
Minimal Time Series Transformer

Joni-Kristian Kämäriinen*
 Department of Computing Sciences
 Tampere University

Abstract

Transformer is the state-of-the-art model for many natural language processing, computer vision, and audio analysis problems. Transformer effectively combines information from the past input and output samples in autoregressive manner so that each sample becomes aware of all inputs and outputs. In sequence-to-sequence (Seq2Seq) modeling, the transformer processed samples become effective in predicting the next output. Time series forecasting is a Seq2Seq problem. The original architecture is defined for discrete input and output sequence tokens, but to adopt it for time series, the model must be adapted for continuous data. This work introduces minimal adaptations to make the original transformer architecture suitable for continuous value time series data.

Jupyter notebook:
https://github.com/kamarain/minimal_time_series_transformer

1 Introduction

In their seminal work, Vaswani *et al.* [6] introduced the transformer architecture for machine translation. The proposed model was trained supervised manner with training data consisting input \mathbf{X} and output sequences \mathbf{Y} . The sequences were corresponding sentences in two languages. During inference, the model is executed in generative manner

$$\begin{aligned}\hat{y}_0 &= \text{Transformer}(\mathbf{X}, \text{SOS}), \\ \hat{y}_1 &= \text{Transformer}(\mathbf{X}, \text{SOS}, \hat{y}_0), \\ \hat{y}_2 &= \text{Transformer}(\mathbf{X}, \text{SOS}, \hat{y}_0, \hat{y}_1), \\ &\dots \\ \hat{y}_N &= \text{Transformer}(\mathbf{X}, \text{SOS}, \hat{y}_0, \hat{y}_1, \dots, \hat{y}_{N-1}) = \text{EOS}\end{aligned}$$

to produce an output sequence $\mathbf{Y} = [\text{SOS}, \hat{y}_0, \dots, \hat{y}_N]$ (translation) corresponding to the input sequence \mathbf{X} , and in which SOS and EOS are special symbols (tokens) 'start-of-sequence' and 'end-of-sequence'. In follow up works, the transformer model has been extended to other modalities and tasks. *Transformer* is undoubtedly the state-of-the-art model for sequential data.

Time series are used in statistics, signal processing, econometrics, control engineering, and largely in any domain of applied science and engineering which involves temporal measurements. *Time series forecasting* is the use of a time series model to predict future values based on previously observed values. The forecasting problem can be cast into the sequence-to-sequence (Seq2Seq) model and generative inference defined in (1). Therefore it is intriguing to adopt transformers for time series forecasting.

Several adaptations have been proposed for time series forecasting. For example, LogSparse Transformer [5], MTS Transformer [8], Informer [9], and ContiFormer [2]. These works address various challenges of time series forecasting such as varying sampling rate and long-term dependencies.

*See https://webpages.tuni.fi/vision/public_pages/JoniKamarainen/

However, it is not clear what are the minimal changes to adapt the ‘vanilla’ transformer [6] for sequences of continuous data. Some adaptation is needed since the vanilla transformer is designed for discrete inputs and outputs, tokens, and does not work with continuous data. The rather complicated adaptations in the previous works raise questions since they have not been compared to any ‘time series transformer baseline’. Moreover, since Zeng *et al.* [7] experimentally demonstrated that an “embarrassingly simple” single layer linear regressor outperforms state-of-the-art time series transformers in multiple benchmarks, it is justified to define a simple time series transformer baseline - *a minimal time series transformer*.

This work investigates minimal adaptations of the vanilla transformer [6] for time series forecasting. To keep things simple and understandable, all experiments are performed on Sinusoid sequences. All code for the time series transformer models and experiments are made publicly available.

