LinTUNet

A Hybrid CNN-Transformer

Architecture for Medical Image Segmentation

PACISE Conference 2025

Created by: Lawrence Menegus

Overview

- **Deep Learning Basics:** Overview of CNNs, Autoencoders, and U-Net for medical images.

- **Transformers & Attention Layers:** How Transformers and attention layers work, plus an intro to Linformer and Sparse attention layer.

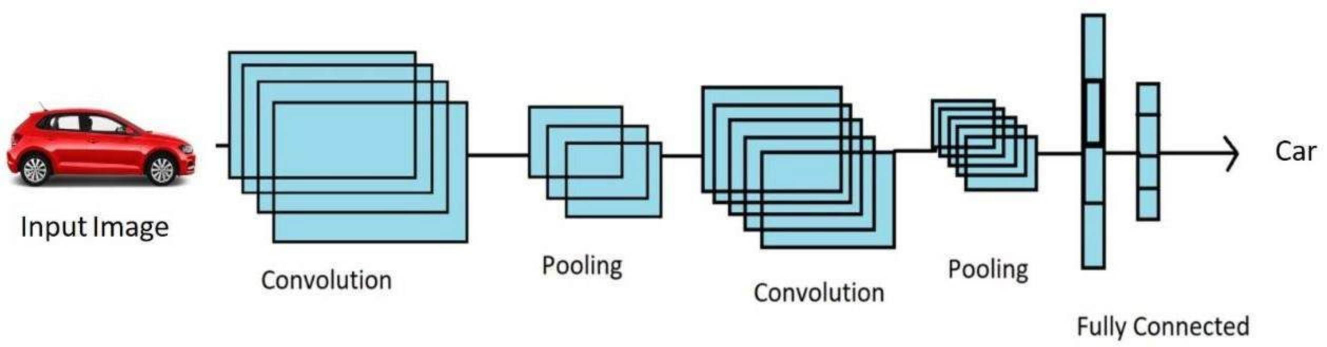
- **LinTUNet Model:** Combining Linformer with U-Net for better image segmentation.
  - **Performance & Metrics:** Evaluating results using IoU and other measures.
  - **Future Impact:** How LinTUNet can improve medical imaging.

Introduction to CNNS (Convolutional Neural Networks)

- **Convolutional Neural Networks (CNN)** uses filters to detect patterns like edges, textures, and shapes in images, making it effective for visual tasks.
  - Basically it analyzes an image by processing it **pixel by pixe**l using filters to detect patterns

- **Lower layers capture simple features** (edges, colors), while deeper layers learn complex patterns (faces, objects).

- **CNN**s are essential for image classification, object detection, facial recognition, and medical image analysis.

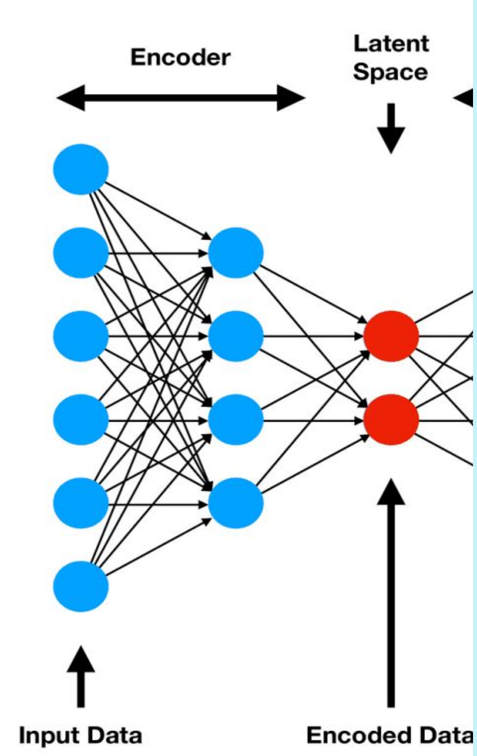


Introduction to Autoencoders (Encoder)

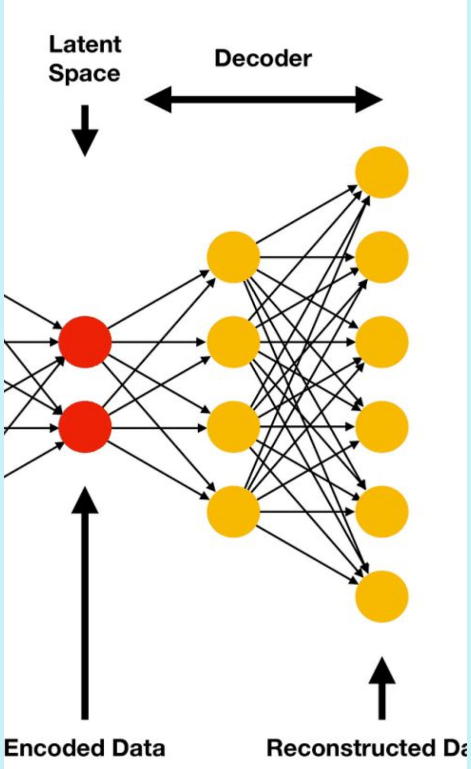
- An autoencoder is a neural network that is broken in **two parts**
    - **Encoder-** compresses input data into a lower-dimensional representation
    - **Decoder** - reconstructs it back to the original form

