Comparative Analysis of CAM and Grad-CAM for Visual Explanation on the Stanford Dogs Dataset using ResNet50

Jiwoong Yang
MODULABS

June 4, 2025

Abstract

The increasing complexity of deep neural networks, particularly in image classification tasks, has amplified the demand for robust interpretability methods. Understanding why a model makes a certain prediction is crucial for building trust and facilitating a deeper analysis of model behavior. This paper presents a comparative study of two prominent visualization-based explainability techniques: Class Activation Mapping (CAM) [1] and Gradient-weighted Class Activation Mapping (Grad-CAM) [2]. We implement and evaluate these methods on the Stanford Dogs dataset [3] using a ResNet50 [4] architecture fine-tuned for dog breed classification. Our study qualitatively analyzes the generated heatmaps, observing that CAM tends to highlight broader, more diffuse object regions, while Grad-CAM, applied to various convolutional layers offers the potential for more localized, albeit sometimes varied, explanations. Quantitatively, we derive bounding boxes from these activation maps and evaluate their localization accuracy using the Intersection over Union (IoU) metric against ground truth bounding boxes. This work contributes to a practical understanding of CAM and Grad-CAM, highlighting their distinct characteristics and performance nuances in a specific fine-grained visual categorization task, thereby informing the selection of appropriate XAI tools for model diagnostics.

1 Introduction

Convolutional Neural Networks (CNNs) have become pivotal in advancing computer vision, yet their "black box" nature often obscures their decision-making processes, necessitating Explainable AI (XAI) for trustworthy and analyzable models. Visual explanations, which identify influential image regions, are particularly crucial. Among these, Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM) are prominent techniques for generating heatmaps that localize class-discriminative areas. Grad-CAM, notably, offers broader applicability by not requiring specific architectural modifications, unlike CAM.

This paper conducts a practical implementation and comparative analysis of CAM and Grad-CAM on the Stanford Dogs dataset using a fine-tuned ResNet50 architecture for dog breed classification. We investigate the qualitative distinctions in their generated heatmaps—applying Grad-CAM to various convolutional layers and quantitatively assess their object localization capabilities via Intersection over Union (IoU) scores derived from heatmap-generated bounding boxes. Our primary contributions involve a direct comparison of these XAI methods in a fine-grained visual categorization context, an analysis of Grad-CAM's layer-specific explanations, and an investigation into notable empirical observations from our study. These observations include an unexpected lower IoU performance for Grad-CAM compared to CAM under our specific experimental conditions, and a tendency for CAM to emphasize

object edges in blended visualizations. This study aims to offer a nuanced understanding of CAM and Grad-CAM, highlighting their distinct characteristics and practical implications for model diagnostics.

The remainder of this paper is organized as follows: Section 2 reviews related XAI and visual explanation methods. Section 3 details our methodology, including the dataset, model, and implementation of CAM and Grad-CAM. Section 4 presents the experimental setup, results, and discussion. Section 5 concludes with key findings and future research directions.

2 Related Work

The rapid advancement of Convolutional Neural Networks (CNNs) has led to significant breakthroughs in various computer vision tasks, particularly in image classification. Despite their impressive performance, the "black-box" nature of many deep learning models has spurred the development of Explainable AI (XAI), which seeks to render model decisions more transparent and interpretable. Within XAI, visual explanation methods are crucial for image-based tasks, providing insights by highlighting image regions that are salient to a model's output.

A variety of visual explanation techniques have been proposed. Some methods, like sensitivity analysis or deconvolution-based approaches, attempt to visualize features learned by neurons or to map activations back to the input space. Other approaches, such as LIME (Local Interpretable Model-agnostic Explanations) [5] and SHAP (SHapley Additive exPlanations) [6], offer *model-agnostic* explanations by perturbing inputs or analyzing feature contributions. However, heatmap-based methods that directly visualize class-discriminative regions have become particularly popular for their intuitive appeal in CNNs.

Class Activation Mapping (CAM) was a seminal technique in this domain. CAM produces a class-specific heatmap by computing a weighted sum of the feature maps from the last convolutional layer, just before the global average pooling (GAP) layer. The weights correspond to those of the final classification layer connected to the GAP output. While effective, a primary limitation of CAM is its requirement for a specific network architecture that incorporates a GAP layer followed by a linear classification layer, often necessitating model modification.

