Medical Image Segmentation Using U-Net with ResNet Backbone and Gated Attention

May 15, 2025

Project Members:

Devyansh Gupta
devyansh_g@mfs.iitr.ac.in
Roll No: 23125011

Adarsh Maurya
adarsh_m@mfs.iitr.ac.in
Roll No: 23125004

Neevadit Verma
neevadit_v@mfs.iitr.ac.in
Roll No: 23125023

Abstract

This project presents a robust medical image segmentation architecture that enhances the classic U-Net by integrating a ResNet backbone for the encoder and a gated attention mechanism to improve focus on the damaged tissue areas. ResNet provides deep feature representations through residual learning, enabling better generalization and context capture. The gated attention module highlights relevant spatial regions by filtering encoder features before merging with decoder layers. Experimental results on the ISIC 2018 skin lesion dataset demonstrate improved segmentation performance in terms of Dice and IoU scores compared to baseline U-Net models.

1. Main Objectives

- Improve segmentation accuracy in medical images using ResNet feature extraction.
- Integrate gated attention to refine skip connections between encoder and decoder.
- Evaluate performance on the ISIC 2018 skin lesion segmentation dataset.

2. Status and Other Details

- **Project Status:** Completed
- **Total time spent on the project:** 3 weeks
- **Evaluation Dataset:** ISIC 2018 Skin Lesion Segmentation Dataset

3. Major Stumbling Blocks

- **Incorporating ResNet Backbone**: Using a pre-trained backbone of ResNet-34 in U-Net's Encoder for feature extraction was a difficult task due to various downsampling and convolution present in both of these two different architectures.
- **High Computational Complexity:** Incorporating a ResNet backbone significantly increases the number of layers and parameters, which leads to higher memory usage and longer training times, especially on limited GPU resources.
- **Integration of Gated and Encoder Attention:** Implementing these attention mechanisms required careful tuning and architectural adjustments to ensure compatibility with the encoder-decoder structure of U-Net. It was challenging to balance the attention gating without disrupting spatial feature integrity.

4. Background and Motivation

U-Net has been widely adopted for semantic segmentation in the biomedical domain due to its symmetric encoder-decoder design with skip connections. However, its CNN-based encoder is often shallow and limited in representation power.

To address this:

- We replace the U-Net encoder with a pretrained ResNet backbone (ResNet-34).
- We introduce attention gates to refine encoder outputs before concatenation with decoder features.

5. Architecture Overview

5.1. U-Net Architecture

U-Net consists of a contracting (encoder) path to capture context and a symmetric expanding (decoder) path for precise localization. The skip connections help preserve spatial information lost during downsampling.

- The encoder applies multiple 3×3 convolutions (commonly referred to as Double Convolution), each followed by Batch Normalization and ReLU activation. These layers progressively extract hierarchical features while preserving spatial context, and are interleaved with 2×2 max-pooling operations to downsample the feature maps.

- The decoder path uses 2×2 transposed convolutions (also known as deconvolutions) for learned upsampling. Each upsampling stage is followed by concatenation with the corresponding encoder feature map (via skip connections), ensuring that spatial information lost during downsampling is effectively recovered.

- The final output layer applies a 1×1 convolution to reduce the number of channels to the number of segmentation classes, producing a pixel-wise classification map that assigns each pixel a label based on learned features.

