Visualizing Decision-Making in Semantic Segmentation Architectures Using Grad-CAM

Jong Hyun Park, Jung Hyuk Lee, and Jong Seok Lee

***Abstract*—This research illustrates the decision-making processes within semantic segmentation architectures like U-Net and SegFormer by using Grad-CAM's visualization capabilities. We examine two aspects: first, tracking the enhancement of model performance in detecting lines, edges, and features with increasing training iterations, and second, visualizing the inner processes of well-known segmentation models using Grad-CAM. Our observations underscore Grad-CAM's role in visualizing the decision-making process of deep learning architectures, enhancing our understanding of their inner mechanisms, and advocating for comprehensible AI systems.**

1. Introduction

Deep learning has revolutionized computer vision and semantic segmentation, enabling machines to understand complex visual scenes at a pixel-level resolution. While architectures like U-Net and SegFormer excel at this task, their decision-making processes often remain obscure in complexity. As these models find application in critical domains such as medical imaging and autonomous driving, understanding how they arrive at their predictions becomes essential.

One powerful tool in understanding the decision-making processes of deep learning architectures is Grad-CAM, which offers visual insights into the regions of an input image that contribute most to a model's prediction. Originally designed for image classification, Grad-CAM's applicability now extends to semantic segmentation, making it a valuable technique for unraveling the decision dynamics of architectures like U-Net and SegFormer.

This paper investigates the interplay between Grad-CAM and semantic segmentation, focusing on U-Net and SegFormer. Our two main goal is to visualize how class activation maps change as the degree of training of segmentation models increases and visualize how the self-attention and encoding mechanisms of Transformer based segmentation models operate through Grad-CAM.

2. Related Work

UNet. U-Net [1] is a special type of convolutional neural network designed for image segmentation tasks, like identifying objects within an image. It's named after its U-shaped architecture.

In U-Net, the network first learns to capture important features from the input image using convolution layers that reduce the image's size while increasing the number of features. This part is called the "encoder." Then, the network uses another set of transposed convolution layers to enlarge and refine the features, gradually creating a segmented image. This part is called the "decoder."

What makes U-Net special are the "skip connections" that link the encoder and decoder. These connections help the network combine both the big-picture features and the fine details, which improves its ability to accurately segment objects in the image.

SegFormer. Unlike traditional convolutional neural networks (CNNs), SegFormer [2] employs a transformer-based architecture for image segmentation. Transformers [3] excel at capturing long-range dependencies and contextual relationships in data, making them well-suited for tasks that require understanding the context of each element within a sequence. SegFormer adapts transformers to image segmentation by using their self-attention mechanisms to operate on image patches instead of text tokens.

Applying the self-attention mechanism on image patches allows the model to focus on relevant parts of the image when making segmentation decisions. This attention-based approach enables the network to effectively capture global context, making it capable of recognizing objects that span large areas in the image.

Many architectures that utilized the attention-mechanism for vision tasks such as the ViT [4] used position encoding to incorporate location information. However, because the resolution of positional encodings is fixed, the encodings had to be interpolated if test resolutions differed from train resolutions, causing accuracy drops. SegFormer proposed that a 3×3 convolution is sufficient to provide positional information, and introduced Mix-FFN which directly uses a 3×3 convolution in the feed-forward network. Comparing the inference results on Cityscapes with two different image resolutions, Transformer encoder with Mix-FFN outperformed the other encoder with positional encoding, proving that their method produces better and more robust results.

Grad-CAM. Grad-CAM [5], short for Gradient-weighted Class Activation Mapping, proposed a technique that generalizes CAM [6] for a wider variety of CNN-based architectures. Its primary purpose lies in uncovering the

decision-making processes of convolutional neural networks and contributes to interpretability and explainability within deep learning.

