ScatterFormer: Locally-Invariant Scattering Transformer for Patient-Independent Multispectral Detection of Epileptiform Discharges

Ruizhe Zheng^{*1,2,3,4,5,6}**, Jun Li**^{*1,6}**, Yi Wang**$^{\dagger 7}$**, Tian Luo**^{7}**, Yuguo Yu**$^{\ddagger 1,2,3,4,5,6}$

[1] Research Institute of Intelligent and Complex Systems, Fudan University
[2] State Key Laboratory of Medical Neurobiology, Fudan University
[3] MOE Frontiers Center for Brain Science, Fudan University
[4] Institutes of Brain Science, Fudan University
[5] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University
[6] Shanghai Artificial Intelligence Laboratory
[7] Department of Neurology, Children's Hospital of Fudan University
rzzheng@fudan.edu.cn, jun_li@fudan.edu.cn, yiwang@shmu.edu.cn, luotian@fudan.edu.cn, yuyuguo@fudan.edu.cn

Abstract

Patient-independent detection of epileptic activities based on visual spectral representation of continuous EEG (cEEG) has been widely used for diagnosing epilepsy. However, precise detection remains a considerable challenge due to subtle variabilities across subjects, channels and time points. Thus, capturing fine-grained, discriminative features of EEG patterns, which is associated with high-frequency textural information, is yet to be resolved. In this work, we propose Scattering Transformer (ScatterFormer), an invariant scattering transform-based hierarchical Transformer that specifically pays attention to subtle features. In particular, the disentangled frequency-aware attention (FAA) enables the Transformer to capture clinically informative high-frequency components, offering a novel clinical explainability based on visual encoding of multichannel EEG signals. Evaluations on two distinct tasks of epileptiform detection demonstrate the effectiveness our method. Our proposed model achieves median AUCROC and accuracy of 98.14%, 96.39% in patients with Rolandic epilepsy. On a neonatal seizure detection benchmark, it outperforms the state-of-the-art by 9% in terms of average AUCROC.

Introduction

Continuous electroencephalography (cEEG) plays an important role in monitoring and diagnosis of epilepsy. Automatic detection of epileptiform discharges and distinguishing these activities from nonepileptiform abnormalities (Yum and Shvarts 2019; Shi et al. 2020) and normal/benign EEG activities (Li et al. 2020) are of great importance in the biomedical field, because that experts with various levels of diagnostic experience usually report discrepant opinions on the EEG records (Xiang et al. 2015). In particular, patient-independent epileptiform detection aims at identifying seizures within an EEG recording without separate fine-tuning efforts on new data to establish a subject-specific detector. However, pitfalls for correct identification of epileptiform discharges include misreading, misinterpretation and overinterpretation of individual EEG signature due to inherent subjectivity, can lead to misdiagnosis of people who do not have epilepsy (Tatum 2012; Shorvon and Schmidt 2016; Tatum and Shellhaas 2020). Therefore, cost-effective, cross-subject algorithms are urgently in need.

So far, a plethora of deep learning-based algorithms based on visual spectral displays of EEG have been developed (Tatum et al. 2018; Rasheed et al. 2020; Saminu et al. 2021). Still, major concerns on the failure in clinical practice remain present due to limited generalizability of deep learning models on out-of-distribution data when conducting patient-independent predictions (Achilles et al. 2018; Li et al. 2020), which requires precise identification of fine-grained features (Jeong et al. 2021) that is clinically interpretable for electroencephalographers with reasonable sensitivity and specificity. However, traditional learning schemes heavily depend on labor-intensive feature selection with regard to time-frequency subbands and electrodes, as well as often inconsistent preprocessing of raw data, which collectively limit the potential to discover a latent, informative representation of signals. Moreover, manual removal of intra- and inter-subject noises and artefacts could potentially eliminate features crucial for a fine-grained analysis of seizure occurrence.

Seizure dynamics is characterized by high-frequency outbursts of epileptiform abnormalities. Recent advances in attention-based deep neural networks, especially Transformers, have prompted augmented representational learning for EEG classification tasks (Bagchi and Bathula 2022; Siddhad et al. 2022; Tao et al. 2021; Sun, Xie, and Zhou 2021). However, the problem of loss of relevant clinical biomarkers of irregularly altered EEG signals has not been explicitly addressed. Invariant scattering transform is a non-linear transform based on a cascade of wavelet transforms and modulus non-linearity. Invariant scattering convolution has energy preservation nature and eliminates translation and rotation variability algorithmically due to its Lipschitz-continuity (Mallat 2012). Therefore, scattering coefficients are sensi-

[*] These authors contributed equally.

[†] Corresponding Author

[‡] Corresponding Author

tive representation of complex, edge-like patterns (Cotter and Kingsbury 2019; Bruna and Mallat 2013; Mallat 2012). These subtle patterns have been proven to be beneficial for texture classification (Singh and Kingsbury 2017). More importantly, invariant scattering transform can retrieve high frequency components lost due to low-pass filtering by cascaded wavelet decomposition. Despite potential benefits, its application has not been investigated in Transformer learning of epileptiform recognition.

In this work, we aim at capturing subtle discriminative features of EEG data by strategically leveraging invariant scattering transform in Transformer architecture design. We apply a frequency-aware attention (FAA) to capture richer contextual dependencies, which incorporates scattering layers to prevent oversmoothing. The design allows the model to precisely capture fine-grained features using invariant scattering transform in combination with dynamic weighting of feature maps. An end-to-end multispectral diagnostic pipeline is established and rigorously tested in classifying spectra calculated directly from raw EEG records on one private and one public dataset. Furthermore, we present our correlation analysis of scattering transform, and show our theoretical result that scattering transform can improve the model generalizability. Our main contributions are summarized as follows:

- We propose an end-to-end diagnostic pipeline targeting patient-independent epileptic seizure detection. The proposed approach is efficient in capturing fine-grained information in seizure-specific multispectral representations while maintaining time- and frequency-shift invariance, presenting a novel multispectral interpretability without compromising performance.
- We first introduce scattering transform to Transformer, which is accomplished through token embedding and frequency-disentangled attention based on locally-invariant scattering layers, leading to synergy between high- and low-frequency attention in order to obtain local discriminative patterns that is critical for epiletiform recognition without increasing computational load.
- We theoretically present our correlation analysis of wavelet transform, and show that the generalizability in patient-independent setting can be improved via reducing upper bound of Gaussian complexity.
- Extensive experimental comparisons based on cross-validation are evaluated over different epilepsy diagnosis datasets in a cross-subject manner. The results demonstrate that our model outperforms state-of-the-art methods.

Related Work

In this section, we briefly review related work on spectral detection of epileptic EEG activities, frequency-aware Transformers and invariant scattering transform, highlighting the recent advances and gaps.

EEG Learning Representation

(Asif et al. 2020) proposed SeizureNet, a diagnostic framework based on multispectral fusion-based encoding of EEG data using short-time Fourier transform (STFT) of multichannel records and saliency mapping of STFT spectrograms, which is showed to be beneficial for capturing fine-grained information. The model generalizes well on unseen patient EEG epochs. In order to optimize DNN-based diagnostic method in terms of capturing fine-grained features as well as their long-range dependencies, (Jeong et al. 2021) proposed an attention-based model that separately captures global attention and fine-grained information for seizure detection. In particular, inductive bias is introduced through stacking of convolutional blocks, which are sensitive to high-frequency components, before self-attention calculation. (Bagchi and Bathula 2022) introduced convolutional feature expansion to Transformer to model the inter-channel similarities. Different from previous research, we propose a mechanistically feasible design to help mitigate low-frequency bias that could cause performance degradation, and to obtain interpretability that highlights clinically informative attributes.