To address this limitation, Selvaraju et al. proposed Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is a generalization of CAM that does not require any changes to the underlying CNN architecture. It uses the gradients of the target class score with respect to the feature maps of any chosen convolutional layer. These gradients are global average pooled to obtain neuron importance weights, which are then combined with the forward activation maps to produce a coarse localization map highlighting important regions for the predicted class. This flexibility allows Grad-CAM to be applied to a wide variety of CNN models and tasks, including those with fully connected layers or structured outputs.

The evaluation of visual explanations itself is an active area of research, with metrics often focusing on faithfulness (how accurately the explanation reflects the model's process) and interpretability (how well humans can understand the explanation). For tasks involving object localization, metrics such as pointing game accuracy or Intersection over Union (IoU) with ground-truth bounding boxes are sometimes used to assess the spatial fidelity of heatmaps, even if the models were not explicitly trained for localization. Our work builds upon these foundational heatmap techniques by providing a focused, comparative study of CAM and Grad-CAM. We specifically investigate their performance nuances on a fine-grained visual categorization task, the Stanford Dogs dataset, and analyze empirical observations such as differences in localization accuracy (IoU) and visual characteristics of the generated heatmaps from different layers in Grad-CAM.

3 Methodology

This section details the experimental methodology employed in our comparative study of Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM). We describe the dataset utilized, the data preprocessing steps, the ar-

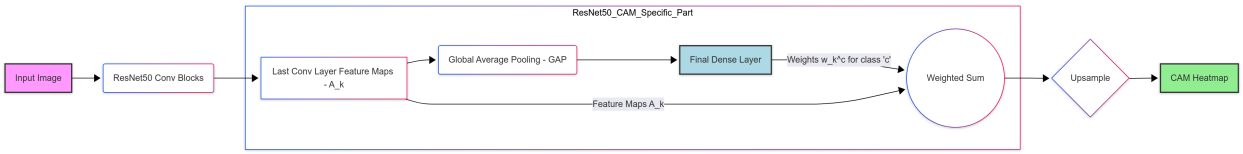


Figure 1: Schematic overview of the Class Activation Mapping (CAM) generation process within the ResNet50 architecture. It highlights the use of feature maps from the final convolutional layer, global average pooling, and weights from the subsequent dense layer to compute the heatmap.

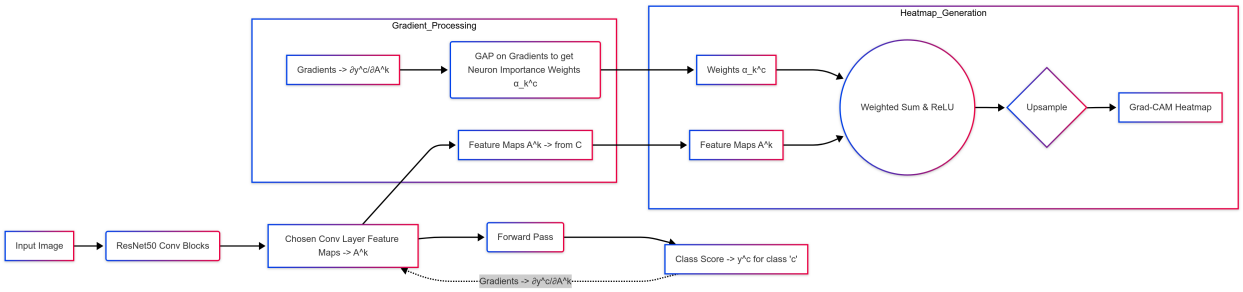


Figure 2: Illustration of the Gradient-weighted Class Activation Mapping (Grad-CAM) process. This diagram shows how gradients from the class score are backpropagated to a chosen convolutional layer to compute neuron importance weights, which then weight the feature maps to produce the final heatmap. This process can be applied to various layers in the network.

chitecture of the base classification model, the specific implementation approaches for CAM and Grad-CAM, the procedure for generating bounding boxes from heatmaps, and the metric used for quantitative evaluation.

3.1 Dataset

For this study, we utilized the Stanford Dogs dataset. This dataset is a challenging fine-grained visual categorization resource, consisting of 20,580 images spanning 120 different breeds of dogs. Each breed is represented by approximately 150-200 images. The dataset provides ground-truth bounding boxes annotating the location of the dog in each image, which are used in our study for the quantitative evaluation of heatmap localization. We used the official train/test splits provided with the dataset.

