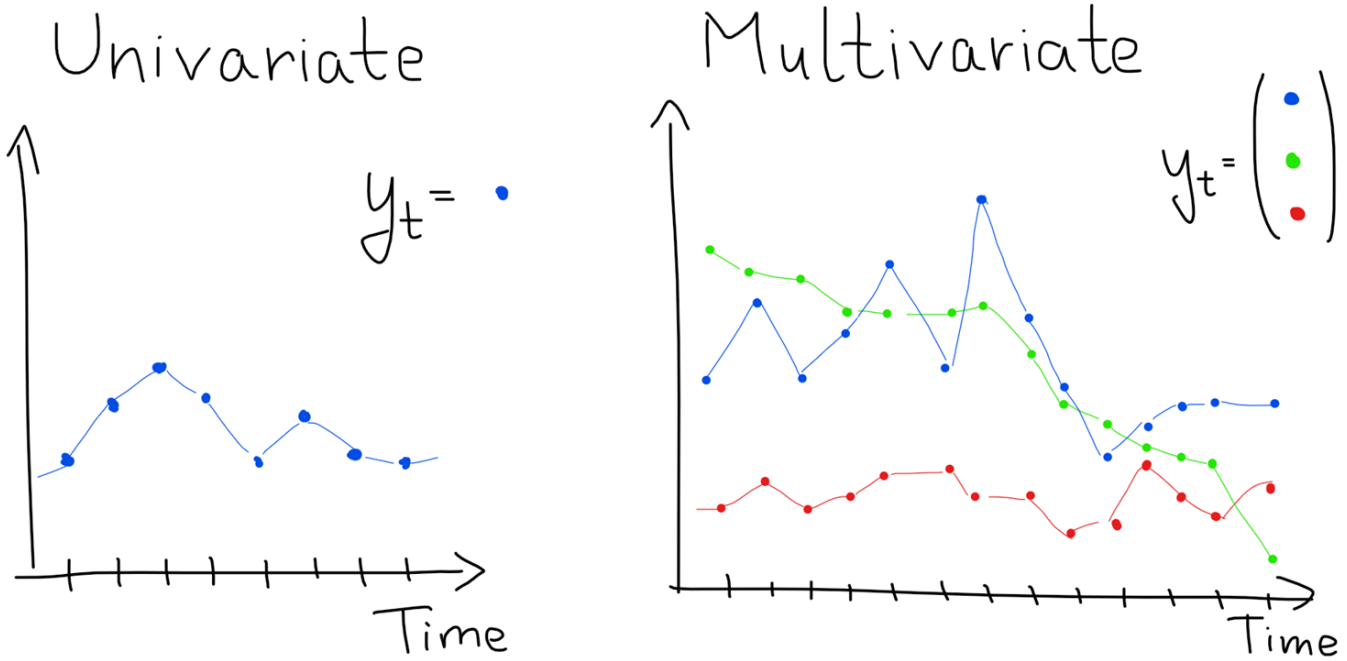
Simple method for long-term time series forecasting

directed by Misha @exxxplainer

Plan

- Intro & motivation
- Problem statement
- Tiny bit of history
- Embarrassingly simple solution
- How to live with it

