Attention-based Deep Tropical Cyclone Rapid Intensification Prediction

Ching-Yuan Bai[1], Buo-Fu Chen[2], and Hsuan-Tien Lin[1]

[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{b05502055@ntu.edu.tw, htlin@csie.ntu.edu.tw}
[2] Department of Atmospheric Sciences,
National Taiwan University, Taipei, Taiwan
bfchen777@gmail.com

Abstract. Rapid intensification (RI) is when a sudden and considerable increase in tropical cyclone (TC) intensity occurs. Accurate early prediction of RI from TC images is important for preventing the possible damages caused by TCs. The main difficulty of RI prediction is to extract important features that are effective for RI prediction, which is challenging even for experienced meteorologists. Inspired by the success of deep learning models for automatic feature extraction and strong predictive performance, we initiate this study that experiments with multiple domain-knowledge guided deep learning models. The goal is to evaluate the potential use of these models for RI prediction. Furthermore, we examine the internal states of the models to obtain visualizable insights for RI prediction. Our model is efficient in training while achieving state-of-the-art performance on the benchmark dataset on HSS metric. The results showcase the success of adapting deep learning to solve complex meteorology problems.

Keywords: Atmospheric Science · Tropical Cyclone · Rapid Intensification · Spatio-temporal Prediction · Deep Learning · Attention.

1 Introduction

Tropical cyclone (TC) is one of the most devastating weather systems on Earth, characterized by intense and rapidly rotating winds around a low-pressure center. In order to reduce and respond to the damages caused by TCs, many efforts have been devoted during the past half-century to improving the forecast of TC track, intensity, and the associated rainfall and flooding. Although TC track forecast has achieved significant improvement during the past few decades, prediction of TC rapid intensification (RI) remains challenging, which affects the subsequent production of TC structure and rainfall forecast [3]. TC intensity is defined as the maximum sustained wind in the TC inner-core region, and rapid intensification (RI) is defined as a TC experiencing an intensity increase surpassing a threshold (25 - 35 knots) per 24hr period.

Accurately predicting the onset of RI is particularly crucial because reacting to an off-shore RI event before the TC landfall requires sufficient time and delayed reaction had caused some of the most catastrophic TC disasters. External environmental forcing [5,6], TC internal dynamical [12], and thermo-dynamical processes [11] simultaneously control its onset. Thus, a successful RI prediction scheme has to accurately depict both environmental conditions (in which a TC is embedded) and vortex-scale features such as the distribution of precipitation or inner-core TC structure. The goal of this paper is to experiment with the plausibility of deep learning on the RI prediction task from only satellite images since the images are known to be feature-rich but were challenging for meteorologist and forecaster to extract in the past.

2 Related Work

Statistical Hurricane Intensification Predictive Scheme (SHIPS) project has developed a series of statistical models for probabilistic prediction of RI [6,15,16]. The SHIPS RI index (SHIPS-RII, [6]) predicted the probability of a TC intensifying at least 25, 30, and 35 kt per 24hr. This scheme uses simple linear discriminant analysis to determine the RI probability based on a relatively small number (< 10) of predictors describing mainly the environmental factors and some limited aspects of TC internal structure observed by meteorological satellite. Candidate predictors ($N \sim 20$) for SHIPS-RII [6] were subjectively determined by human intelligence, and the final predictors used for linear discriminant analysis were basin-dependent.

A subsequent work of [15] used Bayesian inference (SHIPS-Bayesian) and logistic regression (SHIPS-logistic) to predict RI probability. The authors showed that both SHIPS-Bayesian and SHIPS-logistic exhibit forecast skill that generally exceeds the skill of SHIPS-RII and blending these three models further improved the skill. Another study [16] integrated additional 4 6 predictors derived from satellite passive microwave (PMW) observations into the SHIPS-logistic model and a relative skill improvement from 53.5 % to 103.0 % in Atlantic compared to the original model.

In conclusion, proper inclusion of TC convective information into the statistical model is critical to improving the performance. However, determining new predictors that are capable of adequately representing asymmetric convective features is challenging, which is where the feature extraction power of deep learning shines. Convolutional neural networks (CNN), a variant of deep neural networks useful for analyzing images, have been successfully applied to estimating TC intensity [1,2,14] and predicting TC formation [9].

3 Background

3.1 Recurrent neural network

Recurrent neural network (RNN) specializes in dealing with sequentially dependent features. The network maintains a memory state to encode past information as a reference for the current time step. The output for each time step depends on the corresponding input and the memory state. Long short-term memory (LSTM) cells is a variant of RNN cells where gates within the cell facilitate back-propagation and diminish the effect of gradient vanishing for long sequences. LSTM is widely used in multiple deep learning domains and serves as a crucial component in state-of-the-art models for multiple tasks.

Convolutional LSTM (ConvLSTM)[19] replaces all the fully-connected operations in a normal LSTM cell with convolutional operations, thereby preserving spatial information and allows reduction of parameters with efficient convolution kernels replacing dense weights. It has found success in multiple computer vision tasks related to video inputs such as gesture recognition in videos. [7] successfully applied ConvLSTM in predicting TC track on augmented data, which proves the efficacy of ConvLSTM on cyclone datasets.

Previous studies showed that replacing the convolutions in gates of ConvLSTM with fully-connected operations enables better performance in gesture recognition tasks since previous layers have already captured spatial information [20]. Our task coincidentally shares similar input features and output values, that is, given a short video (consecutive frames of satellite images) predict the probability of labels (occurrence of RI). Replacing convolution operation for gates with fully-connected operation allows reduction of parameters, lowering the potential of overfitting.

3.2 Attention

The attention mechanism is a general framework for reweighting input features by importance. Such a mechanism is successful in various machine learning domains including computer vision. Attention mechanism also enables explainability of the model, providing an importance heatmap that is more intuitive to understand compared to hidden feature maps.

Self-attention [17] generates attention with respect to the input itself. Since our input feature is two dimensional, the attention mask is generated via convolution operations without bias and activation before passing into the sigmoid function to output probability distribution. Then the mask is applied to the input feature via Hadamard product.

Sequence attention [8] is designed for reducing the load of memory states of RNNs in the decoding process. Specifically, during the decoding process in the recurrent layer, consider replacing the input at time t X_t with $\hat{X}_t$

$$\hat{X}_t = [X_t, \sum_{\tau=1}^{t-1} a_\tau^t X_\tau], \quad a_\tau^t = S(X_t, X_\tau)$$

where $S : \mathcal{X}^2 \to \mathbb{R}$ is the attention function which calculates the similarity a_τ^t between X_t and X_τ. The original input X_t is then concatenated with a convex combination of previous time frames to form the modified input $\hat{X}_t$, which now contains information that spans across the entire elapsed period. The decoded output now can obtain information from the past not only from the memory state but also $\hat{X}_t$.

4 Proposed Method

4.1 Performance evaluation

RI forecasting is a time series, image-to-probability prediction task. Given a series of cyclone satellite images, the model outputs the probability of RI occurrence for each time frame. Threshold of 0.5 is used to translate prediction probability into binary output. The following metrics evaluate the performance of the model:

- Brier score (BS) is conventionally used to evaluate the performance of probabilistic predictions in meteorology. Brier skill score (BSS) is the improvement measurement of a target forecast ϕ with respect to a reference climatology forecast ψ.

$$\mathrm{BS}(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \mathrm{BSS}(\phi,x,y)_\psi = \frac{\mathrm{BS}(y,\psi(x)) - \mathrm{BS}(y,\phi(x))}{\mathrm{BS}(y,\psi(x))}$$

  where $x = \in \mathcal{X}^n$, $y = \in \{0,1\}^n$ are the labels and $\hat{y} \in [0,1]^n$ are the predictions. Our reference climatology forecast achieves Brier score of 0.3.
- Heidke skill score (HSS) is another widely used score for evaluating relative performances of binary classification tasks in meteorology. Given labels $y \in \{0,1\}^n$ and predictions $\hat{y} \in [0,1]^n$, let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

$$\mathrm{SF} = \frac{\mathrm{TP}+\mathrm{TN}}{n}\cdot\frac{\mathrm{TP}+\mathrm{FN}}{n} + \frac{\mathrm{FP}+\mathrm{TN}}{n}\cdot\frac{\mathrm{FP}+\mathrm{FN}}{n}, \quad \mathrm{HSS} = \frac{\mathrm{ACC}-\mathrm{SF}}{1-\mathrm{SF}}$$

  where standard forecast (SF) is the probability of correct by chance and accuracy (ACC) is simply $\frac{\mathrm{TP}+\mathrm{TN}}{n}$.

4.2 Data

The dataset of TC satellite images arranged by [1,2] is used for our experiment. Each observation event is associated with the intensity change of the following 24hr as the label to define RI occurence. We follow [1]'s preprocessing and augmentation techniques, selecting only the infrared and passive-microwave channels which are channel-wise normalized, randomly rotating each frame in each TC series, and cropping the central 64x64 portion. We split the data into training (1097 TCs, 43404 events) and validation (188 TCs, 7654 events) sets.

4.3 Model architecture

The model is designed to be extremely light-weight (see fig 1) to prevent overfitting on the small dataset. It is mainly inspired by the Advanced Dvorak Technique (ADT) [13] which is the most widely used method for TC intensity prediction. We will explain the connection between the design choices and ADT in the following sections.

CNN backbone The CNN backbone serve as the feature extractor for extracting rich image features from multiple channels of satellite images. We experimented with popular CNN backbone architectures such as multiple variations of the ResNet family[4,18]. However, we opt for a vanilla CNN structure after much experiment. We are unable to utilize large models pretrained on other natural image datasets and the TC dataset size is insufficient to train a complex model from scratch. We discovered that a lightweight vanilla CNN not only is significantly faster to train but also performs no worse than the complex counterparts. The CNN backbone is end-to-end trained with the task but it is possible to initialize the weights by training an autoencoder on TC images in an unsupervised manner.

Recurrent layer Spatio-temporal data naturally calls for ConvLSTM-like [19] structure to facilitate information flow between time steps while maintaining locality of spatial information. We selected the dense-gated ConvLSTM cell [20] due to its ability to decouple temporal and spatial feature, which significantly reduces the number of parameters.

Self-attention The main procedure in an ADT [13] analysis for TC on oceans is performing scene analysis. Specifically, the cloud patterns are classified into one of many pattern types, where each pattern type is associated with its unique analysis sub-procedure. Success of applying the pattern classification rules heavily rely on accurately locating the storm eye, since TC features such as the curved band are defined relative to the eye.

The purpose of incorporating self-attention is to simulate scene analysis where important regions on the feature map associated with the particular scene are emphasized. Self-attention generates the attention map based on the input feature map itself and is applied before the recurrent layer so the attention is determined solely on spatial information such as the location of the storm eye.

Sequence attention ADT analysis outputs several T-numbers (T#) which can be combined and calculated to estimate the current TC intensity as well as forecast 24h intensity. The raw T# represents the estimated purely from analyzing the satellite image the final T# represents estimate that is time-averaged over 6h (originally 12h). The CI# is based on final T# subject to some rule constraint, which mainly serves as the current intensity estimation and the FI# is simply an extrapolation from past time steps to serve as intensity forecast.

In order to mimic the behavior of time-averaging with soft supervision unlike hard coded rules when deriving CI# from final T#, we added the sequence attention module to the recurrent layer. For each time step, we calculate the similarity between the current frame and past time frames to serve as weights to linearly combine past time frames, which is then concatenated to the current time frame. Here, the number of past frames to look back is a hyper-parameter where longer is not necessarily better due to inherent chaos in TCs. Thus, the input to the recurrent layer contains the current state of the TC as well as the concentrated history which can help stabilize predictions. We chose Luong-style attention for convenience but expect other attention mechanisms to work as well.

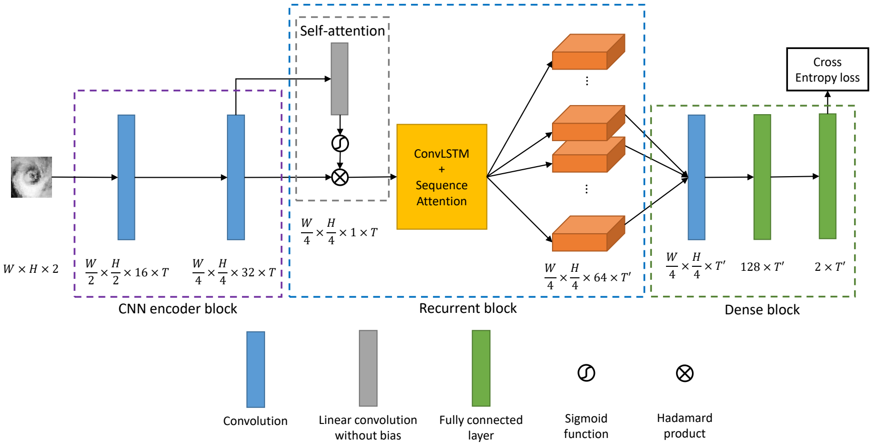
Self-attention
ConvLSTM + Sequence Attention
Cross Entropy loss
$W \times H \times 2$
$\frac{W}{2} \times \frac{H}{2} \times 16 \times T$
$\frac{W}{4} \times \frac{H}{4} \times 32 \times T$
$\frac{W}{4} \times \frac{H}{4} \times 1 \times T$
$\frac{W}{4} \times \frac{H}{4} \times 64 \times T'$
$\frac{W}{4} \times \frac{H}{4} \times T'$
$128 \times T'$
$2 \times T'$
CNN encoder block
Recurrent block
Dense block
Convolution
Linear convolution without bias
Fully connected layer
Sigmoid function
Hadamard product

Fig. 1. Model architecture.

5 Experiment and Analysis

For the experiment, TC satellite image series of length T are first passed through the CNN backbone for feature extraction, then passed into the recurrent layer for combining temporal features from different time frames. Since when calculating sequence attention the length of depended history is fixed, the first $T - T'$ time steps are dropped in the recurrent layer output due to insufficient information. Finally, the three-dimensional features are compressed into T' output of logits for RI corresponding to each time step in the series.

We train each model for 500 epochs and use ADAM optimizer to minimize the class-weighted cross entropy ($\text{pos} : \text{neg} = 20 : 1$) due to the largely unbalanced number of RI events in an entire TC series. Learning rate of $5 \cdot 10^{-4}$ is selected from $[10^{-3}, 10^{-4}, 5 \cdot 10^{-4}, 10^{-5}, 10^{-6}]$ with no decay scheduling. The weights except biases are L2-regularized with factor of 10^{-5}, selected from $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$. All dense layers have dropout rate of 0.9, which do not interfere with batch normalization as they are all placed in the back. We conduct ablation study on the attention techniques to clarify their function.

5.1 Self-attention

Fig 2 and 3 show that the model focuses on the storm eye and some portion of the outer curved band, which matches closely with how ADT would analyze the TC. We observed that self-attention effectively smooths predictions near the actual RI frames. Fig 4 plots the distribution of the span of maximum peak of TC series, The span is defined as the length of frames neighboring the peak, where frames prior to and after the peak decreases in magnitude monotonically. Sharp peaks implies that the predicted probability of RI occurrence perturbs frequently. When RI occurs, the cloud formation generally changes dramatically in short periods of time. Self-attention serves as an importance filter to amplify key regions while muting noises to help the model predict similar intensity for neighboring time frames near high RI probaility regions.

5.2 Sequence attention

Recall the goal of sequence attention is to stabilize the predictions over time, much like how time averaging is used in ADT, which is ideal as RI usually occurs in continuous time span due to the atmosphere being a continuously dynamic system.

We observed that predictions with sequence attention is smoother throughout the entire TC series. In fact, it is so smooth that the smoothness causes the model prediction to be slightly delayed compared to the actual event. Fig 5 plots the distribution of time difference between the predicted RI probability peak and the actual intensity difference peak. We expect that a larger intensity difference is reflected as higher RI probability predicted by the model. A positive time difference implies prediction delay, which is worse than a negative time difference which implies early prediction. According to the graph, sequence attention causes the peak to delay more than the vanilla model. Since the input to the recurrent layer consists of half of current time frame and half of concentrated history, the dependence on the current frame is essentially halved. This causes hints of RI occurrence observed in the current time frame to be not as dominating, which is a trade-off between model stability and sensitivity.

Sequence attention also causes more underestimation of RI probability, aligning with the delayed-peak observation. With the current input combined with attended history, in some sense, the intensity is averaged over the all the past time frames and dilutes the intensity and cause underestimation, which is not all bad since consistent and smoother results is desired. Due to the unbalanced nature of the dataset, where the number of negative samples dominates, it is particularly likely to train a model that underestimates. In conclusion, sequence attention acts as a double-edged sword, where the introduced stability may or may not be beneficial depending on the context. More careful fine-tuning of how much history the sequence attention depends on is essential to find the balancing point.

5.3 Performance

According to table 1, our best performing model for BSS is the self-attention variant (43 % improvement), and the best for HSS is the combined variant (11 % improvement). Interestingly direct fusion of self-attention and sequence attention does not necessarily yield better results, which yields potential for better incorporation of the two attention mechanisms. The reliability diagram (fig 6) is also plotted to show the correlation between predicted probability and true probability. The larger slope is caused by class-weighting of losses, as the model learns to prefer overestimating than underestimating, which shows similar trends with results in previous works [10,16].

Table 1. Skill scores for performance evaluation. Performance is evaluated on the validation set and the two skill scores correspond to the same model.

Model	**Brier skill score**	**Heidke skill score**
Blended SVM+RF+MLP [10]	0.35	0.3
Our model (no attention)	-0.098	0.329
Our model (self-attention)	0.004	0.324
Our model (sequence attention)	-0.13	0.301
Our model (both attentions)	-0.09	0.334

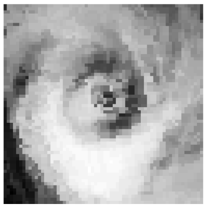

Fig. 2. Satellite image

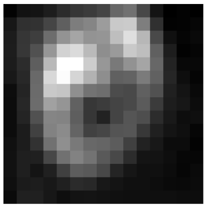

Fig. 3. Attention mask

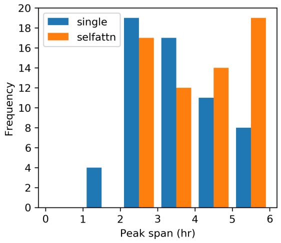

Fig. 4. Peak sharpness

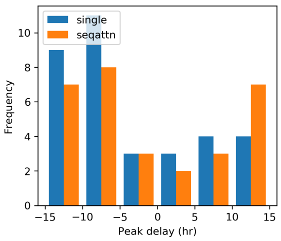

Fig. 5. Peak delay

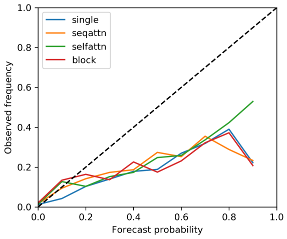

Fig. 6. Reliability diagram

6 Conclusion

We proposed a deep learning model which achieves state-of-the-art performance on the cyclone rapid intensification task relying solely on satellite images, which to the best of our knowledge is the first. In comparison, previous works for RI prediction rely on predictors that requires expensive computation resources and human manual efforts. We thoroughly experimented with different model designs with attention mechanisms inspired by existing successful TC intensity prediction domain knowledge. Our model does not perform that well on BSS due to the metric itself being insensitive to class imbalance while slightly improving in HSS. Furthermore, we explain the predictions by visualizing attention heatmaps and yield insight regarding the influence of each model design on the prediction tendency.

