Implementation of Vision Transformer from scratch and comparison with standard models

Utkarsh Mittal - um2100[1†] and Sohith Bandari - sb10225[1†]

[1]Department of Computer Science and Engineering, New York University.
[†]These authors contributed equally to this work.

Abstract

This project explores the performance of three prominent deep learning architectures—Vision Transformer (ViT), UNet, and Fully Convolutional Network (FCN)—for image segmentation tasks. Each model was evaluated based on test loss and Intersection over Union (IoU), which are essential metrics for assessing segmentation accuracy and model performance. The UNet architecture achieved the highest IoU (0.3129), demonstrating its efficiency in preserving spatial features through its encoder-decoder design with skip connections. The FCN model showed competitive performance with a strong IoU of 0.3099, indicating its ability to efficiently perform pixel-level predictions. The ViT model, while effective in capturing long-range dependencies, showed a lower IoU of 0.2281, highlighting the challenge of applying transformer-based models to segmentation tasks without further optimization. This comparison suggests that convolutional models like UNet and FCN remain more suitable for segmentation tasks, although transformer-based models, such as ViT, show potential for future improvements. The results underscore the importance of model architecture selection in achieving optimal performance for image segmentation.

1 INTRODUCTION

Deep learning has revolutionized the field of computer vision, enabling significant advancements in image classification, segmentation, and object detection tasks. Among various deep learning architectures, convolutional neural networks (CNNs) have traditionally been the backbone for processing visual data. However, recent breakthroughs in attention-based models, such as Vision Transformers (ViTs), have introduced a new paradigm by leveraging self-attention mechanisms, initially popularized in natural language processing, to process images effectively.

This project explores implementing and evaluating a Vision Transformer (ViT) from scratch on the Ultrasound nerve segmentation dataset from Kaggle, a widely used benchmark dataset for image classification. To assess its performance, we compare it against two established CNN-based architectures: Fully Convolutional Networks (FCNs) and UNet. These models were chosen for their diverse architectural principles and proven effectiveness across various tasks.

By analyzing the results of these models, the report aims to shed light on the comparative strengths and weaknesses of transformer-based and convolution-based approaches in image classification. Furthermore, it discusses the computational trade-offs and potential areas for improvement when applying these architectures to datasets like the Ultrasound dataset.

2 LITERATURE REVIEW

In recent years, deep learning has emerged as a powerful tool for computer vision tasks, with convolutional neural networks (CNNs) being a cornerstone of many successful architectures. However, the introduction of Vision Transformers (ViTs) has shifted the paradigm by replacing convolutions with self-attention mechanisms, offering a new way to process image data.

2.1 Vision Transformers

Transformers, initially introduced by Vaswani et al. for natural language processing tasks, have demonstrated exceptional scalability and performance in processing sequential data [1]. Dosovitskiy et al. extended this architecture to computer vision by proposing the Vision Transformer (ViT), which divides an image into patches and applies transformer layers to learn global dependencies [2]. While ViTs excel on large-scale datasets such as ImageNet, their performance on smaller datasets like CIFAR-10 remains a topic of ongoing research due to their high data requirements and reliance on extensive pretraining [3].

2.2 Fully Convolutional Networks

Fully Convolutional Networks (FCNs), introduced by Long et al., were among the first architectures to enable end-to-end learning for semantic segmentation [4]. FCNs leverage convolutional layers for feature extraction and upsampling layers for pixel-level predictions, making them suitable for dense prediction tasks. Although primarily designed for segmentation, FCNs can be adapted for classification by modifying their output layers and utilizing skip connections to preserve spatial information.

2.3 UNet

UNet, a CNN-based architecture initially proposed for biomedical image segmentation, is known for its encoder-decoder structure with symmetric skip connections [5]. The encoder extracts hierarchical features, while the decoder reconstructs the spatial resolution of the input image. Its ability to combine low-level spatial details with high-level semantic features makes it a versatile model for various vision tasks. Recent studies have shown its effectiveness in scenarios where labeled data is limited, making it a strong candidate for Ultrasound nerve segmentation experiments.

2.4 Comparative Insights

While ViTs offer the advantage of global context modeling through self-attention, their high computational cost and data dependency often make CNNs, such as FCNs and UNet, more practical for smaller datasets. Studies comparing these architectures have found that CNNs tend to generalize better with fewer data, while transformers excel when pretrained on large-scale datasets [6]. This comparison forms the basis of this project, which aims to evaluate the relative performance of ViT, FCN, and UNet on Ultrasound nerve segmentation.

