Neuron Explanations for Conformal Prediction (Student Abstract)

Divya Lidder, Kathryn Morse, Bridget Sullivan, Wei Qian, Chenglin Miao, Mengdi Huai*

Department of Computer Science, Iowa State University
Atanasoff Hall B10, 2434 Osborn Dr, Ames, IA 50011
(716) 361-6578
{dslidder, kamorse, bgs2, wqi, cmiao, mdhuai}@iastate.edu

Abstract

Conformal prediction (CP) has gained prominence as a popular technique for uncertainty quantification in deep neural networks (DNNs), providing statistically rigorous uncertainty sets. However, existing CP methods fail to clarify the origins of predictive uncertainties. While neuron-level interpretability has been effective in revealing the internal mechanisms of DNNs, explaining CP at the neuron level remains unexplored. Nonetheless, generating neuron explanations for CP is challenging due to the discrete and non-differentiable characteristics of CP, and the labor-intensive process of semantic annotation. To address these limitations, this paper proposes a novel neuron explanation approach for CP by identifying neurons crucial for understanding predictive uncertainties and automatically generating semantic explanations. The effectiveness of the proposed method is validated through both qualitative and quantitative experiments.

Introduction

Conformal prediction (CP) (Shafer and Vovk 2008; Sesia and Romano 2021; Ndiaye 2022; Li et al. 2024; Sun et al. 2024) has emerged as a widely used technique for uncertainty quantification in deep neural networks (DNNs), offering statistically rigorous uncertainty sets for model predictions. The model-agnostic and distribution-free nature of CP makes it particularly applicable for large neural networks, compared to other uncertainty quantification methods, such as using Bayesian networks (Sun, Chen, and Carin 2017).

Although existing CP methods can offer guaranteed uncertainty sets, they fail to clarify the origins of predictive uncertainties. Identifying the causes of predictive uncertainties is crucial for understanding the model behavior, especially in error-sensitive domains like autonomous driving and medical diagnosis, where knowing the roots of uncertainty can help reduce risks (Chen et al. 2024; Qian et al. 2024). Therefore, explaining predictive uncertainties in CP is essential for DNNs. The interpretability of individual neurons in DNNs has proven to be effective in exploring the internal mechanisms of DNNs (Bau et al. 2020). Neuron-level explanations can be a powerful means of revealing the semantic concepts encoded within neurons, thereby enhancing the comprehension of how models make decisions. However, there is still a gap in explaining CP at the neuron level, which remains largely unexplored. Providing neuron-level explanations for CP is challenging due to the discrete and non-differentiable CP uncertainty results, and the labor-intensive process of semantic annotation.

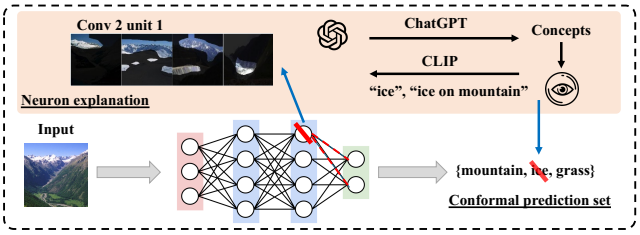

Figure 1: An overview of neuron explanation for conformal prediction with large foundational models.

To address the above challenges, we propose a novel neuron explanation approach for CP, which can identify the neurons associated with the predicted labels in the uncertainty sets and automatically generate semantic explanations for features represented by these neurons. Specifically, in our approach, we design a novel optimization framework to find the most influential neurons for uncertainty predictions, and then we use large foundational models to extract the semantic meanings of the features within these neurons. We validate the effectiveness of our proposed approach through a series of qualitative and quantitative experiments, demonstrating its ability to explain CP with neuron semantics.

Methodology

Consider a training dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$ with $X_i \in \mathcal{X} \subset \mathbb{R}^d$ as the input sample and $Y_i \in \mathcal{Y} = [C]$ as the class label. Let f_{θ^*} denote a DNN classifier trained on the dataset $\mathcal{D}$ using a loss function $\mathcal{L}$ (e.g., cross-entropy), where $\theta^* \in \Theta$ represents the model parameters. Then, we use $f_{\theta^*}^{l,k}$ to denote the k-th neuron in layer l. In CP (Shafer and Vovk 2008; Straitouri et al. 2023; Wang et al. 2024), given a significance level $\varepsilon \in (0, 1)$, we compute a quantile $Q_{1-\varepsilon}$ based on a set of calibration data. Then, for a new test example X_{ts}, we can construct its conformal set $\mathcal{C}(X_{ts})$ with *marginal coverage*,

*indicates corresponding author.

