

# Deep learning enhanced quantitative trading strategies

Stefan Zohren

*Oxford Man Institute of Quantitative Finance*  
*University of Oxford*



July 11, 2023

## Special thanks to:



Associate Member  
**Bryan Lim**



Student  
**Kieran Wood**



Student  
**Daniel Poh**



Visiting Student  
**Fernando Moreno-Pino**



Student  
**Alvaro Arroyo Nuñez**



Post-doctoral Researcher  
**Zihao Zhang**

# Outline

An Overview of Research at the Oxford-Man Institute

Classical Time-Series Momentum Strategies

Deep Momentum Networks

Deep Momentum Networks with Changepoint Detection

Momentum Transformer

# An Overview of Research at the Oxford-Man Institute

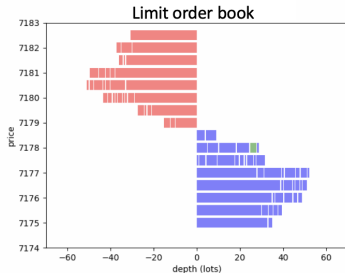
# The Oxford-Man Institute of Quantitative Finance

- A unique collaboration between Oxford University and Man Group plc
  - Opened Summer 2007
  - Focus: Machine Learning in quantitative finance & alternative investments
  - Vision to become world leading research centre
- Man Research Laboratory
  - Embedded commercial research laboratory
  - Undertakes Man AHL's research
  - Opportunity for collaboration and academic publication



# Research Topics: Modelling High-Frequency Data

- Institute intensely engaged in:
  - Modelling of limit order book data
  - Deep learning models for forecasting from HFT observations
  - Key contributions such as DeepLOB have been adopted across industry in market making and execution



Q Search

Bloomberg

Technology

## Man Group-Oxford Quants Say Their AI Can Predict Stock Moves

Journals & Magazines > IEEE Transactions on Signal P... > Volume: 67 Issue: 11

DeepLOB: Deep Convolutional Neural Networks for Limit Order

Publisher: IEEE

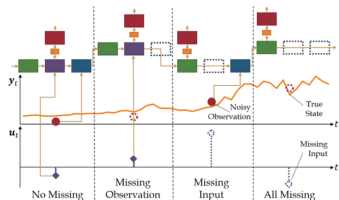
Cite This

PDF

Zihao Zhang ; Stefan Zohren ; Stephen Roberts ; All Authors

# Research Topics: Deep Learning for Time Series

- We have done extensive research on deep learning for time-series modelling with applications in finance and beyond
- Members of the group were involved in one of the first transformer based architectures for time-series, the temporal fusion transformer

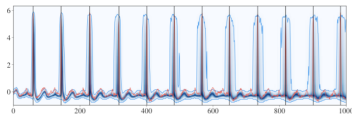


Review articles

## Time-series forecasting with deep learning: a survey

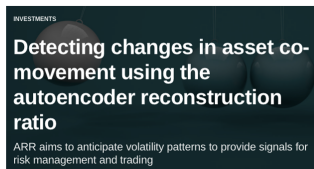
Bryan Lim  and Stefan Zohren

Published: 15 February 2021 | <https://doi.org/10.1098/rsta.2020.0209>



# Research Topics: Deep Learning for Quant Trading

- A major focus has been on deep learning for quant trading, including work on:
  - Portfolio optimisation
  - Reinforcement learning for trading
  - Autoencoder based risk/fragility measures
  - Enhancing traditional quant trading strategies (e.g. time-series momentum and cross-sectional momentum strategies)



## Deep Learning for Portfolio Optimization

Zihao Zhang, Stefan Zohren and Stephen Roberts

The Journal of Financial Data Science Fall 2020, 2 (4) 8-20; DOI: <https://doi.org/10.3905/jfds.2020.1.042>

the journal of  
financial data science

## Building Cross-Sectional Systematic Strategies by Learning to Rank

Daniel Poh, Bryan Lim, Stefan Zohren and Stephen Roberts

The Journal of Financial Data Science Spring 2021, 3 (2) 70-86; DOI: <https://doi.org/10.3905/jfds.2021.1.060>



# Research Topics: Networks and NLP in Finance

- Institute engages in research on:
  - Financial news networks
  - Sentiment analysis and propagation
  - Virality in Social Media (Reddit, Twitter)
  - Financial word embeddings
- Organised some key workshops in field



nature > scientific reports > articles > article

Article | [Open Access](#) | Published: 04 February 2021

## Sentiment correlation in financial news networks and associated market movements

Xingchen Wan , Jie Yang , Slavi Marinov, Jan-Peter Calliess, Stefan Zohren & Xiaowen Dong

*Scientific Reports* **11**, Article number: 3062 (2021) | [Cite this article](#)



# Classical Time-Series Momentum Strategies

# Momentum Strategies

- ▶ Time-series Momentum (TSMOM) (*Moskowitz et al.* [1]) is derived from the philosophy that strong price trends have a tendency to persist.
- ▶ Often known as ‘follow the winner’ because it is assumed that winners will continue to be winners in the subsequent period.
- ▶ TSMOM is a univariate approach as opposed to cross-sectional (*Jegadeesh et al.* [2]) momentum strategies, which trade assets against each other and select a portfolio based on relative ranking.
- ▶ Strategies involve 1) estimation of a trend, and 2) sizing positions accordingly.
- ▶ Momentum strategies are an important part of alternative investments and are at the heart of commodity trading advisors (CTAs).

