A Transformer Based Model for Time Series Forecasting in Finance

Master's Thesis, January 2024

Author: Jacob Thornam Eriksen
UNIVERSITY OF COPENHAGEN

Introduction

The paper tries to investigate the use of Transformer models as a time series tool for asset return prediction and compare these models with other benchmark models and indexes. Moreover, the paper suggests different functions to minimize over when training for the task of portfolio creation.

The process

- Daily return
- Classification Transformer
- 5 groups

- Weekly returns
- Regression Transformer
- Bloomberg data

Data

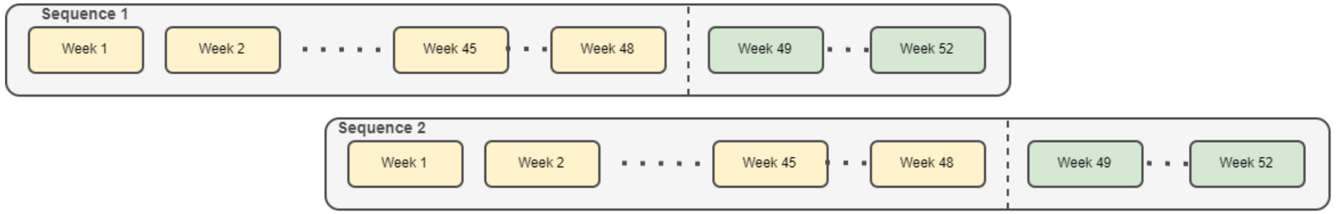
Training data

- Largest Market-Cap stocks
- 2012 – 2017
- ~ 2500 stocks

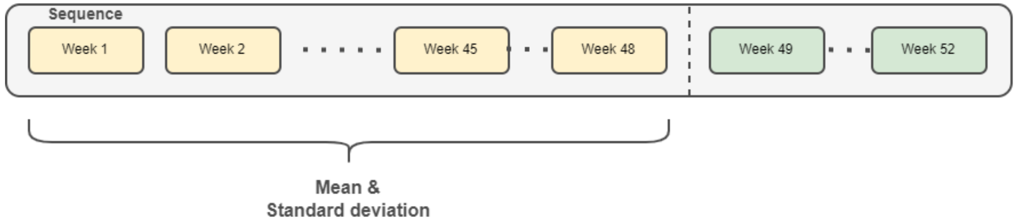
Test data

- S&P 500 stocks
- 2018 – 2023
- ~ 500 stocks

Sequences