Encoder -

- **Compresses** the input data into a smaller representation by capturing its most important features while removing unnecessary details.
- **Reduces** the input size to a compact latent space, making it easier for the decoder to reconstruct the original data while preserving key patterns.



Introduction to Autoencoders (Decoder)

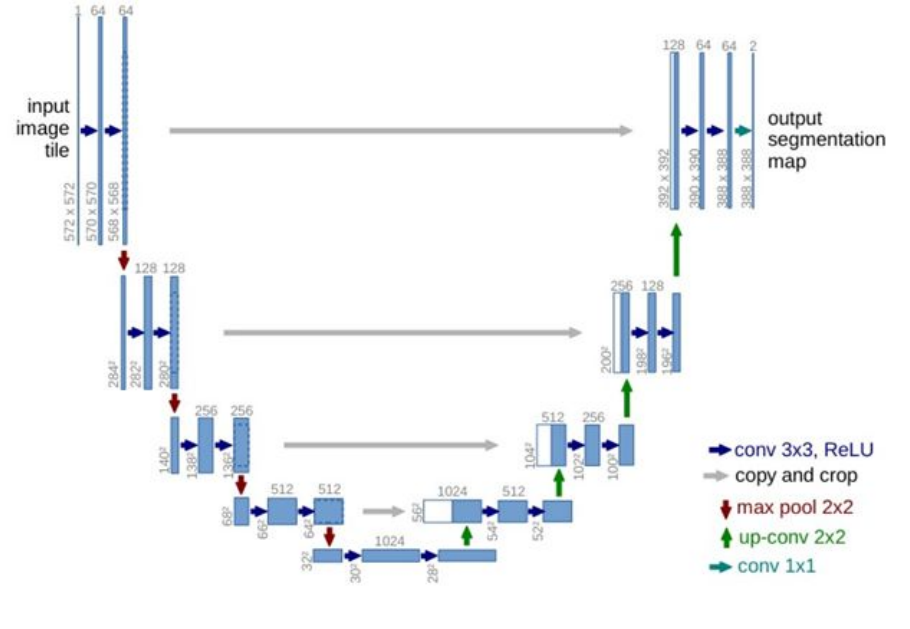


DECODER -

- It takes the compressed representation from the encoder **and transforms it back into the original input format,** aiming to recreate the data as accurately as possible.

- It uses techniques **like upsampling and transposed convolutions to gradually rebuild the data,** refining details as it reconstructs the input from the lower-dimensional representation

