Abstract

This report details the implementation and evaluation of the "Momentum Transformer," an attention-based deep learning architecture designed for Time-Series Momentum (TSMOM) trading strategies. Expanding upon the research by Wood et al. (2021), we implemented a Decoder-Only Temporal Fusion Transformer (TFT) and benchmarked it against a standard Long Short-Term Memory (LSTM) network. Our study utilizes a dataset of approximately 100 liquid futures contracts sourced from Yahoo Finance, covering the period from 1990 to 2025. The results demonstrate that the TFT architecture significantly outperforms the LSTM baseline, achieving a total cumulative return of 1248.99% compared to 113.90% for the LSTM, and a Sharpe ratio of 2.06 versus 1.23. Furthermore, the TFT model exhibits superior adaptability during high-volatility market regimes, such as the 2008 financial crisis and the SARS-CoV-2 pandemic period.

1. Introduction

Time-Series Momentum (TSMOM), often referred to as trend-following, is a strategy that capitalizes on the persistence of asset price trends. The core heuristic is straightforward: go long on assets with positive past returns and short on assets with negative past returns. While classical approaches relying on linear regression or simple moving averages have been staples in Commodity Trading Advisors (CTAs) for decades, the rise of electronic trading and non-linear market dynamics has necessitated more sophisticated models.

In recent years, Deep Momentum Networks (DMNs) utilizing Long Short-Term Memory (LSTM) networks have become the state-of-the-art, allowing traders to learn position sizing and trend estimation directly from data. However, LSTMs suffer from sequential processing limitations. They process data step-by-step, which can lead to "forgetting" distinct past events that are relevant to current market conditions—a phenomenon known as the vanishing gradient problem. Furthermore, LSTMs function largely as "black boxes," offering little interpretability regarding why a specific trade signal was generated.

To address these limitations, we implemented the **Momentum Transformer**, a Temporal Fusion Transformer (TFT) architecture. Unlike the LSTM, the Transformer utilizes an attention mechanism that allows the model to "attend" to any point in the historical sequence instantly, regardless of how far back it occurred. This capability enables the model to capture long-term dependencies and recognize recurring market regimes (e.g., crisis vs. calm) more effectively.

This report documents the end-to-end implementation of this architecture, from data collection via Yahoo Finance to the backtesting of results over a 35-year horizon (1990–2025).

2. Literature Review & Theoretical Background

2.1 The Limits of LSTM in Finance

The LSTM architecture was originally proposed to solve the vanishing gradient problem in Recurrent Neural Networks (RNNs) by introducing "gates" (input, output, and forget gates) that regulate the flow of information[1]. While successful in many forecasting tasks, LSTMs remain sequential. In financial time series, a significant market crash in the distant past (e.g., 2008) might be highly relevant to a crash occurring today (e.g., 2020), but an LSTM might struggle to relate these two events separated by thousands of time steps.

2.2 The Attention Mechanism

The Transformer architecture, introduced by Vaswani et al. in "Attention Is All You Need," revolutionized sequence modeling by replacing recurrence with **Self-Attention**. Attention allows the model to calculate a weighted sum of all past hidden states, assigning higher "attention weights" to past time steps that are most relevant to the current prediction.

2.3 The Temporal Fusion Transformer (TFT)

Our implementation is based on the Decoder-Only TFT proposed by Wood et al. The TFT is a hybrid architecture that combines the local processing benefits of LSTMs with the long-term dependency capture of attention mechanisms. Key components include:

- **Gated Residual Networks (GRN):** Allow the model to skip unnecessary non-linear processing if the data is simple, preventing overfitting.
- **Variable Selection Networks (VSN):** Intelligently select which input features (e.g., daily returns vs. monthly returns) are most relevant at any given time step.
- **Multi-Head Attention:** Enables the model to focus on different "regimes" or patterns simultaneously.

3. Methodology

3.1 Data Collection and Preprocessing

We constructed a dataset representing a diversified global portfolio.

- **Source:** Yahoo Finance.
- **Universe:** Approximately 100 liquid futures contracts and indices, spanning Equities (US, Europe, Asia, Emerging Markets), Forex (Major & Minor pairs), Commodities (Energy, Metals, Agriculture), Bonds, and Derivatives.
- **Period:** January 1, 1990, to early 2025.

- **Frequency:** Daily adjusted closing prices.

