Week 10 Thursday Paper 1: SHAP

Background: 9.5 Shapley Values

Explained by: 9.6 SHAP (SHapley Additive exPlanations)

1. What, intuitively, is a Shapley Value? What does it describe?
Ans: The Shapley Value is a concept from cooperative game theory that provides a fair distribution of payouts to players based on their contribution to the total payoff. Intuitively, it calculates the importance or contribution of each player or feature in the context of machine learning, by considering all possible combinations of players and how the total payoff changes when a particular player joins the game. The Shapley Value of a feature in a prediction model describes the average marginal contribution of that feature across all possible combinations of features, offering insight into how much each feature contributes to the prediction.

2. Explain the Additive Feature Attribution Method. Why is it useful? How is LIME used?
Ans: The Additive Feature Attribution Method refers to a class of methods where the explanation model is a linear function of binary variables. This class unifies several interpretation methods, including LIME. These methods attribute an effect to each feature, and the sum of these effects approximates the output of the original model. LIME specifically approximates the model locally around a given prediction, using a local linear model to understand how the features affect the prediction. It is useful because it provides interpretability to complex models by offering a simple, understandable approximation of how each feature contributes to a specific prediction

3. What are the three special properties of SHAP? Explain briefly why each is important.
Ans:
Local Accuracy: Ensures the explanation model's output matches the original model's output for the simplified input. This property guarantees that the explanations are accurate on a local level, reflecting the true output of the model for the given prediction.

Missingness: If a feature is missing in the simplified input, it should have no attributed impact. This property ensures that only the features present in the input contribute to the explanation, making the attributions more relevant and accurate.

Consistency: If the contribution of a feature to the prediction increases or stays the same, regardless of other features, then its attribution should not decrease. This property ensures fairness and reliability in the attribution process, meaning that as the importance of a feature increases, its attributed importance should not decrease

4. What is the SHAP value? How is it different from a Shapley Value?
Ans: SHAP values are a unified measure of feature importance derived from Shapley values, specifically tailored for model interpretation. SHAP values are the Shapley values of a conditional expectation function of the original model, making them directly applicable to explaining the output of machine learning models. While Shapley values originate from cooperative game theory and describe the fair distribution of total gains among players, SHAP values adapt this concept to the machine learning context, attributing the change in the expected model prediction to each feature. Thus, while Shapley values are a theoretical concept for fair distribution, SHAP values apply this concept to model predictions, offering a way to quantify the contribution of each feature to the prediction

5. Pick one variation of SHAP discussed in the paper and explain what it is and how it is useful compared to the other variations.
Ans: One variation of SHAP discussed is the Kernel SHAP. This method combines ideas from SHAP values and LIME to provide a model-agnostic approximation of SHAP values using weighted linear regression. Kernel SHAP addresses some of the limitations of direct Shapley value computation and LIME by offering a more efficient and theoretically grounded approach to estimating feature importance. It specifies a particular form of the loss function, weighting kernel, and regularization term that aligns with the properties of SHAP values. This method is particularly useful because it requires fewer evaluations of the original model to achieve similar approximation accuracy to direct Shapley value computation, making it more practical for complex models. Compared to other variations, Kernel SHAP offers a balance between computational efficiency and adherence to the desirable properties of SHAP values, making it an attractive choice for feature importance estimation

Week 10 Thursday Paper 2: Grad-CAM

Explained by: Grad-CAM: Visual Explanations from Deep Networks – Glass Box

1. Describe the procedures to obtain the neuron importance weights, and explain what the weights suggest about the image tasks.
Ans: To compute the class-discriminative localization map for a class, Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN. The gradients of the score for the class of interest (before the softmax) with respect to the feature map activations of a convolutional layer are computed. These gradients are then globally averaged over the width and height dimensions to obtain the neuron importance weights. This process captures the "importance" of each feature map for the target class. The weights indicate how much each neuron contributes to the decision of interest, representing a partial linearization of the deep network downstream from the convolutional layer and highlighting which features are important for predicting the class

2. What is *Guided Grad-CAM*? What problem about Grad-CAM does it address?
Ans: Guided Grad-CAM combines the techniques of Grad-CAM and Guided Backpropagation to provide high-resolution, class-discriminative visualizations. While Grad-CAM is able to localize relevant regions within the image, it may not offer fine-grained details. By fusing Grad-CAM with pixel-space gradient visualizations like Guided Backpropagation, Guided Grad-CAM achieves both high-resolution and class-discriminative visualizations. This technique highlights important regions in the image with fine-grained details specific to the class of interest, improving interpretability

3. What are the three evaluation experiments performed by the paper? What do they suggest about Grad-CAM?
Ans:
Weakly-Supervised Localization: Grad-CAM was evaluated on its ability to localize objects in images using off-the-shelf pretrained networks like VGG-16, AlexNet, and GoogleNet. It was found that Grad-CAM significantly outperformed other methods in localization errors without compromising classification performance.

Weakly-Supervised Segmentation: Applying Grad-CAM as a seed for segmentation tasks led to improved Intersection over Union (IoU) scores compared to using CAM maps, demonstrating Grad-CAM's effectiveness in providing better localization cues.

Pointing Game: This experiment evaluated the discriminativeness of visualization methods for localizing target objects. Grad-CAM outperformed c-MWP(contrastive Marginal Winning Probability) by a significant margin, highlighting its precision and recall capabilities in localization.

These experiments suggest that Grad-CAM provides accurate and class-discriminative localizations, significantly improving upon previous methods in terms of both interpretability and effectiveness in localization tasks.

4. What does the paper say about Grad-CAM's capabilities on multi-modal tasks, for example, image captioning?
Ans: Grad-CAM was applied to multi-modal tasks, such as image captioning and Visual Question Answering(VQA), and it produced interpretable visual explanations. For image captioning, despite models not being trained on grounded image-text pairs, Grad-CAM was able to localize discriminative image regions effectively. This demonstrates Grad-CAM's broad applicability and effectiveness in providing insights into how CNN-based models make decisions, even in complex tasks involving both visual and textual components

5. Can Grad-CAM be applied to attention-based networks? If so, what changes or modifications might need to be made?
Ans: The paper suggests that Grad-CAM can be applied to a wide variety of CNN architectures, including those used for structured outputs, multi-modal inputs, and reinforcement learning, without needing architectural changes or re-training. This implies that Grad-CAM can also be applied to attention-based networks by analyzing the feature maps and gradients in layers prior to the attention mechanisms. The generalization capability of Grad-CAM, as it applies to different network architectures and tasks, underscores its versatility in providing visual explanations across different domains.

Depending on the specific type of attention-based network, different additional modifications might be needed. For instance, in Vision Transformers (ViTs), where the input image is divided into patches and processed as a sequence, it might be more straightforward to apply a method similar to Grad-CAM by interpreting the attention weights towards the classification token as indicative of the importance of each patch.