Understanding U-Net

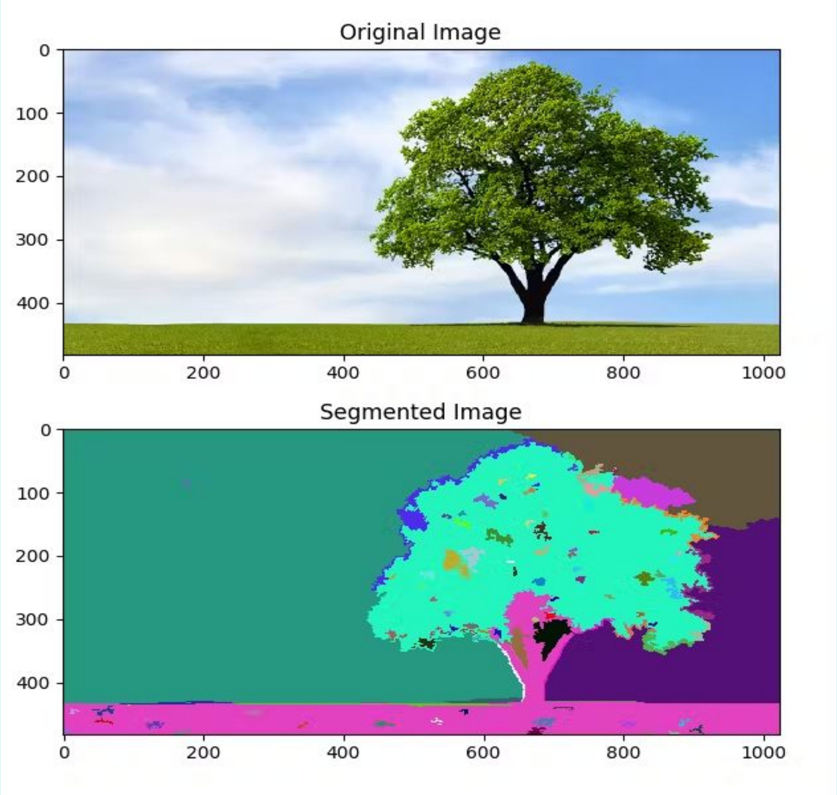


- **U-Net is an CNN-based architecture which mimics as autoencoder,** where the encoder extracts features, and the decoder reconstructs a segmented output, making it effective for pixel-wise predictions.

- **Introduced by Ronneberger et al. in 2015**, in the publication “U-Net: Convolutional Networks for Biomedical Image Segmentation”

- **Named after its U-shaped architecture**, U-Net features **skip connections** that directly link encoder and decoder layers at corresponding levels, preserving spatial information and improving segmentation accuracy.

Image Segmentation



- **Image segmentation** is the process of breaking an image into meaningful parts or regions to make it easier to analyze.

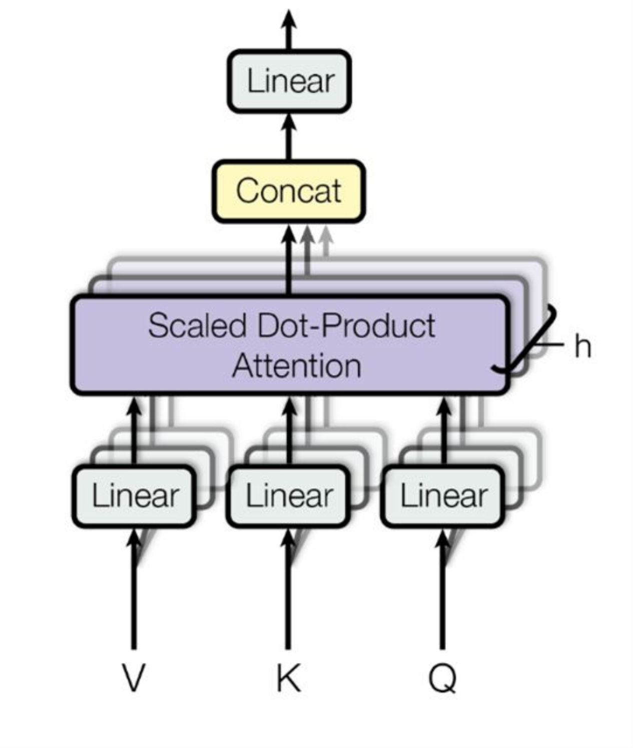
- It helps in identifying and separating objects (**Object Detection**) within an image, commonly used in medical imaging, self-driving cars, and facial recognition.

What is a Transformer in Deep Learning?

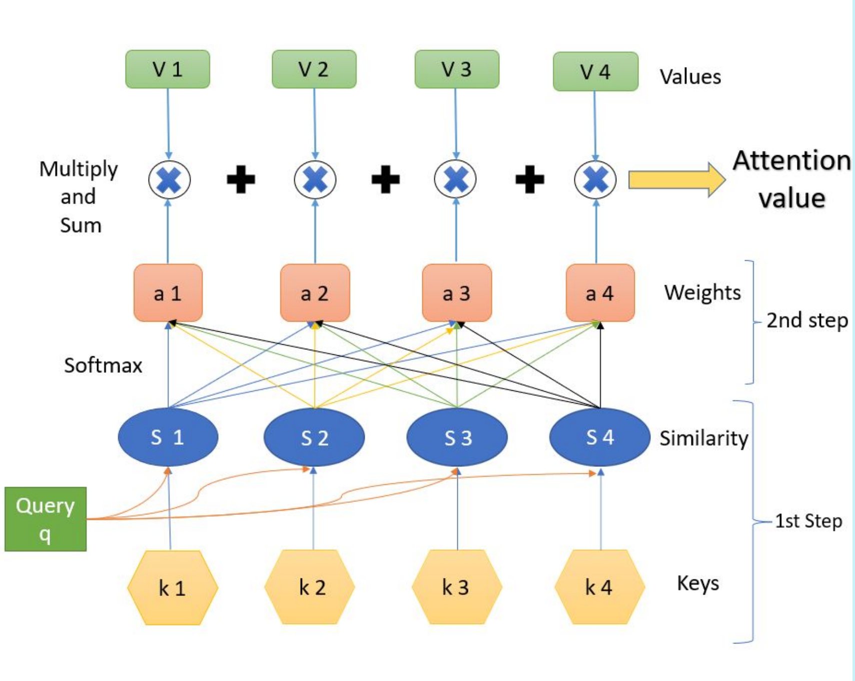
- **Transformers** are deep learning models that handles entire sequences of data at once, making it faster than traditional models like RNNs.