2 Methods

2.1 Vanilla Transformer - Seq2SeqTransformer

Our reference implementation is the experimentally verified `Seq2SeqTransformer` class from [4]. Their class uses the popular PyTorch `torch.nn.Transformer` class as the main building block. This makes the `Seq2SeqTransformer forward()` function easy to follow:

```
.forward()
    # Note: src & tgt default size is (seq_length, batch_num, feat_dim)

    # Token embedding
    src = self.embedding(src) * math.sqrt(self.d_model)
    tgt = self.embedding(tgt) * math.sqrt(self.d_model)

    # Positional encoding - _must_ be seq len x batch num x feat dim
    # Inference often misses the batch num
    if src.dim() == 2: # seq len x feat dim
        src = torch.unsqueeze(src, 1)
    src = self.positional_encoder(src)
    if tgt.dim() == 2: # seq len x feat dim
        tgt = torch.unsqueeze(tgt, 1)
    tgt = self.positional_encoder(tgt)

    # Transformer output
    out = self.transformer(src, tgt, tgt_mask=tgt_mask,
                          src_key_padding_mask=src_key_padding_mask,
                          tgt_key_padding_mask=tgt_key_padding_mask,
                          memory_key_padding_mask=src_key_padding_mask)
    out = self.unembedding(out)

    return out
```

The main processing steps are 1) the source and target sequence embedding using the PyTorch `torch.nn.Embedding` class that maps token ids (integer values) to one-hot encoded vectors and further to an embedding vector, 2) positional encoding implemented using an encoding class that follows the original work [6], and 3) the linear ‘un-embedding’ layer that transforms the transformer outputs to class probabilities. Note that the PyTorch implementation of cross-entropy loss adds the softmax-nonlinearity, and thus it is not explicitly in the forward function. The PyTorch transformer class takes care of masking the future outputs and padding keys, and mask construction is available in the reference code. A general, and often confusing, requirement is that the source and target variables must be three dimensional: *sequence length × number of batches × feature vector dimension*. Inside the class there are some cumbersome steps to maintain the correct shapes, and these steps are often the causes of programming bugs.

2.2 Minimal time series transformer - MiTS-Transformer

The minimal adaptation to the discrete token `Seq2SeqTransformer` is to change the ‘integer-to-vector’ embedding layer (`torch.nn.Embedding`) to a layer that converts continuous value vectors to vectors of the model dimension - ‘vector-to-vector’. A neural network trick to do this is to replace the embedding layer with a linear layer. This can be done by a small change in the original code. The original embedding

```
.init()
```

```

self.embedding = nn.Embedding(num_tokens, d_model)
self.unembedding = nn.Linear(d_model, num_tokens)

```

is replaced by

```

.init()
self.embedding = nn.Linear(d_input, d_model)
self.unembedding = nn.Linear(d_model, d_input)

```

The embedding in the continuous case maps the `d_input`-dimensional sample to `d_model`-dimensional model vector. In the un-embedding step the conversion is made backwards.

2.3 Positional encoding expansion - PoTS-Transformer

In time series forecasting, there are three potential challenges for transformers, *i*) sequences can be long, from thousands to tens of thousands samples, *ii*) temporally close samples can be highly correlated, and *iii*) amount of training data can be limited. The straightforward solutions to the challenges are conflicting. Long sequences require high dimensional models so that there is 'space' for position information. The small amount of data makes it difficult to train large models due to overfitting. Strong correlation between the adjacent samples makes it possible to use small models since each new sample provides only negligible amount of new information. To meet the conflicting requirements, special tricks are needed.

In the existing literature, particularly the long sequences have received attention. For example, Li *et al.* [5] propose logarithmically sparse sampling of sequences so that pose encoding becomes logarithmic reducing the need of high dimensional models. Wav2vec 2.0 is a state-of-the-art audio backbone network for audio feature extraction [1]. It compresses long audio sequences to more compact representation - 'audio tokens' - by first applying convolution filters and then using a nearest-neighbor product quantizer [3].

What is the simplest solution for the challenges? In *PoTS-Transformer* the model size is kept low to avoid overfitting. At the same time, the positional encoding is done in a higher dimensional space to keep it effective for long sequences. This is implemented by wrapping the positional encoder between two linear layers that first perform *positional encoding expansion* and after the encoding step *inverse positional expansion*:

```

.forward()
...
src = self.pos_expansion(src)
src = self.positional_encoder(src)
src = self.pos_inversion(src)
...
tgt = self.pos_expansion(tgt)
tgt = self.positional_encoder(tgt)
tgt = self.pos_inversion(tgt)

```

The expansion and its inverse are defined as

```

.init()
...
# Positional encoding expansion
self.pos_expansion = nn.Linear(d_model, pos_expansion_dim)
self.pos_inversion = nn.Linear(pos_expansion_dim, d_model)

# Positional encoding
self.positional_encoder = PositionalEncoding(d_model=pos_expansion_dim,
                                              dropout=dropout_p)
...

