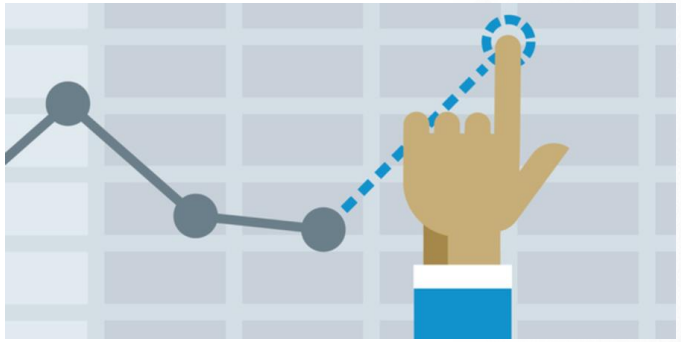
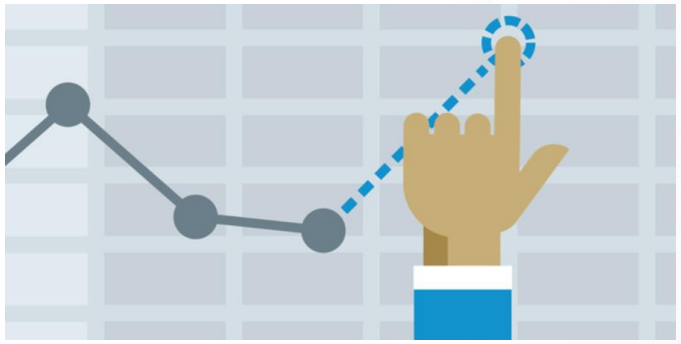


Transformer for time series forecasting

2021 여름 IAB

이예준



Background

Time series forecasting has been...

ARIMA

- Linear assumption
- Limited scalability

RNN (LSTM)

- Difficult to train
- Struggles to capture long-term dependencies

Li et al., “Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting”, NeurIPS, 2019.

Transformer

Vaswani et al., “Attention is all you need”, NeurIPS, 2017.

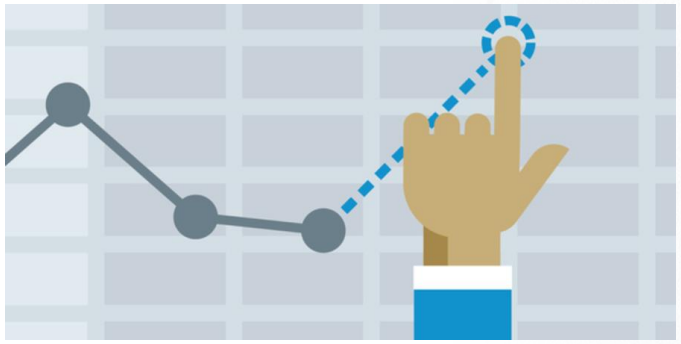
NLP

Vision

Arguably the hottest🔥 algorithm in DL

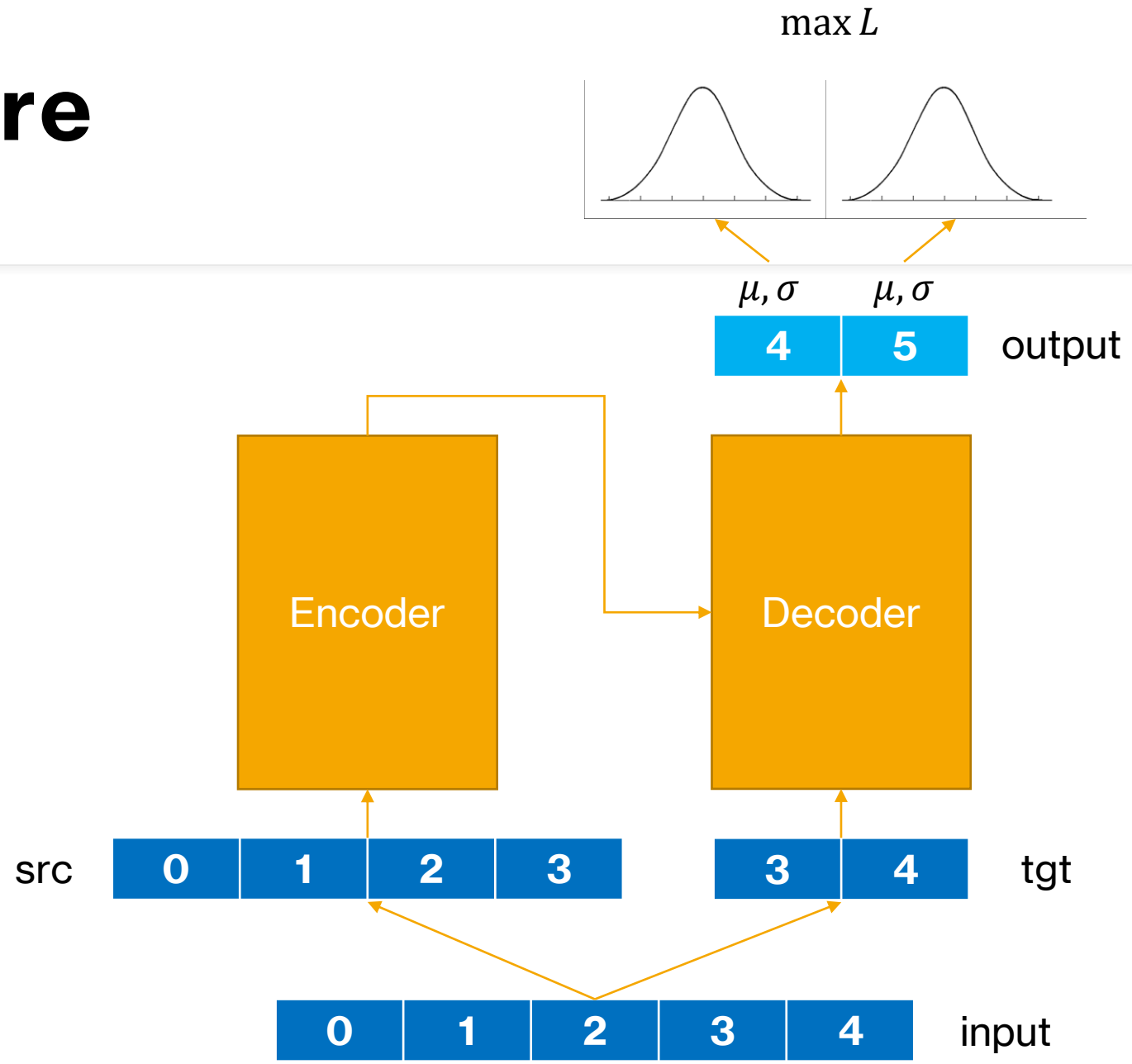
Transformer in time series forecasting

- Li et al., “Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting”, NeurIPS, 2019.
- Wu et al., "Deep transformer models for time series forecasting: The influenza prevalence case.“, ICML, 2020.
- Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting.“, Proceedings of AAAI, 2021.