High-Frequency Components and Attention Mechanism

Transformers are capable of capturing low-frequency components, which are associated with global semantic information. Undesirable low-pass filtering occurs with depth increasing, which could potentially lead to over-smoothing of local textures, thus weakening the modeling capability. Recent works have introduced convolution operators to ViTs to strengthen capability of modeling local dependencies. Moreover, disentanglement of attention for parallel modelling of high- and low- frequencies not only alleviates over-smoothing, but also aggregates richer components that empirically bring performance gain. (Pan, Cai, and Zhuang 2022) proposed HiLo, which separately deals with low and high frequencies in a multi-head self-attention (MHSA) module to achieve disentanglement of feature maps in spectral domain at the same encoder layer without compromising computational efficiency. (Si et al. 2022) proposed Inception Transformer, where high-frequency components are modelled by convolution and max-pooling operations to aggregate features across frequency range. In this work, ScatterFormer aims at more balanced modeling of high- and low-information in a more precise, interpretable manner by introducing frequency-aware operations that decompose and mix tokens in a way that preserves locally invariant high-frequency features.

Preliminaries

Problem Formulation

The EEG activities are recorded as multivariate time series $X = \{X_i\}_{i=1}^{T} \in \mathbb{R}^{T \times C}$, where $X_i = \{x_{i,c}\}_{c \in C} \in \mathbb{R}^{C}$ are recorded signals of multiple channels at the time point i, and X represents an epoch of brain waveforms segmented manually or automatically with appropriate choice of duration T. Domain knowledge is employed to predetermine number of electrodes and EEG montage in order to obtain C channels of signals. Preprocessed EEG epochs are encoded with the proposed saliency-aware multispectral representation. In real-word scenarios, two clinically relevant problems are defined for cross-subject detection of epileptiform discharges based on model P_θ parameterized with θ.

- Given N clinically diagnosed subjects with and without epileptic seizures, n_i EEG epochs and corresponding annotations are available for each individual, which constitutes the training dataset $\{\{X_k, y_k\}_{k=1}^{n_i}\}_{i=1}^{N}$, where $X_k \in \mathbb{R}^{T\times C}$, $y_k \in \{0, 1\}$ is the annotated label. For recently hospitalized M patients who are yet to be diagnosed, m_j EEG epochs are extracted for the j^{th} individual, $1 \leq j \leq M$, the estimate of the label is the probability of seizure occurrence:
$$\{\hat{y}_k^{seizure}\}_{l=1}^{m_j} = P_\theta(y|\{X_l\}_{l=1}^{m_j}, \{\{X_k, y_k\}_{k=1}^{n_i}\}_{i=1}^{N}) \tag{1}$$
- Given N clinically diagnosed subjects with epileptic seizures, n_i EEG epochs and corresponding annotations are available for each individual, which constitutes the training dataset $\{\{X_k, y_k\}_{k=1}^{n_i}\}_{i=1}^{N}$. We aim to predict interictal and ictal activities during continuous EEG monitoring of out-of-domain subjects, who are recently diagnosed with epileptic seizures. For an EEG epoch $X \in \mathbb{R}^{T\times C}$, $y_k \in \{0, 1\}$ recorded at a specific moment, the estimate of the label is the probability of whether it is interictal or ictal:
$$\{\hat{y}_k^{ictal}\}_{l=1}^{m_j} = P_\theta(y|\{X_l\}_{l=1}^{m_j}, \{\{X_k, y_k\}_{k=1}^{n_i}\}_{i=1}^{N}) \tag{2}$$

Vision Transformer

Vision Transformer (ViT) and its variants are highly capable of capturing global contextual information and long-range dependence in attention-based modeling of a plethora of modalities, including EEG spectra (Bagchi and Bathula 2022; Tao et al. 2021), but are not capable enough to learn high-frequency components that are crucial for fine-grained information extraction and classification (Wang et al. 2022; Bai et al. 2022). Recent years have witness significant advance in examination of ViT from spectral domain, which contributes to resolving important gaps such as attention collapse (Wang et al. 2022). Moreover, convolutional neural networks (CNNs) have been re-introduced into Transformer architecture to enhance sensitivity of multi-head self-attention (MHSA) to local features while maintaining reasonable computational load (Wu et al. 2021). In this work, we compute MHSA by cross-covariance attention proposed in (Ali et al. 2021) as follows:

$$\boldsymbol{Attn} = \boldsymbol{V}\mathrm{Softmax}\left(\frac{\hat{\boldsymbol{Q}}^\top\hat{\boldsymbol{K}}}{\tau}\right) \tag{3}$$

where $\boldsymbol{Attn}$ denotes attention maps, $\hat{\boldsymbol{Q}}$, $\hat{\boldsymbol{K}}$ are l_2-normalized query and key embeddings, $\boldsymbol{V}$ is value embedding, τ is a learnable parameter for stabilizing training.

Methodology

Saliency-Aware Multispectral Representation

We propose a novel method to generate a saliency-aware, multi-spectral, multi-channel representation of the EEG records shown in bipolar montage, where the localization of the celebral potential is based on the direction of the waveform between two channels, and a phase reversal is beneficial for easier identification of epileptiform abnormalities (Sazgar and Young 2019). Specifically, continuous wavelet transformation (CWT) to circumvent the problem of non-stationarity.

$$W_\psi[x(a,b)] = \int_{-\infty}^{+\infty} x(t)\bar{\psi}(\frac{t-b}{a})dt \tag{4}$$

where $\psi(\cdot)$ denotes a family of base functions that dilate and contrast with frequency to analyze intricate time-frequency details. Power spectrum $S_i = log(|W_{\psi_i}[x(a,b)]|^2)$ in terms of mother wavelet $\psi_i(\cdot)$. The multispectral representation is formulated as $S = norm(\sum_i S_i)$, where $norm(\cdot)$ denotes l_2 normalization. The spectrogram is stacked with static spectral saliency map SA_1 and fine-grained saliency map SA_2 to incorporate more abundant information that are integral to diagnosis. Differences of Gaussian (DoG), Paul and Morlet are used to compute S.

Invariant Scattering Transformer

We propose frequency-aware attention that employs dual branches of attention calculation and local sensitivity of invariant scattering transform to disentangle high- and low-frequencies.

