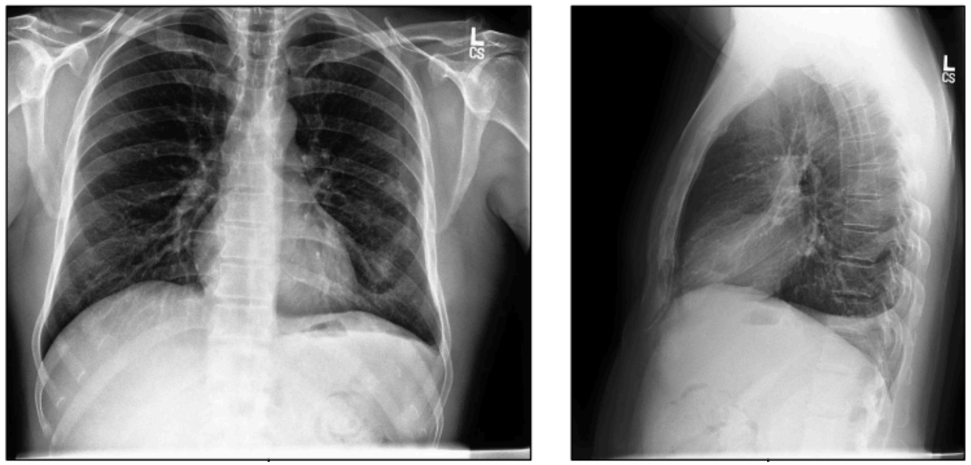
Transformer For Medical AI

Project 6

Zheng Hexing 2023311430
Chang Hwan Kim 2024321234
Maftuna Ziyamova 2024311551
Lee Woo Bin 2025311560

Problem Formulation - Motivation

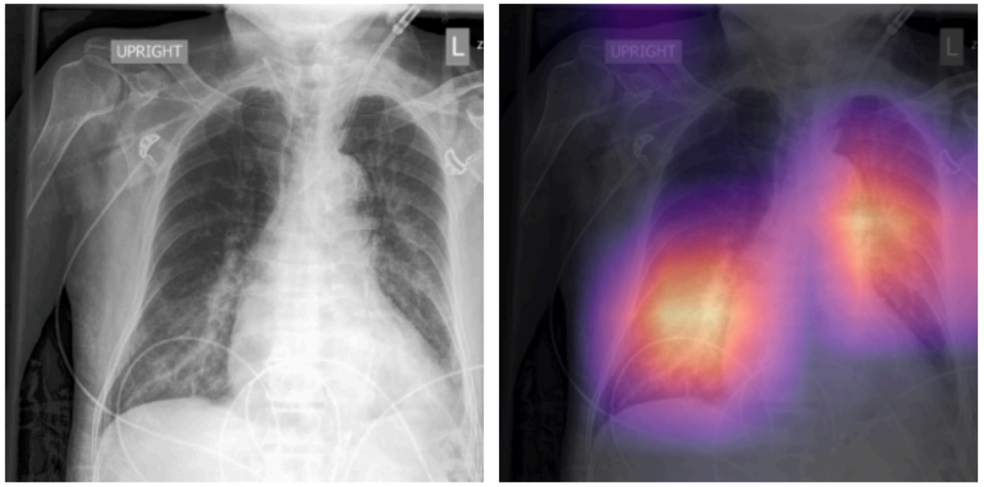
- Chest radiography is the most frequently performed imaging examination globally
- Essential for screening, diagnosing, and managing numerous life-threatening conditions
- Significant potential for automated interpretation systems to match or exceed radiologist accuracy



Problem Formulation - Goal

- Develop a Transformer-based model capable of accurately diagnosing chest radiographs based on 14 labeled observations

