Is attention all you need for intraday Forex trading?

Przemysław Grądzki and Piotr Wójcik

Monthly joint seminar of QFRG and DSLab
21st November 2022

UNIVERSITY OF WARSAW

UNIVERSITY OF WARSAW
Faculty of Economic Sciences

Agenda

- Research objective
- Literature review
- Data
- Neural network architectures
- Results
- Conclusions

Research objectives

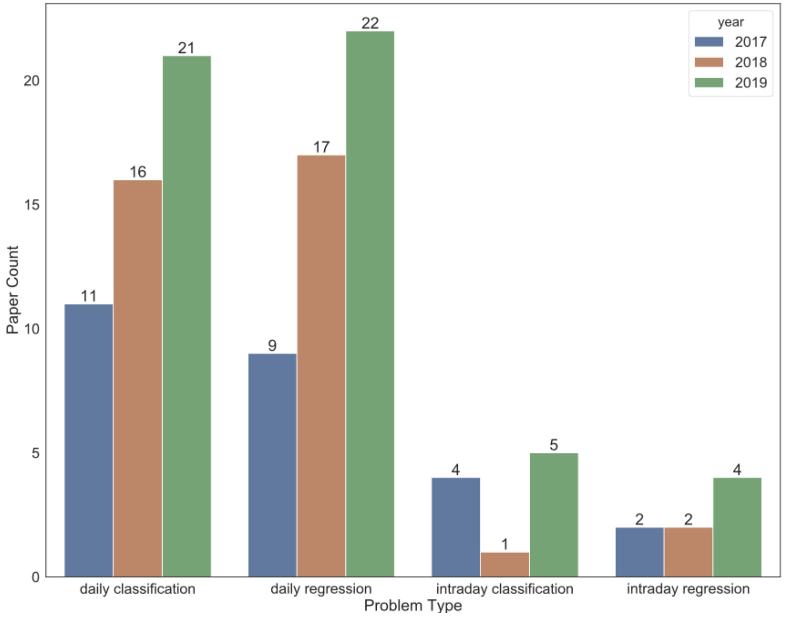
Goal

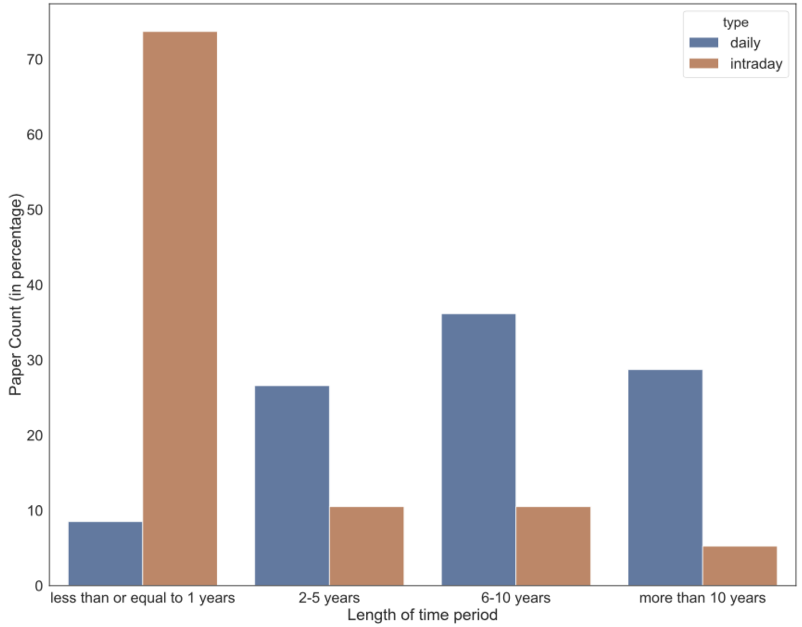
- The **main objective** of our research is to analyze if the Transformer network exhibits predictive capabilities in the context of the Forex market and if so, whether it can outperform other deep learning algorithms that are commonly used in recent financial literature.
- Given that intraday Forex trading based on deep learning is scarcely researched, we first establish what is the state-of-the-art benchmark algorithm that will be compared against the Transformer network.
- Due to more noise in the data of high frequency, as well as the higher impact of transactional costs, we formulate the **research hypothesis** that longer intervals (lower frequency of data) should lead to better trading performance. The analyzed frequencies range from one data point each 60 minutes to one data point each 720 minutes (12 hours).

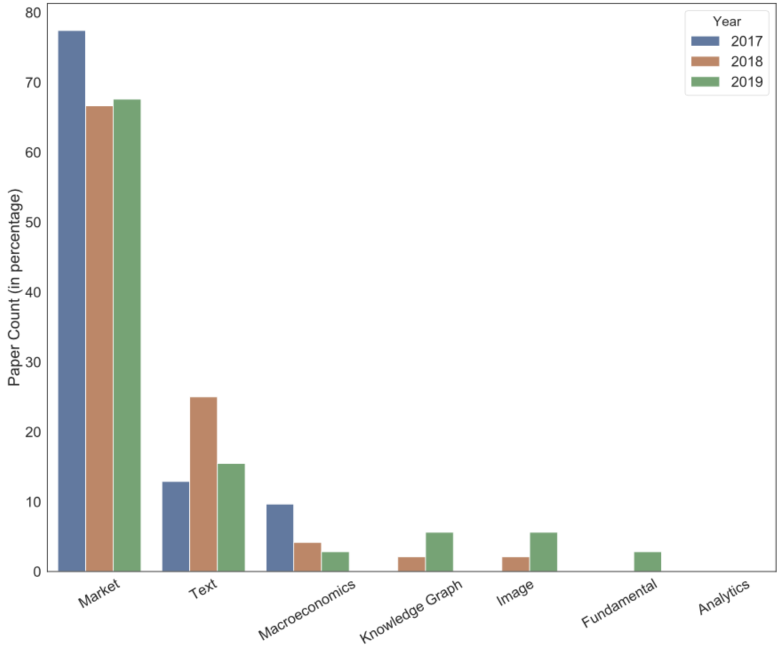
Motivation

- Transformer application in trading remain unexplored. Given its spectacular achievements in different domains focused on NLP, speech recognition, and computer vision, one might expect them to outperform current state-of-the-art deep learning models for trading
- intraday trading on the Forex market based on deep learning algorithms and different time frequencies is not well researched

Literature review - Data







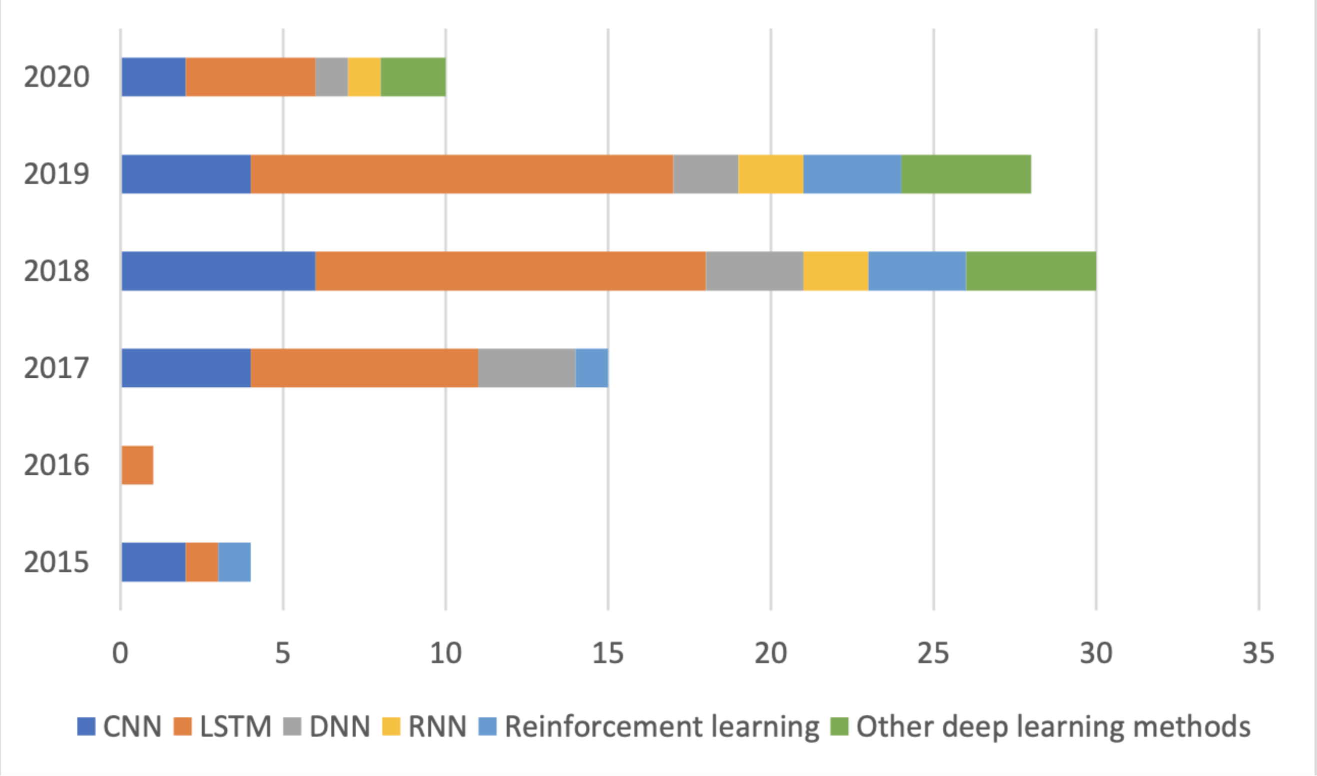
Source: Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. Expert Systems With Applications

Observations:

- Daily data or short period of intraday data may not fully utilize the power of deep learning models that typically shine when provided with a high volume of data.
- Decisions about time aggregation used in the research will often be predicated on data availability, and this leads to a void when it comes to comparison of different time intervals and their predictability
- Researchers most commonly rely on daily data for the stock market and leave other markets overlooked (e.g. Forex)
- Out of dozens of articles evaluated by Hu et al. (2020) only four of them analyzed the forex market and only one considered trading results as a performance metric.

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Literature review - Algorithms



Source: Hu, Z., Zhao, Y., & Khushi, M. (2021) A Survey of Forex and Stock Price Prediction Using Deep Learning. Applied system innovation.