Invariant Scattering Token Embedding In order to accomplish efficient identification and encoding of information that is typically lost in the down-sampling of high-resolution fine-grained multispectral visual representation, invariant scattering operator, which is Lipschitz-continuous to diffeomorphisms that causes significant perturbations to high-frequency components, is proposed to preserve information without over-smoothing. An invariant scattering transform provides locally translation-invariant multiscale coefficients, which characterize the scaling properties of signals. They are computed by iteratively calculating the modulus of complex wavelet coefficients, which are yielded by convolving input signal with mother wavelet $\psi(\cdot)$ and scaling function $\phi(\cdot)$. A dyadic band-pass filter bank is determined for $j \in \mathbb{Z}$ and rotation $r \in G$, G is a finite rotation group in $\mathbb{R}^2$.

$$\psi_{2^j r}(x) = 2^{2j}\psi(2^j r^{-1}x) \tag{5}$$

The 2-D wavelet transform is done by convolving the input with a mother wavelet dilated by 2^j and rotated by θ:

$$\psi_{j,\theta}(x) = 2^{-j}\psi\left(2^{-j}R_{-\theta}x\right) \tag{6}$$

where R is the rotation matrix, $1 \leq j \leq J$ are indexes of the scale. We define $\Lambda_J^1 := \{(j, r) : 1 \leq j \leq J \text{ and } r \in G\}$. For $p := (\lambda_1, \lambda_2, \cdots, \lambda_m) \in \Lambda_J^m := \Lambda_J^1 \times \Lambda_J^1 \times \cdots \times \Lambda_J^1$ (m times), a scattering propagator $U[p] : L^2(\mathbb{R}^2) \rightarrow L^2(\mathbb{R}^2)$ is defined as

$$U[p]f(x) := U[\lambda_m]\cdots U[\lambda_2]U[\lambda_1]f(x) \ \ \forall f \in L^2(\mathbb{R}^2), \tag{7}$$

where $U[\lambda_k]f(x) := |(\psi_{\lambda_k} * f)(x)|$ for $k = 1, 2, \cdots, m$ and $U[\emptyset]f(x) := f(x)$. We compute m times of convolutions and modulus operators along the path of p of length m, the input f is propagated to the m-th layer:

$$S_J[p]f(x) := (\phi_{2^J} * U[p]f)(x). \tag{8}$$

The resulting m^{th}-order scattering coefficients has energy preservation property. Specifically, We define scattering decomposition of input $f(x)$ as

$$S_J[\Lambda_J^1]f(x) := \{S_J[p]f(x)\}_{p\in\Lambda_J^1} \tag{9}$$

Its norm is $\sum_{p\in\Lambda_J^1} \|S_J[p]f(x)\|^2$. Note that $S_J[\Lambda_J^1]$ is contractive:

$$\|S_J f(x) - S_J f(y)\| \leq \|f(x) - f(y)\| \,. \tag{10}$$

It is further proved that (Mallat 2012)

$$\lim_{m_{\max}\to\infty} \sum_{m=m_{\max}}^{\infty} \|S_J[\Lambda_J^m]x\|^2 = 0 \,. \tag{11}$$

The above result implies that invariant scattering transform is energy-preserving in addition to its stability to diffeomorphisms (Mallat 2012; Singh and Kingsbury 2017). Recent approaches have indicated that hybrid neural networks with scattering layers can achieve competitive performance on a range of benchmarks such as ImageNet and CIFAR-10 (Cotter and Kingsbury 2019; Singh and Kingsbury 2018, 2017) at a faster convergence rate. Unlike CNNs, it has not been strategically introduced to Transformer to improve the effectiveness of the Transformers. Typically, second-order invariant scattering transform with optional learnable channel-wise mixing of maps modeled by conventional convolution operation (Cotter and Kingsbury 2019). However, (Oyallon et al. 2018) noted that first-order scattering alone can effectively preserve discriminative information. In light of this, we adopt a hierarchical design in ScatterFormer by applying multiple invariant scattering layers to incorporate more high-frequency structures.

Frequency-aware Attention (FAA) The tokens are split apart before being separately processed in dual-branch attention module. For high-frequency branch, we propose to use an invariant scattering transform projects feature maps to queries $\boldsymbol{Q}$. Multiple invariant scattering layers located at the beginning of each stage acting as down-sampling and patch-merging module together with those reside within the encoder intuitively preserves high-frequency scattering coefficients required to discriminate between images. Moreover, first-order transform is able to capture important attributes with minimal energy loss (Bruna and Mallat 2013; Oyallon et al. 2018). Thus, higher-order coefficients could offer only marginal gains while significantly increasing memory and computational cost in the high-frequency branch of self-attention. In conformity with reduced resolution of token maps, the projection of keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ uses 3×3 convolution with a stride of 2×2, which further reduces the computational cost. The high-frequency attention is formally expressed as

$$\boldsymbol{Q}_h = \mathrm{BatchNorm}\left(\mathrm{Inv}\left(\boldsymbol{X}_h\right)\right) \tag{12}$$

$$\boldsymbol{K}_h = \mathrm{BatchNorm}\left(\mathrm{Conv}_K\left(\boldsymbol{X}_h\right)\right) \tag{13}$$

$$\boldsymbol{V}_h = \mathrm{BatchNorm}\left(\mathrm{Conv}_V\left(\boldsymbol{X}_h\right)\right) \tag{14}$$

for $\boldsymbol{X}_h \in \mathbb{R}^{\frac{C_i}{2}\times H_i \times W_i}$ at i^th stage, where Inv denotes invariant scattering layer, Conv_K and Conv_V denote convolution, $\mathrm{BatchNorm}$ denotes batch normalization, which is used to stablize the training procedure. Following (Ali et al. 2021), $\boldsymbol{Q}_k$ and $\boldsymbol{K}_h$ are normalized to obtain channel-wise attention score matrix.

$$\boldsymbol{X}_h = \boldsymbol{V}_h \mathrm{Softmax}\left(\boldsymbol{Q}_h^\top \boldsymbol{K}_h\right) \tag{15}$$

Low-frequency attention is calculated in the separate path without Transformer encoder is demonstrated to be inherently sensitive to low-frequency components (Wang et al. 2022). We use 3×3 convolution for token embedding. Linear projection layer for two groups of attention maps is used to mixing the separately learned attention in order to obtain a common representation. The locally-enhanced relative position encoding (LePE) (Dong et al. 2021) is applied on $\boldsymbol{X} \in \mathbb{R}^{\frac{C_i}{2}\times H_i \times W_i}$ before attention calculation and added on the fused attention maps $\boldsymbol{X}_{fused}$:

$$\boldsymbol{X}_{fused} = \mathrm{Linear}(\mathrm{Concat}((X_l), \mathrm{Upsample}(\boldsymbol{X}_h))) \tag{16}$$

$$\boldsymbol{X}_{fused} = \boldsymbol{X}_{fused} + \mathrm{DWConv}(\boldsymbol{X}) \tag{17}$$

where Linear denotes linear projection, $\mathrm{Upsample}$ denotes bilinear upsampling module, DWConv denotes depthwise convolution, Concat denotes concatenation.

Generalization Capacity Analysis

In this section, we theoretically prove that invariant scattering transform can enhance the generalization capacity of a network.

We first introduce the Gaussian complexity,

$$\hat{G}_N(F) = \mathbb{E}[sup_{f\in F} \frac{2}{N} \sum_{i=1}^{N} g_i f(x_i)] \tag{18}$$

where $g_i, i = 1, 2, \cdots, N$ are independent standard normal random variables. Gaussian complexity in Eq. (18) measures the capacity of a functional class F. Gaussian complexity (and the closely-related Rademacher complexity) is often linked with the generalization analysis of deep learning, as its definition does not rely on the number of model parameters.

Then, we link correlation analysis that examines locality of features with the generalization capacity of transformer. As recent work (Naseer et al. 2021) shows that increase locality of features can compensate the high-frequency features in transformers, strong feature correlation is expected to improve the generalization capacity of transformers. The following theorem shows that feature correlation is closely related to an upper bound of Gaussian complexity.

