TransUNet: Transformers Make strong Encoders for Medical Image Segmentation

Problem they are trying to solve / Purpose of method

In medical image segmentation UNet was/is the defacto choice. UNet has limitations in long-range relations due to the locality of convolution operations. In other words, it struggles to capture global context. TransUNet is the propoposed solution to this, by adding a Transformer in the UNet architecture.

In medical image segmentation it is important to get as detailed and accurate predictions as possible, since they are used in diagnosis or other medical applications.

How does it differ from other methods?

Compared to UNet, TransUNet introduces a Transformer in the decoder. Otherwise they are very similar.

Other methods apply self-attention to CNNs, or use only Transformers to capture global context.

TransUNet combines both CNN and Transformer to get the benefits of both.

How the method works

TransUNet, similar to UNet, consists of a encoder decoder architecture. The encoder is made up of a CNN followed by a Transformer. Throughout the decoder, the input image is downscaled.

In the decoder we upsample and combine with skip connections from the CNN in the encoder.

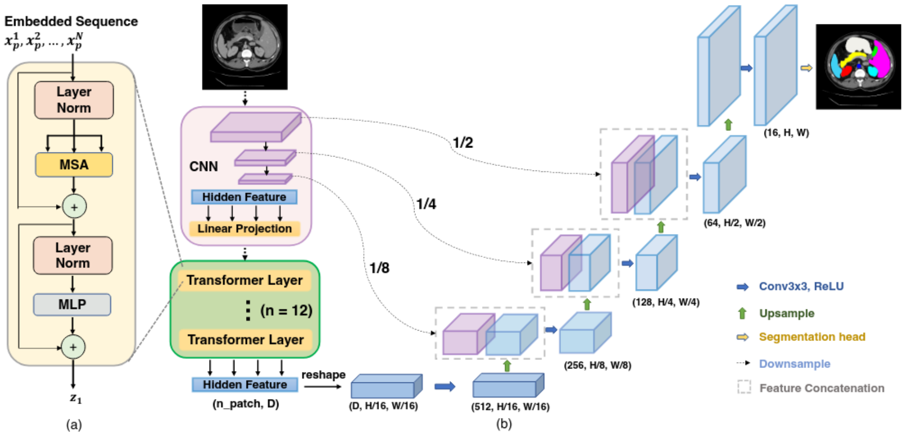

Figure 1: TransUNet architecture