Patch-aware Long-term Weather Forecasting

Aslan Feng*

Tsinghua International School Beijing, China

*aslanfxr06@gmail.com

Abstract. This paper explores the prediction of weather condition and employs long-sequence time-series forecasting techniques, specifically the Transformer model. Contrary to traditional methods that examined the model based on prediction length, we focus on the analysis of the relationship between patch value and prediction accuracy of the PatchTST Transformer model. The statistical analysis involves identifying specific words or phrases in news titles and correlating them with view counts. Through experiments, we demonstrate that the Transformer model effectively predicts the popularity of news articles based on weather information, yielding accurate results. We hope this work illuminates the untapped potential of utilizing weather data to forecast public engagement with news content and uncovers novel insights into the intricate relationship between weather conditions and public attention.

Keywords: Time series prediction, PatchTST, weather

1 Introduction

The popularity of news articles plays a crucial role in today’ s information-driven society. Under- standing the factors that contribute to the popularity of news can provide valuable insights for content creators, journalists, and online platforms. In this research paper, we propose the use of PatchTST (Patch Time Series Transformer), an advanced algorithm and model, to predict news popularity. The prediction of news popularity holds significant importance in various domains, including media analytics, content recommendation systems, and online advertising. By accurately forecasting the popularity of news articles, media organizations can optimize their content creation strategies, improve user engagement, and enhance their revenue generation.

Previous approaches for predicting news popularity have been characterized by several limitations. Traditional methodologies often struggle to capture the dynamic nature of news articles, frequently neglecting temporal patterns and abrupt shifts in popularity. Additionally, the inability to detect or accommodate patches or anomalies within the time series data further compounds these challenges. In an effort to surmount these issues, we introduce a novel framework, PatchTST. Distinct from preceding models, PatchTST capitalizes on the advancements in transformer-based architectures and attention mechanisms. Through a sophisticated analysis of temporal patterns, the model is able to consider both the historical context and the presence of

X. Ding et al. (eds.), *Proceedings of the 2023 4th International Conference on Big Data and Social Sciences (ICBDSS 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 12,
https://doi.org/10.2991/978-94-6463-276-7_50

patches or sudden fluctuations in the time series data. By addressing the inadequacies inherent in conventional methods, PatchTST fosters a more comprehensive understanding of news popularity dynamics, thereby leading to predictions that are both more accurate and reliable.

At a high level, our approach involves training a PatchTST model on a large dataset of news articles, capturing the temporal dynamics and patches within the popularity time series. The PatchTST model utilizes attention mechanisms to focus on relevant patches, enabling it to extract meaningful features and relationships from the data. The intuition behind PatchTST is to exploit the temporal patterns and anomalies present in news popularity. By considering the historical context and identifying patches of significant changes in popularity, the model can learn to make accurate predictions and capture the underlying factors that contribute to news popularity.

Before delving into the specifics of our approach, it is essential to highlight the significance of this research. Understanding the factors that drive news popularity can have far-reaching implications in various fields. Content creators can optimize their strategies, journalists can focus on producing relevant and impactful news, and online platforms can enhance their recommendation systems to provide users with engaging content.

By addressing the drawbacks of previous approaches and introducing PatchTST, we aim to provide a more robust and accurate method for predicting news popularity. Through this research, we hope to contribute to the field of media analytics and enable stakeholders to make informed decisions based on advanced predictive models. In the subsequent sections of this paper, we will present a detailed explanation of our PatchTST model, the experimental setup, and the evaluation of its performance in predicting news popularity.

2 Related Work

Because news is prevalent and a part of many people's lives, forecasting news is explored and studied in a variety of ways. One of the methods to predict news events is Markov logic networks [1]. It calculates the probability of certain events by transforming the existing textual news articles into readable data and learning the cause-and-effect relationship between the knowledge, consequently evaluating the authenticity of the news. However, this method fails to take public attention and views into account and lacks analysis of the general trend in society.

Natural Language Processing, also known as NLP, has also been suggested to predict future news events [2]. Information in headlines including subjectivity, sentiments, context, and named entities can be extracted using NLP, helping classify the news contents and anticipate their DJIA (Dow Jones Industrial Average) index. Although it is convenient concerning the headlines themselves and their attributes, it has difficulty building the connection between them and the trend.

In order to recommend news—similar to predicting news—in people's APPs based on their personal preferences, scholars established NRS (News Recommender System) to generate connections between the news articles and people's predilections.