## Classical Strategies

- ▶ Volatility scaling has been proven to play a crucial role in the positive performance of TSMOM strategies (*Kim et al.* [3]).
- ▶ Where  $X_t^{(i)}$  is position size of the  $i$ -th asset,  $N$  the number of assets in our portfolio,  $\sigma_t^{(i)}$  the ex-ante volatility estimate and  $\sigma_{\text{tgt}}$  the target volatility,

$$R_{t+1}^{\text{TSMOM}} = \frac{1}{N} \sum_{i=1}^N R_{t+1}^{(i)}, \quad R_{t+1}^{(i)} = X_t^{(i)} \frac{\sigma_{\text{tgt}}}{\sigma_t^{(i)}} r_{t+1}^{(i)}. \quad (1)$$

- ▶ *Moskowitz et al.* [1], selects position as  $X_t^{(i)} = \text{sgn}(r_{t-252,t})$ , where we are using the volatility scaling framework and  $r_{t-252,t}$  is annual return.
- ▶ Moving Average Convergence Divergence (*MACD*) is a volatility normalised trend-following momentum indicator that describes the relationship between two moving averages of a security's price, functioning as a trigger for buy and sell signals (see *Baz et al.* [4]).

# Deep Momentum Networks

# Deep Momentum Networks

- ▶ Previous Deep learning approaches either only learnt the trend or alternatively performed classification, taking a maximum long or short position.
- ▶ In our first work with B. Lim and S. Roberts, we proposed a framework termed as *Deep Momentum Networks* (DMNs) which resulted in significantly better risk-adjusted-returns.
- ▶ Rather than estimating the trend and then using a rule based approach to size positions, DMNs learn the trend in a data-driven manner and directly output position sizes.
- ▶ A squashing function  $\tanh(\cdot)$  directly outputs positions  $X_t^{(i)} \in (-1, 1)$
- ▶ DMNs also benefit from the volatility scaling framework.
- ▶ Inputs are normalised returns at different timescales and different MACD indicators.

# Deep Momentum Networks

- ▶ We found that the LSTM, a type of Recurrent Neural Network used for sequence modelling, produced the best results.
- ▶ Since we want to maximise risk-adjusted-returns, DMNs use a Sharpe Ratio Loss function

$$\mathcal{L}_{\text{sharpe}}(\theta) = -\frac{\sqrt{252} \mathbb{E}_{\Omega} [R_t^{(i)}]}{\sqrt{\text{Var}_{\Omega} [R_t^{(i)}]}}. \quad (2)$$

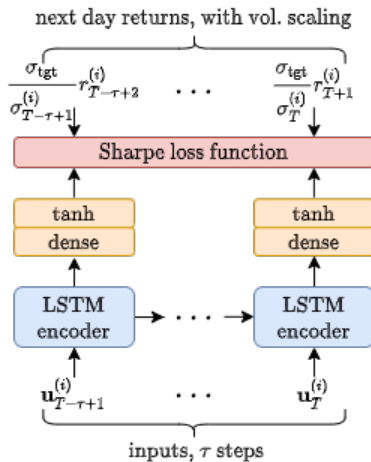
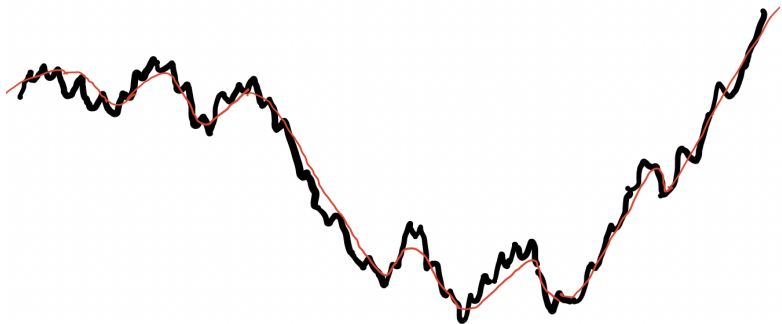


Figure: LSTM Deep Momentum Network architecture

## Interpreting the results of DMNs

- ▶ DMN's strong performance attributed to learning simultaneously to exploit slow momentum and fast reversion
- ▶ Following small reversions in regimes of strong trend can lead to larger transaction costs

