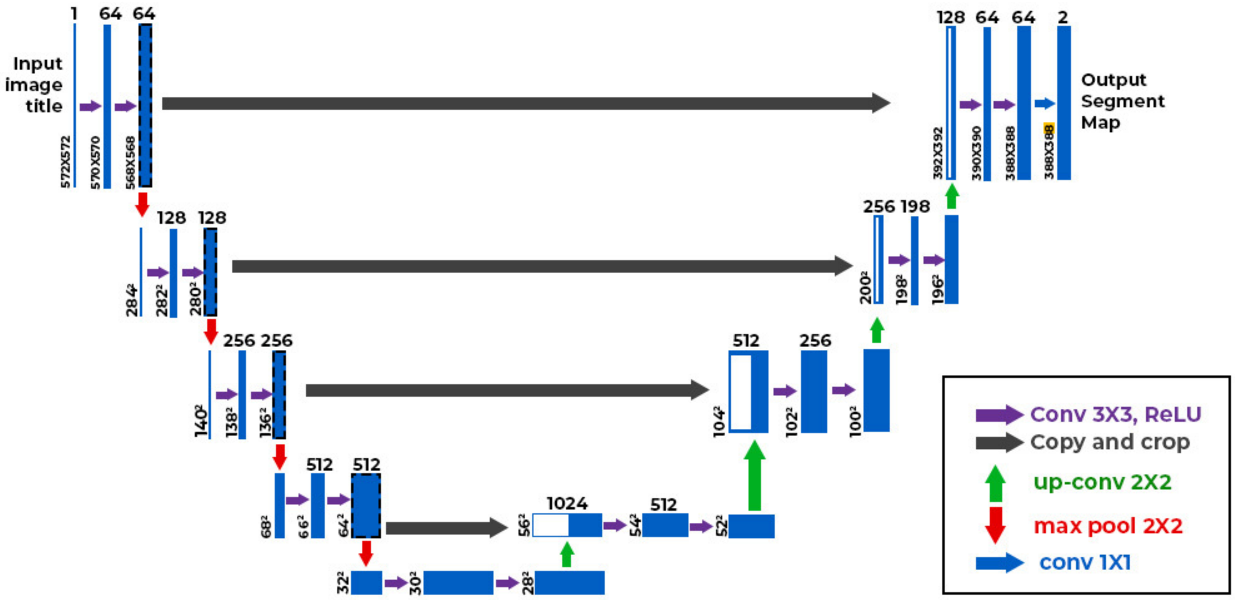
TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

Introduction

In this blog post, we'll explore **TransUNet**, an innovative architecture that enhances the well-known **U-Net** model by integrating a transformer encoder, specifically the Vision Transformer (**ViT**). We will start by understanding the basic structure of U-Net and the improvements made through these adjustments. By breaking down the architecture and explaining key concepts, I aim to provide a clear and comprehensive understanding of how TransUNet works and why it represents a significant advancement in medical image segmentation.

U-Net

In this diagram, we see the **U-Net architecture**, which is designed for **biomedical image segmentation**. This model is known for its distinctive "U" shape, resulting from an **encoder-decoder setup** with **skip connections** that help preserve important details.



2D UNet Architecture

Encoder:

The left side of the U-Net is the encoder, which utilizes a convolutional Neural Network (CNN) similar to those used in image classification. It works by progressively reducing the **spatial dimensions** of the input image (making it smaller in size) while increasing the **depth** (number of channels). This process outputs a **feature map** that captures the information extracted from the image. Each step down in the encoder involves:

- A **convolution operation** (purple arrows), which extracts features from the image.
- A **ReLU** activation, which helps the model to learn non-linear features.
- A **max-pooling** operation (red arrows), which reduces the dimensionality of the feature maps by selecting the most prominent features of the previous one.