```

Performing the transformer model computations with 8-dimensional vectors and the pose encoding with 128-dimensional vectors increases the number of learnable parameters in PoTS-Transformer from 1,433 to 3,473 (2.4 \times). For comparison, increasing the model size of MiTS-Transformer from 8- to 128-dimensional spaces increases the number of parameters from 1,289 to 204,689 (158 \times). The difference is nearly two orders of magnitude.

2.4 Training details

All transformer models were trained with the standard PyTorch Adam optimizer (`torch.optim.Adam`) using the initial learning rate 0.023. A multistep scheduler was used to decay the learning rate of each parameter group by gamma once the number of epoch reaches one of the milestones. The gamma was fixed to 0.1 and the milestones were manually optimized for each case separately, and the goal losses documented to facilitate replication of all experiments. However, similar results can be achieved with the fixed learning rate 0.023, no scheduling, and training for 2000 epochs. The training of the largest model in our experiments took only a few minutes on a standard laptop without a GPU.

3 Data

The Seq2Seq data for the experiments were generated by sampling the sine function

$$y(t) = \sin 2\pi f t$$

f is the since frequency and t defines the time instant. For discrete signals, it is convenient to define the frequency as discrete frequency of waveforms per number of samples:

$$f = \frac{w}{L}$$

where $w \in \mathbb{R}$ is the number of waves and L the number of samples. For example, $f = 2/31$ means 2 waveforms per 31 samples. Sinusoids with varying frequencies are shown in Figure 1. Figure 1(d) shows sampling errors that become more dominant as the frequency increases (note that the top of each wave looks different).

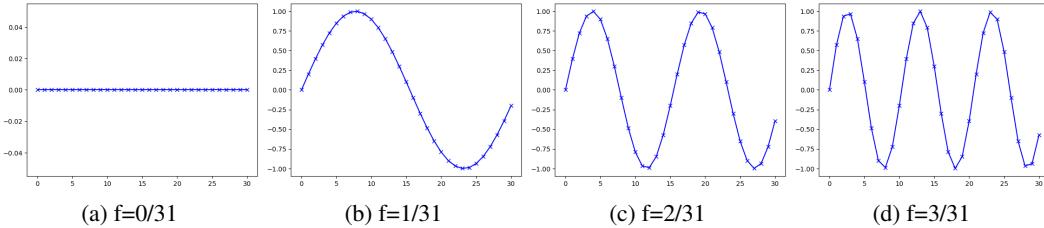


Figure 1: Example sinusoids used in the experiments.

Data types. Three types of data were generated:

1. Type 1: Single sequence (sinusoid with $f = 1/L$)
2. Type 2: A fixed number of sequences (e.g., $f = 0/L, 1/L, 2/L, 3/L$)
3. Type 3: Arbitrary number of sequences ($f \in U(0, f_{max})$)

$U()$ is the uniform random distribution. Type 1 is used to sanity check the models. Type 2 is an easy case where there are a small number of different sequences. Type 3 is the most difficult case as the difference between two sinusoids of the nearly same frequency can be subtle. In the experiments the signal length is set to 31 which is divided to 19 input (source) samples \mathbf{X} and 12 output (target) samples \mathbf{Y} (Figure 2).

4 Experiments

4.1 Code sanity check with a single fixed sequence (Type 1 sequences)

The sanity checks were done with a single sinusoid of the lenght $L = 31$ and frequency $f = 1/31$. The signal was repeated 100 times in the training and test sets to mimic standard training.

MiT-S-Transformer. The model converged after 200 epochs with the fixed learning rate 0.023 and obtained the final MSE error 0.23. See Figure 3 for examples.

