Comparing Ante-hoc and Post-hoc Interpretability Methods: SENN, LIME, and Integrated Gradients on CIFAR-10

Mattia Rampazzo[1]

[1]mattia.rampazzo@studenti.unitn.it

Github: https://github.com/mattia-rampazzo/SENN

Abstract— **This work evaluates the explanations generated by a Self-Explaining Neural Network (SENN) trained on CIFAR-10, comparing them to post-hoc methods LIME and Integrated Gradients (IG). The explanations are assessed across three criteria: intelligibility, faithfulness, and stability. The study reveals that SENN does not produce truly interpretable explanations in practice, while LIME and IG offer complementary strengths and weaknesses.**

Introduction

Deep learning algorithms have achieved remarkable success in a variety of complex tasks, leading to a widespread adoption in a large number of domains including high-stakes ones like finance and health-care. However, this surge in popularity has been accompanied by growing concerns over their lack of transparency. Often referred to as "black boxes," deep learning systems typically provide little insight into how or why they arrive at a particular prediction or decision, undermining the trust in their outputs. More importantly this opacity raises significant ethical questions, particularly when these decisions have a direct impact on human lives.

To address these issues, a growing subfield within artificial intelligence known as explainable AI (XAI) has emerged. Unlike traditional approaches that focus solely on performance, XAI aims to make machine learning models more transparent and their decision-making processes more interpretable. Over the years, this field has seen a proliferation of different methods which can be categorized mainly into two categories: ante-hoc methods and post-hoc methods.

Ante-hoc methods build interpretability directly into the model architecture. These models are designed to produce explanations as part of their output, offering a more seamless and often more computationally efficient interpretability process. A prominent example is the Self-Explaining Neural Network (**SENN**) [1], which aligns learned internal concepts with human-understandable coefficients.

Post-hoc methods, on the other hand, attempt to interpret a model's predictions after it has been trained. These approaches typically require additional computations and do not alter the original model. Examples include **LIME** [2], a model-agnostic technique, and **Integrated Gradients (IG)** [3], a gradient-based method that attributes input features to output predictions

In this work, I conduct a comparison of one ante-hoc method (SENN) and two post-hoc methods (LIME and Integrated Gradients), using the CIFAR-10 dataset, a widely used image classification benchmark in the machine learning community. By applying all three methods to the same task, I evaluate the quality of their explanations according to three key criteria:

- **Intelligibility (Explicitness):** How clear and understandable the explanation is to a human.
- **Faithfulness:** How accurately the explanation reflects the model's true reasoning.
- **Robustness (Stability):** How stable the explanation remains under small perturbations to the input.

I. Explanation methods

In this section, I introduce the three explanation methods under investigation: Self-Explaining Neural Networks (SENN), LIME, and Integrated Gradients.

A. *Self-Explaining Neural Networks (SENN)*

Self-Explaining Neural Networks (**SENN**) were first introduced by Alvarez-Melis and Jaakkola in 2018 as a way to build interpretability directly into the architecture of a neural network. Their approach is based on the idea that linear models are inherently interpretable, since their predictions can be directly traced back to individual input features and their corresponding weights. SENN generalizes this principle by extending the linear model into a more flexible, modular architecture, where each component is implemented as a separate neural network.

The key innovation in SENN lies in replacing the fixed terms of a linear model with learnable, input-dependent functions. The final model architecture is reported in Figure 1 and consists of three main components:

Conceptizer: This module is responsible for mapping the input data into a set of interpretable, high-level concepts. It is usually implemented as a shallow neural network. The goal is to extract a compact and meaningful representation of the input. To encourage interpretability and reduce redundancy, the concepts are often trained with a sparsity constraint, meaning that only a small number of them are activated for a given input.

Parameterizer: This module generates the relevance scores (or weights) associated with each concept, depending on the input. It is typically a deeper network than the conceptizer. Importantly, the authors introduce a robustness regularization term in the loss function to ensure that the relevance scores do not change too drastically in response to small changes in the input. This helps improve the stability of the explanations.

Aggregator: The final component combines the concepts and their corresponding relevance scores to produce the final prediction. This step is kept simple and interpretable by design and it is usually implemented as a linear combination of the concepts and weights.

SENN is trained end-to-end using a **composite loss function** that balances multiple objectives: accurate classification, meaningful concept representations, and stable, robust explanations. This makes the training process more complex, but results in a model that is both effective and inherently interpretable.

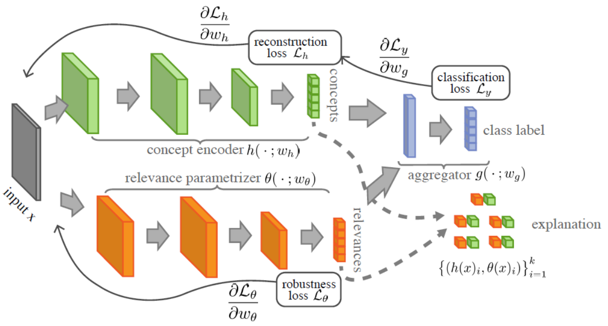
$\frac{\partial \mathcal{L}_h}{\partial w_h}$
reconstruction loss $\mathcal{L}_h$
$\frac{\partial \mathcal{L}_y}{\partial w_g}$
classification loss $\mathcal{L}_y$
concepts
class label
concept encoder $h(\cdot; w_h)$
relevance parametrizer $\theta(\cdot; w_\theta)$
aggregator $g(\cdot; w_g)$
input x
relevances
explanation
$\{(h(x)_i, \theta(x)_i)\}_{i=1}^k$
robustness loss $\mathcal{L}_\theta$
$\frac{\partial \mathcal{L}_\theta}{\partial w_\theta}$

Fig. 1 SENN architecture

B. *Local Interpretable Model-Agnostic Explanations (LIME)*

LIME is a widely used post-hoc explanation method that explains individual predictions by locally approximating the model with an interpretable one, such as a sparse linear model or decision tree. The core idea is to perturb the input data near the instance of interest and observe how the black-box model's predictions change. LIME then fits a simple, interpretable model to this locally sampled data, yielding feature importances that are meant to reflect the original model's behavior in that local region.

Notably, LIME is model-agnostic, making it flexible and widely applicable. However, it has been criticized for its lack of faithfulness (the explanation may not reflect the model's true reasoning) and instability (different runs can produce different explanations due to randomness in perturbation sampling).

C. *Integrated Gradients (IG)*

Integrated Gradients (IG) is a post-hoc explanation technique, designed for differentiable models like neural networks. It assigns feature importance by integrating gradients along a straight path from a baseline input to the actual input. For a model F, input x, and baseline x′, the attribution for the i-th feature is:

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial x_i}{\partial F(x' + \alpha(x - x'))} d\alpha$$

IG was designed to satisfy three key axioms: sensitivity, implementation invariance, and completeness. In practice, it satisfies completeness and implementation invariance by construction, but only a weaker form of sensitivity.

II. Methodology

To compare the explanations generated by different methods, a Self-Explaining Neural Network (SENN) was first trained on the CIFAR-10 dataset. After training, the same model was used as the basis for generating post-hoc explanations using LIME and Integrated Gradients. Training was conducted on Google Colab using a single NVIDIA T4 GPU. However, reproducing the reported results proved to be challenging, as noted in other studies as well [4]. In particular, the learned concepts were not as coherent or interpretable as those presented in the original work. To address this, extensive experimentation was carried out, including modifications to the model architecture and training settings, in an effort to improve the quality of the learned concepts. Despite these efforts, the results remained inconsistent. Ultimately, the final model was trained for 10 epochs on 5 concepts using the architecture and hyperparameters closely aligned with the original SENN paper: learning rate of 2×10^{-4}, sparsity regularization (ξ) set to 2×10^{-5}, and robustness regularization (λ) set to 1×10^{-4}.

To evaluate the quality of explanations produced by SENN in comparison to LIME and Integrated Gradients, I focused on three key properties, as proposed by the SENN authors:

Intelligibility: This refers to how understandable the explanations are to a human observer. My evaluation of intelligibility was purely qualitative and conducted on a small subsample of 10 images, one for each class.

Faithfulness: This measures how accurately an explanation reflects the model's actual decision-making process. Faithfulness as the author of SENN proposed, was measured conducting a series of ablation analysis. For SENN, ablation was performed on the entire test set. For the post-hoc methods (LIME and Integrated Gradients), I used a subsample of 100 images, ablating superpixels for LIME and the most important pixels (by percentage) for Integrated Gradients. While 100 images are not sufficient for statistically significant results, this approach still provides valuable insights into the explanations.

Stability: This captures how stable an explanation remains under small perturbations to the input. Stability was measured by adding varying levels of noise (from 0.001 to 0.01) to the images and then observing the similarity between the resulting explanations. Again for the post-hoc methods stability was evaluated on the subsample of 100 images.

III. Results

This section presents the evaluation results of explanations produced by the Self-Explaining Neural Network (SENN) and compares them to those generated by the post-hoc methods LIME and Integrated Gradients, based on the three key desiderata: intelligibility, faithfulness, and stability.

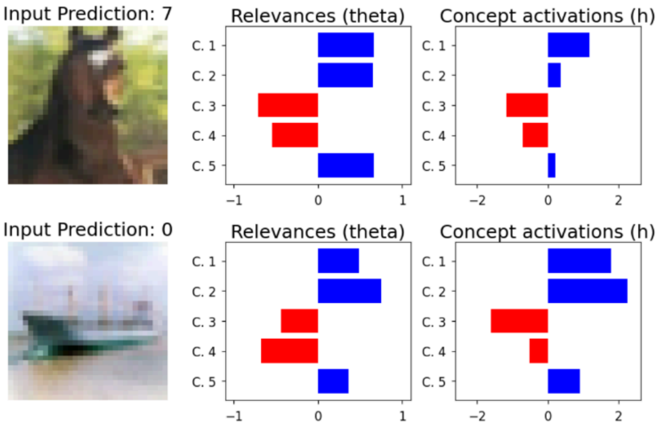

Fig. 2 SENN explanations

4.1 SENN Explanations

Intelligibility

SENN generates explanations by combining learned concept prototypes with input-dependent relevance scores. In theory, these should provide a clear and interpretable decomposition of the model's decision. However, in practice, the learned concepts were often not semantically meaningful.

As illustrated in Figure 2, explanations for multiple test examples appear strikingly similar. The figure shows two instances belonging to different classes: in the first case the model produces the correct prediction, while in the second it misclassifies the input. Nevertheless, for both examples and more broadly for the entire dataset the same three concepts (C1, C2, and C5) consistently receive large positive relevance scores, whereas other concepts (C3 and C4) are assigned negative relevance. This consistency occurs irrespective of the actual content of the input images.

Such repetitive patterns suggest that the model may not have learned distinct, informative concepts. Instead, it appears to rely on a fixed scoring mechanism that is detached from true semantic understanding.

Further investigation of the concept prototypes, using the evaluation methods proposed by the authors, confirmed that these learned concepts lack clear meaning. While some concepts, such as C5 (representing primarily white objects on a black background) or C2 (depicting airplanes on a white background), may seem interpretable, this interpretation is subjective and vulnerable to individual bias.

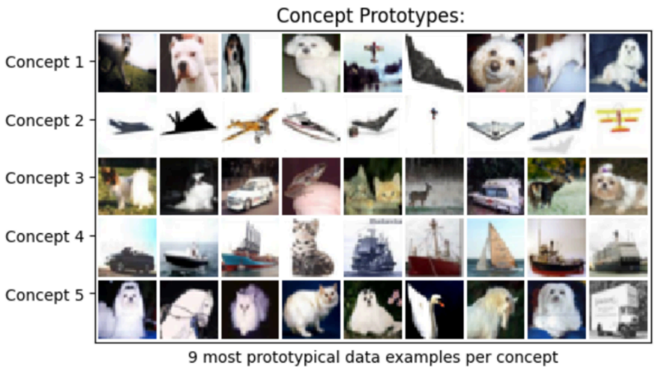

Fig. 3 SENN concepts

Faithfulness

Despite the disappointing results in terms of intelligibility, I proceeded to evaluate faithfulness: following the method proposed in the SENN paper. Specifically, I performed a concept ablation analysis: for each test image, I zeroed out one concept at a time and recorded the resulting drop in prediction confidence for the predicted class. In principle, concepts with higher relevance scores should lead to greater drops in confidence when ablated. However, the correlation between relevance scores and confidence drops was weak and inconsistent (0.139 ± 0.486), with high variance across samples. This suggests that the assigned relevance scores do not reliably reflect the actual contribution of each concept to the model's prediction.

Stability

To test stability, I perturbed each test image with a small amount of Gaussian noise and re-ran the explanation. A robust explanation should remain consistent under such minor input changes.

In the examined sample, I observed that the relevance scores generally remained stable, with only a few exceptions. Interestingly, for the dog category example, perturbations caused the relevance scores to shift dramatically, highlighting completely different concepts compared to the original, unperturbed image.

4.2 LIME Explanations

Intelligibility

By construction, LIME explanations are inherently interpretable, as they highlight input features, like superpixels in images, that contribute most to the model's prediction. Figure 4 shows LIME explanations for two different test examples, along with their associated masks.

For the "Car" class example, LIME accurately highlights relevant parts of the car, such as the windshield, which aligns well with human intuition. Interestingly, for the "Aircraft" class, LIME highlights background elements like grass and sky rather than the object itself. This suggests that the model may be relying on confounding factors or spurious correlations rather than the aircraft itself to make its prediction, providing valuable insight into the model's inner workings.

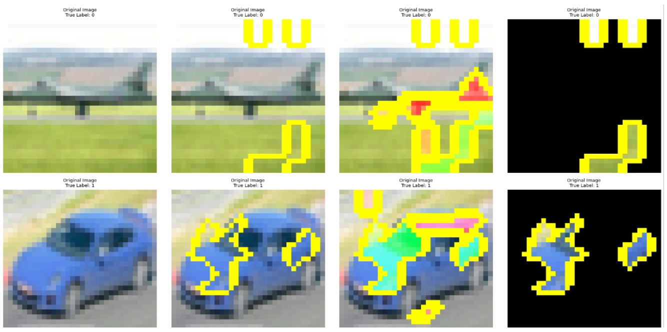

Fig. 4 LIME explanations

Faithfulness

To assess the faithfulness of LIME explanations, the top k% of superpixels with the highest absolute attribution were progressively removed, and the resulting changes in model behavior were measured. The results show a consistent, monotonic relationship between superpixel removal and model degradation.

Removing the top 5% led to a mean confidence drop of 0.0423 and a prediction flip rate of 12.00%. At 20% removal, confidence drop was 0.1386 with a 21.00% flip rate, and at 50%, these values rose to 0.3169 and 43.00%, respectively.

These trends suggest that LIME explanations are relatively faithful, with the most important superpixels having significant influence on the model's predictions

Stability

Stability analysis was performed on the subsample by applying five perturbations per image at increasing levels of noise and computing the Jaccard similarity of the top 10 most important superpixels. The results reveal considerable instability in LIME explanations. An overall Jaccard similarity of 0.1631 ± 0.1064 indicates substantial variation in the sets of important features identified for the same image across different perturbations. This variability is further underscored by 379 prediction changes, suggesting that the model's outputs are sensitive to even minor input alterations. Moreover, explanation stability declines as perturbation noise increases, with the mean Jaccard similarity dropping from 0.2218 (at σ = 0.01) to 0.1301 at (at σ = 0.1).

