Can Conformer algorithm be effectively used for weather prediction?

Kelvin Susanto Gunawan

November 2025

Abstract

This template includes the format and typical sections for the report. The sections shown in the template are for guidance, this content is required but if the sections headings are different then this is acceptable (for example, some reports may combine the Introduction and Background or include a Literature Review or Related Work section).

1 Introduction

Weather predictions have always been a critical task in the meteorology field and highly relevant for human's daily life, either to be alerted of natural disaster or to simply plan an outing/activity based on the fore-casted weather, ss such, weather prediction has been a common tasks among Artificial Intelligence (AI)/Machine Learning (ML) tasks. However, weather predictions is a challenging tasks. Starting with Numerical Weather Prediction (NWP) methods in their early stages, This method proves to be expensive with limitations in capturing patterns [1].

In recent years, alternatives has emerge for weather forecasting utilizing deep learning techniques. Models such as Recurrent neural networks (RNNs), short-term memory networks (LSTMs), or even combination of two models like convolutional neural networks (CNNs) and LSTMs to gain better performance. These modern methods have been proven to be able to outperform conventional statistical baselines [2, 3]. Furthermore, a newly attention mechanism learning approach using Transformer models have further expanded the field with their capabilities to capture long-range temporal dependencies and multi variable patterns between the weather's features although are still outperformed by RNN in term of capturing short-term forecasting [4].

A variant of Transformer models, Conformer, has been developed with the original proposition to enhance the capabilities of speech recognition tasks. The main idea behind the architecture of conformer is to allow the models to capture both local and global dependencies in sequential data utilizing CNNs and Transformer respectively [5]. In this pilot study, the author wants to leverage the usage of attention models for weather prediction tasks. Given the tendency of weather data to be enormous and will continue to grow overtime, optimizing attention mechanism that is capable to capture global connection in between dataset will result in a great improvement for future weather prediction services. The author aims to provide a clear knowledge on the effect of implementing a convolution layer in the transformer model specifically in the weather forecasting field while acknowledging the limitation of time and resource in doing so.

2 Literature Review

2.1 Traditional Weather Prediction Methods

As discussed in the introduction section, early stages of weather predictions are filled with NWP that is capable to solve complex physical equations which describe atmospheric dynamics. Some of the well known models within this field are Global Forecast System (GFS) and ECMWF, both are known to be able to provide reliable forecast. These methods however, scales badly with fine spatial resolutions resulting in high computational cost. Moreover, NWP was also known to struggle with micro features and rapid-evolving phenomenons which include localized rainfall and wind gusts caused by the chaotic nature of the events [1, 6].

2.2 Deep Learning

With the keep increasing datasets for meteorology, deep learning has emerge to be an alternative to the conventional NWP that utilized physic based learning. Some of the widely used deep learning models include RNNs, LSTMs, and GRU for temperature, wind speed, humidity, and many more and studies have shown that they are capable to outperform the statistical baselines and simpler machine learning and they strive in predicting short-term weather in which temporal dependencies work best [4, 7, 8].

Hybrid deep learning architectures have also been implemented like CNNs + LSTMs and demonstrate improved performance in temperature and rainfall forecasting by capturing spatial patterns and temporal sequences. Studies also show that deep learning architecture scale better with the increment of data quality in time-series data, allowing the model to learn better overtime [9].

2.3 Transformer

With the recent shift towards the use of attention mechanism for deep learning models, the implementation has been more and more integrated towards a lot of tasks. Originally developed in Natural Language Processing (NLP), Transformer models has branched to another deep learning section such as speech recognition, image processing, and including but not limited to weather prediction [10, 11, 5, 12, 13]. A study shows that transformer models that includes Informer, iTransformer, Former, and PatchTST has been proven to outperform RNN-based models in term of overall accuracy [4].

2.4 conformer

Introduced by Gulati et al. [5], conformer combines convolution layer with transformer's attention mechanism, allowing the model to capture both local and global patterns/dependencies within the data's sequence. After its initial discovery, conformer kept on getting attention for some specific fields in deep learning especially speech recognition [14].

3 Methodology

The proposed workflow is designed to fairly compared the performance capabilities of transformer and conformer model. The same data preprocessing method will be conducted for both model.

3.1 Data Collection and Preprocessing

Our proposed methodology use the Seattle Weather Dataset provided by ANANTH R and can be obtained from kaggle [15]. The dataset is sufficiently large and includes relevant features such as temperature, precipitation, wind, and weather conditions. During the data prepossessing method, "date" feature is dropped due to the irrelevancy in correlation with other features. The author use the standard 80% - 20% train-test-split ratio and apply robust scaler to reduce the influence of outliers. The X_features shape will be 4 consisting of "precipitation", "temp_max", "temp_min", and "wind" while the y_target will be "weather". To fit with the transformer and conformer model input, the data is reshaped to 3D tensor format (Samples, Features, Target) resulting in (1168, 4, 1) final shape.

3.2 Model Architecture and Training

This study implements two attention-based deep learning models for weather classification: a Transformer model and a Conformer model. Both architectures operate on the same preprocessed 3D input tensor and are designed to capture temporal dependencies in weather sequences. The following subsections summarise the essential structure of each model.

3.2.1 Transformer Model

The Transformer architecture is built using three main components:

(1) Positional Encoding: Since the Transformer does not naturally encode order information, positional embeddings are added to the input sequence to represent temporal structure.

(2) Transformer Block: Each block contains a multi-head self-attention layer followed by a feed-forward network. The attention mechanism enables the model to capture long-range temporal dependencies between weather features, which is valuable for forecasting tasks.