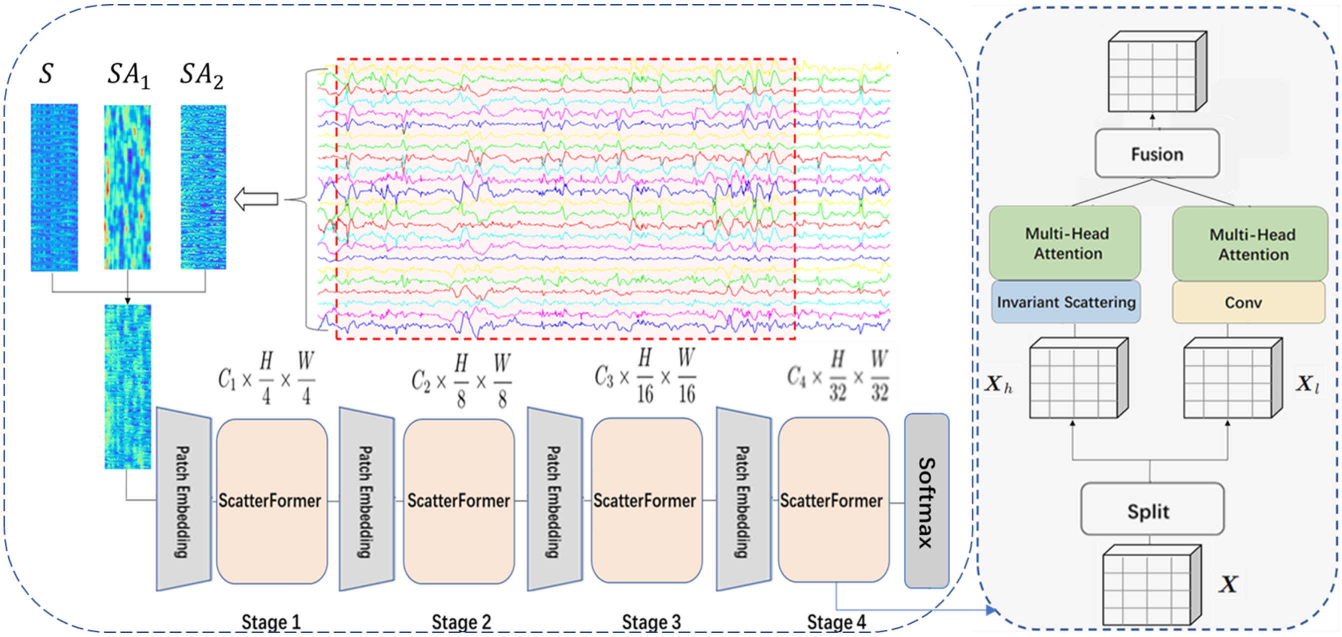


Figure 1. Left: Architecture of ScatterFormer and diagnostic pipeline based on multispectral visual representation of EEG epochs for epileptiform identification. Input features are the raw EEG signals processed by multispectral representation. On top of the hierarchical stacking of ScatterTransformers is a Softmax classifier to predict epileptiform discharges.Right: Design of frequency-aware attention (FAA). The convolution (Conv) encoding of low-frequency tokens and invariant scattering encoding of high-frequency tokens are processed in separate pipelines. Attention maps are followed by channel mixing using pointwise convolution. The outputs are fused feature representations. Within the multi-head attention, we utilize cross-covariance attention to reduce computational cost.

Theorem 1 [(Li et al. 2017)] Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is a contraction mapping. Define the class computed by one convolutional layer followed by one fully connected layer with 2-norm constraint as:

$$F = \{\boldsymbol{x} \to \sum_i v_i \sigma(\boldsymbol{w}_i)\boldsymbol{x} : ||\boldsymbol{v}||^2 \leq 1, ||\boldsymbol{w}||_1 \leq B\} \quad (19)$$

For any $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N \in \mathbb{R}^d$, we have

$$\hat{G}_N(F) \leq \frac{cB(\ln d)^{1/2}}{N} \max_{\boldsymbol{j}-\boldsymbol{j}' \in \mathcal{N}} \sqrt{\sum_{i=1}^{N} ||\boldsymbol{x}_i(\boldsymbol{j}) - \boldsymbol{x}_i(\boldsymbol{j}')||^2} \quad (20)$$

where $\mathcal{N} \subset \mathbb{Z}^p$ defines the shape of the convolution filter as:

$$(\boldsymbol{w}_i * \boldsymbol{x})(\boldsymbol{k}) = \sum_{\boldsymbol{j} \in \mathcal{N}} \boldsymbol{w}_{i,\boldsymbol{j}} \boldsymbol{x}[\boldsymbol{k}](\boldsymbol{j}) \quad (21)$$

where $\boldsymbol{j}$ is the integer index vector that denotes the index shift. It is easy to see that minimize $\sum_{i=1}^{N} ||\boldsymbol{x}_i(\boldsymbol{j}) - \boldsymbol{x}_i(\boldsymbol{j}')||^2$ is equivalent to maximize the feature correlation $cov(\boldsymbol{x}_i(\boldsymbol{j}), \boldsymbol{x}_i(\boldsymbol{j}'))$. Therefore, strong feature correlation can improve the network generalization in some sense. In the following, we show that Scatter transform or Fourier transform can increase feature correlation, thus may improve the model generalizability.

Generalization Capacity Analysis of Wavelet Transform When tackling with scattering transform, we characterize it in the case of Morlet wavelet ψ:

$$\psi(\boldsymbol{x}) = \|C_1(e^{i\boldsymbol{x}\xi} - C_2)e^{-|\boldsymbol{x}|^2/(2\sigma^2)}\| \quad (22)$$

The following theorem shows that Morlet wavelet ψ can also increase the feature correlation.

Theorem 2 Suppose the input feature $\boldsymbol{x}$ is normalized, i.e. $||\boldsymbol{x}|| = 1$, then Morlet wavelet in Eq. (22) can increase feature correlation by appropriately choosing C_1, C_2, ξ.

Proof.

$$\begin{aligned}
&\|\psi(\boldsymbol{x}_i(\boldsymbol{j})) - \psi(\boldsymbol{x}_i(\boldsymbol{j}'))\| \\
\leq &\|C_1\| \cdot \|(e^{i\boldsymbol{x}_i(\boldsymbol{j})\xi} - C_2)e^{-|\boldsymbol{x}_i(\boldsymbol{j})|^2/(2\sigma^2)} \\
&-(e^{i\boldsymbol{x}_i(\boldsymbol{j}')\xi} - C_2)e^{-|\boldsymbol{x}_i(\boldsymbol{j}')|^2/(2\sigma^2)}\| \\
\leq &\|C_1\|\big(\|i\xi e^{i\boldsymbol{\eta}\xi}e^{-|\boldsymbol{\eta}|^2/(2\sigma^2)}\| \\
&\quad + \|(e^{i\boldsymbol{\eta}\xi} - C_2)e^{-|\boldsymbol{\eta}|^2/(2\sigma^2)}\frac{|\boldsymbol{\eta}|}{\sigma^2}\|\big) \cdot |\boldsymbol{x}_i(\boldsymbol{j}) - \boldsymbol{x}_i(\boldsymbol{j}')| \\
\leq &\|C_1\|\big(\|\xi\| + (\|C_2\| + 1)\frac{2}{\sigma^2}\big)\|\boldsymbol{x}_i(\boldsymbol{j}) - \boldsymbol{x}_i(\boldsymbol{j}')\| \quad (23)
\end{aligned}$$

As long as $\|C_1\|\big(\|\xi\| + (\|C_2\| + 1)\frac{2}{\sigma^2}\big) < 1$, we can increase the feature correlation, thus reduce the upper bound of Gaussian complexity, as shown in Theorem 1.

