weights, while the imaginary part captures frequency-related contributions. This preserves the hypercomplex structure and minimizes loss of information. In contrast, ScaleMax normalizes by min-max range without exponential weighting, losing probabilistic sharpness. NormMax reduces to squared norm, which emphasizes energy but discards distributional detail. In Transformer architectures, attention weights are computed via SoftMax. The component-wise variant enables simultaneous modeling of amplitude and frequency in multidimensional space, improving the ability to capture long-range dependencies in time-series data like EEG. Theoretically, this relates to the entropy:

$$H = -\sum_k p_k \log p_k$$

the standard SoftMax yields entropy over the real part, but incorporating the imaginary component reduces effective entropy, allowing sharper focus on frequency-dependent features-explaining the observed accuracy gains.

In Table 2 we compare our proposed model with multiple neural networks of different architectures: convolution-based (ConvNet [Schirrmeister, 2017], EEGNet [Lawhern et al., 2018]), FBCNet [Mane, 2021] spatial data filtering, DRDA [Zhao et al., 2021a] and others.

We can observe that the average accuracy of our Conformer model with the complex-valued attention is 1.2% higher than that of Conformer with the real-valued one and 5% higher than the average accuracy of the convolution-based models. Moreover, the model with dual-valued attention achieves the average accuracy, which is by 0.2% even higher.

The average accuracy of Conformer with the complex-valued attention on BCI Competition IV Dataset 2b (Table 3) is 2.1% higher than that of Conformer with the real-valued one; our model exceeds other methods in accuracy for almost all

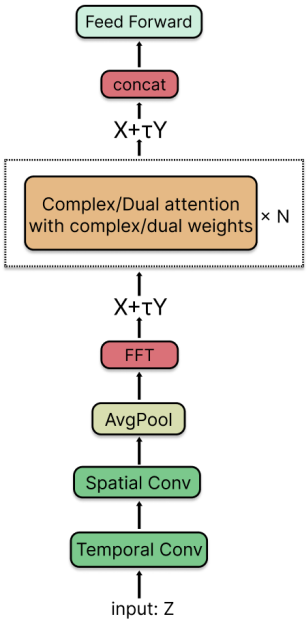

Figure 2: The architecture of Conformer with hypercomplex-valued attention.

the subjects. The model with the dual-valued attention also surpasses all the others, achieving a new state-of-the-art accuracy value of 87%.

4 Conclusion

In this study, we generalize the Transformer architecture to the algebra of dual and complex numbers. Specifically, we suggest and implement a new hypercompex attention mechanism. Then we integrate it into Transformer model, that we designed for the EEG classification task. We compare accuracy of the models with the hypercomplex attention against the value of the same metric shown by the corresponding models with the traditional real-valued attention with the same number of parameters.

Then, we built the novel type of hypercomplex attention mechanism into Conformer - a specific Transformer-based model architecture. Our experiments show that on BCI_competition_IV2a dataset our Conformer model overcomes the state-of-the-art real-valued network by 1.2% in accuracy metric for complex algebra and 1.4% for dual algebra. On BCI_competition_IV2b dataset the model with the complex-valued Attention surpasses the real-valued one by 2.1% in accuracy. At the same time, Conformer with the dual-valued Attention reaches the state-of-the-art metrics value on this dataset.

Novel hypercomplex polarization functions are designed, revealing that for Conformer models with hypercomplex attention the best accuracy value can be achieved using the Component-wise Softmax function.

We conclude that extending self-attention mechanism to the space of hypercomplex numbers within the same model's architecture proves its efficiency, since it leads to better metrics than the original network.

5 Limitations

The discussion of hypercomplex transformers in this study is limited to the Conformer model applied to the EEG signal classification task. Although promising preliminary results were obtained, the proposed method has not yet been tested across diverse domains or alternative architectures.

As part of the further research, we intend to extend the developed hypercomplex attention mechanism for solving other tasks by integrating it into a wide variety of models, including modern LLMs. We expect to achieve improvements in quality metrics, such as accuracy, perplexity, and others.