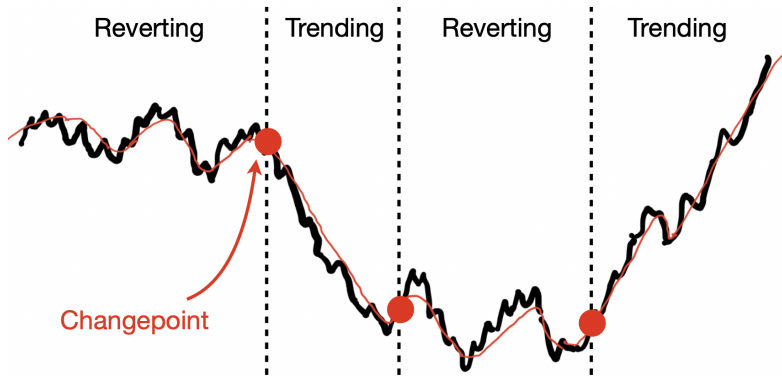




# Deep Momentum Networks with Changepoint Detection

# Momentum Turning Points and Changepoint Detection

- ▶ Ideally identify regimes – trending and reverting market – then learn to switch to exploit each state.
- ▶ Can be done with changepoint detection as proposed in earlier work with K. Woods and S. Roberts



## Motivation from Momentum Turning Points

- ▶ Immediately after momentum turning points, where a trend reverses from an uptrend (downtrend) to a downtrend (uptrend), time-series momentum (TSMOM) strategies are prone to making bad bets.
- ▶ We require an approach which is a balancing act between being quick enough to respond to turning points, but not over-reacting to noise.
- ▶ *Garg et al.* [5] proposed an *Intermediate* strategy where a slow momentum signal based on a long lookback window, such as one year, is blended with a fast momentum signal based on a short lookback window, such as one month.

$$X_t = (1 - w) \operatorname{sgn}(r_{t-252,t}) + w \operatorname{sgn}(r_{t-21,t}). \quad (3)$$

- ▶ *MACD* can often produce false positives and signal a possible reversal without one actually happening.

# Changepoint Detection

- ▶ Changepoint detection (CPD) is a field which involves the identification of abrupt changes in sequential data.
- ▶ To enable us to respond to CPD in real time, we require an ‘online’ algorithm, which processes each data point as it becomes available.
- ▶ First introduced by *Adams et al.* [6], Bayesian approaches to online CPD, which naturally accommodate to noisy, uncertain and incomplete time-series data, have proven to be very successful.
- ▶ We focus on approaches using Gaussian Processes (GPs) (*Williams et al.* [7]) a Bayesian non-parametric model which has a proven track record for time-series forecasting, is principled and is robust to noisy inputs.

# Changepoint Detection with Gaussian Processes

- ▶ For daily return  $\hat{r}_t^{(i)}$ , normalised over some look-back window (we'll revisit this), we define the GP as a distribution over functions,

$$\hat{r}_t^{(i)} = f(t) + \epsilon_t, f \sim \mathcal{GP}(0, k_\xi), \epsilon_t \sim \mathcal{N}(0, \sigma_n^2), \quad (4)$$

where  $\epsilon$  is an additive noise process and  $\mathcal{GP}$  is specified by a covariance function  $k_\xi(\cdot, \cdot)$ , which is in turn parameterised by a set of hyperparameters  $\xi$ . Noise variance  $\sigma_n$ , helps to deal with noisy outputs which are uncorrelated.

- ▶ The Matérn 3/2 kernel is a good choice of covariance function for noisy financial data, with kernel hyperparameters  $\xi_M = (\lambda, \sigma_h, \sigma_n)$ , with  $\lambda$  the input scale and  $\sigma_h$  the output scale.
- ▶ A changepoint can either be a drastic change in covariance, a sudden change in the input scale, or a sudden change in the output scale

## Changepoint Kernel

- ▶ *Garnett et al.* introduced the **Region-switching kernel**, where it is assumed there is a drastic change, or changepoint, at  $c \in \{t - l + 1, t - l + 2, \dots, t - 1\}$ , after which all observations before  $c$  are completely uninformative about the observations after this point,

$$k_{\xi_R}(x, x') = \begin{cases} k_{\xi_1}(x, x') & x, x' < c \\ k_{\xi_2}(x, x') & x, x' \geq c \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

- ▶ The lookback window (LBW)  $l$  for this approach needs to be prespecified and it is assumed to contain a single changepoint.
- ▶ A more flexible approach is the **Changepoint kernel**, where  $c \in (t - l, t)$  is the changepoint location,  $s > 0$  is the steepness parameter and  $\sigma(x, x') = \sigma(x)\sigma(x')$ ,  $\bar{\sigma}(x, x') = (1 - \sigma(x))(1 - \sigma(x'))$ ,

$$k_{\xi_C}(x, x') = k_{\xi_1}(x, x')\sigma(x, x') + k_{\xi_2}(x, x')\bar{\sigma}(x, x'). \quad (6)$$

- ▶ We use Matérn 3/2 for the left and right kernels.

