CV Topic for Paper

Group members:

- Alam López
- Mario Rosas Otero
- Javier González Béjar
- Alberto Becerra Tomé

Topic

Explainability techniques for CNNs

Explainability techniques for Convolutional Neural Networks (CNNs) aim to provide insights into the model's decision-making process, helping users understand why the network produces a particular output. Some of these techniques are:

1. **Saliency Maps:**
    - Saliency maps highlight important regions in the input image that contribute most to the model's prediction. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) use gradient information to determine the importance of each pixel.
2. **Layer-wise Relevance Propagation (LRP):**
    - LRP assigns relevance scores to each neuron in the network, indicating their contribution to the final prediction. This backward-pass technique helps trace the impact of input features on the output.
3. **Feature Visualization:**
    - Feature visualization generates images that maximally activate specific neurons in the network. By visualizing what each neuron is responsive to, users can gain insights into the learned features.
4. **LIME (Local Interpretable Model-agnostic Explanations):**
    - LIME perturbs input data and observes the model's response, building a local interpretable model around the instance of interest. This provides insight into how the model behaves in the vicinity of a specific input.
5. **SHAP (SHapley Additive exPlanations):**
    - SHAP values provide a unified measure of feature importance. It is based on cooperative game theory and calculates the average contribution of each feature to all possible coalitions, offering a comprehensive view of feature impact.
6. **Attention Mechanisms:**

- Attention mechanisms, commonly used in natural language processing, can also be applied to CNNs. They highlight specific parts of the input image that the network focuses on during the decision-making process.

7. **Integrated Gradients:**
    - Integrated Gradients considers the integral of the gradients along the path from a baseline (e.g., an image with all pixel values set to zero) to the input image. This helps attribute the model's prediction to different features.
8. **Class Activation Maps (CAM):**
    - CAM highlights discriminative image regions by identifying the most informative parts for a particular class. It is often used for visualizing where the network looks to make its predictions.

These techniques offer different perspectives on understanding CNN decisions, providing users with insights into the model's internal workings and aiding in building trust and transparency in AI systems.

Not all of them will be covered in the presentation, but we will try to cover the most important ones.