Even for one class of algorithm implementation options are limitless. Example for CNNs:

- Network architecture: vanilla, LeNet, AlexNet, VGG, Inception, Resnet etc.
- Depth on neural network
- 1D, 2D, 3D convolutions
- Number of filters and their size
- Pooling layers
- Size of an ‘image’
- Regularization
- Batch size
- Learning rate
- Optimizer
- others

Look at the data

Distribution of target variable (share of "buy") in training dataset

Interval	EURGBP	EURCHF	EURUSD	EURPLN	USDJPY	USDCHF
60 min	0.49	0.49	0.49	0.49	0.50	0.50
120 min	0.49	0.49	0.49	0.49	0.50	0.50
240 min	0.50	0.49	0.49	0.49	0.50	0.50
480 min	0.50	0.50	0.49	0.49	0.51	0.50
720 min	0.49	0.50	0.50	0.48	0.51	0.50

Fraction of movements of magnitude lower than transactional cost

Interval	EURGBP	EURCHF	EURUSD	EURPLN	USDJPY	USDCHF
60 min	31.5%	38.1%	17.3%	72.3%	18.5%	26.7%
120 min	22.9%	29.7%	12.3%	61.9%	13.2%	19.3%
240 min	16.4%	22.4%	8.4%	50.3%	9.0%	13.4%
480 min	10.5%	16.2%	5.3%	27.5%	5.9%	8.8%
720 min	7.7%	13.3%	4.1%	29.3%	4.6%	6.5%

Train/ Validation/ Test split:

- Train: 2010-01-01 to 2019-12-31
- Validation: 2020-01-01 to 2020-12-31
- Test: 2021-01-01 to 2021-12-31

Sequences:

- Training length: 96
- Forecast horizon: 1 (classification of price going up or down)

Cost assumption - spread in pips:

- EURCHF: 2
- USDCHF: 2
- EURGBP: 2
- EURUSD: 1.5
- USDJPY: 1.5
- EURPLN: 30

VARSOVIENSIS

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

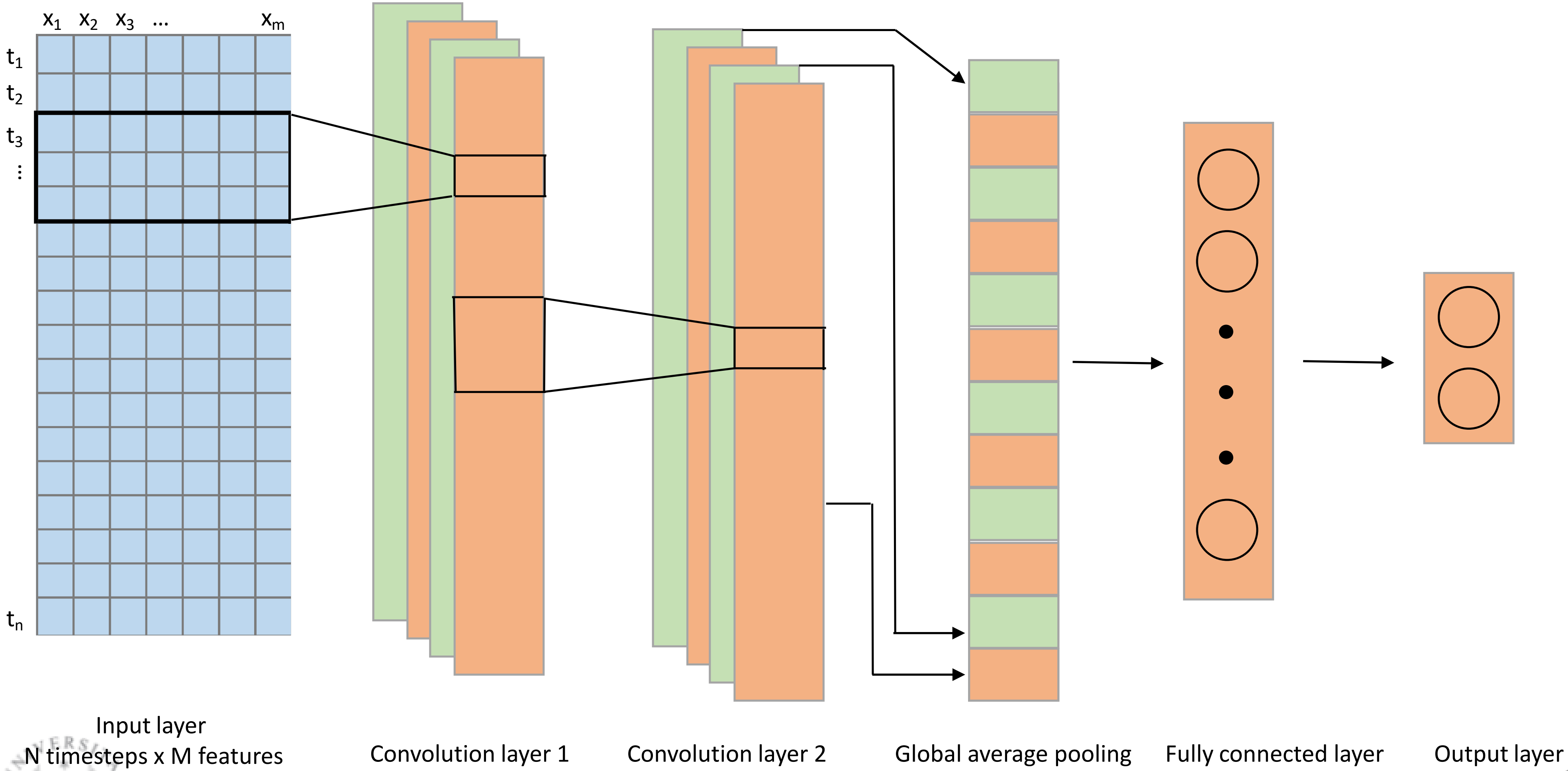
Feature engineering

- Price: open, close, high, low
- Simple, exponential moving averages and standard deviation for close prices of different lengths
- Stochastic Oscillator
- RSI (relative strength index) for length 14, 10
- MACD (Moving Average Convergence / Divergence)
- Williams %R
- Bollinger bands for 2 standard deviations and period length 5
- Historical returns between consecutive periods
- Time transformations: sin and cos of hour and weekday

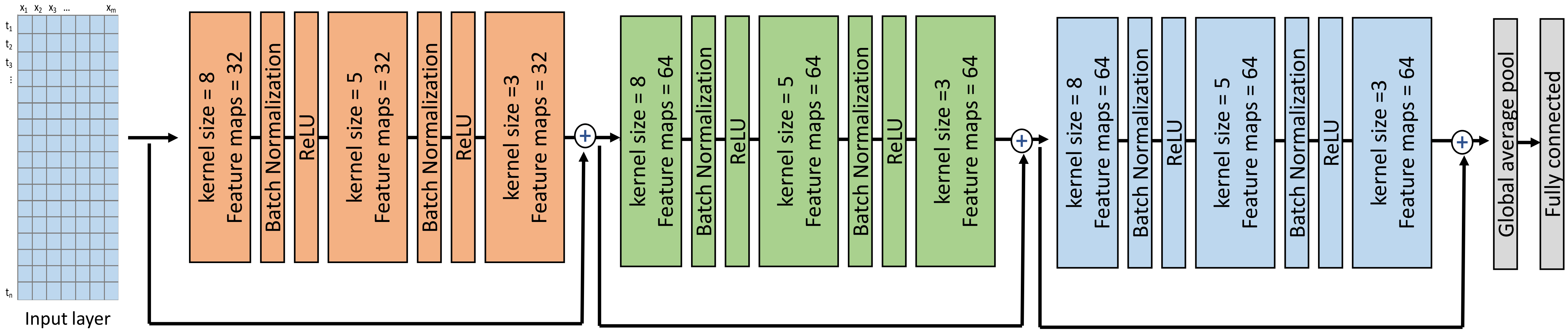
In total we used 40 different features, hence one training sample has dimension of 96x40. All features before entering the model are standardized by removing the mean and scaling to unit variance.

Source of data: histdata.com

CNNs 1D are emergent time series approach

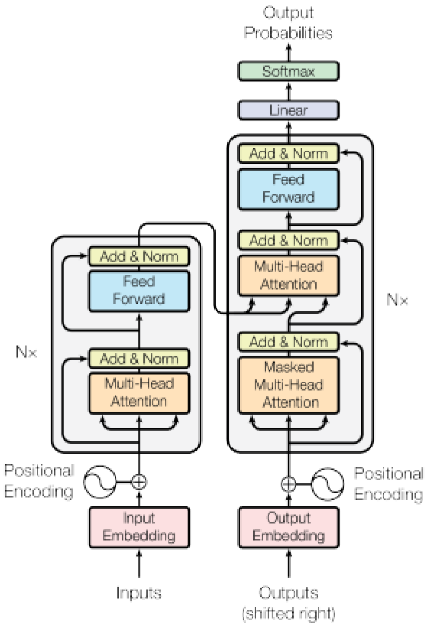


Resnet 1D – benchmark model



Transformers – brief intro

Transformers



Source: Vaswani et al. (2017) Attention is All you Need

State-of-the-art results

- Natural language processing:
    - GPT2, GPT3
    - BERT
- Protein folding - AlphaFold 2
- Image classification – Vision Transformer
- DALL-E

Time series application

- Still early stages of research, achieveing state-of-the-art results in some domains [Liu *et al.*, 2021][1] and underperforming in others [Zeng et al. 2022][2]
- Great opportunity for research in finance domain

1. Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting.
2. Ailing Zeng1, Muxi Chen1, Lei Zhang2, Qiang Xu. Are Transformers Effective for Time Series Forecasting?

