**Assignment: Feature attribution methods and their evaluation**

**Due: Nov 21, 2024**

**Objective**The goal of this assignment is to explore how various feature attribution methods provide explanations for model predictions and to assess their effectiveness using quantitative metrics.

**Dataset**

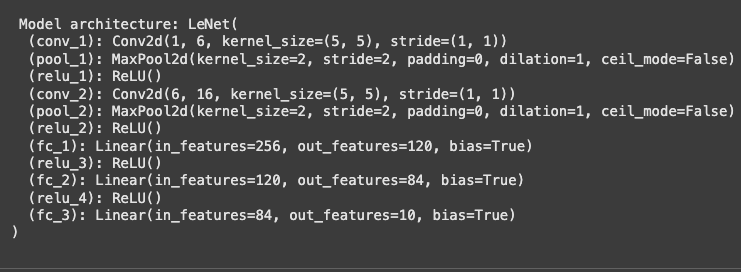
**Image Dataset:** Used the Fashion MNIST (FMNIST) dataset, which is similar to the MNIST dataset but with 10 target classes representing different fashion items.

**Model**

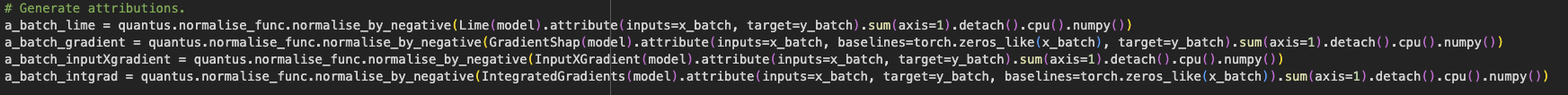
For FMNIST, used a straightforward LeNet model from https://github.com/ChawDoe/LeNet5-MNIST-PyTorch.

**Tasks**

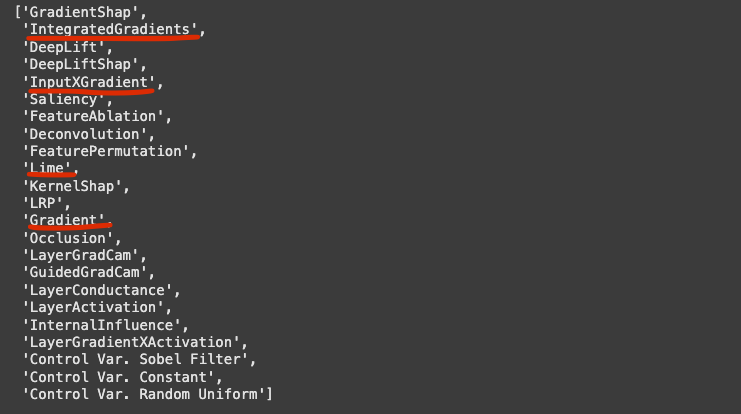
1. **Model Training:** Prepared and trained the model on the chosen dataset.



1. **Explainability Analysis:**
   1. Evaluated explanations generated by the following methods: LIME, Gradient, InputXGradient, and IntegratedGradients.
   2. Implemented these methods.