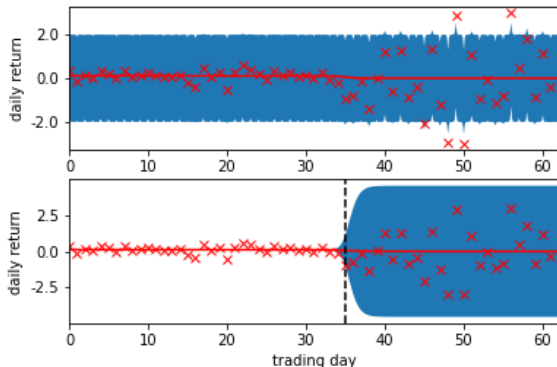
## Changepoint Detection Module Outputs

- ▶ We consider the series  $\{r_{t'}^{(i)}\}_{t'=t-l}^t$ , with lookback horizon  $l$  from time  $t$ . For every CPD window, where  $\mathcal{T} = \{t-l, t-l+1, \dots, t\}$ , we standardise our returns for consistency.
- ▶ For each time step, our changepoint detection module outputs,
  1. changepoint detection location  $\gamma_t^{(i)} \in (0, 1)$ , indicating how far in the past the changepoint is, and,
  2. changepoint score  $\nu_t^{(i)} \in (0, 1)$ , which measures the level of disequilibrium, measured by the reduction in negative log marginal likelihood achieved via the introduction of the changepoint kernel hyperparameters.

$$\nu_t^{(i)}(l) = 1 - \frac{1}{1 + e^{-(\text{nlmn}_{\xi_C} - \text{nlmn}_{\xi_M})}}, \quad \gamma_t^{(i)}(l) = \frac{c - (t - l)}{l}, \quad (7)$$

- ▶ Both values are normalised to help improve stability and performance of our LSTM module.

# Changepoint Kernel



**Figure:** Plots of daily returns for S&P 500, composite ratio-adjusted continuous futures contract during the first quarter of 2020, where returns have been standardised. The top plot fits a GP, using the Matérn 3/2 kernel and the bottom using the Changepoint kernel specified in (6).



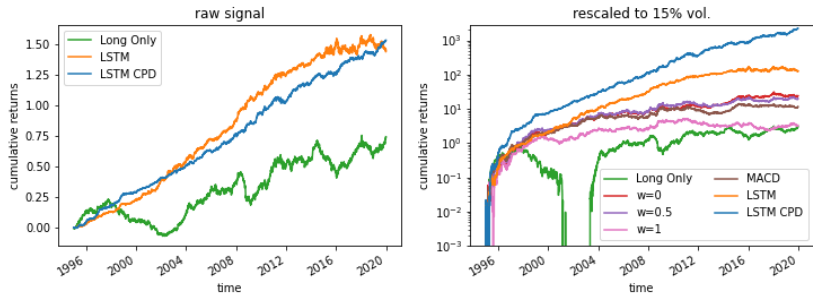
## DMNs with Changepoint Detection Model

- ▶ In our second paper with K. Wood and S. Roberts, we introduce online CPD based on GPs into DMNs.
- ▶ Precisely, the input  $u_t^{(i)}$  for each time-step of LSTM sequence consists of past returns, MACD signals, as well as changepoint location and severity scores:
  1.  $\{r_{t-t',t}^{(i)} / \sigma_t^{(i)} \sqrt{t'} \mid t' \in \{1, 21, 63, 126, 252\}\}$ ,
  2.  $\{\text{MACD}(i, t, S, L) \mid (S, L) \in \{(8, 24), (16, 28), (32, 96)\}\}$ ,
  3.  $\nu_t^{(i)}(l)$  and  $\gamma_t^{(i)}(l)$  for  $l \in \{10, 21, 63, 126, 252\}$
- ▶ The LSTM is not complex enough to handle multiple CPD lookback-windows (LBW) and we optimise  $l$  as part of the hyperparameter tuning process.
- ▶ Later work also demonstrates that multiple LBWs (short and long) work well in conjunction with a variable selection network.

