Enhancing Explainability in Convolutional Neural Networks Using Entropy-Based Class Activation Maps

Éric Boketsu-Boulu, Ghazal Rouhafzay
Department of Computer Science, University of Moncton, NB ,Canada

ABSTRACT

With the emergence of visual sensors and their widespread application in intelligent systems, precise and interpretable visual explanations have become essential for ensuring the reliability and effectiveness of these systems. Sensor data, such as that from cameras operating in different spectra, LiDAR, or other imaging modalities, is often processed using complex deep learning methods, whose decision-making processes can be unclear. Accurate interpretation of network decisions is particularly critical in domains such as autonomous vehicles, medical imaging, and security systems. Moreover, during the development and deployment of deep learning architectures, the ability to accurately interpret results is crucial for identifying and mitigating any sources of bias in the training data, thereby ensuring fairness and robustness in the model's performance. Explainable AI (XAI) techniques have garnered significant interest for their ability to reveal the rationale behind network decisions. In this work, we propose leveraging entropy information to enhance Class Activation Maps (CAMs). We explore two novel approaches: the first replaces the traditional gradient averaging scheme with entropy values to generate feature map weights, while the second directly utilizes entropy to weigh and sum feature maps, thereby reducing reliance on gradient-based methods, which can sometimes be unreliable. Our results demonstrate that entropy-based CAMs offer significant improvements in highlighting relevant regions of the input across various scenarios.

BACKGROUND

The rapid advancements in deep learning have produced highly complex models, which are increasingly utilized in critical applications. However, these models often operate as "black boxes," with difficult-to-interpret inner workings. This has driven the need for Explainable AI (XAI), a field focused on elucidating the decisions and outputs of AI systems. Particularly, Convolutional Neural Networks (CNNs), widely employed in image classification, present unique interpretability challenges. Researchers have developed methods to reveal the factors driving CNN classification, highlighting influential image features.

Interestingly, XAI methods share objectives with weakly supervised object detection, which aims to locate object regions in images without exhaustive labeling. The techniques developed for XAI can thus assist in weakly supervised tasks. Despite significant progress, XAI faces challenges, particularly when handling multiple objects of the same class. Addressing these limitations is essential to enhance the robustness and trustworthiness of AI systems.

METHODS

Grad-CAM [1] highlights influential regions by averaging gradients across spatial dimensions. Equations (1) and (2) outline this process, where gradients with respect to the final convolutional layer $\frac{\partial y^c}{\partial A_{ij}^k}$ are averaged to compute weights for each feature map.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

$$L_{Grad-CAM}^c = ReLU\sum_k \alpha_k^c\, A^k \quad (2)$$

While effective, this averaging can overlook fine details, leading to coarse heatmaps. To improve precision, we propose entropy-based CAMs, using entropy as a measure of "disorder" in feature maps to capture richer patterns and improve spatial detail.

1- Grad Entropy-CAM replaces gradient averaging with entropy to weigh each feature map:

$$E_k^c = entropy(\frac{\partial y^c}{\partial A_{ij}^k}) \quad (3)$$

$$L_{Grad\_Entripy-CAM}^c = ReLU\sum_k E_k^c\, A^k \quad (4)$$

2- Feature Entropy-CAM reduces reliance on gradients by summing feature maps weighted by entropy directly:

$$L_{Frature\_Entripy-CAM}^c = ReLU\sum_k entropy(A^k)\, A^k \quad (5)$$

Entropy-based CAMs provide a more detailed understanding of network focus by capturing diverse and intricate features from input data, potentially enhancing interpretability in critical applications.

RESULTS

Figure 1 shows explanation heatmaps generated by Grad-CAM [1], HiResCAM [2], Respond-CAM [3], Grad Entropy CAM, and Feature Entropy CAM, illustrating the image regions most influential in the network's correct classifications (e.g., "goldfish," "Japanese spaniel," "Maltese dog," "Shih-Tzu," "Labrador retriever," "strawberry," "banana").

Entropy-based heatmaps, especially, offer a finer, more nuanced view of feature importance compared to Grad-CAM. These maps effectively expand critical regions to cover all object instances (e.g., rows 1 and 8) and more precisely localize other relevant objects (e.g., rows 6 and 7). By capturing the variability in gradient and feature maps, entropy-based methods provide richer feature representations, enhancing the clarity and accuracy of visual explanations.

CONCLUSIONS

In conclusion, this work introduces two entropy-based CAM visualization techniques that utilize the amount of information contained in gradients and feature maps to generate heatmaps. Comparative visual evaluation indicates that these methods offer enhanced precision in determining the exact importance and localization of relevant regions in the input. By incorporating pixel-level entropy, these techniques provide more detailed and accurate explanations of model behavior, making them particularly beneficial in applications requiring fine-grained analysis of input features.

REFERENCES

[1] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.
[2] Draelos, R.L., & Carin, L. (2020). Use HiResCAM Instead of Grad-CAM for Faithful Explanations of Convolutional Neural Networks. arXiv preprint, arXiv:2011.11293.
[3] Zhao, G., Zhou, B., Wang, K., Jiang, R., & Xu, M. (2020). Respond-CAM: Analyzing Deep Models for 3D Imaging Data by Visualizations. IEEE Transactions on Visualization and Computer Graphics, 27(1), 16–26