Decoder:

The right side is the decoder, which performs the opposite function of the encoder. It progressively restores the spatial dimensions to match the size of the original image, aiming to pinpoint the exact location of objects. Each step up in the decoder includes:

- An **up-convolution operation** (green arrows), which enlarges the feature maps.
- A **concatenation** with the corresponding encoder feature map (gray arrows). This step is crucial and is facilitated by the **skip connections** that directly bring features from the encoder to the decoder. These connections help the model **remember precise details about the image structure, which might be lost during the encoding process.**

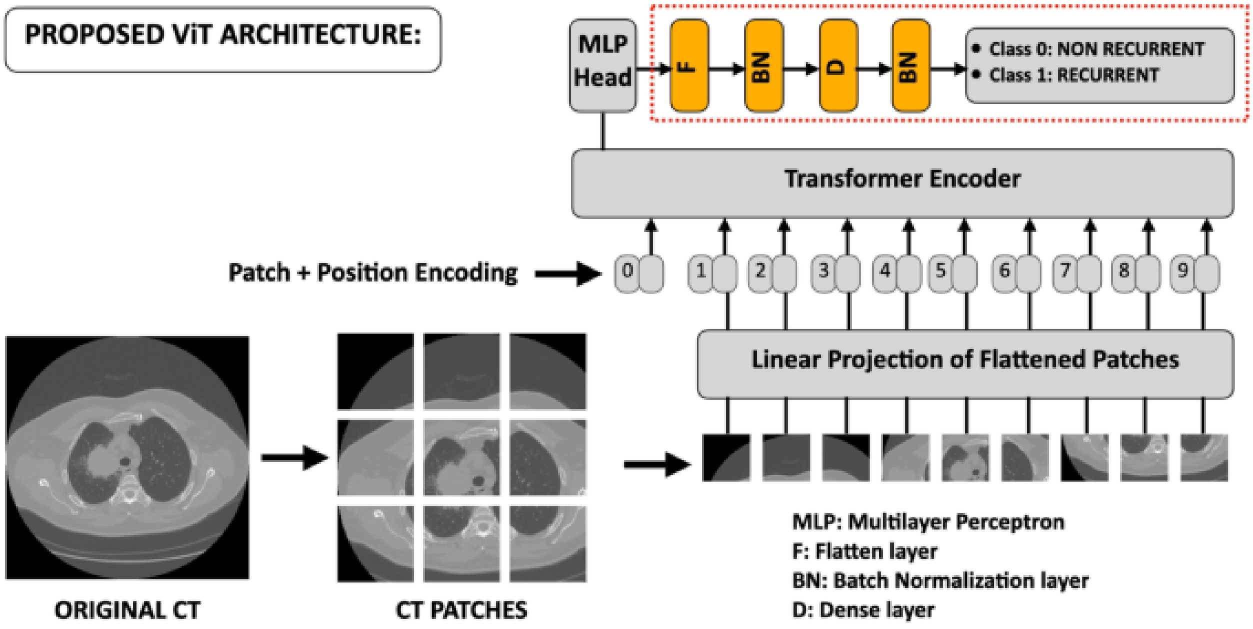
At the end, a **convolution** converts the feature map into the final **segmentation map**, which classifies each pixel of the input image into different categories based on the model’s training.

Limitations:

While U-Net is powerful for many tasks, it has limitations. It struggles with modeling **long-range dependencies** since it focuses predominantly on local details and may lose global information. Additionally, its performance can be weak for target structures with large **inter-patient variations**, such as in medical images where tumors may differ significantly in texture, shape, and size across different patients.

Transformers

Originally designed for natural language processing (NLP), transformers have been adapted for image processing as **Vision Transformers (ViT)**. ViT adapts the transformer architecture for visual data by treating images as sequences of pixels or small image patches.


PROPOSED ViT ARCHITECTURE:
MLP Head
F
BN
D
BN
Class 0: NON RECURRENT
Class 1: RECURRENT
Transformer Encoder
Patch + Position Encoding
0
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
ORIGINAL CT
CT PATCHES
MLP: Multilayer Perceptron
F: Flatten layer
BN: Batch Normalization layer
D: Dense layer

ViT Architecture

Key Components of ViT:

1- Self-Attention Mechanisms:

- The core strength of ViT is its self-attention mechanism, which dynamically weighs the importance of various parts of an image. This enables ViT to capture the “global context,” highlighting key features necessary for specific tasks.

2- MSA (Multihead Self-Attention):

- ViT uses **MSA** to focus on multiple image parts simultaneously. By dividing the input into several “heads,” it processes different aspects in parallel, thus capturing a broader range of information like feature types and spatial relationships.

3- MLP (Multilayer Perceptron):

- Post self-attention, ViT processes features through an MLP, which consists of fully connected layers capable of learning non-linear feature combinations. The MLP integrates information across attention heads, enhancing the model’s ability to discern complex patterns and relationships.

4- Transferability:

- A major advantage of ViT is its transferability, allowing a model trained on a large dataset to be effectively adapted to new, related tasks with minimal adjustment, thanks to its rich, generalized feature representations.