Mainly used algorithm CBF (content-based filtering) compares the attributes in the user and item profiles, solving the problem of updating the contents based on the changing interests or focuses of users; however, it relies on the amount of information the user provides and fails to deal with a large number of anonymous or temporary users [4].

3 Our Approach

3.1 Problem Definition

We define the problem as follows: given a list of news titles, categories, and popularity P (characterized by the number of views) A:}, where each xi represents the ith news in the list and has attributes as described, we aim to forecast the popularity list B: {P1, ..., PN} of the input data news.

3.2 Transformer-based Time Series Encoder

We use a Transformer Encoder to manage the inputs of titles. It works by breaking down a sentence or phrase into words and transferring them into integers tokens. These tokens are then converted into embedding vectors, which are common practices in neural machine language translation, to let the transformer extracts useful information. Positional encoding is involved to ensure the program knows word positions without recurrence.

Specifically, in supervised learning, the data matrices D: {x1, ..., xn} are first input as series, put to go through the instance normalization operator, and segmented into patches as tokens. The patches are mapped to the Transformer dimension D latent space through a trainable linear projection Wp, and a learnable additive position encoding Wposis applied to monitor the temporal order of patches.

The encoder first process the matrices with multi-head attention (for h = 1, ..., H) and transform

them into query matrices $Q_h^i = \left(x_h^i\right)^T W_h^Q$, key matrices $K_h^i = \left(x_h^i\right)^T W_h^K$, and value matrices $V_h^i = \left(x_h^i\right)^T W_h^V$. And finally, the output would be [6]:

$$\left(O_h^i\right)^T = \text{Attention}\left(Q_h^i, K_h^i, V_h^i\right) \tag{1}$$

where the Attention function gives the desired result with the following Softmax function.

4 Experiment

In this section, we performed an experiment to train the transformer with the MSE loss function using supervised learning. We use news and behaviors data sources from PatchTST and summarized contents with evaluated popularity from News Summari-

zation. The latter source contains a greater time series, which, therefore, is likely to give a more stable outcome.

We evaluate and aggregate the performances of the transformer as shown in the table below. It can be seen that the loss values are all small enough to be considered accurate. However, when the time period becomes longer, the prediction model shows greater and greater deviation from the actual data.

Table 1. Long-term forecasting results from PatchTST weather dataset using patch value of 16.

Models		**PatchTST/64**	
Metric		**MAE**	**MSE**
Weather	96	0.199	0.152
	192	0.242	0.196
	336	0.284	0.249
	720	0.334	0.318

4.1 Dataset

3 datasets including MIND, News Summarization, and PatchTST are used to train and experiment with the model.

MIND (MIcrosoft News Dataset) [3]: Containing about 15 million impression logs and 160,000 news articles with specific details including title, abstract, and category, this dataset provides a basis for future studies to further enhance its algorithms when labeling news articles. However, its time span is limited at present, requiring modifications to be applied in news public attention prediction.

News Summarization [5]: It collects three major datasets, including Xsum, CNN/Daily Mail, and Multi-News, and provides 871,520 news articles' sources and summaries.

PatchTST dataset [6]: It includes a variety of news, including aspects of weather, traffic, illness, and so on.

4.2 Metric

As shown in Table 1 and Table 2, we incorporated two error indices to evaluate the prediction model's accuracy. For one, we use the Mean Absolute Error (MAE), which is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{P_N} - P\right| \tag{2}$$

It can be referred from the formula that MAE is the arithmetic mean of the sum of the absolute value of errors in each predicted value. It preserves the same units of measurement as the data under analysis and gives all individual errors the same weights.

For another, we use the Mean Squared Error (MSE) to estimate the possible error of predictions compared with the true values. The MSE function is featured as follows

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{P_N} - P\right)^2 \tag{3}$$

where $\widehat{P_N}$ is the predicted popularity and P is the true value. In other words, MSE measures the sum of the square of errors, which is able to avoid negative error values and mutual cancellation of errors. In addition, it is able to enlarge the outliers within the predicted result to find out if there is any particular case that deviates the most from reality.