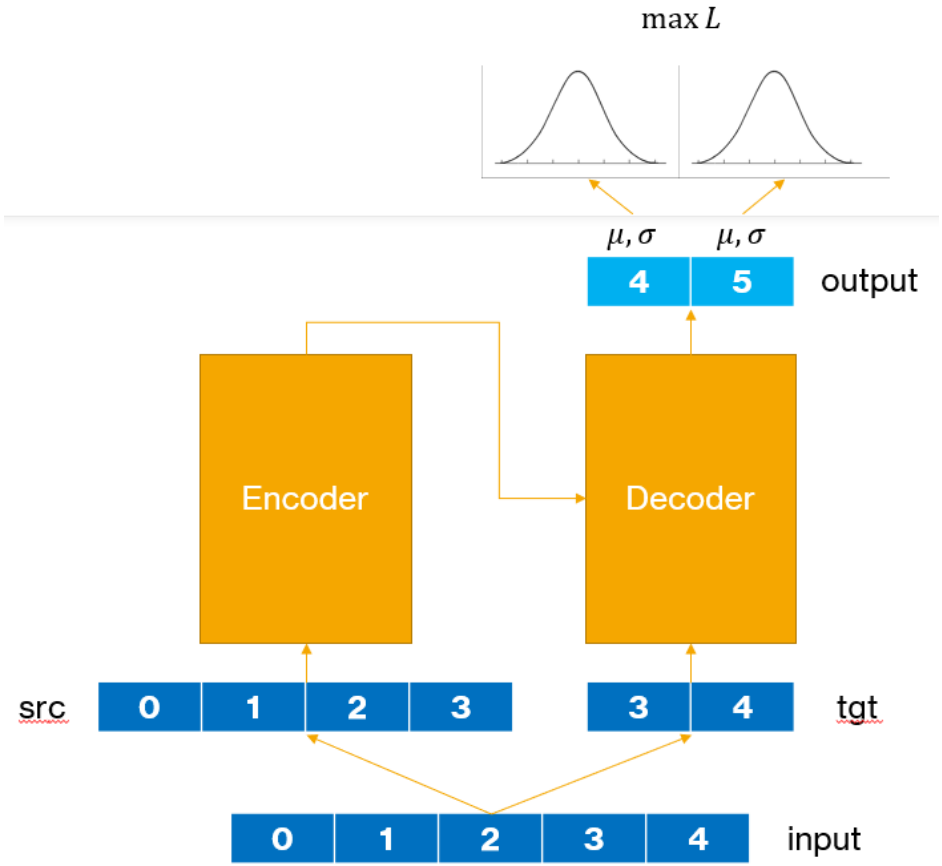


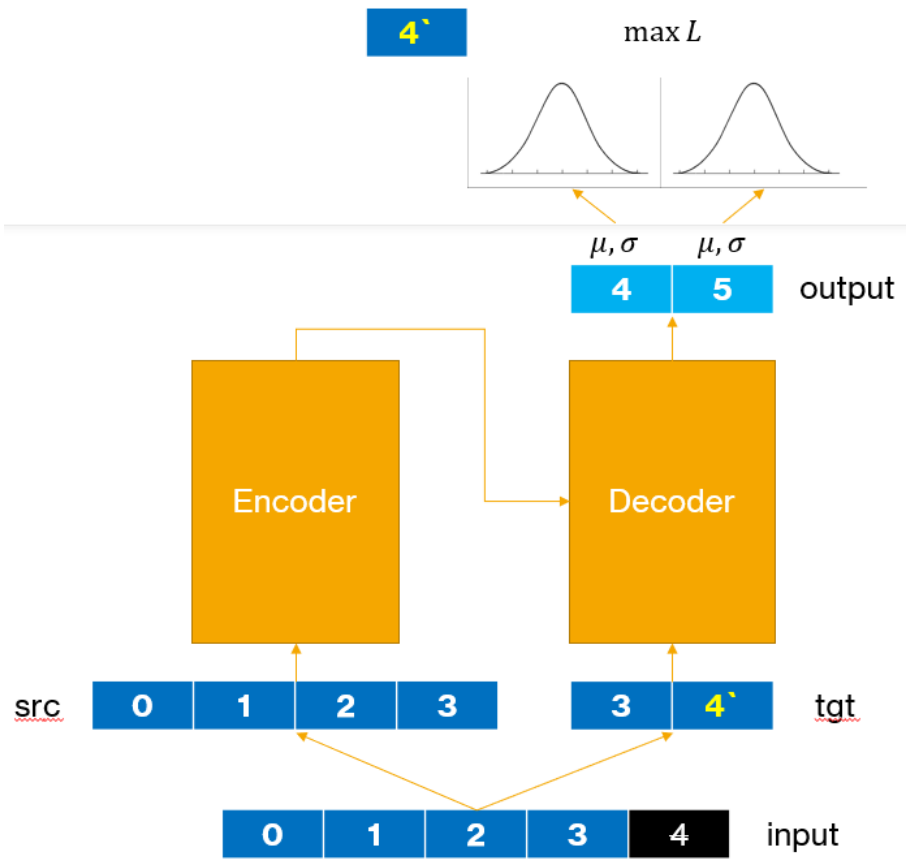
Overview

Model structure



Short-term vs Long-term

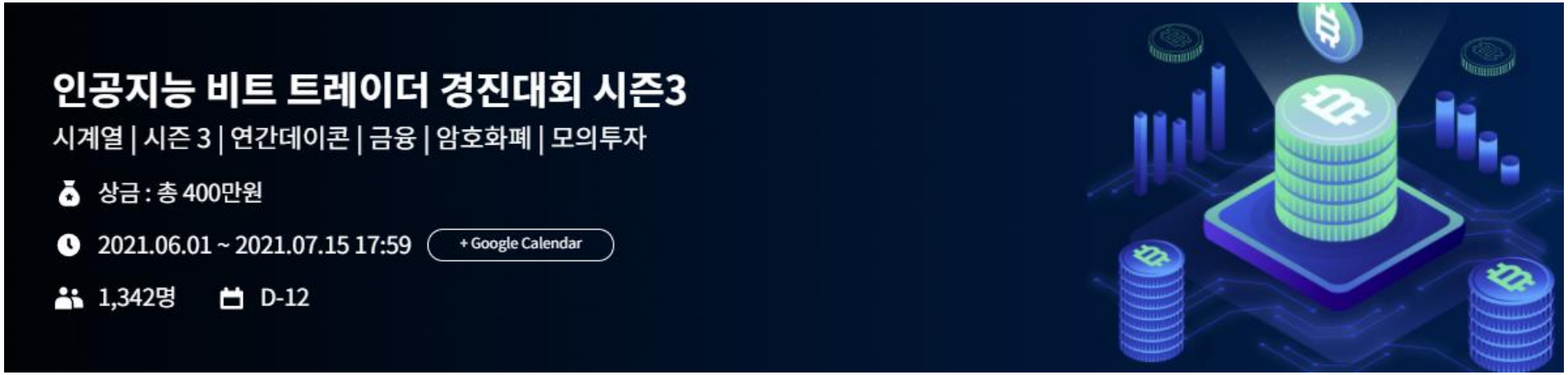




Goal

1. Build a pipeline.
2. Improve the model by tweaking the model structure and hyperparameters.

Coin dataset