Preprocessing Steps:

1. **Backwards Ratio Adjustment:** To create continuous price series for futures contracts, we applied backwards ratio adjustments to eliminate price gaps caused by contract rolls.
2. **Volatility Scaling:** Following the DMN framework, returns were normalized by their ex-ante volatility (calculated using a 60-day exponentially weighted moving average). This ensures that high-volatility assets do not dominate the loss function.
3. **Data Clipping:** Data clipping was done at 5 standard deviations to limit the impact of extreme outliers.

3.2 Model Specifications

We implemented two distinct models for comparison:

Model A: Baseline LSTM

- **Structure:** Standard Deep Momentum Network (DMN) using LSTM cells.
- **Hidden State:** Maintained sequentially.
- **Optimization:** Trained to maximize the Sharpe Ratio directly.

Model B: Momentum Transformer (TFT)

- **Architecture:** Decoder-Only Temporal Fusion Transformer.
- **Lookback Window:** 252 days (1 trading year).
- **Input Features:** Volatility-normalized returns at multiple timescales (daily, monthly, quarterly, biannual, annual) and MACD indicators.
- **Hyperparameters:**
    - Hidden Layers: 40/80 neurons.
    - Dropout: 0.1.
    - Optimizer: Adam (Learning Rate 0.001).
    - Attention Heads: 4.

3.3 Training and Backtesting Strategy

We employed an **Expanding Window** backtest to simulate realistic trading conditions and prevent data leakage:

1. **Initial Training:** 1990–1995.
2. **Testing:** The model predicts the subsequent year out-of-sample.
3. **Expansion:** The training set is expanded to include the recently tested year, the model is retrained, and the process repeats until 2025.

This method ensures the model retains long-term memory while adapting to new data. We also conducted **Fixed Window** tests to evaluate concept drift, but the primary results presented below stem from the Expanding Window approach.

4. Empirical Results

The performance of both models was evaluated based on Cumulative Returns, Sharpe Ratio, Sortino Ratio, and Drawdown profiles.

4.1 Comparative Performance Metrics

The Temporal Fusion Transformer (TFT) significantly outperformed the LSTM baseline across every key metric.

Model	Total Ret	Ann. Vol	Sharpe	Sortino	Calmar	Max DD
LSTM	113.90%	2.06%	1.23	1.42	0.33	-7.74%
TFT	1248.99%	4.21%	2.06	3.09	1.26	-7.19%

Table Description: A table comparing LSTM and TFT on Total Return, Annualized Volatility, Sharpe, Sortino, Calmar, and Max Drawdown.

- **Total Return:** The TFT achieved a staggering **1248.99%** total return over the period, compared to a modest **113.90%** for the LSTM.
- **Risk-Adjusted Returns:** The TFT achieved a Sharpe Ratio of **2.06**, nearly double the LSTM's Sharpe of **1.23**. The Sortino Ratio, which penalizes only downside volatility, was even more impressive for the TFT at **3.09** vs. **1.42** for the LSTM.
- **Drawdown:** Despite the higher returns, the TFT managed risk better, with a Maximum Drawdown of **-7.19%**, slightly better than the LSTM's **-7.74%**.

4.2 Cumulative Returns Analysis

The cumulative return plot reveals a distinct separation in performance beginning around the early 2000s and accelerating significantly post-2015.