Generalization Capacity Analysis of Fourier Transform We further analyze the generalization capacity of Fourier transform. Suppose a Fourier transform γ featurize input coordinate with a set of sinusoids as follows:

$$\begin{aligned}\gamma(\boldsymbol{x}) =&[a_1 cos(2\pi \boldsymbol{b}_1^\top \boldsymbol{x}), a_1 sin(2\pi \boldsymbol{b}_1^\top \boldsymbol{x}), \cdots, \\ & a_m cos(2\pi \boldsymbol{b}_m^\top \boldsymbol{x}), a_m sin(2\pi \boldsymbol{b}_m^\top \boldsymbol{x})] \end{aligned} \tag{24}$$

where m is dimension of input feature x, a_i, b_i are parameters in the Fourier transform γ.

Then the inner product of Fourier features can be represented as follows:

$$\gamma(\boldsymbol{x}_1)^\top \gamma(\boldsymbol{x}_2) = \sum_{k=1}^{m} a_k^2 cos\big(2\pi \boldsymbol{b}_k^\top (\boldsymbol{x}_1 - \boldsymbol{x}_2)\big) \tag{25}$$

The following theorem shows that Fourier features can increase the feature correlation.

Theorem 3 Suppose the input feature $\boldsymbol{x}$ is normalized, i.e. $||\boldsymbol{x}|| = 1$, then the Fourier transform in Eq. (24) can increase feature correlation by appropriately choosing $a_1, a_2, \cdots, a_m$.

Proof.

$$\begin{aligned}&\sum_{i=1}^{N} ||\gamma(\boldsymbol{x}_i(\boldsymbol{j})) - \gamma(\boldsymbol{x}_i(\boldsymbol{j}'))||^2 \\ =&\sum_{i=1}^{N} \Big[\sum_{k=1}^{m} 2a_k^2 - 2a_k^2 cos\big(2\pi \boldsymbol{b}_k^\top (\boldsymbol{x}_i(\boldsymbol{j}) - \boldsymbol{x}_i(\boldsymbol{j}'))\big)\Big] \\ \leq& \sum_{i=1}^{N} \Big[\sum_{k=1}^{m} 4a_k^2\Big] = 4N \sum_{k=1}^{m} a_k^2 \end{aligned} \tag{26}$$

Therefore, we can increase the feature correlation by selecting $a_1, a_2, \cdots, a_m$ such that

$$4N \sum_{k=1}^{m} a_k^2 < \sum_{i=1}^{N} ||\boldsymbol{x}_i(\boldsymbol{j}) - \boldsymbol{x}_i(\boldsymbol{j}')||^2 \tag{27}$$

According Theorem 1, Fourier transform can reduce the upper bound of Gaussian complexity, thus may improve the model generalization capacity.

Experiments

Model Architecture

ScatterFormer adopts a hierarchical architecture illustrated in Figure 1, which is demonstrated to be effective for a range of tasks on two-dimensional signals due to its capability to generate multi-scale representation. Several variants are investigated. Specifically, main variants denoted by ConvScatter-1, ConvScatter-2 and ScatterFormer are examined, which correspond to networks with convolutional token embedding, with scattering token embedding positioned at the initial stage, with scattering token embedding positioned at all stages, respectively. In particular, the initial token embedding achieves 4×4 downsampling using a second-order invariant scattering layer. At later stages, a first-order invariant scattering layer is used. In order to investigate the frequency characteristics of ScatterFormer, we establish FourierFormer, where scattering layers are replaced with Local Fourier Unit (LFU). Moreover, we establish ProtoFormer, which calculates MHSA using naive setting of cross-covariance attention.

Experimental Setup

Datasets Extensive experiments are performed on two datasets for evaluation of proposed diagnosing models, with performance on epileptiform discharges detection and neonatal seizure or ictal epileptiform discharges detection being evaluated, respectively.

BECTS/Rolandic epilepsy dataset is a private dataset comprised of 110 patients (average age: 133.7 ± 27.4 months) with BECTS/Rolandic epilepsy, a common child epilepsy, and 170 normal controls (average age: 131.6±25.3 months). Collection and use of the data was approved by the Ethics Review Committee of the Children Hospital of Fudan University (No. 522-2020), and all subjects provided written informed consent. The annotated EEG epochs eligible for this work last for 446.55 h. Preprocessed data will be available at https://github.com/albertcheng19/scatterformer.

Helsinki University dataset is a public dataset that contains 39 patients with consensus annotation, 22 patients that are diagnosed as free of seizure in their records and 18 patients with no consensus annotation of ictal epileptiform discharges. Additional description can be referenced in (Stevenson et al. 2019). Data is available at https://zenodo.org/record/4940267.

Data Collection and Preprocessing Routine EEG recordings were performed on all subjects underwent continuous ($>$ 3 hours) sleep state with Nicolet EEG system (Natus Medical, Incorporated, San Carlos, CA, USA) with Ag/AgCl electrodes in the EEG examination room at the Hospital. Patients were seizure free $\leq$ 1 hours before study and receiving stable doses of medication. EEG electrodes were placed according to the conventional 10–20 EEG system and the guideline of American Clinical Neurophysiology Society. The impedance of the electrodes was calibrated under 3 $k\Omega$. The EEG signals were amplified and digitized at a sampling rate of 250 Hz, and then filtered at 0.1 Hz high-pass, 100 Hz low-pass and notch filter of 50 Hz. The classical Rolandic epileptic and normal EEG periods were labeled by two senior clinical neurophysiologists. The raw EEG data were saved in EDF format. EEG signals were preprocessed to remove the linear trend and eye movement artifacts using independent component analysis (ICA). Bipolar montage is adopted in our experiments.

Regularization Two types of augmentation is applied in training to alleviate over-fitting. The first one is termed channel reshuffling in our work. The feature representation of each channel is resized into a $3 \times 32 \times 256$ array, and 24 multi-spectral feature arrays in total are randomly rearranged into a $3 \times 768 \times 256$ array with different channels located in different positions of the array. The proposed method excludes the potential bias introduced by the specific arrangement of channels. After reshuffling, the input data are subject to further augmentation using MixUp (Zhang et al. 2017).

Training Details AdamW optimizer with weight decay of 0.05 is used. The initial learning rate is set to 5e-4 and progressively decays after each interaction by a cosine scheduler. During training, exponential moving average (EMA) with decay at 0.9999 is used to smooth the updating of weights. Unless otherwise stated, all models are trained with an 768×256 input size. The maximum training epochs are set at 50 with early stopping. All experiments are conducted on 2 NVIDIA A100 GPUs.

Experimental Results

Evaluation results on BECTS/Rolandic dataset are reported in Table 1 (a). Results of several variants of ScatterFormer is presented. Trained with the same initialization, data augmentation, learning rate tuning and regularization, the results suggest that both patch embedding implemented with invariant scattering layer and convolutional layer could improve the performance, but applying scattering transform only at the first stage leads to drop in AUCROC. In particular, ScatterFormer outperforms other variants and a strong baseline CNN model. Results on Helsinki University dataset are reported in Table 1 (b). In comparison to previous work, our method significantly surpasses the state-of-the-art given the constraints on data availability and cross-validation. Additional evaluation is reported in Table 3. ScatterFormer achieves satisfactory sensitivity to epileptic EEG segments in patient-independent classification. Moreover, as can be observed in Figure 2, performance on different folds follows skewed distribution. More concentrated distribution of ScatterFormer suggests a more robust overall performance across all folds. Therefore, ScatterFormer is a more robust model for generalization on heterogeneous individuals.