Time series



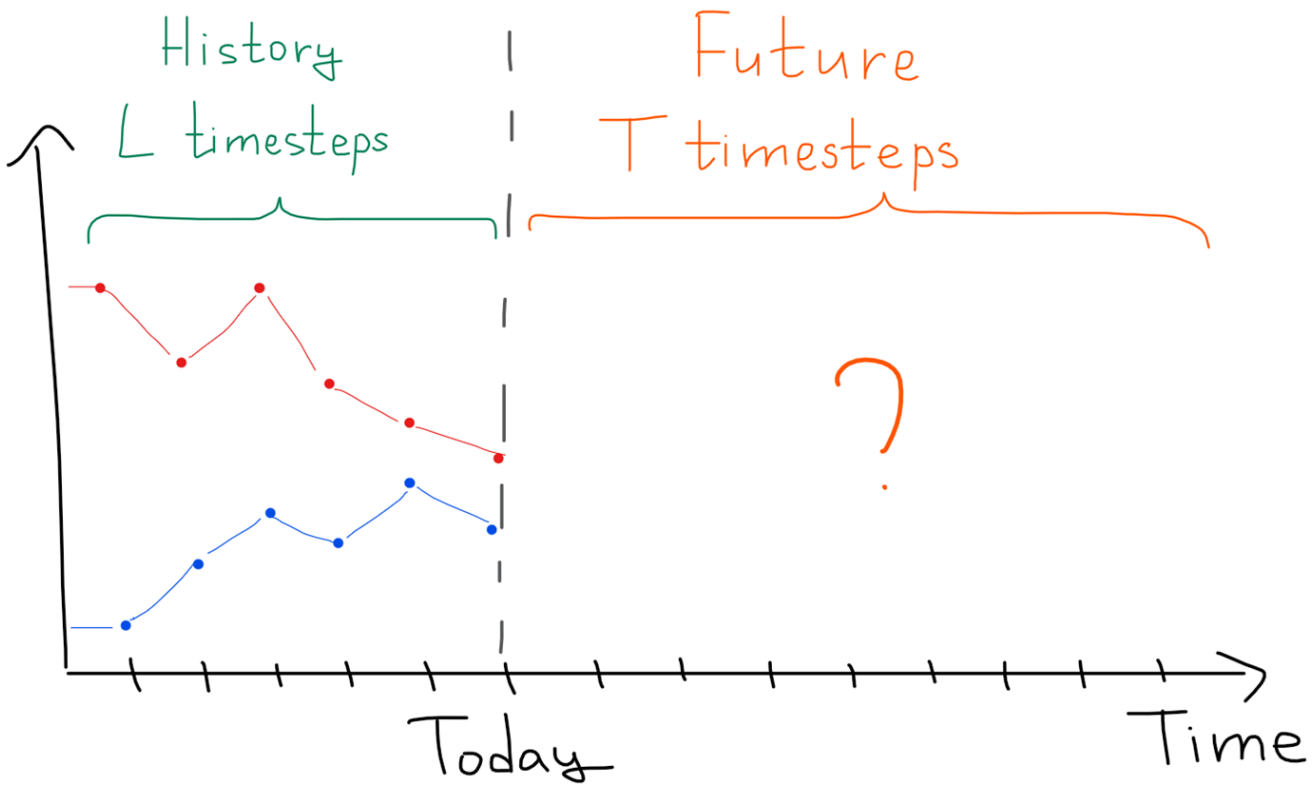
- Time series is a sequence {y_t}, t = 1, 2, …
- y_t could be vectors

Motivation

People want to forecast

- demand (Samokat, X5 retail group)
- traffic (Yandex Maps)
- weather
- diseases (million of papers, rarely actually works)
- …

Problem statement



How to solve?

Simplest

- Naive (Repeat)
- Seasonal Naive
- Drift
- Simple Moving Average
- Exponential MA