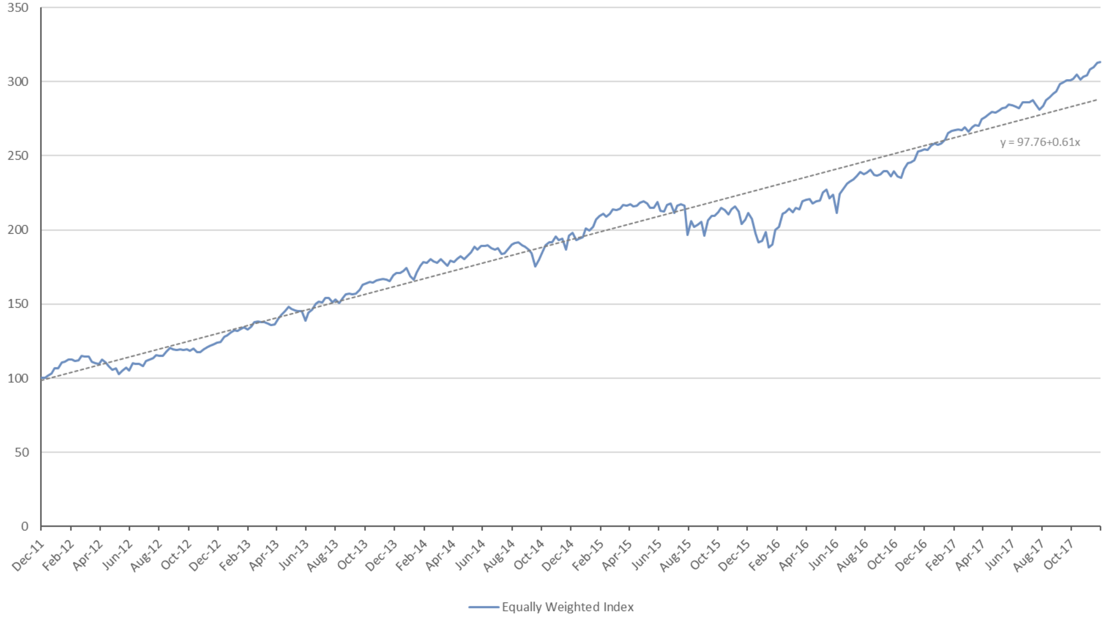
Scaling



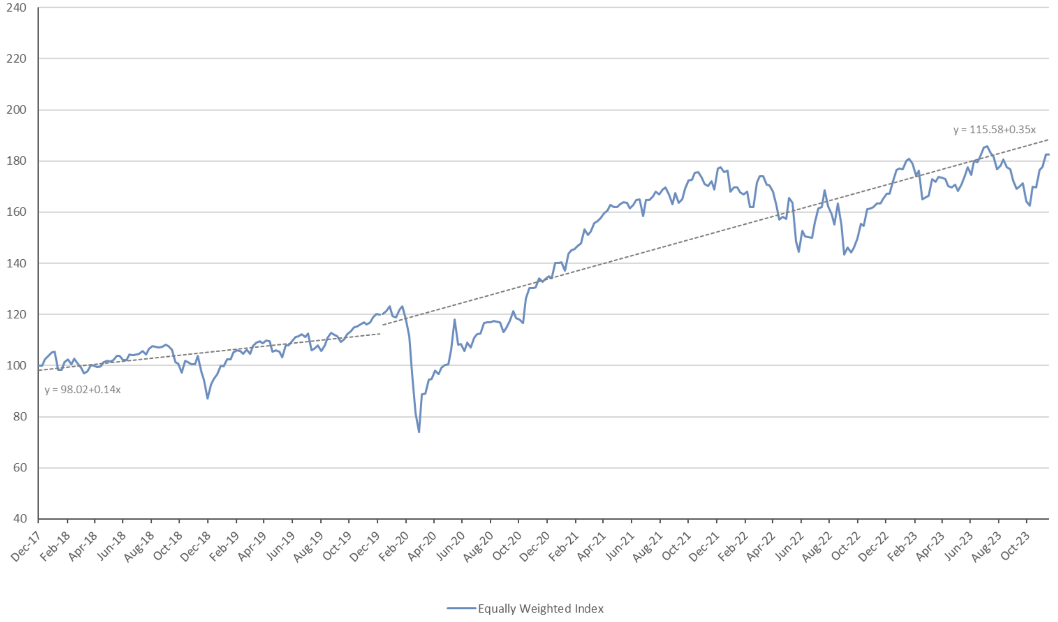
Data

Equally weighted index

Training period



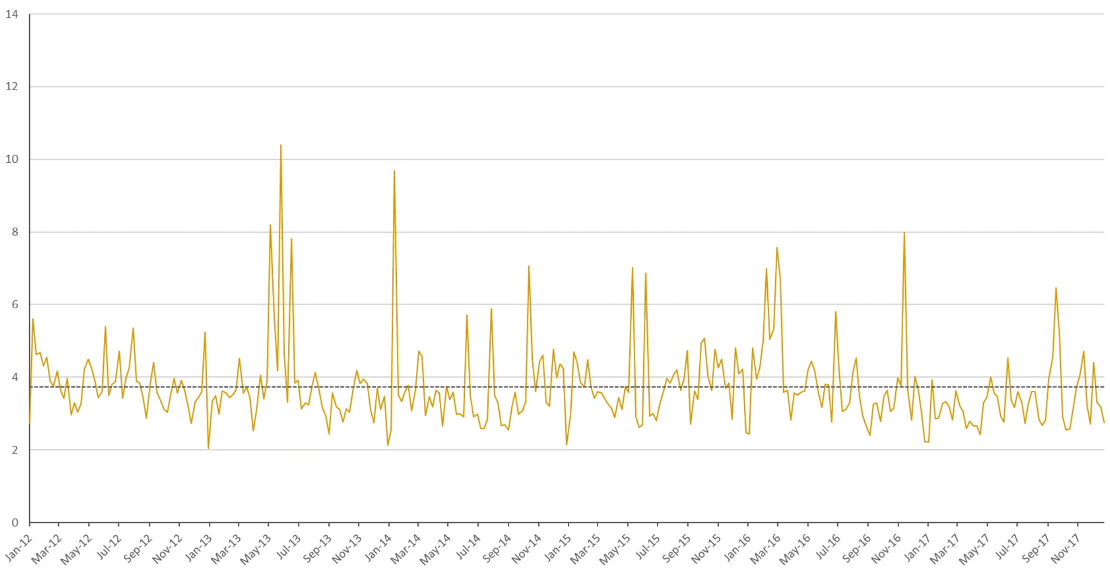
Test period



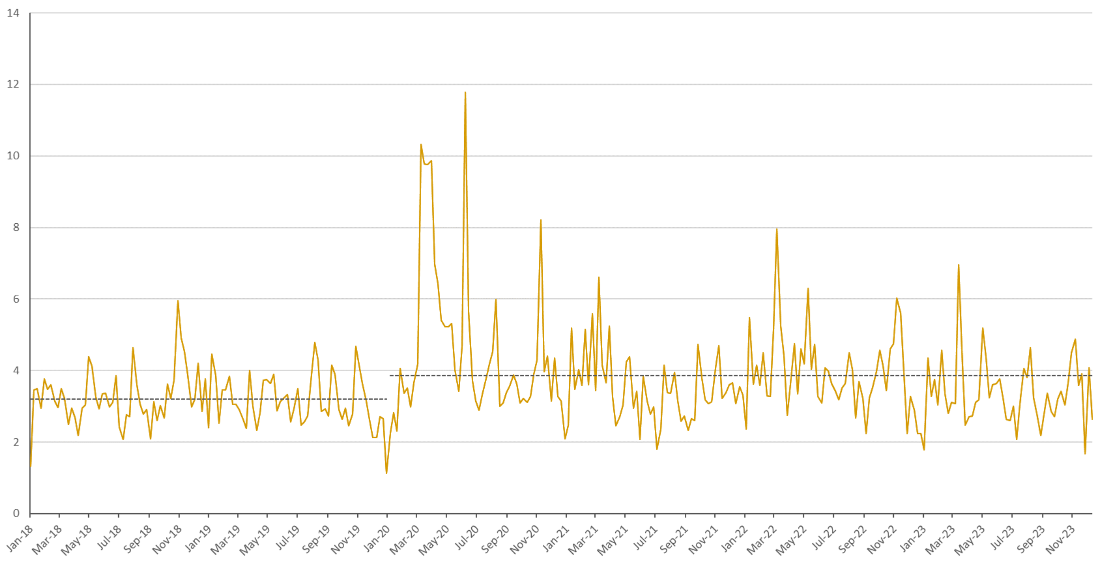
Data

Standard deviation of weekly returns

Training period



Test period



Data

Average positive and negative weekly returns

Training period

Test period

Transformer

Encoder

$$\tilde{X}_1 = X + PE(X)$$

$$\tilde{X}_2 = Normalization(\tilde{X}_1 + multihead(\tilde{X}_1))$$

$$\tilde{X}_3 = Normalization(\tilde{X}_2 + FFN(\tilde{X}_2))$$

Feed-Forward

$$FFN(X) = \mathrm{Max}(0, XW_1 + b_1)W_2 + b_2$$

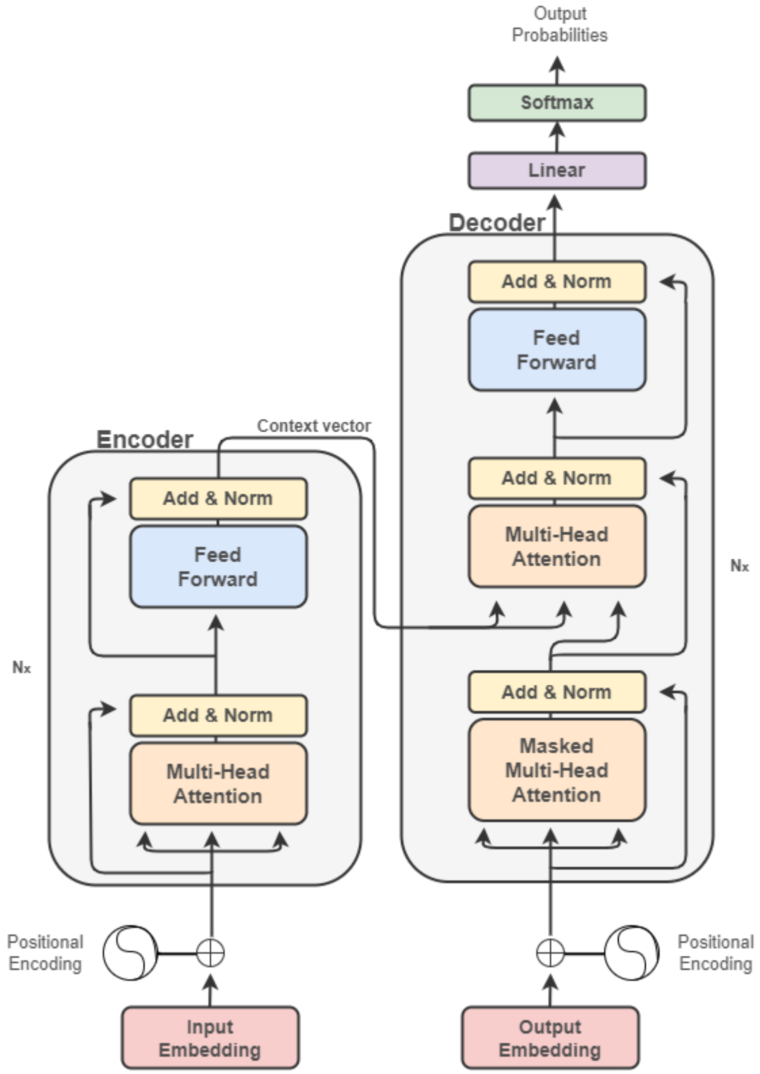
Positional Encoding

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

Linear Layer

$Linear(X) = XH^T \quad , \quad H \in \mathbb{R}^{1 \times d_{model}}$