3.2 Data Preprocessing

All images from the Stanford Dogs dataset were preprocessed before being fed into the neural network. Initially, each image was resized to 224x224 pixels to match the input requirements of the ResNet50 architecture. The pixel values of the images were then normalized to a range of [0, 1] by dividing by 255. This normalization step is standard practice and helps in stabilizing the learning process.

3.3 Base Model Architecture

The base model for our classification task and subsequent XAI analysis is ResNet50. We utilized a ResNet50 model pre-trained on the ImageNet dataset. The original top classification layer of the ResNet50 model was removed. In its place, a Global Average Pooling (GAP) layer was applied to the output of the last convolutional block, followed by a new dense classification layer with a softmax activation function. This final dense layer was configured with 120 output units, corresponding to the number of dog breeds in the Stanford Dogs dataset. The model was then fine-tuned on the Stanford Dogs training set for the dog breed classification task.

3.4 Class Activation Mapping (CAM)

CAM provides a way to visualize the regions of an image that are important for a particular class prediction, but it requires a specific network architecture, typically one ending with a GAP layer followed by a dense layer, as employed in our base model. To generate the CAM for a given image and target class, we first extract the feature maps from the convolutional layer immediately preceding the GAP layer. The weights connecting the GAP layer's output to the target class node in the final dense layer are then retrieved. The CAM is computed as a weighted sum of the extracted feature maps, where each feature map is weighted by the corresponding class-specific weight (see Figure 1). The resulting heatmap is then upsampled to the original input image resolution using bilinear interpolation to visualize the class-discriminative regions.

3.5 Gradient-weighted Class Activation Mappping (Grad-CAM)

Grad-CAM offers a more general approach for producing visual explanations and does not require any specific model architecture. For a given image, target class, and a chosen convolutional layer, Grad-CAM computes the gradient of the score for the target class with respect to the feature maps of that chosen layer. These gradients are then global average pooled channel-wise to obtain the neuron importance weights for each feature map. The final Grad-CAM heatmap is a weighted linear combination of the forward-pass feature maps, followed by a ReLU activation to retain only features that have a positive influence on the class of interest (see Figure 2). In this study, to analyze how the choice of layer affects the resulting visualizations, we generated Grad-CAM heatmaps by targeting three distinct convolutional layers at varying depths within the ResNet50 architecture. These layers are specifically the final output of the third, fourth, and fifth residual blocks (denoted as conv3_block3_out, conv4_block3_out, and conv5_block3_out, respectively, following common ResNet naming conventions). These layers represent feature hierarchies from intermediate to deeper stages

of the network. The generated heatmaps were also upsampled to the original input image resolution using bilinear interpolation.

3.6 Bounding Box Generation from Heatmaps

To quantitatively evaluate the localization ability of CAM and Grad-CAM, we converted the generated heatmaps into bounding boxes. This process involved several steps. First, the normalized heatmap (with values between 0 and 1) was thresholded at a predefined value (e.g., 0.05 in our experiments) to create a binary mask, isolating the most salient regions. Any pixel values below this threshold were set to zero. Contours were then identified in the binarized heatmap. Among these, the contour enclosing the largest area was selected, assuming it corresponds to the main object of interest. Finally, a minimum area rotated rectangle was fitted to this largest contour, and its axis-aligned bounding box coordinates were extracted to serve as the predicted bounding box.

3.7 Evaluation Metric

The primary metric used to quantitatively assess the localization performance of the bounding boxes derived from CAM and Grad-CAM heatmaps was Intersection over Union (IoU). IoU is a standard metric for comparing the similarity between two arbitrary shapes, in this case, the predicted bounding box(B_p) and the ground-truth bounding box(B_{gt}) provided by the dataset. It is calculated as the ratio of the area of overlap between the two boxes to the area of their union.

$$\text{IoU} = \frac{\text{Area}(B_p \cap B_{gt})}{\text{Area}(B_p \cup B_{gt})}$$

An IoU score ranges from 0 (no overlap) to 1 (perfect overlap). Higher IoU scores indicate better localization of the object by the heatmap.