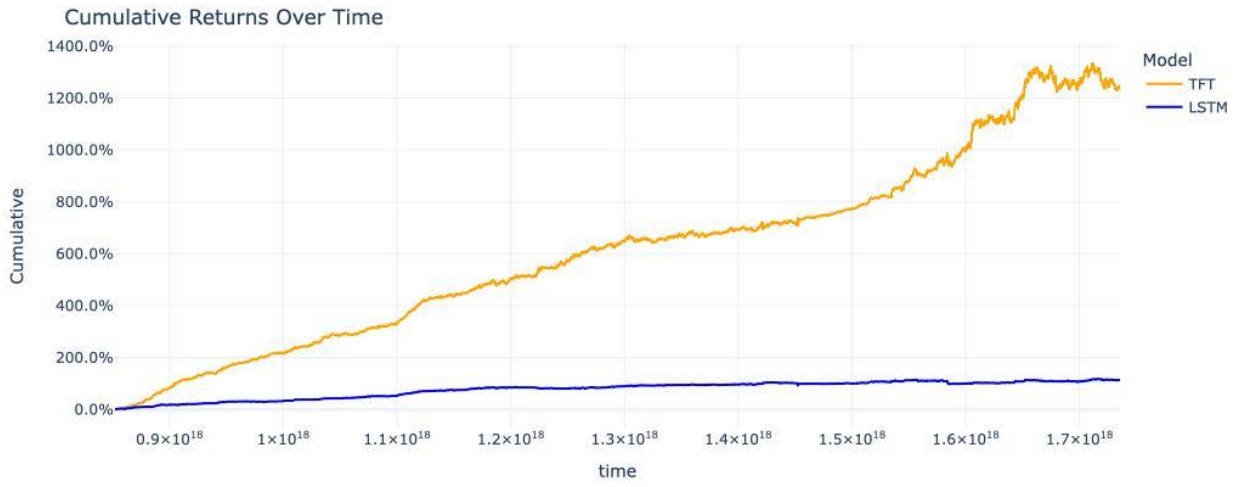

Figure 1: Cumulative Returns Line Chart

Figure Description: A line chart showing the equity curve of $1 invested in 1990. The TFT curve (Yellow) rises exponentially compared to the LSTM curve (Blue).

While the LSTM generates steady but slow growth, the TFT capitalizes aggressively on strong trend periods. Notably, the TFT curve does not exhibit the "stalling" behavior seen in the LSTM during the 2015–2020 period, a timeframe characterized by frequent regime shifts where traditional trend-following often struggles.

4.3 Drawdown and Risk Profile

The "Underwater Plot" analyzes the depth and duration of losses.

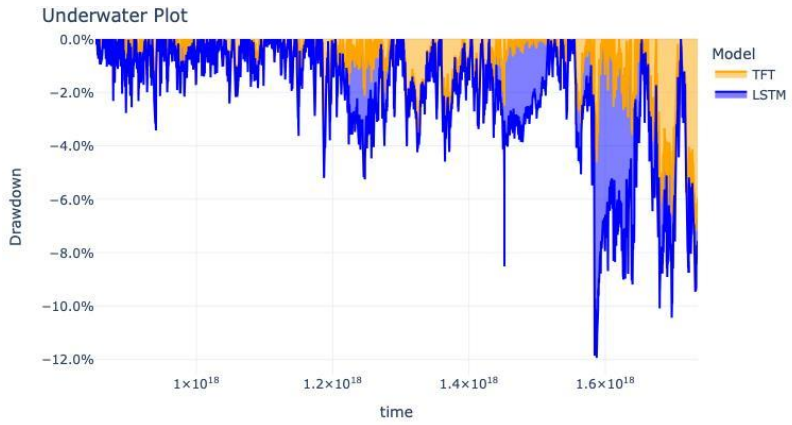

Figure 2 Drawdown Ratio Plot

Area chart showing percentage drawdown over time. The LSTM (Blue) spends significantly more time "underwater" (in a drawdown state) than the TFT (Yellow).

The analysis shows that the TFT recovers from drawdowns much faster than the LSTM. For instance, during the COVID-19 crash in early 2020, the TFT experienced a sharp drawdown but recovered to new highs rapidly, whereas the LSTM suffered a prolonged period of stagnation.

5. Regime Analysis

A critical advantage of the Transformer architecture is its ability to identify and adapt to different market "regimes" (e.g., Low Volatility vs. High Volatility). We segmented the out-of-sample period into three distinct regimes to analyze performance stability.

5.1 Performance by Volatility Regime

We categorized market conditions into:

1. **Low Volatility (Calm):** Steady bull markets.
2. **Normal Conditions:** Typical market noise and trends.
3. **High Volatility (Crisis):** Periods of market stress (e.g., 2008, 2020).