Statistical

- ARIMA
- Holt-Winters
-
- Theta
- TBATS
- GARCH

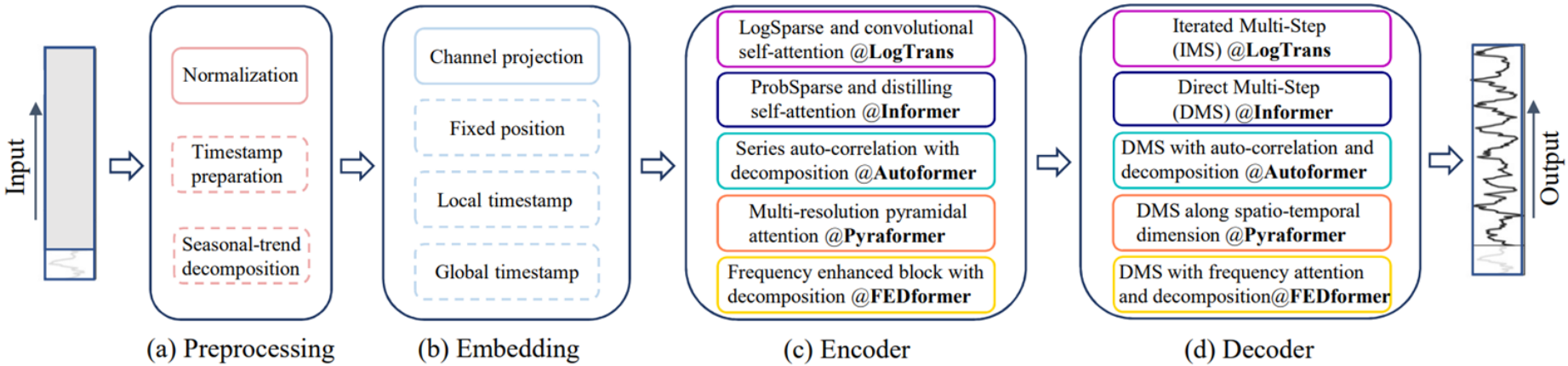
ML & DL

- Boostings
- MLP
- RNN
- Convolutions
- Transformers
- S4

Creative

- Conditional Diffusion
- Topological Data Analysis
- Neural ODE
- Convert time series to images

Transformers



Are Transformers Effective for Time Series Forecasting?

Authors propose 3 baselines:

- Linear
- DLinear
- NLinear

Linear layer is **shared between variates**.

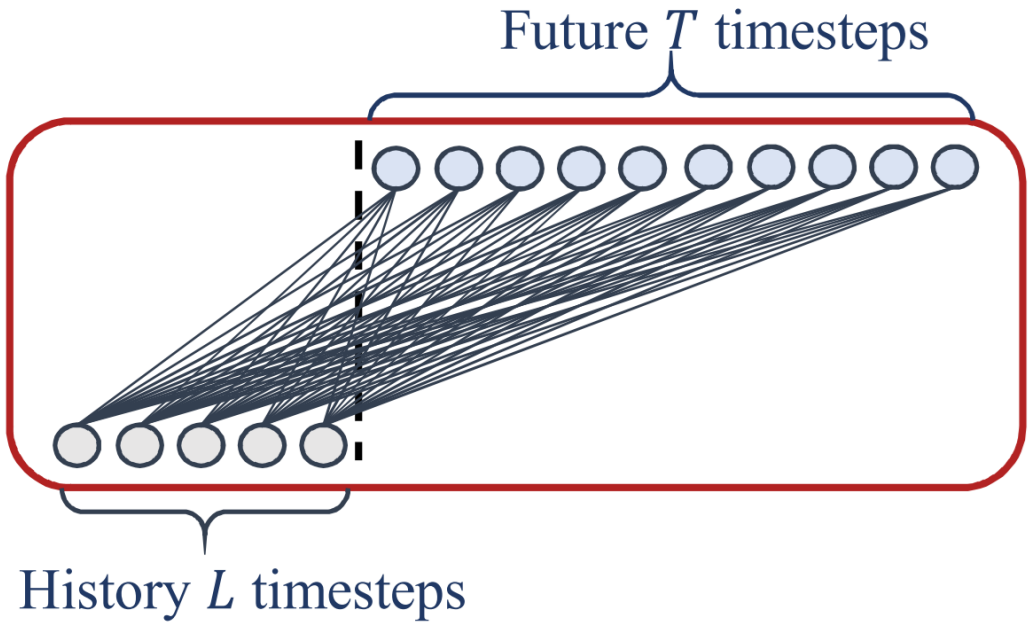


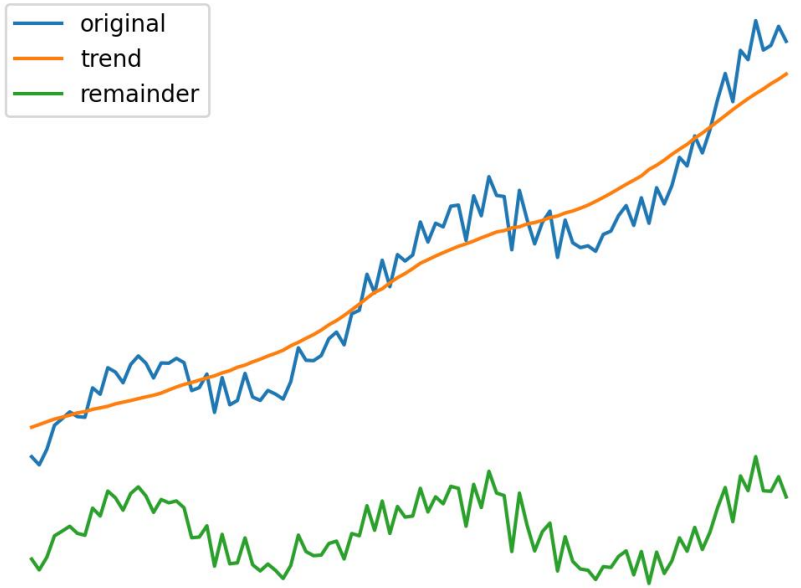
Figure 2. Illustration of the basic linear model.

