Typhoon Intensity Prediction with Vision Transformer

Huanxin Chen
South China University of Technology
Guangzhou, China

Pengshuai Yin
Guangdong Laboratory of Artificial Intelligence and Digital Economy
Shenzhen, China

Huichou Huang
City University of Hong Kong
Hong Kong, China

Qingyao Wu
South China University of Technology
Guangzhou, China

Ruirui Liu
Brunel Univeristy London
London, United Kingdom

Xiatian Zhu*
CVSSP, University of Surrey
Guildford, United Kingdom

Abstract

Predicting typhoon intensity accurately across space and time is crucial for issuing timely disaster warnings and facilitating emergency response. This has vast potential for minimizing life losses and property damages as well as reducing economic and environmental impacts. Leveraging satellite imagery for scenario analysis is effective but also introduces additional challenges due to the complex relations among clouds and the highly dynamic context. Existing deep learning methods in this domain rely on convolutional neural networks (CNNs), which suffer from limited per-layer receptive fields. This limitation hinders their ability to capture long-range dependencies and global contextual knowledge during inference. In response, we introduce a novel approach, namely "Typhoon Intensity Transformer" (**Tint**), which leverages self-attention mechanisms with global receptive fields per layer. Tint adopts a sequence-to-sequence feature representation learning perspective. It begins by cutting a given satellite image into a sequence of patches and recursively employs self-attention operations to extract both local and global contextual relations between all patch pairs simultaneously, thereby enhancing per-patch feature representation learning. Extensive experiments on a publicly available typhoon benchmark validate the efficacy of Tint in comparison with both state-of-the-art deep learning and conventional meteorological methods. Our code is available at https://github.com/chen-huanxin/Tint.

1 Introduction

In recent decades, global warming has led to both the intensification and expansion of typhoons [1]. Typhoons are severe or even extreme weather systems that originate from warm tropical oceans and gradually build up their own power when approaching nearby lands. Upon making landfall, they pose significant threats to lives and properties in their vicinity [2, 3]. Different typhoon intensities correspond to varying levels of economic losses and environmental devastation [4]. The intensity of a typhoon is closely associated with the maximum sustained surface wind speed near its center, making it a key factor in disaster management.

*Xiatian Zhu (xiatian.zhu@surrey.ac.uk) is the corresponding author.
Acknowledgement: This work is supported by China Postdoctoral Science Foundation (2022M721182)

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

Typhoon satellite imagery plays a pivotal role in predicting typhoon intensity due to its rich and timely real-time data, enabling us to dynamically monitor the typhoon's structure, cloud patterns, and environmental conditions [5]. Ample studies focus on predictive regressions using the information, features, and parameters extracted from these remote sensing images [6, 7].

In recent years, to overcome the limitations of regression-based methods, researchers have resorted to deep learning methods [8] for estimating typhoon intensity using satellite imagery [9]. One of the most prevailing methods in this field relies heavily on Convolutional Neural Networks (CNNs) [10], which are renowned for their proficiency in capturing local image features and intricate image structures. However, the over-concentration on local information may deteriorate the model's overall performance as it neglects the global context that is essential for improving predictive accuracy.

In this study, we address the above critical limitations by proposing a novel method called the *Typhoon Intensity Transformer* (**Tint**). The Tint employs self-attention mechanisms with expansive global fields in each layer. This is achieved by partitioning input images into fixed-size patches and transforming them into one-dimensional feature vector sequences. Self-attention mechanisms are then used for establishing the connections among these patches so as to extract comprehensive global features and contextual information spanning the entire image. Moreover, The Tint further refines and consolidates image features through the incorporation of multiple Transformer layers, through which these features are forwarded to an output layer for typhoon intensity predictions. The superior performance of the Tint can be attributed to its capacity in treating images analogously to text data from a sequential perspective, i.e., it is able to extract extensive global relations and contextual cues beyond the local features captured by the CNNs.

2 Methodology

The Tint adopts the same input-output configuration as Vision Transformers[11, 12, 13] with a key difference that it outputs a continuous integer as the typhoon intensity estimation. Overview of our Tint is given in Figure 1.

Image to sequence Our method takes a 1D sequence of feature embeddings $Z \in \mathbb{R}^{L\times C}$ as inputs with L as the preset sequence length and C as the hidden channel size. This is implemented as follows: (1) We first cut a given input image into a grid of patches; (2) Then an embedding block is employed to encode each individual patch into a feature embedding; (3) Positional embedding is further added on top of each content embedding.

