Multi-Window Autoformer for Dynamic Systems Modelling

Autoformer++

Sergio Vanegas[1] Lasse Lensu[1] Fredy Ruiz[2]

[1]Department of Computational Engineering (CopE)
Lappeenranta-Lahti University of Technology LUT, Finland

[2]Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)
Politecnico di Milano (PoliMi), Italy

Workshop on Nonlinear System Identification Benchmarks, 8th Edition, April 25, 2024

Outline

Outline

Transformer

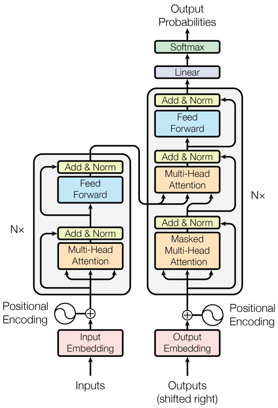

Figure 1: Transformer Architecture [1]

- Dates back to 2017 [1], sparking a revolution in predictive and generative models
    - Original *attention* concept from Bahdanau, Cho, and Bengio [2]
- Separation of input into context and fed-back output (*query**)
- Embeds sample/temporal order by adding a positional tensor to the embedded input

Attention Scoring

- Point-wise similarity [2] between two sequences (development focus of time-series-oriented attention mechanisms [3])
- Non-linearity introduced by pooling dot-product similarity scores
- Output sequence generated from a linear combination between nonlinear weights and a third input sequence

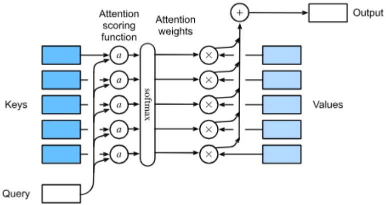

Figure 2: Attention Scoring Mechanism [4]

$$\text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)\mathbf{V}$$

Autoformer

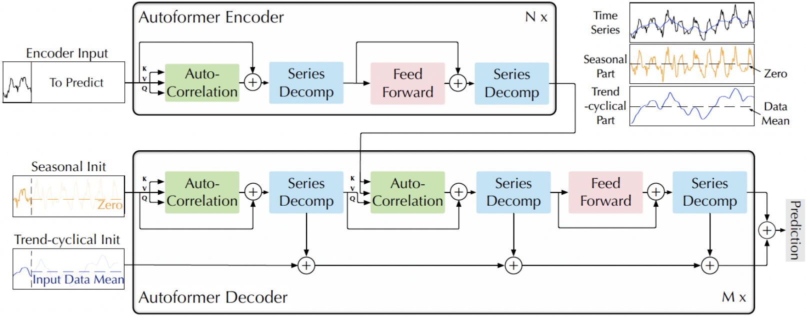

Figure 3: Autoformer Architecture [5]

Correlation-based Attention

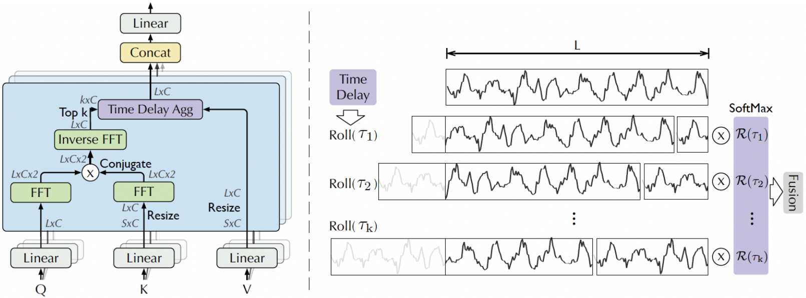

Figure 4: *Auto-Correlation** (left) and Time-Delay Aggregation (right) [5]. k function of the sequence length and the *attention sampling factor*.

