LinTUNet

A Hybrid CNN-Transformer Architecture for Medical Image Segmentation

PACISE 2025

Lawrence Menegus, DongSheng Che

East Stroudsburg University, Computer Science Department

Overview

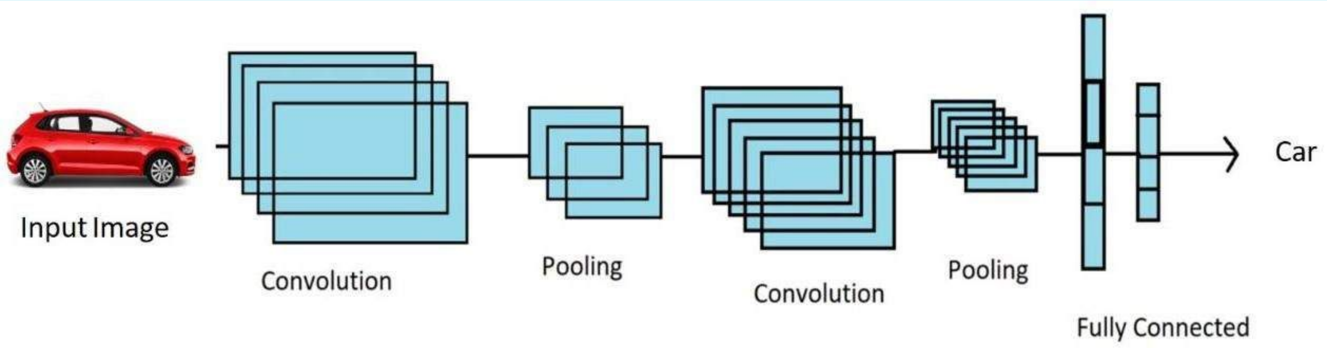
- **Deep Learning Basics:** Overview of CNNs, Autoencoders, and U-Net for medical images.
- **Transformers & Attention Layers:** How Transformers and Self- Attention Layers work, plus an intro to Linformer.
- **LinTUNet Model:** Combining Linformer with U-Net for better image segmentation.
    - **Performance & 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

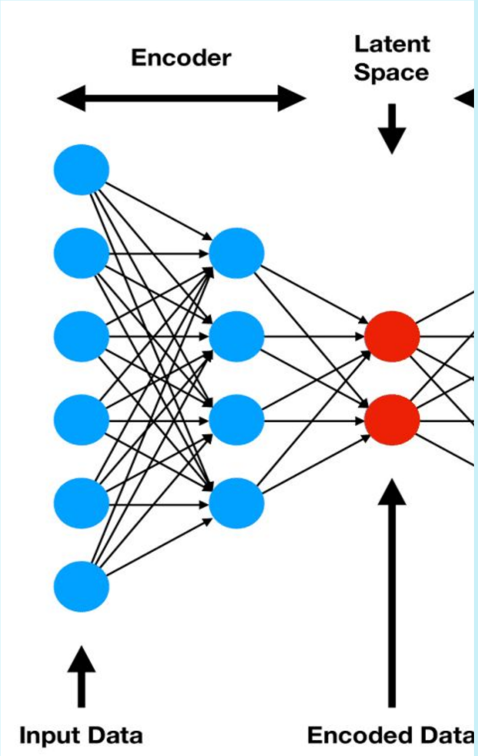
- **Lower layers capture simple features** (edges, colors), while deeper layers learn complex patterns (faces, objects).

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



Introduction to Autoencoders (Encoder)