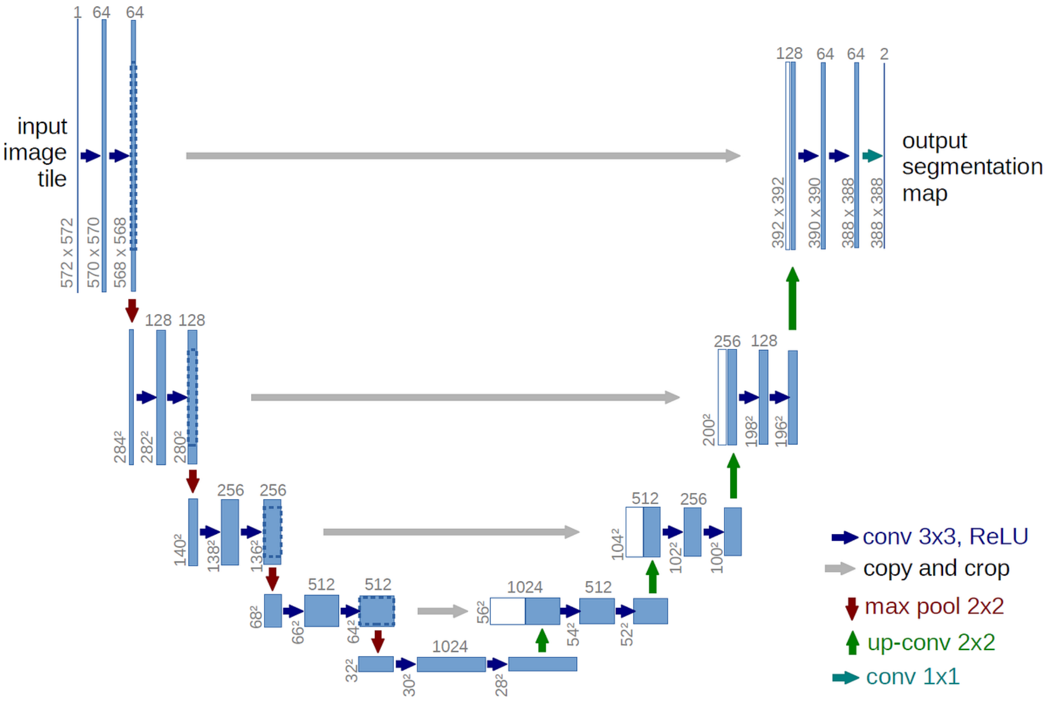

Figure 1: Original U-Net Architecture

5.2. ResNet Architecture (Encoder)

ResNet is a deep convolutional neural network that introduces residual connections to mitigate the vanishing gradient problem in deep networks. A residual block is defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, W) + \mathbf{x}$$

Where:

- $\mathcal{F}$ is typically a sequence of 3×3 Conv → BN → ReLU layers.
- The identity connection $\mathbf{x}$ allows gradient flow directly to earlier layers.

Architecture Components:

- **Conv1**: Initial 7×7 convolution + max pool.
- **Conv2_x to Conv5_x**: Each stage contains several residual blocks depending upon the variant of ResNet.
- **Classification Head**: At last the feature maps of final Convolution Layer are flattened and passed to a Multi Layer Perceptron (MLP) head to predict the desired class in the Image.

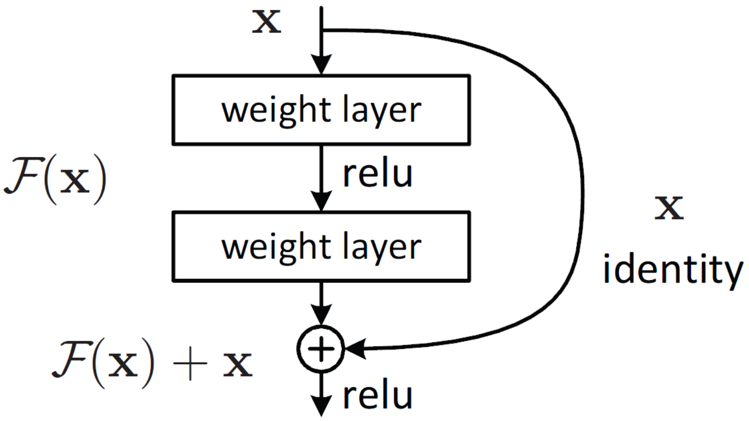
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2: Residual Block in ResNet

5.3. Integration with U-Net

We replace the U-Net encoder with a pretrained ResNet backbone (ResNet-34). The feature maps from intermediate residual stages (after downsampling) are passed as skip connections to the decoder, preserving high-level context.