Model architecture To strike a balance between computational efficiency and model performance, we choose to implement the Tiny-ViT architecture [11]. Our model comprises four stages with progressively decreasing resolutions. The patch embedding block consists of two convolutional layers using a kernel size of 3, a stride of 2, and a padding size of 1. In the initial stage, we employ lightweight and efficient MBConvs [14] along with downsampling blocks. This choice is motivated by the efficiency of early convolutional layers in learning low-level features [15].

The subsequent three stages are constructed using Transformer blocks that employ window attention mechanisms to reduce computational overhead. We incorporate attention bias [13] and apply a separable convolution of 3×3 depth between the attention and MLP components to capture local information [16], emphasizing the intrinsic ability to learn global context information. The residual connections [10] are used in each block of the first stage also within the attention and MLP blocks. For activation functions, we employ GELU [17]. Furthermore, BatchNorm [18] is applied to the normalization layers of the convolutional layers, and LayerNorm [19] is used for the linear layers.

The architectural selection aims to balance computational efficiency and model performance. Formally, given the embedding sequence Z as an input, the Transformer with M layers of multi-head self-attention (MSA) and Multilayer Perceptron (MLP) blocks is employed to learn feature representations. At each layer l, the input to self-attention is a triplet (`query`, `key`, `value`) computed from the input $Z^{l-1} \in \mathbb{R}^{L\times C}$ as $Z^{l-1}\mathbf{W}_q$, $Z^{l-1}\mathbf{W}_k$, and $Z^{l-1}\mathbf{W}_v$, respectively, where $\mathbf{W}_q/\mathbf{W}_k/\mathbf{W}_v \in \mathbb{R}^{C\times d}$ are the parameters of the learnable linear projections, and d is the dimension of (`query`, `key`, `value`). Self-attention (SA) is then formulated as:

$$SA(Z^{l-1}) = Z^{l-1} + \text{softmax}(\frac{Z^{l-1}\mathbf{W}_q(Z^{l-1}\mathbf{W}_k)^\top}{\sqrt{d}})(Z^{l-1}\mathbf{W}_v). \quad (1)$$

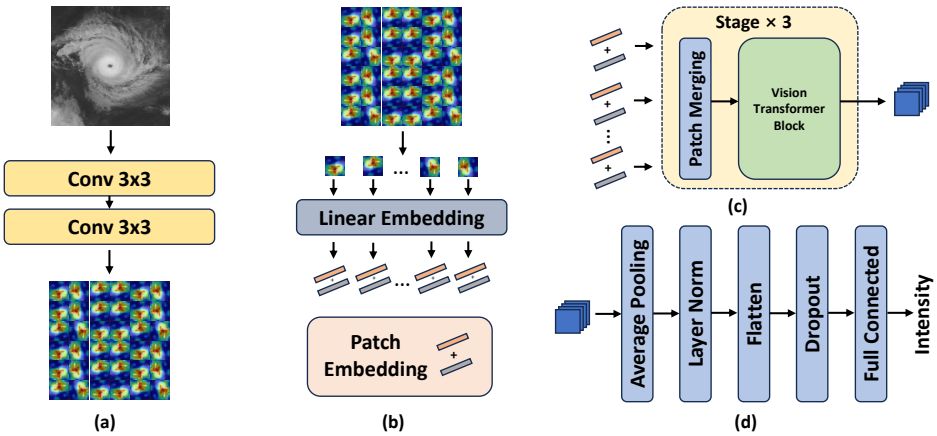

Figure 1: Overview of our proposed *Typhoon Intensity Transformer* (Tint) method. Our approach begins with a sensor image, which undergoes the following key steps: (a) An input image is initially divided into a grid of patches. (b) We then perform patch embedding and integrate positional embeddings, resulting in a sequence of feature embeddings, which serve as our initial image representations. (c) Subsequently, we employ a lightweight vision Transformer to process this sequence of feature embeddings. This step enables both local and global information processing, yielding a comprehensive representation for each patch. (d) Finally, we apply average pooling to aggregate the patch-level feature representations, creating an image-level representation. This is followed by a fully connected layer, which is responsible for predicting the typhoon intensity.

The Multi-Head Self-Attention (MSA) design including m independent SA operations is deployed. It projects a concatenated output in the form of $MSA(Z^{l-1}) = [SA_1(Z^{l-1});\ SA_2(Z^{l-1});\ \cdots;\ SA_m(Z^{l-1})]\mathbf{W}_O$, where $\mathbf{W}_O \in \mathbb{R}^{md \times C}$ and d is set to C/m. The output of the MSA is then transformed by an Multi-Layer Perceptron (MLP) block with a residual skip as the layer output $Z^l = MSA(Z^{l-1}) + MLP(MSA(Z^{l-1})) \in \mathbb{R}^{L \times C}$. The layer normalization applied prior to the MSA and MLP blocks is excluded here for simplicity.