- **It focuses** on the most important parts of the input, improving tasks like translation, text generation, and image processing.

- Transformers are **the foundation of advanced** AI models like ChatGPT, BERT, etc.



The Role of the Attention Layer in Transformers



- **Attention layers** helps the model highlight relevant parts of the input instead of treating everything equally.

- **It uses these vectors to calculate attention scores,** determining how much focus each input should get.
  - Higher scores **give more influence to important inputs**, improving the model's understanding of context.

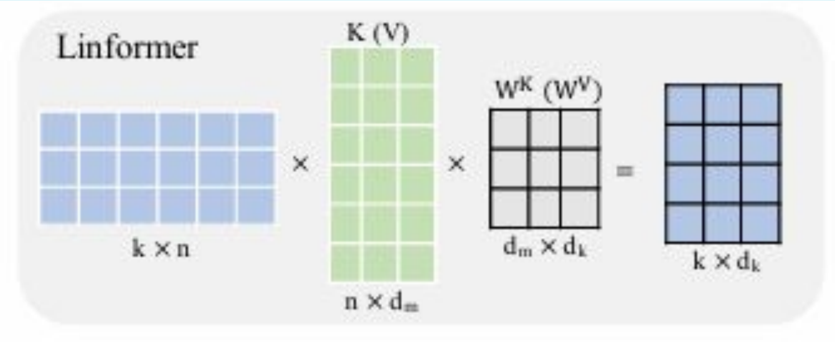
- This enables **the model to understand connections** between distant elements, improving performance in NLP and vision tasks.

Introducing Linformer: A Efficient Transformer

- Linformer was introduced by researchers at Facebook AI, including **Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma in 2020.**

- **Linformer** reduces memory and computation by **projecting Key and Value matrices into a lower-dimensional space** before computing attention.

- Traditional **Transformers with O(n²) complexity, Linformer achieves O(n) complexity,** making it more efficient for long sequences.

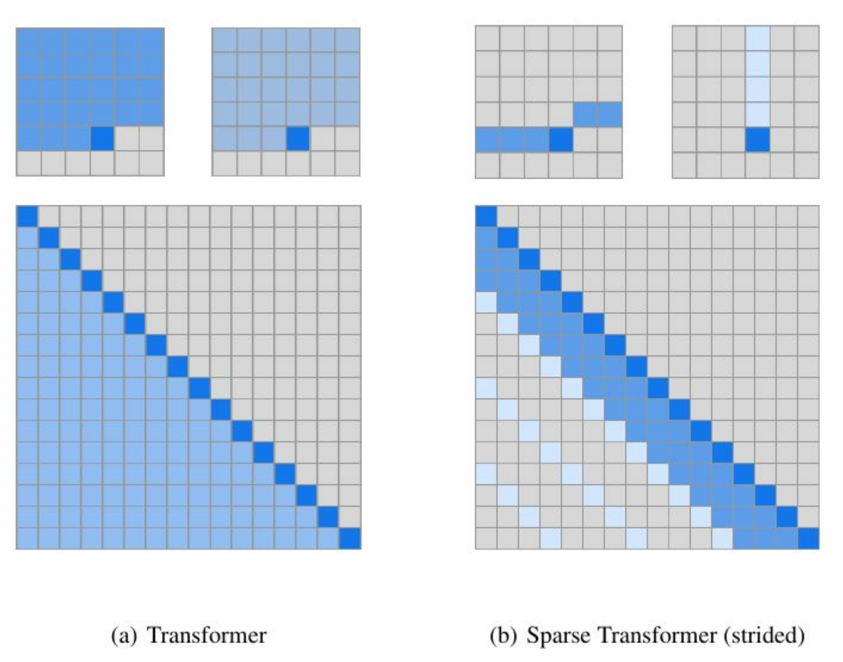


Sparse Attention Mechanism

- The Sparse Attention Mechanism was introduced by **researchers at Google Research, including Rewon Child, Scott Gray, and others in 2019**

- **Sparse attention attends to only a subset of key-value pairs,** reducing memory and computation costs.
  - **Uses predefined attention patterns** like **local windows** or **strided attention** to capture important dependencies efficiently.
  - Reduces operations needed for attention, making it more scalable for long sequences and large datasets.