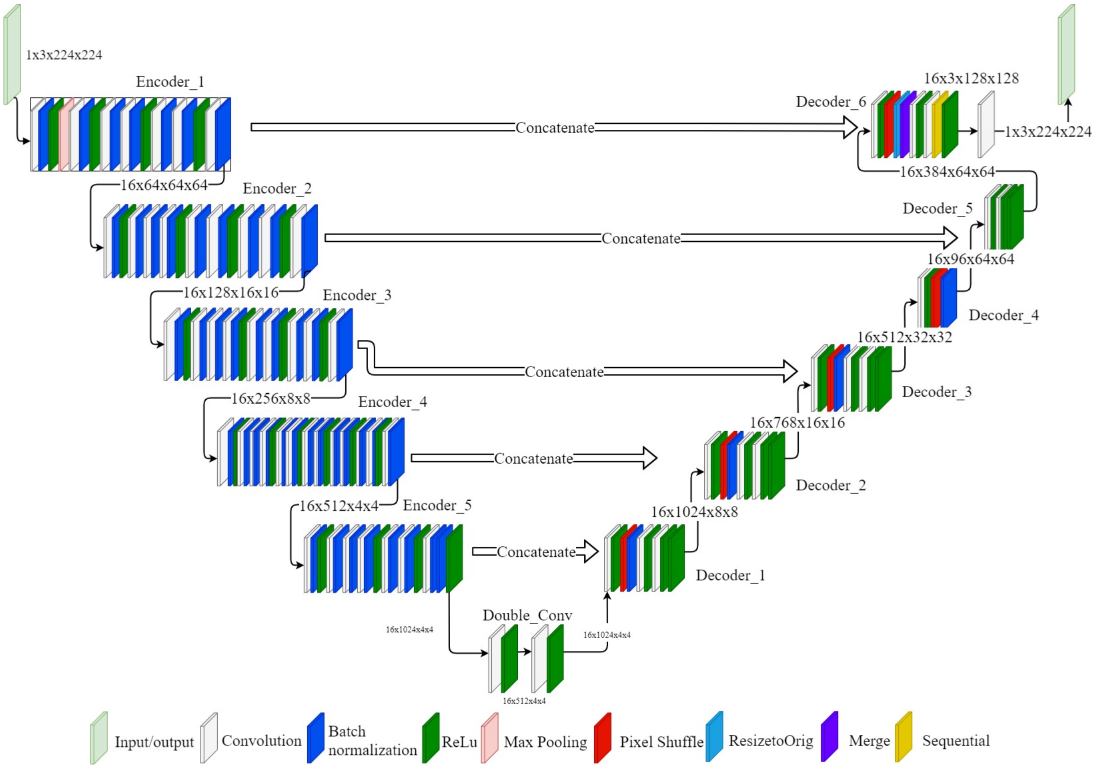

Figure 3: U-Net with ResNet Encoder

6. Gated Attention Module

6.1. Need for Attention

In medical images, regions of interest (e.g., lesions) are often small and ambiguous. Standard skip connections may introduce irrelevant background noise into the decoder. Attention gates help by selectively allowing important features to pass.

6.2. Mechanism

Given encoder features x and decoder gating signal g, the gated attention output is:

$$\begin{aligned} x' &= Conv(x), \quad g' = Conv(g) \\ f &= \text{ReLU}(x' + g') \\ \alpha &= \sigma(W_f f) \\ \hat{x} &= \alpha \cdot x \end{aligned}$$

Where:

- σ is the sigmoid activation for producing attention weights α.
- The output $\hat{x}$ is the gated feature passed to the decoder.