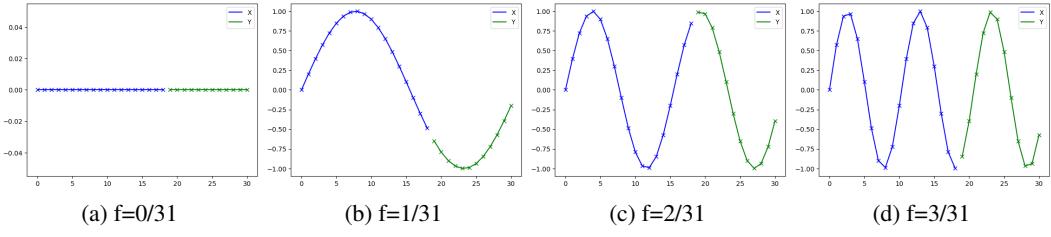


Figure 2: Sinusoids (Type 2 sequences) of 31 samples divided into source **X** (samples 0-18 in blue) and target **Y** parts (samples 19-30 in green).

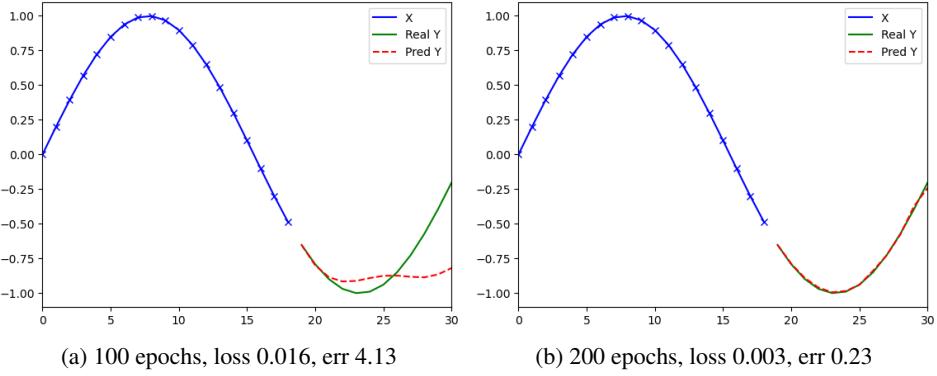


Figure 3: Single sequence sanity check (Type 1 data) of the MiTS-Transformer implementation (see Jupyter notebook).

4.2 Multiple sequences (Type 2)

MiTS-Transformer. The results and illustrations in Figure 4 demonstrate good learning performance. There is virtually no difference between the single and multiple (4) sequence accuracy.

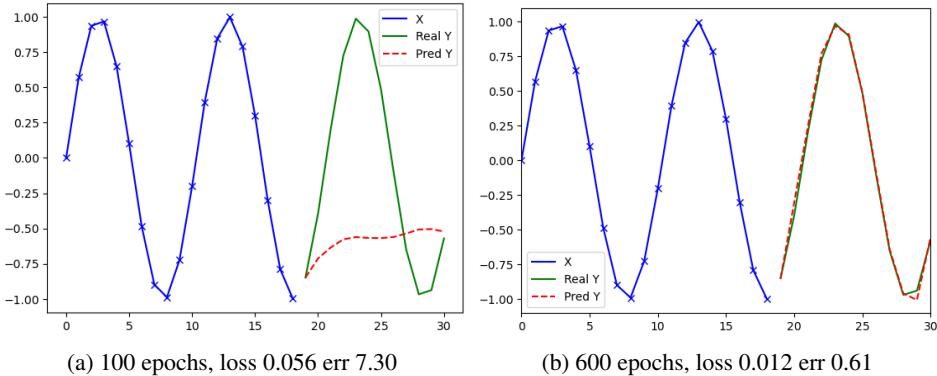


Figure 4: Four sequence (Type 2) results of MiTS-Transformer (model parameters: $d_model=8$, $dim_feedforward=8$, the total of 1,289 learnable parameters). The signal frequencies were $0/31$, $1/31$, $2/31$, and $3/31$.

4.3 Arbitrary sequences (Type 3)

MiTS-Transformer. The results and sample illustrations from three different runs with the same training and test data are in Figure 5. The results verify that the minimal time series transformer

works for data with arbitrarily small changes in the sequences. The results verify certain *interpolation capability* of the model since many test sequences do not appear in training data. The results for the arbitrary sequences are almost order of magnitude worse than for the fixed Type 1 and 2 sequences.