Grad-CAM achieves this through an integration of gradients, capturing the sensitivity of the model's output to features of the input image. By calculating these gradients and applying them to the feature maps generated in the network, Grad-CAM generates a "heat map." This heat map highlights the spatial locations in the input image that played a crucial role in the network's decision-making process.

3. Methodology

We trained our models on the IIIT Oxford Pet Dataset, which is a collection of labeled images that focuses on pets, primarily cats and dogs. Each image is manually labeled with pixel-level annotations to identify the boundaries of the pets, making it especially useful for segmentation tasks.

For classification tasks, to obtain the class-discriminative localization map, we need to compute the gradient of the score for class c, y^c before the softmax. However, outputs of architectures for semantic segmentation are pixel-level labels that assign each pixel in an input image to a specific class. Therefore, to observe the class activation map for a certain class, we need to inspect the model output, pick a pixel that has the highest score for the wanted class and use it to compute the gradient.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \tag{1}$$

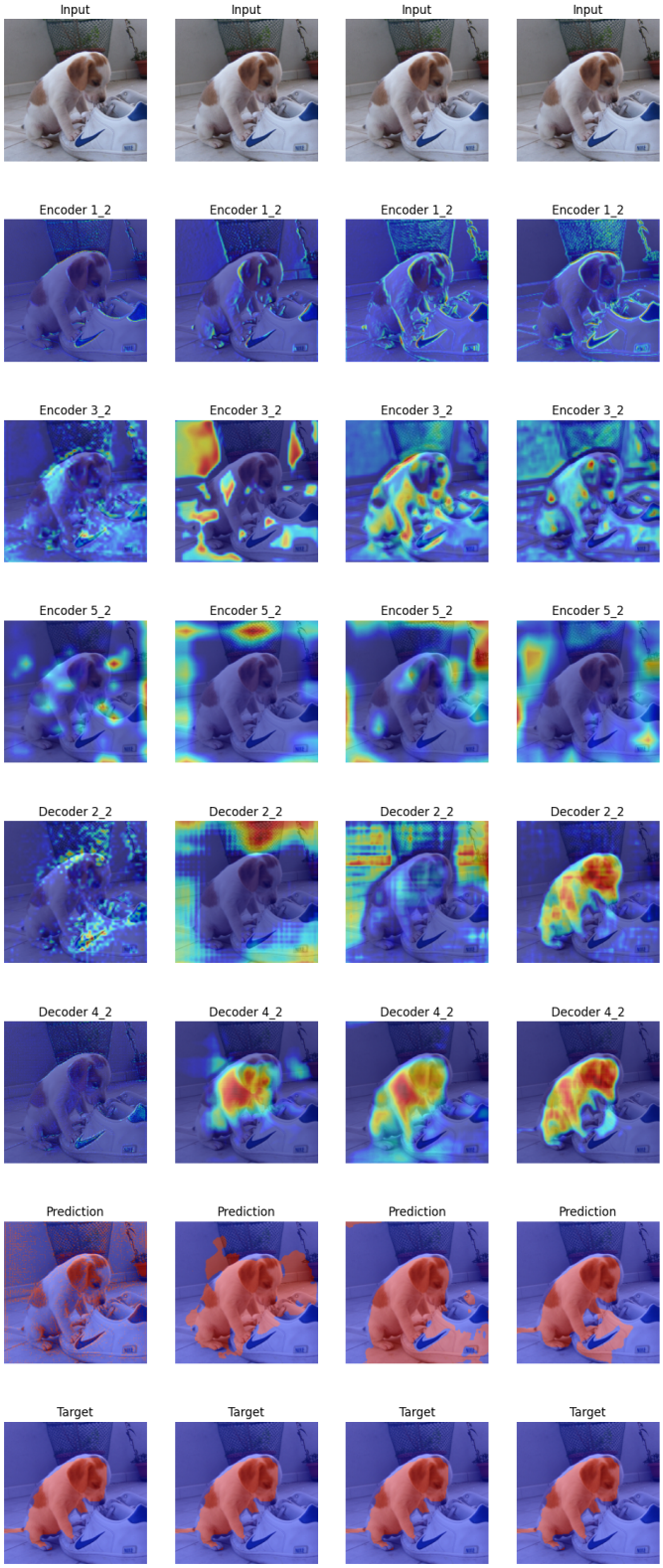
This weight α_k^c represents the importance of feature map k for a target class c and performing a multiplication to the forward activation maps followed by a ReLU, it produces a heatmap the size of the model output.

