Towards Interpretable Deep Learning for Accelerating Scientific Discoveries in Climate Science

Anh N. Nhu
University of Maryland, College Park
anhu@terpmail.umd.edu
SRC: Undergraduate | ACM Member Number: 1908799

Yiqun Xie
University of Maryland, College Park
xie@umd.edu
Research Advisor

ABSTRACT

While deep learning models have high representation power and promising performances, there is often a lack of evidence to interpret potential reasons behind the predictions, which is a major concern limiting their usability for scientific discovery. We propose a Neural Additive Convolutional Neural Network (NA-CNN) to enhance the interpretability of the model to facilitate scientific discoveries in climate science. To investigate the interpretation quality of NA-CNN, we perform experiments on the El Niño identification task where the ground truth for El Niño patterns is known and can be used for validation. Experiment results show that compared to Spatial Attention and state-of-the-art post-hoc explanation techniques, NA-CNN has higher interpretation precision, remarkably improved physical consistency, and reduced redundancy. These qualities provide an encouraging ground for domain scientists to focus their analysis on potentially relevant patterns and derive laws governing phenomena with unknown physical processes.

CCS CONCEPTS

• **Applied computing**; • **Computing methodologies** → **Machine learning**; *Spatial and physical reasoning*;

KEYWORDS

Explainable AI, Deep Learning, Scientific Discovery

ACM Reference Format:
Anh N. Nhu and Yiqun Xie. 2023. Towards Interpretable Deep Learning for Accelerating Scientific Discoveries in Climate Science. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3589132.3628369

1 INTRODUCTION

In recent years, deep learning has been widely applied in various Earth Science domains, including climate science, geology, and oceanography. With the emergence of eXplainable Artificial Intelligence (XAI), deep learning's usability extends beyond traditional classification and predictive tasks. For instance, [8] used patterns detected by XAI to calibrate model trust. Most prominently, prior work [9] investigated the usability of XAI for scientific discovery by attempting to re-discover known scientific patterns using XAI explanations. [9] applied Layer-wise Relevance Propagation (LRP) [2], a gradient-based post-hoc XAI method, to uncover learned patterns from trained CNN and re-discover the ENSO's Sea Surface Temperature (SST) patterns. Their results demonstrated that – when applied to the study of unknown phenomena – XAI holds promises to help domain scientists narrow down hypotheses and accelerate the detection of scientifically meaningful yet undiscovered patterns. Despite the advancements, model-agnostic post-hoc XAI paradigms such as LRP [2], CAM [3, 7, 10], and DeepSHAP [4] are shown to have a number of limitations, including shattering gradients and disentangling positive/negative contributions, impacting the trustworthiness of their interpretations [5]. Our experiments in Fig. 2 also indicate that post-hoc XAI algorithms produce highly varied interpretations even for identical model weights. This instability further limits the credibility of the resulting interpretation.

2 PROPOSED APPROACH

Inspired by the Neural Additive Model [1], we introduce the Neural Additive Convolutional Network (NA-CNN) whose design is more naturally interpretable. NA-CNN's primary advantage is that its interpretation is aligned with how we interpret simple linear models (e.g., logistic regression), which are generally appreciated by their interpretability. This design allows domain scientists to interpret the individual contribution of each feature in a straightforward and intuitive manner. Moreover, since NA-CNN does not use gradients for producing interpretation, it is less vulnerable to problems such as gradient shattering and saturation [5].

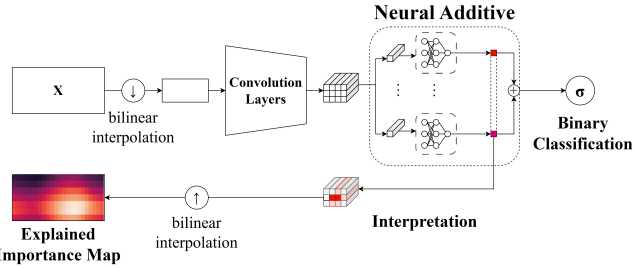

Figure 1: Neural Additive CNN Architecture (image-level binary classification with intermediate interpretation).

NA-CNN consists of two primary building blocks: (1) convolutional layers for spatial feature extraction and (2) neural additive networks, each acting as a shape function for individual spatial locations. Formally, convolutional layers can be defined as $f_{conv} : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{c'\times h'\times w'}$, then there are $h'\times w'$ independent fully-connect networks: $f^i_{fc} : \mathbb{R}^{c'} \rightarrow \mathbb{R}$, where $i \in [1, h'\times w']$. Each fully connected network maps the representation in each spatial location to a scalar value. The mapped scalars from all locations are

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11.
https://doi.org/10.1145/3589132.3628369

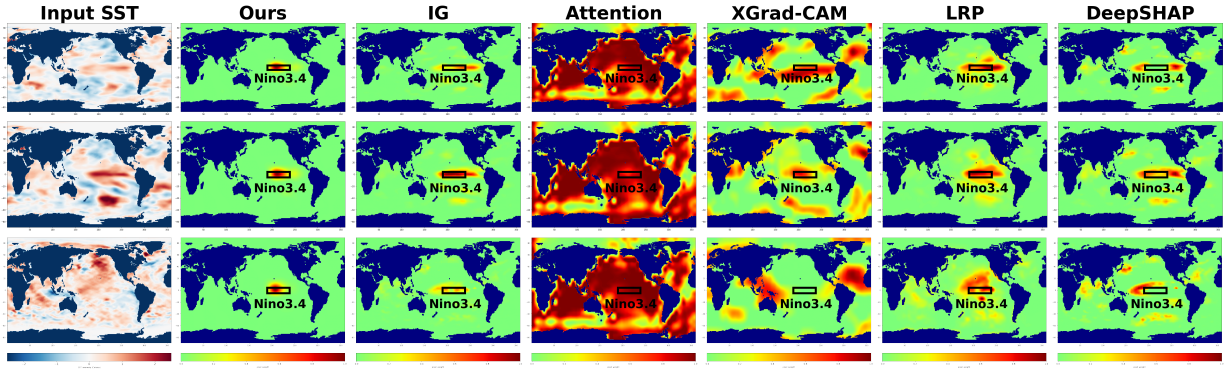

Figure 2: Output interpretations of candidate methods for sampled El Niño cases. Note that the model is trained with scene-level binary classification, and the maps are intermediate interpretation results.