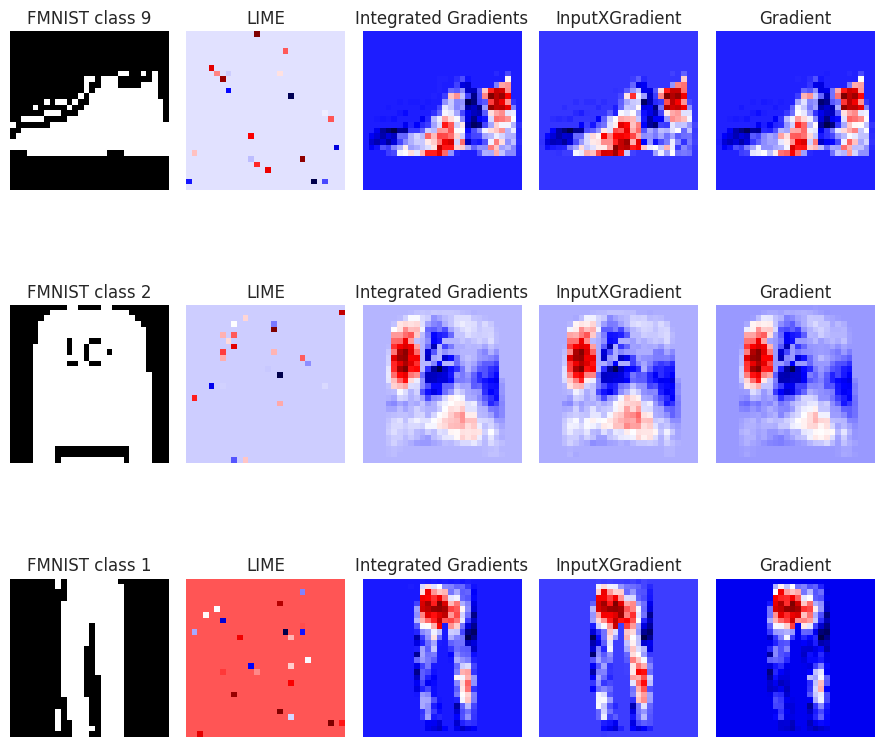


* 1. Confirmed the availability of these methods with the following command: quantus.AVAILABLE\_XAI\_METHODS\_CAPTUM.



* 1. **Qualitative Analysis**

Generated heatmaps for each explanation method and arrange them in columns to visually compare the differences.

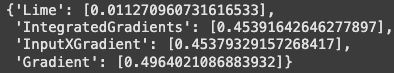


**Observations**

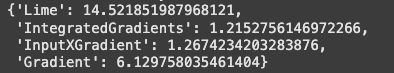
1. **LIME**:
   * LIME focuses on individual, sparse pixels rather than larger regions of interest.
   * It fails to highlight meaningful areas relevant to the model’s prediction, often selecting random-looking pixels.
   * This makes it ineffective for identifying global features or structural outlines. For instance, it doesn’t capture the contours of "ankle boots" or "trousers" but instead highlights scattered areas unrelated to the object’s overall shape.
2. **Integrated Gradients**:
   * This method highlights **broader, continuous regions**, capturing global dependencies and the overall shape of the object in the image.
   * For example, it smooths over large areas like the full trouser outline or the main body of a pullover.
   * It effectively balances relevance across the entire object, making it ideal for understanding the general importance of various parts of an image.
3. **InputXGradient**:
   * InputXGradient emphasizes **edges and fine details**, such as the heel of an "ankle boot" or the neckline of a "pullover."
   * It provides sharper, more localized heatmaps, which are useful for identifying specific regions critical to the prediction.
   * While it captures finer aspects better than Integrated Gradients, the added noise slightly diminishes its interpretability in some cases.
4. **Gradient**:
   * Gradients focus on **edge detection** and highlight similar regions as InputXGradient but with **higher noise** and less smoothness.
   * It captures outlines like the trouser shape or the boot’s heel, but the scattered nature of the heatmap reduces its reliability.
   * Compared to InputXGradient, its focus on details is less precise and harder to interpret due to the additional noise.

Integrated Gradients excels in highlighting global shapes and broader regions with smooth and comprehensive heatmaps that capture critical object features. InputXGradient focuses on finer details and edges, offering precise explanations but introducing slight noise. LIME, however, produces sparse and scattered heatmaps with disconnected pixels, failing to capture the overall shape or key features effectively. Gradient-based explanations emphasize edges but are noisier and less interpretable than Integrated Gradients and InputXGradient. Each method showcases unique aspects of the image—whether global structure, fine details, or sparse points—but their effectiveness depends on balancing detail, smoothness, and relevance.