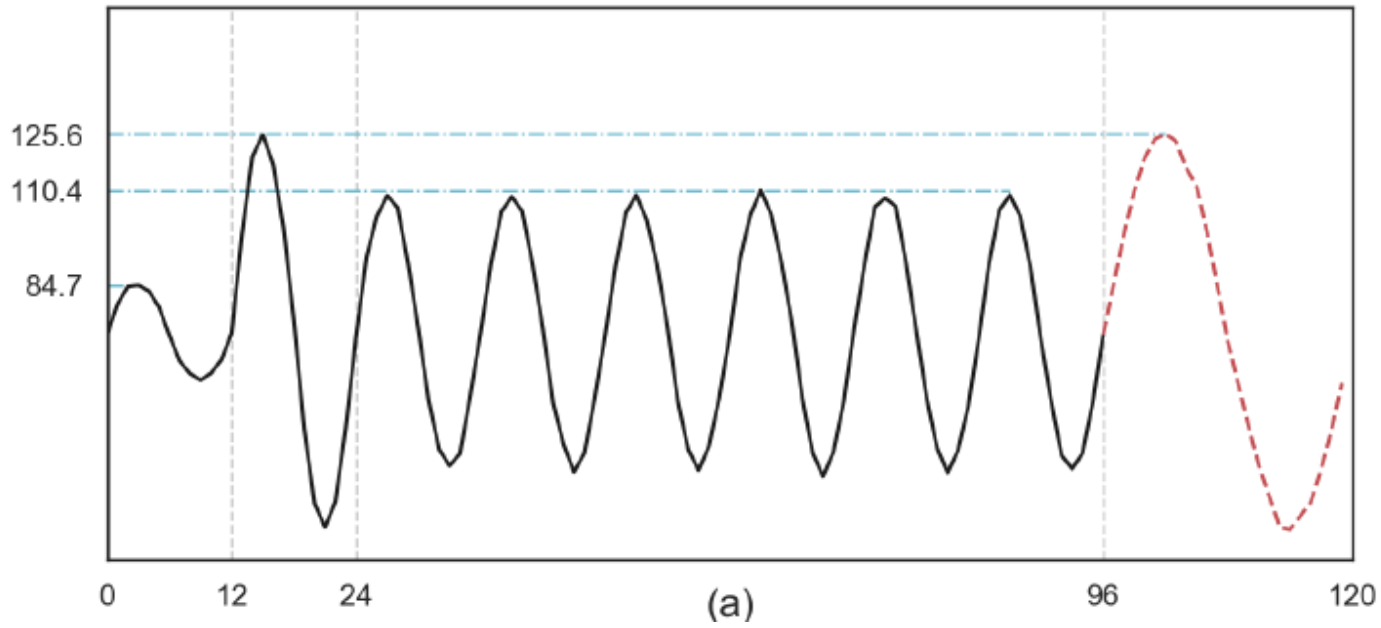


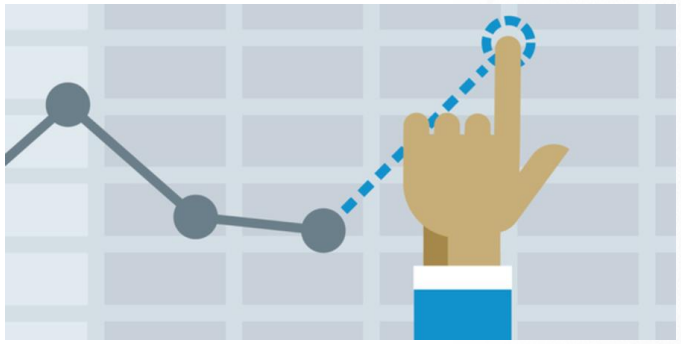
- Data provided per minute
- ‘open’ as a time series
- seq length = 1500 → source length = 300, target length = 10

Synthetic dataset

Li et al., “Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting”, NeurIPS, 2019.

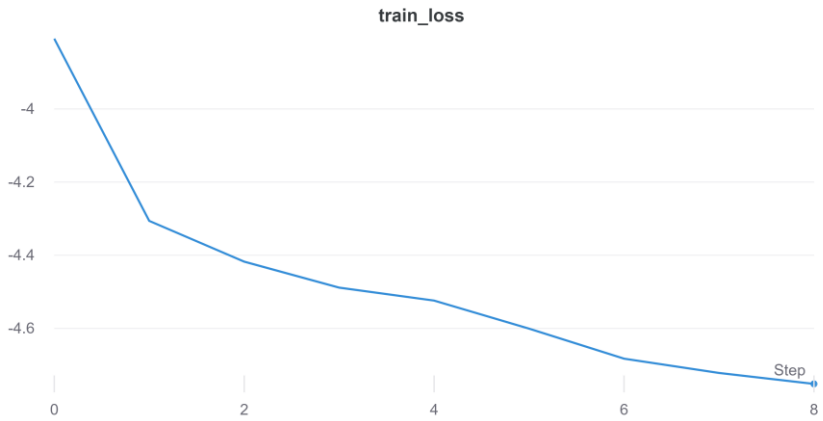
source length = 96,
target length = 24

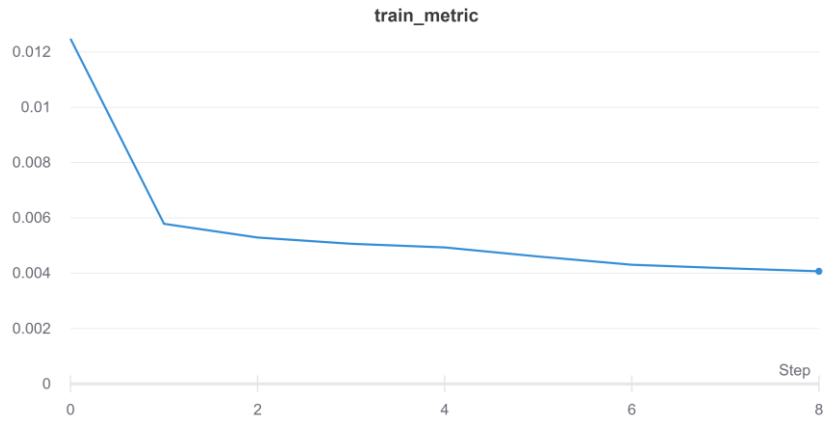


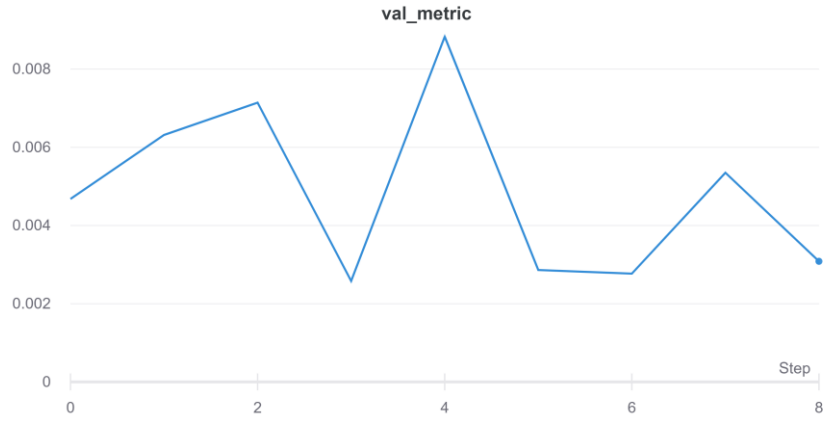


Results

Coin dataset – training curve



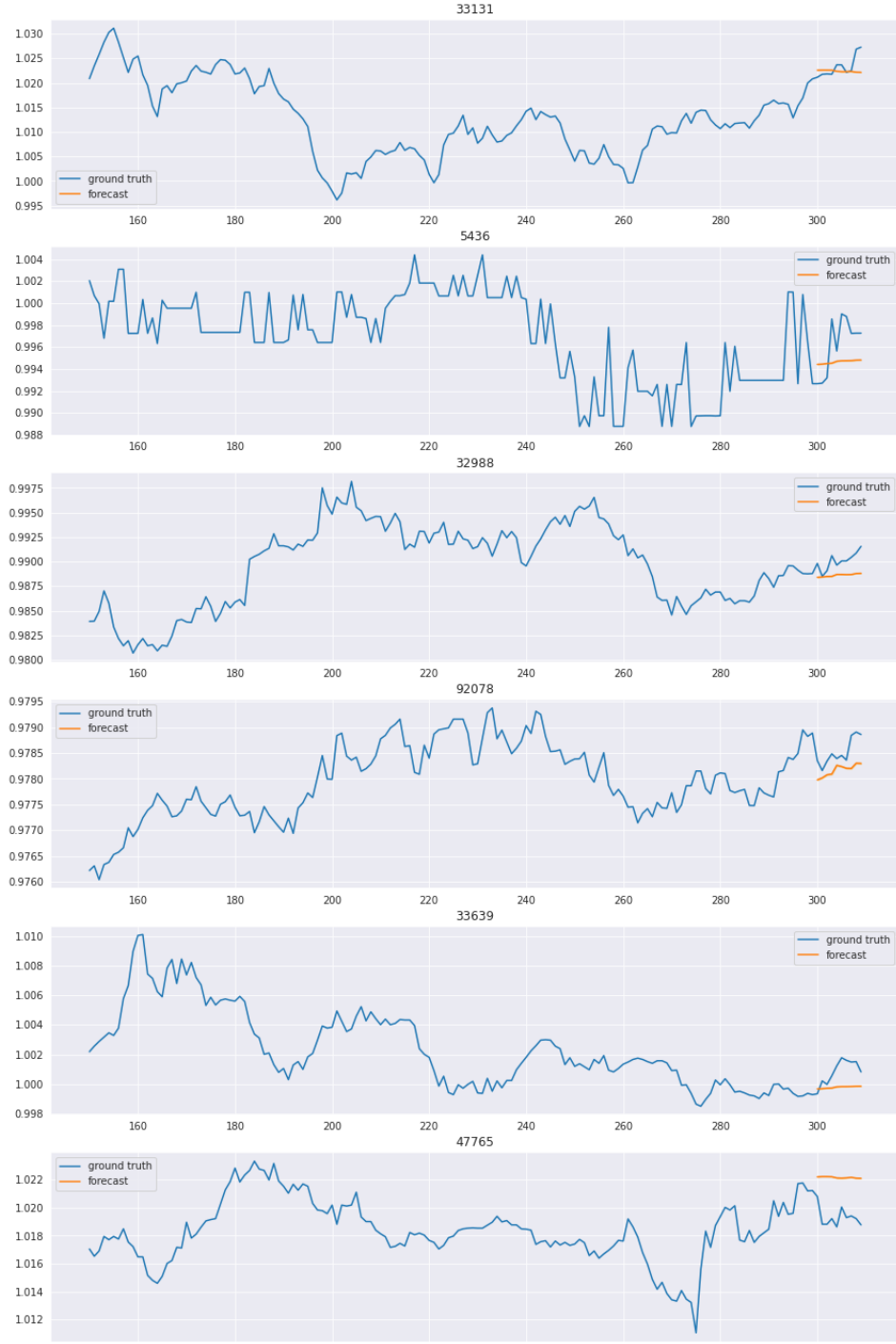


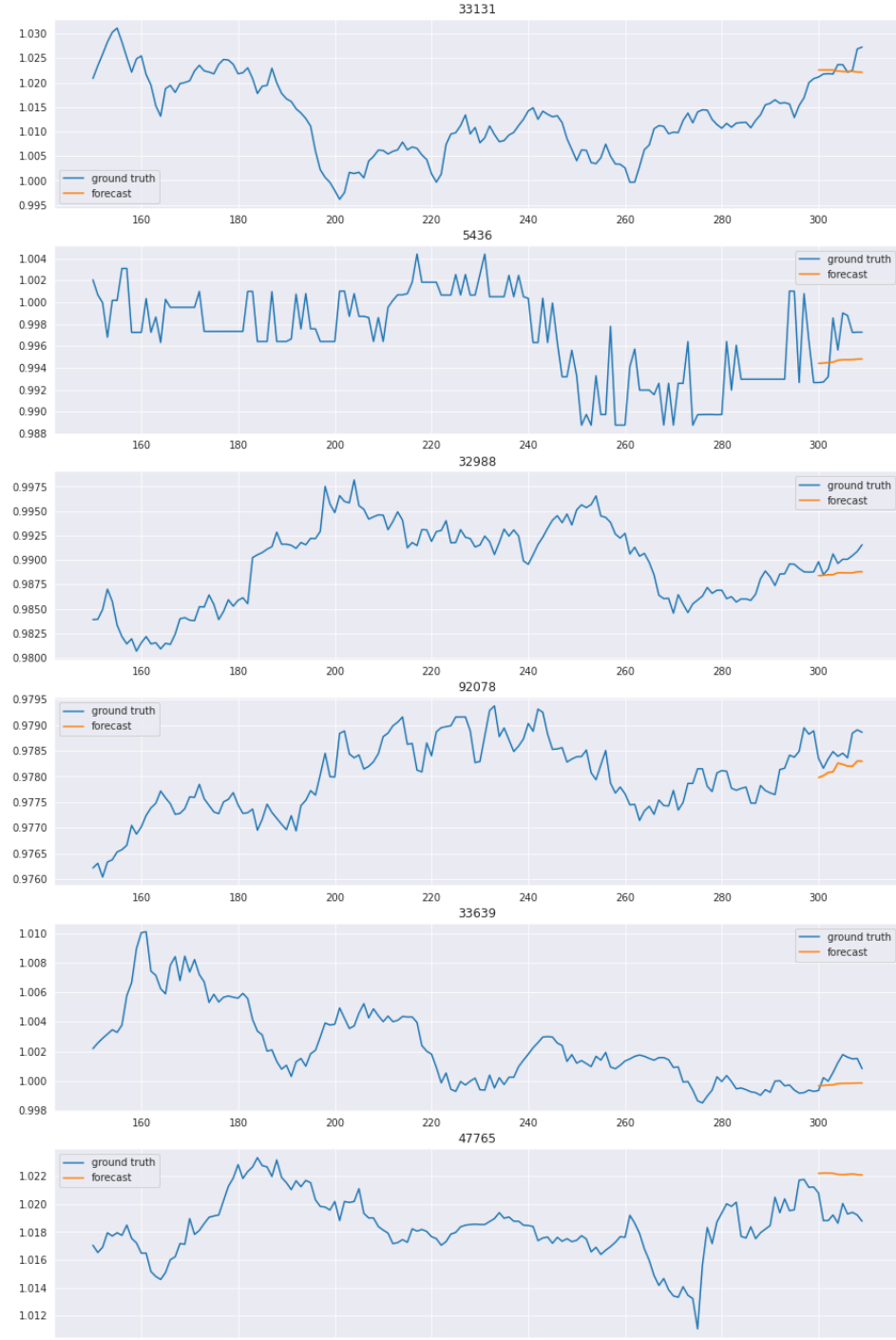