Ablation Studies

Effects of Invariant Scattering Layer To validate the efficacy of the frequency-aware attention dependent on dual-branch token embedding mechanism that simultaneously utilizes convolution and invariant scattering transform, we investigate its role in learning. Detection accuracy is increased by 3.52% compared to ProtoFormer (Table 1). The substitution by fast Fourier convolution results in drop in performance (Table 4) despite slightly reducing the number of parameters.

Effects of Activation ScatterFormer uses Mish activation. To investigate the effects of activation function, Mish is replaced with Swish, which leads to only negligible drops in several clinically important evaluation metrics with similar latency, as indicated in Table 3. Choice of nonlinear activation function does not seem to have significant influence on the expressivity and inference time of the network.

Interpretability

In this section, we investigate the interpretability and robustness of ScatterFormer. In particular, we discuss the clinical translatability of the associated research results.

(a) Results on BECTS/Rolandic Dataset

Model	Params (M)	LT (μs)	AUCROC	ACC
RegNet-Y-8G (Radosavovic et al. 2020)	37	179	$94.52_{2.22}$	$88.11_{18.98}$
Swin-B (Liu et al. 2021)	88	540	$97.81_{0.65}$	$94.04_{1.21}$
ProtoFormer	24	79	$97.91_{3.45}$	$92.87_{21.82}$
ConvScat-1	39	300	$98.15_{1.59}$	$96.33_{3.55}$
ConvScat-2	39	441	$96.85_{0.84}$	$95.10_{1.70}$
ScatterFormer	42	365	$98.14_{2.04}$	$96.87_{2.46}$

(b) Results on Helsinki University Dataset

Model	Params (M)	LT (μs)	AUCROC	ACC
ScatterFormer	42	365	$96.38_{5.66}$	$90.55_{11.75}$
SVM (Isaev et al. 2020)	-	-	$92.3_{12.1}$	-
ScatterFormer	42	365	90.31	89.67
SWT-FCN (Frassineti et al. 2020)	-	-	81	82
SWT-CNN (Frassineti et al. 2020)	-	-	77	79

Table 1: Results of ScatterFormer on patients with BECTS/Rolandic epilesy (a) and neonatal seizures (b). Compared with the existing deep learning models with similar amount of parameters, our method achieved superior outcomes and moderate inference time. We report evaluation metrics using median with interquartile range (IQR) as subscript. In addition, to compare our results of previous work, we also report values using mean value on Helsinki University dataset. Latency (LT) and number of parameters (Params) are reported for our experiments.

Disentanglement of Attention To empirically corroborate the effectiveness of dual-branch design, we visualize the Fourier heat maps of high- and low-frequency branches of attention module, respectively, and quantitatively analyze the magnitude-frequency relationship. Observably, disentangled attention achieves slower decay of magnitude spectrum as frequency increases than the prototypical architecture and captures more high-frequency information signals (Figure 3). The response curve suggests concentration of energy primarily in low-frequency range, which is consistent with the results revealed in (Wang et al. 2022). Noticeably, invariant scattering transform achieves more competitive performance than Fourier transform (Table 4). This performance gap could be attributed to energy preservation in scattering propagation, which corresponds to narrower frequency bands (Bruna and Mallat 2013) and could potentially avoid oversmoothing while alleviating noises.

Clinical Significance In Figure 4, we inspect the feature maps learned by different variants to analyze which features were actually used for discriminative prediction. The results suggest that ScatterFormer learns more fine-grained spec-

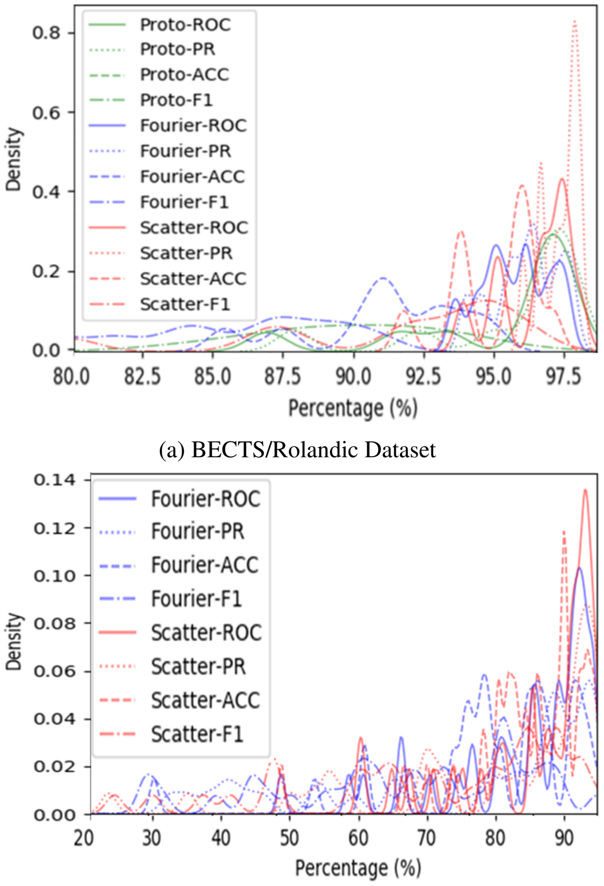
(a) BECTS/Rolandic Dataset

Model	AUCPR	F1
Proto	98.18$_{0.84}$	89.12$_{0.82}$
Scatter	98.88$_{1.13}$	93.93$_{6.42}$
Fourier	97.44$_{4.72}$	86.26$_{9.75}$

(b) Helsinki University Dataset

Model	AUCPR	F1
Scatter	92.32$_{22.63}$	79.90$_{28.88}$
Fourier	89.96$_{37.35}$	70.05$_{35.21}$

Table 2: Results of AUCPR and F1-score metrics on two datasets. ScatterFormer achieves significantly higher performance, indicating improved sensitivity to ictal samples.

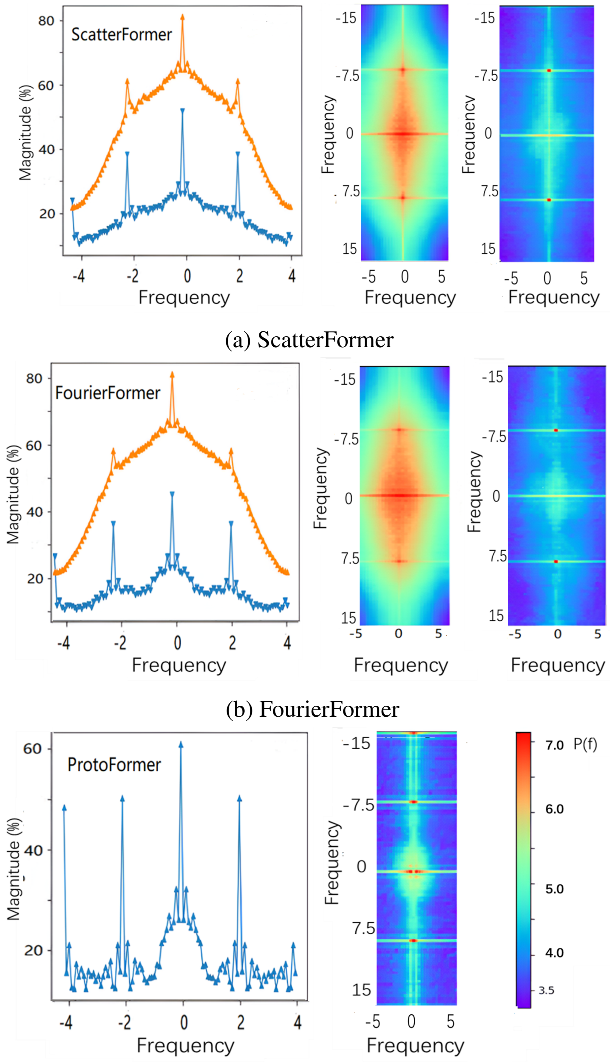


(a) BECTS/Rolandic Dataset

(b) Helsinki University Dataset

Figure 2. Probability distribution of evaluation metrics across cross-validation folds. Density is estimated using kernel method. ScatterFormer have more concentrated probability distribution of various metrics on cross-validation folds. The results suggest that ScatterFormer achieves higher performance and generalizability than other models in cross-validation setting.