References

K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang. 2012. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers Neurosci.".*

Martin Arjovsky, Amar Shah, and Yoshua Bengio. 2015. Unitary Evolution Recurrent Neural Networks. 48.

C. Brunner, R. Leeb, G. R. Müller-Putz, A. Schlögl, and G. Pfurtscheller. 2008. BCI Competition 2008 – Graz data set A.

Hyun Woong Cho, Sungdo Choi, Young Rae Cho, and Jongseok Kim. 2021. Complex-Valued Channel Attention and Application in Ego-Velocity Estimation with Automotive Radar.

Jesper Sören Dramsch, Mikael Lüthje, and Anders Nymark Christensen. 2021. Complex-valued neural networks for machine learning on non-stationary physical data. *Computers and Geosciences*, 146.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *Preprint*, arXiv:2008.00264.

Dmitry Kozlov, Mikhail Bakulin, Stanislav Pavlov, Aleksandr Zuev, Mariya Krylova, and Igor Kharchikov. 2023. Learning Properties of Holomorphic Neural Networks of Dual Variables. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2023-June.

Dmitry Kozlov, Stanislav Pavlov, Alexander Zuev, Mikhail Bakulin, Mariya Krylova, and Igor Kharchikov. 2022. Dual-valued Neural Networks. *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. 2018. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *J. Neural Eng.".*.

R. Leeb, C. Brunner, G. R. Müller-Putz, A. Schlögl, and G. Pfurtscheller. 2008. BCI Competition 2008 – Graz data set B.

Qiuchi Li, Benyou Wang, Yudong Zhu, Christina Lioma, and Qun Liu. 2023. Adapting Pre-trained Language Models for Quantum Natural Language Processing.

R. Mane. 2021. Fbcnet: A multi-view convolutional neural network for brain–computer interface. *arXiv preprint arXiv:2104.01233*.

Tohru Nitta. 2003. The computational power of complex-valued neuron. In *Joint International Conference ICANN/ICONIP*, pages pp. 993–1000.

Yuto Okawa and Tohru Nitta. 2021. Learning Properties of Feedforward Neural Networks Using Dual Numbers. In *Proceedings, APSIPA Annual Summit and Conference*, pages 187–192.

Stanislav Pavlov, Dmitry Kozlov, Mikhail Bakulin, Aleksandr Zuev, Andrey Latyshev, and Alexander Beliaev. 2023. Generalization of neural networks on second-order hypercomplex numbers. *Mathematics*, 11(18).

S. Sakhavi, C. Guan, and S. Yan. 2018. Learning temporal information for brain–computer interface using convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.".*.

R. T. Schirrmeister. 2017. Deep learning with convolutional neural networks for eeg decoding and visualization: Convolutional neural networks in eeg analysis. *Hum. Brain Mapping.".*.

Yonghao Song, Bingchuan Liu, and Xiaorong Gao. 2023. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING*.

Yi Tay, Aston Zhang, Anh Tuan Luu, Jinfeng Rao, Shuai Zhang, Shuohang Wang, Jie Fu, and Siu Cheung Hui. 2020. Lightweight and efficient neural natural language processing with quaternion networks. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1494–1503.

Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, and Sandeep Subramanian. 2018. Deep complex networks. *ICLR arXiv:1705.09792*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. 2020. Encoding word order in complex embeddings. *ICLR arXiv:1912.12333v2*.

Muqiao Yang, Martin Q. Ma, Dongyu Li, Yao-Hung Hubert Tsai, and Ruslan Salakhutdinov. 2020. Complex Transformer: A Framework for Modeling Complex-Valued Sequence. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2020-May, pages 4232–4236.

Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. 2021. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with 1/n parameters. *Preprint*, arXiv:2102.08597.

H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng. 2021a. Deep representationbased domain adaptation for nonstationary eeg classification. *IEEE Trans. Neural Netw. Learn. Syst.".*.

Shengkui Zhao and Bin Ma. 2023. D2former: A fully complex dual-path dual-decoder conformer network using joint complex masking and complex spectral mapping for monaural speech enhancement. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Shengkui Zhao, Trung Hieu Nguyen, and Bin Ma. 2021b. Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6648–6652.

