Deep Dive into Transformer Segmentation: A Comparative Architectural Study

Anjel Patel*, **Eshaan Raj Sharma** *, **Meet Oswal***

New York University

Link to Project GitHub:
https://github.com/eshaanraj25/Deep-Dive-into-Transformer-Segmentation-A-Comparative-Architectural-Study/tree/main

Abstract

Image segmentation, a pivotal task in computer vision, facilitates the precise delineation and classification of objects or regions within images. In this comparative study, we investigate the efficacy of transformer-based models such as ViT (Vision Transformer) and UNet Transformer, alongside the CNN model SegNet, across various architectural configurations for image segmentation tasks. Leveraging the OxfordIIIT pet dataset, we meticulously train and evaluate these models, presenting detailed analyses and performance comparisons. These comparisons clearly indicate the promising nature of Ensembling UNet (with ResNet50, DenseNet121 and MobileNet v2) while ViT based models perform spectacularly poor all accross given the high number of parameters.

Introduction

In the rapidly evolving domain of computer vision, the critical tasks of image segmentation, object localization, and classification are pivotal across various advanced applications from autonomous vehicles to detailed medical diagnostics. Traditionally, these tasks have been dominated by convolutional neural networks (CNNs), particularly demonstrated by the effectiveness of architectures like ResNets. However, the advent of transformer architectures in image processing presents a novel paradigm that could potentially redefine benchmark standards in the field. This paper aims to delve into a comparative analysis of transformer-based models on the Oxford Pet dataset [6], focusing on their application in the semantic segmentation of animals in images. We investigate whether the newer, potentially more flexible transformer models, including Vision Transformers, SegNet-Basic With Standard Convolutions, SegNet-Basic With Depthwise Separable Convolutions, and U-Net Architecture, can match or exceed the performance of traditional CNNs in terms of accuracy and efficiency in real-world imaging scenarios.

Literature Review

The field of computer vision, pivotal in applications ranging from autonomous driving to medical diagnostics, has significantly advanced through deep learning techniques, especially convolutional neural networks (CNNs). Traditional CNN architectures like ResNets have demonstrated remarkable efficacy in complex image processing tasks due to their deep structures and ability to combat vanishing gradients [4].

However, transformer architectures, originally designed for natural language processing, have been adapted to image processing. Vision Transformers (ViT) apply transformer blocks to sequences of image patches, learning contextual relationships vital for nuanced tasks like semantic segmentation [3]. Similarly, SegNet variants, known for their encoder-decoder structure, offer optimized computational efficiency, making them suitable for real-time applications [1].

Additionally, the U-Net architecture, particularly noted for its efficient use of data through extensive augmentation and a novel network structure conducive to precise localization and segmentation in biomedical imaging, represents a significant advancement in the field. The U-Net architecture, with its contracting and expansive paths, has been proven to work with very few training images while outperforming traditional methods in segmentation tasks [7].

SegNet architectures, known for their distinct encoder-decoder structure, are especially valued for their computational efficiency, making them well-suited for real-time image processing applications. This efficiency stems from a unique approach where the encoder network maps an input image down into a smaller, denser representation, which the decoder network then expansively maps back to the original dimensions to complete the segmentation task. This structure not only helps in reducing the computational load but also preserves important spatial hierarchies between the image segments, which is crucial for tasks requiring precise localization and contextual awareness [1]. This study explores these models' performance on the Oxford Pet Dataset, a challenging set for semantic segmentation due to its variety in animal poses and scales. The aim is to evaluate whether transformers, advanced CNN variants, and the U-Net model can surpass traditional models in accuracy and efficiency, crucial for real-world imaging scenarios.

Methodology

In this study, we employed four distinct transformer architectures for image segmentation, alongside Mask R-CNN, to establish a comparative framework. This approach allowed us to explore the efficacy of transformer-based meth-

*These authors contributed equally.

ods relative to the well-established Mask R-CNN in handling complex image segmentation tasks. We implemented four transformer-based architectures, each designed to address the challenges of image segmentation through distinct mechanisms and structural innovations. These models were evaluated against the Mask R-CNN, a model that is traditionally celebrated for its robust performance in object detection and segmentation tasks. This comparative analysis aims to highlight the strengths and potential limitations of transformer architectures in achieving high precision and efficiency in segmenting complex images from the Oxford-IIIT Pet dataset.