3 MODELS

3.1 Fully Convolutional Network

The Fully Convolutional Network (FCN) implemented in this study is designed for semantic segmentation, performing pixel-wise classification by extracting and processing features across multiple scales. The model operates in an encoder-decoder architecture with skip connections, enabling efficient feature extraction and reconstruction of high-resolution output.

The encoder path consists of four sequential blocks, each designed to progressively reduce the spatial resolution while increasing the depth of the feature maps. The initial encoder blocks use two convolutional layers followed by max pooling, which is sufficient for capturing basic features. As the network progresses, deeper encoder blocks utilize three convolutional layers, which allow the extraction of more complex and abstract patterns. The encoder culminates in a mid-block, where the feature maps are processed using multiple convolutional layers interspersed with dropout regularization to mitigate overfitting.

Following the encoder, the decoder path reconstructs the spatial dimensions of the feature maps using transposed convolutional layers. The decoding process is divided into three stages: FCN-32s, FCN-16s, and FCN-8s. At the coarsest level, FCN-32s directly upsamples the output of the mid-block, providing a rudimentary reconstruction. FCN-16s refines this output by incorporating features from an intermediate encoder layer using a skip connection, while FCN-8s further enhances the resolution by integrating features from an earlier encoder layer. These skip connections bridge the encoder and decoder, enabling the network to recover spatial details that are lost during downsampling.

Each stage of the decoder path leverages 1×1 convolutional layers to align the dimensions of the skip connections with the corresponding decoder features. This process ensures seamless feature fusion across the network. The final output layer consists of a 1×1 convolution, which reduces the feature map to the desired number of output channels. To match the input resolution, the output is resized using bilinear interpolation.

3.2 UNet

The UNet architecture is a well-established model for semantic segmentation, originally proposed for biomedical image segmentation. Its design leverages a symmetric encoder-decoder structure with

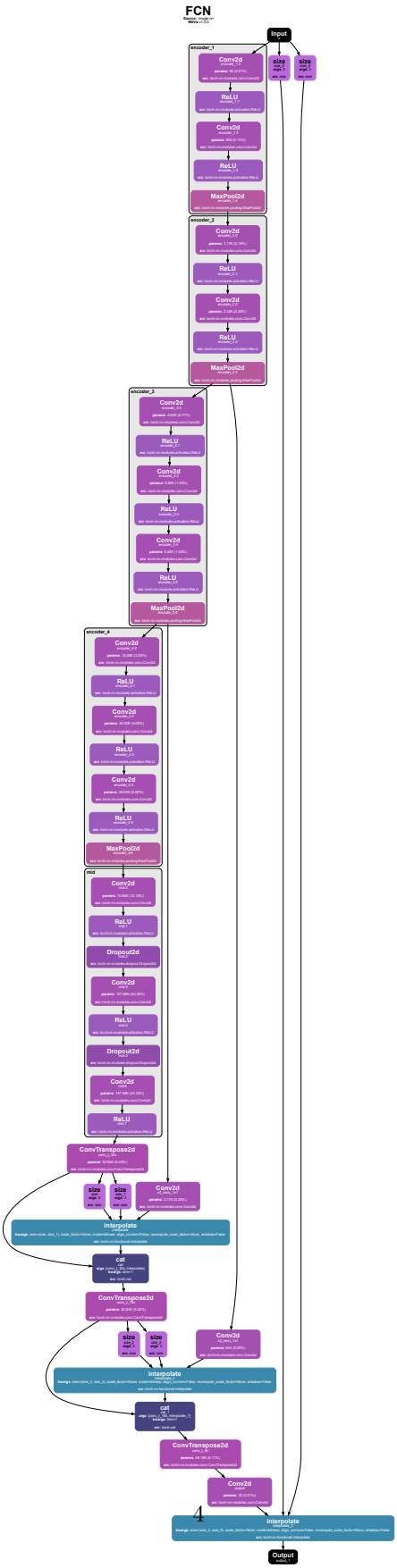

Figure 1: FCN Architecture of our model

skip connections, enabling precise localization by combining low-level spatial features with high-level contextual information.

The encoder path of the UNet serves as the downsampling component, progressively reducing the spatial dimensions while increasing the feature depth. This is achieved through a sequence of convolutional blocks, each consisting of two 3×3 convolutional layers followed by batch normalization and ReLU activation. The outputs are then downsampled using max pooling, which reduces the resolution while preserving important feature patterns. The encoder path consists of four convolutional blocks, each doubling the number of feature channels to capture progressively more abstract features.