- Loss: Negative log likelihood
- Metric: Mean Absolute Error(MAE)

Val loss is not falling 😢

Short-term





Long-term

Coin dataset - summary

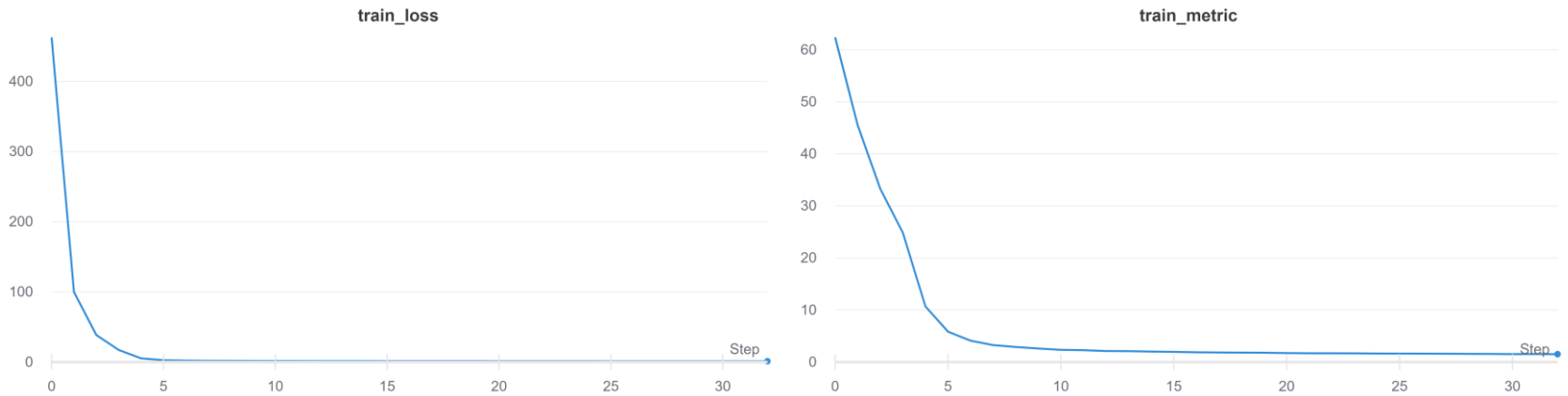
*Only use 1/10 for long-term forecasting
**Value for long-term forecasting