Dataset

The Oxford-IIIT Pet dataset served as the foundational dataset for our experiments. This dataset is renowned for its application in pet classification, detection, and segmentation tasks. It encompasses a diverse collection of cat and dog images, each supplemented with annotated bounding boxes and detailed segmentation masks. For the purposes of our study, these annotations were instrumental in generating binary masks that precisely delineate the pet figures in the images. These binary masks, combined with the original images, formed the comprehensive training dataset for our models. Evaluation of the models was rigorously performed using the designated test set, enabling us to assess and compare the segmentation performance of each model accurately.

Vision Transformer

The model architecture is based on the Vision Transformers framework. The input consists of images with a batch size of 64, 3 channels, and dimensions of 128 x 128 pixels. Initially, the images undergo batch normalization and are divided into patches of 16 x 16 pixels, which are then transformed into embeddings of dimension 768. Positional information is incorporated into these patches to preserve spatial context. Subsequently, the patches are fed into a sequence of Transformer Encoder Blocks, each comprising a Self-Attention mechanism and an MLP block. These blocks process the patch sequence data. Following the passage through 14 Transformer Encoder Blocks, the output is directed to the Output head, where the vector embeddings of the patches are amalgamated to generate a mask of the image, serving as the model's output. The model architecture can be found in Figure 1. The total number of parameters of the trained model amounts to 100.46 million.

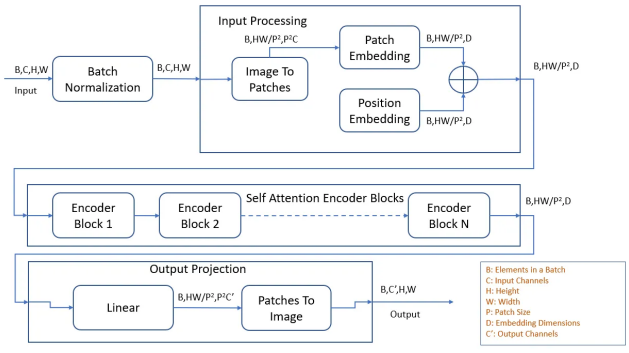

Figure 1: Vision Transformer For Image Segmentation [5]

Hyperparameter tuning was conducted to optimize model accuracy, encompassing various parameters:

- **Batch size**: Experimentation with batch sizes of 32, 64, and 128 revealed superior performance with a batch size of 64.
- **Patch size**: Evaluation of patch sizes 8, 16, and 32 indicated that a patch size of 8 did not generalize well, while a patch size of 32 did not localize effectively. Therefore, a patch size of 16 was deemed optimal.
- **Self-Attention Head Count**: Different configurations with 8, 12, and 14 self-attention heads were examined.
- **Epochs**: The model underwent training for 20, 50, and 100 epochs, with 50 epochs yielding a balanced performance in terms of training and test accuracy.

U-Net

The U-Net model consists of an encoder-decoder structure. The encoder uses convolutional layers to capture context and max-pooling layers to reduce spatial dimensions. The decoder applies transposed convolutions to achieve precise localization by upscaling the feature maps. Skip connections are utilized to carry information directly across the network to recover spatial details in the output segmentation mask. Input and Processing: Each input image is resized to 128x128 pixels with 3 channels. The images are normalized to have pixel values between 0 and 1. Hyperparameter tuning was conducted to optimize model accuracy, encompassing various parameters:

- **Batch size**: Tested with 32, 64, and 128, with 32 found to be optimal.
- **Dropout rates**: Varied between 0.1 and 0.5 to mitigate overfitting, with 0.3 used in the final architecture to balance regularization and network training dynamics.
- **Number of filters**: Begins at 32 in the first layer and doubles with each subsequent layer of the encoder, optimizing the network's ability to process and learn from image features.
- **Epochs**: Models were initially tested with 20, 50, and 100 epochs; 20 epochs were chosen for efficiency and to avoid overfitting, given the robustness of the training data.

