PathFormer

Multi-scale Transformers with Adaptive Pathways for Time Series Forecasting

Yifei Ding

November 19, 2024

Contents

Introduction

Introduction

Motivations:

- Transformer calls for better designs and adaptations to fulfill its potential.
- Temporal resolution and temporal distance need to be considered.

Challengs:

- *Incompleteness of multi-scale modeling*;
- *Fixed multi-scale modeling*.

Related Work

Related Work

Time series forecasting:

- Deep learning methods: GNNs, RNNs, DeepAR, CNN, TimesNet, LLM-based methods, etc.
- Transformer models: Informer, Triformer, Autoformer, FEDformer, PatchTST, etc.

Multi-scale modeling for time series:

- N-HiTS, Pyraformer, Scaleformer, etc.

Methodology

Methodology

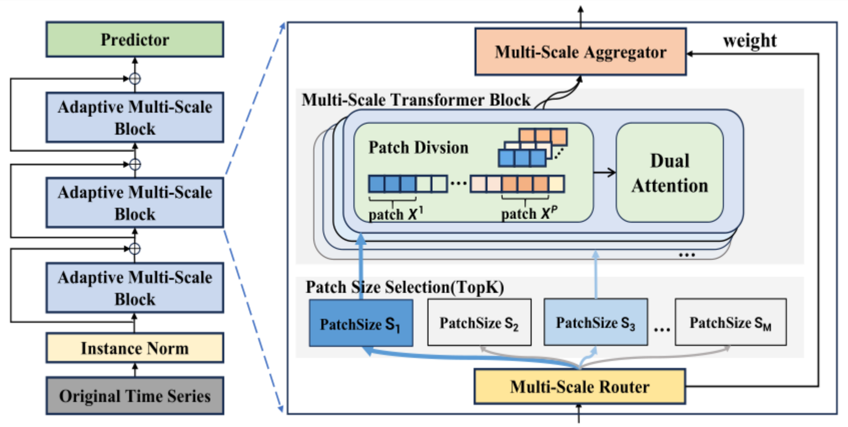

Figure 1: The Architecture of PathFormer.

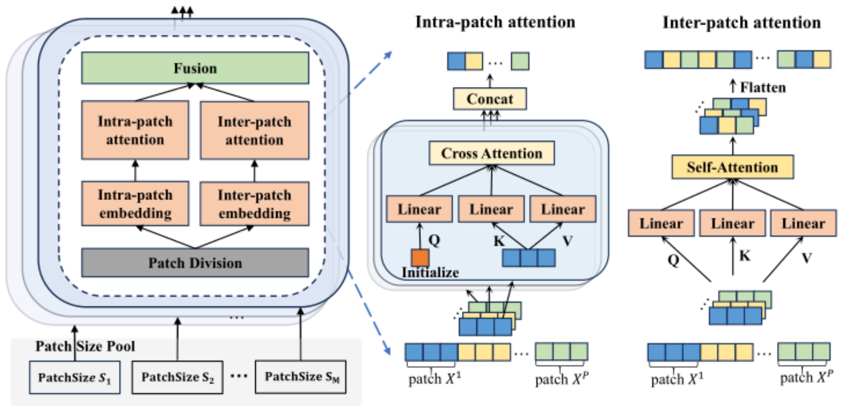

Figure 2: Multi-scale Transformer Block.

Multi-scale Transformer Block

1.Multi-scale Division.

- Define a collection of M patch size values as $\mathcal{S} = \{S_1, S_2, \cdots, S_M\}$;
- Define the input time series as $X \in \mathbb{R}^{H \times d}$, where H represents the length of the time series and d represents the dimension of features.

For each $i \in [1, M]$, divide X into $(X^1, X^2, \cdots, X^{P_i})$, $P_i = H/S_i$, $X^j \in \mathbb{R}^{S_i \times d}$, $j \in [1, P_i]$.

Multi-scale Transformer Block

2.Dual Attention.

Intra-patch Attention:

- Embed the patchs along the feature dimension d to get $X^j_{\text{intra}} \in \mathbb{R}^{S_i \times d_m}, \forall j \in [1, P_i]$;
- Perform trainable linear transformations on X^j_{intra} to get $K^j_{\text{intra}}, V^j_{\text{intra}} \in \mathbb{R}^{S_i \times d_m}$;
- Employ a trainable query matrix $Q^j_{\text{intra}} \in \mathbb{R}^{1 \times d_m}$.

$$\text{Attn}^j_{\text{intra}} = \text{Softmax}\left(Q^j_{\text{intra}}(K^j_{\text{intra}})^T / \sqrt{d_m}\right) V^j_{\text{intra}},$$

$$\text{Attn}_{\text{intra}} = \text{Concat}\left(\text{Attn}^1_{\text{intra}}, \cdots, \text{Attn}^{P_i}_{\text{intra}}\right).$$

Multi-scale Transformer Block

2.Dual Attention.