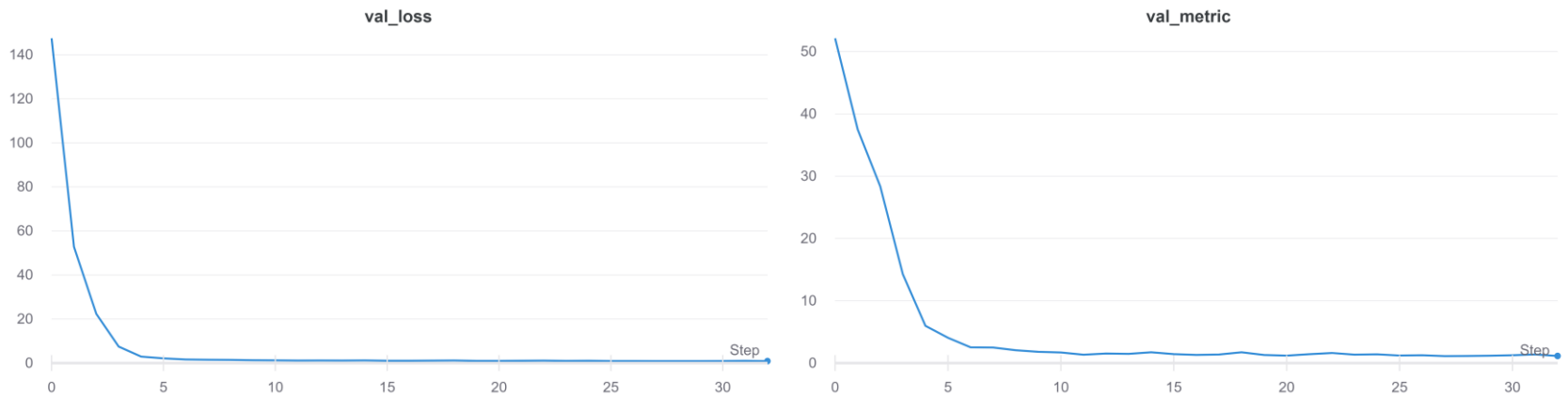
	Train (608K)	Val (152K)	Test (95K*)
Loss (NLL)	-4.68	-5.12	-5.13 (-5.11**)
MAE	0.0043	0.0027	0.0025 (0.0026**)

- Metrics above and plots do show some signs of underfitting.
- Nonetheless, more training leads to significant overfitting in which model simply *copies* the previous value.
- Then long-term forecasting turns into disaster where model simply repeats the upward/downward direction.

- Hyperparameter optimization was not done enough due to limited resources.
- **Lack of patterns for model to learn**