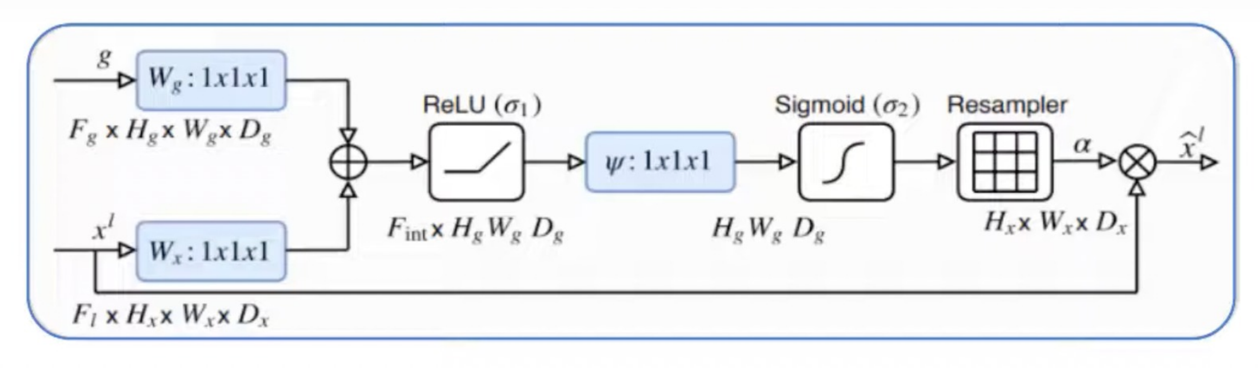

Figure 4: Gated Attention Module

7. Proposed Approach: Enhancing Spatial Attention through Encoder Integration

7.1. Encoder Attention

$$\begin{aligned}
x_q &= W_q x \\
x_k &= W_k x \\
x_v &= W_v x \\
x_{\text{att}} &= \text{softmax}(x_q x_k^\top) x_v \\
\hat{x} &= \text{GatedAttention}(x_{\text{att}}, g)
\end{aligned}$$

7.2. Need for Encoder Attention

While existing gated attention mechanisms primarily deliver spatial attention via skip connections, our proposed method advances this concept by integrating spatial attention directly within the encoder pathway of a U-Net architecture. By embedding attention modules into the encoder and leveraging skip connections, we harness a dual advantage: the encoder enriches feature representations with spatial context early in the network, and the skip connections further preserve and transmit these spatially attentive features to the decoder. This synergy results in spatially richer and more context-aware decoder outputs, thereby enhancing overall performance in tasks demanding fine-grained spatial understanding.

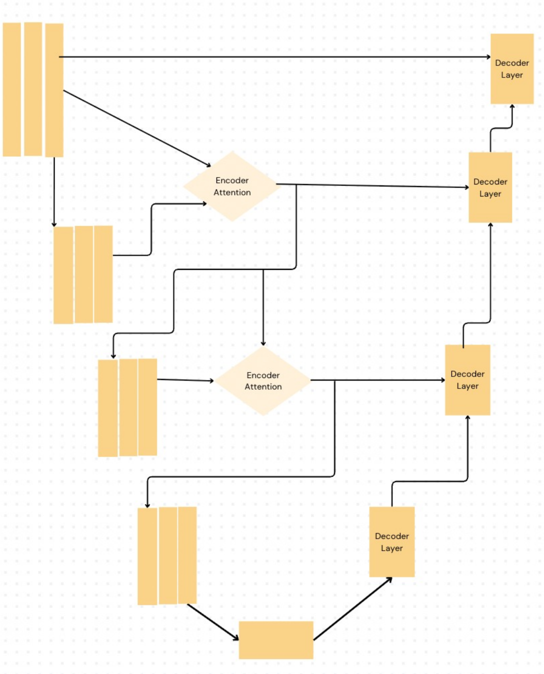

Figure 5: Encoder Attention U-Net Architecture

8. Full Architecture Pipeline

The proposed architecture, **AttenResUnet**, integrates an attention-augmented U-Net with residual connections derived from a ResNet34 backbone, facilitating accurate medical image segmentation tasks. The pipeline is structured into the following stages:

8.1. Data Preparation

Initially, images and their corresponding segmentation masks from the ISIC2018 dataset undergo preprocessing. This includes resizing and augmentation operations such as horizontal and vertical flipping, enhancing the dataset's variability and promoting improved model generalization.

8.2. Encoder Path (Downsampling)

The encoder utilizes a pre-trained ResNet34 architecture, partitioned into sequential stages:

- **Initial Encoding:** An input image first passes through initial convolutional, batch normalization, and ReLU activation layers inherited from ResNet34, forming a foundational feature extraction step.
- **Encoder Stages:** Subsequently, the model employs four successive residual encoding blocks (layers 1 to 4 from ResNet34), each block systematically downsamples and enriches feature maps. Between these stages, customized *Encoder Attention* modules are integrated:
  - **EncoderAttention Modules:** These modules apply encoder-attention to refine intermediate encoder feature maps. The encoder-attention mechanism involves query (W_q), key (W_k), and value (W_v) transformations. The resulting self-attention features are modulated via a secondary attention gate.