Sparse Attention Scheme



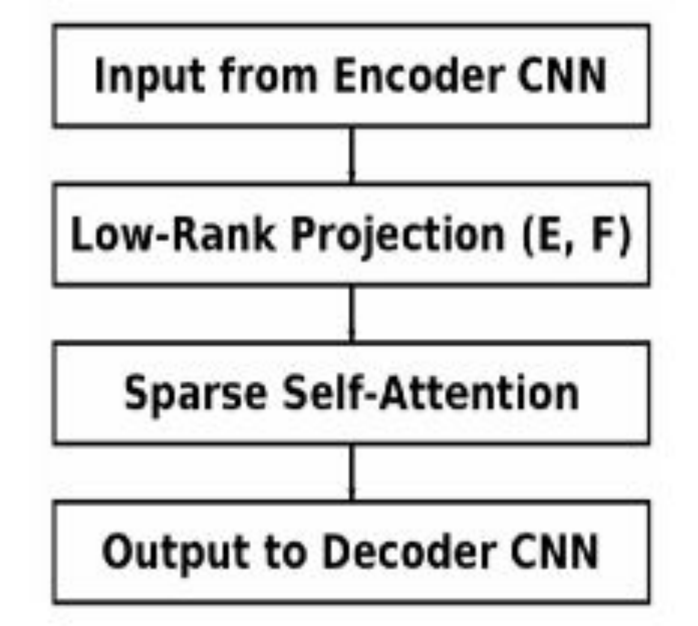
(a) Transformer

(b) Sparse Transformer (strided)

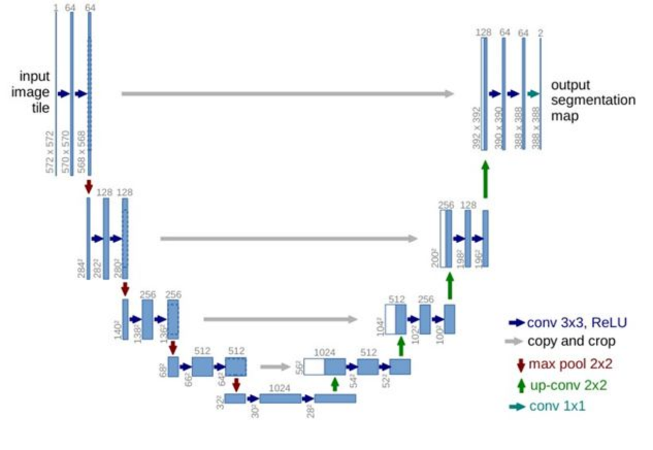
Integrating into U-Net: The Architecture

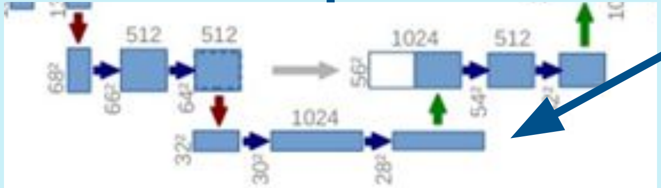
- **LinTUNet retains U-Net’s encoder-decoder structure but enhances it with Linformer and Sparse Attention** in the bottleneck for efficient computations and better feature representation.

- **Sparse Attention reduces connections**, improving efficiency, while **Linformer enables scalable attention**, enhancing long-range dependency handling for tasks like medical image segmentation.

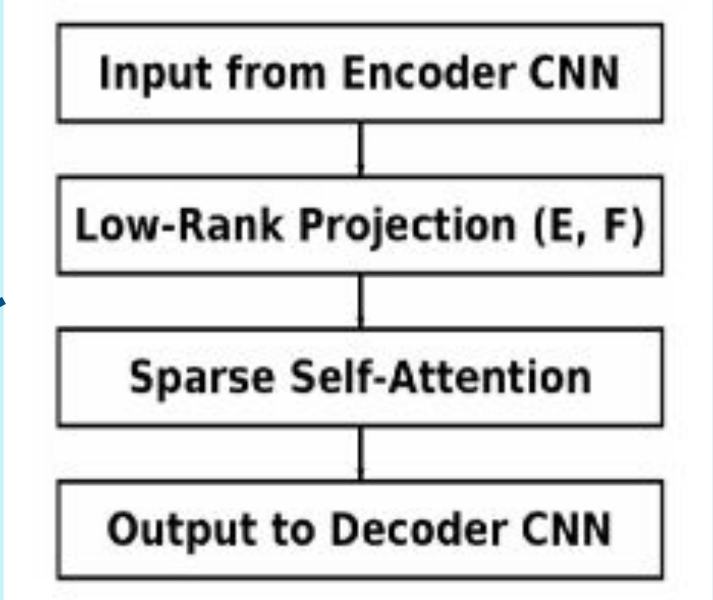


Integrating into U-Net: The Architecture (cont)





