UDC 004.05

Forecasting Information Operations with Hybrid Transformer Architecture

Anatolii Feher[1]

[1] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», 37, Prosp. Beresteiskyi, Kyiv, 03056, Ukraine*

Abstract

Proactive decision-making in all processes is difficult to imagine without forecasting methods, especially in the field of cybersecurity where the speed and quality of response are often critical. For this reason, we proposed a unique methodology based on a new hybrid architecture Transformer that perfectly captures long-term dependencies and an adaptive algorithm ACWA that quantifies historical patterns. Thus, the described approach considers short-term fluctuations, long-term trends, and seasonal patterns more effectively than traditional forecasting models, as demonstrated by the application of Information Operations and Disinformation occurrences time series forecasting.

Keywords: time series forecasting, transformer models, adaptive contextual weighted average, OSINT

Introduction

Difficult to underestimate the importance of information security in the industry of our lives, and as an industry associated with high risks at every level of operation in our lives, where the priority has always been the need to be able to identify patterns, trends, and anomalies in advance, this is what the forecasting area helps with. Traditional methods like Autoregressive Integrated Moving Average (ARIMA) have been actively and efficiently used for decades [1], but due to their specific nature, such analytical approaches often face limitations, variability, and complexity, which significantly reduces the reliability of forecasts.

Recent advancements in sequence modeling, particularly transformer architectures, have revolutionized processings by leveraging tokenization and attention mechanisms to interpret contextual dependencies [2]. While recurrent neural networks like Long Short-Term Memory (LSTM) have been common for time series analysis, transformers excel in processing long sequences without recurrence but struggle with overfitting on smaller datasets [3]. Weighted average approaches, offer an adaptive solution by weighting historical data based on confidence and time decay, making them effective for capturing non-stationary phenomena with long-term dependencies and sudden changes [4].

However, the inherent volatility and non-linearity of real-world captured datasets demand more sophisticated tools capable of balancing short and long-term accuracy and interpretability. To combine the benefits of broad contextual awareness and adaptive weighting, a hybrid approach integrating Adaptive Contextual Weighted Average (ACWA) with Transformer models was developed, where Transformers provide a powerful mechanism for detecting long-term dependencies, ACWA promotes contextual weighting of historical events, thereby enhancing already short-term responsiveness. This integration not only improves predictive accuracy but also enables dynamic adjustments in the face of rapidly changing patterns, making it suitable for applications requiring high resilience.

Experiments on collected Open Source Intelligence (OSINT) datasets demonstrated that the hybrid architecture delivers a robust, accurate, and flexible forecasting system. Initial systems testing was made on circular functions and achieved accurate results bringing us to perform an actual application on actual Cybersecurity related datasets. Where that related scenario can be strongly characterized by uneven temporal distributions and abrupt changes in event frequency, reflecting the noise in real-world challenges of anomaly detection and trend forecasting providing a challenging environment for analysis and method's research.

1. Methodology

Selected time series for this study represents a complex relationship between the occurrence of specific events, gathered using Open Source Intelligence technology, over one-year, two-years, and three-years intervals, creating lesser 334, 700, and bigger 1064 days observed datasets which were allocated for training the models, while the remaining 30 days, representing the final month, were reserved for prediction and subsequent analysis. The high Hurst exponent values that were calculated with R/S analysis for both series confirmed strong persistence, indicating predictability based on historical trends. The values varied from a smaller H≈0.71 and H≈0.76 to a bigger dataset up to H≈0.77.

Collected events inside formed datasets include values for the frequency of mentions of terms such as "Information Operation" and "Disinformation" within the context of Ukraine in online sources like news articles, blogs, and other internet platforms. Data for the time series was sourced primarily from Infostream resources for statistical analysis, additionally, preprocessing steps ensured uniformity in data intervals while mitigating noise caused by sudden surges or gaps in event reporting.

Inspired by tokenization approaches in natural language processing [5], the method transformed the time series into sequences by dividing values into unequal ranges using quantiles, with an adaptive upper bound ensuring robustness to future data variations, employed ACWA method as an attention-inspired pattern search, assigning dynamic weights based on frequency and recency, normalized to emphasize significant patterns [6]. A custom attention mechanism with dynamic ACWA weights improved pattern recognition, while residual connections, feedforward layers, and attention dropout enhanced stability and temporal dependency capture.

By merging raw data with ACWA-derived attention, the methodology provides a robust framework for addressing the challenges of time series forecasting, showcasing the synergy between statistical pattern recognition and transformer architectures.

Presented approach was benchmarked against ARIMA, LSTM, and regular Transformer models, revealing consistent superiority in identifying and predicting nuanced temporal patterns critical for operational decision-making.