8.3. Bridging Layer

The bridging (bottleneck) layer consists of a *Double Convolution* (*DoubleConv*) module that processes the deepest encoder features, distilling the most salient information prior to decoding.

8.4. Decoder Path (Upsampling)

The decoder path involves iterative spatial reconstruction steps aided by attention-based skip connections from corresponding encoder layers:

- **Upsampling and Concatenation:** Four decoder blocks progressively restore spatial resolution. Each decoder unit includes a transposed convolution for upsampling, followed by attention-guided fusion with skip-connected features from the encoder path.
- **Attention-Gated Skip Connections:** Decoder blocks integrate attention gates to selectively emphasize encoder features, significantly improving boundary preservation and detail accuracy in segmentation outputs.

8.5. Final Reconstruction and Output

In the final reconstruction step, the decoder output undergoes an additional transposed convolutional layer for spatial refinement. The resultant feature maps are further processed via a convolutional layer, batch normalization, and a ReLU activation, resulting in segmentation maps aligned with input image dimensions.

8.6. Attention Visualization and Interpretability

Optionally, the model can produce intermediate attention maps, providing insights into which regions contribute significantly to segmentation decisions. These attention visualizations enhance interpretability, a crucial aspect for clinical and diagnostic decision-making processes.

Overall, the proposed architecture strategically combines the robust feature-extraction capabilities of residual networks with precision-enhancing attention mechanisms, demonstrating promising performance enhancements for medical image segmentation tasks.

9. Experimental Setup

9.1. Loss Function

In the conducted experiments, the training of the proposed AttenResUnet model utilized the Binary Cross-Entropy (BCE) with Logits loss function. This loss function is specifically selected for its effectiveness in binary segmentation tasks, providing numerically stable gradient updates by combining the sigmoid activation function directly into the loss calculation. It enables precise pixel-level classification, crucial for accurate delineation in medical image segmentation scenarios.

9.2. Optimizer

For optimization, the Adam optimizer was employed due to its adaptive learning capabilities, which dynamically adjust learning rates for each model parameter. The initial learning rate was set to 1×10^{-3}. Adam's adaptability to sparse gradients and robustness in convergence has been widely recognized in deep learning, making it suitable for efficiently training deep, attention-augmented architectures like AttenResUnet.

9.3. Batch Size

A batch size of 16 was chosen, balancing computational constraints with stable convergence behavior. This moderate batch size ensures efficient GPU utilization, reduces training variance, and enhances model stability without excessive memory consumption, an important consideration in resource-constrained environments.

9.4. Model Training

The model was run for a total of 10 epochs, sufficient to achieve convergence given the complexity of the dataset. The model is fine-tuned over an NVIDIA Tesla P100 for 1 hour.

9.5. Experiments Conducted

The experiments were implemented using the PyTorch deep learning framework. We evaluated four distinct variants of our proposed architecture to systematically investigate the effectiveness of attention mechanisms in different stages of the model. Specifically, these variants were: (I) **Model 1**, employing gated attention exclusively in the decoder segment. (II) **Model 2**, applying both encoder attention and gated attention comprehensively in both encoder and decoder stages; (III) **Model 3**, integrating encoder attention and gated attention exclusively within the encoder; and (IV) **Model 4**, utilizing encoder attention up to layer 3 alongside gated attention; These variations allowed us to thoroughly assess the individual and combined contributions of encoder and decoder attention mechanisms toward segmentation accuracy.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|P \cap G|}{|P| + |G|}, \quad \mathcal{L}_{\text{BCE}} = -[y \log \hat{y} + (1-y) \log(1-\hat{y})], \quad \mathcal{L}_{\text{FL}} = (p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t)$$

10. Experimental Results

Model	**Dice Coefficient**	**IoU**	**Focal Loss**
U-Net + ResNet + LimitEncoderAttention	**0.8660**	**0.795**	**0.245**
U-Net + ResNet + EncoderAttention	0.8615	0.789	0.257
U-Net + ResNet + GatedAttention	0.856	0.782	0.263
U-Net + ResNet + EncoderAttention + GatedAttention	0.8529	0.778	0.270

Table 1: Comparison of segmentation models using Dice Coefficient, IoU, and Focal Loss

Analysis:

- The incorporation of the **ResNet34 backbone** significantly enhances the model's ability to extract deep, meaningful features. Specifically, the residual connections mitigate the vanishing gradient problem by allowing direct gradient flow through identity shortcuts, which improves model generalization and training stability.

- The addition of **gated and encoder attention mechanisms**, further refines segmentation outputs by effectively suppressing irrelevant or noisy features. Specifically, these gates dcompute attention coefficients that highlight important spatial regions, ensuring that decoder layers focus primarily on informative areas identified during the encoder phase. This is evidenced by improved segmentation boundaries and reduced false-positive detections observed in the model's segmentation outputs.

- The final proposed model, integrating both the ResNet backbone and the limit encoder attention mechanism at critical encoder-decoder junctions, demonstrates superior performance when compared to the baseline Attention U-Net architecture. The enhanced performance is quantitatively validated by higher Dice Coefficient and Intersection over Union (IoU) scores achieved on the ISIC 2018 dataset.

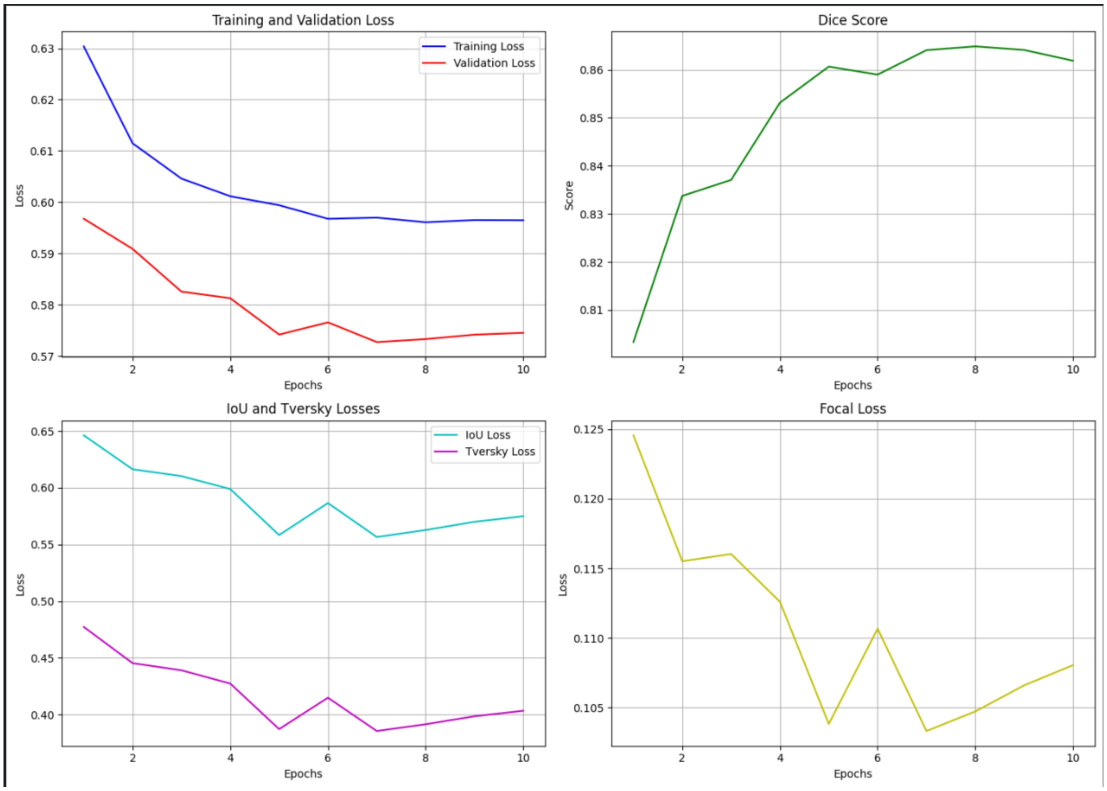

Figure 6: Graph Plots of Various Losses

Attention Maps of some Trained Models:

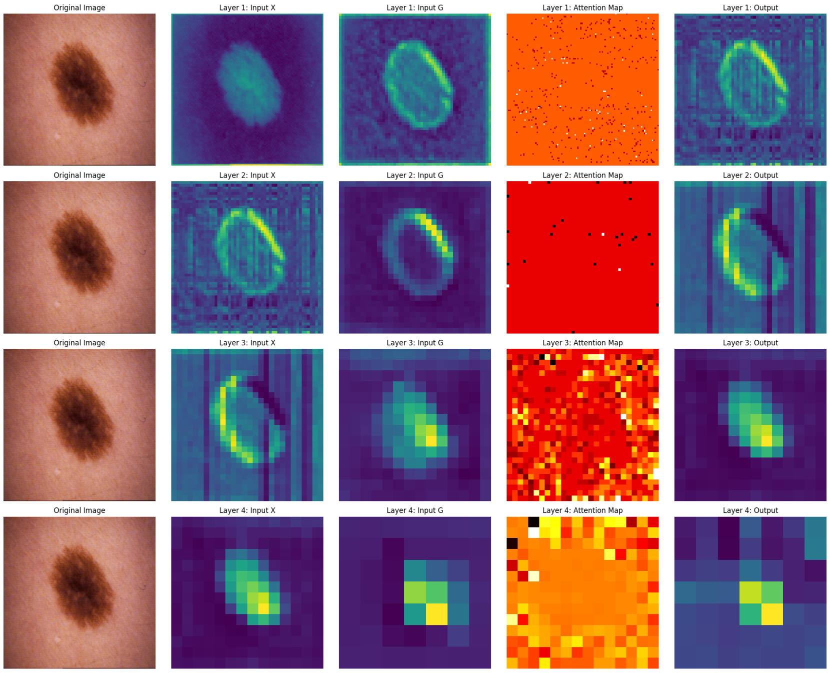

Figure 7: Visualization of Attention Maps in Encoder Attention

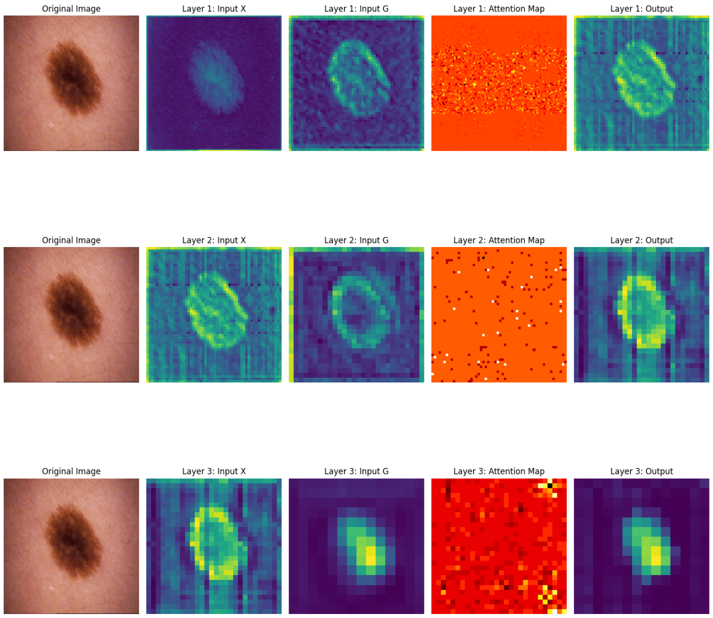

Figure 8: Visualization of Attention Maps in Limit Encoder Attention

11. Conclusion

We proposed an advanced medical image segmentation network based on U-Net with a ResNet encoder and gated attention modules. The model captures deep semantic features and focuses on relevant spatial regions, resulting in improved performance on ISIC 2018 data. Future extensions may include medical image segmentation of 3D images (like Tomograms), using a Resnet-50 or Resnet-101 backbone for Encoder.

12. References

1. Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation.* arXiv preprint arXiv:1505.04597.

2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

3. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M. P., Kainz, B., Glocker, B., & Rueckert, D. (2019). *Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images.* Medical Image Analysis, 53, 197–207.

4. Islam, M. H., Zhang, Y., Ren, H., Ren, P., & Hasan, M. K. (2020). *Gated Attention for Deep Multi-modal Medical Image Segmentation.* arXiv preprint arXiv:2007.03172.