Head design The final layer's feature output is represented as $Z^m \in \mathbb{R}^{L \times C}$. We initially perform an average pooling operation on Z^m, generating $\mathbf{v} = avg(Z^m)$. Subsequently, a fully connected layer uses $\mathbf{v}$ to derive the final prediction $y = \mathbf{v}\mathbf{w} + b$ with $\mathbf{w}$ and b as trainable parameters.

3 Experiments

3.1 Experiment Settings

Dataset Our experiments are conducted using the Tropical Cyclone for Image-to-Intensity Regression (TCIR) dataset [20]. This dataset serves as an important open benchmark for evaluating typhoon intensity estimation models with a fair assessment setting. See the Supplementary Dataset for details.

Performance metric Given that typhoon intensity estimation is formulated as a regression problem [20], we evaluate the model's performance using the Root Mean Squared Error (RMSE).

Training strategy Details of our implementation can be found in the `Supplementary Material`.

Competitors We conduct a comparative analysis of our Tint model with conventional meteorological models and recent deep-learning methods (See more details of these competitors in the `Supplementary Material`).

3.2 Evaluation results

We present the results from typhoon intensity estimation using various methods in Table 1a (validation set) and Table 1b (testing set). Our analysis starts with a comprehensive examination of the validation set, offering the following insights:

Table 1: Compare with other models. IR: Infrared; PMW: Passive Microwave; WV: Water Vapor. The unit of RMSE is knots

(a) Performance evaluation on TCIR validation dataset.

Model	Modality	RMSE
ADT [6]	IR	11.79
AMSU [21]	IR	14.10
SATCON [22]	IR+PMW	**9.21**
KT [23]	IR+PMW	13.20
FASI [7]	IR	12.70
Improved DAV-T [24]	IR	12.69
TI index [25]	IR	9.34
MLRM [26]	IR	12.01
TDCNN [9]	IR	10.00
Deep CNN [27]	IR	10.18
CNN-TC [20]	IR	10.59
GAN-CNN [28]	IR	10.45
ResNet32 [10]	IR	11.63
Tint	IR	9.63
Tint	IR+PMW	**9.54**

(b) Performance evaluation on TCIR testing dataset.

Model	Modality	RMSE
ADT [6]	IR	12.19
SATCON [22]	IR+PMW	9.21
CNN-TC [20]	IR+PMW	10.13
GAN-CNN [28]	IR+WV	10.45
ResNet32 [10]	IR	10.42
ResNet32 [10]	IR+WV	10.39
ResNet32 [10]	IR+PMW	10.32
Tint	IR	9.35
Tint	IR+WV	9.33
Tint	IR+PMW	**9.00**

Meteorological Modeling: Surprisingly, simple linear regressions show considerable performance, surpassing more complex retrieval-based methods (e.g., KT [23] and FASI [7]), as well as brightness temperature gradient-based approaches (e.g., Improved DAV-T [24]). Notably, the TI index as one of the latter [25] provides valuable incremental information but is slightly outperformed by the composite approach that leverages multiple models and meteorological expertise [22].

Deep Learning: Most of the deep learning methods exhibit competitive performance or even surpass their conventional counterparts, highlighting the promising potential of the data-driven methods. In particular, ResNet [10], as one of the most widely used CNN architectures, shows notable performance comparable to traditional linear regressions. CNN models with specific task-oriented designs, such as regularization [27] and multi-modality exploitation [20], exhibit improved performance. Still, these developments fail to overcome the limitations imposed by restricted receptive fields. While it is clear that our Tint model characterized by per-layer global receptive fields outperforms significantly. This supportive evidence validates our model assumptions and structure design.

Our findings remain consistent and qualitatively unchanged on the test set, confirming the stability and generality of our proposed model. In addition, the empirical results suggest that incorporating more sensor data tends to improve the predictive performance. In the test set, our model achieves state-of-the-art after adding WV information, and more so with PMW information.

Qualitative evaluation: We also provide attention visualization for comparing ResNet32 with our Tint model (refer to Figure 2 in the `Supplementary Material` for details). The visualization clearly indicates that Tint captures a broader spatial context and empowers more comprehensive feature representation learning that results in significant improvement in prediction accuracy.