## Data and Experimental Setting

- ▶ Portfolio consisting of 50 of the most liquid, ratio-adjusted continuous futures contracts over the period 1990–2020.
- ▶ Includes daily Commodities, FX, Fixed Income and Equities data, extracted from the Pinnacle Data Corp CLC database.
- ▶ We use an expanding window approach, where we start by using 1990–1995 for training/validation, then test out-of-sample on the period 1995–2000. With each successive iteration, we expand the training/validation window by an additional five years.
- ▶ We use a 90%/10% split for training/validation data, training on the Sharpe loss function via minibatch Stochastic Gradient Descent (SGD), using the validation set to tune the hyper-parameters and for early stopping.
- ▶ The outer optimisation loop tunes dropout rate, hidden layer size, minibatch size, learning rate, max gradient norm and CPD LBW length, with 50 iterations of random grid search.

# Performance Results



**Figure:** Benchmarking DMNs against *Intermediate* strategy  $w \in \{0, 0.5, 1\}$ , *Long Only* and *MACD*.

# Performance Results

	Returns	Vol.	Sharpe	Downside Deviation	Sortino	MDD	Calmar	% of +ve Returns	Ave. P Ave. L
<b>Reference</b>									
Long Only	2.30%	5.22%	0.44	3.59%	0.64	3.12%	0.79	52.45%	0.975
MACD	2.65%	3.58%	0.77	2.57%	1.09	2.56%	0.95	53.34%	1.002
<b>TSMOM</b>									
$w = 0$	4.41%	4.80%	0.94	3.44%	1.32	3.22%	1.35	54.28%	0.990
$w = 0.5$	3.29%	3.78%	0.89	2.80%	1.23	2.70%	1.16	53.88%	0.998
$w = 1$	2.17%	4.71%	0.48	3.29%	0.68	3.24%	0.67	51.48%	1.026
<b>LSTM</b>									
	3.53%	2.52%	1.62	1.71%	2.46	1.72%	2.79	55.23%	1.075
<b>LSTM w/ CPD</b>									
10 day LBW	3.04%	1.57%	1.77	<b>1.07%</b>	2.74	1.09%	2.78	55.50%	1.096
21 day LBW	<b>3.68%</b>	1.81%	2.04	1.21%	3.07	1.08%	<b>3.75</b>	<b>56.43%</b>	1.095
63 day LBW	3.51%	<b>1.72%</b>	2.08	1.10%	3.27	<b>1.06%</b>	3.58	55.61%	1.140
126 day LBW	3.37%	2.28%	1.75	1.59%	2.66	1.52%	2.88	54.95%	1.117
252 day LBW	2.81%	2.24%	1.45	1.57%	2.19	1.54%	2.32	54.00%	1.101
LBW Optimised	3.64%	1.73%	<b>2.16</b>	1.17%	<b>3.33</b>	1.14%	3.50	56.22%	<b>1.133</b>

Figure: Strategy performance benchmark for raw signal output.

# Slow Momentum with Fast Reversion

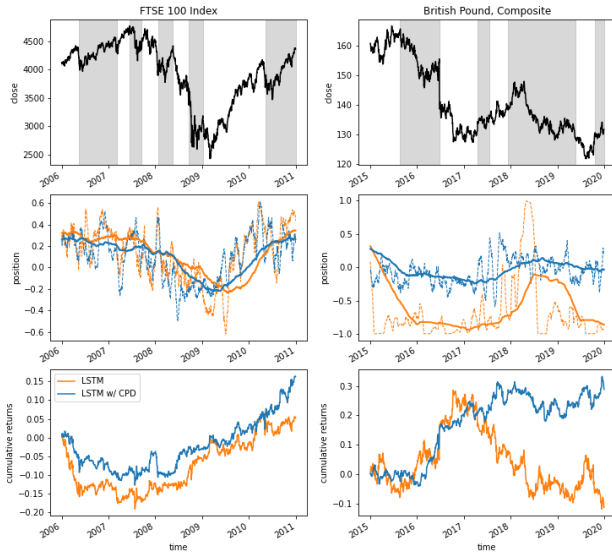


Figure: Slow momentum and fast reversion happening simultaneously.

# Momentum Transformer

# Transformers

- ▶ Based on the concept ‘attention is all you need’, doing away with convolutions and recurrent neural networks (RNNs).
- ▶ The attention-based architecture allows the network to focus on significant time steps in the past and longer-term patterns
- ▶ Have led to state-of-the-art performance in diverse fields, such as of natural language processing, computer vision, and speech processing (see *Lin et al.* [8]).
- ▶ Have recently have been harnessed for time-series modelling (*Li et al.* [9] *Lim et al.* [10], *Zhuo et al.* [11]).
- ▶ Naturally adapts to new market regimes, such as during the SARS-CoV-2 crisis.

# Attention and Self-Attention

- ▶ From a time series modelling perspective, attention provides two key benefits.
  1. Networks with attention are able to directly attend to any significant events that occur.
  2. Attention-based networks can also learn regime-specific temporal dynamics, by using distinct attention weight patterns for each regime (*Lim. et. al.* [10]).