4.3 Baselines

PatchTST, which stands for Patch Time Series Transformer, is a powerful algorithm and model used in machine learning and time series analysis. It is specifically designed to analyze and make predictions on time series data that may contain patches or anomalies. PatchTST utilizes a transformer-based architecture to effectively capture temporal patterns and relationships within the time series data. By incorporating attention mechanisms, it can focus on relevant patches and effectively model the dynamics of the time series. Specifically, it starts with an initial patch-based representation, where the input image is divided into patches. These patches are linearly projected to obtain patch embeddings. The patch embeddings are then combined with positional embeddings to capture spatial information. To incorporate token-based representations, the patch embeddings are transformed into tokens through a linear projection. These tokens are then fed into a transformer encoder, as explained in section Ⅲ, consisting of self-attention and feed-forward networks [8]. This hybrid structure allows the PatchTST transformer to effectively leverage both patch-based and token-based information for improved performance in vision tasks. This innovative approach makes PatchTST suitable for a wide range of applications, such as anomaly detection, forecasting, and pattern recognition in various domains, including finance, healthcare, energy, and more. By leveraging PatchTST, researchers can gain valuable insights from time series data and make informed decisions based on accurate predictions and anomaly detection.

4.4 Configuration

PatchTST for predicting news popularity involves configuring several key components and parameters. The architecture of PatchTST is based on transformers, which consist of multiple layers of self-attention and feed-forward neural networks. The number of layers and hidden size can be adjusted based on the complexity of the problem. To capture patches or anomalies in the news popularity time series, PatchTST employs a patch extraction mechanism, allowing for granularity control. Attention mechanisms are incorporated to focus on relevant patches and learn temporal relationships. Configurable parameters such as learning rate, batch size, and optimizer are set during training to optimize model performance. The choice of loss

function, whether mean squared error or binary cross-entropy, depends on the specific prediction task. Hyperparameter optimization techniques can be applied to find the optimal values for dropout rates, regularization terms, and the number of attention heads. The configuration of PatchTST for predicting news popularity is tailored to the dataset, problem complexity, and desired performance, with experimentation and refinement being essential for accurate and robust predictions.

In this research, patch values are adjusted based on the original model to explore the optimal value. This is because the patch value directly influences the granularity at which the time series data is analyzed. By varying the patch value and observing its impact on the prediction accuracy metrics such as mean squared error (MSE) loss and mean absolute error (MAE) loss, we can assess the relationship between the patch value and the model's performance. This allows us to identify the patch value that strikes the right balance, capturing meaningful patterns in the data while avoiding noise or overfitting. Experimentally testing different patch values enables us to optimize the model's predictive capabilities and ultimately enhance the accuracy of our predictions.

4.5 Main Results

The patch value, which refers to the size or granularity of the patches used in PatchTST, can have an impact on the prediction accuracy of the model. The selection of an appropriate patch value is crucial as it determines the level of detail at which the time series data is analyzed. When the patch value is too large, the model may overlook important temporal patterns or sudden changes in the data. It may fail to capture the fine-grained dynamics of the time series, leading to less accurate predictions. In such cases, the model may average out or overlook significant variations within the patches, resulting in decreased prediction accuracy. On the other hand, if the patch value is too small, the model may become overly sensitive to noise or minor fluctuations in the data. It may start to capture irrelevant or insignificant patterns, leading to overfitting and decreased generalization performance. This can result in less accurate predictions of unseen or real-world data.

Discovering Optimal Patch Value.

In order to determine the optimal patch value that minimizes both the mean squared error (MSE) loss and mean absolute error (MAE) loss, we conducted an empirical analysis by adjusting the patch value. Specifically, we examined the impact of three different patch values, namely 16, 24, and 32, on the accuracy of the predictions. By systematically varying the patch value, we aimed to investigate the relationship between the accuracy metrics and the granularity of the patches. To better analyze the result, we also compared with the result generated in paper[6] displayed in Table 3.

We analyze the relationship between the loss function value and patch value inside the transformer.

Table 2. Long-term forecasting results from PatchTST using patch values of 16, 24, and 32 respectively