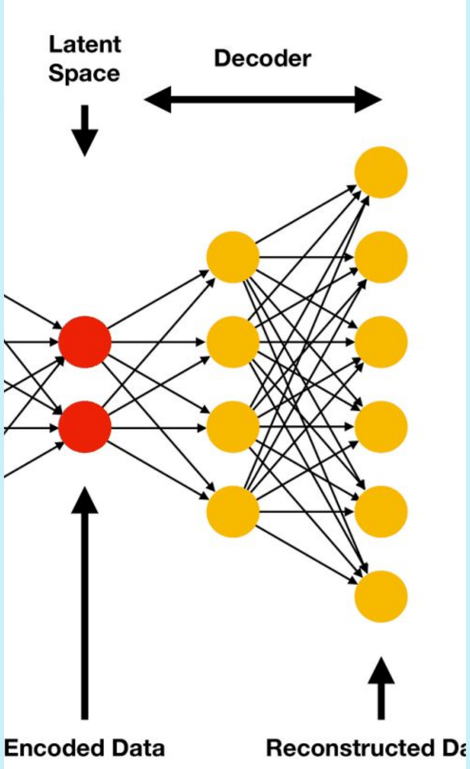
- An autoencoder is a neural network that is broken in **two parts**
  - **Encoder-** compresses input data into a lower-dimensional representation
  - **Decoder** - reconstructs it back to the original form



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)