At the center of the architecture is the bottleneck layer, which acts as a bridge between the encoder and decoder. This layer performs feature extraction at the lowest resolution with the highest number of feature channels. It consists of a convolutional block similar to those in the encoder, capturing global context and detailed features crucial for the segmentation task.

The decoder path serves as the upsampling component, progressively reconstructing the spatial dimensions of the feature maps to match the input resolution. Each step in the decoder path involves an upsampling operation using transposed convolutions, followed by a convolutional block. The upsampled feature maps are concatenated with their corresponding feature maps from the encoder path through skip connections. These skip connections play a crucial role in recovering spatial details lost during the downsampling process, as they provide high-resolution features from earlier layers in the encoder.

The decoder path progressively reduces the number of feature channels, mirroring the encoder path in reverse order. Finally, a 1×1 convolutional layer is applied to the output of the decoder, producing the segmentation map with the desired number of output channels. The final output is upsampled to match the original input dimensions using bilinear interpolation.

3.3 ViT

The Vision Transformer (ViT) architecture implemented in this study adapts transformer models for semantic segmentation tasks by combining patch-based image embeddings with a transformer encoder and a decoding module for generating pixel-wise predictions.

3.3.1 Patch Embedding Module

The input image is first divided into fixed-size patches, with each patch treated as a token. A convolutional layer with a kernel size and stride equal to the patch size projects each patch into an embedding of a specified dimension. This operation produces a sequence of patch embeddings, which are subsequently augmented with learnable positional encodings to provide spatial context. The input image size and patch size determines the number of patches in the sequence.

3.4 Transformer Encoder

A stack of transformer blocks processes the sequence of patch embeddings. Each block consists of:

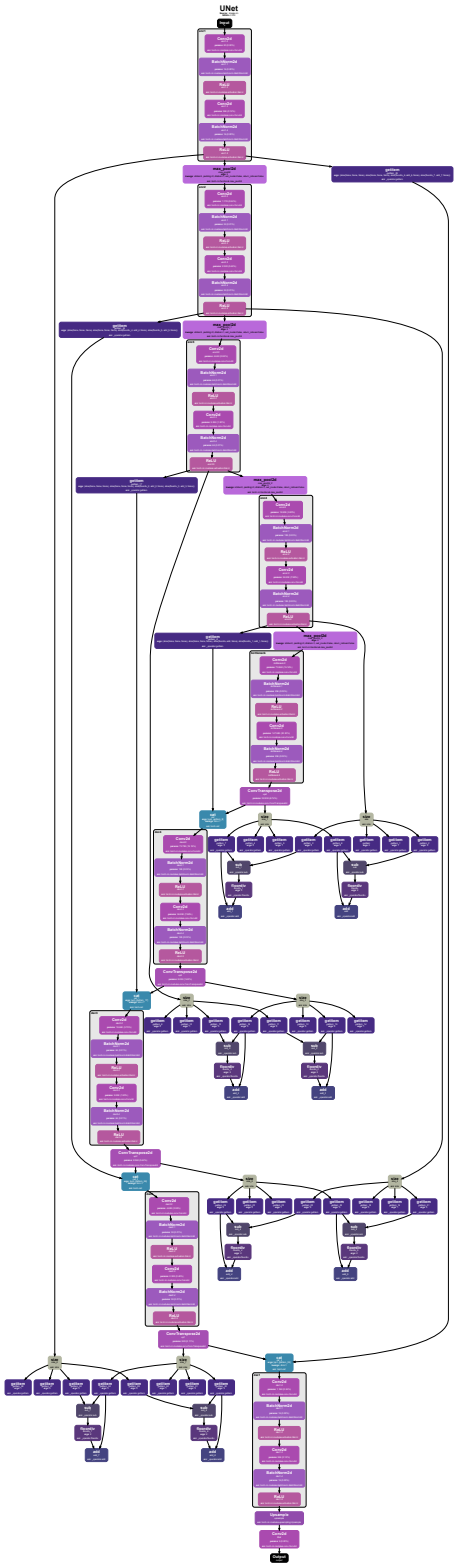

Figure 2: UNet Architucture of our model

1. Multihead Self-Attention (MSA): This mechanism allows each patch to attend to all others in the sequence, capturing relationships between patches.