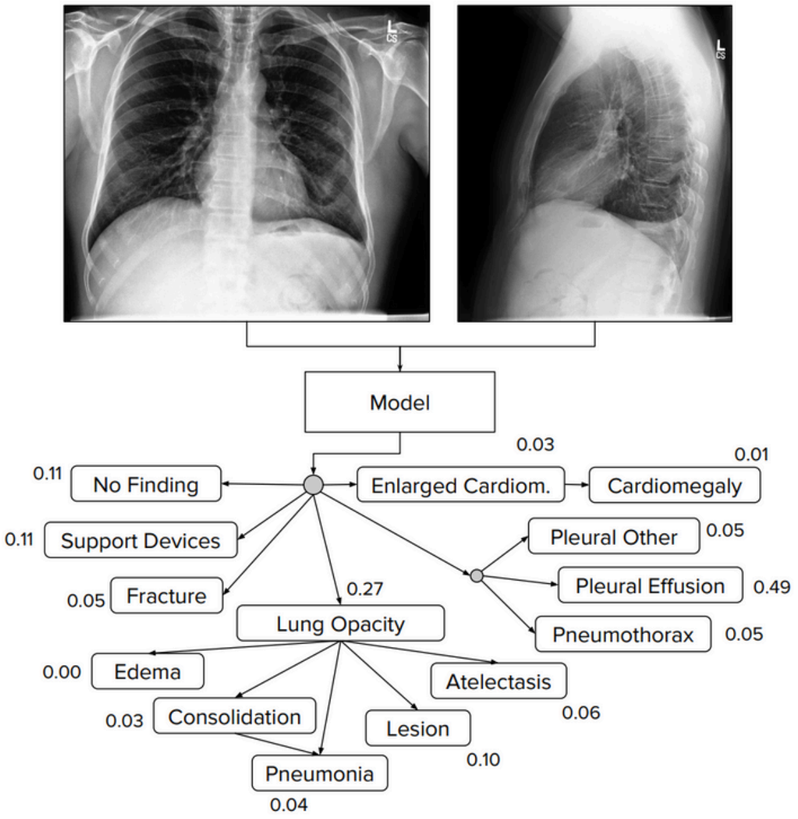
- Generate interpretable heatmap visualizations highlighting model attention areas to support clinical decision-making