(3) Final Model Structure: Multiple Transformer blocks are stacked, followed by a global average pooling layer and a dense output layer with softmax activation for multi-class classification.

The model is trained using the Adam optimizer with categorical cross-entropy loss, and evaluated using accuracy, precision, and recall metrics. Training is conducted for 100 epochs with a batch size of 32.

3.2.2 Conformer Model

The Conformer model extends the Transformer by incorporating convolutional operations to model both global and local dependencies:

(1) Conformer Block: Each block contains two feed-forward modules, a multi-head self-attention module, and a convolutional module. The convolutional component (including depthwise convolutions) enhances the model's ability to capture short-term and local weather patterns.

(2) Final Model Structure: Similar to the Transformer, several Conformer blocks are stacked, followed by global average pooling and a softmax output layer.

The training setup mirrors that of the Transformer to ensure fair comparison.

3.3 Result

Both models are evaluated on the test set using loss, accuracy, precision, and recall. The Conformer achieves slightly higher overall performance, demonstrating the benefit of combining attention with convolutional operations.

In addition to numerical metrics, confusion matrices are generated for both models to analyse prediction errors across weather classes. These results further confirm that the Conformer provides a modest but consistent improvement over the baseline Transformer model.

Model	Loss	Accuracy	Precision	Recall
Transformer	0.5308	82.93%	84.80%	81.91%
Conformer	0.4867	83.95%	83.90%	83.61%

3.4 Model Performance Comparison

The following figure shows a side-by-side comparison of the performance of the Transformer and Conformer models, plotting their accuracy and loss during training:

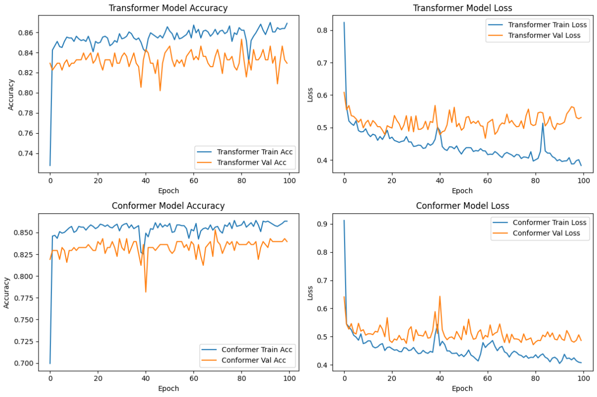

Figure 1: Model Performance Comparison: Accuracy and Loss for Transformer vs. Conformer

3.5 Discussion and Anaysis

This study utilized a publicly available data from kaggle containing unadulterated weather data designed specifically for weather study without containing sensitive information (governed by Kaggle's licensees) and pilot study methodology is also independent in nature. In this study, author compared the performance of Transformer and Conformer models for weather prediction. The result indicated that the Conformer model outperformed the Transformer in terms of accuracy, precision, and recall, although the difference was small. A paired t-test was conducted The results indicated that the difference in accuracy between the two models. The t-statistic was -1.27 (degrees of freedom = n - 1), with a p-value of 0.21. Given that the p-value exceeds the common significance threshold of 0.05, the null hypothesis cannot be rejected as the difference between the model is not significant statistically. Within the nature of an independent pilot study, this research is limited with time and hardware capability. A moderately big dataset that is used might not be enough to represent the capabilities of attention based deep learning models.

3.6 Future Work

As discussed previously about the limitation, Future works could improve research qualities by employing large dataset (e.g. ERA5 dataset) to allow both models capturing dependencies in between data. Another improvement is to try out a wider configurations of layer variations for each model to allow them achieve their best result while still using the same data split and input. The source code for the models, data preprocessing, and evaluation scripts is available on GitHub: https://github.com/itskelv/Weather$_predictions$.

References

[1] Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, Cambridge, 2002.

[2] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

[3] Stephan Rasp, Peter D. Dueben, Sandro Scher, Jonathan A. Weyn, Samia Mouatadid, and Nils Thuerey. Weatherbench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems (JAMES)*, 12(11):e2020MS002203, 2020.

[4] Rogerio Pereira Dos Santos, João P. Matos-Carvalho, and Valderi R. Q. Leithardt. Deep learning in time series forecasting with transformer models and rnns. *PeerJ Computer Science*, 11:e3001, 2025.

[5] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[6] Catherine O. de Burgh-Day and Tennessee Leeuwenburg. Machine learning for numerical weather and climate modelling: a review. *Geosci. Model Dev.*, 16(22):6433–6477, 2023.

[7] Xiaobao Song, Liwei Deng, Hao Wang, Yaoan Zhang, Yuxin He, and Wenming Cao. Deep learning-based time series forecasting. *Artificial Intelligence Review*, 58(1):23, 2024.

[8] Jianbin Zhang, Meng Yin, Pu Wang, and Zhiqiu Gao. A method based on deep learning for severe convective weather forecast: Cnn-bilstm-am (version 1.0). *Atmosphere*, 15(10):1229, 2024.

[9] Yuhao Gong, Yuchen Zhang, Fei Wang, and Chihan Lee. Deep learning for weather forecasting: A cnn-lstm hybrid model for predicting historical temperature data. *Applied and Computational Engineering*, 99:168–174, 2024.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[11] Leonardo Câmara and Maurício Corrêa. Residues for maps generically transverse to distributions. *arXiv preprint arXiv:1804.07139*, 2018.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiao-

hua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[14] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. *arXiv preprint arXiv:2105.03889*, 2021.

[15] ananthr1. Weather prediction. https://www.kaggle.com/datasets/ananthr1/weather-prediction, 2025. Accessed: 2025-12-02.