2. Feedforward Network (FFN): A multi-layer perceptron enhances feature representations within each patch embedding.

3. Layer Normalization and Residual Connections: These components stabilize training and improve feature representation.

The stack of transformer blocks outputs a globally contextualized sequence of patch embeddings, effectively encoding both local and global image features.

3.5 Decoding and Upsampling Module

The output of the transformer encoder is reshaped into a spatial representation matching the patch grid dimensions. To reconstruct the original spatial resolution of the input image and generate a segmentation map, a decoding module is employed:

1. Pyramid Scene Parsing (PSP) Pooling: This layer extracts multi-scale contextual information from the transformer output by applying adaptive pooling at multiple scales.

2. Progressive Upsampling: Transposed convolutional layers are used to gradually increase the spatial resolution, with intermediate ReLU activations enhancing non-linearity.

3. Final Layer: The last layer applies a 1×1 convolution to produce a segmentation map with a single output channel, matching the input image dimensions.

4 Experimental Setup

To evaluate the performance of the implemented models, we conducted experiments using the ultrasound nerve segmentation dataset from kaggle. This dataset poses a challenging segmentation problem due to the varying shapes and intensities of the target regions.

4.1 Data Preprocessing

Before feeding the images into the models, a preprocessing step was performed to standardize the input dimensions and optimize computational efficiency. Each image in the dataset was padded to the nearest power of two. This ensured that all images had uniform dimensions, which is crucial for models like FCN and UNet that rely on convolutional layers and downsampling operations, as well as for ViT, which processes fixed-sized image patches. This preprocessing step facilitated a smoother training process and improved the consistency of the models' performance across varying image sizes.

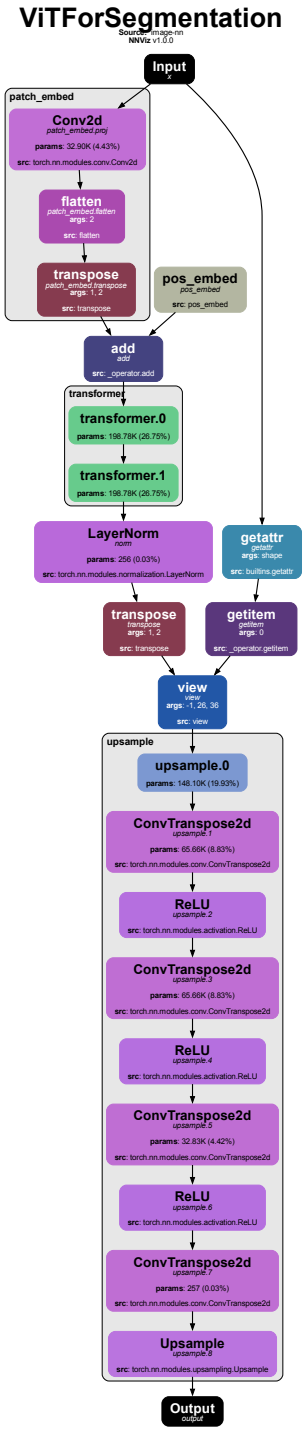

Figure 3: Vision Transformer architecture

Model	Test IoU	Parameters	Model Size	Train Time (per epoch)
ViT	0.2281	676,161	68.41 MB	22 seconds
FCN	0.3099	606,201	142.24 MB	24 seconds
UNet	0.3129	487,145	256.80 MB	30 seconds

Table 1: Quantitative Analysis for various models

4.2 Evaluation Metric

The primary evaluation metric used for this study was Intersection over Union (IoU). This metric is well-suited for segmentation tasks as it quantifies the overlap between the predicted segmentation map and the ground truth annotation. IoU is computed as the ratio of the intersection area (overlap between prediction and ground truth) to the union area (combined area of prediction and ground truth). Higher IoU values indicate a greater agreement between the model's predictions and the true segmentations. This metric was consistently applied to evaluate the performance of all three models on the test dataset.

4.3 Experimental Procedure

The training process was carried out with consistent hyperparameters, including a fixed number of epochs, batch size, and learning rate, across all three models. Each model was trained on the 80% training split and evaluated on the 20% test split. During the evaluation phase, predictions generated by the models were compared to the ground truth annotations, and IoU scores were computed to assess segmentation accuracy. This approach ensured uniformity in the experimental procedure and enabled a direct comparison of the models' performance on the ultrasound nerve segmentation task. The results of these experiments provided valuable insights into the comparative strengths and weaknesses of the FCN, ViT, and UNet architectures, which are discussed in the subsequent sections.