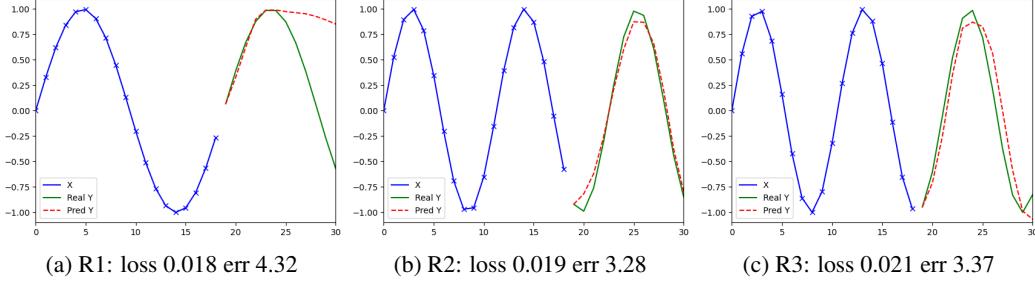


Figure 5: Arbitrary sequences (Type 3) results for MiTS-Transformer ($d_{model}=8$, $dim_feedforward=8$, 1289 params). Data consists of arbitrary sequences of Sinusoids with the frequency in $(0/31, 3/31]$.

The learning capacity of MiTS-Transformer can be increased by increasing the model dimension (d_{model}). In Figures 6 and 7 are results for the models with the model dimension set to 16 and 32. The size 16 model is systematically better than the size 8 model, but the size 32 model starts to occasionally overfit to the training data. The total number of learnable parameters in the 8, 16, and 32 models are 1289 to 4097 and 14321, respectively.

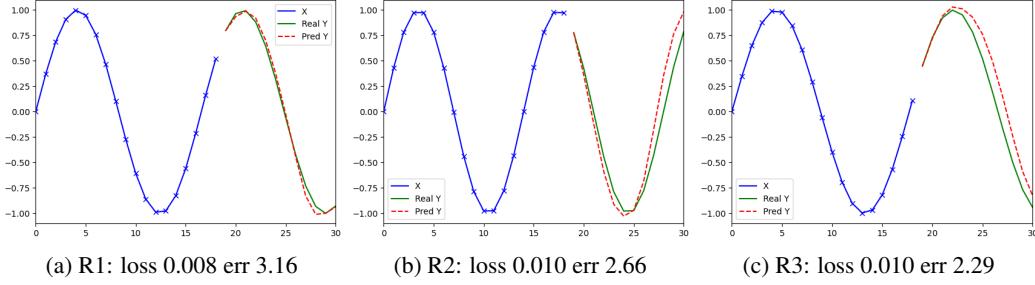


Figure 6: Arbitrary sequences (Type 3) results for MiTS-Transformer ($d_{model}=16$, $dim_feedforward=8$, 4097 params).

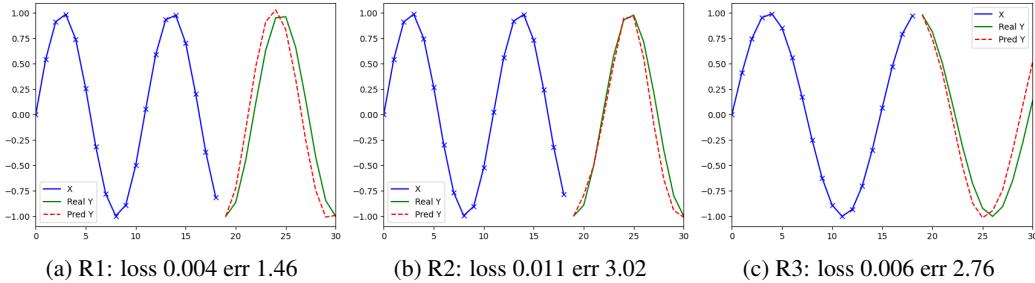


Figure 7: Arbitrary sequences (Type 3) results for MiTS-Transformer ($d_{model}=32$, $dim_feedforward=8$, 14321 params).