Decoder

$$\tilde{Y}_1 = Y + PE(Y)$$

$$\tilde{Y}_2 = Normalization(\tilde{Y}_1 + masked\text{-}multihead(\tilde{Y}_1))$$

$$\tilde{Y}_3 = Normalization(\tilde{Y}_2 + multihead(\tilde{X}_3, \tilde{X}_3, \tilde{Y}_2))$$

$$\tilde{Y}_4 = Normalization(\tilde{Y}_3 + FFN(\tilde{Y}_3))$$

Multihead-Attention

$$Multihead\text{-}Attention(X) = concat(head_1, ..., head_h)W^O$$

$$where \quad head_i = Attention(Q_i, K_i, V_i)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_{model}}}\right)V$$

Embedding

$InputEmbedding(X) = XM \quad , \quad M \in \mathbb{R}^{Sequence\_length \times d_{model}}$

Multihead-Attention Layer

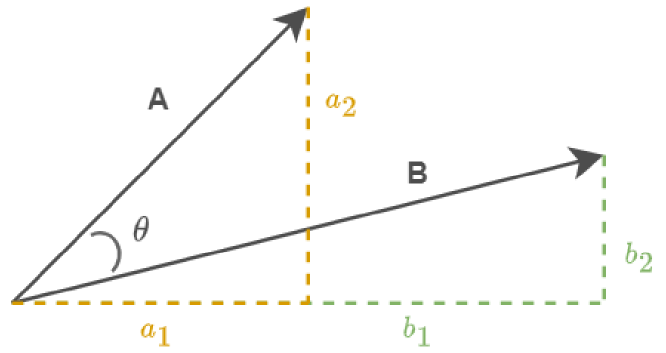
Math

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_{model}}}\right)V$$

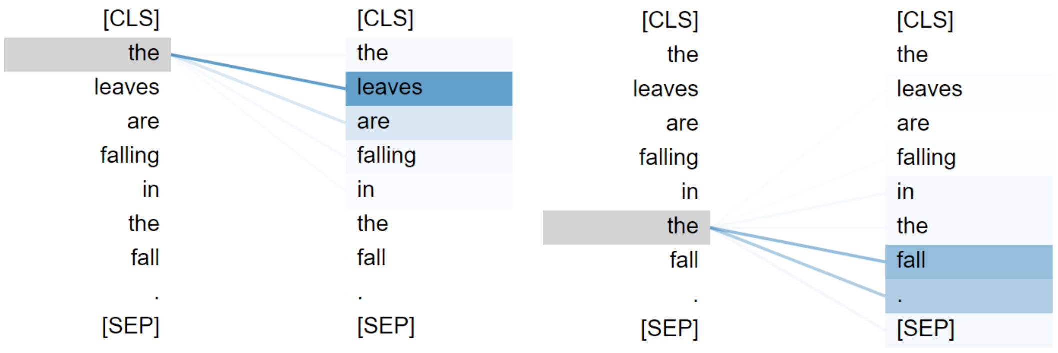
$$A \cdot B = AB^T = ||A|| * ||B|| * cos(\theta)$$