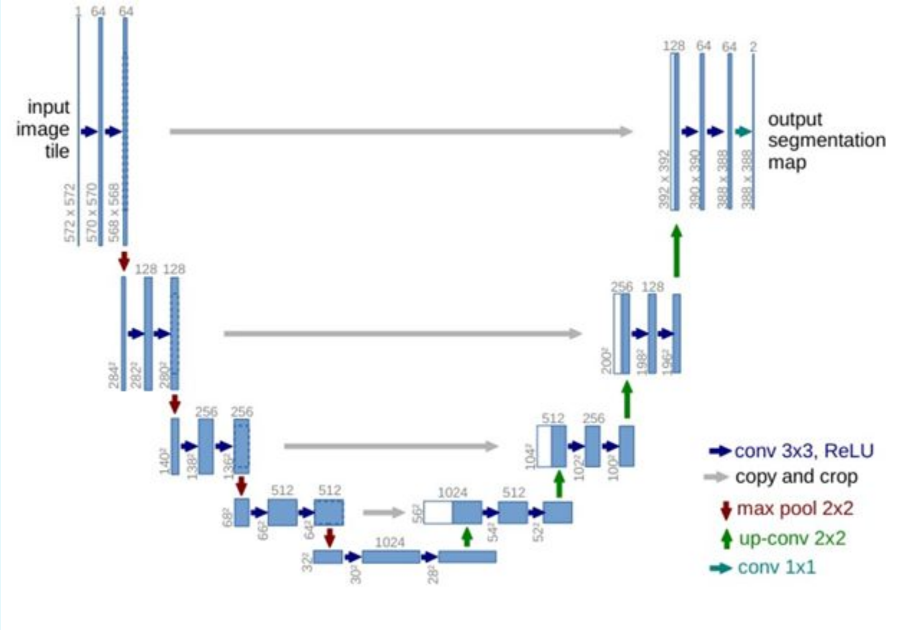


DECODER -

- 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

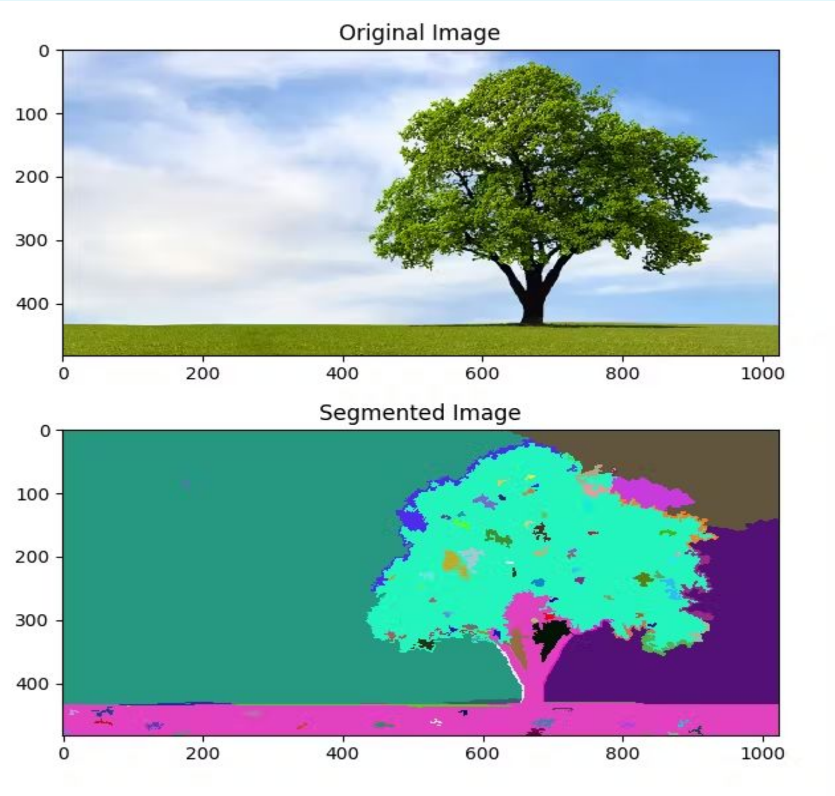


- **U-Net is an CNN-based architecture which mimics as autoencoder,** where the encoder extracts features, and the decoder reconstructs a segmented output, making it effective for pixel-wise predictions.

- **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.

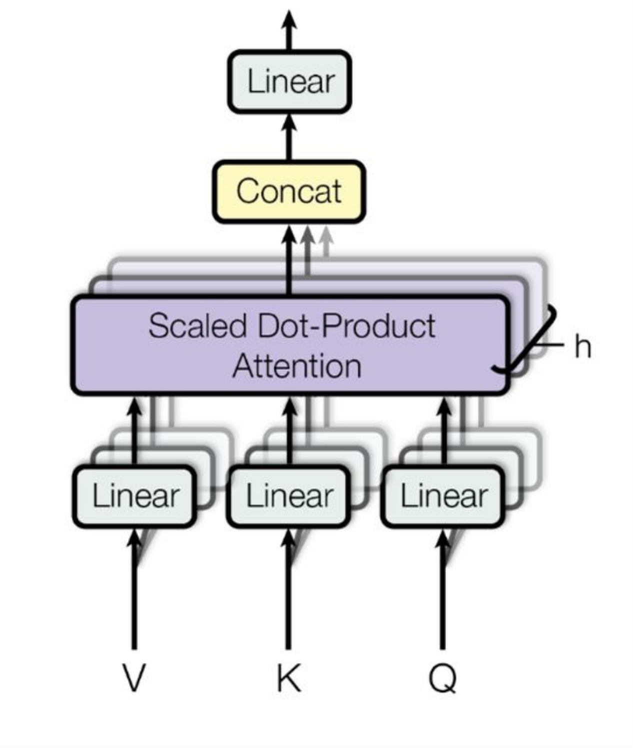
- It helps in identifying and separating objects (**Object Detection**) within an image, commonly used in medical imaging, self-driving cars, and facial recognition.

What is a Transformer in Deep Learning?

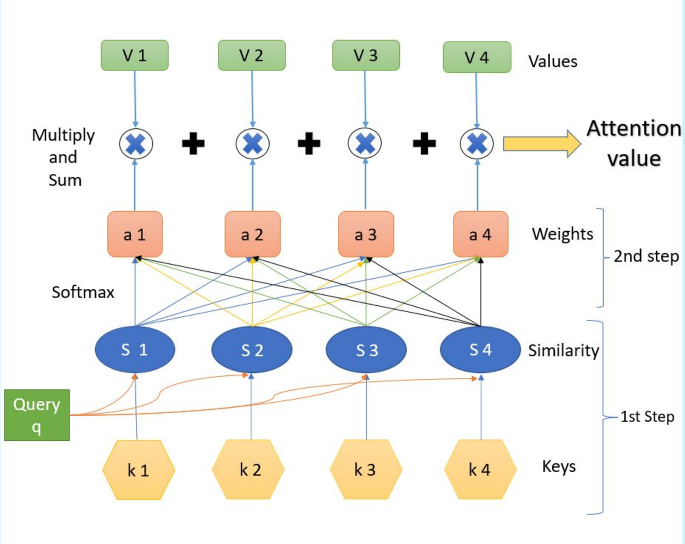
- **A Transformer** is a type of deep learning model designed to **handle sequential data** by processing the entire sequence at once, rather than step-by-step like traditional models such as RNNs (Recurrent Neural Networks)

- It uses a mechanism called **self-attention** to weigh the importance of different elements within the sequence.