Limitations

- Single periodicity assumption

$$\mathcal{X}_t = \mathrm{AvgPool}\left(\mathrm{Padding}\left(\mathcal{X}\right)\right)$$
$$\mathcal{X}_s = \mathcal{X} - \mathcal{X}_t$$
$$\mathcal{X}_s, \mathcal{X}_t = \mathrm{SeriesDecomp}(\mathcal{X})$$

- Input limited to past information + placeholders

$$\mathcal{X}_{\mathrm{enc},s}, \mathcal{X}_{\mathrm{enc},t} = \mathrm{SeriesDecomp}\left(\mathcal{X}_{\mathrm{enc}}\left[\frac{I}{2} : I\right]\right)$$
$$\mathcal{X}_{\mathrm{dec},s} = \mathrm{Concat}\left(\mathcal{X}_{\mathrm{enc},s}, \mathcal{X}_0\right)$$
$$\mathcal{X}_{\mathrm{dec},t} = \mathrm{Concat}\left(\mathcal{X}_{\mathrm{enc},t}, \mathcal{X}_{\mathrm{mean}}\right)$$

Outline

Multiple Seasonality Assumption

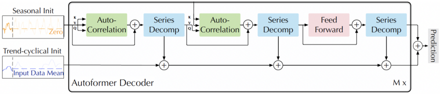

Figure 5: Decoder Information Flow

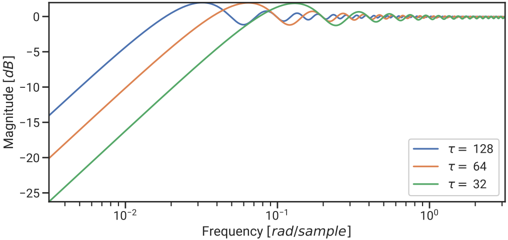

Figure 6: Seasonal magnitude response for different averaging windows

Controllable Future Assumption

System assumed to be auto-regressive with external inputs:

$$\widehat{\mathbf{y}}[t+O|t] = f(\mathbf{y}[t], \ldots, \mathbf{y}[t-I+1]; \mathbf{u}[t+O], \ldots, \mathbf{u}[t], \ldots, \mathbf{u}[t-I+1]$$

Placeholder input modified by applying SeriesDecomp (with the largest window) to the future control sequence:

$$\begin{aligned}
\mathcal{X}_{\text{enc},s}, \mathcal{X}_{\text{enc},t} &= \text{SeriesDecomp}\left(\mathcal{X}_{\text{enc}}\left[t - \frac{I}{2} : t\right]\right) \\
\mathcal{X}_{\text{dec},s} &= \text{Concat}_t\left(\mathcal{X}_{\text{enc},s}, \text{Concat}_c\left(\mathcal{U}_s[t+1 : t+O], \mathcal{X}_0\right)\right) \\
\mathcal{X}_{\text{dec},t} &= \text{Concat}_t\left(\mathcal{X}_{\text{enc},t}, \text{Concat}_c\left(\mathcal{U}_t[t+1 : t+O], \mathcal{X}_{\text{mean}}\right)\right)
\end{aligned}$$

Outline

Dataset Description

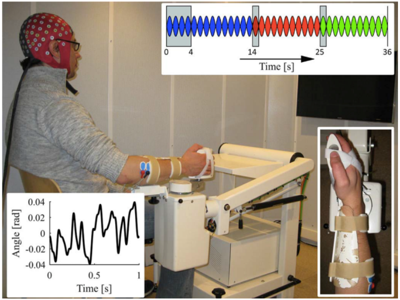

Figure 7: Cortical Responses Evoked by Wrist Joint Manipulation - Setup and Overview [6]

- Input: Angle of the manipulation (normalized)
- Output: EEG response (normalized)
- Medium dataset ($\sim$ 440 MB) consisting of 30 105 600 samples:
  1. 10 patients
  2. 7 realizations/patient
  3. 210 periods/realization
  4. 2 048 samples/period