Bottleneck



Input from Encoder CNN

Low-Rank Projection (E, F)

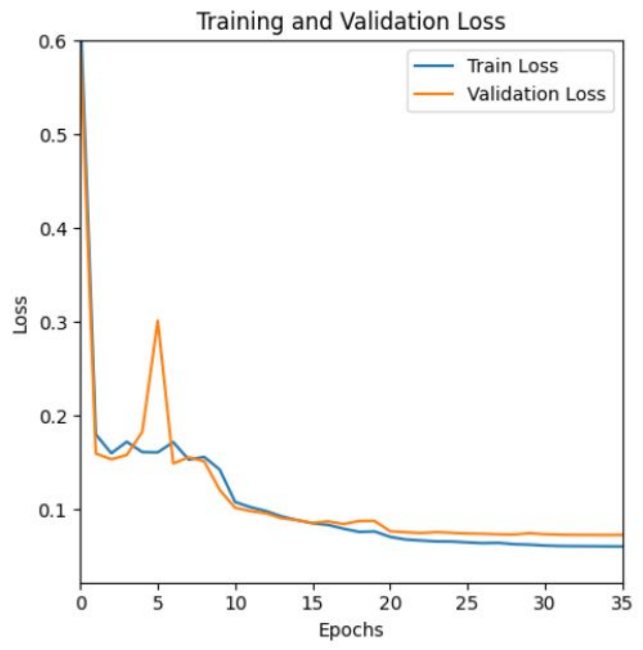
Sparse Self-Attention

Output to Decoder CNN

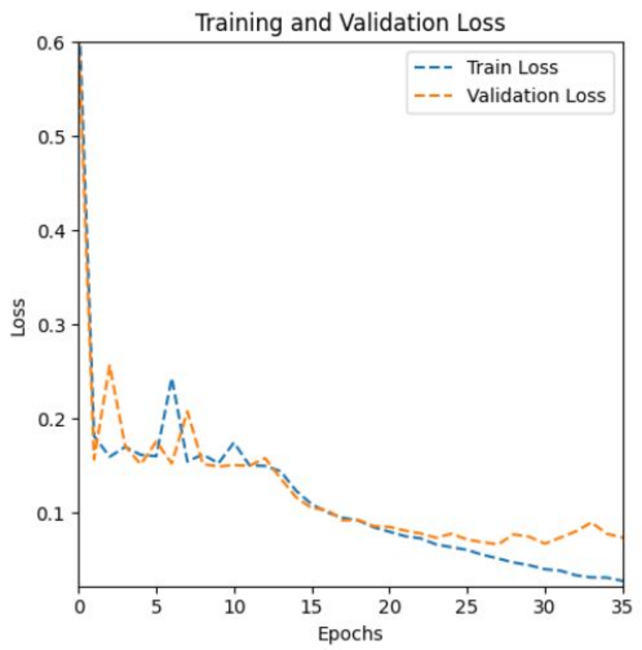
Performance Evaluation : Loss

- The following results are from one of several trial runs in my study, comparing the performance between the traditional U-Net CNN and my proposed model, **LinTUNet**, a hybrid CNN-Transformer for image segmentation.