Inter-patch Attention:

- Embed feature along the feature dimension d to d_m;
- Rearrange the data to combine the two dimensions of S_i and d_m, making $X_{\text{inter}} \in \mathbb{R}^{P_i \times d'_m}$, $d'_m = S_i \cdot d_m$;
- Obtain $Q_{\text{inter}}, K_{\text{inter}}, V_{\text{inter}} \in \mathbb{R}^{P_i \times d'_m}$ by linear mapping on X_{inter}.

$$\mathrm{Attn}_{\text{inter}} = \mathrm{Softmax}\left(Q_{\text{inter}}(K_{\text{inter}})^T / \sqrt{d'_m}\right) V_{\text{inter}}.$$

Multi-scale Transformer Block

2.Dual Attention.

Final Output of Dual Attention:

- Rearrange the outputs of intra-patch attention to $\mathrm{Attn}_{\mathrm{intra}} \in \mathbb{R}^{P_i \times S_i \times d_m}$ by performing linear transformations on the patch size dimension from 1 to S_i;
- Add $\mathrm{Attn}_{\mathrm{intra}}$ with $\mathrm{Attn}_{\mathrm{inter}} \in \mathbb{R}^{P_i \times S_i \times d_m}$ to obtain the final output of dual attention $\mathrm{Attn} \in \mathbb{R}^{P_i \times S_i \times d_m}$.

Adaptive Pathways

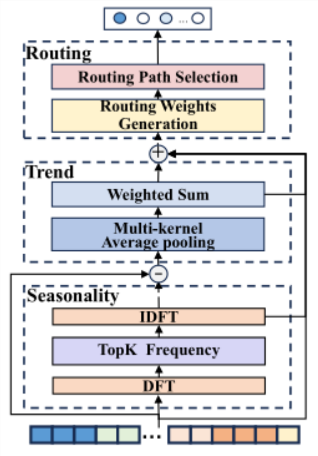

Figure 3: Multi-scale Router.

Adaptive Pathways

1.Multi-scale Router.

Seasonality Decomposition:

- Utilize DFT to decompose X into Fourier basis;
- Select the K_f basis with the largest amplitudes;
- Obtain $X_{\text{sea}} = \text{IDFT}\left(\{f_1, f_2, \cdots, f_{K_f}\}, A, \Phi\right)$, where Φ and A represent the phase and amplitude of each frequency from $\text{DFT}(X)$, $\{f_1, f_2, \cdots, f_{K_f}\}$ represent the frequencies with top K_f amplitudes.

Adaptive Pathways

1.Multi-scale Router.

Trend Decomposition:

- Get the remaining part after the seasonality decomposition $X_{\text{rem}} = X - X_{\text{sea}}$;
- Obtain the result from average poolings with different kernels and a weighted operation:

$$X_{\text{trend}} = \text{Softmax}(L(X_{\text{rem}})) \cdot (\text{Avgpool}_{\text{kernel}_1}(X_{\text{rem}}), \cdots, \text{Avgpool}_{\text{kernel}_N}(X_{\text{rem}}));$$

Adaptive Pathways

1.Multi-scale Router.

Final Result of Multi-scale Router:

- Add $X_{\text{sea}}, X_{\text{trend}}$ with X and perform a linear mapping $\mathrm{Linear}(\cdot)$ to transform and merge them along the temporal dimension to get $X_{\text{trans}} \in \mathbb{R}^d$.
- Generate pathway weights:

$$R(X_{\text{trans}}) = \mathrm{Softmax}(X_{\text{trans}} W_r + \varepsilon \cdot \mathrm{Softplus}(X_{\text{trans}} W_{\text{noise}})), \varepsilon \sim \mathcal{N}(0, 1);$$

- Perform top$K$ selection on the pathway weights, keeping the top K pathway weights and setting the rest weights as 0, and denote the final result as $\overline{R}(X_{\text{trans}})$.

Adaptive Pathways

2.Multi-scale Aggregator.

- Let X_{out}^{i} denote the output of the multi-scale Transformer with the patch size S_i;
- Define $T_i(\cdot)$ as a transformation function to align the temporal dimension from different scales;
- Get the final output of AMS block:

$$X_{\text{out}} = \sum_{i=1}^{M} \mathcal{I}\left(\overline{R}\left(X_{\text{trans}}\right)_i > 0\right) R(X_{\text{trans}})_i T_i(X_{\text{out}}^{i}).$$

Experiments

Experiments

Time Series Forecasting:

- The best performance in 81 cases and the second-best performance in 5 cases out of the overall 88 cases;
- Demonstrate a significant improvement when compared with PatchTST;
- Outperform when compared with strong linear models NLinear.

Transfer Learning:

- Can provide effictive lightweight transfer learning for time series forecasting.

Experiments

Ablation Studies:

- Verying the Number of Adaptively Selected Patch Sizes:
    - Adaptively modeling critical multi-scale characteristics improves accuracy;
    - Distinct time series samples benefit from feature extraction using varied patch sizes, but not all patch sizes are equally effictive.
- Visualization of Pathways Weights:
    - Underscore PathFormer's adaptability, emphasizing its ability to discern and apply the optimal patch size combinations for the diverse seasonality and trend patterns across samples.

Acknowledgement

Thank you!