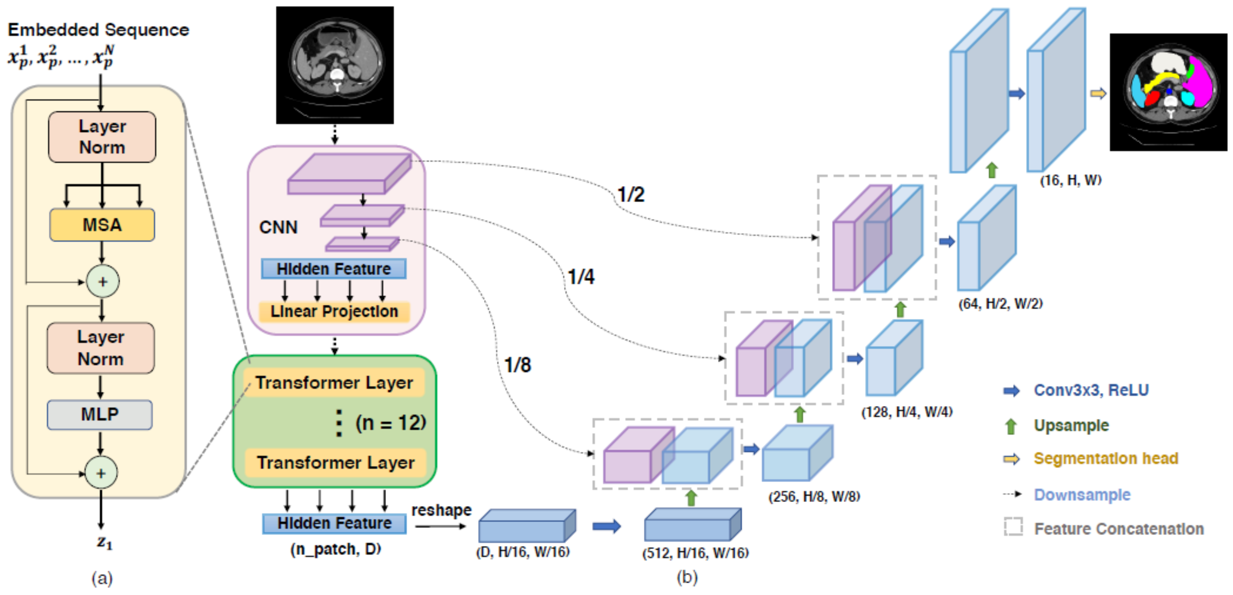
Limitations and Integration with CNNs:

Despite its strengths, ViT struggles with **localization**, often failing to pinpoint exact object locations within an image due to its 1D sequence processing approach. This leads to **lower resolution outputs** when used solely with CNN decoders.

The TransUNet architecture addresses this by **combining ViT’s global focus with the local precision of CNNs.**

TransUnet

TransUNet is a hybrid architecture that combines CNN and Transformer technologies to leverage both high-resolution spatial information from CNNs and global contextual insights from Vision Transformers (ViTs). This integration allows for precise image segmentation, particularly useful in medical imaging.



TransUNet Architecture

Architecture Components:

- **Encoder:** The encoder features a dual approach, starting with a CNN encoder that processes the input image into a feature map. This feature map is then **tokenized into patches**, which are processed by a Transformer encoder. This setup captures both local details and global patterns effectively.
- **Decoder:** The decoder uses the features encoded by the Transformer and **upsamples** them, integrating them back with the high-resolution feature maps from the CNN. This step is crucial for precise localization of the segmentation areas.

Detailed Encoder Process:

1. **Initial CNN Processing:** The input image is first processed by a CNN, which generates a **feature map** of dimensions H×W. This feature map captures localized features from the image, providing a rich basis for further processing.

2. Tokenization:

Thefeature map is divided into patches of size P×P. Each patch is then flattened into a 1D sequence, resulting in an array of N tokens, where

$$N = \frac{H * W}{P * P}$$

The tokens represent parts of the original image but in a format suitable for transformer processing. Each token from the patches is denoted as

$$x_i^p \in \mathbb{R}^{P^2 * C}, i \in [1, N]$$

indicating its position in the sequence.