Synthetic dataset – training curve



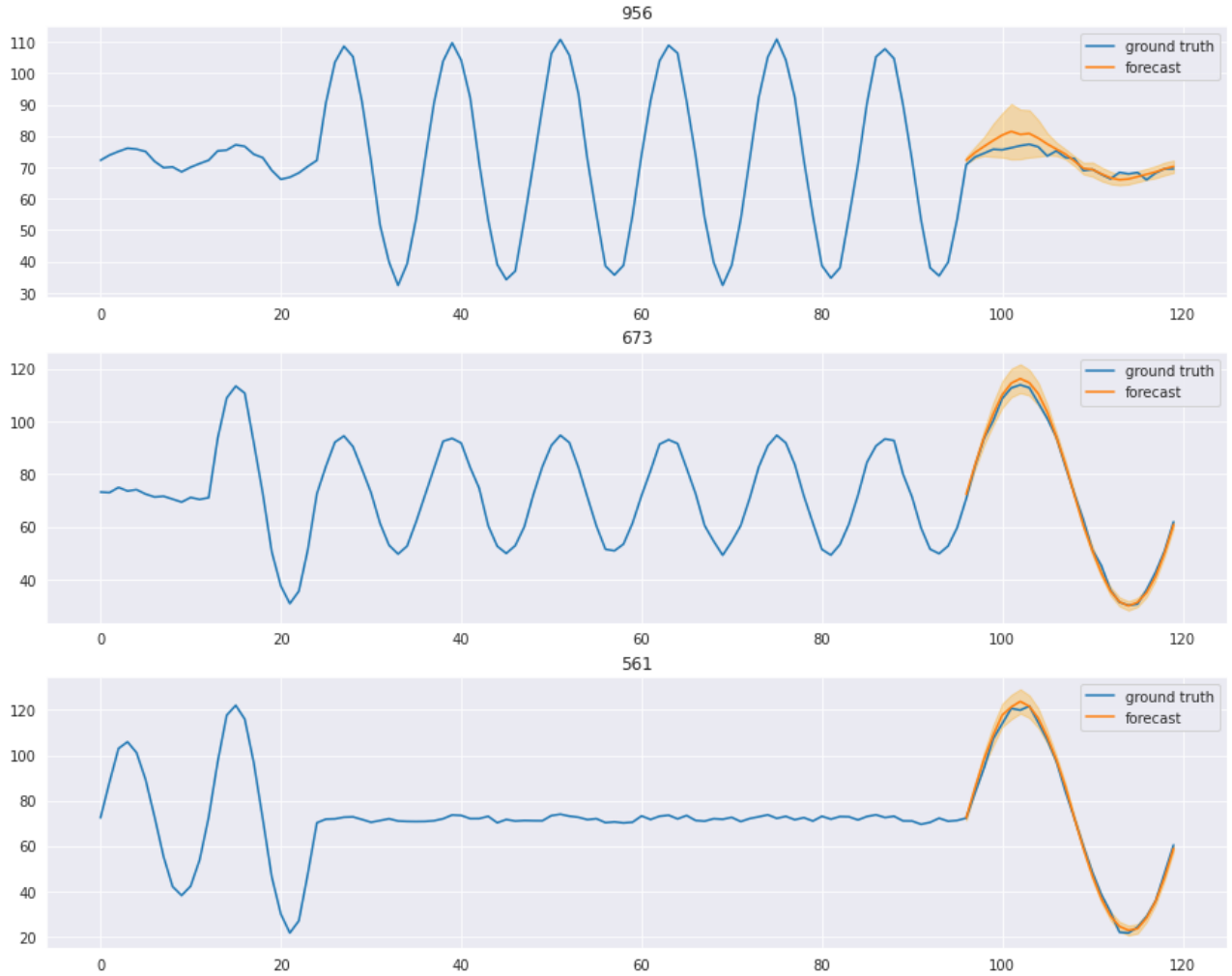


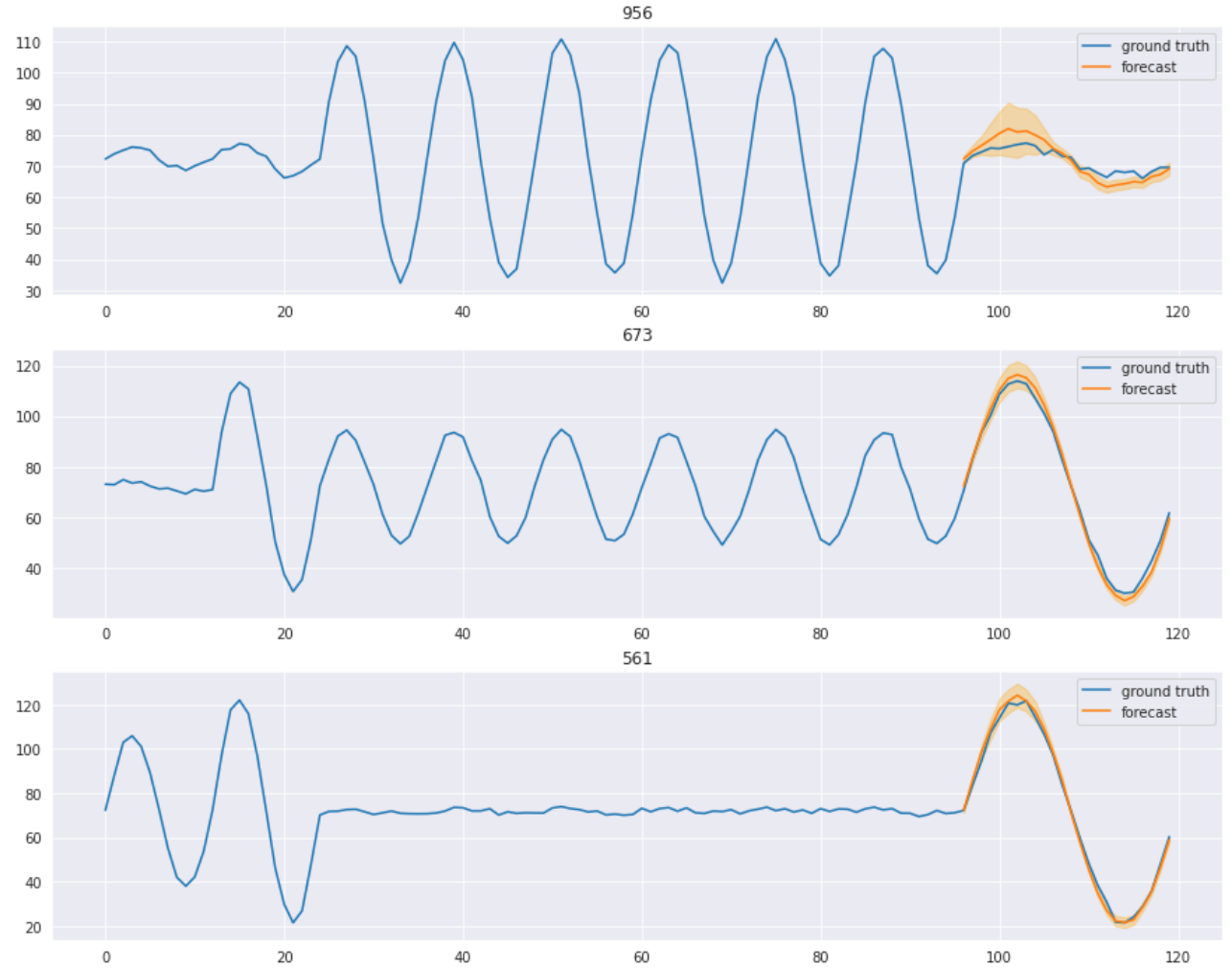
- Loss: Negative log likelihood
- Metric: MAE

Seems good 👍

Short-term

Long-term





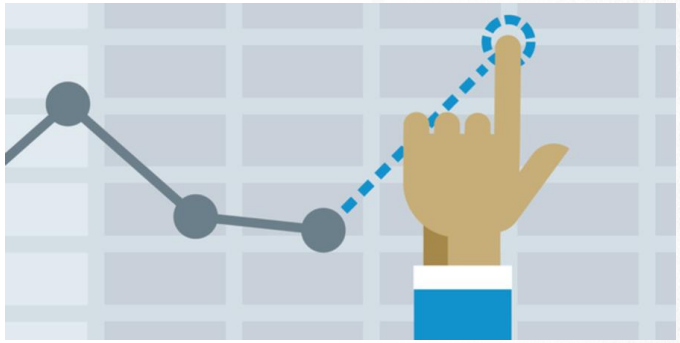
prediction quantile = 50

Synthetic dataset - summary

*Value for long-term forecasting

	Train (4.5K)	Val (0.5K)	Test (1K)
Loss (NLL)	1.16	0.88	0.88 (1.13*)
MAE	1.59	1.10	1.09 (1.39*)

- Overall good performance.
- Performance drops in long-term forecasting (would be surprising if it didn't), but not significantly.



More: Wavenet for time series embedding

Embedding before attention

- Information of a single pixel or time series value is very limited.
- Convolution is a common choice in vision.

➢ Xu et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML, 2015.

➢ Ascoli et al., “ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases”, ICML, 2021

- Causal convolution is a common choice in time series.

➢ Li et al., “Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting”, NeurIPS, 2019.

➢ Zhou et al., “Informer: Beyond efficient transformer for long sequence time-series forecasting”, Proceedings of AAAI, 2021.

Wavenet

- Oord et al., "Wavenet: A generative model for raw audio”, 2016.