4 Conclusion

In summary, our "Typhoon intensity Transformer" (**Tint**) model significantly improves the predictive accuracy of typhoon intensity, which is of great value for real-world disaster management. The innovative use of self-attention mechanisms is attributable to the superior performance of the Tint, as they succeed in featuring global receptive fields per layer, thereby substantially enhancing its ability in capturing the long-range dependencies in sensor observations. This also makes it adaptive across diverse domains with dynamic high-dimensional data. The results from the experiments conducted on a benchmark dataset demonstrate Tint's superiority over both deep learning and meteorological methods and highlight Tint's potential in innovating disaster management in response to meteorological disasters such as typhoons.

References

[1] Yuan Sun et al. “Impact of ocean warming on tropical cyclone size and its destructiveness”. In: *Scientific reports* 7.1 (2017), p. 8154.

[2] Edward N Rappaport. “Loss of life in the United States associated with recent Atlantic tropical cyclones”. In: *Bulletin of the American Meteorological Society* 81.9 (2000), pp. 2065–2074.

[3] Robert Mendelsohn et al. “The impact of climate change on global tropical cyclone damage”. In: *Nature climate change* 2.3 (2012), pp. 205–209.

[4] Alice R Zhai and Jonathan H Jiang. “Dependence of US hurricane economic loss on maximum wind speed and storm size”. In: *Environmental Research Letters* 9.6 (2014), p. 064019.

[5] Christopher Velden et al. “The Dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years”. In: *Bulletin of the American Meteorological Society* 87.9 (2006), pp. 1195–1210.

[6] Christopher S Velden, Timothy L Olander, and Raymond M Zehr. “Development of an objective scheme to estimate tropical cyclone intensity from digital geostationary satellite infrared imagery”. In: *Weather and Forecasting* 13.1 (1998), pp. 172–186.

[7] Gholamreza Fetanat, Abdollah Homaifar, and Kenneth R Knapp. “Objective tropical cyclone intensity estimation using analogs of spatial features in satellite data”. In: *Weather and forecasting* 28.6 (2013), pp. 1446–1459.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[9] Jeffrey Miller, Manil Maskey, and Todd Berendes. “Using deep learning for tropical cyclone intensity estimation”. In: *AGU fall meeting abstracts*. Vol. 2017. 2017, IN11E–05.

[10] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[11] Kan Wu et al. “TinyViT: Fast Pretraining Distillation for Small Vision Transformers”. In: *European conference on computer vision (ECCV)*. 2022.

[12] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[13] Benjamin Graham et al. “Levit: a vision transformer in convnet’s clothing for faster inference”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 12259–12269.

[14] Andrew Howard et al. “Searching for mobilenetv3”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314–1324.

[15] Tete Xiao et al. “Early convolutions help transformers see better”. In: *Advances in neural information processing systems* 34 (2021), pp. 30392–30400.

[16] Kan Wu et al. “Rethinking and improving relative position encoding for vision transformer”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10033–10041.

[17] Dan Hendrycks and Kevin Gimpel. “Gaussian error linear units (gelus)”. In: *arXiv preprint arXiv:1606.08415* (2016).

[18] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).

[20] Boyo Chen, Buo-Fu Chen, and Hsuan-Tien Lin. “Rotation-blended CNNs on a new open dataset for tropical cyclone image-to-intensity regression”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 90–99.

[21] Stanley Q Kidder et al. “Satellite analysis of tropical cyclones using the Advanced Microwave Sounding Unit (AMSU)”. In: *Bulletin of the American Meteorological Society* 81.6 (2000), pp. 1241–1260.

[22] CS Velden and D Herndon. “Update on the SATellite CONsensus (SATCON) algorithm for estimating TC intensity”. In: *Poster session I. San Diego* (2014).

[23] Jim P Kossin et al. “A globally consistent reanalysis of hurricane variability and trends”. In: *Geophysical Research Letters* 34.4 (2007).

[24] Elizabeth A Ritchie et al. “Satellite-derived tropical cyclone intensity in the North Pacific Ocean using the deviation-angle variance technique”. In: *Weather and forecasting* 29.3 (2014), pp. 505–516.

[25] Chung-Chih Liu et al. “A satellite-derived typhoon intensity index using a deviation angle technique”. In: *International Journal of Remote Sensing* 36.4 (2015), pp. 1216–1234.

[26] Yong Zhao et al. “A multiple linear regression model for tropical cyclone intensity estimation from satellite infrared images”. In: *Atmosphere* 7.3 (2016), p. 40.

