Project

Problem Statement:

In this project, we have to select one of the tasks from image denoising, classification and segmentation apply a deep Learning model. For deep learning, first we have to survey three papers of the task that we want to solve and choose one from them.

Introduction:

In this project, I have selected the task of segmentation for medical image analysis. For this purpose, I have selected following papers to survey:

1. Fast-SCNN: Fast Semantic Segmentation Network [1].
2. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers (SETR) [2].
3. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers [3].

Fast-SCNN:

State-of-the-art semantic segmentation DCNNs combine two separate modules: the encoder and the decoder [4] (Figure 1). The encoder module uses a combination of convolution and pooling operations to extract DCNN features. The decoder module recovers the spatial details from the sub-resolution features, and predicts the object labels (i.e., the semantic segmentation). Fully convolution networks (FCN) [4] started the trend of bilinear up-sampling along with skip connection for recovering spatial details. More recently, two branch systems were introduced [5], [6]. They learn global context with a deep branch using small input and boundaries are learnt using a shallow branch but full-size input image. Fast-SCNN merges the two-branch setup with encoder-decoder framework and proposes improvements to make the semantic segmentation prediction above real-time (figure 2).

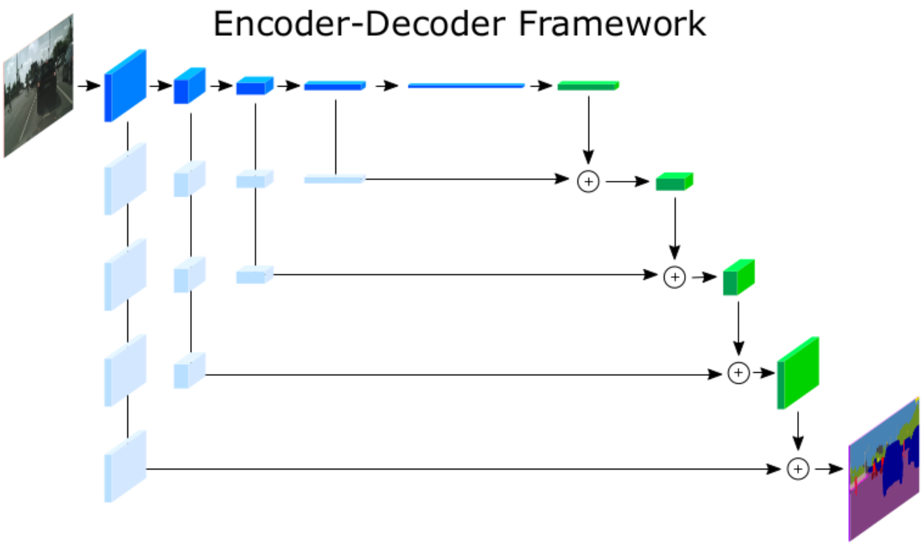


Figure 1: Encoder Decoder Architecture

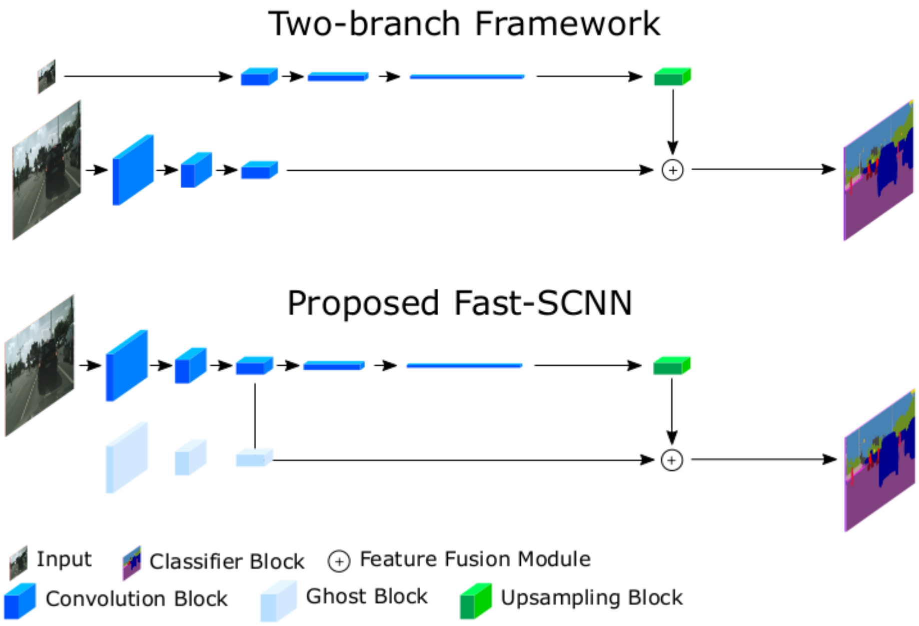


Figure 2: Two branch and proposed method

As stated earlier that two branch networks require two different resolution images for feature extraction. Therefore, rather than employing a two-branch approach with separate computation, authors introduce learning to downsample, which shares feature computation between the low and high-level branch in a shallow network block. Fast-SCNN network can be divided into four major parts:

1. **Learning to Down sample:** here they employ only three convolutional layers with stride two for downsampling. First layer is simple Conv2D while the remaining two are depthwise separable convolutional layers.
2. **Global Feature Extraction:** The global feature extractor module is aimed at capturing the global context for image segmentation. In contrast to common two-branch methods which operate on low resolution versions of the input image, their module directly takes the output of the learning to downsample module.
3. **Feature Fusion Module:** In this module features from above two modules are added up.
4. **Classifier:** In the classifier, authors employ two depthwise separable convolutions (DSConv) and one pointwise convolution (Conv2D). They found that adding few layers after the feature fusion module boosts the accuracy.

For improving inference performance, authors use depthwise separable convolutions introduced in MobileNet [7] along with network quantization and network compression using pruning technique. Fast-SCNN was designed with efficiency in mind so it is a low-capacity network with 1.11 million parameters. They empirically show that small capacity networks don't get significant benefit from ImageNet pretraining. Instead, aggressive data augmentation and more training epochs provide similar results.

SETR (SEgmentation TRansformer):

Semantic segmentation models mainly contain encoder-decoder architecture where encoder downsample the feature maps and extract semantic features from the input image and decoder upsample the feature maps and produces pixel level classification. CNN based models need to stack several layers of decreasing spatial dimension to obtain large enough receptive field for understanding the semantics of the image and making prediction. However, learning long-range dependency information, critical for semantic segmentation in unconstrained scene images remains challenging due to still limited receptive fields. A number of approaches are introduced, some includes using large kernel sizes [8], atrous convolutions [9] while others apply attention modules into FCN architecture.

This paper takes completely different approach and replaces all CNN layers with transformer [10]. Previously vision transformer (ViT) [11] has been used for classification task. This paper took this approach even further and use it for semantic segmentation. Encoder of the transformer is similar to that of ViT [?] .i.e., image is divided into several patches then these patches are flattened and further mapped into embedding space using a linear projection. Spatial information embedding is also added with the patch embedding before passing it into the transformer encoder (figure 3).