U-Net

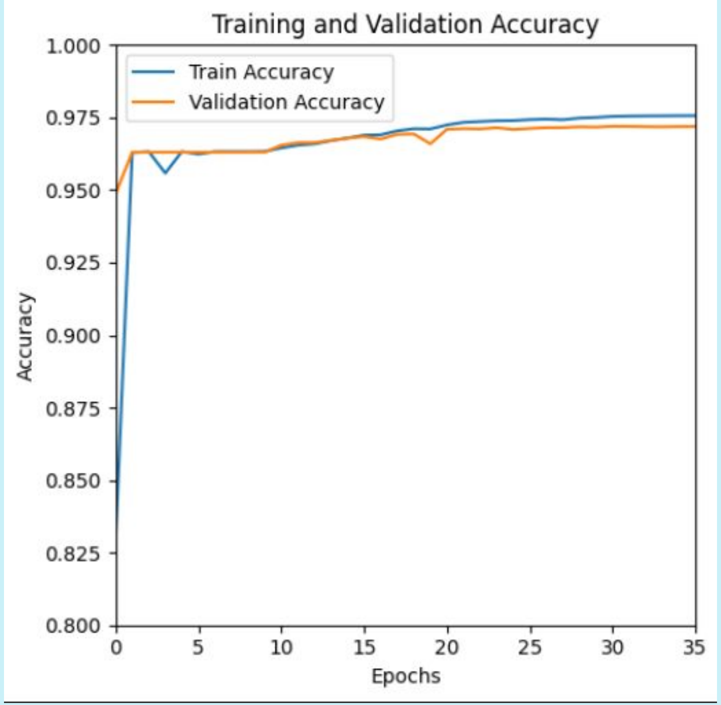


LinTUNet

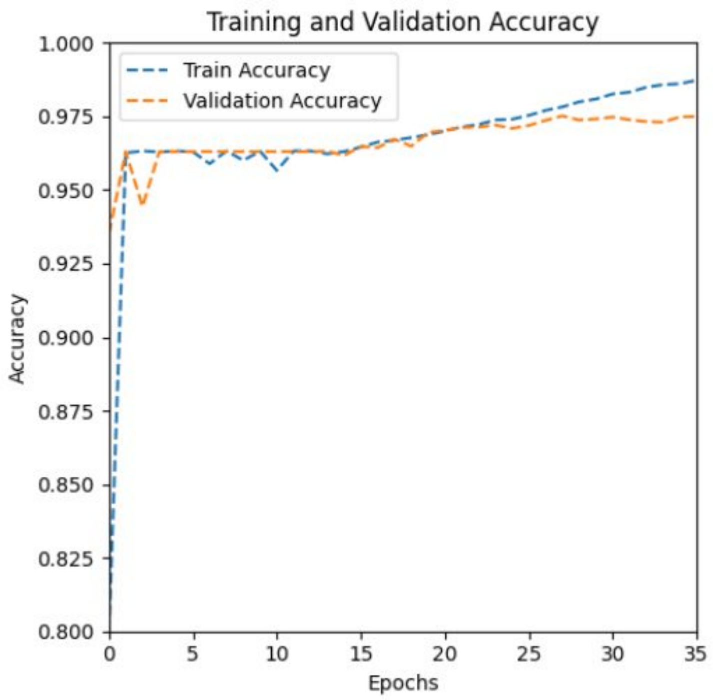


Performance Evaluation: Accuracy

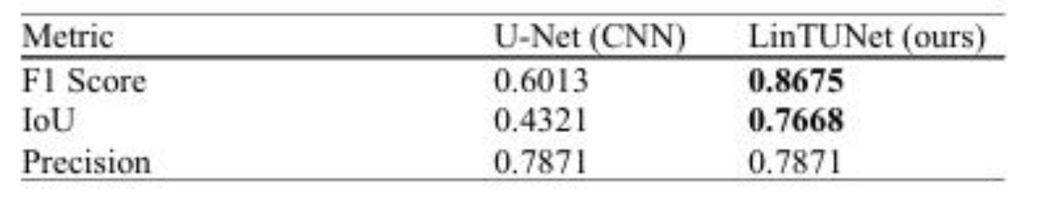
U-Net



LinTUNet



Performance Evaluation Metrics



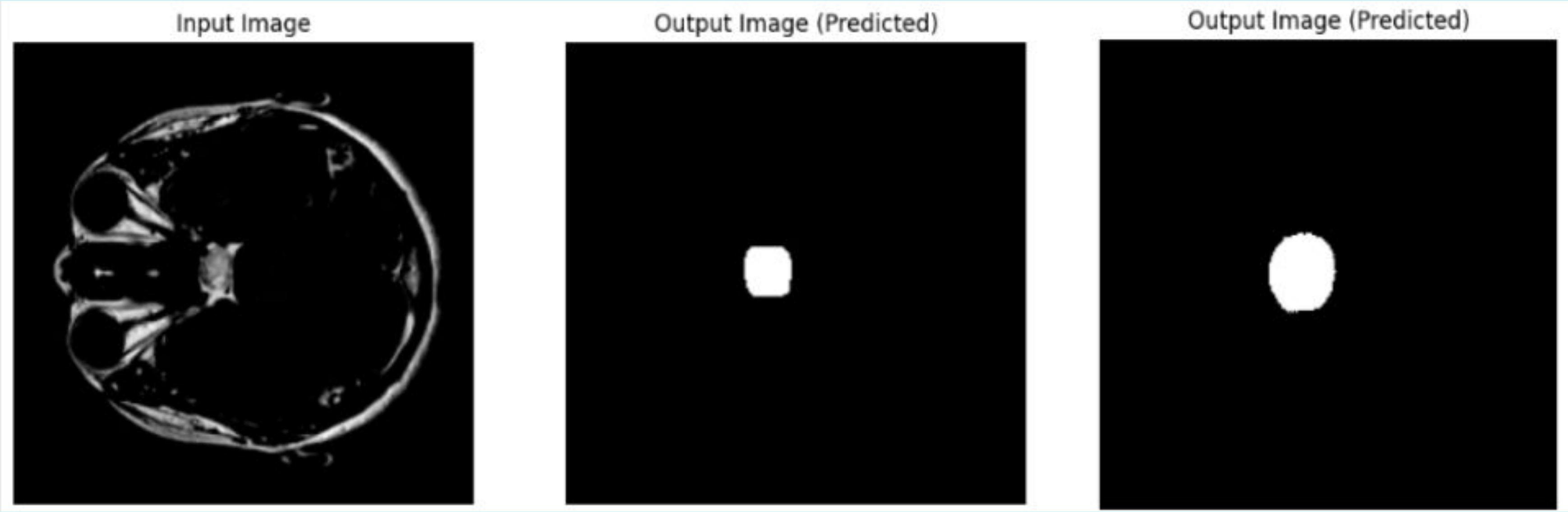
Metric	U-Net (CNN)	LinTUNet (ours)
F1 Score	0.6013	**0.8675**
IoU	0.4321	**0.7668**
Precision	0.7871	0.7871