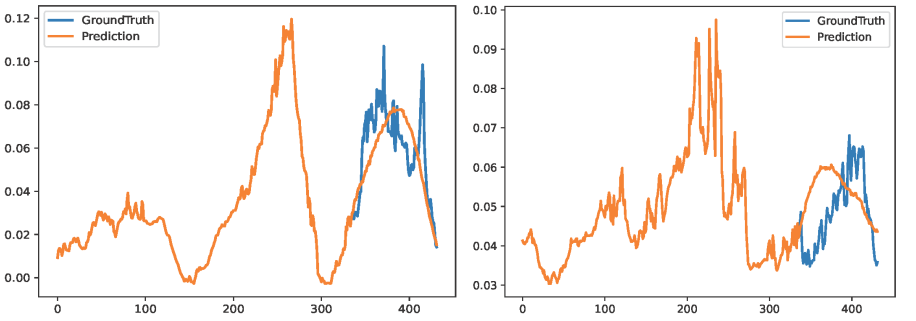
Models		**PatchTST/64**					
Patch Value		**16**		**24**		**32**	
Metric		MAE	MSE	MAE	MSE	MAE	MSE
Weather	96	0.199	0.152	0.203	0.154	0.205	0.155
	192	0.242	0.196	0.243	0.197	0.244	0.201
	336	0.284	0.249	0.282	0.247	0.281	0.246
	720	0.334	0.318	0.333	0.316	0.331	0.314

Table 3. Long-term forecasting results from PatchTST using patch values of 16, data reported from "A Time Series Is Worth 64 Words: Long-term Forecasting With Transformers" by Yuqi, N.[6].

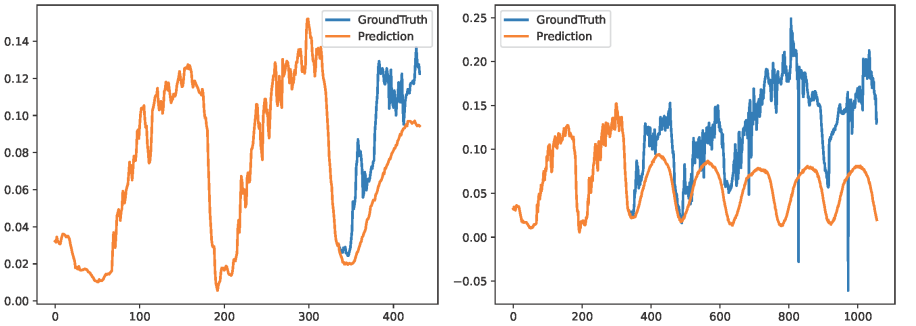
Models		**PatchTST/64**		**Informer**		**LogTrans**	
Patch Value		**16**		**16**		**16**	
Metric		MAE	MSE	MAE	MSE	MAE	MSE
Weather	96	0.198	0.149	0.405	0.354	0.490	0.458
	192	0.241	0.194	0.434	0.419	0.589	0.658
	336	0.282	0.245	0.543	0.583	0.652	0.739
	720	0.334	0.314	0.705	0.916	0.675	1.004

Larger Patch Values Capture Long-term Information.

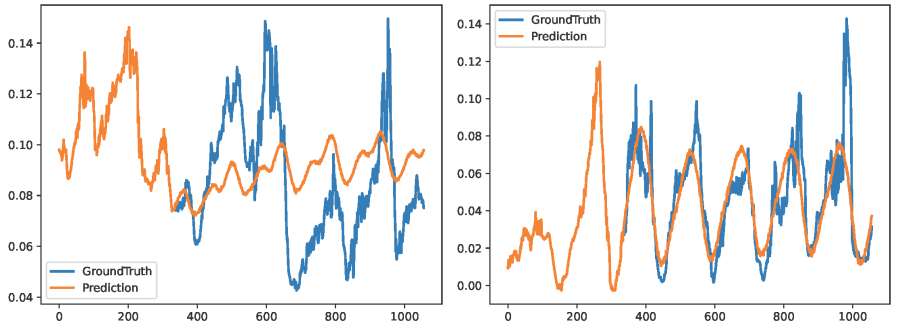
Upon analyzing the graph, it becomes evident that variations in the model' s accuracy are directly influenced by changes in the manipulated patch value during the experiment. Notably, our observations indicate that an increase in the patch value corresponds to an upward trend in both the Mean Absolute Error (MAE) and Mean Squared Error (MSE) values when the prediction period is either 96 or 192. Conversely, when the prediction period lengthens to either 336 or 720, there is a noticeable decrease in the error indicators. This insightful finding highlights the importance of selecting an appropriate patch value based on the desired prediction timeframe. As shown in Figure 1, specifically, larger patch values are more suited for long-term predictions, whereas smaller patch values should be implemented to accurately predict short-term information. By considering these factors, we can optimize the accuracy and reliability of our predictive models.