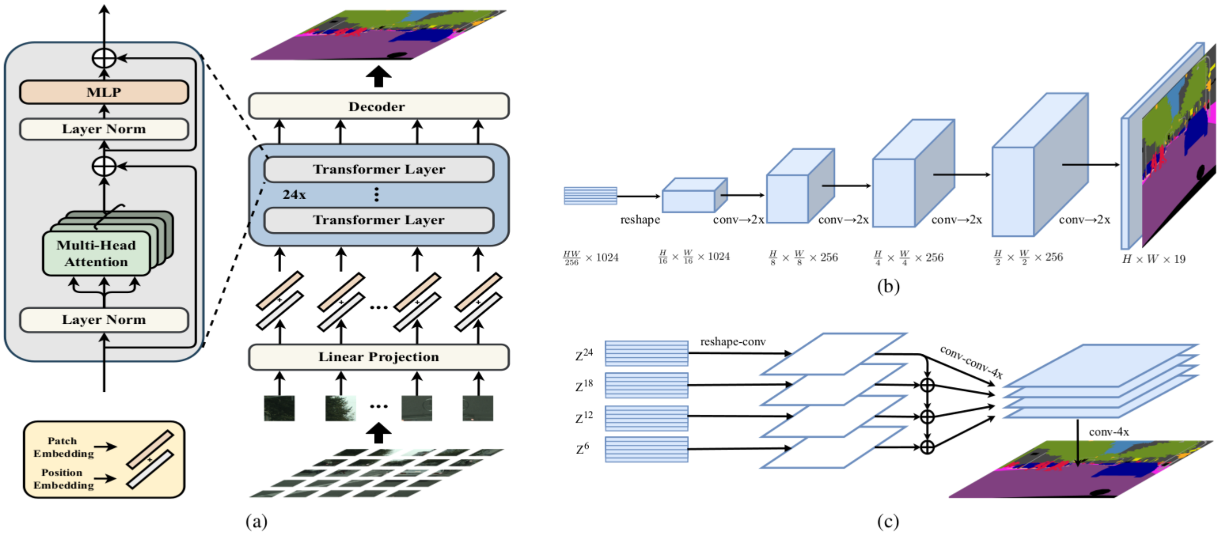


Figure 3: SETR Architecture

As for decoding, authors discuss three strategies which are as follows:

- **Naïve upsampling:** features from transformer encoder are transformer using a 2-layer network and upsampled using bilinear upsampling.
- **Progressive UPsampling (PUP):** authors use a progressive upsampling strategy that alternates conv layers and upsampling operations (figure 2(b)).
- **Multi-Level feature Aggregation (MLA):** here authors use a concept similar to feature pyramid network but modify it to incorporate transformer architecture as all transformer layers share same resolution feature representation.

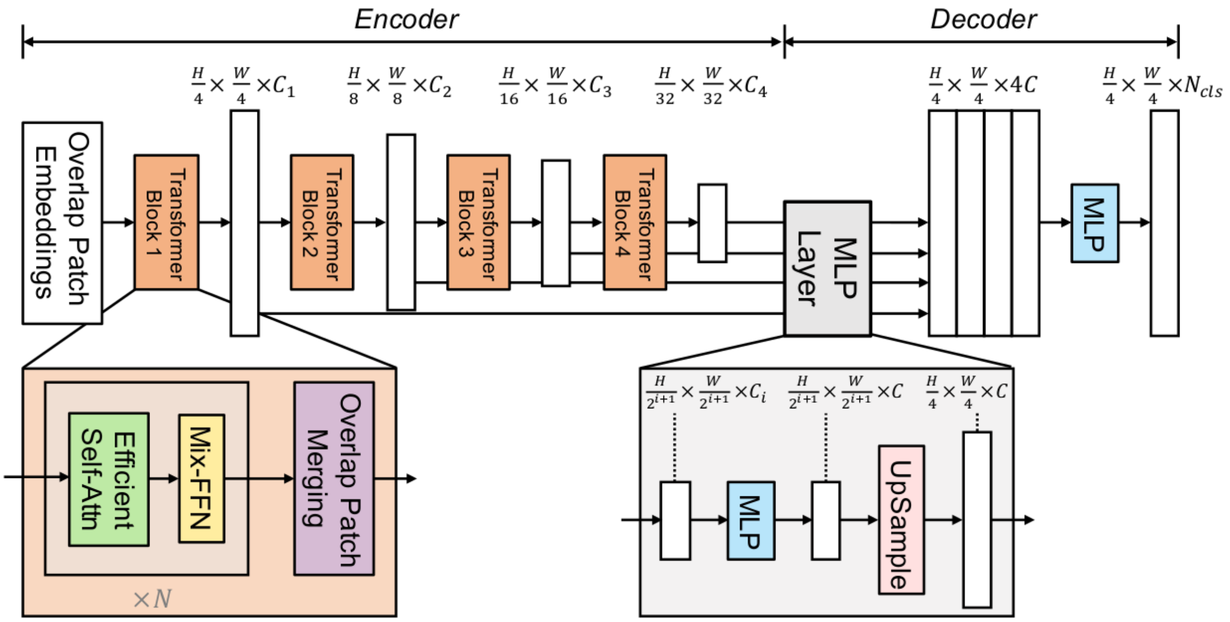
SegFormer:

SegFormer builds upon SETR [2] paper and propose improvements to make it more efficient, accurate and robust. SETR adopts ViT as a backbone and incorporates several CNN decoders to enlarge feature resolution. Despite the good performance, ViT has some limitations:

1. ViT outputs single-scale low resolution features instead of multi-scale ones.
2. It has high computation cost on large images.