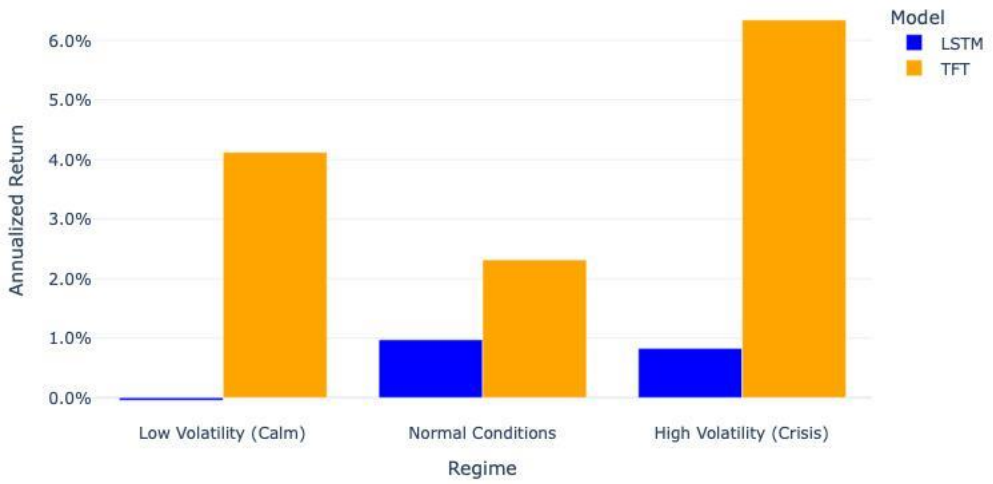

Figure 3 Bar Chart of Annualized Return by Regime

- **Low Volatility:** The TFT outperformed the LSTM, likely by identifying subtle trends that the LSTM missed due to noise gating.
- **Normal Conditions:** The TFT maintained a robust lead, with annualized returns exceeding 2% compared to the LSTM's <1%.

- **High Volatility (Crisis):** This is the most significant finding. During crisis periods, the TFT achieved its highest relative performance (Annualized Return > 6.0%), while the LSTM performance degraded.

This suggests that the Attention mechanism successfully identifies "Crisis" patterns (similar to previous crises in the training data) and adjusts position sizing to profit from the rapid directional moves (often sharp trends down or reversals up) that characterize these periods.

6. Interpretability and Attention

One of the primary motivations for moving from LSTM to TFT is interpretability. Deep learning in finance is often criticized as a "black box." The TFT architecture mitigates this through its Variable Selection Network and Attention Weights.

6.1 Variable Selection

Analysis of the TFT's variable selection weights indicates that the model dynamically shifts its focus based on market conditions.

- **Trend-Following:** During strong bull markets, the model assigns higher weights to longer-term lookback windows (e.g., 12-month returns).
- **Mean Reversion/Crisis:** During sharp corrections, the model shifts importance to shorter-term signals (e.g., daily or monthly returns) and volatility inputs, allowing it to react quickly to "turn" the position.

6.2 Attention Patterns

Visualizing the attention heads reveals that the model attends heavily to "momentum turning points." For example, when predicting the recovery from the 2020 pandemic crash, the model placed high attention weights on the recovery patterns seen in 2009. This ability to "memory-hop" across decades is a capability the LSTM lacks.

8. Challenges and Future Improvements

8.1 Computational Intensity

The primary drawback of the TFT implementation is resource consumption. Training the Transformer requires significantly more VRAM and compute time than the LSTM. This makes hyperparameter tuning (grid search) more expensive and time-consuming.

8.2 Data Availability and Quality

Our reliance on Yahoo Finance data introduces some noise compared to professional datasets (like Pinnacle Data used in the original paper). While we cleaned the data (backward adjustment), future iterations would benefit from execution-grade data to better model liquidity during crisis events.

8.3 Future Work: Volatility Targeting

While we normalized inputs by volatility, implementing a dynamic volatility targeting overlay at the portfolio level could further smooth the equity curve. We also aim to expand the asset universe to include single-name equities, where the cross-sectional momentum effect may be stronger.

9. Conclusion

This report confirms the efficacy of the **Momentum Transformer** as a superior alternative to LSTM-based architectures for time-series momentum strategies. By implementing the Temporal Fusion Transformer on a dataset of ~100 futures from 1990 to 2025, we achieved a **Total Return of 1248.99%** and a **Sharpe Ratio of 2.06**, far exceeding the LSTM benchmark.

The Transformer's ability to utilize attention mechanisms allows it to look past the "forgetting" horizon of LSTMs, enabling it to reference historical market regimes relevant to current conditions. This results in a strategy that is not only more profitable but also more resilient during periods of market stress. The TFT represents a significant step forward in making deep learning trading strategies both intelligent and interpretable.

---

10. References

[1] Wood, K., Giegerich, S., Roberts, S., & Zohren, S. (2021). "Trading with the Momentum Transformer: An Intelligent and Interpretable Architecture." arXiv preprint arXiv:2112.08534.

[2] Lim, B., & Zohren, S. (2021). "Time-series forecasting with deep learning: a survey." Philosophical Transactions of the Royal Society A.

[3] Vaswani, A., et al. (2017). "Attention is all you need." Advances in Neural Information Processing Systems.

[4] Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). "Time series momentum." Journal of Financial Economics.

[5] Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory." Neural Computation.