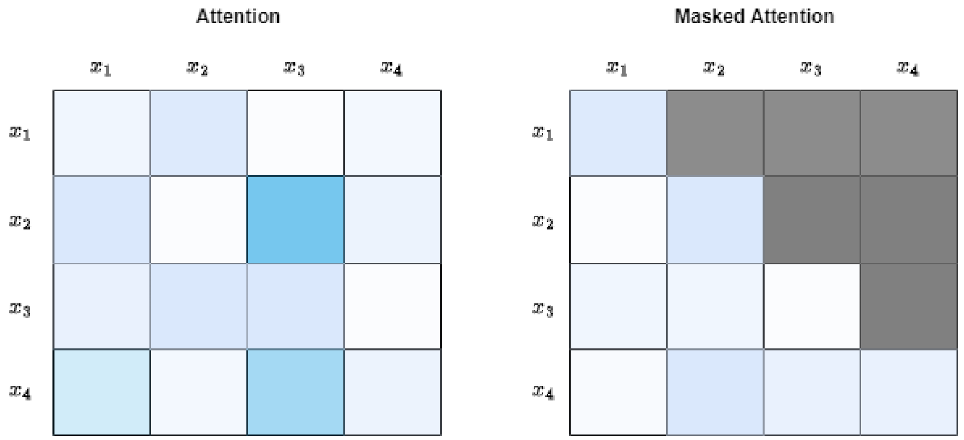
$$||A|| = \sqrt{a_1^2 + a_2^2}$$

$$||B|| = \sqrt{b_1^2 + b_2^2}$$



Intuition





Positional Encoding

$$PositionalEncoding(X) = X + PE(X)$$

The Original

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

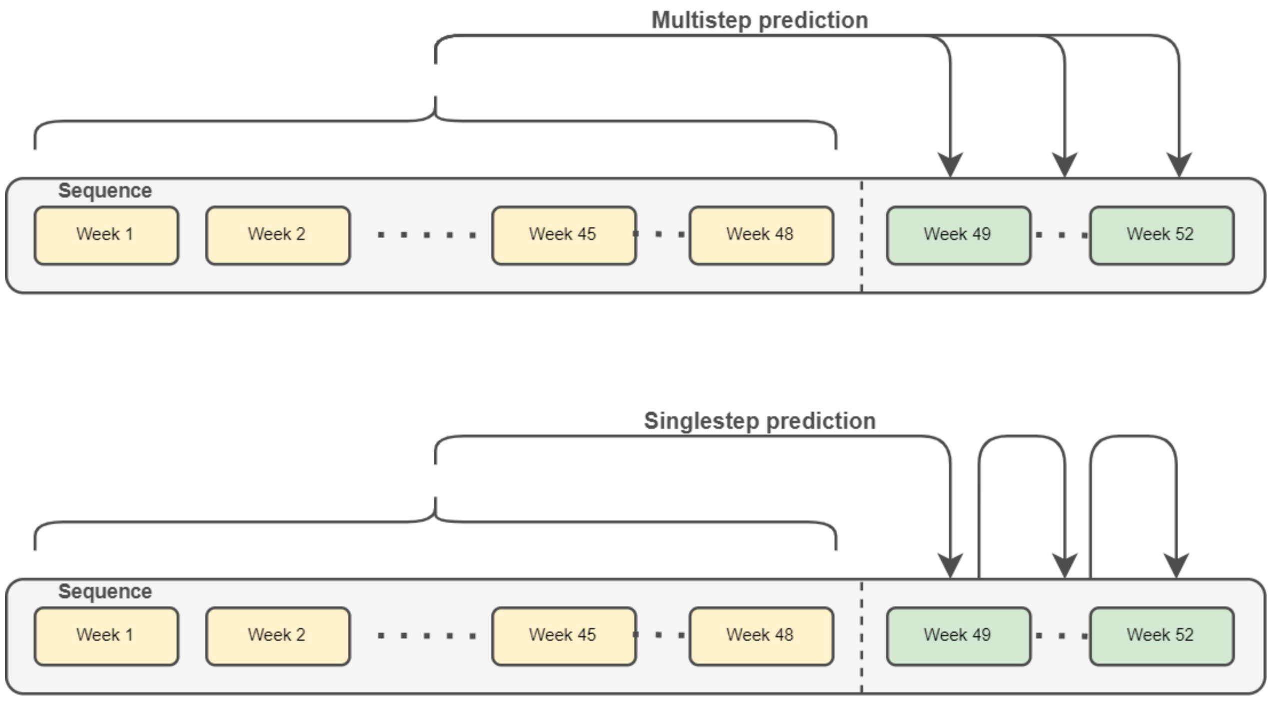
Time2Vector

$$t2v(\tau)[i] = \begin{cases} \omega_i \tau + \phi_i, & \text{if } i = 0 \\ Sin(\omega_i \tau + \phi_i), & \text{if } 1 \le i \le k \end{cases}$$

Model

Parameters

- 4 Encoder blocks
- 4 Decoder block
- 8 Heads
- d_model: 512
- Batch size: 128 (1024)
- Learningrate: 0.0001, decreasing
- 100 Epochs



Loss Functions

Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{it} - \hat{y}_{it})^2$$

Adjusted MSE

$$\text{AdjMSE} = \frac{a \times (y - \hat{y})^2}{1 + a - \frac{(a-0.5)}{1+e^{(100*y*\hat{y})}}}$$

With a equal 2.5

Weighted Mean Squared Error

$$WMSE(y, \hat{y}) = \frac{(1 + |y_i|)}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Negative Correlation

$$Corr(y, \hat{y}) = -\frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}}$$

Portfolios

Equally weighted Index

- All stocks
- Weight $\frac{1}{n}$
- Rebalance every 4th week

Momentum Portfolios

- 20 top stocks
- Based on 3-, 6- and 11-month average weekly return
- Rebalance every 4^{th} week