Data and Methods

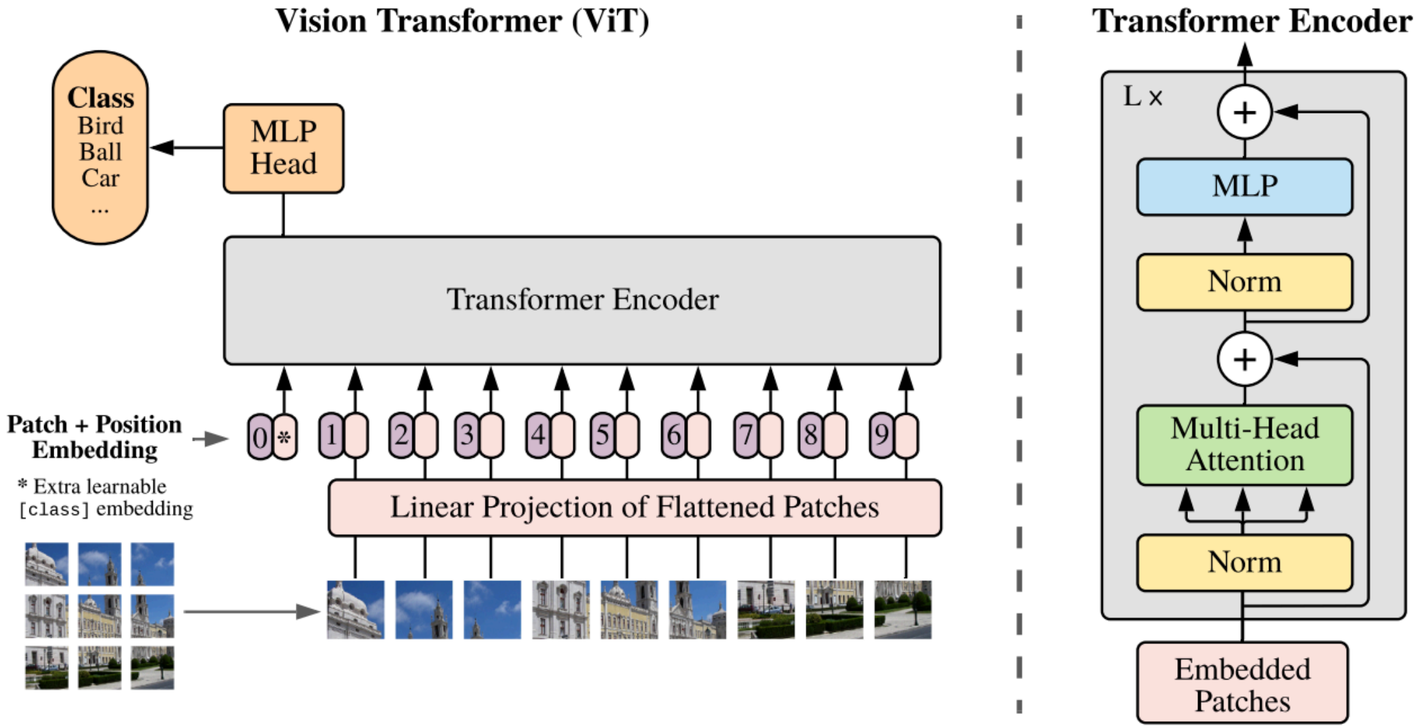
Dataset: CheXpert

- 224,316 chest radiographs from 65,240 patients
- 14 labeled clinical observations
- Automated labeling system designed to detect and classify observations, including inherent uncertainties
- Validation set of 200 radiographic studies that was manually annotated by 3 board-certified radiologists
- *Currently used **CheXpert-v1.0-small** that is a smaller, downsampled version of the original dataset*



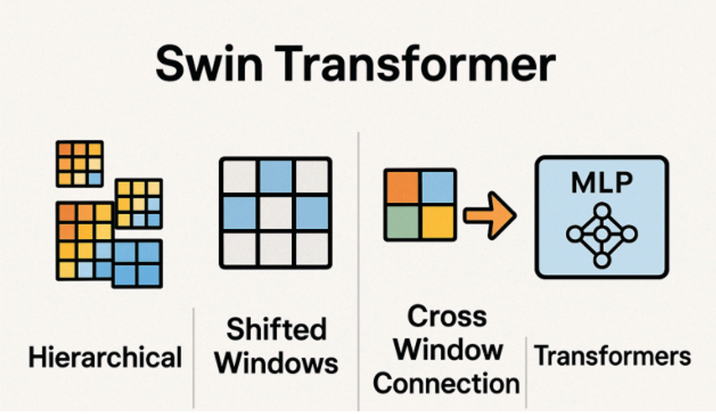
Data and Methods

We used 3 different architectures: Vision Transformer (ViT), Swin Transformer, BEiT Transformer