- **Transformers** highly effective for tasks like machine translation, text generation, and image processing.



Self-Attention Layer Mechanisms



- **Self-Attention** weighs the importance of different elements within the sequence, allowing the model to consider the relationships between all tokens in the input.

- **Tokens** represent smaller units of the input, like words or subwords or pixels, and are essential for processing and understanding the data in sequence-based tasks.

- Each **token is converted** into a d-dimensional vector and transformed into three matrices:
  - Query (Q): What the token is looking for
  - Key (K): What information the token contains.
  - Value (V): The token's actual content.

- The attention layer **compares Q and K** to c**alculate focus scores**, giving more weight to important inputs.

- This helps the model **highlight relevant parts** and capture long-range dependencies

Sparse Attention Mechanism

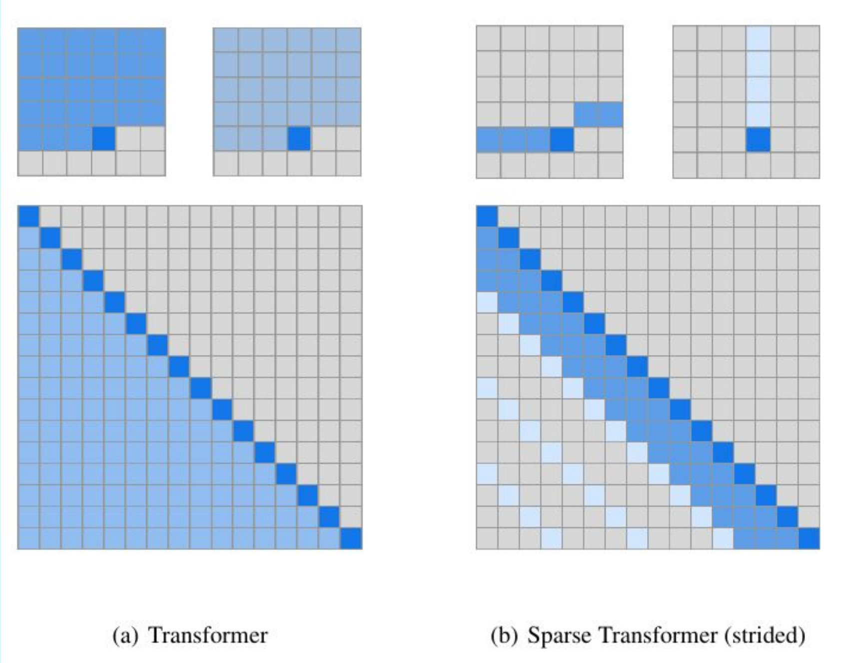
- The Sparse Attention Mechanism was introduced by **researchers at Google Research, including Rewon Child, Scott Gray, and others in 2019**

- **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



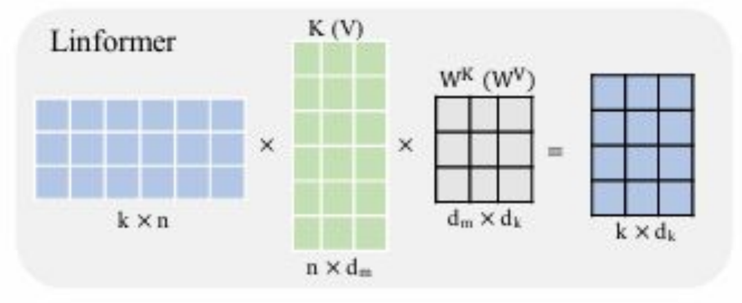
Introducing Linformer: A Better Transformer

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

- **Linformer compresses these vectors/matrices using low-rank projection,** reducing their size while preserving key information.

    - Instead of full n × d matrices, **Linformer** multiplies K and V by a **smaller learned projection matrix, shrinking** them to k × d

- **This removes redundancy** while keeping essential information, making self-attention faster and more memory-efficient.