Wavenet

Vanilla convolution

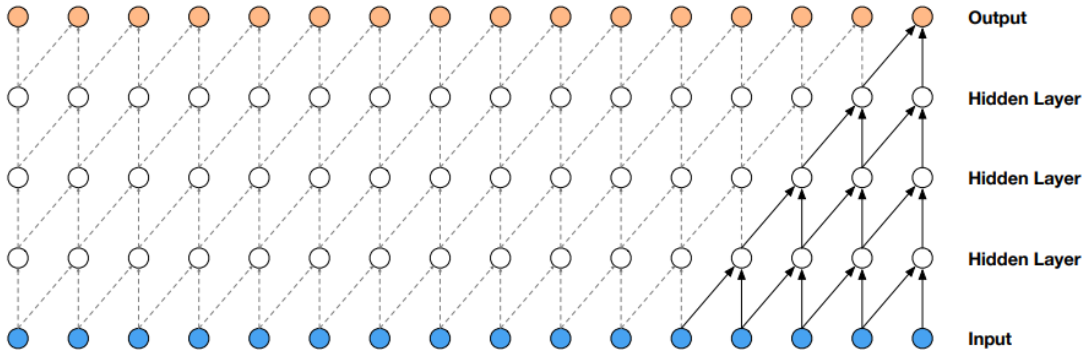


Figure 2: Visualization of a stack of causal convolutional layers.

Wavenet

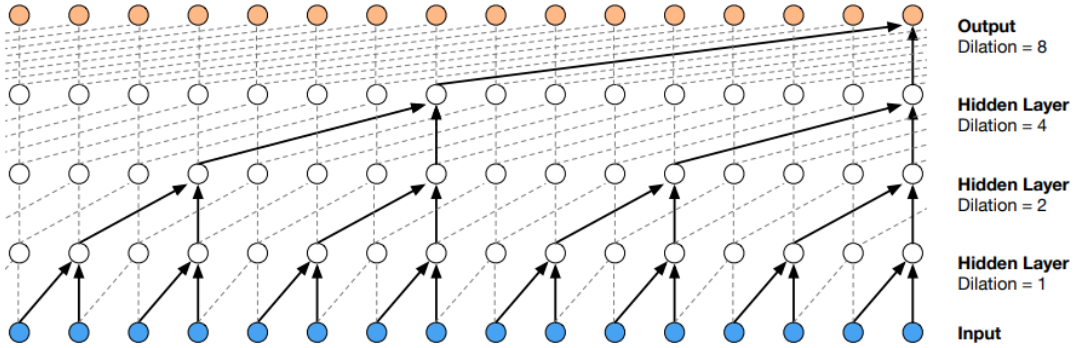


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

“enable networks to have very large receptive fields” without expense of complexity.

Wavenet for time series

- Motivation:

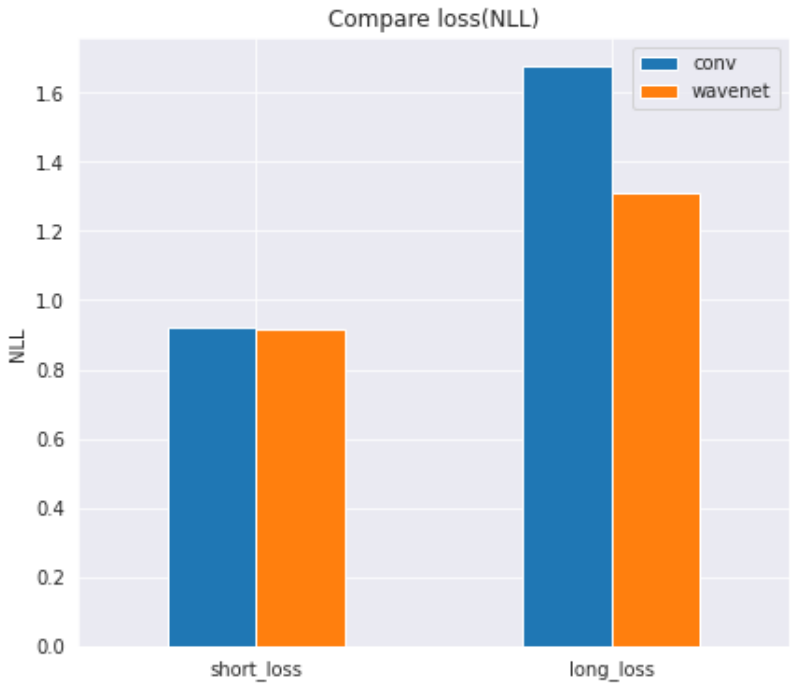
Audio could be seen as an extreme of time series, at least in an approach taken by wavenet.

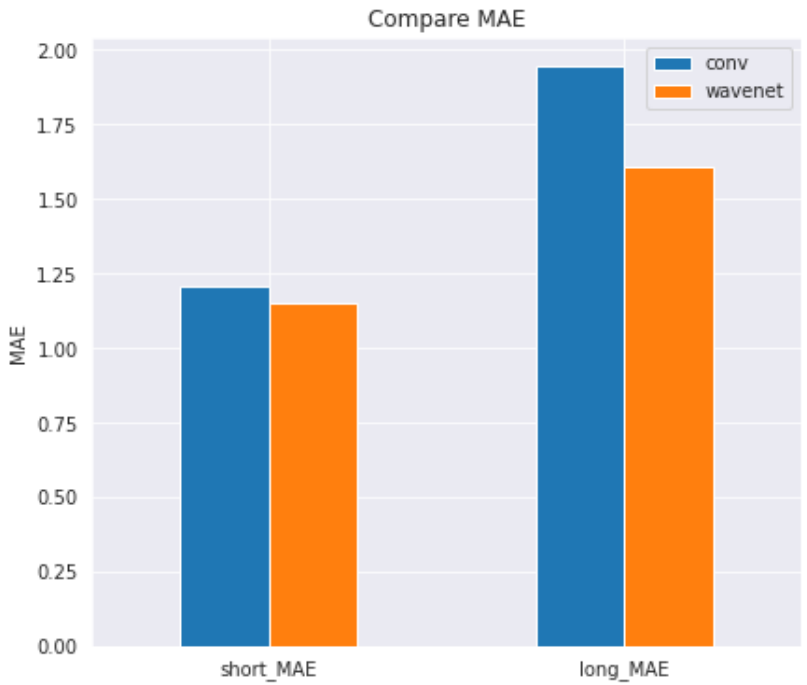
- Related work:

Couldn’t find in top-tier conferences, but few works exist on the Internet.

- Use synthetic dataset.

Ablation study on wavenet as time series embedding





conv: kernel size = 3, receptive field = 9;
wavenet: kernel size = 2, receptive field = 16

Share tuned hyperparameters. Mean of each top 5 runs.

Improves long-term performance(24%) while reducing the training time(7%).

Code

- https://github.com/Wittgensteinian/Transformer_for_time_series

Thank You!