2. Framework

Transformers, renowned for their attention mechanisms, excel at handling long dependencies without explicit repetition, however, when dealing with small or noisy datasets, they risk overfitting, necessitating a reduction in layers and heads. To address this, we applied a simplified transformer architecture, modifying the self-attention mechanism to incorporate ACWA-based weights, thereby prioritizing critical time steps, and called such approach - ChronoTensor.

In turn, adaptive pattern weighting approaches, especially ACWA, have proven promising in non-stationary environments, as ACWA algorithm systematically assigns dynamic weights that depend on researched pattern frequency and time decay, allowing it to quickly adapt to new modes or signals.

Such a method systematically collects pattern occurrences from historical data and assigns dynamic weights based on the frequency and time decay of each pattern. The adaptive nature of ACWA has brought tremendous value to time series use, enabling the model to focus on current or relevant historical intervals most likely to shape future behavior.

Thus, the hybrid approach described above combined two distinct but complementary components, by tokenizing the data, assigning robust contextual weights using ACWA, and feeding this knowledge into the transformation pipeline, we combined local reactivity with global contextual awareness. This synergy proved particularly effective for OSINT collected data, where sudden changes of captured mentions of researched frequency for "Information Operations" and "Disinformation" can have the same impact as underlying multi-week or multi-month trends.

2.1. Pattern Recognition

Tokenization, inspired by text processing in NLP, partitions the time series $X = \{x_1, x_2, \ldots, x_T\}$ into n unequal ranges (tokens) such that each token contains an approximately equal number of data points. Token boundaries are determined using quantiles:

$$q_i = F^{-1}\left(\frac{i}{n}\right), \quad (1)$$

where F^{-1} is the inverse Cumulative Distribution Function (CDF), each value in the

series is assigned to the corresponding token. An adaptive upper bound ensures robustness to new data values, calculated as:

$$upper\_bound = max(X) + k\sigma \quad (2)$$

where σ is the standard deviation of the series and k is a scaling coefficient. This approach ensures that the model remains resilient to outliers and future extreme values.

The ACWA method extends the tokenization process by identifying patterns in historical data. For a given value x_T, relevant patterns are searched using tokens $\tau(x_i)$ corresponding to each value x_i. Matched patterns are weighted dynamically, with weights w_i assigned based on confidence scores and recency:

$$w_i = c_p \times e^{-\lambda \Delta t_i}, \quad (3)$$

where c_p is a confidence score derived from the frequency of pattern p, Δt_i is the time difference between the pattern's occurrence and the current time, and λ is a decay rate parameter prioritizing recent patterns. Weights are normalized to ensure they sum to 1:

$$w_i = \frac{w_i}{\sum_{j=1}^{N} w_j}, \quad (4)$$

The ACWA prediction $\hat{x}^{\mathrm{ACWA}}$ is computed as a weighted average of subsequent values following matched patterns:

$$\hat{x}^{\mathrm{ACWA}} = \sum_{i=1}^{N} w_i x'_i, \quad (5)$$

where x'_i represents the value immediately following the *i*-th matched pattern.

Multiple pattern lengths (3, 5, 7, 9) were tested iteratively to capture varying temporal dependencies. The model was retrained and evaluated for each configuration to identify the optimal pattern length, ensuring the approach remained adaptable to diverse forecasting scenarios and applications

Additionally, the tokenization process was adjusted to handle multi-dimensional data where it's not used within the chosen dataset it has been foreseen to use when it's applicable, dictionary was modified to accommodate extended patterns.

2.2. Attention Mechanism

To utilize the transformer's capabilities for sequential data processing, the ACWA method was integrated into a simplified transformer architecture, where model was adjusted to mitigate overfitting by reducing the number of layers and tuning hyperparameters such as the learning rate and dropout rate. Enhancements were introduced to the attention mechanism to incorporate ACWA predictions effectively.

The model input combines normalized time series values and ACWA predictions, expressed as:

$$X_{input} = [x_t, \hat{x}_t^{\mathrm{ACWA}}], \qquad t = 1, 2, \ldots, T, \quad (6)$$

where x_t is the normalized time series value and $\hat{x}_t^{\mathrm{ACWA}}$ is the corresponding ACWA forecast.

These inputs are first converted through a token embedding layer, mapping them to a higher-dimensional space optimized for use with the transformer model. To encode temporal relationships, positional information is added to these embeddings, compensating for the transformer's inability to inherently recognize sequence order. This combined embedding is formulated as:

$$E_t = Embedding\left(X_{input,t}\right) + PositionalEncoding(t), \quad (7)$$

The transformer's multi-head self-attention mechanism was modified to integrate dynamic pattern weights derived from the ACWA method, enabling it to better capture significant temporal dependencies. The self-attention computation was adjusted as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}} + W,\right)V, \quad (8)$$

where Q, K, and V are query, key, and value matrices, d_k is the dimension of the key vectors, and W incorporates ACWA-based weights.