Forecasting Portfolios

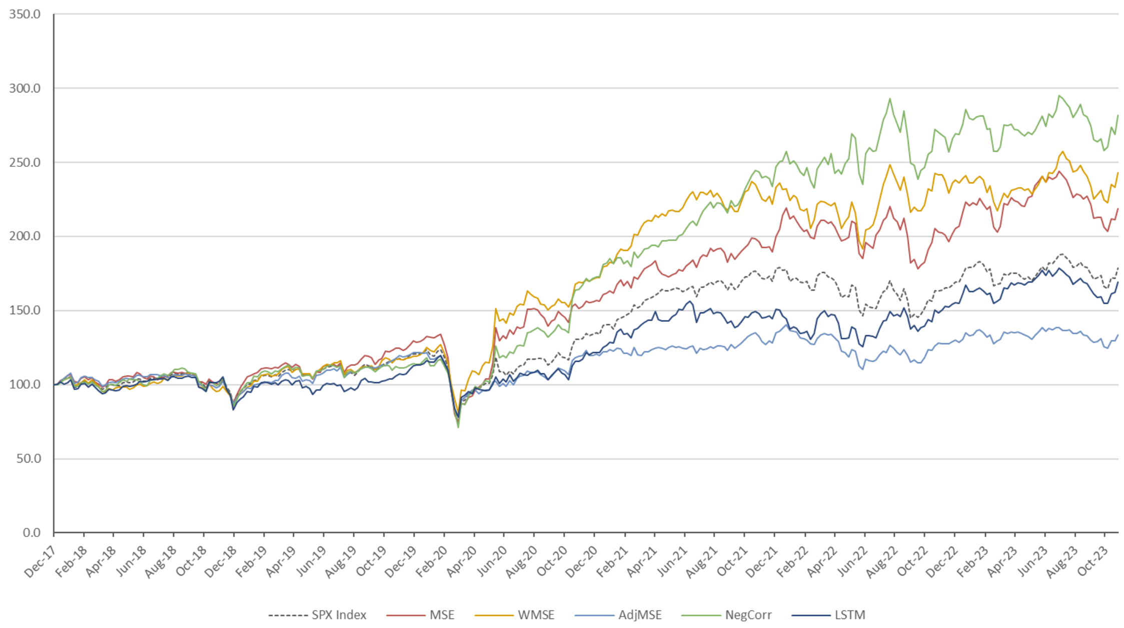
- 20 top stocks
- Based on model forecasts
- Rebalance every 4th week