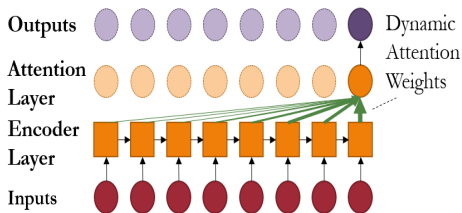


Figure: Attention model.



## Base Architectures Tested in the Momentum Transformer

- ▶ **Transformer:** (*Vaswani et al.* [12]) consists of encoder and decoder – each consisting of  $l$  identical layers of a (multi) self-attention mechanism, followed by a position-wise feed-forward network and a residual connection between these two components.
- ▶ **Decoder-Only Transformer:** (*Li et al.* [9]) only the decoder side.
- ▶ **Convolutional Transformer:** *Li et al.* [9] incorporates convolutional and log-sparse self-attention.
- ▶ **Informer Transformer:** *Zhuo et al.* [11] replaces the naive sparsity rule of the Conv. Transformer with a measurement based on the Kullback-Leibler divergence to distinguish essential queries, referred to as *ProbSparse* self-attention.
- ▶ **Decoder-Only Temporal Fusion Transformer (TFT):** an attention-LSTM hybrid which uses recurrent LSTM layers for local processing and interpretable self-attention layers for long-term dependencies. We consider the Decoder-Only version of the original TFT (*Lim et al.* [10]).

# Momentum Transformer (Decoder-Only TFT)

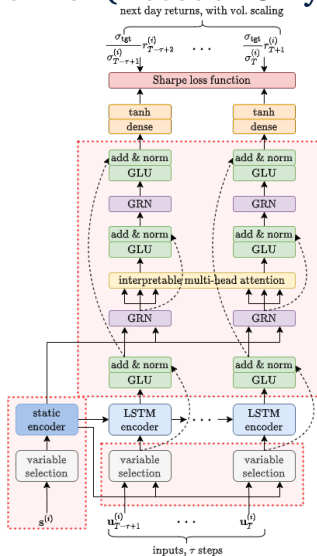


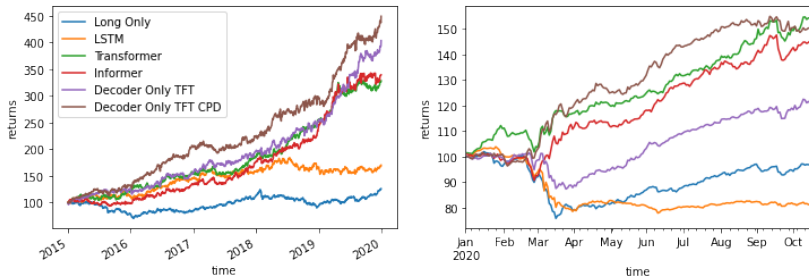
Figure: Decoder-Only TFT

# Results

	Returns	Vol.	Sharpe	Downside Deviation	Sortino	MDD	Calmar	% of +ve Returns	Ave. P Ave. L
<u>Average 2015–2020</u>									
Long-Only	1.73%	5.00%	0.37	3.59%	0.51	11.41%	0.15	51.97%	0.982
TSMOM	0.97%	4.38%	0.24	3.19%	0.33	8.25%	0.12	52.82%	0.931
LSTM	1.23%	1.85%	0.82	1.32%	1.19	3.55%	0.66	53.38%	1.004
Transformer	1.98%	1.29%	1.53	0.85%	2.32	1.07%	1.86	54.76%	1.071
Decoder-Only Trans.	1.37%	1.97%	0.72	1.37%	1.03	2.63%	0.60	52.76%	1.012
Conv. Transformer	1.85%	1.92%	0.98	1.30%	1.47	3.14%	0.77	52.93%	1.056
Informer	1.67%	1.09%	1.51	0.72%	2.30	1.17%	1.44	54.39%	1.089
Decoder-Only TFT	1.99%	1.23%	1.71	0.82%	2.61	1.17%	2.06	55.72%	1.073
Decoder-Only TFT CPD	<b>2.06%</b>	<b>1.02%</b>	<b>2.00</b>	<b>0.66%</b>	<b>3.10</b>	<b>0.82%</b>	<b>2.53</b>	<b>55.74%</b>	<b>1.120</b>
<u>SARS-CoV-2</u>									
Long-Only	-1.46%	6.73%	-0.19	5.64%	-0.22	12.32%	-0.12	57.28%	0.720
TSMOM	0.90%	4.73%	0.21	3.14%	0.32	4.17%	0.22	50.00%	1.041
LSTM	-4.15%	2.82%	-1.50	2.52%	-1.67	5.35%	-0.78	52.29%	0.643
Transformer	4.42%	<b>1.28%</b>	<b>3.38</b>	<b>0.83%</b>	<b>5.55</b>	<b>0.84%</b>	7.31	<b>64.85%</b>	1.066
Decoder-Only Trans.	<b>8.02%</b>	2.58%	3.01	1.42%	<b>5.55</b>	1.05%	<b>8.56</b>	58.83%	<b>1.243</b>
Conv. Transformer	3.13%	1.99%	1.81	1.40%	2.74	1.61%	3.17	57.48%	1.058
Informer	4.30%	1.60%	2.71	1.00%	4.45	1.07%	4.28	59.61%	1.137
Decoder-Only TFT	1.81%	1.75%	1.22	1.37%	1.74	2.14%	1.57	60.39%	0.831
Decoder-Only TFT CPD	3.39%	1.51%	2.47	1.03%	4.08	1.15%	5.92	59.90%	1.068