People proposed new architectures [12]–[14]to mitigate these problems but they all mainly focused on the encoder part, neglecting the contribution of decoder. Main contribution of this paper are as follows (figure 4):

- A novel positional-encoding-free and hierarchical Transformer encoder.
- A lightweight All-MLP decoder design that yields a powerful representation without complex and computationally demanding modules.



Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
$\times N$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

Figure 4: SegFormer Architecture

Due to positional-encoding-free encoder, it can easily adapt to arbitrary test resolutions without impacting the performance. The hierarchical part enables the encoder to generate both high-resolution fine features and low-resolution coarse features, this is in contrast to ViT that can only produce single low-resolution feature maps with fixed resolutions. Light weight decoder takes advantage of both local and global attention by aggregating the information from different layers, the MLP decoder combines both local and global attention.

Part2:

For this project, I used Fast-SCNN for my comparison studies. Input of this model is the four channel MRI image and the network outputs a segmentation map. For experimentation I used the official implementation of this paper which can be found through this link (https://github.com/Tramac/Fast-

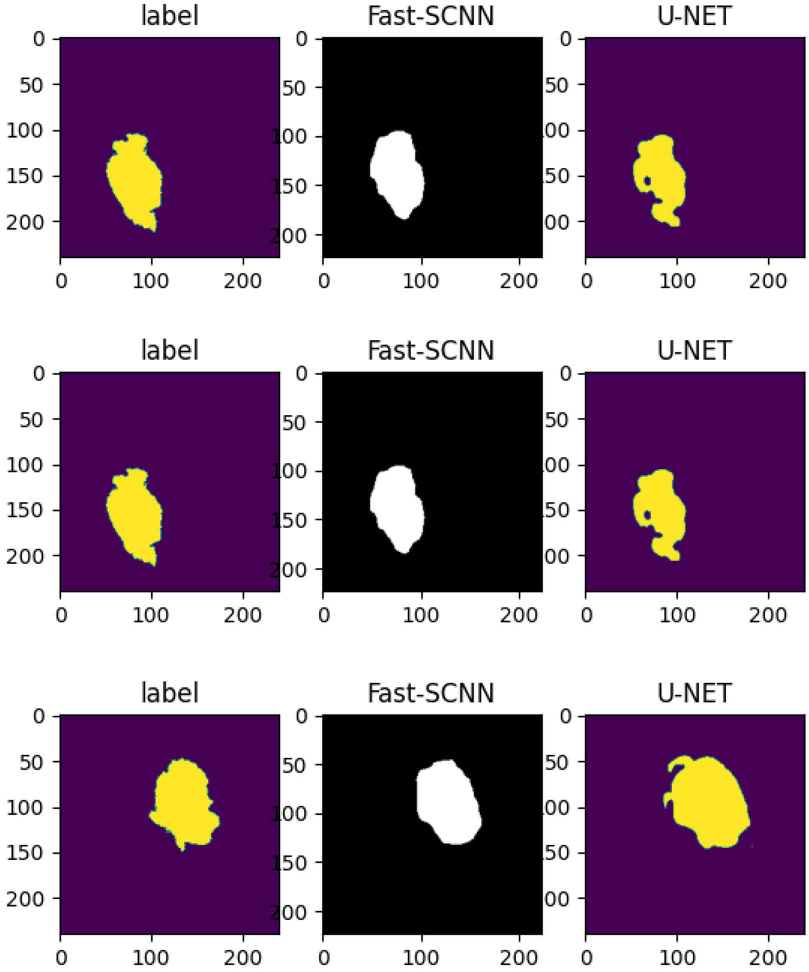
SCNN-pytorch). Fast-SCNN was trained on whole training dataset and dice score was calculated using all of the test dataset.