PoTS-Transformer. The experiments were run using the positional expansion dimension 64 and 2,385 learnable parameters. The results in Figure 8 demonstrate systematically better performance than the MiTS-Transformer.

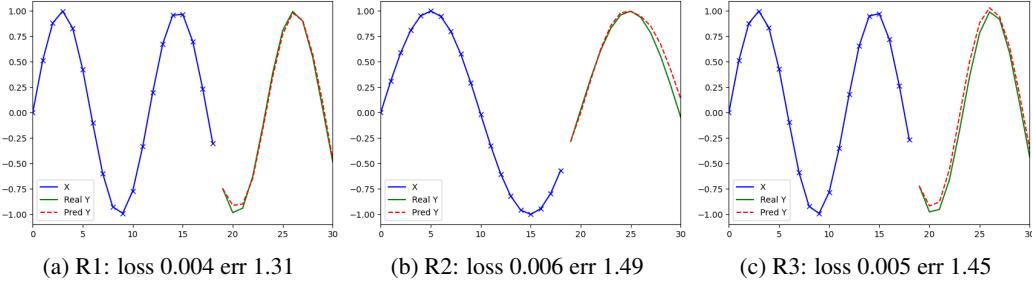


Figure 8: Arbitrary sequences (Type 3) results for the PoTS-Transformer ($d_{model}=8$, $dim_{feedforward}=8$, $pos_expansion_dim=64$, 2,385 params).

5 Conclusions

The purpose of this work and the accompanying Jupyter notebook was to study how the “Attention is All You Need” (vanilla) transformer for discrete tokens can be adapted for time series (continuous) data. The minimal adaptation is to change the token embedding layer to a linear layer. This was implemented in our minimal time series transformer (MiTS-Transformer) which learned sinusoids very well. The learning capacity of MiTS-Transformer can be adjusted by changing the model dimension. This however quickly leads to model size explosion in the terms of learnable model parameters and the model easily overfits. For this problem a simple model called positional encoding expansion time series transformer (PoTS-Transformer) was proposed. It combines positional encoding of long sequences in the expanded space and a low dimensional model avoiding overfitting. The results in our experiments were convincing and pave way for similar minimal and simple tricks for transformer based time series forecasting.

Acknowledgments and Disclosure of Funding

This work would not be possible without my tenure at Tampere University. Tenure allows me to allocate time to projects that interest me. To thank my institute and the Finnish government, I share my results and findings with everyone interested in machine learning and programming.

If you find my writings or code helpful, please cite this work. The work is available as an ArXiv article. Mention the author name(s), title, and the ArXiv URL in your project report, Web page, thesis, etc.

Most universities worldwide are ideal places for curious and playful people like me. I hope the corporative style management will not destroy them as similar people to me have been developing them for the past thousand years.

Keep on Rockin’ in the Free World!

References

- [1] A. Baevski et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020.
- [2] Y. Chen et al. “ContiFormer: Continuous-Time Transformer for Irregular Time Series Modeling”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023. URL: <https://arxiv.org/abs/2402.10635v1>.
- [3] J. Jegou, M. Douze, and C. Schmid. “Product Quantization for Nearest Neighbor Search”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 33.1 (2011).
- [4] Joni-Kristian Kämäräinen. *Introduction to Sequence Modeling with Transformers*. 2025. arXiv: 2502.19597 [cs.LG]. URL: <https://arxiv.org/abs/2502.19597>.
- [5] S. Li et al. “Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2019.

- [6] Ashish Vaswani et al. “Attention Is All You Need”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2017.
- [7] A. Zeng et al. “Are Transformers Effective for Time Series Forecasting?” In: *The AAAI Conference on Artificial Intelligence (AAAI)*. 2023.
- [8] George Zerveas et al. “A Transformer-based Framework for Multivariate Time Series Representation Learning”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*. 2021. DOI: 10.1145/3447548.3467401.
- [9] H. Zhou et al. “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting”. In: *The AAAI Conference on Artificial Intelligence (AAAI)*. 2021.