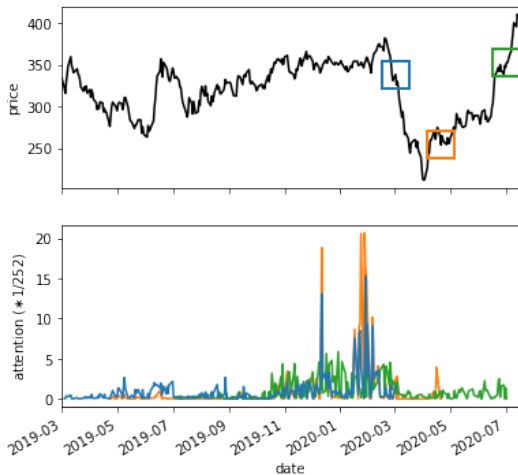
Figure: Strategy Performance Benchmark – Raw Signal Output

# Results



**Figure:** These plots benchmark our strategy performance for the 2015–2020 scenario (left) and the SARS-CoV-2 scenario (right). For each plot we start with \$100 and we re-scale returns to 15% volatility. Since we ran each experiment five times, we plot the repeat which resulted in the median Sharpe ratio, across the entire experiment.

## Attention Patterns



**Figure:** Lumber future price during SARS-CoV-2 crisis and the associated attention pattern when making a prediction at 1 March 2020 (blue), 21 April 2020 (orange), and 2 July 2020 (green).

# Attention Patterns

- ▶ We observe significant structure in attention patterns.
- ▶ The attention on momentum turning points is pronounced, segmenting the time series into regimes.
- ▶ Our model focuses on previous time-steps which are in a similar regime.

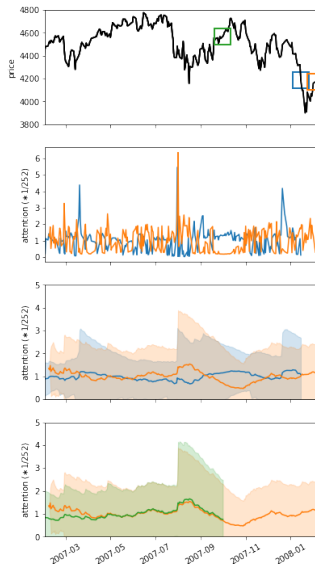
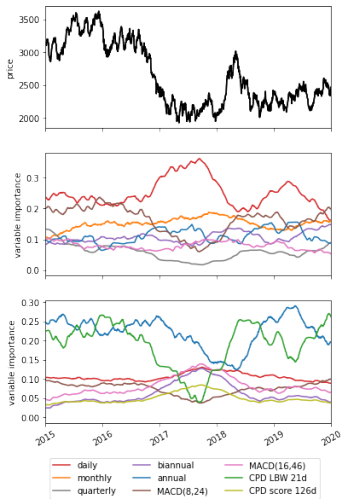


Figure: FTSE 100 future prior to 2008.

# Variable Importance

- ▶ Our model intelligently blends different classical strategies at different points in time.
- ▶ We observe that the strategy changes with the addition of CPD, placing left emphasis on returns at timescales in between daily (shortest) and annual (longest).