Intersection over Union (IoU)

- **Intersection over Union (IoU)** is a metric used to evaluate the performance of image segmentation models. It measures the overlap between two sets: the predicted segmentation and the ground truth.

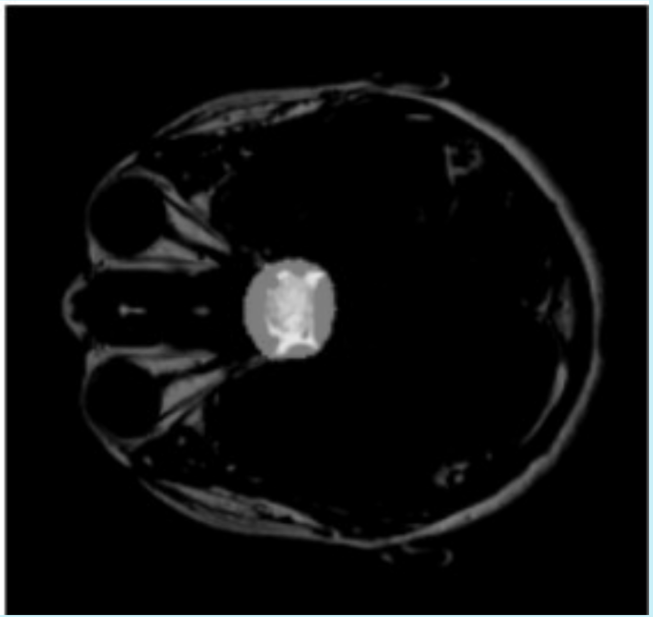
LinTUNet

U-Net

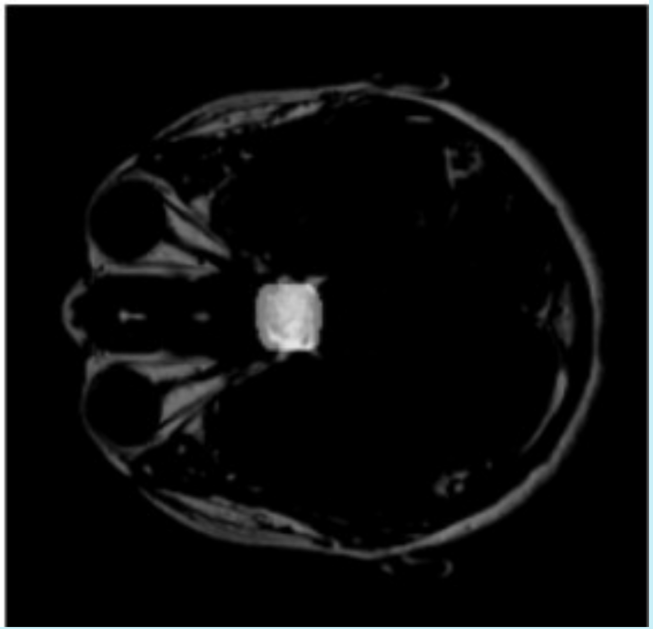


Intersection over Union (IoU) (cont)

U-Net



LinTUNet



Performance Evaluation: Execution Time

- LinTUNet is 7× faster than traditional U-Net

- Processing time:
  - LinTUNet: 0.0002 sec (0.2 ms)
  - U-Net: 0.0014 sec (1.4 ms)

- Speed boost due to the attention layer, which improves information extraction efficiency.

Future Implications of LinTUNet

1. LinTUNet enhances accuracy in complex medical images by combining **CNNs and Transformers.**

2. Works well with high-resolution images for applications like **healthcare, satellite imaging, and autonomous driving.**

3. **It is suitable for real-time diagnostics, crop analysis, and road scene segmentation**, with strong performance across different data types

Any Questions?