Integrating into U-Net: The Architecture

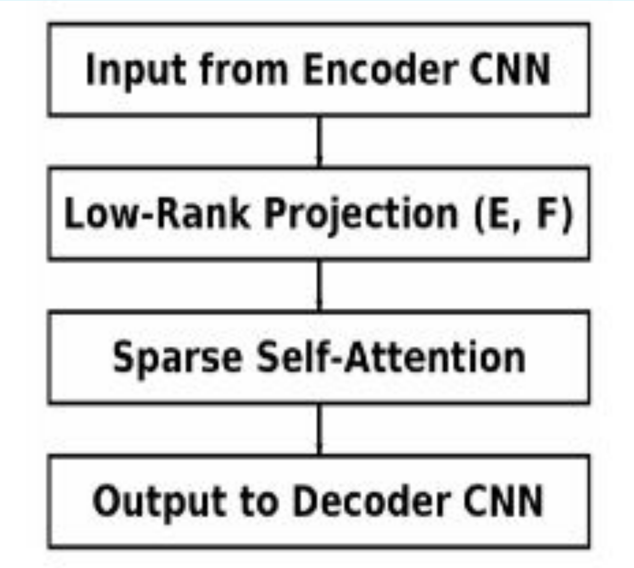
- **LinTUNet retains U-Net's encoder-decoder structure but enhances it with Linformer** in the bottleneck portion of U-Net Architecture.

- **Linformer takes the input image** that has been processed through the encoder.

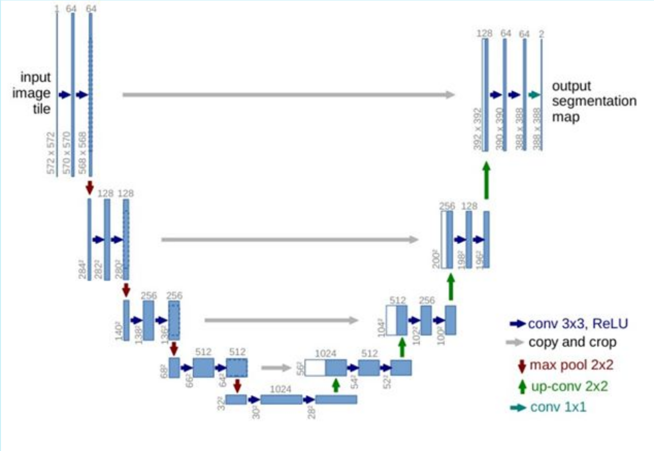
- It then compresses the feature vectors, **reducing their size.**

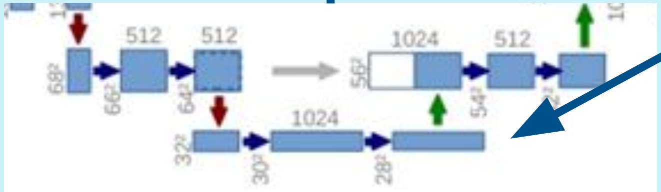
- This compression **creates a sparse attention layer,** where only the **most relevant parts of the data (from the compressed matrices)** are used when focusing on specific tokens.

- **The relevant information** is passed through the decoder of the U-Net to reconstruct the segmented output.

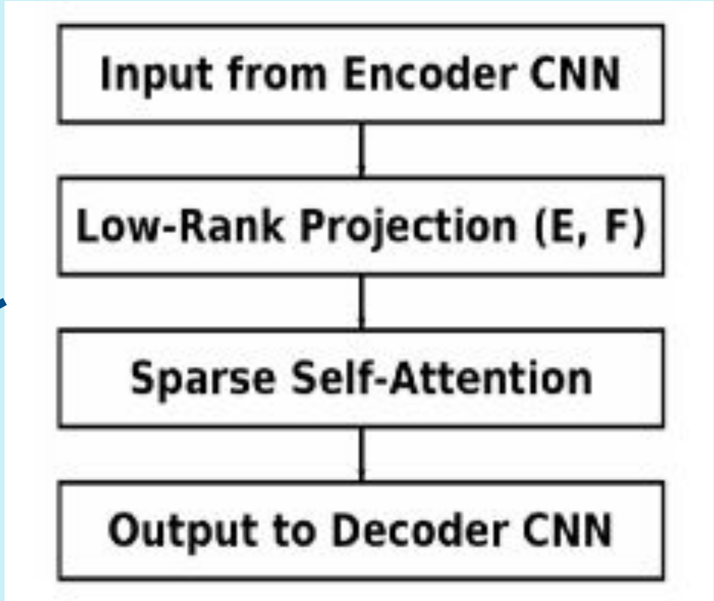


Integrating into U-Net: The Architecture (cont)





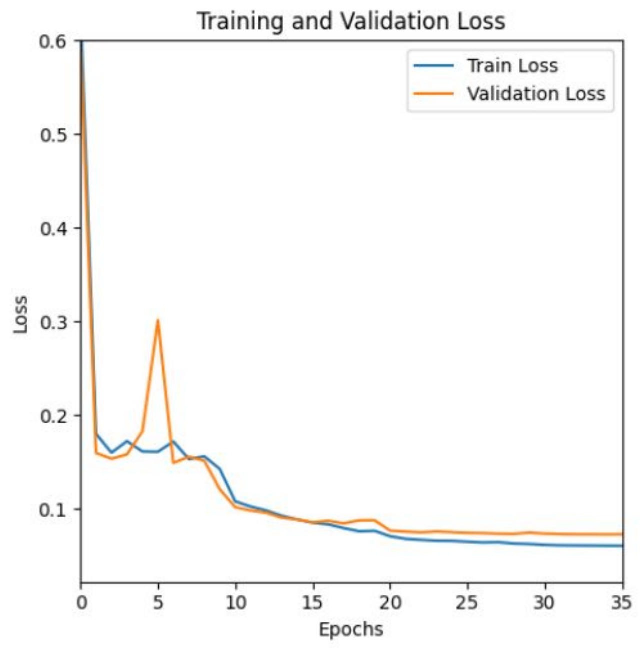
Bottleneck



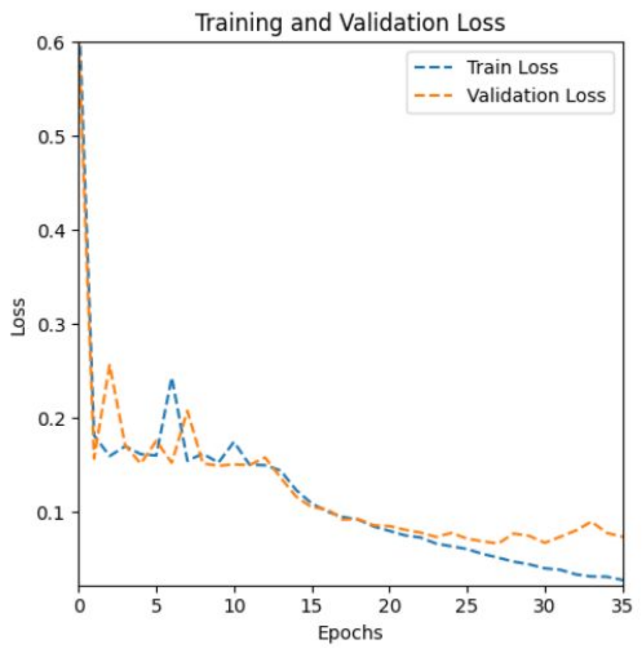
Performance Evaluation : Loss

- 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

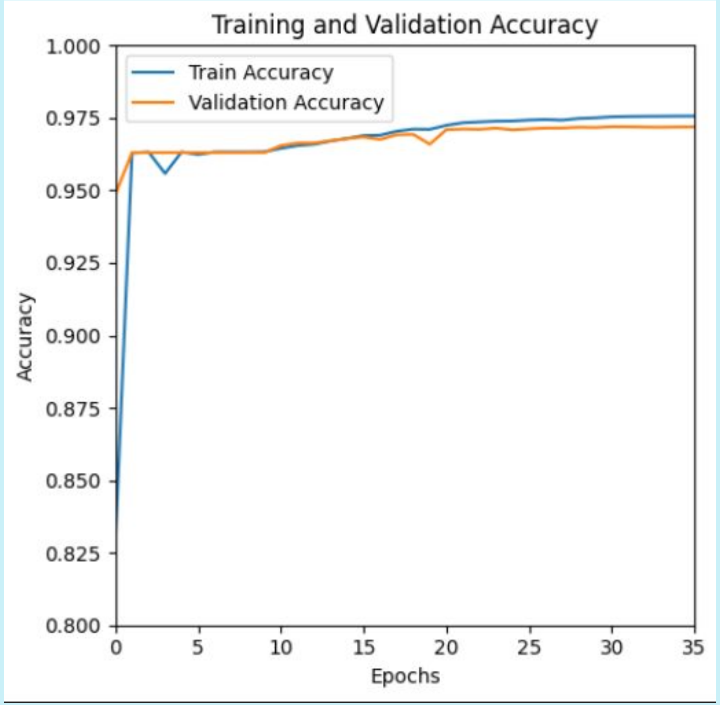


LinTUNet

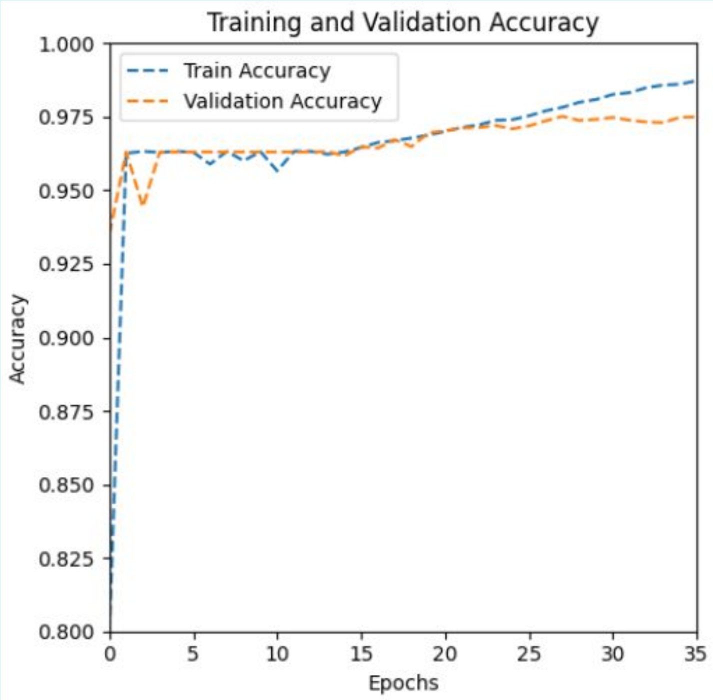


Performance Evaluation: Accuracy

U-Net



LinTUNet