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{2}$$

Although Grad-CAM maps from earlier layers become worse and harder to understand what the model is paying attention to, it gives us insight to the different mechanisms used in various architectures. Therefore, we observed the class activation maps from various levels. If the heat map size of the layer's output differed from the input size, we bilinear upsampled the heat map to match the input image.

To track the enhancement of model performance as the training iterations increase, we compared the Grad-CAM maps produced by four UNet models, each trained to achieve mean IoU of 30%, 50%, 70%, 90% respectively.

To gain deeper insights into how the self-attention mechanism and Mix-FFN operates within the context of computer vision, we first trained the MiT-B0 SegFormer model provided by Hugging Face to achieve 90% mean IoU. Through a thorough analysis of the Grad-CAM maps within self-attention and Mix-FFN layers, we discerned notable similarities and differences by comparing these maps with those generated by UNet.



(a) mIoU: 30% (b) mIoU: 50% (c) mIoU: 70% (d) mIoU: 90%

Fig. 1: Grad-CAM maps of encoder, decoder convolution layers in UNet trained to achieve mean IoU of (a) 30%, (b) 50%, (c) 70%, (d) 90%

4. Results

Fig. 1 illustrates Grad-CAM maps for the target class ‘Pet’ from different layers of the UNet architectures, each trained to varying degrees. Grad-CAM maps produced by encoding layer 1_2 of UNet indicates that earlier convolution layers tend to detect lower-level features such as edges, corners, and basic texture. These layers capture simple patterns that are common across different types of images. Although UNet model with mean IoU of 30% is still able to detect these low-level features, it is evident that a model with higher mIoU creates a more detailed class activation map.

As the network's depth increases, deeper convolution layers progressively learn to detect more complex features, which are specific to the task at hand. These features might correspond to higher-level shapes, object parts, and even entire objects. Even so, if the model lacks enough training to learn these complex features, it is noticeable in Fig. 1 (a) that deeper convolution layers in undertrained models still focus on capturing lower-level features.

Another notable aspect is that although the Grad-CAM maps in Fig. 1 was created by the class activation maps for the target class ‘Pet’ itself, the maps from Encoder 5_2 in (b), (c), (d) and Decoder 2_2 in (b), (c) indicates that the model is focusing on the background rather than the object. This suggests that, for segmentation tasks, it is crucial not only to detect the target class itself but also to identify parts that belong to other classes.

In Fig. 2, the ‘Before Attention’ illustrates the Grad-CAM maps of the layer normalization layer preceding self-attention, while ‘After Attention’ presents maps of the dropout layer after self-attention. Unlike convolution layers, which operate on local receptive fields, the self-attention mechanism can capture global context by considering relationships between all positions in an input feature map. This leads to a broader understanding of the overall scene and enables the model to make more informed decisions.

Fig. 3 shows the Grad-CAM maps of the depth-wise convolution in Mix-FFN used to provide positional information to the features extracted by self-attention. It is observable that although depth-wise convolution is used for computational efficiencies instead of normal convolutions, it is still able to gradually capture lower-level features to higher-level features. However, it’s worth noting that because the feature maps created by self-attention is smaller in width and height than in the input image, although an earlier convolution layer still captures lower-level features, the thickness increases in contrast to earlier layers in UNet.

5. Conclusion

In conclusion, our investigation into the application of Grad-CAM to semantic segmentation architectures has given insights into the decision-making processes. We could visualize that when the model is trained more intensely, it becomes noticeably better at recognizing lines and objects in the images fed into the model. This link between how much the model is trained and its ability to detect specific patterns, edges or textures highlights how learning happens in these models and how they get better over time through repeated weight adjustments. Equally important is understanding the mechanism of self-attention and using convolution as a tool for positional encoding. With Grad-CAM, we were able to see the strategies they employ for context and feature extraction.

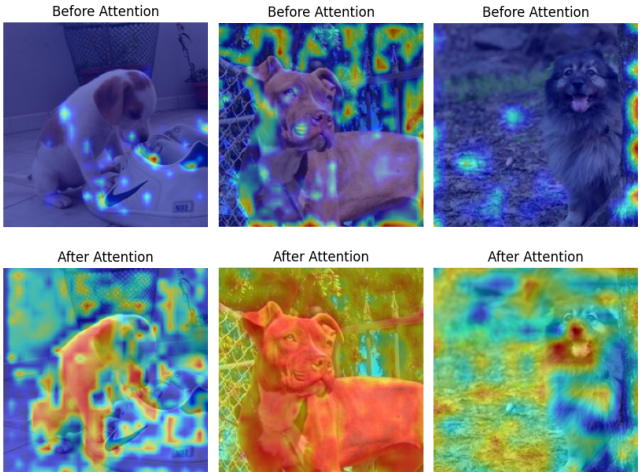


Fig. 2: Grad-CAM maps of layer normalization layers before self-attention, and dropout layer after self-attention.

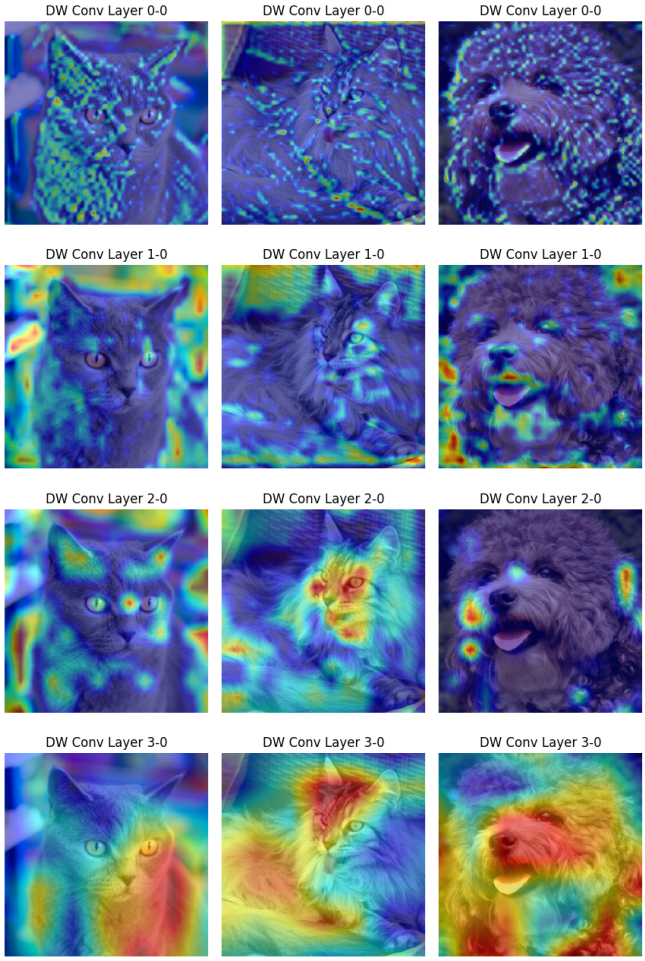


Fig. 3: Grad-CAM maps of the first depth-wise convolution layers in the four hierarchical encoder blocks of SegFormer.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI* 2015

[2] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS* 2021

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR* 2021

[5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *ICCV*, 2017.

[6] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.