A Appendix with attendant tables

Model	Size of model, MB	Inference time, s	Batch size
Real	790	0.09	72
Complex	793	0.11	72
Dual	793	0.1	72

Table 1: Comparison of Conformer models by the number of parameters and inference time on CPU

Method	S01	S02	S03	S04	S05	S06	S07	S08	S09	Avg
FBCSP [Ang et al., 2012]	76.0	56.5	81.3	61.0	55.0	45.3	82.8	81.3	70.8	67.8
ConvNet [Schirrmeister, 2017]	76.4	55.2	89.2	74.7	56.9	54.2	92.7	77.1	76.4	72.5
EEGNet [Lawhern et al., 2018]	85.8	61.5	88.5	67.0	55.9	52.1	89.6	83.3	86.8	74.5
C2CM [Sakhavi et al., 2018]	87.5	**65.3**	90.3	66.7	62.5	45.5	89.6	83.3	79.5	74.5
FBCNet [Mane, 2021]	85.4	60.4	90.6	76.4	**74.3**	53.8	84.4	79.5	80.9	76.2
DRDA [Zhao et al., 2021a]	83.2	55.1	87.4	75.3	62.3	57.2	86.2	83.6	82.0	74.7
Conformer [Song et al., 2023]	88.2	61.5	93.4	78.1	52.1	65.3	92.4	88.2	**88.9**	78.7
Complex Conformer	87.8	59.4	94.4	77.4	68.0	62.8	92.8	87.5	86.9	79.7
Dual Conformer	**88.8**	63.2	**94.8**	**78.5**	61.1	**65.7**	**95.4**	**88.2**	88.8	80.1

Table 2: Comparison With State-of-the-Art Methods on BCI Competition IV Dataset 2a

Method	S01	S02	S03	S04	S05	S06	S07	S08	S09	Avg
FBCSP [Ang et al., 2012]	70.0	60.4	60.9	97.5	93.1	80.6	78.1	92.5	86.6	80.0
ConvNet [Schirrmeister, 2017]	76.6	50.0	51.6	96.9	93.1	85.3	83.8	91.6	85.6	79.4
EEGNet [Lawhern et al., 2018]	75.9	57.6	58.4	98.1	81.3	88.8	84.1	93.4	89.7	80.5
DRDA [Zhao et al., 2021a]	81.4	62.9	63.6	95.9	93.6	88.2	85.0	**95.3**	90.0	83.9
Conformer [Song et al., 2023]	82.5	65.7	63.8	98.4	86.6	90.3	87.8	94.4	92.2	84.6
Complex Conformer	**83.8**	64.6	**74.0**	98.0	96.5	88.4	89.3	94.0	91.3	86.7
Dual Conformer	83.1	65.7	73.8	98.0	97.2	**90.3**	**91.5**	**95.0**	91.3	**87.0**

Table 3: Comparison With State-of-the-Art Methods on BCI Competition IV Dataset 2b

Dataset	Model	Polarization	Avg Accuracy, %
BCI Competition IV 2a	Complex	ScaleMax	78.9
		Component-wise Softmax	**79.7**
		NormMax	74.4
	Dual	ScaleMax	79.7
		Component-wise Softmax	**80.1**
		NormMax	75.0
BCI Competition IV 2b	Complex	ScaleMax	86.3
		Component-wise Softmax	**86.7**
		NormMax	85.5
	Dual	ScaleMax	86.3
		Component-wise Softmax	**87.0**
		NormMax	86.0

Table 4: Results of experiments with different polarization types

While our main focus was EEG classification, we also evaluated the model on automatic music transcription task using the MusicNet dataset. Our experiments show the applicability of our model in sequential symbolic domains, which share structural parallels with NLP tasks such as sequence labeling and multi-label classification. The results are available in Table 5.

Weights	Real Avg Precision	Dual Avg Precision	Real Time	Dual Time
Real	70.5	72.1	9.6	43.7
Dual	N/A	**73.7**	N/A	**41.4**

Table 5: Comparison of average precision (%) and inference time (ms) of real- and dual-valued Transformer models with real and dual weights on MusicNet dataset