4 Experiments and Results

This section outlines the experimental setup for evaluating CAM and Grad-CAM, presents the qualitative and quantitative results obtained, and provides a discussion of these findings. Our experiments were designed to compare the visual explanations generated by CAM and Grad-CAM and to assess their ability to localize objects of interest within the Stanford Dogs dataset.

4.1 Experimental Setup

4.1.1 Implementation Details

The models and algorithms were implemented using Python with the TensorFlow and Keras libraries. The base ResNet50 model, pre-trained on ImageNet, was fine-tuned on the Stanford Dogs dataset for 4 epochs using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 and a batch size of 16. Model checkpoints were saved based on the best validation loss observed during training. For Grad-CAM, visualizations were generated from three distinct convolutional layers: conv3_block3_out, conv4_block3_out, and conv5_block3_out, representing varying depths within the network, as detailed in Section 3.5. The threshold for binarizing heatmaps to generate bounding boxes (as described in Section 3.6) was set to 0.05.

4.1.2 Evaluation Protocaol

or qualitative analysis, we visually inspected the heatmaps generated by CAM and Grad-CAM overlaid on original images, focusing on the extent and specificity of the highlighted regions. For quantitative analysis, we used the Intersection over Union (IoU) metric, as defined in Section 3.7, to compare the bounding boxes derived from the heatmaps against the ground-truth bounding boxes provided with the Stanford Dogs dataset. This evaluation was performed on the test split of the dataset.

Table 1: Experimental Setup and Hyperparameters

Parameter	Value
Base Model	ResNet50 (pre-trained on ImageNet)
Dataset	Stanford Dogs
Optimizer	SGD
Learning Rate	0.01
Batch Size	16
Epochs (Fine-tuning)	4
Grad-CAM Layers	conv3_block3_out, conv4_block3_out, conv5_block3_out
Heatmap Threshold	0.05
Evaluation Metric	IoU

4.2 Qualitative Results

In this subsection, we present representative visual examples of CAM and Grad-CAM heatmaps overlaid on images from the Stanford Dogs test set.

Overall, our qualitative observations indicate that CAM generally highlights larger, more encompassing regions of the target object. In contrast, Grad-CAM's visualizations vary significantly with the choice of layer. Deeper layers (e.g., conv5_block3_out) tend to produce more semantically focused heatmaps, pinpointing discriminative features of the dog breed. Intermediate layers (e.g., conv4_block3_out) offer a balance, while earlier layers (e.g., conv3_block3_out) often highlight more generalized, lower-level features or textures across the object. We also observed instances where CAM visualizations appeared to emphasize the edges or contours of the object more strongly than the object's core.

4.3 Quantitative Results

To illustrate the localization performance of CAM and Grad-CAM on a concrete example, we present the Intersection over Union (IoU) scores for the bounding boxes derived from the heatmaps generated for the representative image of a [Spaniel] shown in Figure 3 and Figure 4. The IoU scores for this specific instance are summarized in Table 2

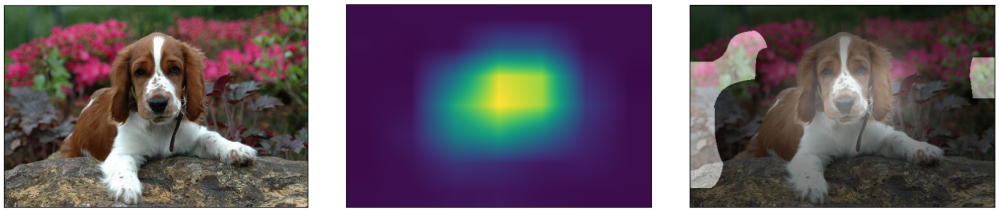
For this particular example image, CAM achieved an IoU score of 0.67. The Grad-CAM variations yielded scores of 0.57 for conv3_block3_out, 0.58 for conv4_block3_out, and 0.56 for conv5_block3_out. In this specific instance, CAM's heatmap resulted in a bounding box with a higher IoU score compared to those derived from the Grad-CAM heatmaps. It is important to note that these scores reflect performance on a single illustrative example and are not indicative of average performance across the entire dataset.