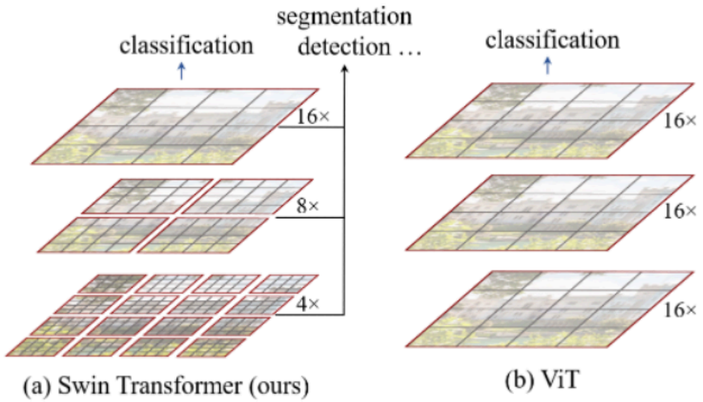


Vision Transformer (ViT) Overview

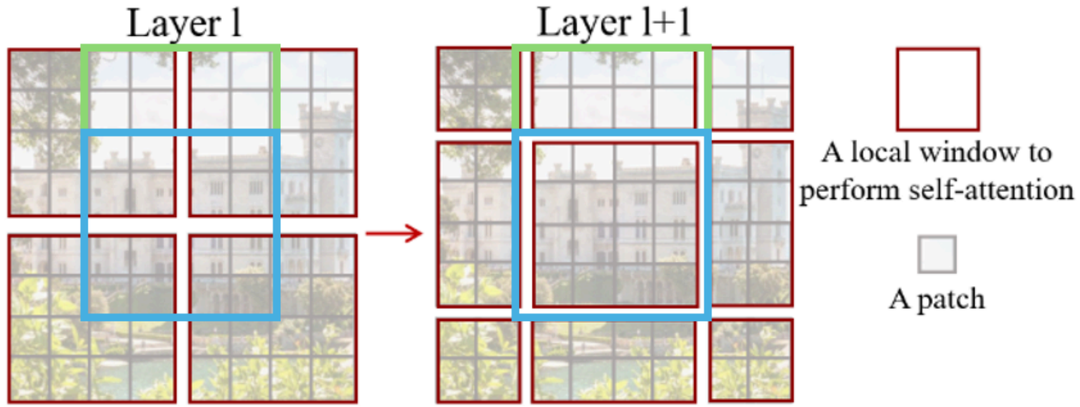
Data and Methods



Hierarchical architecture:



Cross window connection and Shifted Windows :

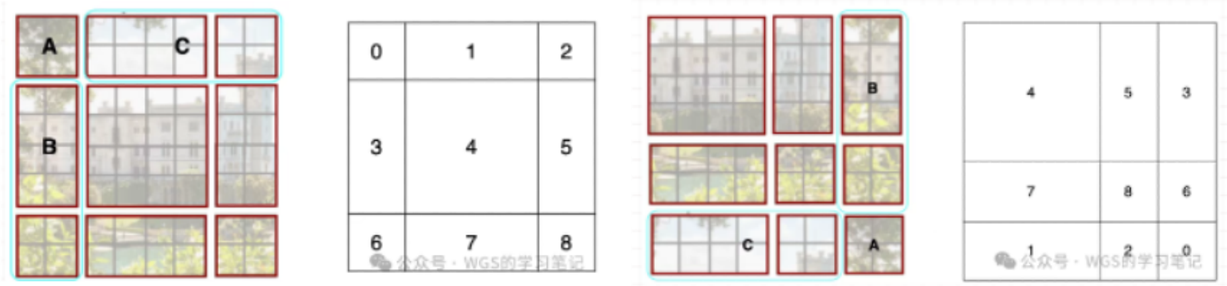


Shifted Windows

- Improved global modeling
- Better context aggregation
- Preserves locality

Cross window connections

- Stronger contextual learning
- Better spatial coherence
- Handles object boundaries well