- Model MSE
- Model AdjMSE
- Model WMSE
- Model NegCorr
- Model LSTM

Results

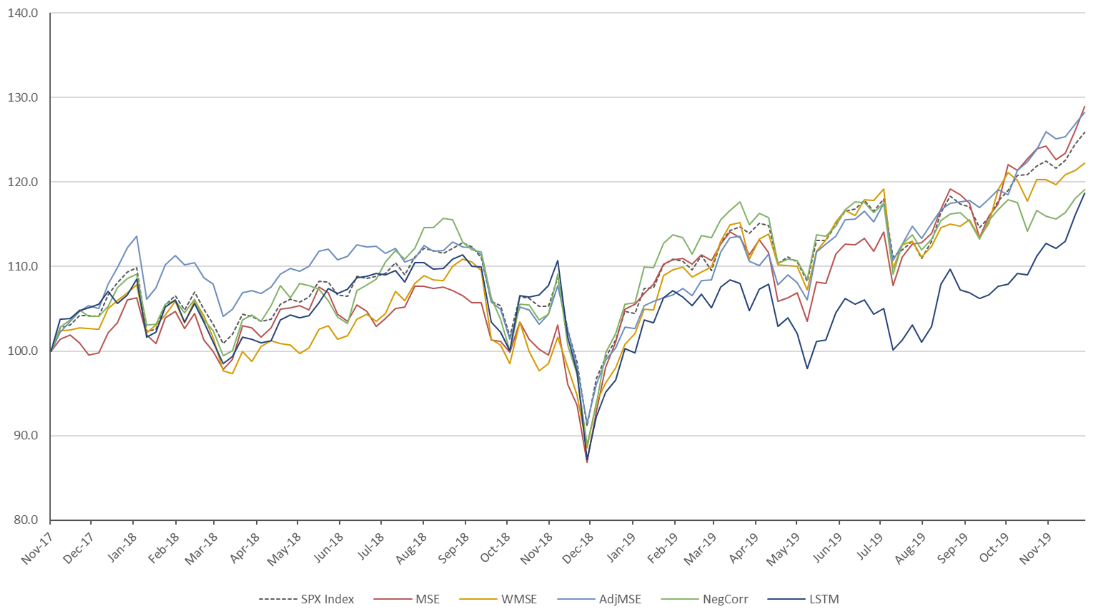
Momentum Portfolios

Forecasting Portfolios

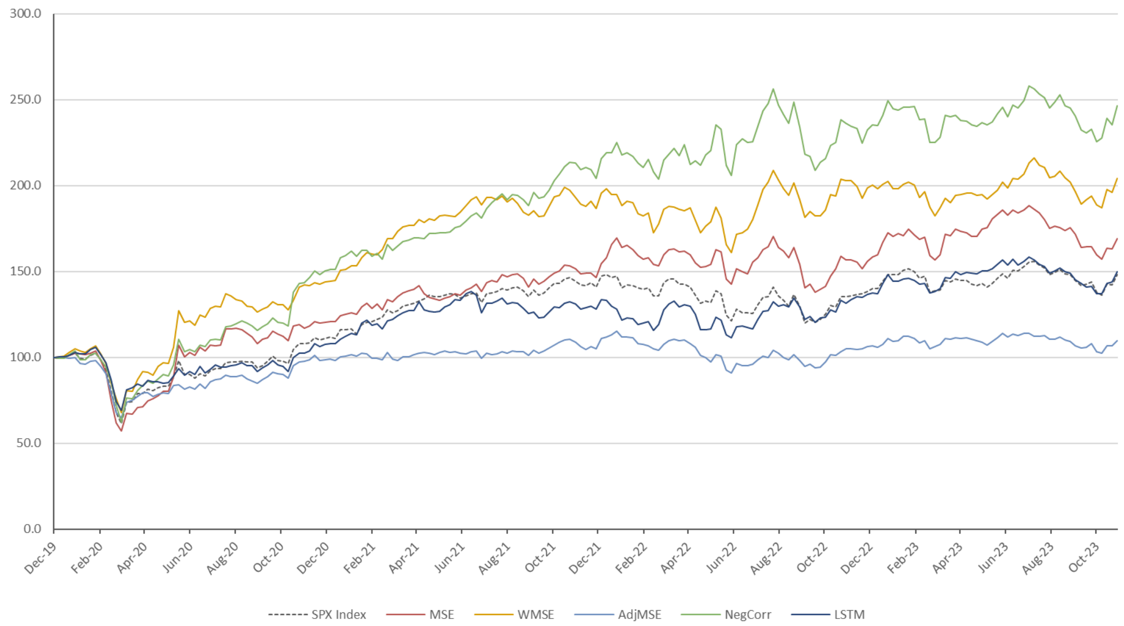


Results

Period 1

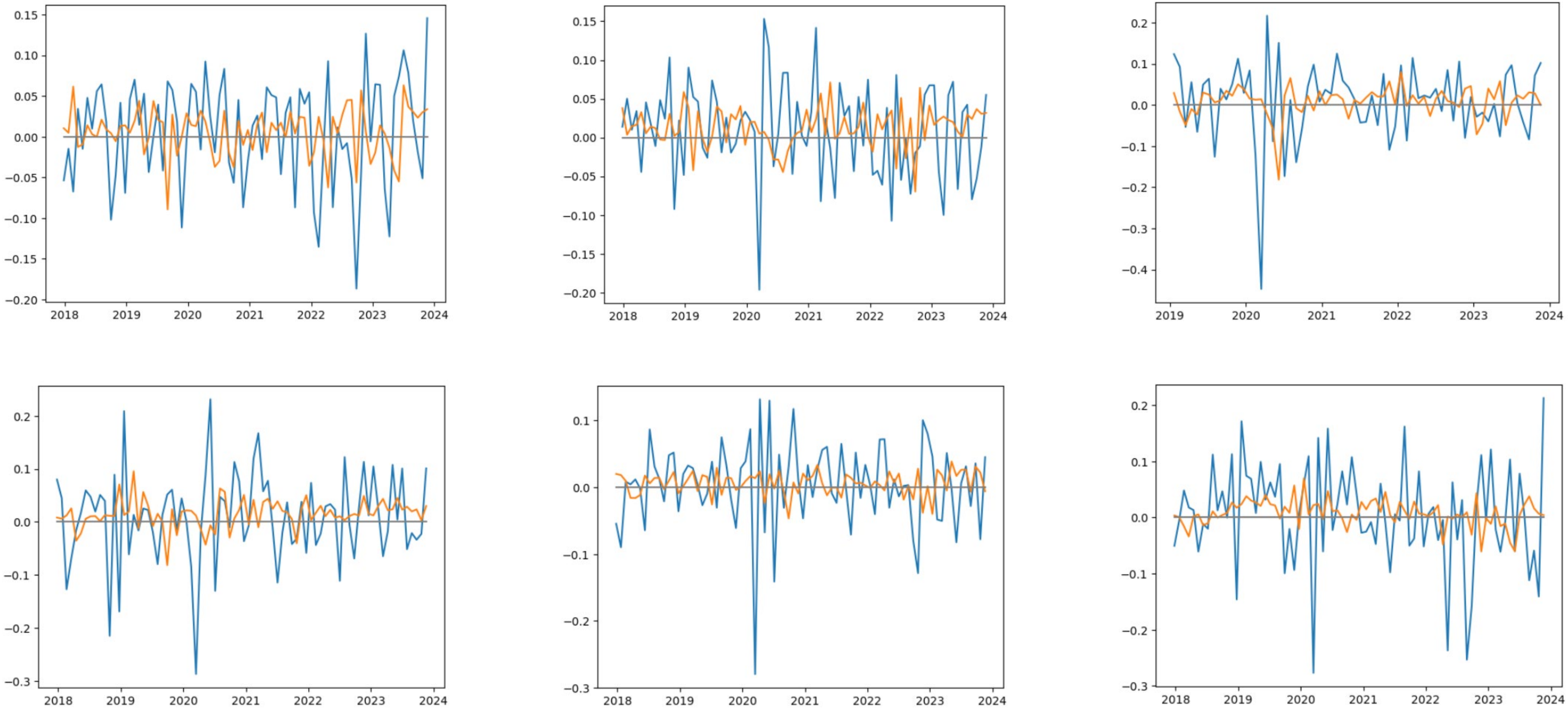


Period 2

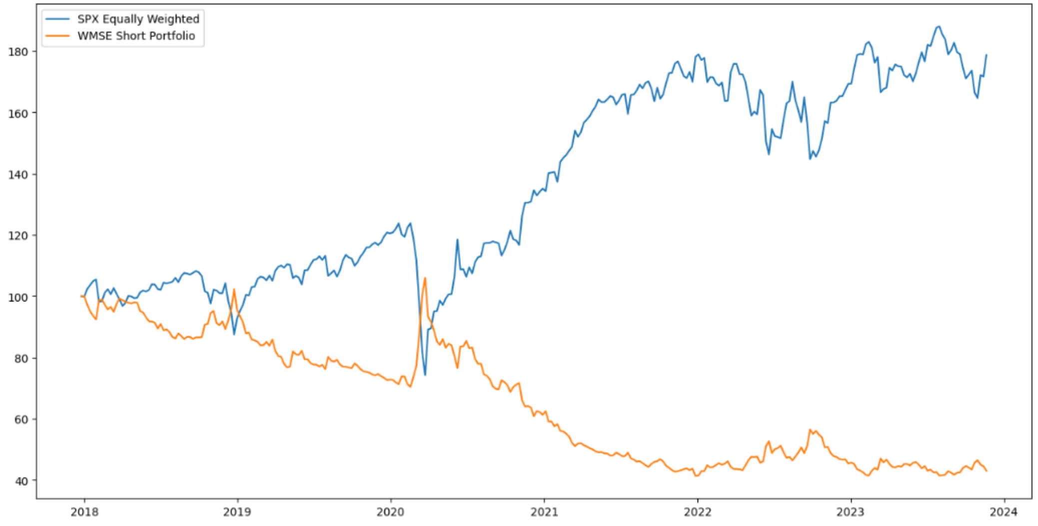