Pre-Processing

- According to the original authors, all other coordinates being the same, constant signal across all periods:
    - Assumption partially forgone for sample preservation
- Noise filtered using median-over-window and

$$\hat{x}(i,j,k,l) = \alpha x(i,j,k,l) + \frac{(1-\alpha)}{N_{\text{periods}}} \sum_{k'=1}^{N_{\text{periods}}} x(i,j,k',l),$$

    - $\alpha \in [0,1]$: period's relative relevance
        - $\alpha = 0.25$ used in this work
    - i, j, k, l: patient, realization, period, and sample coordinates

Experiment Setup

- 70/30 train/validation split
- Prediction horizon of 64 samples
- Tuned hyperparameters
  - Context length
  - Encoding depth
  - Dropout rate
  - FF dimensionality
  - Number of Encoder/Decoder blocks
  - Attention sampling factor
  - Averaging window lengths
- Same split used to train similarly configured
  - LSTM [7]
  - Informer [3]
  - Canonical Autoformer [5]

Numerical Results

Architecture	Weight Count	Exec. Time [s]	MSE
LSTM	297 505	~~5 293.99 ± 1.29~~	0.444
Informer	226 273	**14.08 ± 0.07**	0.377
Autoformer	216 380	27.70 ± 0.07	0.440
Autoformer++	216 380	33.00 ± 0.08	**0.350**

Table 1: Benchmark - Numerical Results

- Networks configured by grabbing the equivalent parameters from the optimal hyperparameters tuned for the Autoformer++
- Execution time calculated over 10 iterations of the forecast for the entire dataset
- MSE calculated over the normalized validation dataset by averaging all periods (other coordinates untouched)

Graphical Results

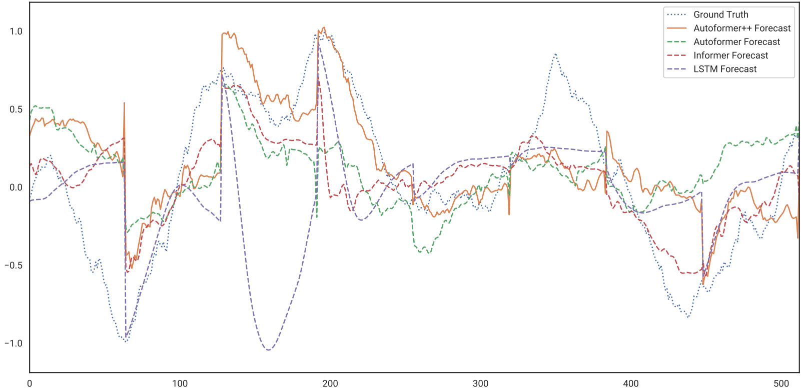

Figure 8: Benchmark results over a single realization (8 prediction horizons)

Outline

Conclusions

- The features that render transformer-like architectures desirable for Natural Language Processing (parallelism in training/deployment, context discrimination, modular design) also make them excel in dynamical system modelling over extended prediction horizons compared to more traditional recursive approaches.
- In the same time-complexity ($\mathcal{O}(N \log N)$), the *Auto-Correlation* mechanism achieves better accuracy than the Probabilistic-Sparse mechanism in Zhou, Zhang, Peng, et al. [3] by exploiting signal analysis theory and calculating time-correlation rather than approximating it.
- As expected, accuracy is improved when providing the model with future control inputs and independent averaging windows. Since the decoder stack already expects a placeholder sequence, structural changes are limited to the decoder-input initialization layer.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[3] H. Zhou, S. Zhang, J. Peng, et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in Proceedings of the AAAI conference on artificial intelligence, vol. 35, 2021, pp. 11 106–11 115.

[4] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, Dive into Deep Learning. Cambridge University Press, 2023, `https://D2L.ai`.

[5] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," Advances in neural information processing systems, vol. 34, pp. 22 419–22 430, 2021.

[6] M. P. Vlaar, G. Birpoutsoukis, J. Lataire, et al., "Modeling the nonlinear cortical response in eeg evoked by wrist joint manipulation," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 1, pp. 205–215, 2017.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.