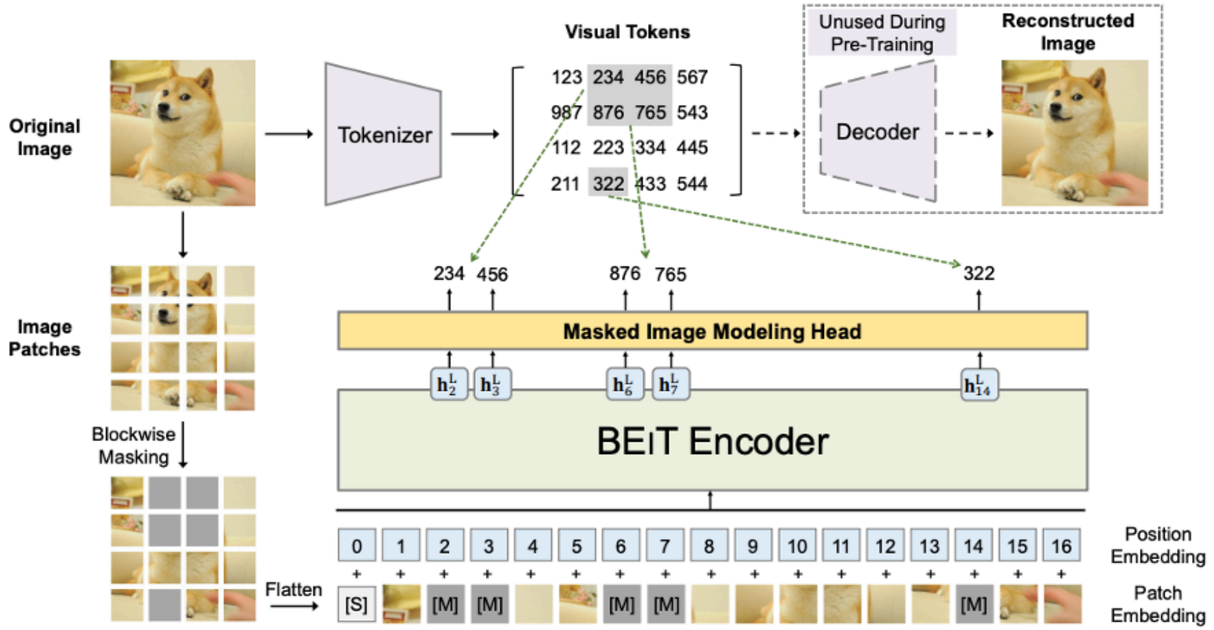


YONSEI UNIVERSITY

Data and Methods

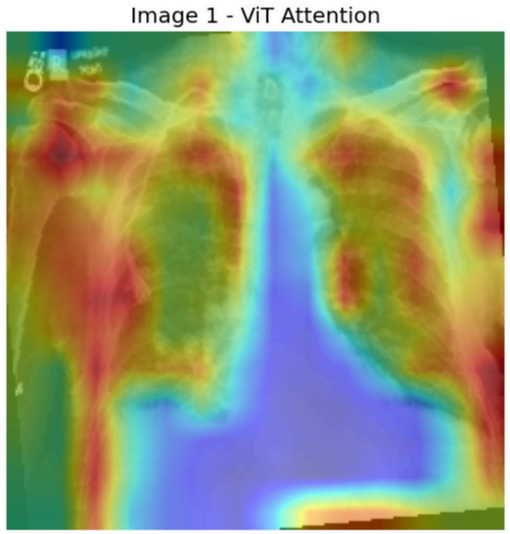
- **BEiT Transformer** model is a vision transformer based on self-supervised learning
- It applied BERT's Masked Language Modeling to images by predicting visual tokens

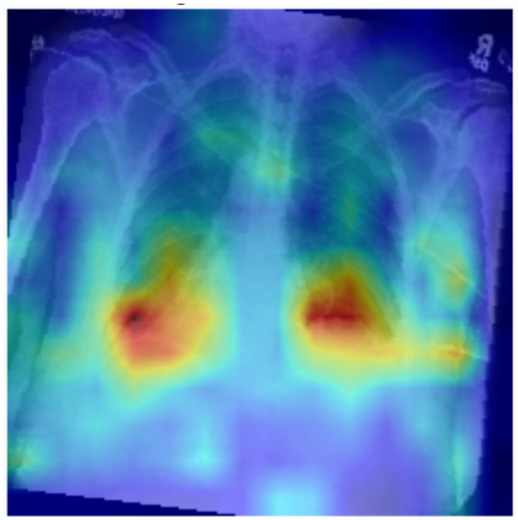
- It uses a Discrete VAE to convert images into codebook-based visual tokens and predicts the masked parts
- BEiT shows strong pretraining performance and is effective for various downstream vision tasks



Data and Methods

- Last layer of the model's CLS token – trained to gather "information that represents the entire image" during the learning process
- Converting "the attention score value that the CLS token gives to each patch in the image" into "a 2D heatmap"
- Using a **heatmap**, we can understand intuitively where the Vision Transformer model is focusing on the image





