CS 512 - F24

Implementation of DeIT - Data efficient Image Transformer

Team members:
Vignesh Ram Ramesh Kutti (A20548747)
Aravind Balaji Srinivasan (A20563386)

ILLINOIS TECH | Discover. Create. Solve.

1

ILLINOIS TECH

Introduction

- **Transformers in NLP vs. Vision**: While Transformers excel in NLP, Vision Transformers (ViT) face challenges due to their need for large datasets to perform well.
- **DeiT's Innovation**: DeiT, or Data-Efficient Image Transformer, addresses this issue by introducing a distillation method, allowing effective training on smaller datasets.
- **Project Goal**: This project aims to implement DeiT and compare its performance with ViT, evaluating its efficiency and potential as a data-efficient image classifier.

2

ILLINOIS TECH

Objectives

- **Implementation:** To implement ViT, DeiT and perform augmentation.
- **Analysis:** To analyse the performance of all three models using multiple metrics (e.g., Accuracy, AUC, F1 score, Top-1, Top-5, Precision, Recall).
- **Report:** Document our inference and show the results in a report.

3

ILLINOIS TECH

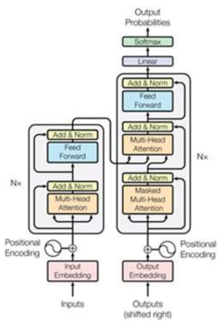
Background

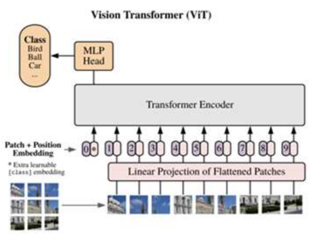
- **NLP:** Attention is all you need - Transformers - 2017
- **CV:** Vision Transformer ViT - 2020
- Data Efficient Image Transformer - 2021
- Why DeiT on top of ViT? - JFT300M, ImageNet
- Significant Performance improvement in smaller datasets.

4

ILLINOIS TECH

Background





5

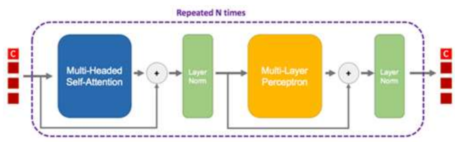
ILLINOIS TECH

Class Token

Purpose: A special token that gathers information from all image patches to predict the image class.
How It Works:

- Added to the input sequence along with image patches.
- Interacts with patches across transformer layers, building a summary of the image.

Benefit: Enables classification without additional pooling layers.



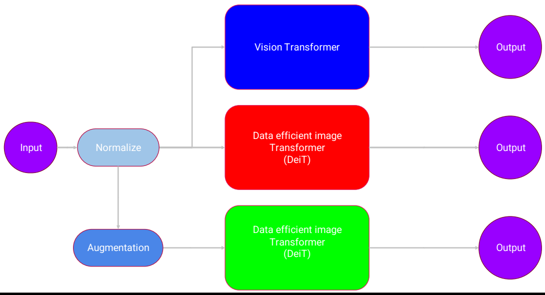
6

Dataset

- We used CIFAR-10 dataset which has 50,000 train images and 10,000 test images each of size 32 x 32.
- Well-labeled and well-researched dataset.
- It has many pre-trained models for us to choose from which provide state-of-the-art accuracy and performance.

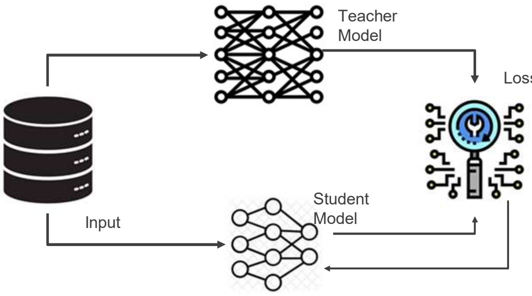
7

Methodology



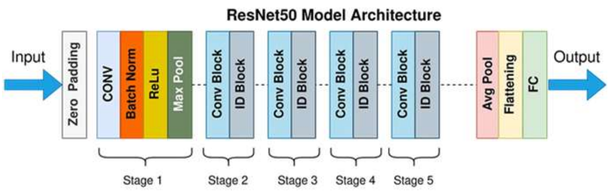
8

DeiT Architecture



9

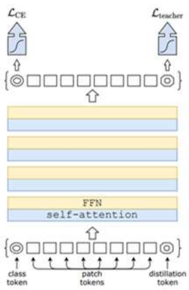
Pre-Trained Teacher Model: ResNet-50



This model provides accurate results for CIFAR-10 dataset, it has 50 layers and computationally cheap.

10

Methodology



Distillation: the process where a smaller student model (DeiT) learns to mimic the outputs of a larger teacher model (e.g., ResNet-50) by minimizing a loss function that combines both the teacher's predictions and the ground truth labels.

Soft Distillation: This is a technique where a student model learns from the teacher model by matching its "soft" probability distributions (or logits) for each class, rather than just the hard class labels, allowing the student to capture nuanced information about class similarities.

11

Soft Distillation Loss Calculation

$$\mathcal{L}_{\rm global} = (1-\alpha)\mathcal{L}_{\rm CE}(\psi(Z_s), y) + \alpha\tau^2 \mathrm{KL}(\psi(Z_s/\tau), \psi(Z_t/\tau)).$$

α - the coefficient balancing the Kullback–Leibler divergence loss (KL)

$\mathcal{L}_{\rm CE}$ - Cross-entropy loss on ground truth labels y

Z_s - Student Logits

Z_t - Teacher Logits

ψ - Softmax Function

τ - Temperature

KL Divergence: It is a measure of how one probability distribution differs from a reference distribution, quantifying the "distance" between them by calculating how much information is lost when approximating the reference with the other distribution.

12

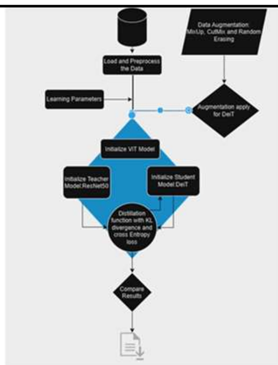
ILLINOIS TECH

Significance of Distillation token

- **Captures Teacher Knowledge**: Learns from a teacher model (e.g., ResNet), embedding insights typically gained from larger datasets.
- **Enhances Generalization**: Combines classification and distillation tokens, blending supervised learning with teacher guidance.
- **Boosts Accuracy**: Aligns student predictions with the teacher's features for higher accuracy.
- **Uses Cosine Similarity**: Matches student features to the teacher's distribution via cosine similarity.
- **Increases Data Efficiency**: Reduces data needs by transferring teacher knowledge, enabling effective learning on smaller datasets.

13

ILLINOIS TECH



14

ILLINOIS TECH

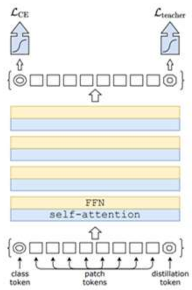
Implementation

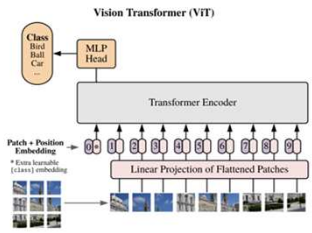
- **ViT Model Structure**: Implements **patch embedding** for input images and **attention layers** to capture spatial relationships, forming the basis of transformer-based vision processing.
- **DeiT Model with Distillation Token**: Extends ViT by adding a **distillation token**, allowing knowledge transfer from a pretrained **teacher model** (ResNet-50), improving efficiency.
- **Teacher-Student Distillation**: DeiT's **teacher-student setup** enables the student model to learn both from labeled data and from the teacher's output, enhancing generalization.
- **Tracked Metrics**: During training, essential metrics such as **accuracy, AUC, F1 score, precision,** and **recall** are recorded to assess performance across various dimensions.
- **Data Augmentation**: CutMix, MixUp, and other augmentations enhance learning, especially with the distillation setup, contributing to improved model robustness and accuracy.

15

ILLINOIS TECH

Structure





16

ILLINOIS TECH

Training and Data Augmentation

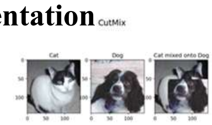
- The training setup:
  - We ran each model for 20 epochs.
  - We used Adam optimizer with learning rate 0.001
  - Batch_size = 64
  - ResNet-50 from torchvision.models is used as the teacher model for DeiT.
  - Top-1 Accuracy, Top-5 Accuracy, AUC, Precision, Recall, and F1 Score are logged for each paradigm for plotting the results.
- Plots showcasing the models performance on the validation set are displayed in the following slides.

17

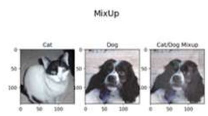
ILLINOIS TECH

Data Augmentation

CutMix: it cuts a patch from one image and pastes it into another, creating new examples with different regions from different images.



MixUp: Combines two images by taking weighted averages of both the images and their labels, increasing dataset diversity and improving model robustness.



18

ILLINOIS TECH

Data Augmentation

Horizontal Flip: Flipping the image horizontally (with a 50% probability) introduces different object orientations, helping the model generalize across various image flips.

Colour Jitter: Alters brightness, contrast, and saturation, simulating lighting changes.

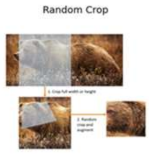
19

ILLINOIS TECH

Data Augmentation

Random Erasing: Masks portions of the image to simulate occlusion.

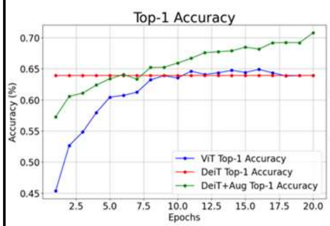
Random Crop: Randomly cropping a portion of the image ensures object robustness despite variations in object position and scale

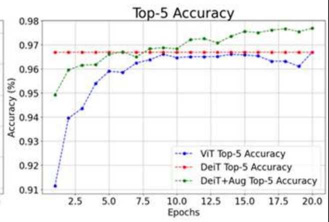


20

ILLINOIS TECH

Results



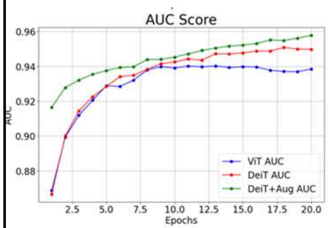


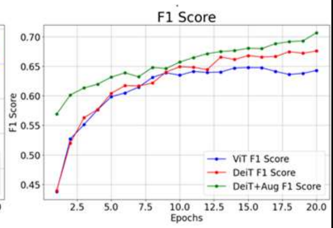
Constant line of DeiT is because of the Distillation process.

21

ILLINOIS TECH

Results

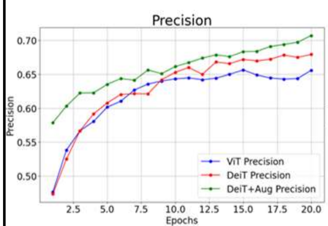


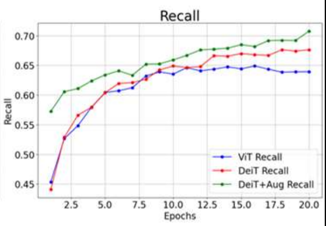


22

ILLINOIS TECH

Results





23

ILLINOIS TECH

Results

- From the Plots above we can observe that DeiT performs better than vanilla ViT.
- And, DeiT with Data Augmentation provides a significant boost to the performance of the DeiT.
- DeiT with augmentation also reduces Overfitting which is evident from the validation accuracy graph.

24

Problems Faced

- **Problem:** DeiT model's validation loss was high.
- **Solution:** We implemented Data Augmentation.

- **Problem**: computation of CutMix and MixUp for the whole batch took too much time.
- **Solution:** We first tried to do it to one image but we were unsuccessful, so we reduced the probability of performing these operations on batches to 1%, which improved computation time, and improved performance greatly.

25

Inference

- **Performance Comparison**: Both ViT and DeiT achieved similar Top-1 (63.9%) and Top-5 (96.7%) accuracy on CIFAR-10, with DeiT showing better results in F1 score (0.675 vs. 0.642) and AUC (0.949 vs. 0.938).

- **Impact of Data Augmentation**: Adding data augmentation to DeiT significantly boosted all metrics, including Top-1 accuracy (70.7%) and F1 score (0.706).

- **Key Insight**: The combination of DeiT's efficient design and data augmentation enhances performance, making it well-suited for real-world applications demanding high accuracy and balanced metric scores.

26

References

1. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.12877
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2010.11929
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

27