tral representations of eletrophysiological biomarkers such as spike-and-waves and sharp waves. Features generated by ProtoFormer are smoother.



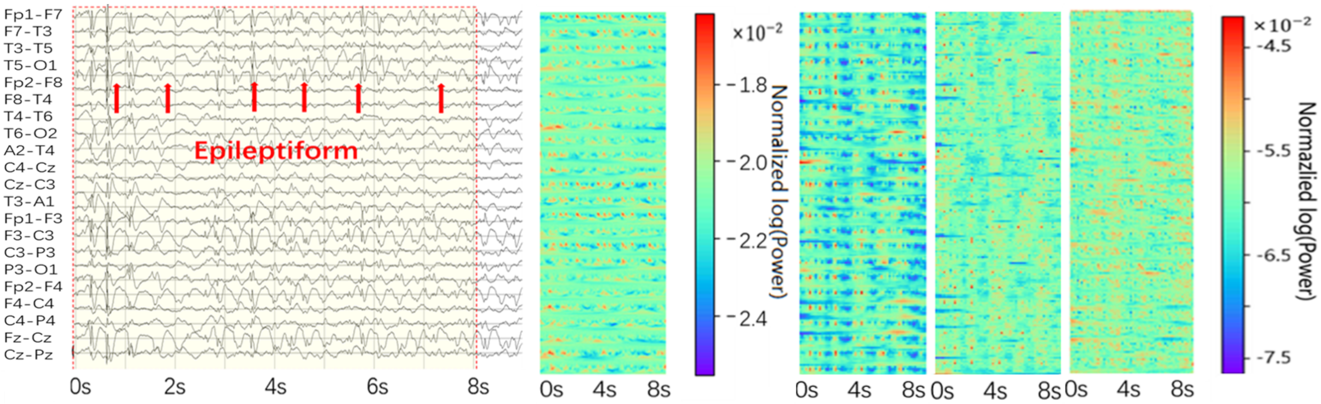
(a) ScatterFormer

(b) FourierFormer

(c) ProtoFormer

Figure 3. Fourier spectrum analysis of attention maps of three main variants investigated in this work. The results are averaged across 100 random samples at the 4^{th} layer from a randomly selected attention head. Both invariant scattering (a) and fast Fourier convolution (b) reduces the low-frequency preferability observed in (c), where prototypical attention tends to have sharper frequency response. Magnitude-Frequency responses (d) show that high-frequency components are enhanced by the dual-branch design. Spectra of high- and low-frequency branches are denoted by yellow and blue curves, respectively, in ScatterFormer and FourierFormer.

In addition, fast Fourier convolution could lead to aliasing of distinguishing spectral patterns that correspond to epileptic spikes, and between channels, as shown in Figure 4, which may explain the deteriorated inference-time performance. ScatterFormer enables more low-level features, especially regions associated with higher spectral power,



(a) Raw EEG Waveform (b) Input (c) Scatter (d) Fourier (e) Proto

Figure 4. Visualization of an EEG segment of epileptiform abnormalities and its corresponding visual representations. **(a)**: An illustrative EEG segment that contains continuous spike-and-wave during sleep (CSWS) characteristic of BECTS/Rolandic epilepsy. Epileptiform patterns are marked by red arrows for one channel with typical epileptic waves. **(b)**: Multispectral features. The color bar indicates normalized log wavelet power spectrum. **(c) (d) (e)**: Intermediate features learned by various Transformers. ScatterFormer (c) preserves more local spectral details that corresponds to spike-and-wave and sharp wave patterns appeared in the EEG than FourierFormer (d) and ProtoFormer (e). The latter both suffer oversmoothing of high-frequency information to various extents. The color bar indicates normalized intermediate feature values that do not represent power spectrum of raw data.

Activation	LT (μs)	AUCROC	AUCPR	ACC
Mish	365	$98.14_{2.04}$	$98.88_{1.13}$	$96.87_{2.46}$
Swish	358	$97.55_{1.90}$	$98.39_{1.29}$	$96.39_{2.00}$

Table 3: Ablation on the effects of activation functions. Two popular activation functions are investigated for ScatterFormer on BECTS/Rolandic dataset. No significant difference is observed.

(a) BECTS/Rolandic Dataset

Model	Params (M)	LT (μs)	AUCROC	ACC
Scatter	42	365	$98.14_{2.04}$	$96.87_{2.46}$
Fourier	39	239	$96.97_{1.83}$	$92.06_{2.96}$

(b) Helsinki University Dataset

Model	Params (M)	LT (μs)	AUCROC	ACC
Scatter	42	365	$96.38_{8.27}$	$90.05_{11.14}$
Fourier	39	239	$94.52_{8.92}$	$89.04_{11.93}$

Table 4: The Frequency-aware attention (FAA) module is replaced with Fast Fourier convolution to ablate its effects. The invariant scattering convolution outperforms the former by a median accuracy of 1.01% with respect to neonatal seizure detection task (b) without significant increase in parameters. The unit of latency is μs.

to be captured. Noise and artifacts are suppressed at initial phase. It is suggested that our approach achieves better cross-subject accuracy because it works in a manner similar to expert electroencephalographers, who distinguish epileptiform abnormalities from other clinically irrelevant activities by detecting of fine-grained edge-like patterns.

Conclusion

In this work, we propose Scattering Transformer (ScatterFormer), an invariant scattering transform-based hierarchical Transformer that distinguishes subtle variations associated with high-frequency textural information for detecting epileptiform discharges. We theoretically prove that wavelet transform can increase features correlations to compensate the high-frequency features in transformer. Strong feature correlations of scattering transform or Fourier transform lead to reduced upper bound of Gaussian complexity, thus may improve model generalizability. Improvement of generalizability by transformation of features to scattering domain is further validated in experiments.

Moreover, the scattering energy preservation feature allows more high-frequency information in different spatial, time, and frequency scales of epileptic EEG to be faithfully represented, which is reinforced by frequency-aware attention (FAA) that disentangles high- and low-frequency components. Therefore, the approach is able to differentiate subtle differences between spectral features converted from cEEG waveform records. We achieve optimal prediction AUCROC and accuracy in cross-subject detection epileptiform discharges in patients with BECTS/Rolandic epilepsy and neonates with heterogeneous etiologies for seizure, suggesting promise in accelerating clinical decision-making in various scenarios. Further analysis demonstrates the capability of ScatterFormer to extract discriminative patterns that are associated with epilepsy-specific EEG abnormalities, thereby offering clinical interpretability for supporting early and accurate identification of seizures. Our code is available at https://github.com/albertcheng19/scatterformer.

