Activation-Based Methods in Explainable AI (XAI)

Introduction

Activation-based methods in Explainable AI (XAI) are used to understand how a neural network model makes decisions by analyzing the activations of neurons in hidden layers. These methods focus on which parts of the input (features) are important based on the neuron activations.

Neuron Activation: Refers to the output of a neuron after applying an activation function (e.g., ReLU, GELU). It represents the "response" of the neuron to an input.

Logit: The raw output from the model before applying any activation function in the output layer (e.g., before softmax).

Key Definitions:

- **Activation:** The response value of a neuron after applying an activation function (ReLU, GELU, etc.).
- **Logit:** The raw score output before applying softmax or any other activation at the final layer.

Example of Difference Between Activation and Logit

When we say "activation" of a neuron, we are referring to the value after the activation function has been applied. The **logit value** is the raw score calculated from the weighted sum of inputs to that neuron.

Example:

- **Weighted sum (logit):** A neuron calculates a raw value based on weights and biases.
- **Activation:** After applying the activation function (e.g., ReLU, GELU), the weighted sum is transformed.

Answer to Question: Does a High Logit Always Mean High Activation?
- **No.** A neuron can have a high logit but a low activation, depending on the activation function.
- For example, with **ReLU**, if the logit is negative, the activation becomes zero (if using ReLU), so even a large negative value will not lead to a high activation.

Activation-Based Methods and Their Application in Transformers

In **transformers**, activation-based methods track how neurons or attention heads respond to different inputs. For example:
- A neuron with a negative value might have its activation squashed by ReLU, and this could cause an **activation-based method** to miss the importance of certain features if we focus purely on activations.
- **Class Activation Mapping (CAM)** and **Layer-wise Relevance Propagation (LRP)** help visualize which features (words in a sentence or parts of an image) the model relies on.

Key Points about Activation Functions:
1. **ReLU** activation may cause neurons with negative pre-activation values to become "dead," as their activations are set to zero.
2. **Leaky ReLU, GELU, and Tanh** allow negative values to persist, offering better insight into which features the model focuses on.

Can Negative Activations Be More Important Than Positive Ones?
Yes, a model can focus more on **negative values** within hidden layers, especially when those

negative activations **reverse** or **modify** the meaning of other activations. For instance, in sentiment analysis:

- A word like **"not"** can produce strong negative activations, making it more important for understanding negative sentiment.

Conclusion:

- In activation-based methods, it's essential to consider **both positive and negative activations** because negative values can hold critical information, especially when dealing with sentiment reversal or other key features.