Visualization vs. Interpretability

3. Topics. CAM vs. Grad-CAM

June 11th, 2025

Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a popular **visual explanation technique** for interpreting the decisions of **Convolutional Neural Networks (CNNs)**, especially in **image classification and computer vision**tasks.

Q1. What does Grad-CAM do?

- Grad-CAM helps you visualize **which regions of an input image** were most influential in a CNN's decision for a particular class.
- It produces a **heatmap** over the image, highlighting the "important" areas the network used to make its prediction.

Example Use Case

Suppose a CNN predicts "dog" for an image. Grad-CAM can show you **which parts of the image (e.g., dog's face or tail)** led to that prediction.

Grad-CAM

Q2. How does Grad-CAM work?

1. **Forward Pass**: Run the input image through the CNN to get the prediction.

2. **Backward Pass**: Compute the gradient of the score for the target class (e.g., "cat") with respect to the **feature maps** in the last convolutional layer.

3. **Weight Computation**: These gradients are global-average-pooled to obtain **importance weights** for each feature map.

4. **Weighted Sum**: Multiply each feature map by its corresponding weight and sum them to get a **class-discriminative heatmap**.

5. **ReLU**: Apply ReLU to focus only on the features that positively influence the class of interest.

Q3. CAM vs. Grad-CAM

🔁 Summary Table: CAM vs Grad-CAM

Feature	CAM (Class Activation Mapping)	Grad-CAM (Gradient-weighted CAM
Introduced in	2016	2017
Requires model modification?	✅ Yes – requires a special architecture	❌ No – works with most CNNs as-is
Architecture required	Global Average Pooling (GAP) before softmax	Any CNN with convolutional layers
How it works	Uses the weights of the final FC layer over feature maps	Uses gradients of class score w.r.t. feature maps
Flexibility	Limited – only works with specific CNN architectures	Flexible – works with pre-trained networks like ResNet, VGG
Uses gradients?	❌ No	✅ Yes
Layer used	Last convolutional layer before GAP	Any convolutional layer (usually last)
Interpretability	Good (but less flexible)	Better and more general

Q4. Conceptual Differences

CAM

- Requires modifying the model so that the final feature maps go directly to a **Global Average Pooling** layer, then to softmax.
- CAM-friendly networks, only
- The class activation map is computed directly using the **learned weights** from the classification layer.
- **Analogy**: **CAM** is like using the model's built-in roadmap.

Grad-CAM

- Works **without modifying the architecture**.
- Uses the **gradient of the class score** with respect to the feature maps to compute importance.
- More **general-purpose**, and works with most modern CNN architectures out of the box.
- **Analogy: Grad-CAM** "If you want more of this class, where should you look in the image?" — and it answers based on gradients.

Q5. When to Use Which?

- Use **CAM** if you're designing a model from scratch and can control the architecture.
- Use **Grad-CAM** if you're working with existing pre-trained models like **ResNet**, **VGG**, or **Inception**, and need **explainability without retraining**.

Visualization vs. Interpretability

4. Topics. Salience Map vs. a Grad-CAM

Both are methods to **visualize which parts of an input image influence a model's prediction**, but they differ significantly in how they **compute** and **display** this influence.

June 15th, 2025

1. Salience Map (Gradient-based Saliency)

- **Definition**: Shows how sensitive the model's output is with respect to changes in each input pixel.

- **How it works**:

  - Computes the **gradient of the output** (e.g., **class score**) w.r.t. **input image**.

  - These gradients indicate which pixels would most affect the output if changed slightly.

- **Visualization**: Typically a grayscale or heatmap image highlighting the most influential pixels.

- **Interpretation**: Tells you "**which pixels in the input image** were most influential" for the decision.

🔥2. Grad-CAM (Gradient-weighted Class Activation Mapping)

- **Definition**: Produces a coarse localization map showing the **regions of the image** that were important for a particular class decision.

- **How it works**:

  - Computes the **gradient of the output** (e.g., **class score**) w.r.t. **feature maps of a convolutional layer**.

  - These gradients are used to **weight the** **importance of each channel** (each **feature/convolution**) in the feature map.

  - The weighted feature maps are then combined and upsampled to the input size.

Pros vs. Cons

Salience Map (Gradient-based Saliency)

Pros:

- **Pixel-level** precision.
- Fast to compute.

Cons:

- Often **noisy** and **hard to interpret visually**.
- Sensitive to small perturbations.

Grad-CAM (Gradient-weighted Class Activation Mapping)

Pros:

- More visually intuitive and less noisy.
- Highlights **regions** rather than individual pixels.
- Works well for CNNs.

Cons:

- **Less precise** than saliency maps (since it uses a **lower-resolution conv layer**).
- Depends on the choice of convolutional layer.

🧠 In Short:

Feature	Salience Map	Grad-CAM
Based on	Input gradients	Gradients w.r.t. convolutional features
Output	Pixel-level map	Coarse region-level heatmap
Interpretability	Low (noisy)	High (clear region emphasis)
Use Case	Sensitive pixel analysis	Object localization and explainability
Works well with	Any model	CNNs with conv layers