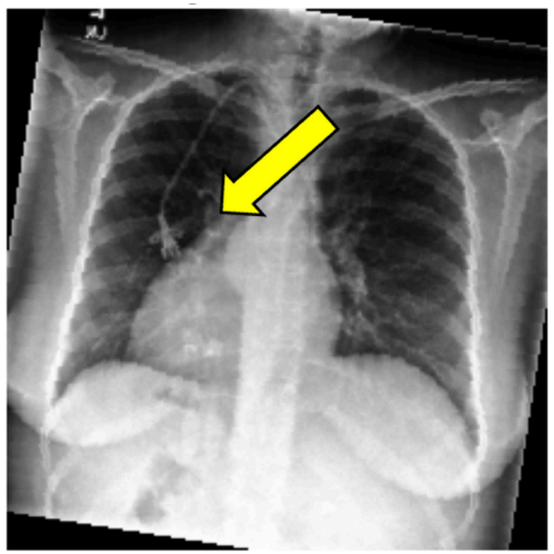
An example of 2D heatmap from our previous baseline model
: since the model is **not fully trained** yet, it's **not accurate**

An example of 2D heatmap our current baseline model
: the model **properly focuses** on the part used to determine the disease

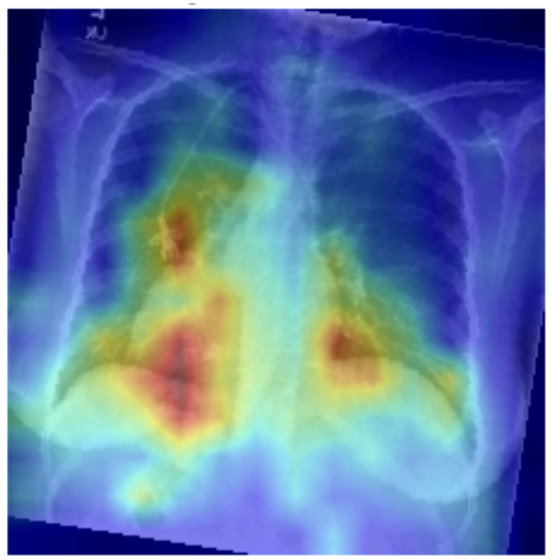