Method	Average IoU Score
CAM	0.67
Grad-CAM (conv3_block3_out)	0.57
Grad-CAM (conv4_block3_out)	0.58
Grad-CAM (conv5_block3_out)	0.56

Table 2: This Table presents the Intersection over Union (IoU) scores for CAM and Grad-CAM (applied to different layers) on the specific example image of a [Spaniel] shown in Figure 3. The IoU measures the overlap between the bounding box derived from the heatmap and the ground-truth bounding box for this single instance.

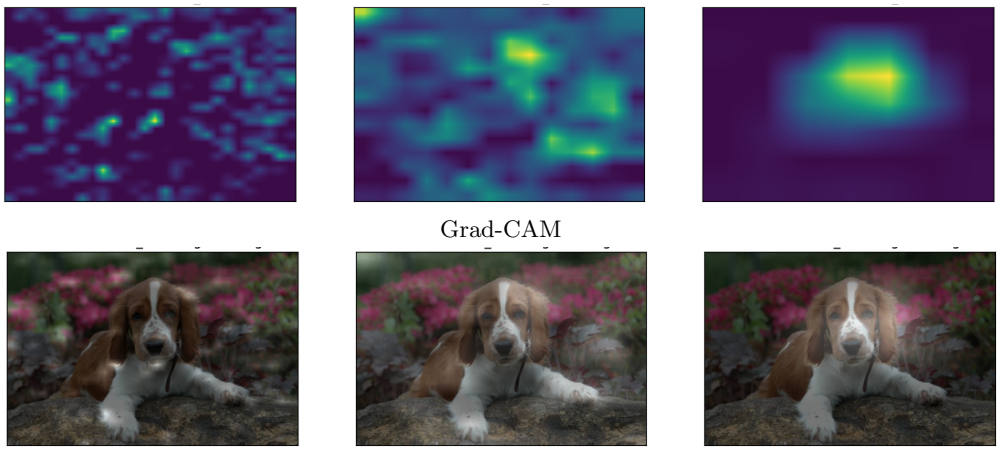
4.4 Discussion

The findings from our qualitative analysis and the illustrative quantitative IoU scores for a representative example image (as presented in Sections 4.2 and 4.3) offer insights into the application of CAM and Grad-CAM for understanding model decisions on the Stanford Dogs dataset. It is important to preface this



Original Image CAM CAM + Original Overlay

Figure 3: Class Activation Map (CAM) visualization for a representative image of a [Spaniel]. The heatmap highlights the image regions identified by CAM as most influential for the model's classification of this breed. CAM typically localizes discriminative object parts, often covering broader areas of the subject.



Grad-CAM

Grad-CAM + Original Overlay

Figure 4: Gradient-weighted Class Activation Map (Grad-CAM) visualizations for the same image of a [Spaniel] shown in Figure 3. Heatmaps are generated from the convolutional layers conv3_block3_out, conv4_block3_out, and conv5_block3_out in order. This illustrates how the choice of layer impacts the visual explanation, with deeper layers generally focusing on more semantically specific regions.

discussion by emphasizing that the quantitative IoU results pertain to a single image instance and thus serve as an illustrative case study rather than a basis for generalizable performance claims.

Qualitatively, for the selected example image (Figure 3 and 4), CAM produced a heatmap that broadly covered the dog, highlighting a significant portion of its silhouette. Grad-CAM, when applied to different layers for the same image, exhibited varied localization patterns. The heatmap from conv5_block3_out appeared to focus on more specific, potentially discriminative, parts of the dog, while conv3_block3_out yielded a more diffuse heatmap, and conv4_block3_out offered an intermediate level of detail and coverage. The tendency for CAM to sometimes highlight object contours was also observable in certain qualitative examples.

For this specific illustrative image, the IoU scores (Table 2) were 0.67 for CAM, 0.57 for Grad-CAM from conv3_block3_out, 0.58 from conv4_block3_out, and 0.56 from conv5_block3_out. In this instance, CAM resulted in the highest IoU score. This could be attributed to CAM's broader heatmap aligning well with the full extent of the dog in this particular image, which might have been more congruent with the ground-truth bounding box than the more focused or fragmented heatmaps from Grad-CAM. The bounding box generation method, which favors larger contiguous regions, might also contribute to this outcome for CAM in this example.