Building upon the baseline U-Net model, the UNet-Ensemble model employs three distinct U-Net configurations utilizing pretrained ResNet50, DenseNet121, and MobileNetV2 as the backbone encoders. These models are chosen for their robust feature-extraction capabilities and are integrated with custom decoder blocks to ensure detailed and

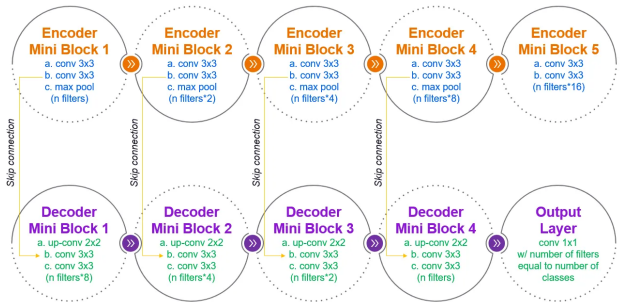

Figure 2: U-Net Transformer For Image Segmentation [2]

accurate segmentation. The ResNet50 U-Net uses ResNet50 layers for feature extraction with additional decoder blocks to refine segmentation details. The DenseNet121 U-Net leverages DenseNet121, known for its dense connectivity pattern, improving feature propagation and reuse. The MobileNetV2 U-Net incorporates MobileNetV2, optimized for mobile devices, ensuring a balance between accuracy and computational efficiency. The ensemble approach is not just limited to leveraging multiple architectures but also extends to using a combination of loss functions to optimize training:

- Sparse Categorical Crossentropy: Provides a baseline for comparison.
- Dice Loss: Improves performance on imbalanced data, ensuring that the model pays more attention to the segmentation regions.
- Categorical Focal Loss: Focuses learning on hard examples that are misclassified.
- Hybrid Losses: Combinations such as focalLoss + 3*diceLoss and (2*focalLoss + 1.7*diceLoss)/2 are used to tailor the model's sensitivity to the dataset's specific characteristics.

Each model is compiled with Adam optimizer and evaluated using accuracy, IoU (Intersection over Union), and F-score, which provide a comprehensive overview of model performance across various thresholds. The ensemble is trained using customized data pipelines that involve extensive data augmentation and batch processing to enhance the model's ability to generalize across different data distributions. The final model output is derived by strategically weighting the predictions from each individual network, thereby harnessing the strengths of each underlying architecture: Weighted Ensemble Predictions: Each model's predictions are weighted and combined, with weights tuned to optimize performance based on validation metrics. Operational Efficiency: Given the computationally efficient backbones like MobileNetV2, the ensemble is not only accurate but also practical for deployment in resource-constrained environments. By integrating these sophisticated architectures and hybrid loss functions, the UNet-Ensemble stands out for its ability to deliver highly precise segmentation results, making it ideal for challenging applications where accuracy and efficiency are paramount.

SegNet

SegNet, a modified convolutional neural network architecture tailored for image segmentation tasks, employs an encoder-decoder framework with downsampling and upsampling stages. The input image of size 128x128 pixels undergoes batch normalization to standardize pixel value distributions across the batch. The normalized image is then fed through a series of downsampling blocks comprising convolutional layers, batch normalization, ReLU activations, and max-pooling operations. The max-pooling indices and shapes are stored for subsequent upsampling during decoding. The encoded feature maps from the downsampling stage are then passed through upsampling blocks in the decoder, where they are convolutionally processed, batch normalized, and ReLU activated. The final output is a segmentation map classifying each pixel into distinct semantic categories based on the learned features, representing the model's prediction of the input image's semantic segmentation.

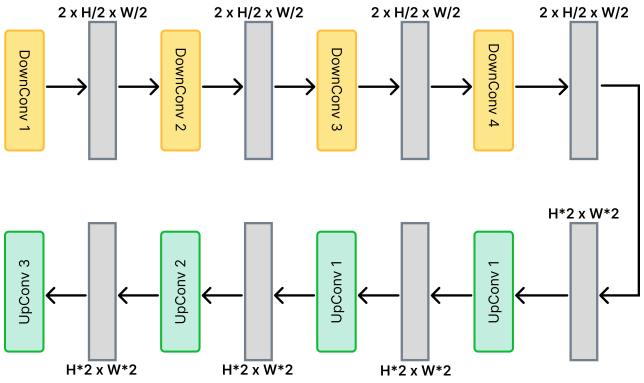

Figure 3: SegNet for Image Segmentation

Hyperparameter tuning was conducted to optimize model accuracy, encompassing various parameters:

- **Batch size**: Experimentation with batch sizes of 32, 64, and 128 revealed superior performance with a batch size of 64.
- **Patch size**: Evaluation of patch sizes 4, 8, and 16. A patch size of 4 was deemed optimal.
- **Dropout**: Dropout rates ranging from 0.2 to 0.4 were tested.
- **Self-Attention Head Count**: Different configurations with 2, 5, and 8 self-attention heads were examined.
- **Epochs**: The model underwent training for 20, 50, and 100 epochs, with 50 epochs yielding a balanced performance in terms of training and test accuracy.
- **Embedding Dimension**: Experimented with embedding dimensions of 64, 512, and 320. Where the dimension of 320 gives the best result.

Model	Params	Epochs	Pixel Acc	IoU
Baseline-UNet	8.6M	10	90.15%	-
ResNet50-UNet	20.6M	10	89.91%	71.39%
DenseNet121-UNet	16.4M	10	90.15%	77.32%
MobileNet-v2-UNet	10M	10	**90.39**%	75.32%
Ensemble-UNet	10M	10	-	**78.04**%
ViT-4	29.58M	60	84.47%	71.40%
ViT-14	100.46M	60	86.04%	74.83%
ViT-16	114.46M	60	86.20%	74.84%
SegNet-S	15.27M	20	85.24%	57.94%
SegNet-D	1.75M	20	84.50%	60.82%

Table 1: Results

Results

The results, summarized in Table 1, demonstrate the performance variations across all models above. Among the UNet-based architectures, MobileNet-v2-UNet achieved the highest pixel accuracy of 90.39%, while the ensemble model yielded the optimal IoU score of 78.04%, indicating the best performance overall. The ViT models exhibited subpar pixel accuracy and IoU score, with ViT-16 barely reaching 86.20%, at the cost of astronomical parameter counts making them the least suitable for image segmentation tasks. Conversely, the lightweight SegNet-D model, with 1.75M parameters, attained a competitive IoU score of 60.82%, outperforming the more parameterized SegNet-S variant. These findings underscore the trade-offs between model complexity, pixel accuracy, and IoU performance for segmentation tasks.

Future Directions

To continuously advance the field of image segmentation, we propose the following strategic directions for future research and development.

1. **Advanced Transformer Architectures:** Explore newer transformer architectures like Swin Transformers and CrossViT for better feature capture in segmentation tasks.
2. **Few-Shot and Zero-Shot Learning:** Investigate learning techniques that require minimal data annotation for generalization in image segmentation.
3. **Real-Time Segmentation:** Experiment with real-time segmentation to apply models in scenarios requiring immediate results.
4. **Hybrid Architectures:** Develop models that combine CNNs and transformers to leverage the strengths of both in achieving superior segmentation accuracy.
5. **Domain Adaptation:** Focus on methods to enhance model robustness and adaptability across different environments without extensive retraining.
6. **Edge AI Deployment:** Optimize models for deployment on edge devices with limited computational resources.
7. **Explainable AI (XAI):** Incorporate explainable AI techniques to enhance understanding and trust in AI decisions.
8. **Dataset Diversification:** Validate models on diverse datasets to test robustness and generalization capabilities.
9. **Augmentation Strategies:** Explore advanced data augmentation techniques to help models learn robust features.
10. **Loss Function Innovations:** Experiment with various loss functions to improve segmentation accuracy, especially at boundaries.

References

[1] Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.

[2] Bhatia, V. 2021. U-Net Implementation from Scratch Using TensorFlow. https://medium.com/geekculture/u-net-implementation-from-scratch-using-tensorflow-b4342266e406.

[3] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[4] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[5] Matani, D. 2023. Efficient Image Segmentation Using PyTorch Part 4. https://towardsdatascience.com/efficient-image-segmentation-using-pytorch-part-4-6c86da083432.

[6] Oxford Robotics Institute. 2012. The Oxford-IIIT Pet Dataset.

[7] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. Springer.

Appendix

Link to the Project Github: https://github.com/eshaanraj25/Deep-Dive-into-Transformer-Segmentation-A-Comparative-Architectural-Study/tree/main