YONSEI UNIVERSITY

Data and Methods

- Modified the attention layer that is used to output the results



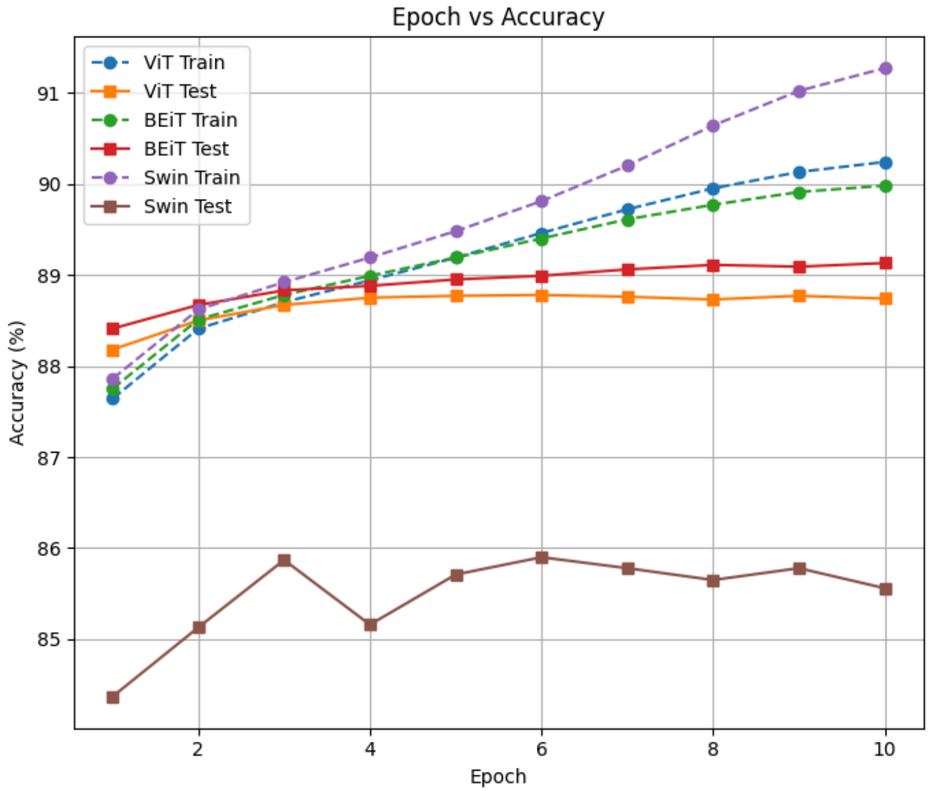
Original image

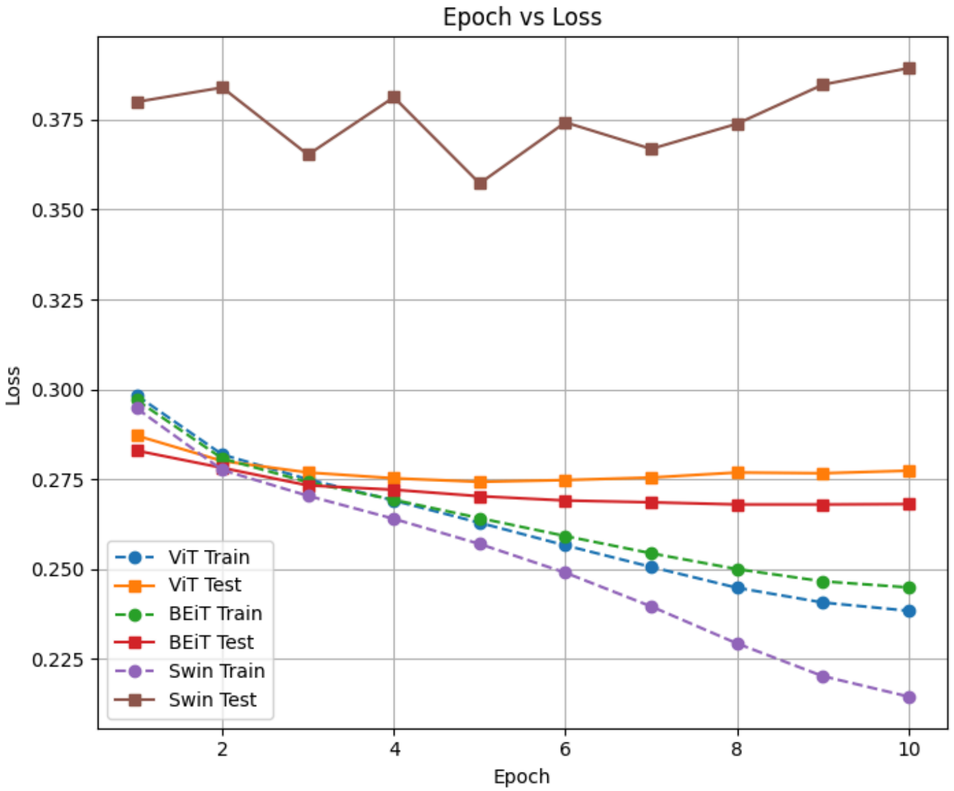


Heatmap image

Another example of 2D heatmap from our baseline model
: correctly classifies support device