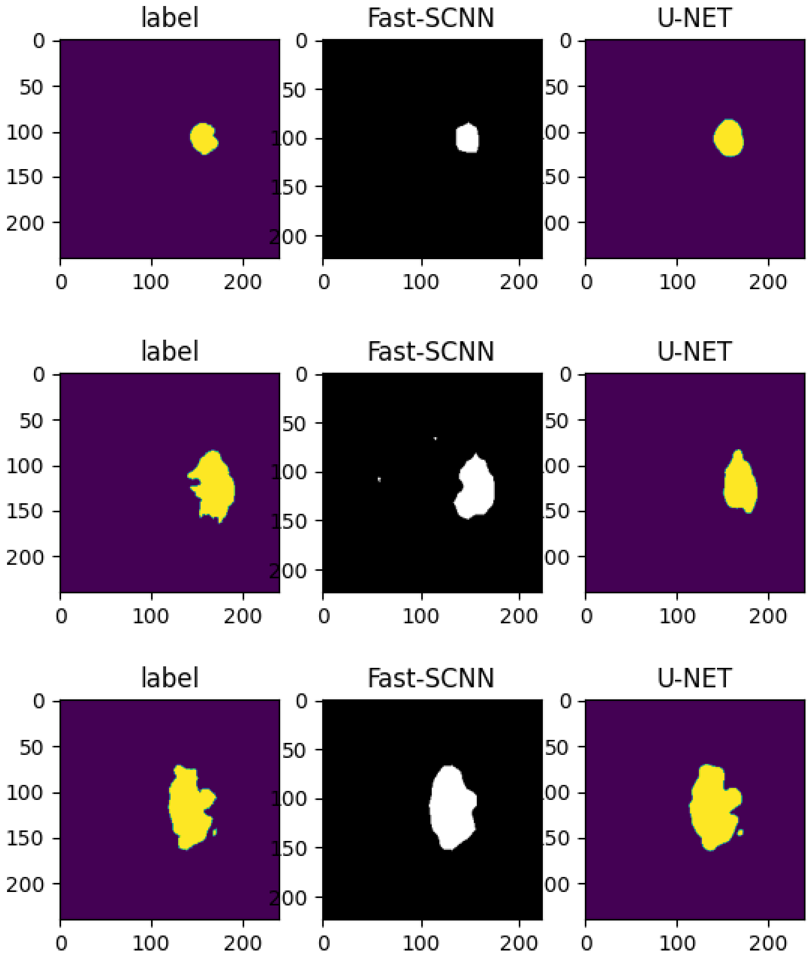
Model was trained for 5 epochs due to limited data. Image data was resized to the size of (224x224) to make the local and global features of same size. SGD optimizer with momentum was used along with learning rate schedular. A customized SoftMax cross-entropy loss was used for training.

Results:

Compared to simple U-Net model, Fast-SCNN outperforms it by a healthy margin. In my testing Fast-SCNN on average got 5% higher dice score.

Model	Fast-SCNN	U-NET
Dice Score	**0.85**	0.78





References:

[1] R. P. K. Poudel, S. Liwicki, and R. Cipolla, “Fast-SCNN: Fast Semantic Segmentation Network,” Feb. 2019.

[2] S. Zheng *et al.*, “Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers,” Dec. 2020.

[3] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” May 2021.

[4] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” May 2016.

[5] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, “ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-time,” May 2018.

[6] D. Mazzini, “Guided Upsampling Network for Real-Time Semantic Segmentation,” Jul. 2018.

[7] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017.

[8] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large Kernel Matters -- Improve Semantic Segmentation by Global Convolutional Network,” Mar. 2017.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” Jun. 2016.

[10] A. Vaswani *et al.*, “Attention Is All You Need,” Jun. 2017.

[11] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Oct. 2020.

[12] X. Chu *et al.*, “Twins: Revisiting the Design of Spatial Attention in Vision Transformers,” Apr. 2021.

[13] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” Mar. 2021.

[14] W. Wang *et al.*, “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions,” Feb. 2021.