[27] Ritesh Pradhan et al. “Tropical cyclone intensity estimation using a deep convolutional neural network”. In: *IEEE Transactions on Image Processing* 27.2 (2017), pp. 692–702.

[28] Boyo Chen, Buo-Fu Chen, and Yun-Nung Chen. “Real-time tropical cyclone intensity estimation by handling temporally heterogeneous satellite data”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 17. 2021, pp. 14721–14728.

Appendix

Dataset

TCIR gathers tropical cyclone data from satellite images across four channels: infrared (IR), water vapor (WV), passive microwave (PMW), and visible light (VIS). Each frame contains 201 $\times$ 201 data points, alongside corresponding wind speed information.

For our experiment setup, we employ typhoon images spanning the years 2003 to 2014 as the training dataset. Data from 2015 to 2016 are designated for validation, and the data from 2017 are reserved for testing. The training dataset comprises 40,348 frames corresponding to 730 typhoons, while the validation dataset encompasses 7,569 frames associated with 131 typhoons. The test set comprises 4,580 frames covering 94 typhoons.

Training strategy

Before feeding an image into the model, we apply preprocessing steps, which include resizing the image to 224x224 pixels, performing a random rotation within the range of $[0, 20^{\circ}]$, and randomly applying horizontal or vertical flips with a 50% probability.

To address potential overfitting, the Tint model's backbone is pretrained on ImageNet. We utilize the Mean Squared Error (MSE) as the loss function for network training. Our training setup employs a batch size of 32 and an initial learning rate of 0.00001. The learning rate undergoes a 10-fold decay at the 50th and 75th epochs, respectively. The training process continues for a total of 100 epochs.

Qualitative results

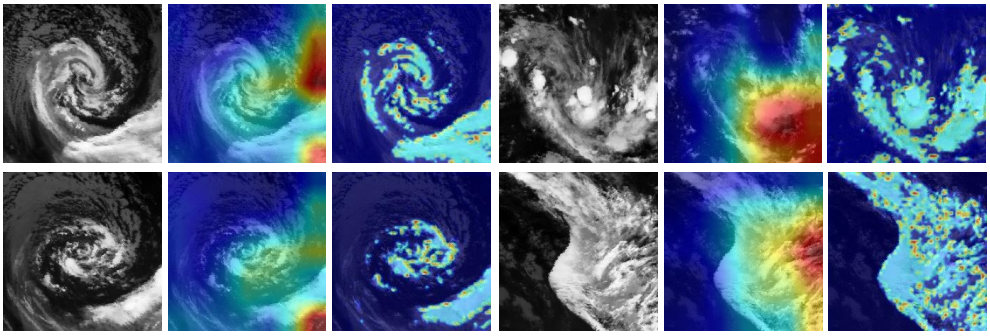

Figure 2: The Grad-CAM visualization of ResNet32 and Tint.

Competitors

Among the conventional meteorological models, we include: Linear Regression-based Advanced Dvorak Technique (ADT) [6]. Advanced Microwave Sounding Unit (AMSU) [21], which relies on near-earth orbit satellites and operates only during satellite passages through typhoons. SATellite CONsensu (SATCON) [22], which is a heuristic combination of ADT and AMSU, widely used in forecasting practice, in particular relying on low-earth orbit satellite observations and expert inputs. Kossin Technique (KT) [23], which improves predictive performance by constructing a global record of typhoon/hurricane intensity from existing data. Feature Analogs in Satellite Imagery (FASI) [7], which employs a k-nearest-neighbor algorithm to determine intensity based on the ten closest typhoons. Improved DAVT [24], which predicts typhoon intensity through statistical analysis of the gradients of the IR brightness temperatures. TI index [25], which leverages image edge processing techniques to examine meaningful discontinuity characteristics and calculate the brightness temperature gradients for typhoon intensity prediction.

We also consider the recently developed deep learning methods for our task: Multiple Linear Regression Model (MLRM) [26], which designs multiple features, including eyewall slope and brightness temperatures, and aggregates multiple regression models using different features for typhoon intensity prediction. Transferring Deep Convolutional Neural Networks (TDCNN) [9], which introduces a simple Convolutional Neural Network (CNN). Deep CNN [27], which proposes a deep neural network with regularization techniques. CNN-TC [20], which presents a multi-modality deep network that takes both IR and PMW inputs meanwhile incorporating rotation-blending and sequence-smoothing techniques. GAN-CNN [28], which employs Generative Adversarial

Networks (GANs) to generate high-quality images, reducing the reliance on PMW data. ResNet [10], which is a widely adopted CNN architecture with numerous successful applications in various domains.