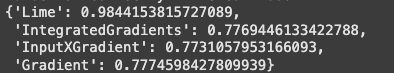
* 1. **Quantitative Analysis**
     1. For each explanation method, evaluated the following metrics:
        1. **Faithfulness Correlation:** Measures how well the explanation aligns with the model's predictions. Link: <https://quantus.readthedocs.io/en/latest/docs_api/quantus.metrics.faithfulness.faithfulness_correlation.html>



* + - 1. **Relative Input Stability:** Evaluates the robustness of explanations under slight input perturbations. The lower, the better. Link:  
         <https://quantus.readthedocs.io/en/latest/docs_api/quantus.metrics.robustness.html>



* + - 1. **Sparsity:** Assesses the simplicity of explanations by checking how many features contribute significantly. Link:   
         <https://quantus.readthedocs.io/en/latest/docs_api/quantus.metrics.complexity.sparseness.html>



* + 1. After computing these metrics, analyze the results. Consider the following questions:
       1. Which explanation method scores highest for faithfulness, stability, and sparsity?

**Faithfulness**: **IntegratedGradients** scores the highest with a value of **0.4539**, closely followed by **InputXGradient** (0.4538). Both methods demonstrate strong alignment between the model’s predictions and the explanation.

**Stability (Relativity)**: **Lime** exhibits the highest instability with a **Relativity** score of **14.52**, indicating the least stable explanation method. On the other hand, **IntegratedGradients** performs best in terms of stability, with a **Relativity** of **1.22**, followed by **InputXGradient** (1.27), suggesting that these methods provide more stable and consistent explanations.

**Sparsity**: **Lime** again scores the highest in sparsity with **0.9844**, indicating it produces more sparse explanations compared to the others. In contrast, **IntegratedGradients** achieves the lowest **Sparsity** score of **0.7769**, meaning it generates more focused and less scattered explanations.

* + - 1. Is any method consistently the best across all metrics, or do trade-offs exist?

No method is consistently the best across all metrics, indicating trade-offs:

* **IntegratedGradients** performs the best in **stability** and **sparsity**, while also showing strong **faithfulness** (0.4539). This makes it a reliable choice for balanced performance across metrics.
* **InputXGradient** excels in **faithfulness** (0.4538) but shows a slightly higher **Relativity** (1.27) than IntegratedGradients, indicating slightly less stability. It is a strong choice for highlighting finer details.
* **Gradient** performs well in **faithfulness** (0.4964), but suffers from lower **stability** (6.13) and **sparsity** (0.7775), suggesting that it produces noisier explanations and less focused attribution maps.
* **LIME** performs the worst across all metrics, with **poor faithfulness** (0.0113), the highest **instability** (14.52), and the most **sparse** explanations (0.9844), making it less effective overall.
  + - 1. Does the quantitative evaluation align with the qualitative observations?

Yes, the quantitative evaluation aligns with qualitative observations:

* **IntegratedGradients** is observed to generate smooth and comprehensive explanations, which is reflected in its high **stability** and **low sparsity** scores.
* **InputXGradient** is noted for capturing fine details, aligning with its high **faithfulness** score, though it does show slightly higher instability than IntegratedGradients.
* **Gradient**, although demonstrating high **faithfulness**, suffers from instability and sparse explanations, which matches its lower **stability** and **sparsity** scores.
* **LIME**, which often produces scattered and sparse explanations, aligns with its poor **faithfulness** and high **instability** scores.
  1. **Points to remember:**
     1. Always perform analyses on correctly classified samples to ensure meaningful and reliable insights.
     2. When calculating metrics, average the results over 500/1000 correctly classified samples to obtain robust quantitative evaluations.

**Submission**

Submit a report with your findings, including heat maps or tables, and a detailed analysis of each explanation method's effectiveness based on both qualitative and quantitative metrics.