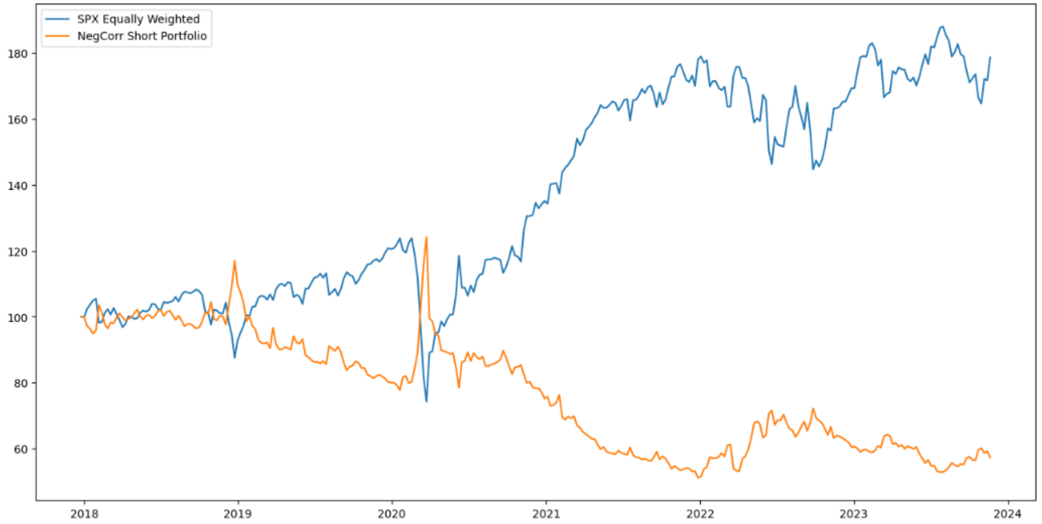
Presentation Appendix 1

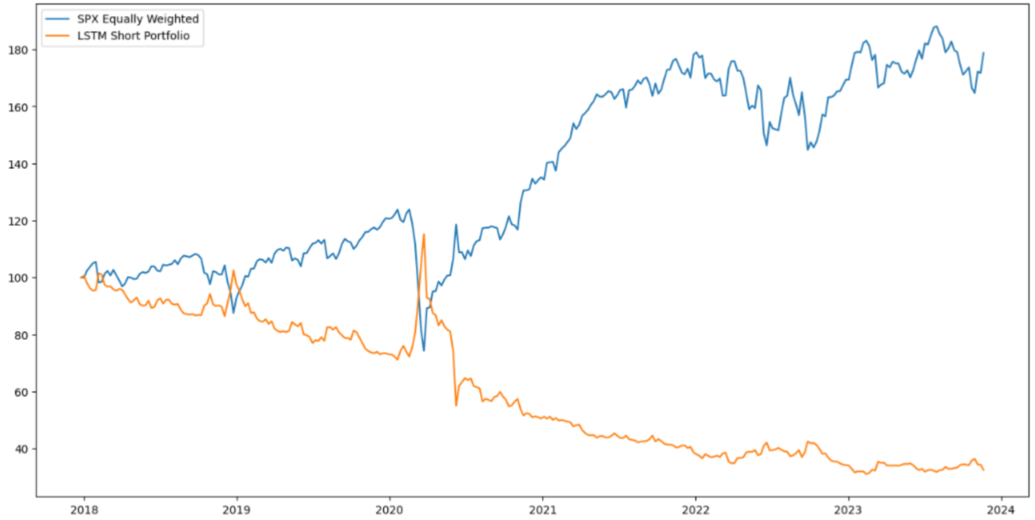
4 weeks real and predicted returns



Presentation Appendix 2



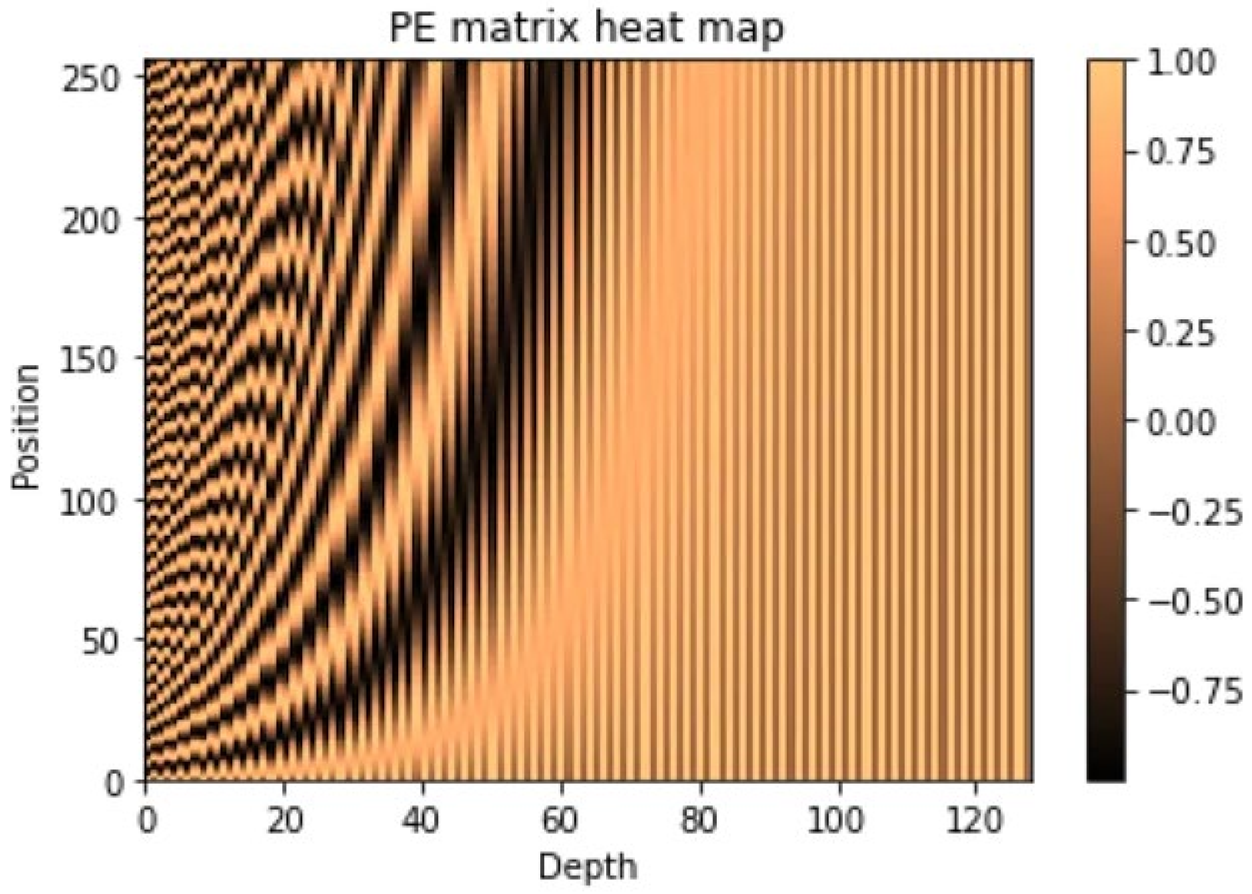




Presentation Appendix 3

$$\mathbf{z}_i = \sum_j \mathrm{softmax}\left(\frac{\mathbf{q}_i^T \mathbf{k}_j}{\sqrt{d_q}}\right) \mathbf{v}_j.$$

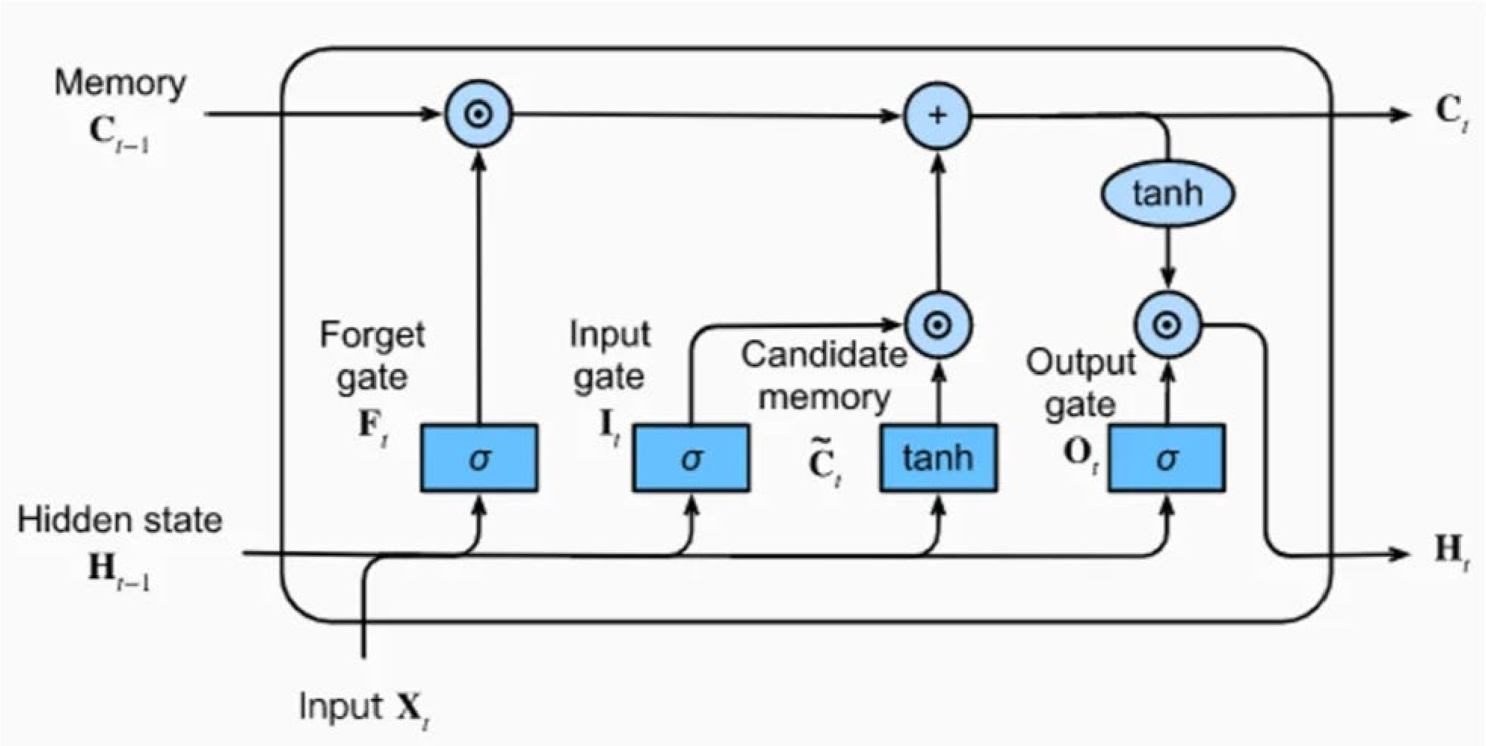
Ref.: Ahmed et al., 2022



Ref.: https://towardsdatascience.com/master-positional-encoding-part-i-63c05d90a0c3

Presentation Appendix 4

LSTM Model



Ref.: https://d2l.ai/chapter_recurrent-modern/lstm.html