For future works, we wish to explore improved designs of attention mechanisms to capture intrinsic properties among time frames tailored for the RI prediction task. The stability that sequential attention provides demonstrates the potential of reducing false positives with proper modification to the vanilla mechanism. We also wish to explore the possibility of augmenting training data with generative adversarial networks to improve robustness.

References

1. Chen, B., Chen, B.F., Lin, H.T.: 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. pp. 90–99. ACM (2018)
2. Chen, B.F., Chen, B., Lin, H.T., Elsberry, R.L.: Estimating tropical cyclone intensity by satellite imagery utilizing convolutional neural networks. Weather and Forecasting **34**(1), 447–465 (2019)
3. Gall, R., Franklin, J., Marks, F., Rappaport, E.N., Toepfer, F.: The hurricane forecast improvement project. Bulletin of the American Meteorological Society **94**(3), 329–343 (2013)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Hendricks, E.A., Peng, M.S., Fu, B., Li, T.: Quantifying environmental control on tropical cyclone intensity change. Monthly Weather Review **138**(8), 3243–3271 (2010)
6. Kaplan, J., DeMaria, M., Knaff, J.A.: A revised tropical cyclone rapid intensification index for the atlantic and eastern north pacific basins. Weather and forecasting **25**(1), 220–241 (2010)
7. Kim, S., Kang, J.S., Lee, M., Song, S.k.: Deeptc: Convlstm network for trajectory prediction of tropical cyclone using spatiotemporal atmospheric simulation data (2018)
8. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)

9. Matsuoka, D., Nakano, M., Sugiyama, D., Uchida, S.: Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model. Progress in Earth and Planetary Science **5**(1), 80 (2018)
10. Mercer, A., Grimes, A.: Atlantic tropical cyclone rapid intensification probabilistic forecasts from an ensemble of machine learning methods. Procedia computer science **114**, 333–340 (2017)
11. Miyamoto, Y., Nolan, D.S.: Structural changes preceding rapid intensification in tropical cyclones as shown in a large ensemble of idealized simulations. Journal of the Atmospheric Sciences **75**(2), 555–569 (2018)
12. Nolan, D.S., Moon, Y., Stern, D.P.: Tropical cyclone intensification from asymmetric convection: Energetics and efficiency. Journal of the Atmospheric Sciences **64**(10), 3377–3405 (2007)
13. Olander, T.L., Velden, C.S.: The advanced dvorak technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. Weather and Forecasting **22**(2), 287–298 (2007). https://doi.org/10.1175/WAF975.1, `https://doi.org/10.1175/WAF975.1`
14. Pradhan, R., Aygun, R.S., Maskey, M., Ramachandran, R., Cecil, D.J.: Tropical cyclone intensity estimation using a deep convolutional neural network. IEEE Transactions on Image Processing **27**(2), 692–702 (2017)
15. Rozoff, C.M., Kossin, J.P.: New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. Weather and Forecasting **26**(5), 677–689 (2011)
16. Rozoff, C.M., Velden, C.S., Kaplan, J., Kossin, J.P., Wimmers, A.J.: Improvements in the probabilistic prediction of tropical cyclone rapid intensification with passive microwave observations. Weather and Forecasting **30**(4), 1016–1038 (2015)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
18. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
19. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802–810 (2015)
20. Zhang, L., Zhu, G., Mei, L., Shen, P., Shah, S.A.A., Bennamoun, M.: Attention in convolutional lstm for gesture recognition. In: Advances in Neural Information Processing Systems. pp. 1953–1962 (2018)