Regarding the Grad-CAM variations for this single image, the slightly higher IoU for conv4_block3_out (0.58) compared to conv5_block3_out (0.56) might suggest that for this specific image and its ground-truth annotation, an overly granular focus (as potentially provided by conv5_block3_out) on a small discriminative part did not translate to a better bounding box covering the entire object. The conv5_block3_out layer, while perhaps identifying key semantic features, might have produced a heatmap leading to a bounding box that was too tight. The conv3_block3_out layer's heatmap, likely capturing more general features, also resulted in a lower IoU, possibly due to its lack of specificity. Thus, for this particular image, conv4_block3_out may have offered a more effective balance in terms of feature abstraction for the IoU-based localization task.

However, it is crucial to reiterate that these quantitative observations are derived from a single image. While they serve to illustrate the calculation of IoU and potential differences between the methods on an anecdotal basis, they cannot be generalized to conclude on the overall superiority of one method or layer over another across the entire dataset. A comprehensive quantitative evaluation averaging results over a large, diverse set of images from the test set would be necessary to draw statistically significant conclusions about the general localization performance of these XAI techniques. Such an evaluation was beyond the scope of the illustrative analysis presented here.

Despite this limitation, this case study highlights several important considerations. It demonstrates how different XAI methods and their configurations (like layer choice in Grad-CAM) can produce visually distinct explanations for the same image. It also underscores the challenge of aligning qualitative visual appeal with quantitative metrics like IoU, as a visually "good" or "focused" heatmap does not invariably lead to a higher IoU score, especially when ground-truth bounding boxes encompass the entire object. This suggests that the utility of an explanation can be context-dependent and tied to the specific aspect of model behavior one aims to understand or evaluate.

Future research should prioritize a large-scale quantitative comparison of CAM and Grad-CAM on this dataset, potentially exploring a wider range of layers for Grad-CAM and different hyperparameter settings for bounding box generation (e.g., heatmap thresholds). Comparing these methods across different datasets, model architectures, and with alternative or more sophisticated localization-oriented evaluation metrics would also be highly valuable. Investigating the impact of model training strategies on the characteristics of the generated explanations remains another important avenue for exploration.

5 Conclusion

This paper presented a comparative study of two prominent visual explanation techniques, Class Ac-

tivation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM), applied to a ResNet50 model fine-tuned for fine-grained image categorization on the Stanford Dogs dataset. We conducted both qualitative analysis of the generated heatmaps and an illustrative quantitative assessment of localization performance using Intersection over Union (IoU) for a representative image example.

Our qualitative findings revealed distinct characteristics for each method. CAM typically produced broader heatmaps that encompassed significant portions of the target object, sometimes highlighting its contours. In contrast, Grad-CAM's visualizations were highly dependent on the selected convolutional layer, with shallower layers (e.g., conv3_block3_out) often capturing more general, textural features, while deeper layers (e.g., conv5_block3_out) tended to focus on more specific, semantically discriminative regions. The intermediate layer (conv4_block3_out) often provided a balance between these extremes. For the single illustrative image example analyzed quantitatively, CAM yielded a higher IoU score (0.67) compared to the Grad-CAM variations applied to conv3_block3_out (0.57), conv4_block3_out (0.58), and conv5_block3_out (0.56), suggesting that for this particular instance, CAM's broader heatmap aligned better with the ground-truth bounding box.

The primary takeaway from this study is that different XAI methods, and their configurations, can offer varied insights and may excel in different aspects of explanation. The visual interpretability or semantic focus of a heatmap, particularly for a single instance, does not always directly correlate with its performance on a specific quantitative localization metric like IoU. This underscores the importance of carefully considering the specific goals of an XAI analysis and the nature of the explanations sought when choosing a method. The quantitative results presented were illustrative for a single case and highlight the complexity of evaluating these methods.

While this study provides useful initial observations, it is subject to limitations, primarily the restriction of quantitative IoU analysis to a single image example and the focus on one dataset and model architecture. Future research should prioritize a comprehensive quantitative evaluation across the entire test set to draw more generalizable conclusions about the localization performance of CAM and Grad-CAM. Further investigations could also explore these methods on diverse datasets and architectures, examine the impact of different bounding box generation techniques, and incorporate alternative or complementary evaluation metrics beyond IoU to assess the multifaceted nature of visual explanations.

References

[1] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[3] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Ad-*

vances in neural information processing systems, 30, 2017.