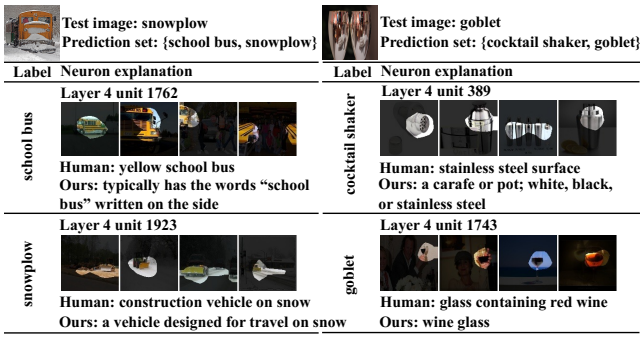

Figure 2: Illustrations of neuron explanation for uncertainty.

where $\mathcal{C}(X_{ts}) = \{Y \in [C] : \mathcal{S}((X_{ts}, Y); \theta^*) \leq Q_{1-\varepsilon}\}$, where $\mathcal{S}(\cdot)$ is the non-conformity measure. Let Y_{ts} denote the true label for X_{ts}. Note that when the involved data are exchangeable, the marginal coverage guarantee implies that $\mathbb{P}(Y_{ts} \in \mathcal{C}(X_{ts})) \geq 1 - \varepsilon$, meaning that $\mathcal{C}(X_{ts})$ will contain the true label with probability at least $1 - \varepsilon$.

Based on the above, our goal is to design a novel neuron explanation approach for CP, as outlined in Figure 1. First, we aim to construct a neuron semantic pool for the neurons in the pre-trained model θ^*. To achieve this, we first extract a set of activated image patches associated with the features of each neuron $f_{\theta^*}^{l,k}$ (Zhou et al. 2015; Bau et al. 2020; Zhao et al. 2023). Then, we propose promoting large language models (e.g., ChatGPT) to obtain comprehensive semantic concepts (e.g., phrases) about features in all class labels. After that, we adopt vision language models (e.g., CLIP (Radford et al. 2021)) to align the concepts with the image patches. Thus, this process allows us to generate meaningful explanations to understand the neuron semantics.

Next, we focus on identifying the neurons responsible for the predicted labels in the uncertainty set and leveraging the semantics in neurons to explain the origins of uncertainty. Given X_{ts} with its uncertainty set $\mathcal{C}(X_{ts})$, for a specific target label $Y_{tg} \in \mathcal{C}(X_{ts})$, we aim to find the minimum set of neurons in layer l of the network such that, after masking these neurons, Y_{tg} is excluded from the new uncertainty set $\tilde{\mathcal{C}}(X_{ts})$ (i.e., $Y_{tg} \notin \tilde{\mathcal{C}}(X_{ts})$), while the remaining labels in $\mathcal{C}(X_{ts})$ are retained in the new set (i.e., $(\mathcal{C}(X_{ts}) \setminus Y_{tg}) \subseteq \tilde{\mathcal{C}}(X_{ts})$). Thus, we propose

$$\arg \min_{M_l \subseteq U_l} |M_l| \tag{1}$$
$$s.t. \quad \max \mathcal{S}((X_{ts}, Y_{tg}); \theta^*_{M_l}) - \tilde{Q}_{1-\varepsilon},$$

where U_l represents all the neurons $\{f_{\theta^*}^{l,k}\}_{k=1}^{|U_l|}$, $\theta^*_{M_l}$ denotes the model after masking the neurons M_l, and $\tilde{Q}_{1-\varepsilon}$ denotes the quantile computed on the calibration data using $\theta^*_{M_l}$. Here, neurons can be represented as a binary vector, with M_l indicating which neurons are inactive during forwarding. Note that directly solving Eq. (1) is infeasible due to its discrete nature and the non-differentiability of the quantile. To address this, we propose to solve the above optimization problem via an empirical search.

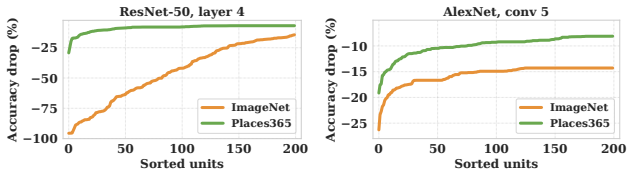

Figure 3: Impact of the neurons on class accuracy.