3. Patch Embedding:

Each token x_{i,p} is transformed into a D-dimensional embedding space using a **trainable linear projection**, commonly referred to as the embedding matrix

$$E \in \mathbb{R}^{P^2 * C * D}$$

essentially reshaping and projecting the input tokens into a higher-dimensional space. Positional embeddings Epos are added to these embeddings to incorporate the spatial relationships between different patches, forming the initial sequence for the transformer:

$$E_{pos} in R^{N * D}$$

$$z0 = [x_{1p} * E, x_{2p} * E, ...] + E_{pos}$$

This sequence z0 now represents the image in a form that preserves both content and positional context, facilitating the global contextual processing by the transformer.

4. Transformer Processing:

The sequence of embedded patches z0 is fed into the transformer's layers, each comprising Multihead Self-Attention (MSA) and Multilayer Perceptron (MLP) blocks:

- **MSA Layer:** At each layer L , the sequence undergoes self-attention where the model computes attention scores between all patches, allowing it to focus on different parts of the image based on their relevance to the segmentation task. The output of the MSA layer for each token is a combination of its own features and features from other tokens, weighted by their calculated attention scores:

$$zl' = MSA(NL(z_{l-1})) + z_{l-1}$$

- **MLP Layer:** Following the self-attention, each token's features are passed through an MLP, which can further refine features by applying nonlinear transformations:

$$zl = MLP(NL(zl')) + zl'$$

Detailed Decoder Process:

The paper discusses the possibility of using a 'naive' approach with the decoder by leveraging Bilinear Upsampling and shows that this **leads to loss of low level details**. Thus cascade upsampler (CUP) is used in TransUNet.

Cascade Upsampler (CUP):

con sists of multiple upsampling steps to decode the hidden feature for outputting the final segmentation mask. After reshaping the sequence of hidden feature we instantiate CUP by cascading multiple upsampling blocks for reaching the full resolution from H*P/ W*P to H*W. where each block consists of:

- Two upsampling operators.
- A 3×33 \times 33×3 convolutional layer.
- ReLU activation function.

Parameters Used

The study strategically selects specific parameter settings to optimize the TransUNet's performance:

- **Default Image Size:** 224×224 pixels
- **Patch Size:** 16
- **Batch Size:** 24

Why these choices?

These parameters are chosen based on their ability to balance detailed image processing with computational efficiency. However, you might wonder about **the implications of altering these parameters**.

Impact of Configuration Variations:

1- Skip Connections:

Increasing the number of skip connections improves model performance. This enhancement occurs because more connections allow better integration of local details and global context.

Optimal results are achieved by inserting skip connections at all three intermediate upsampling stages of the Cascade Upsampler (CUP), excluding the output layer.

This can be improved by using **additive transformers** but due to their computational cost they weren't adopted in TransUNet architecture.

2- Input Resolution:

Using higher resolutions betters the details available for segmentation, which can lead to more accurate outcomes. However, this advantage comes with **increased computational demands** and **greater resource consumption**, necessitating a careful balance.

3- Patch Size:

Reducing patch size **increases the number of patches** and, consequently, the **number of connections** within the model, which enhances its capability to capture complex details. This leads to improved segmentation accuracy.

4- Model Scaling:

Enlarging the model—by adding more layers or increasing parameters—generally results in better performance. However Larger configurations demand significantly more computational resources, making them less feasible for certain applications.

TransUNet is available in both Base and Large configurations to cater to different computational capabilities.

Conclusion

The TransUNet architecture predominantly processes images in 2D, managing each image slice independently before synthesizing the segmented outputs. This

2D approach, while efficient, may lead to information loss—especially noticeable in applications like tumor segmentation, where continuity between slices can indicate the presence of tumors.

To address this limitation, a 3D handling approach has been considered, which processes data in sub-volumes or patches. This method not only preserves spatial continuity but also manages computational costs effectively. The development and details of the 3D TransUNet architecture will be explored in future blog posts.

For a deeper understanding and further insights, including comparisons with R50 Unet, AttnUnet, and ViT-CUP, the full article is available for those interested. It provides extensive results and discussions that enhance our understanding of TransUNet's capabilities. I highly encourage all readers to delve into the article for a comprehensive exploration of these innovative segmentation solutions.

References:

U-Net: Convolutional Networks for Biomedical Image Segmentation

Architecture U-Net

Comparison between vision transformers and convolutional neural networks to predict non-small lung cancer recurrence

TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

Attention gated networks: Learning to leverage salient regions in medical images

An Image is worth 16*16 words: Transformers for Image Recognition at Scale

3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers