Attribution Techniques in Deep Learning Models

Jayin Khanna
Research Internship under Dr. Niloy Ganguly
Indian Institute of Technology, Kharagpur

1 Problem Statement: Explainability in AI

In deep learning, while we often know the output of a model, we do not always understand why the model made a particular decision.

Which input features helped the neural network decide?

The goal is to explain a model's decision by assigning **importance scores** to input features—i.e., determining which parts of the input were responsible for a specific output.

Examples

- In vision: Which pixels made the model classify an image as a cat and not a dog?
- In sentiment analysis: Which words contributed most to a positive sentiment?
- In healthcare: Which symptoms or signals led to a particular diagnosis?

2 What Is a Good Attribution Technique?

Two important properties of a good attribution method:

1. Sensitivity Axiom

If two inputs x and x' differ in only one feature and the output $F(x) \neq F(x')$, then the differing feature must be assigned a non-zero attribution.

2. Implementation Invariance Axiom

If two networks compute the same function $F(x)$ for all inputs (i.e., are functionally equivalent), then the attribution method should yield identical attributions for both.

3 From Gradients to Integrated Gradients

For a linear model:

$$F(x) = w^\top x$$

The attribution of each input feature is simply its corresponding weight: w_i.
In complex models like neural networks, we generalize this using gradients:

$$\frac{\partial F(x)}{\partial x_i}$$

However, gradients are local and unstable. Hence, we use:

Integrated Gradients (IG) [?]

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a function representing the neural network, and let x be the input with a baseline x'. The integrated gradient along the i^{th} dimension is defined as:

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

Completeness Axiom

Integrated Gradients satisfy:

$$\sum_{i=1}^{n} \text{IG}_i(x) = F(x) - F(x')$$

This means the total attribution equals the difference between the model's prediction at input and baseline.

4 Path Methods

Integrated Gradients are a special case of **path methods**, which integrate gradients along any path γ from x' to x:

$$\text{Attribution}_i(x) = \int_0^1 \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \cdot \frac{d\gamma_i(\alpha)}{d\alpha} \, d\alpha$$

IG corresponds to the straight-line path: $\gamma(\alpha) = x' + \alpha(x - x')$.

5 Random Path Integrals (RPI) for LLMs

While Integrated Gradients work for vision and small models, they struggle in large models like LLMs. RPI addresses this with two changes:

1. **Attention Attribution**: RPI computes gradients over internal attention scores rather than just raw input embeddings.
2. **Random Baseline Sampling**: The baseline x' is not fixed, but drawn from a distribution B, allowing sampling of multiple integration paths.

RPI Methodology

Repeat for $r = 1$ to R:

- Sample random baseline $b_r \sim B$
- Compute attribution map m_r using IG along the path from b_r to x

This yields a set of R candidate attribution maps:

$$\{m_1, m_2, \ldots, m_R\}$$

Selecting the Best Attribution Map

Using a task-specific metric $\psi(\cdot)$ (e.g., **sufficiency** or **comprehensiveness**), choose:

$$m_{\text{RPI}} = \arg\max_{m_r} \psi(m_r)$$

This allows tailoring the explanation to the evaluation metric—something IG lacks.

6 Baselines and Distribution B

Sturmfels et al. (2020) explored many baseline representations, but found no clear winner. Hence, a natural choice is to let B be a mixture of distributions. RPI simplifies this by sampling $b \sim \mathcal{N}(\mu, \sigma^2 I)$, where the Gaussian is defined by a diffusion process as in [**?**].

7 Computational Complexity

- IG requires n interpolations per input.
- RPI requires $n \times R$ evaluations (for R baselines).

With a GPU that supports batch size $n \cdot R$, both methods have comparable runtime. However, RPI consistently outperforms IG across models, metrics, and datasets even with small R.

8 Conclusion

Attribution is a crucial step toward transparent, interpretable, and trustworthy AI. Techniques like Integrated Gradients and Random Path Integrals help demystify neural network decisions by identifying key input features.

RPI extends IG by leveraging attention scores and random baseline sampling, offering metric-adaptive explanations that are both accurate and theoretically grounded.

References

References

1. Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*. Available at: https://arxiv.org/abs/1703.01365
2. Sturmfels, P., Lundberg, S., Lee, S.-I. (2020). *Visualizing and Understanding Baselines for Attribution Methods*. Available at: https://arxiv.org/abs/2405.09800
3. Shapiro, A., et al. (2024). *Improving LLM Attributions with Randomized Path Integration*. Available at: https://cris.iucc.ac.il/en/publications/improving-llm-attributions-with-randomized-path-integration-2