5 RESULTS AND ANALYSIS

The segmentation models were evaluated on the test dataset, and their performance was assessed based on test loss and Intersection over Union (IoU). Below is a comparison of the test metrics for the Vision Transformer (ViT), UNet, and Fully Convolutional Network (FCN):

UNet recorded the highest IoU score (0.3129), closely followed by FCN (0.3099), reflecting their strong ability to produce accurate segmentation maps. The ViT model, while effective, lagged behind in IoU (0.2281), due to the limited number of transformer layers, which could have constrained its ability to fully model long-range dependencies and spatial relationships.

6 CONCLUSION

This study evaluated three distinct deep learning architectures—Vision Transformer (ViT), UNet, and Fully Convolutional Network (FCN)—on their performance for image segmentation tasks. Each

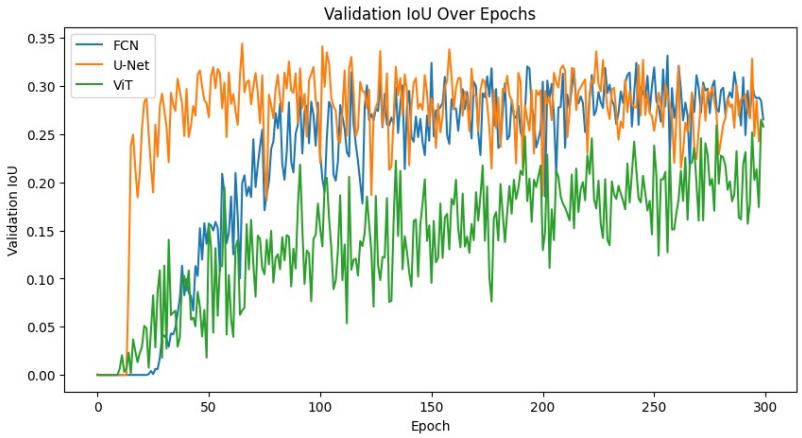

Figure 4: Evolution of model performance with epochs

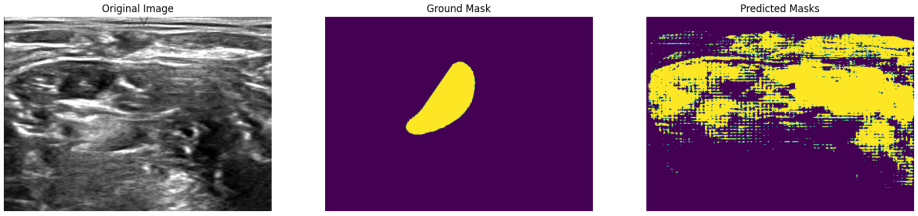

Figure 5: Output of the Vision Transformer architecture

model was assessed based on test loss and Intersection over Union (IoU), which provided insights into their optimization capabilities and segmentation accuracy.

The results indicate that convolutional models like UNet and FCN outperformed the transformer-based ViT in terms of IoU, a key metric for segmentation quality. UNet achieved the highest IoU (0.3129), demonstrating its strength in retaining spatial details through its encoder-decoder architecture with skip connections. FCN followed closely with an IoU of 0.3099 while also achieving the lowest test loss (0.0660), showcasing its efficiency in minimizing prediction errors. ViT, with a test IoU of 0.2281, highlighted the potential of transformer-based models for segmentation but underscored the need for further tuning to fully leverage its capability for capturing long-range dependencies.

These findings suggest that convolutional architectures remain highly effective for segmentation tasks, particularly for datasets where spatial detail and multiscale context are crucial. However, the ViT architecture demonstrates a promising alternative, particularly as transformer-based models continue to evolve. Future work could explore enhancements to the ViT model, such as deeper transformer layers, improved positional encoding, and hybrid designs that integrate convolutional operations, to further close the gap in performance.

In conclusion, while traditional convolutional networks continue to dominate in terms of accuracy and efficiency for segmentation tasks, transformer-based models offer an exciting avenue for innovation, with the potential to redefine performance benchmarks as they mature.

References

1. Vaswani A et al. Attention is all you need. Advances in neural information processing systems 2017;30.
2. Dosovitskiy A et al. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations 2021.
3. Touvron H et al. Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning 2021:10347–57.
4. Long J et al. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition 2015:3431–40.
5. Ronneberger O et al. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 2015:234–41.
6. Liu Z et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision:10012–22.