Acknowledgments

This work was supported from the Science and Technology Innovation 2030 - Brain Science and Brain-Inspired Intelligence Project (2021ZD0201301 and 2021ZD0200204), the National Natural Science Foundation of China (U20A20221), the Shanghai Municipal Science and Technology Major Project (2018SHZDZX01 and 2021SHZDZX0103) and ZJLab, Shanghai Municipal Science and Technology Committee of Shanghai outstanding academic leaders plan (21XD1400400).

References

Achilles, F.; Tombari, F.; Belagiannis, V.; Loesch, A. M.; Noachtar, S.; and Navab, N. 2018. Convolutional neural networks for real-time epileptic seizure detection. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3): 264–269.

Ali, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; et al. 2021. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34: 20014–20027.

Asif, U.; Roy, S.; Tang, J.; and Harrer, S. 2020. SeizureNet: Multi-spectral deep feature learning for seizure type classification. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, 77–87. Springer.

Bagchi, S.; and Bathula, D. R. 2022. EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. *Pattern Recognition*, 129: 108757.

Bai, J.; Yuan, L.; Xia, S.-T.; Yan, S.; Li, Z.; and Liu, W. 2022. Improving Vision Transformers by Revisiting High-frequency Components. *arXiv preprint arXiv:2204.00993*.

Bruna, J.; and Mallat, S. 2013. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1872–1886.

Cotter, F.; and Kingsbury, N. 2019. A learnable scatternet: Locally invariant convolutional layers. In *2019 IEEE International Conference on Image Processing (ICIP)*, 350–354. IEEE.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2021. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*.

Frassineti, L.; Ermini, D.; Fabbri, R.; and Manfredi, C. 2020. Neonatal Seizures Detection using Stationary Wavelet Transform and Deep Neural Networks: Preliminary Results. In *2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON)*, 344–349.

Isaev, D. Y.; Tchapyjnikov, D.; Cotten, C. M.; Tanaka, D.; Martinez, N.; Bertran, M.; Sapiro, G.; and Carlson, D. 2020. Attention-based network for weak labels in neonatal seizure detection. *Proceedings of machine learning research*, 126: 479.

Jeong, S.; Jeon, E.; Ko, W.; and Suk, H.-I. 2021. Fine-grained Temporal Attention Network for EEG-based Seizure Detection. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*, 1–4.

Li, Q.; Gao, J.; Zhang, Z.; Huang, Q.; Wu, Y.; and Xu, B. 2020. Distinguishing epileptiform discharges from normal electroencephalograms using adaptive fractal and network analysis: a clinical perspective. *Frontiers in Physiology*, 828.

Li, X.; Li, F.; Fern, X.; and Raich, R. 2017. Filter Shaping for Convolutional Neural Network. In *International Conference on Learning Representations*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Mallat, S. 2012. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10): 1331–1398.

Naseer, M. M.; Ranasinghe, K.; Khan, S. H.; Hayat, M.; Shahbaz Khan, F.; and Yang, M.-H. 2021. Intriguing Properties of Vision Transformers. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 23296–23308. Curran Associates, Inc.

Oyallon, E.; Belilovsky, E.; Zagoruyko, S.; and Valko, M. 2018. Compressing the input for cnns with the first-order scattering transform. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 301–316.

Pan, Z.; Cai, J.; and Zhuang, B. 2022. Fast Vision Transformers with HiLo Attention. *arXiv preprint arXiv:2205.13213*.

Radosavovic, I.; Kosaraju, R. P.; Girshick, R.; He, K.; and Dollár, P. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10428–10436.

Rasheed, K.; Qayyum, A.; Qadir, J.; Sivathamboo, S.; Kwan, P.; Kuhlmann, L.; O'Brien, T.; and Razi, A. 2020. Machine learning for predicting epileptic seizures using EEG signals: A review. *IEEE Reviews in Biomedical Engineering*, 14: 139–155.

Saminu, S.; Xu, G.; Shuai, Z.; Abd El Kader, I.; Jabire, A. H.; Ahmed, Y. K.; Karaye, I. A.; and Ahmad, I. S. 2021. A Recent Investigation on Detection and Classification of Epileptic Seizure Techniques Using EEG Signal. *Brain Sciences*, 11(5): 668.

Sazgar, M.; and Young, M. G. 2019. Overview of EEG, electrode placement, and montages. In *Absolute epilepsy and EEG rotation review*, 117–125. Springer.

Shi, Q.; Zhang, T.; Miao, A.; Sun, J.; Sun, Y.; Chen, Q.; Hu, Z.; Xiang, J.; and Wang, X. 2020. Differences between interictal and ictal generalized spike-wave discharges in childhood absence epilepsy: a MEG study. *Frontiers in Neurology*, 1359.

Shorvon, S.; and Schmidt, D. 2016. The right and the wrong with epilepsy and her science. *Epilepsia Open*, 1(3-4): 76–85.

Si, C.; Yu, W.; Zhou, P.; Zhou, Y.; Wang, X.; and Yan, S. 2022. Inception Transformer. *arXiv preprint arXiv:2205.12956*.

Siddhad, G.; Gupta, A.; Dogra, D. P.; and Roy, P. P. 2022. Efficacy of Transformer Networks for Classification of Raw EEG Data. *arXiv preprint arXiv:2202.05170*.

Singh, A.; and Kingsbury, N. 2017. Dual-tree wavelet scattering network with parametric log transformation for object classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2622–2626. IEEE.

Singh, A.; and Kingsbury, N. 2018. Generative scatternet hybrid deep learning (g-shdl) network with structural priors for semantic image segmentation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2991–2995. IEEE.

Stevenson, N. J.; Tapani, K.; Lauronen, L.; and Vanhatalo, S. 2019. A dataset of neonatal EEG recordings with seizure annotations. *Scientific data*, 6(1): 1–8.

Sun, J.; Xie, J.; and Zhou, H. 2021. EEG classification with transformer-based models. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, 92–93. IEEE.

Tao, Y.; Sun, T.; Muhamed, A.; Genc, S.; Jackson, D.; Arsanjani, A.; Yaddanapudi, S.; Li, L.; and Kumar, P. 2021. Gated transformer for decoding human brain eeg signals. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 125–130. IEEE.

Tatum, W.; Rubboli, G.; Kaplan, P.; Mirsatari, S.; Radhakrishnan, K.; Gloss, D.; Caboclo, L.; Drislane, F.; Koutroumanidis, M.; Schomer, D.; et al. 2018. Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. *Clinical Neurophysiology*, 129(5): 1056–1082.

Tatum, W. O. 2012. EEG interpretation: common problems. *Clinical Practice*, 9(5): 527.

Tatum, W. O.; and Shellhaas, R. A. 2020. Epileptiform discharges. *Neurology*, 94(20): 862–863.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22–31.

Xiang, J.; Li, C.; Li, H.; Cao, R.; Wang, B.; Han, X.; and Chen, J. 2015. The detection of epileptic seizure signals based on fuzzy entropy. *Journal of neuroscience methods*, 243: 18–25.

Yum, A.; and Shvarts, V. 2019. *Ictal and Interictal Epileptiform Electroencephalogram Patterns*, 239–250. Cambridge University Press.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.