Experiments

Experimental Settings. In experiments, we utilize ImageNet and Places365 (Zhou et al. 2017) datasets. We analyze the features of neurons in the convolutional layers of ResNet-50 and AlexNet. Two widely adopted conformal methods are employed: HPS (Lei, Robins, and Wasserman 2013), which is based on softmax output, and APS (Romano, Sesia, and Candes 2020), which offers marginal coverage. We use human annotations as the baseline. *More experimental details and results can be found in the full paper.*

Experimental Results. First, we evaluate the performance of neuron explanations for CP. We use GPT-3.5 to generate feature descriptions for classes in ImageNet and Places365, and we employ the ViT-B/32 model from CLIP to label neuron descriptions for the activated image patches. The patch generated for each unit (neuron) is shown on four maximally activating images. Figure 2 presents the explanation results. As shown, our proposed approach successfully identifies the neurons that are most influential on the predicted labels in the prediction set. Additionally, our proposed approach efficiently captures the features represented in these neurons and describes the corresponding semantic meanings, leveraging the large models. For instance, the inclusion of the label "school bus" in the prediction set of the "snowplow" image is driven by unit 1762 in layer 4, which shows the related features of the "school bus" with the image patch, and the neurons' descriptions accurately reflect the semantics of these features. We observe that our descriptions are highly aligned with the baseline of human annotations. Next, we validate the influence of neurons on model predictions by computing the maximum accuracy drop of individual classes after masking the units. The results for the 200 sorted units are shown in Figure 3, highlighting the relative importance of neurons that capture features in influencing model predictions. Therefore, our method demonstrates the capability to identify important neurons and generate comprehensive semantic meanings without human annotations.

Conclusion

This paper presents a novel neuron explanation approach for conformal prediction. The proposed approach identifies the crucial neurons responsible for predictive uncertainty sets and automatically generates semantic explanations without requiring human interventions. Extensive experiments have been performed to validate the effectiveness of the proposed approach. We believe that this approach will serve as a valuable tool for the research community, enabling understanding of the uncertainty in conformal prediction.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grant CNS-2350332. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *In Proceedings of the National Academy of Sciences (PNAS)*.

Chen, A.; Li, Y.; Qian, W.; Morse, K.; Miao, C.; and Huai, M. 2024. Modeling and Understanding Uncertainty in Medical Image Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 557–567. Springer.

Lei, J.; Robins, J.; and Wasserman, L. 2013. Distribution-free prediction sets. *Journal of the American Statistical Association (JASA)*.

Li, Y.; Chen, A.; Qian, W.; Zhao, C.; Lidder, D.; and Huai, M. 2024. Data Poisoning Attacks against Conformal Prediction. In *Forty-first International Conference on Machine Learning*.

Ndiaye, E. 2022. Stable conformal prediction sets. In *International Conference on Machine Learning*. PMLR.

Qian, W.; Zhao, C.; Li, Y.; Ma, F.; Zhang, C.; and Huai, M. 2024. Towards Modeling Uncertainties of Self-Explaining Neural Networks via Conformal Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14651–14659.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*.

Sesia, M.; and Romano, Y. 2021. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34: 6304–6315.

Shafer, G.; and Vovk, V. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*.

Straitouri, E.; Wang, L.; Okati, N.; and Rodriguez, M. G. 2023. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, 32633–32653. PMLR.

Sun, J.; Jiang, Y.; Qiu, J.; Nobel, P.; Kochenderfer, M. J.; and Schwager, M. 2024. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. *Advances in Neural Information Processing Systems*, 36.

Sun, S.; Chen, C.; and Carin, L. 2017. Learning structured weight uncertainty in bayesian neural networks. In *Artificial Intelligence and Statistics*, 1283–1292. PMLR.

Wang, J.; Zhao, C.; Lyu, L.; You, Q.; Huai, M.; and Ma, F. 2024. Bridging Model Heterogeneity in Federated Learning via Uncertainty-based Asymmetrical Reciprocity Learning. *arXiv preprint arXiv:2407.03247*.

Zhao, C.; Qian, W.; Shi, Y.; Huai, M.; and Liu, N. 2023. Automated natural language explanation of deep visual neurons with large models. *arXiv preprint arXiv:2310.10708*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2015. Object detectors emerge in deep scene cnns. *International Conference on Learning Representations (ICLR)*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.