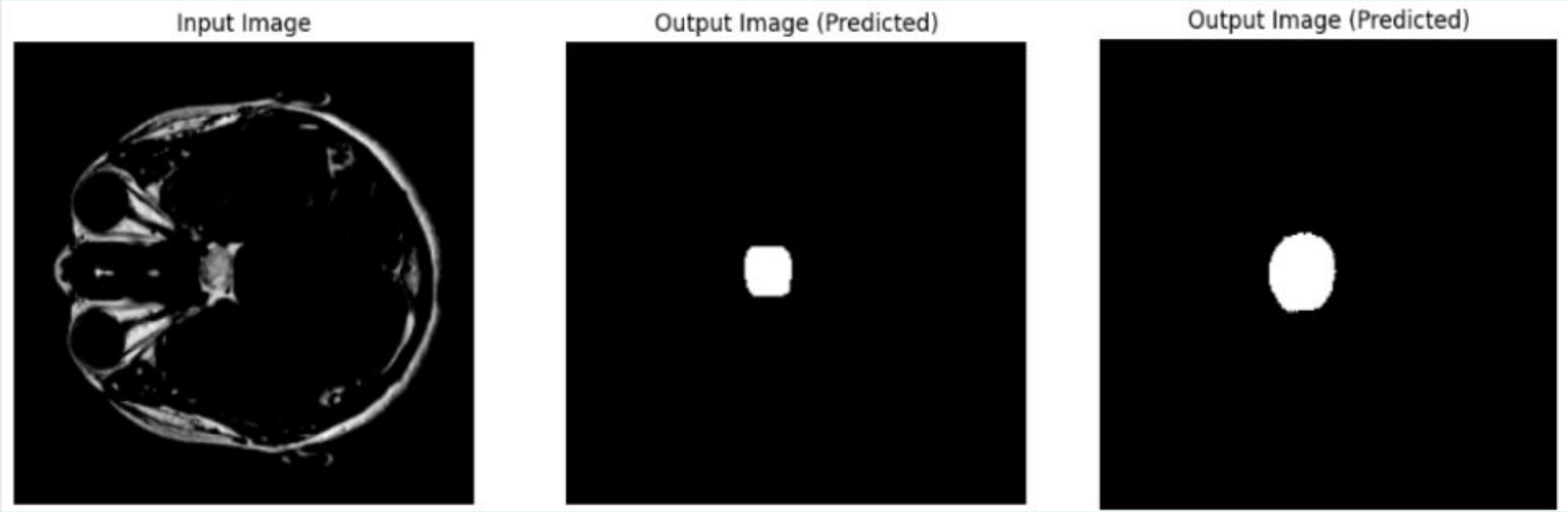
Performance Evaluation Metrics

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)

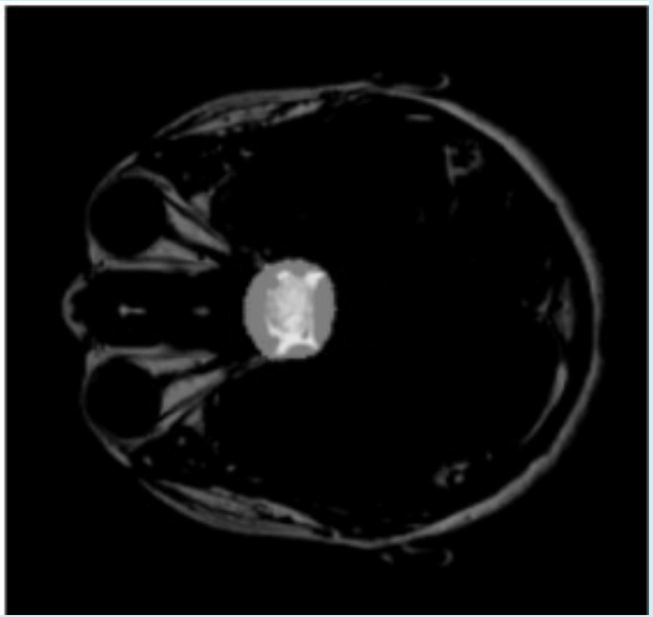
- **Intersection over Union (IoU)** is a metric used to evaluate the performance of image segmentation models. It measures the overlap between two sets: the predicted segmentation and the ground truth (input image).