(a) Patch value 8 predicting weather news with period 96 (b) Patch value 16 predicting weather news with period 96



(c) Patch value 24 predicting weather news with period 96 (d) Patch value 8 predicting weather news with period 720



(e) Patch value 24 predicting weather news with period 720 (f) Patch value 24 predicting weather news with period 720

Fig. 1. Graphs of comparisons between predicted popularity compared with ground truth in different patch values and prediction period, graph generated using the data from PatchTST [6].

In our analysis, we generated several graphs comparing the real values and predicted values of the news popularity across different patch values. As we increased the

time period represented in the graphs, a common phenomenon emerged: a deviation between the real and predicted values became more apparent. Initially, for shorter time periods, the predicted values closely aligned with the real values, indicating accurate predictions. However, as the time period increased, the predicted values gradually deviated from the real values. This phenomenon can be attributed to the inherent complexity of long-term forecasting. Over an extended period, various factors such as changing trends, evolving user preferences, and external events might influence news popularity. The deviation observed in the graphs highlights the challenge of accurately capturing and predicting these long-term dynamics. It emphasizes the need for further research and refinement of models to enhance their ability to capture and forecast the complex dynamics of news popularity over extended time periods.

5 Conclusion

This research paper has presented a comprehensive analysis of predicting news popularity based on weather information using long-sequence time-series forecasting techniques and the Transformer model (SOTA). Through an examination of a dataset comprising categorized news articles, we have successfully demonstrated the efficacy of the Transformer model in accurately forecasting the popularity of news articles. The experimentation process has emphasized the significance of adjusting the patch value to optimize the model's performance and achieve improved prediction accuracy.

Moreover, our investigation has revealed an intriguing phenomenon observed when comparing the real and predicted values over varying time periods. As the time period increased, a noticeable deviation between the predicted and real values became more pronounced. This finding under- scores the intricate challenges associated with long-term forecasting and the complexity inherent in capturing and predicting the dynamic nature of news popularity over extended periods.

The observed deviation can be attributed to the interplay of numerous factors that influence news popularity, such as shifting trends, evolving user preferences, and external events. These complex dynamics, combined with the intrinsic limitations of predictive models, contribute to the widening gap between the predicted and real values as the time horizon expands. This phenomenon highlights the need for further research and advancements in forecasting models to enhance their ability to capture and forecast the intricate long-term dynamics of news popularity accurately.

By incorporating weather information into the prediction models, this research has offered valuable insights into the integration of external factors in forecasting models. The exploration of the relationship between weather conditions and public attention provides a novel perspective in under- standing the interplay between contextual factors and news popularity. This knowledge can empower content creators, journalists, and online platforms to optimize their strategies, deliver more engaging content, and effectively engage their target audience.

In summary, this research contributes to the advancement of media analytics by demonstrating the effectiveness of the Transformer model in predicting news popular-

ity based on weather information. The findings highlight the importance of adjusting the patch value for optimal performance and shed light on the challenges associated with long-term forecasting. By leveraging external factors, this study enriches our understanding of the dynamics driving news popularity and offers practical implications for content creators and online platforms seeking to enhance user engagement and improve content recommendation systems. Future research endeavors should focus on refining predictive models, considering additional contextual factors, and exploring innovative approaches to address the complexities of long-term forecasting in the realm of news popularity prediction.

References

1. Dami, S., Barforoush, A. A., & Shirazi, H. (2018). News events prediction using Markov logic networks. Journal of Information Science, 44(1), 91–109. https://doi.org/10.1177/0165551516673285.
2. Velay, M., & Daniel, F. (2018). Using NLP on news headlines to predict index trends. ArXiv, abs/1806.09533.
3. Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., & Zhou, M. (2020). MIND: A Large-scale Dataset for News Recommendation. Proceed- ings of the 58th Annual Meeting of the Association for Computational Linguistics, 3597–3606. https://doi.org/10.18653/v1/2020.acl-main.331.
4. Raza, S., Ding, C. (2022). News recommender system: a review of recent progress, challenges, and opportunities. Artif Intell Rev 55, 749–800. https://doi.org/10.1007/s10462-021-10043-x.
5. Narayan, S., Cohen, S., Lapata, M., See, A., Liu, P., Manning, C., Fabbri, A., Li, I., She, T., & Radev, D. (2022). News Summarization. Kaggle.
6. Yuqi, N., Nam, N., Phanwadee, S., & Jayant, K. (2023). A Time Series Is Worth 64 Words: Long-term Forecasting With Transformers. ICLR. https://doi.org/10.48550/arXiv.2211.14730.