DLinear

$$y_k = t_k + S_k + \varepsilon_k$$

- Decompose series into trend and remainder
- Train 2 Linear models, one for each component
- Sum predictions

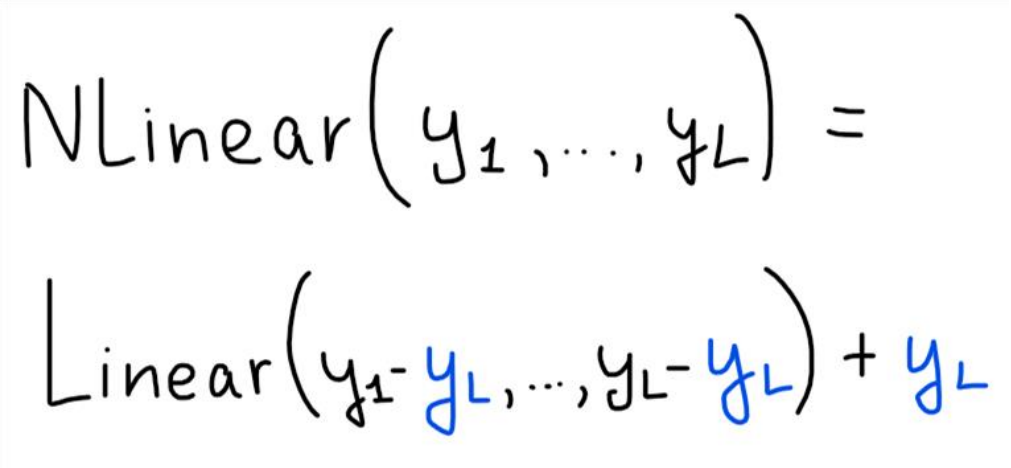
Helps when there is a clear trend in the data.



NLinear

- Subtract last value from inputs
- Pass through Linear model
- Add subtracted value back

Helps when there is a distribution shift.



Results

- Surpasses SOTA Transformers by 20-50% on 9 datasets
- x7-10 less VRAM
- x10-100 less parameters
- x10-100 faster

Some details

- They use MSE/MAE on normalized series
  - Probably not the best metric (however, I think, results will hold)
- The datasets aren't small
  - So don't forget about simplest baselines
- They essentially use StepLR schedule
  - Overall, if you plan to use them, check the repo

Datasets	ETTh1&ETTh2	ETTm1 &ETTm2	Traffic	Electricity	Exchange-Rate	Weather	ILI
Variates	7	7	862	321	8	21	7
Timesteps	17,420	69,680	17,544	26,304	7,588	52,696	966
Granularity	1hour	5min	1hour	1hour	1day	10min	1week

Table 1. The statistics of the nine popular datasets for the LTSF problem.

How to live with it

- Some stick to Transformers (they managed to catch on!)
  - patches & channel independence
  - more complicated attention & loss modifications (SOTA as of 20 May 2023)
    (about 5% better on small datasets, 10 and 15% better on big ones)
- Some boost Linear further
  - reversible normalization & non-shared layers on big datasets
    (win ~5% back)
- Some develop other models
  - Google published model based on MLP
    (beats DLinear, but worse than SOTA)

Unanswered questions

- Why Linear models are so strong?
  - Low-data regime?
  - Dependencies are really linear?
- Where do they fail? How to fix it?
  - Apparently, they struggle with trends and multiperiodicity
- Why treating channels independently is often better?
  - Overfitting?

Possible research directions

- Preprocessing
- Augmentations
- Pretraining
- Modeling inter-variate dependence
- Loss functions

Links

- Are Transformers Effective for Time Series Forecasting?
- A Time Series is Worth 64 Words: Long-term Forecasting with Transformers

Forecasting is trying to drive
by looking only at your car’s rear mirrors.