then additively combined and activated by the sigmoid function, which is shown in Fig. 1. Suppose $\mathbf{X}$ is the input SST map, the end-to-end algorithmic flow is as follows:

$$\mathbf{M}^{c' \times h' \times w'} = f_{conv}(\mathbf{X}^{C \times H \times W}) \tag{1}$$

$$o^i = f^i_{fc}(\mathbf{M}_i) \quad \forall i \in [1,\ h' \times w'] \tag{2}$$

$$\hat{y} = sigmoid(\sum_{i=1}^{h' \times w'} o^i) \tag{3}$$

The linear additive operation in Neural Additive architecture makes it straightforward to identify high-contribution locations in the SST map. Furthermore, as shown in [1], we can also visualize specific shape function f^i_{fc} to gain concrete insights into how SST anomalies variations at a particular region relevant to El Niño.

3 EXPERIMENT RESULTS

To illustrate the effectiveness of NA-CNN, we validate its capability to discover known ground-truth patterns of the El Niño-Southern Oscillation (ENSO). Specifically, we train our model on the global monthly Sea Surface Temperature (SST) anomalies to perform **scene-level binary classification** of El Niño (class 1) and La Niña (0). SST anomaly maps are from the NOAA PSL Climate Data Repository [6], covering 1578 monthly observations from 1891 to 2020. To label El Niño and La Niña phases, we use the Niño3.4 index - spatial average of equatorial Pacific Ocean SST anomalies ($5°N - 5°S$, $170°W - 120°W$) - where cases with Niño3.4 > 0.5 and Niño3.4 < -0.5 correspond to El Niño and La Niña, respectively.

NA-CNN's predictive performance is comparable to the performance of conventionally black-box models with overall accuracy, precision, and recall of 99.8%, 99.98%, and 99.7%, respectively. To compare the interpretability of NA-CNN with other XAI approaches, we show their interpretations on 3 randomly selected SST maps of anomalies. All post-hoc XAI techniques (Integrated Gradient, LRP, XGrad-CAM, DeepSHAP) are applied on the same base CNN. We qualitatively evaluate the interpretation based on 3 primary criteria: **(1) Pattern alignment**: The interpretation should highlight locations with high overlap with Niño3.4 region ($5°N - 5°S$, $170°W - 120°W$), which is the true pattern based on domain knowledge. **(2) Physical consistency**: Since El Niño is defined by only SST anomalies within the Niño3.4 region, the interpretation should consistently highlight attributions at the Niño3.4 region across different observations. Such consistency is critical to guide where domain scientists should focus their pattern analysis for scientific discovery. **(3) Reduced distraction**: Interpretation should minimize identifying less useful or irrelevant features, avoiding providing unimportant noises and distracting domain scientists from the real signals. In this case study, the significance of a location is inversely proportional to its distance from the Niño3.4 region. Fig. 2 shows that NA-CNN's interpretations has the highest level of pattern alignment and reduced distracting patterns compared to other methods. It consistently highlights interpretation in the same area, which is important for pinpointing major hypotheses.

4 CONCLUSION AND FUTURE WORK

We proposed the NA-CNN framework to enhance model interpretability to facilitate scientific discoveries. Our experiments on climate science datasets for El Niño confirmed the improvements of NA-CNN in consistently aligning the interpretable evidence with known physical patterns. Our future work will integrate this with other architectures (e.g., ViT) and domain problems. Furthermore, NA-CNN focuses on interpreting spatial patterns and lacks the ability to interpret more complex spatio-temporal patterns. In future work, we will extend NA-CNN to interpret spatio-temporal models.

ACKNOWLEDGMENTS

This work was supported by NSF awards 2147195, 2105133 & 2126474.

REFERENCES

[1] Rishabh Agarwal et al. 2021. Neural Additive Models: Interpretable Machine Learning with Neural Nets. In *NeurIPS 2021*.
[2] Sebastian Bach et al. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One* 10, 7 (2015), e0130140.
[3] Ruigang Fu et al. 2020. Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. arXiv:2008.02312
[4] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *NeurIPS* (2017).
[5] Antonios Mamalakis et al. 2022. Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience. *AMS AI for the Earth Systems* 1, 4 (2022), e220012.
[6] NOAA. 2019. NOAA PSL Climate Data Repository. www.psl.noaa.gov/repository/entry/show?entryid=f45cf25c-bde2-44bd-bf3d-c943d92c0dd8
[7] Ramprasaath R Selvaraju et al. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of IEEE ICCV*. 618–626.
[8] Maike Sonnewald and Redouane Lguensat. 2021. Revealing the Impact of Global Heating on North Atlantic Circulation Using Transparent Machine Learning. *AGU JAMES* 13, 8 (2021), e2021MS002496.
[9] Benjamin A. Toms et al. 2020. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *AGU JAMES* 12, 9 (2020).
[10] Bolei Zhou et al. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE CVPR*. 2921–2929.