**Figure:** Variable importance for Cocoa future for Decoder-Only TFT (middle) and with CPD (bottom).

# Transaction Cost Impact

$C$	0bps	0.5bps	1bps	1.5bps	2bps	2.5bps	3bps
<b>LSTM</b>							
Indv. CM	0.12	0.09	0.05	0.01	-0.02	-0.06	-0.10
Indv. EQ	<b>0.37</b>	0.32	0.27	0.22	0.16	0.11	0.06
Indv. FI	0.09	-0.11	-0.32	-0.53	-0.74	-0.94	-1.15
Indv. FX	0.11	0.01	-0.08	-0.18	-0.27	-0.37	-0.46
Portfolio	0.82	0.51	0.20	-0.12	-0.43	-0.74	-1.05
<b>Transformer</b>							
Indv. CM	0.27	0.23	0.19	0.15	0.11	0.08	0.04
Indv. EQ	<b>0.37</b>	<b>0.33</b>	<b>0.28</b>	<b>0.23</b>	<b>0.19</b>	<b>0.14</b>	<b>0.10</b>
Indv. FI	0.23	0.03	<b>-0.16</b>	<b>-0.35</b>	<b>-0.55</b>	<b>-0.74</b>	<b>-0.93</b>
Indv. FX	-0.17	-0.24	-0.32	-0.39	-0.47	-0.55	-0.62
Portfolio	1.53	1.26	0.99	0.72	<b>0.45</b>	<b>0.18</b>	<b>-0.09</b>
<b>Informer</b>							
Indv. CM	0.28	0.24	0.19	0.15	0.10	0.06	0.01
Indv. EQ	0.34	0.28	0.22	0.16	0.10	0.04	-0.02
Indv. FI	0.08	-0.13	-0.35	-0.56	-0.78	-0.99	-1.20
Indv. FX	-0.14	-0.24	-0.33	-0.43	-0.53	-0.62	-0.72
Portfolio	1.51	1.17	0.83	0.49	0.15	-0.19	-0.53
<b>Decoder-Only TFT</b>							
Indv. CM	0.44	0.40	0.35	0.31	0.26	0.22	0.17
Indv. EQ	0.25	0.19	0.13	0.07	0.02	-0.04	-0.10
Indv. FI	<b>0.30</b>	<b>0.05</b>	-0.20	-0.45	-0.69	-0.94	-1.18
Indv. FX	<b>0.28</b>	<b>0.18</b>	<b>0.08</b>	<b>-0.02</b>	<b>-0.12</b>	<b>-0.22</b>	<b>-0.32</b>
Portfolio	1.71	1.36	1.01	0.67	0.32	-0.03	-0.37
<b>Decoder-Only TFT CPD</b>							
Indv. CM	<b>0.55</b>	<b>0.50</b>	<b>0.45</b>	<b>0.40</b>	<b>0.35</b>	<b>0.30</b>	<b>0.25</b>
Indv. EQ	0.18	0.12	0.05	-0.01	-0.07	-0.14	-0.20
Indv. FI	0.23	-0.03	-0.29	-0.55	-0.81	-1.07	-1.33
Indv. FX	0.24	0.13	0.02	-0.09	-0.20	-0.30	-0.41
Portfolio	<b>2.00</b>	<b>1.61</b>	<b>1.22</b>	<b>0.83</b>	0.44	0.04	-0.35

Figure: Transaction cost impact on Sharpe over 2015–2020 for individual assets, averaged by asset class, and for diversified portfolio.



## Conclusions

- ▶ Deep Momentum Networks are novel models which directly output trading signals which are optimised for Sharpe ratio
- ▶ The original deep Momentum Networks based on LSTMs perform well by exploiting a blend of momentum and mean reversion
- ▶ We introduce Changepoint detection to this model to more intelligently adapt to changes from trending to more reverting regimes
- ▶ We further improve the model by considering transformer based architectures
- ▶ The attention-based architectures, which we tested, are robust to significant events, such as during the SARS-CoV-2 market crash and tend to focus less on mean-reversion and more on longer term trends.

Thanks!

Papers:

Time Series Forecasting With Deep Learning:

A Survey [2004.13408]

Deep Momentum Networks [1904.04912]

DMNs with Changepoints [2105.13727]

Momentum Transform [2112.08534]

stefan.zohren@eng.ox.ac.uk

- [1] T. J. Moskowitz, Y. H. Ooi, and L. H. Pedersen, "Time series momentum," *Journal of Financial Economics*, vol. 104, no. 2, pp. 228 – 250, 2012. Special Issue on Investor Sentiment.
- [2] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of Finance*, vol. 48, no. 1, pp. 65–91, 1993.
- [3] A. Y. Kim, Y. Tse, and J. K. Wald, "Time series momentum and volatility scaling," *Journal of Financial Markets*, vol. 30, pp. 103 – 124, 2016.
- [4] J. Baz, N. Granger, C. R. Harvey, N. Le Roux, and S. Rattray, "Dissecting investment strategies in the cross section and time series," *SSRN*, 2015.
- [5] A. Garg, C. L. Goulding, C. R. Harvey, and M. Mazzoleni, "Momentum turning points," *Available at SSRN 3489539*, 2021.
- [6] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.
- [7] C. K. Williams and C. E. Rasmussen, "Gaussian processes for regression," 1996.
- [8] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of Transformers," *arXiv preprint arXiv:2106.04554*, 2021.
- [9] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of Transformer on time series forecasting," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 5243–5253, 2019.
- [10] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, 2021.
- [11] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient Transformer for long sequence time-series forecasting," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, vol. 35, pp. 11106–11115, AAAI Press, 2021.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.