LinTUNet U-Net

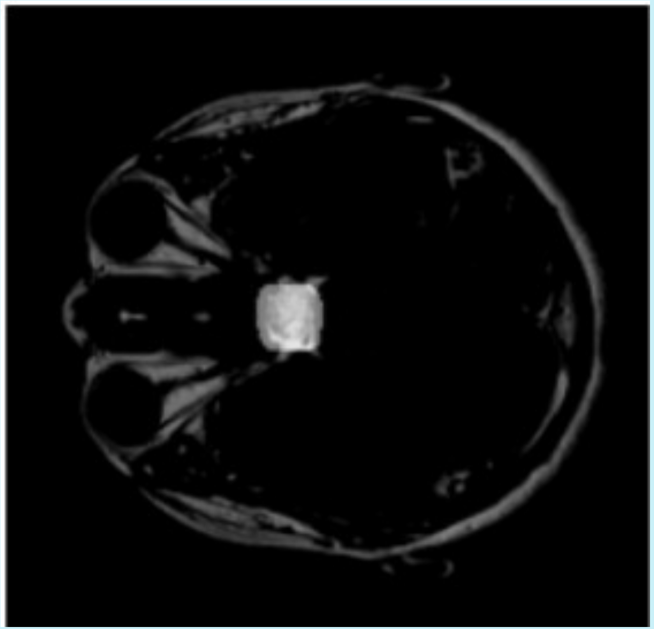


Intersection over Union (IoU) (cont)

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

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

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

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

Any Questions?