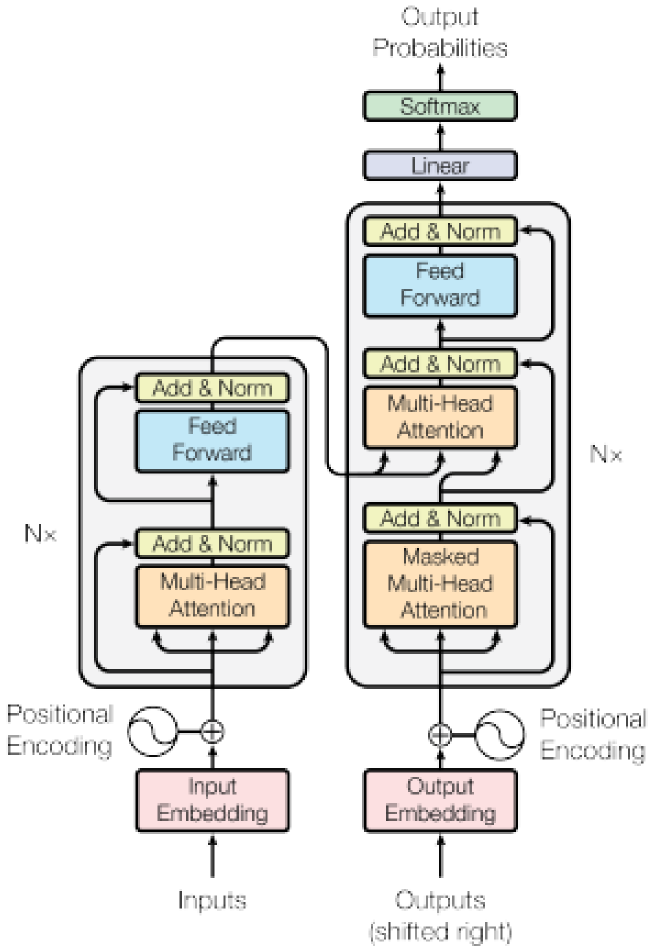
Advantages of Transformers

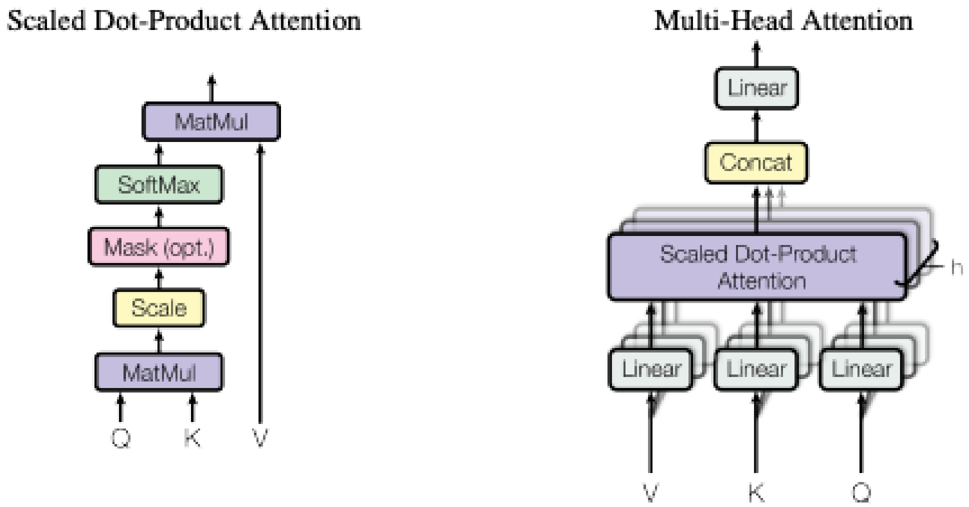
Why Transformers are significant

- Transformers excel at modeling sequential data, such as natural language.
- Unlike the recurrent neural networks (RNNs), Transformers are parallelizable. This makes them efficient on hardware like GPUs and TPUs. The main reasons is that Transformers replaced recurrence with attention, and computations can happen simultaneously. Layer outputs can be computed in parallel, instead of a series like an RNN.
- Unlike RNNs (like seq2seq, 2014) or convolutional neural networks (CNNs) (for example, ByteNet), Transformers are able to capture distant or long-range contexts and dependencies in the data between distant positions in the input or output sequences. Thus, longer connections can be learned. Attention allows each location to have access to the entire input at each layer, while in RNNs and CNNs, the information needs to pass through many processing steps to move a long distance, which makes it harder to learn.
- Transformers make no assumptions about the temporal/spatial relationships across the data. This is ideal for processing a set of objects (for example, StarCraft units).

Source: https://www.tensorflow.org/text/tutorials/transformer

Transformers architecture

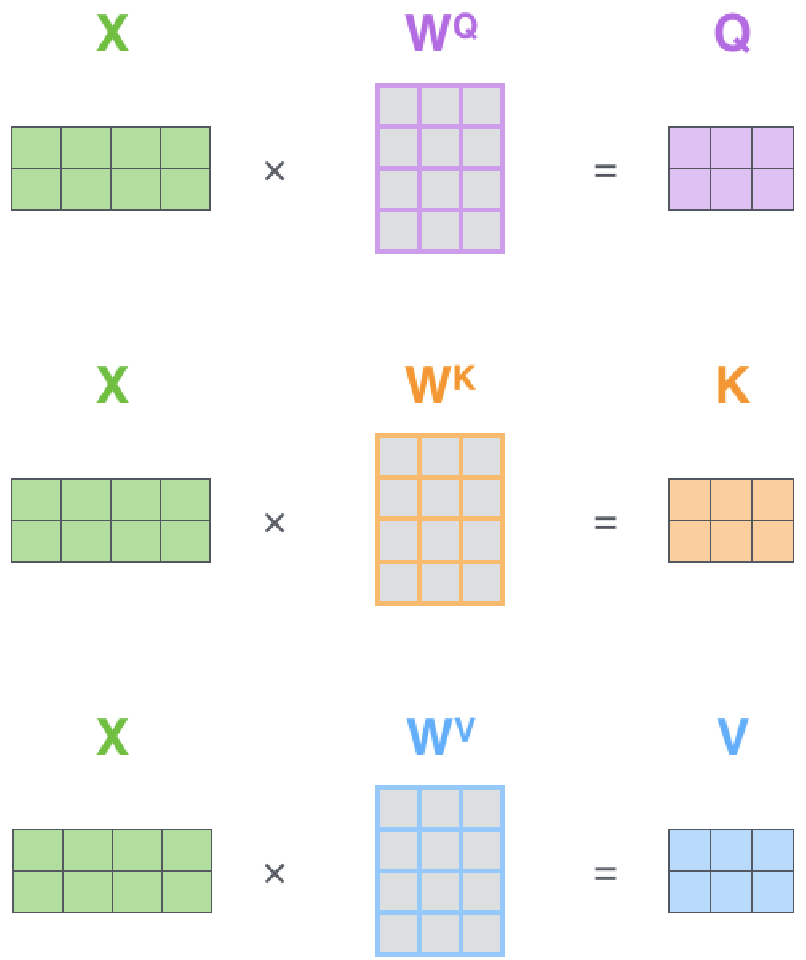




(left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Illustrated Transformer (1/6)

The first step is to calculate the Query, Key, and Value matrices. We do that by packing our embeddings into a matrix X, and multiplying it by the weight matrices we've trained (WQ, WK, WV).

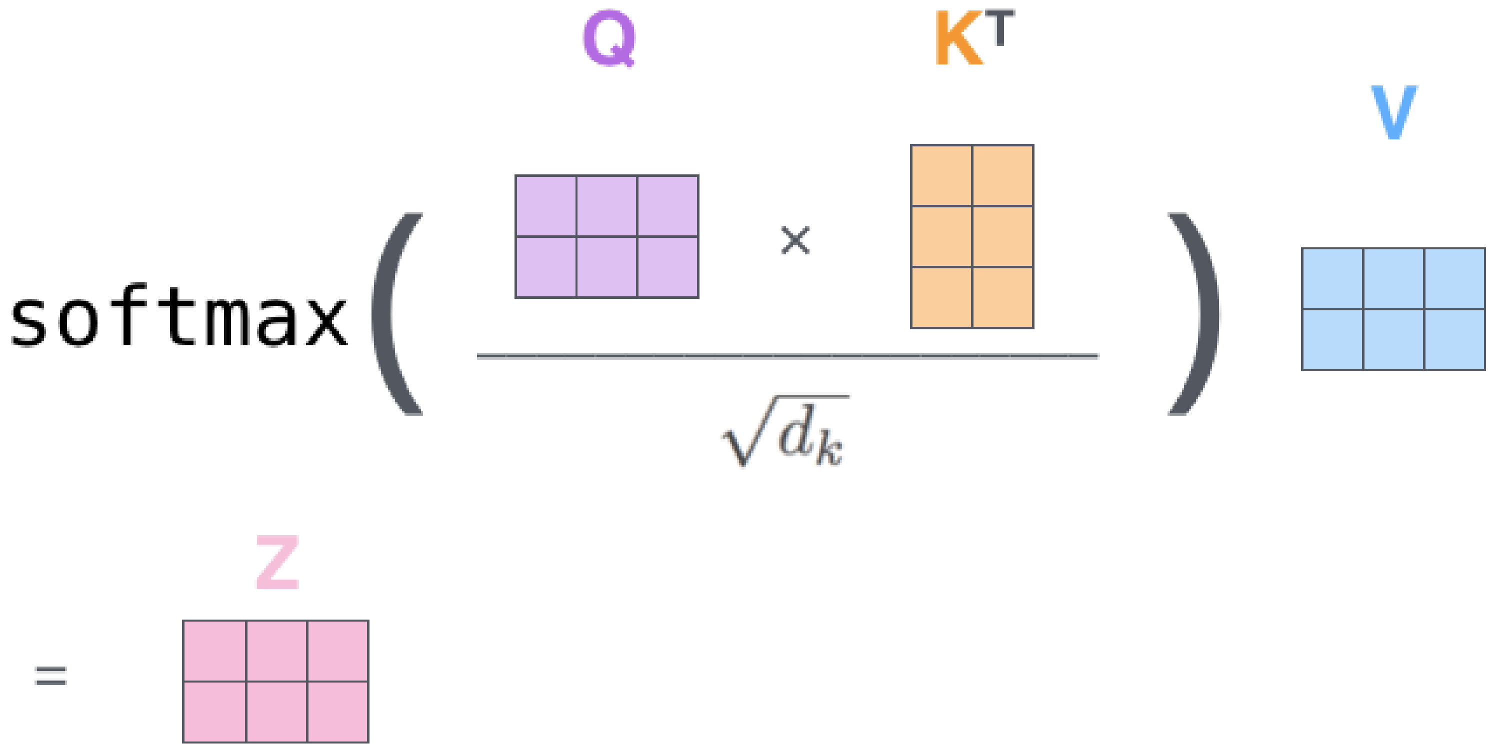


Every row in the X matrix corresponds to a word in the input sentence. We again see the difference in size of the embedding vector (512, or 4 boxes in the figure), and the q/k/v vectors (64, or 3 boxes in the figure)