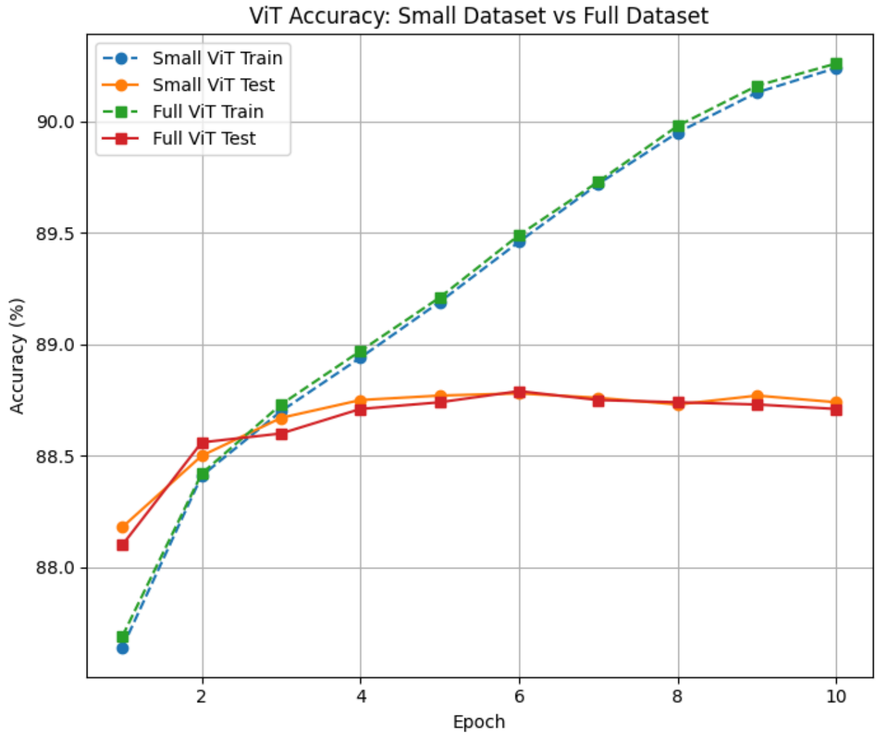
Results

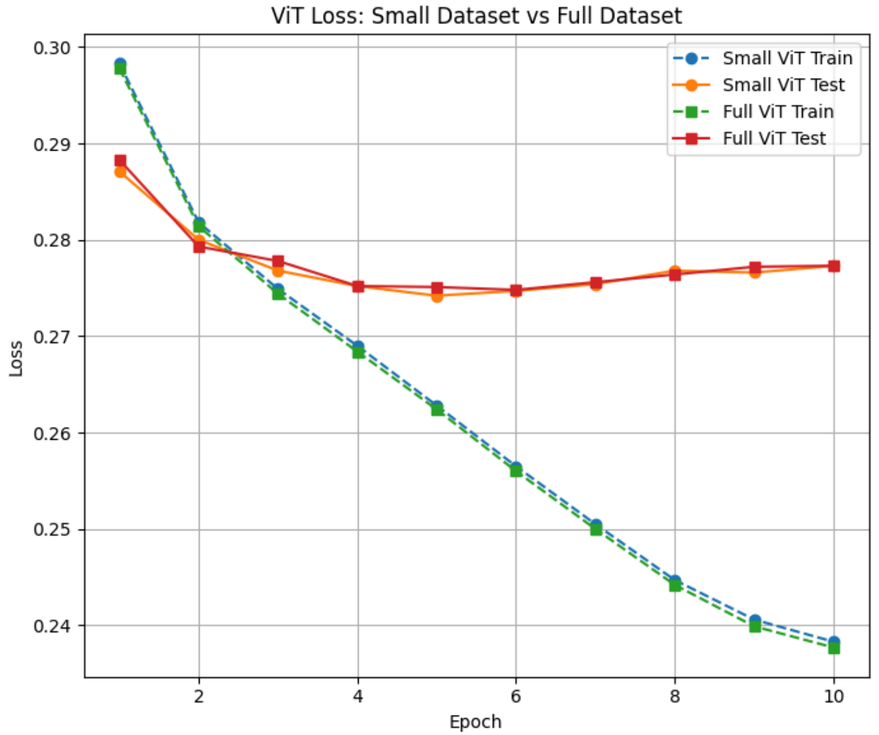




- Swin transformer performs best among those models

Results





- No difference in results between training full and light datasets

Thank you for attention!