These weights prioritize patterns with higher confidence scores and recency, aligning the model's focus with the most relevant sequences. This adjustment enhances the model's ability to recognize critical patterns and improve forecasting accuracy.

2.3. Model Processing

In addition an attention dropout layer was incorporated to improve generalization and reduce overfitting by randomly deactivating certain attention connections during training. The normalized attention output is processed by a simplified position-wise feedforward network:

$$MLP(X) = ReLU (XW_1 + b_1)\, W_2 + b_2, \quad (9)$$

where W_1, W_2 are weight matrices, and b_1, b_2 are weights and biases of the feedforward layers. Residual connections propagate benefits across the encoder, enhancing learning of complex representations. For time series forecasting, an encoder-only architecture simplifies the model by omitting the decoder stack, focusing on input dependencies without autoregressive decoding.

The output from the final encoder layer is passed through a linear projection layer to map the model's hidden states back to the target dimension, then combine the transformer's prediction with the weighted ACWA prediction to produce the final forecast:

$$\hat{x}_{T+1}^{\text{transformer}} = Linear\,(EncoderOutput), \quad (10)$$

The final transformer output, $\hat{x}_{T+1}^{\text{transformer}}$, is blended with the ACWA prediction, $\hat{x}_{T+1}^{\text{ACWA}}$, to produce the final forecast:

$$\hat{x}_{final} = \alpha\, \hat{x}_{T+1}^{\text{transformer}} + (1 - \alpha)\, \hat{x}_{T+1}^{\text{ACWA}}, \quad (11)$$

where α is a blending coefficient determined through validation. This combination leverages the global sequence modeling of the transformer and the local adaptability of ACWA.

Gradient clipping (max_norm = 1.0) was applied to prevent exploding gradients, and *L*2 regularization (weight_decay = 1 × 10−5) minimized overfitting. The encoder-only transformer design, omitting the decoder, reduced complexity and focused on dependency modeling within the input sequence.

These optimizations preserved the adaptive nature of ACWA while harnessing the transformer's representational power, enabling robust and accurate time series forecasting.

3. Results and Discussion

The software for each method was developed with algorithms tailored to analyze OSINT data and forecast trends. Time series data was prepared through cleaning and formatting for compatibility with all models. Figure 1 shows prediction results for smaller (364, one-year) and larger (1094, three-years) datasets, providing a clear visual comparison between predicted within actual series and between methods.

The inherent noise in observed OSINT collected datasets, characterized by sporadic spikes driven by real-world events, posed a challenge for traditional models. ACWA integration mitigates this by emphasizing recent and frequent patterns and dynamically discounting outdated signals. This allows for adaptation in real-time, offering a tailored approach well-suited to threat intelligence forecasting needs.

The process was repeated multiple times and refined to ensure the generation of the most accurate median predictions, enabling a comprehensive efficiency comparison across the models represented in Table 1, which provides averaged over 20 runs results of Root Mean Squared Error (RMSE) for each observed model.

Table 1
RMSE Comparison Across Models

	One-year	Two-years	Three-years
ARIMA	72.943	69.555	66.910
LSTM	77.763	77.900	78.997
Transformer	75.362	76.813	77.775
ChronoTensor	59.087	61.070	62.038

From the achieved we see that ChronoTensor demonstrates improvements in short-range forecasts estimating an average enhancement of 18.64% emphasizing the synergy between ACWA's local adaptability and the transformer's global sequence modeling. Such advantage is particularly evident in scenarios with abrupt shifts or spikes, such as changes in "Disinformation" and "Information Operation" activity patterns.

For larger datasets ChronoTensor retains the lowest RMSE, although its relative improvement over standard transformers diminishes, this trend suggests that longer time horizons and increased noise levels may reduce the localized weighting advantage, particularly with higher embedding dimensions or additional layers.

Figure 1: The rows show applied forecast methods by chosen dataset: actual time series – blue, predicted – red, last month segment – grey cut, and truncated previous years for better representation – light blue.

Nonetheless, the ability to generalize effectively across temporal scales underscores the strength of combining global attention mechanisms with adaptive local pattern weighting, which has demonstrated even more remarkable improvements in accuracy when applied to specific datasets and applications as were initially mentioned while testing on circular functions.

While ChronoTensor incorporates gradient clipping and layer normalization to address these issues [7], future iterations may benefit from advanced stabilization techniques, such as tailored weight initialization or adaptive optimizers, to maintain performance on more complex architectures, as well as highlight the need for additional automatization for hyperparameter tuning [8] that would level up developed system. Furthermore, integrating meta-learning frameworks could enhance adaptability across varying temporal datasets.

Conclusions

Presented a hybrid forecasting framework ChronoTensor which combined the Adaptive Contextual Weighted Average method with a Transformer architecture to address OSINT time series collections with long memory and abrupt shifts, demonstrated that ACWA with its adaptive nature brought remarkable value in highlighting and weighting relevant historical patterns, while the transformer effectively modeled global sequence dependencies. By blending both outputs, we achieved better predictive accuracy in comparison to reviewed regular Transformer, standard deep learning (LSTM), and statistical (ARIMA) baselines, especially on datasets where strong persistence coexisted with noised and spiky fluctuations.

Developed research showed that the uniqueness of merging ACWA's pattern-driven approach with transformer-based attention offered a robust solution for cybersecurity threat intelligence for example predicting Disinformation in Ukraine frequencies, among other domains that demanded both long-term memory and short-term adaptability.

Moreover, the outcome suggested that the transformer's mechanism of assigning relevance to different parts of a sequence aligned naturally with ACWA's adaptive weighting of past occurrences, reinforcing the notion that "where we look" in the time series should be guided by pattern confidence and time decay.

With this real-world application, we demonstrated the clear advantages of the applied method, showing its ability to outperform traditional techniques in adaptive relevance assignment. We also discussed ways to enhance the solution's efficiency, with plans to validate these improvements in future research. Specifically, we aim to explore ChronoTensor's potential by integrating frequency decompositions to uncover hidden cycles or regularities, potentially addressing performance drops on longer datasets [9]. ACWA's multi-token or probabilistic extensions could further refine sequence processing, while advanced ensembling approaches may better capture complex nonlinear interactions [10].

We also plan to extend the architecture to support multivariate and multimodal datasets [11], enabling ChronoTensor to process diverse data types and contextual nuances. Incorporating OSINT sources through entity extraction and semantic networks would add richer contextual overlays to the framework [12]. These advancements could unlock features such as advanced weighting functions and real-time deployment, significantly improving the system's ability to predict and model the occurrence and evolution of Information Operations with exceptional precision.

In doing so, we believed that this synergy of time series within the context attention would continue to transform the landscape of forecasting not only in cybersecurity but even beyond established boundaries.

Acknowledgments

Profoundly grateful to Dmitry Lande, my scientific advisor, for his exceptional guidance, unwavering support, and inspiring encouragement throughout the prediction and forecasting research journey.

Deep expertise and thoughtful mentorship have been pivotal in shaping the direction and outcomes of this work.

References

[1] Time Series Analysis: Forecasting and Control / G.E.P. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung. - 5th. - 2015. - 712 p.

[2] Attention is All You Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin // Advances in Neural Information Processing Systems (NeurIPS). - 2017. - P. 5998–6008.

[3] Bryan Lim Stefan Zohren S. R. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting // International Journal of Forecasting. - 2021.

[4] Makridakis S., Spiliotis E., Assimakopoulos V. The M4 Competition: Results, Findings, and Conclusions // International Journal of Forecasting. - 2020. - P. 1–26.

[5] Yoshua Bengio Oriol Vinyals N. J. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks // NeurIPS. - 2015. - P. 1–9.

[6] Noh K., Kim S., Kim J. CAWA: Correlation-Based Adaptive Weight Adjustment via Effective Fields for Debiasing // 2024 Joint 13th International Conference on Soft Computing and Intelligent Systems - 2024.

[7] Pascanu R., Mikolov T., Bengio Y. On the Difficulty of Training Recurrent Neural Networks // Proceedings of the 30th International Conference on Machine Learning (ICML). - 2013.

[8] Bergstra J., Bengio Y. Random Search for Hyper- Parameter Optimization // Journal of Machine Learning Research. - 2012. Vol. 13. - P. 281– 305.

[9] The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis / N.Huang, Z. Shen, S. Long, M.Wu, H.Shih, Q. Zheng, N. Yen, C.Tung, H.Liu // Proceeding of the Royal Society.- 1998. Vol. 454. P. 903–995.

[10] Graves A. Sequence Transduction with Recurrent Neural Networks // International Conference of Machine Learning (ICML) Workshop on Representation Learning 2012.

[11] Mahmoud A., Mohammed A. Leveraging Hybrid Deep Learning Models for Enhanced Multivariate Time Series Forecasting // Neural Processing Letters. - 2024. - P. 1–10.

[12] D. Lande L. Strashnoy GPT Semantic Networking: A Dream of the Semantic Web - The Time is Now. - ISBN 978-966-2344-94-3. - 2023. - 168 p.