http://jalammar.github.io/illustrated-transformer/

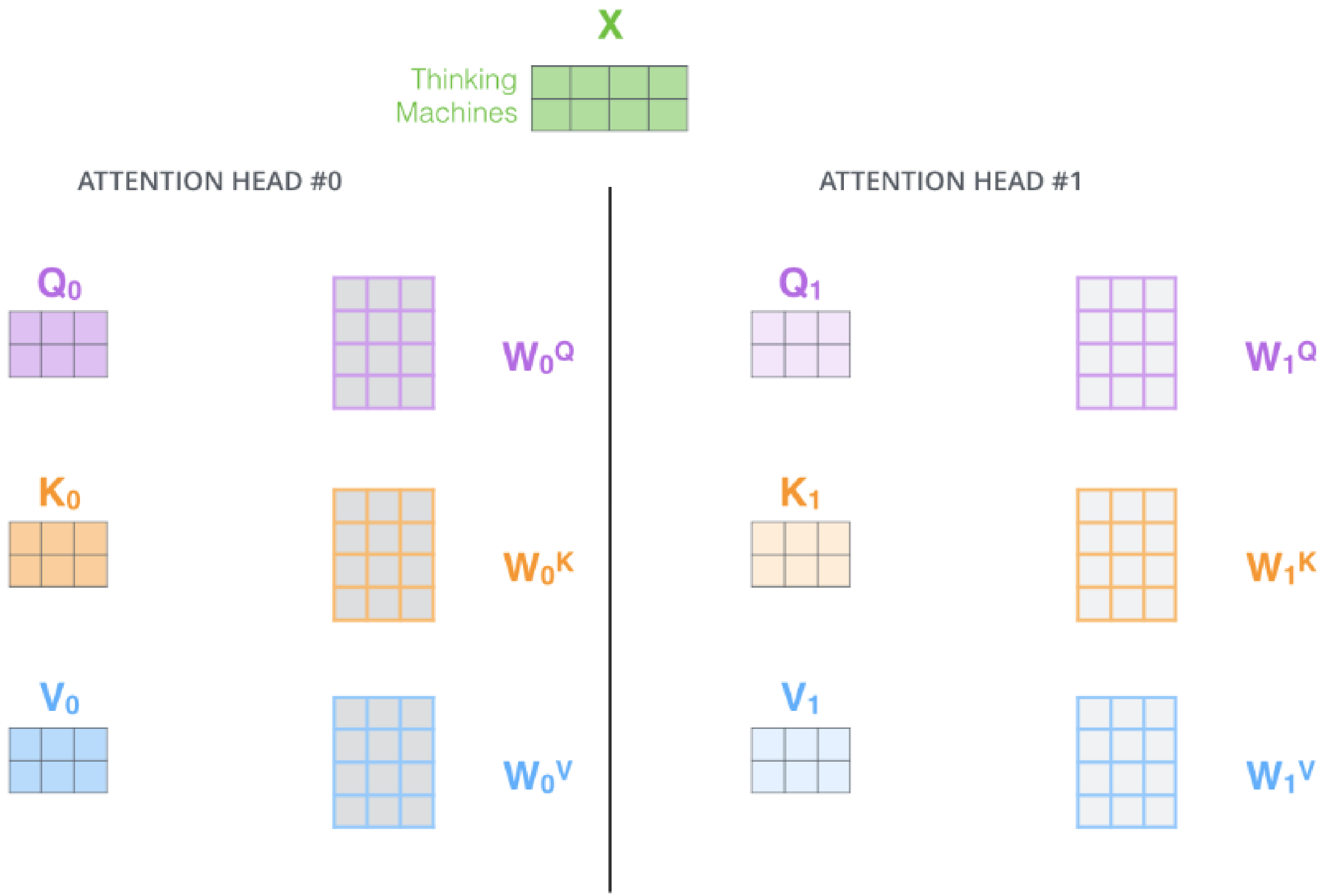
Illustrated Transformer (2/6)

Finally, since we're dealing with matrices, we can condense steps two through six in one formula to calculate the outputs of the self-attention layer.



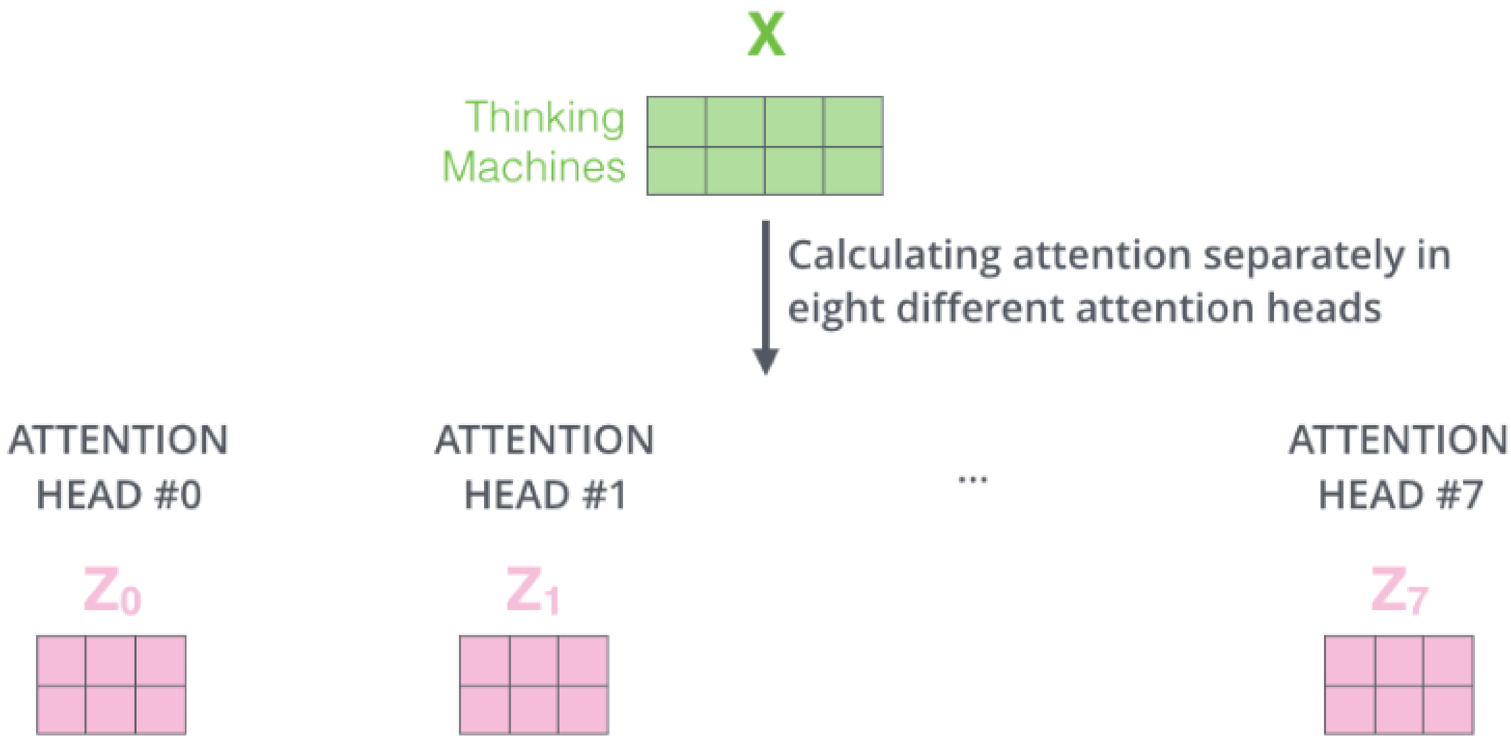
The self-attention calculation in matrix form

Illustrated Transformer (3/6)



With multi-headed attention, we maintain separate Q/K/V weight matrices for each head resulting in different Q/K/V matrices. As we did before, we multiply X by the WQ/WK/WV matrices to produce Q/K/V matrices.

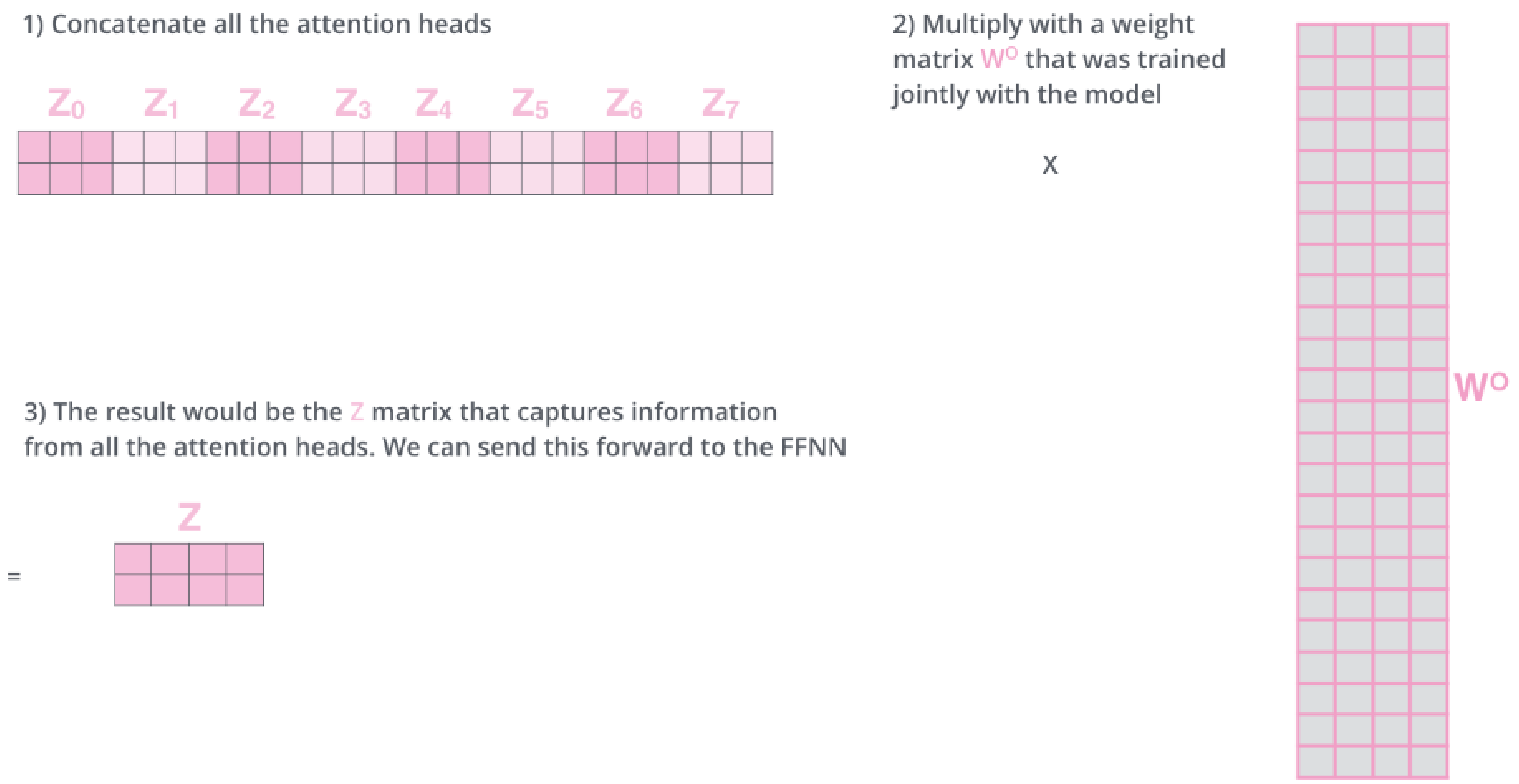
Illustrated Transformer (4/6)



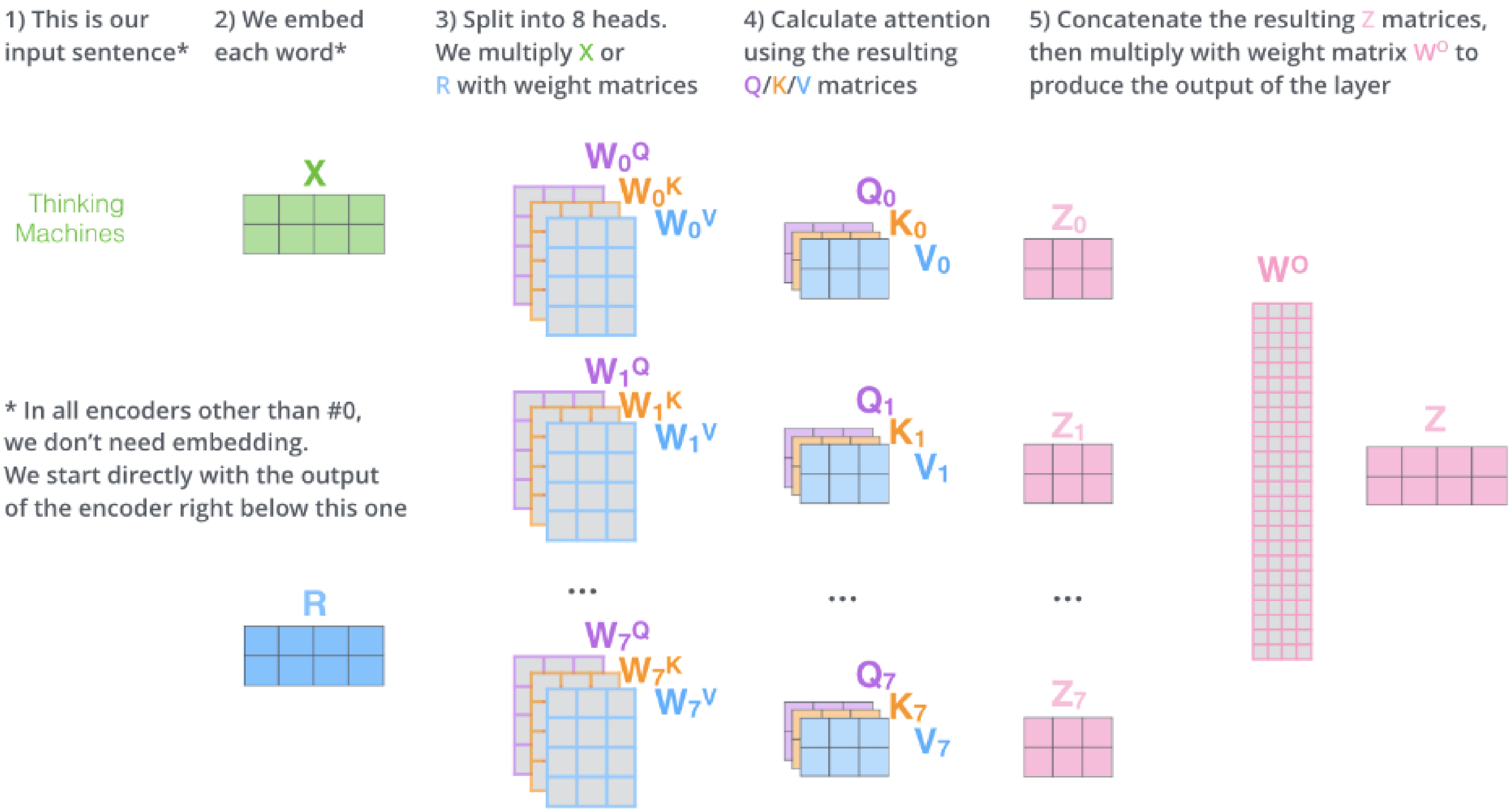
Illustrated Transformer (5/6)

This leaves us with a bit of a challenge. The feed-forward layer is not expecting eight matrices – it's expecting a single matrix (a vector for each word). So we need a way to condense these eight down into a single matrix.

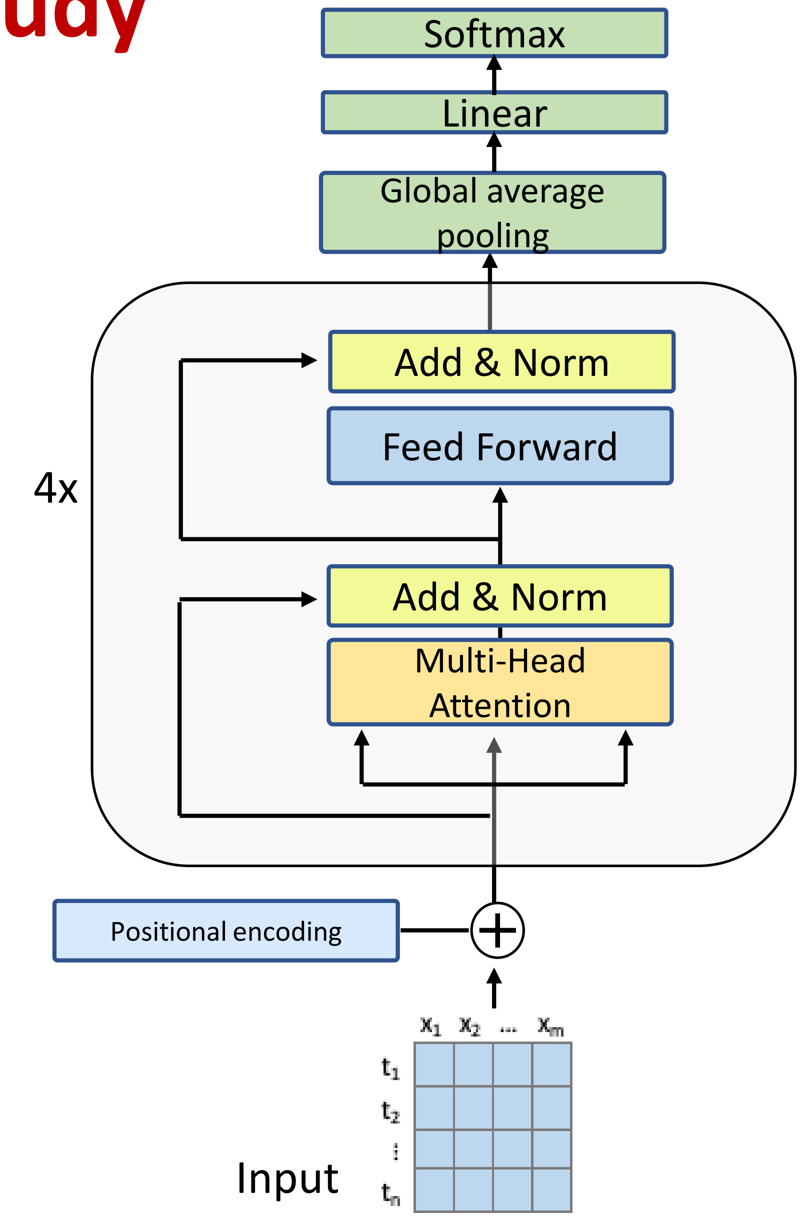
How do we do that? We concat the matrices then multiply them by an additional weights matrix WO.



Illustrated Transformer (6/6)



Transformer used in our study



Models comparison

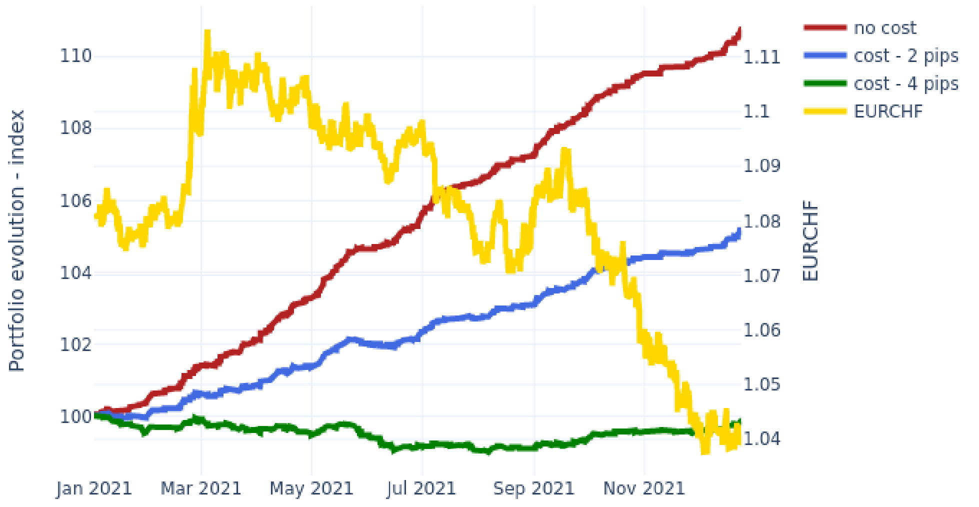
		Transformer		Resnet_LSTM		Resnet	
	Interval	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
EURCHF	60 min	**0.55**	**0.56**	0.54	0.54	0.54	0.54
	120 min	0.55	0.56	**0.57**	**0.57**	**0.57**	**0.57**
	240 min	0.52	0.52	**0.53**	**0.54**	0.52	0.52
	480 min	**0.56**	**0.56**	0.54	0.53	0.54	0.54
	720 min	**0.60**	**0.60**	0.54	0.55	0.58	0.58
USDJPY	60 min	**0.51**	0.51	**0.51**	**0.52**	0.50	0.51
	120 min	**0.50**	**0.52**	**0.50**	0.50	**0.50**	0.51
	240 min	**0.52**	**0.52**	0.50	0.51	0.51	0.50
	480 min	**0.57**	**0.56**	0.49	0.49	0.49	0.51
	720 min	**0.53**	**0.53**	0.51	**0.53**	0.48	0.48
EURUSD	60 min	**0.53**	**0.53**	0.52	**0.53**	0.51	0.52
	120 min	0.51	0.51	**0.52**	**0.52**	0.51	**0.52**
	240 min	0.50	0.51	**0.51**	**0.52**	**0.51**	**0.52**
	480 min	0.52	0.52	**0.54**	**0.53**	0.52	0.50
	720 min	0.49	0.49	**0.51**	**0.51**	0.49	0.50
EURGBP	60 min	0.52	**0.53**	**0.53**	**0.53**	0.52	0.52
	120 min	0.52	0.53	0.52	0.53	**0.54**	**0.54**
	240 min	**0.53**	0.50	0.52	**0.53**	**0.53**	0.52
	480 min	**0.54**	0.51	0.51	**0.52**	0.53	0.50
	720 min	0.52	0.51	0.48	0.48	**0.57**	**0.57**
EURPLN	60 min	**0.55**	0.55	**0.55**	**0.56**	0.54	0.55
	120 min	0.56	0.57	**0.57**	**0.58**	0.55	0.57
	240 min	0.54	0.55	0.53	0.54	**0.58**	**0.59**
	480 min	**0.55**	**0.54**	**0.55**	**0.54**	0.53	0.53
	720 min	**0.57**	**0.56**	0.51	0.50	0.54	0.53
USDCHF	60 min	0.52	0.52	0.53	0.53	**0.54**	**0.54**
	120 min	0.53	0.54	0.53	0.54	**0.55**	**0.55**
	240 min	0.49	0.51	0.51	0.52	**0.52**	**0.53**
	480 min	**0.53**	**0.54**	**0.53**	**0.54**	**0.53**	**0.54**
	720 min	**0.59**	**0.60**	0.58	0.58	0.56	0.56

Results for EURCHF

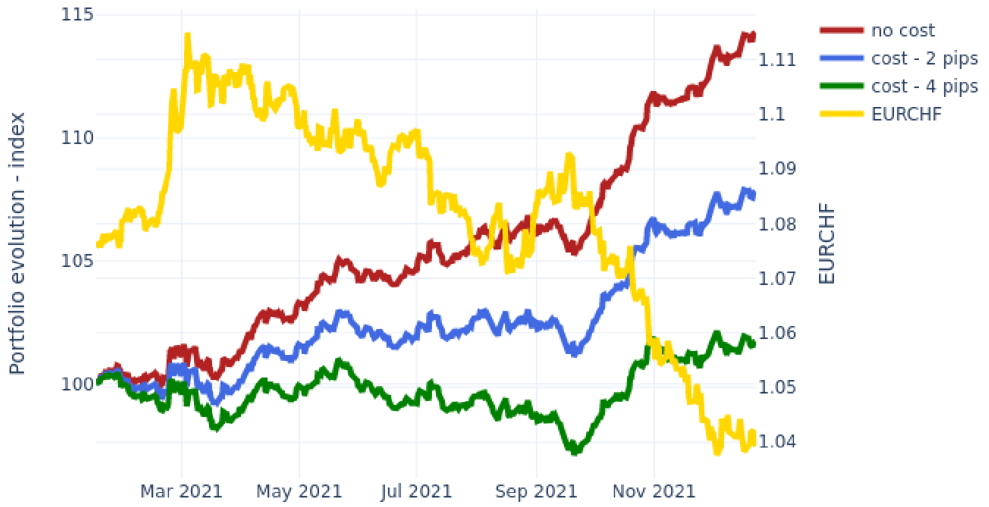
Threshold	Interval	60	120	240	480	720
	AUROC	0.56	0.56	0.52	0.56	0.60
	F1	0.49	0.56	0.54	0.53	0.57
0.5 - 0.5	accuracy	0.55	0.55	0.52	0.56	0.60
	%time	1.00	1.00	1.00	1.00	1.00
	Net results	98.2	98.1	91.9	99.9	**107.8**
0.45 - 0.55	accuracy	0.63	0.64	0.54	0.57	0.72
	%time	0.27	0.36	0.49	0.63	0.21
	Net results	101.1	**106.5**	94.5	99.7	104.0
0.4 - 0.6	accuracy	0.71	0.69	0.56	0.60	0.77
	%time	0.11	0.21	0.23	0.41	0.07
	Net results	105.0	105.1	97.1	100.7	101.6
0.35 - 0.65	accuracy	0.79	0.79	0.65	0.62	0.69
	%time	0.06	0.13	0.10	0.23	0.02
	Net results	**105.2**	105.3	**100.7**	100.0	100.4
0.3 - 0.7	accuracy	0.84	0.83	0.71	0.66	–
	%time	0.03	0.07	0.04	0.13	–
	Net results	102.9	103.1	100.5	101.2	–
0.25 - 0.75	accuracy	0.88	0.91	0.68	0.74	–
	%time	0.01	0.03	0.02	0.06	–
	Net results	101.1	101.8	100.0	**101.4**	–

Note: the best result in each column is in bold.

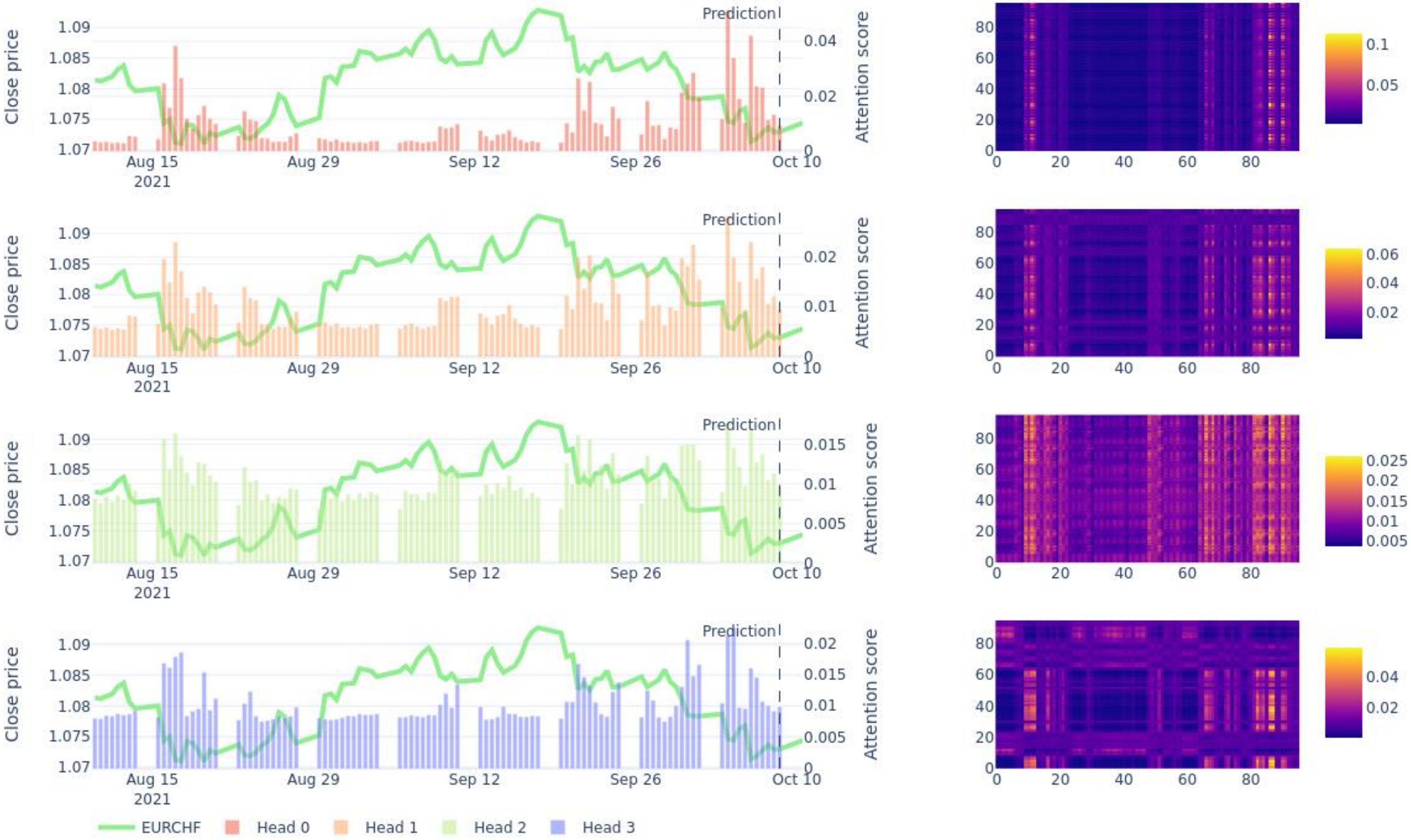
60 min



720 min

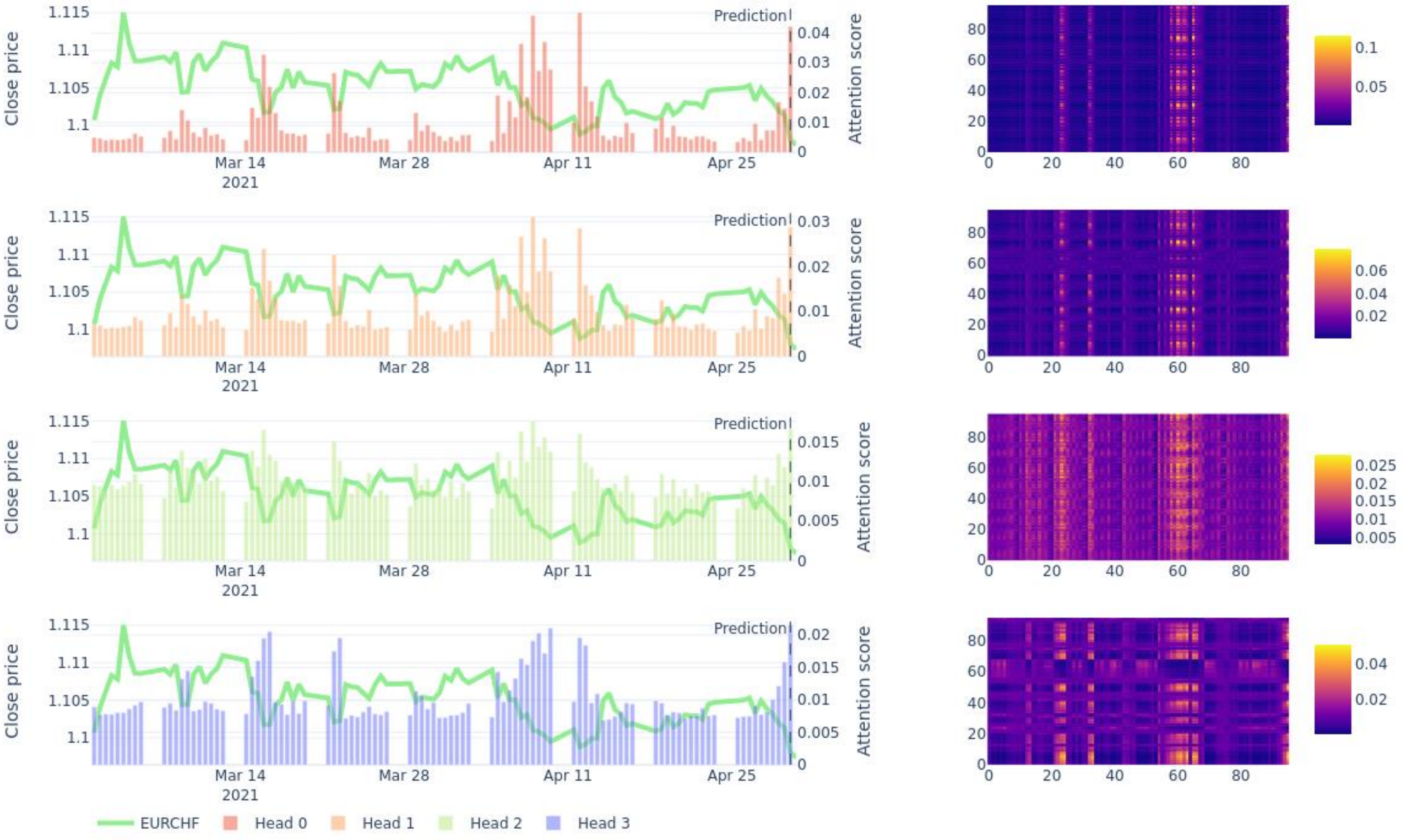


Attention scores – move up



Example of correctly predicted upward movement within the next 12 hours with a probability of 60.8%

Attention scores – move down

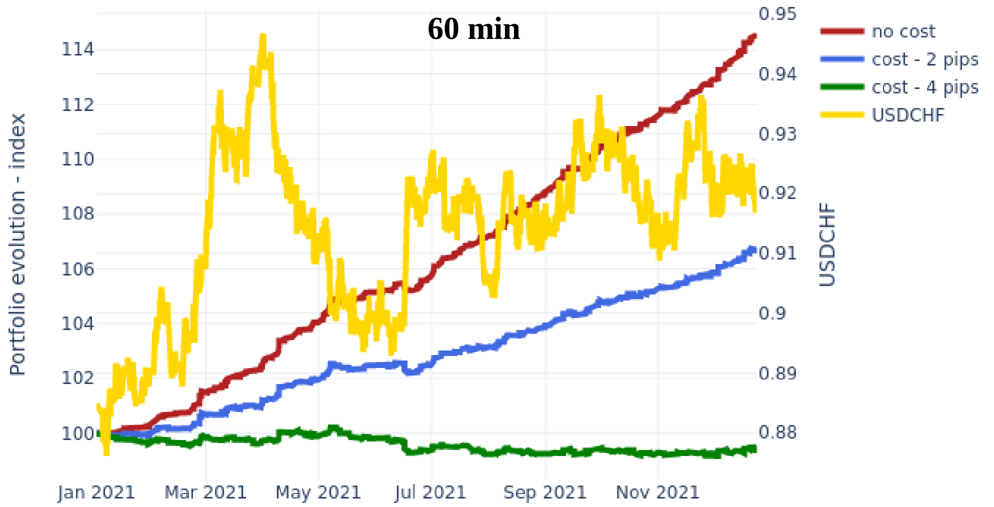


Example of correctly predicted a drop within the next 12 hours with a probability of 58%

Results for USDCHF

Threshold	Interval	60	120	240	480	720
	AUROC	0.52	0.54	0.51	0.54	0.60
	F1	0.39	0.42	0.20	0.41	0.46
0.5 - 0.5	accuracy	0.52	0.53	0.49	0.54	0.60
	%time	1.00	1.00	1.00	1.00	1.00
	Net results	70.5	87.5	91.2	99.4	**112.7**
0.45 - 0.55	accuracy	0.68	0.59	0.51	0.81	0.74
	%time	0.14	0.22	0.13	0.09	0.25
	Net results	105.1	96.5	96.4	**103.1**	104.7
0.4 - 0.6	accuracy	0.77	0.70	0.60	0.82	0.85
	%time	0.06	0.06	0.03	0.01	0.09
	Net results	**106.7**	**101.9**	100.1	100.3	104.1
0.35 - 0.65	accuracy	0.79	0.78	0.81	1.00	–
	%time	0.03	0.02	0.01	0.00	–
	Net results	103.7	101.7	**100.6**	100.1	–
0.3 - 0.7	accuracy	0.79	0.80	0.80	–	–
	%time	0.01	0.01	0.00	–	–
	Net results	101.6	100.8	100.1	–	–
0.25 - 0.75	accuracy	0.73	0.67	0.67	–	–
	%time	0.00	0.00	0.00	–	–
	Net results	100.0	100.0	99.9	–	–

Note: the best result in each column is in bold.



Results for EURPLN and USDJPY

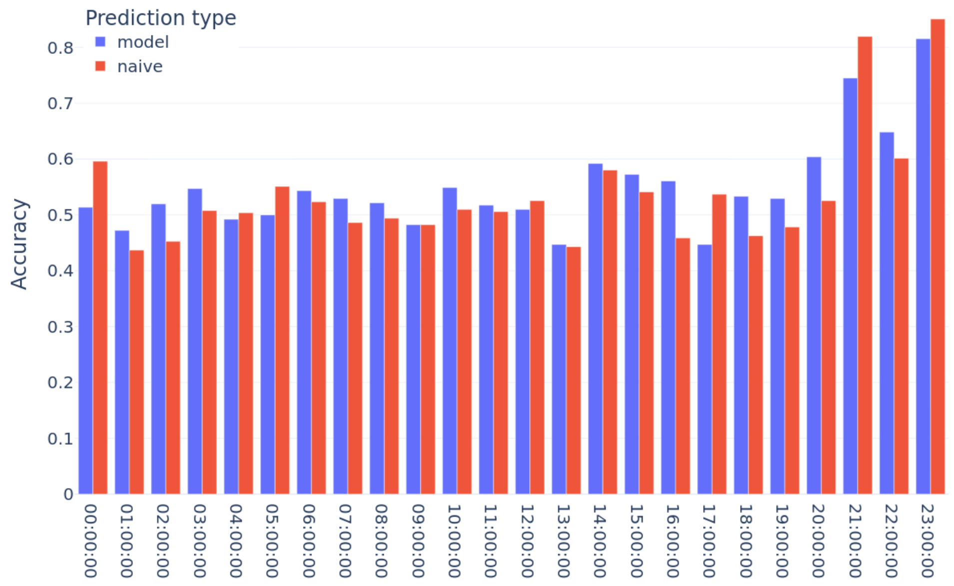
		EURPLN					USDJPY				
Threshold	Interval	60	120	240	480	720	60	120	240	480	720
	AUROC	0.55	0.572	0.55	0.54	0.56	0.51	0.52	0.52	0.56	0.53
	F1	0.51	0.433	0.33	0.34	0.35	0.54	0.44	0.53	0.64	0.51
0.5 - 0.5	accuracy	0.55	0.562	0.54	0.55	0.57	0.51	0.50	0.52	0.57	0.53
	%time	1.00	1,00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Net results	54.1	86.4	88.8	95.7	96.4	81.6	88.1	94.4	**105.9**	**104.0**
0.45 - 0.55	accuracy	0.64	0.693	0.57	0.58	0.60	0.56	0.53	0.53	0.64	0.59
	%time	0.20	0.22	0.44	0.46	0.68	0.26	0.21	0.14	0.14	0.38
	Net results	81.6	93.35	87.8	93.5	99.1	94.4	95.8	98.8	100.8	102.0
0.4 - 0.6	accuracy	0.75	0.871	0.64	0.68	0.66	0.60	0.63	0.54	0.90	0.68
	%time	0.04	0.02	0.12	0.11	0.33	0.10	0.07	0.03	0.01	0.11
	Net results	98.0	**101,0**	97.3	99.2	98.9	**100.9**	**100.9**	99.9	101.2	102.4
0.35 - 0.65	accuracy	0.82	1,00	0.73	0.67	0.62	0.67	0.74	0.67	0.00	0.78
	%time	0.01	0,00	0.01	0.02	0.10	0.04	0.02	0.00	0.00	0.02
	Net results	99.8	100.2	99.9	99.0	99.9	100.8	101.2	**100.4**	99.9	100.5
0.3 - 0.7	accuracy	0.84	1,00	0.67	–	0.88	0.69	0.82	–	–	–
	%time	0.00	0,00	0.00	–	0.01	0.02	0.01	–	–	–
	Net results	99.7	99.99	**100.2**	–	100.6	100.3	100.5	–	–	–
0.25 - 0.75	accuracy	–	–	1.00	–	–	0.80	0.89	–	–	–
	%time	–	–	0.00	–	–	0.01	0.00	–	–	–
	Net results	–	–	100.1	–	–	100.5	100.4	–	–	–

Results for EURUSD and EURGBP

		EURUSD					EURGBP				
Threshold	Interval	60	120	240	480	720	60	120	240	480	720
	AUROC	0.53	0.51	0.51	0.52	0.49	0.53	0.53	0.50	0.51	0.52
	F1	0.49	0.51	0.64	0.51	0.60	0.47	0.60	0.00	0.09	0.43
0.5 - 0.5	accuracy	0.53	0.51	0.50	0.52	0.49	0.52	0.52	0.53	0.54	0.52
	%time	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Net results	91.1	92.3	93.8	95.3	96.0	71.4	78.6	**106.6**	**104.5**	97.4
0.45 - 0.55	accuracy	0.57	0.52	0.25	0.51	0.53	0.61	0.57	–	0.59	0.56
	%time	0.19	0.26	0.01	0.17	0.22	0.18	0.37	–	0.02	0.48
	Net results	94.9	94.1	99.1	**100.3**	100.0	93.4	89.5	–	100.2	**101.6**
0.4 - 0.6	accuracy	0.64	0.53	0.00	0.57	0.25	0.70	0.68	–	0.38	0.57
	%time	0.05	0.07	0.00	0.01	0.01	0.08	0.16	–	0.01	0.26
	Net results	99.6	97.7	100.0	100.2	99.7	99.4	98.7	–	99.4	**101.6**
0.35 - 0.65	accuracy	0.68	0.54	0.00	0.00	0.00	0.84	0.72	–	0.33	0.53
	%time	0.01	0.03	0.00	0.00	0.00	0.04	0.08	–	0.01	0.11
	Net results	100.0	99.0	100.0	100.0	99.9	**101.8**	99.4	–	99.3	100.4
0.3 - 0.7	accuracy	0.60	0.50	–	–	–	0.87	0.76	–	0.25	0.53
	%time	0.00	0.01	–	–	–	0.01	0.04	–	0.01	0.03
	Net results	99.9	99.7	–	–	–	100.9	99.8	–	99.5	100.3
0.25 - 0.75	accuracy	0.75	–	–	–	–	0.86	0.84	–	0.50	0.40
	%time	0.00	–	–	–	–	0.00	0.02	–	0.00	0.01
	Net results	100.0	–	–	–	–	100.2	100.6	–	99.6	99.7

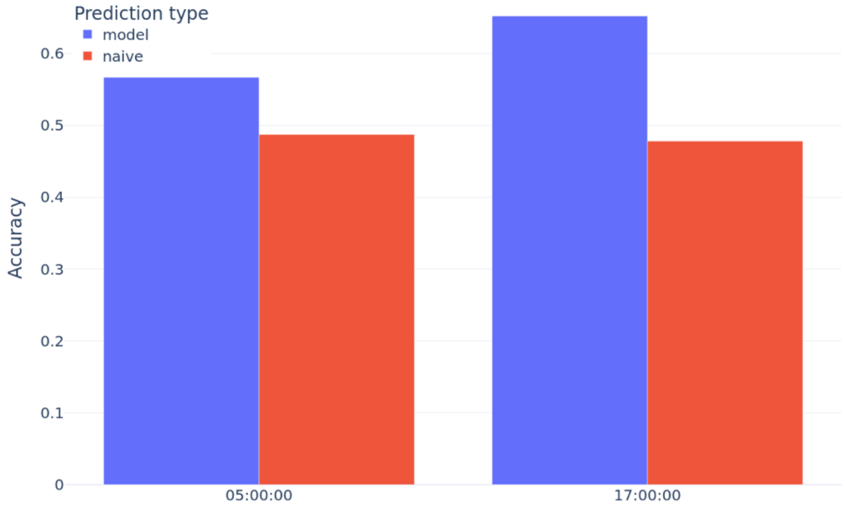
Issues with OHLC data for Forex

Model for EURCHF built on 60 min interval



Note: London time zone

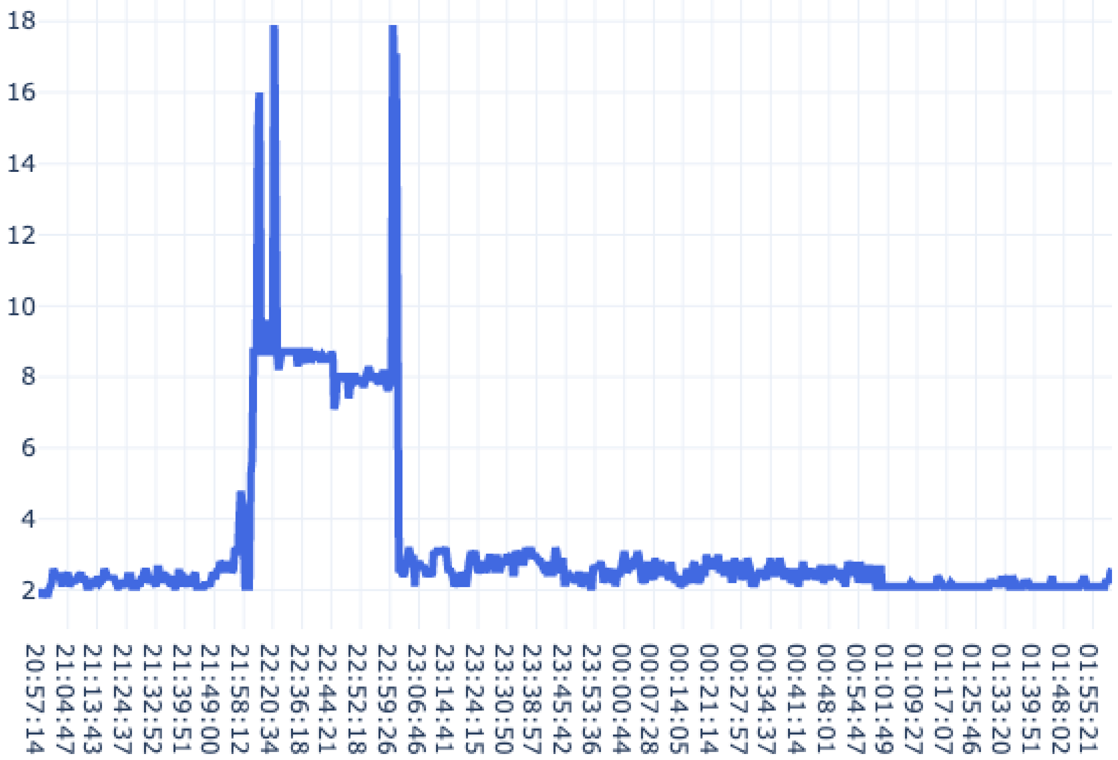
Model for EURCHF built on 720 min interval



Naïve predictions were created based on an asymmetric distribution of difference in close prices between consecutive hours. Concretely, it is the mode of the direction of price changes in a particular time interval in the training set.

Spread can be volatile depending on time of the day

Spread at night for EURCHF measured in pips



Comparison of trading performance for EURCHF after excluding hours with high spread

Threshold	Interval	60	120	60	120
		All hours (initial results)		High spread hours excluded	
0.5 - 0.5	accuracy	0.55	0.55	0.52	0.50
	%time	1.00	1.00	0.87	0.83
	Net results	98.2	98.1	85.5	85.1
0.45 - 0.55	accuracy	0.63	0.64	0.55	0.51
	%time	0.27	0.36	0.17	0.21
	Net results	101.1	106.5	90.5	95.5
0.4 - 0.6	accuracy	0.71	0.69	0.55	0.45
	%time	0.11	0.21	0.04	0.08
	Net results	105.0	105.1	97.0	96.6
0.35 - 0.65	accuracy	0.79	0.79	0.65	0.48
	%time	0.06	0.13	0.01	0.03
	Net results	105.2	105.3	99.6	98.8
0.3 - 0.7	accuracy	0.84	0.83	0.67	0.39
	%time	0.03	0.07	0.00	0.01
	Net results	102.9	103.1	99.9	99.7
0.25 - 0.75	accuracy	0.88	0.91	1.00	0.33
	%time	0.01	0.03	0.00	0.00
	Net results	101.1	101.8	100.0	100.0

VARSOVIENSIS

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Conclusions (1/2)

- The Transformer network exhibits predictive power in the context of intraday Forex movements with performance metrics varying greatly between different currency pairs and frequencies of data
- From the practical (trading) standpoint, the Transformer model offers improvement over the ResNet-LSTM model, which presents a challenging benchmark and, in some cases still might be a better choice
- Switch from positional encoding introduced in the original Transformer network to trainable encodings improved the performance of the model
- Four currency pairs (EURCHF, USDJPY, EURGBP, USDCHF) it is possible to devise a profitable trading strategy based on the proposed model and we offer insights with respect to optimal time aggregation for those currencies
- Applying transfer learning between currencies is a viable strategy to improve the results as we can take advantage of the fact that for some currencies it is easier to train neural networks.
- Relying on typical accuracy metrics and simple backtesting can be very misleading in the Forex market when using OHLC data. We showed that for high frequencies high accuracy was achieved when a spread cost was the highest.
- Intervals 240, 480, and 720 min are more suitable for Forex trading based on deep learning models than 60 and 120 min.
- Length of the training sample in the network did not play a critical role, whereas feature selection did.

Conclusions (2/2)

The **main contribution** of this article is the application of state-of-the-art Transformer network to different currency pairs within different time aggregations, which is followed by robust analysis of the practical aspect of our findings based on realistic assumptions regarding trading costs, something which is rarely the case in the empirical literature.

The innovative elements of this study are:

- Application of the Transformer network to financial forecasting and devising a trading strategy based on its predictions.
- Research of currently state-of-the-art deep learning models for intraday Forex trading to ensure the robustness of comparative models.
- The search for an optimal (from the trading perspective) time aggregation frequency for different currency pairs.
- Thorough analysis of the results obtained that shows how the performance of deep learning models for intraday Forex trading can overestimate its performance and consequently lead to mistaken conclusions.
- Application of transfer learning between different currency pairs, which combined with the importance of features derived from technical analysis discussed in the sensitivity analysis chapter, give some credence to this method today popular among traders.

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

THANK YOU FOR YOUR ATTENTION!