4.2 IG Explanations

Intelligibility

Integrated Gradients explanations were hard to interpret for this dataset because green (positive) and red (negative) pixels were often mixed together. This made it difficult to see clear patterns. Still, some larger green areas could be noticed, showing where the model was focusing. Interestingly, in Figure 5, IG highlighted some pixels related to the aircraft object as important for the prediction, contrary to the LIME explanation..

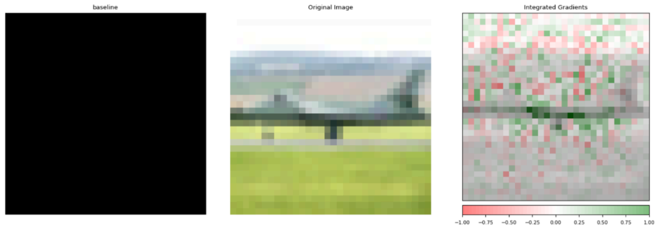

Fig. 5 IG explanations

Faithfulness

To assess the faithfulness of Integrated Gradients (IG) explanations, a systematic pixel removal strategy was applied to the selected subsample of test images. IG attributions were first computed, and then the top k% of pixels (ranked by highest attribution magnitudes) were progressively removed by zeroing out their values. The effects of this removal were measured using two key metrics: the reduction in model confidence for the originally predicted class and the rate of prediction flips to alternative classes.

Results showed a clear, monotonic relationship: as more pixels were removed, the model's confidence steadily declined (from 0.082 at 5% to 0.284 at 50%) and prediction flip rates rose from 22% to 44%. Remarkably, removing just the top 5% of pixels caused prediction changes in over 20% of cases, highlighting IG's strong faithfulness in identifying influential features.

Stability

To evaluate the stability of Integrated Gradients (IG) explanations, each image in the subsample was perturbed five times using Gaussian noise with varying standard deviations (σ = 0.01 to 0.1). For each perturbed image, a new IG heatmap was generated, and stability was quantified using the Structural Similarity Index (SSIM) between the original and perturbed explanations.

Results show that IG explanations are highly stable under input noise. At the lowest noise level (σ = 0.01), the average SSIM was 0.9417, indicating strong similarity. Even at the highest noise level (σ = 0.1), the SSIM remained relatively high at 0.6205, suggesting that IG consistently identifies core features despite perturbations.

IV. Conclusions

The study explored and compared explanations generated by three different methods, representing two categories of interpretability techniques. Several key observations emerged from this analysis.

First, SENN did not produce highly interpretable results. The concepts it generated were not semantically meaningful, suggesting that the model likely does not rely heavily on these learned concepts for its predictions. This conclusion is further supported by the results of the faithfulness evaluation. However, the robustness regularization applied during training contributed to relatively stable explanations.

In contrast, the two post-hoc methods exhibited different strengths and weaknesses. LIME produced explanations that were interpretable and relatively faithful but proved to have high instability. Integrated Gradients, on the other hand, generated less intelligible explanations due to the low-level, pixel-wise nature of the attributions. Despite this, it performed well in terms of both faithfulness and robustness, providing consistent and reliable explanations across perturbed inputs

These results illustrate a clear trade-off between the different strategies, underscoring the value of employing multiple interpretability methods to achieve a more comprehensive understanding. However, it's important to note that the computed metrics are based on a limited sample of only 100 images, and are therefore not statistically significant.

References

[1] Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*. https://doi.org/10.48550/arxiv.1806.07538

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?" Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 1135–1144. https://doi.org/10.1145/2939672.2939778

[3] Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks*. Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 3319–3328. PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html

[4] Open Omar Elbaghdadi, Aman Hussain, Christoph Hoenes, and Ivan Bardarov. Self explainingm neural networks: A review. Technical report, University of Amsterdam, 2020. https://github.com/AmanDaVinci/SENN.git