Comparative Analysis of CNN and Transformer Architecture for Image Classification

Karthik Kumar Kaiplody
Illinois Institute of Technology
kkumarkaiplody@hawk.iit.edu

Shrey Jaradi
Illinois Institute of Technology
sjaradi@hawk.iit.edu

***Abstract*—This project aims to undertake a comprehensive comparative analysis between Convolutional Neural Networks (CNNs) and Transformer architectures, specifically in the context of image classification. The primary objective is to construct, train, and assess two distinct deep learning models: one founded on a CNN and the other on a Transformer, leveraging identical datasets. The central aim is two fold: firstly, to acquire a nuanced understanding of these architectural variances and, secondly, to discern the optimal circumstances warranting the utilization of each. Through this exploration, we aim to unravel the distinctive traits, strengths, and scenarios conducive to the efficacious implementation of CNNs and Transformers in image classification tasks.**

***Index Terms*—machine learning, deep learning, CNN, ResNet, Vision Transformer, Transformer**

I. Introduction

In the domain of computer vision, the task of image classification plays a pivotal role, especially in applications related to waste management. This study embarks on a comprehensive exploration, aiming to conduct an intricate comparative analysis between two powerful architectures for image classification: Convolutional Neural Networks (CNNs) and Transformers. Our focus is not just on understanding their technical disparities but also on discerning their optimal applications in handling waste classification scenarios.

Automated waste classification, a critical step toward sustainable waste management, demands robust image classification models. For our analysis, we leverage a curated dataset encompassing images of Organic (denoted as "O") and Recyclable (denoted as "R") waste. The dataset provides a rich reservoir to test the efficacy of these architectures in distinguishing between these crucial waste categories.

II. Proposed Solution

A. Overview

To delve into this endeavor, we design and evaluate four distinct models — two each based on CNNs and Transformers—trained under identical conditions. Specifically, our methodology involves:

- Custom CNN Model (ResNet-18): Constructing a CNN architecture, ResNet-18, from scratch, providing a granular understanding of its intricate layers and mechanisms, and training it on waste classification dataset.
- Pre-trained ResNet-18 Model: Utilizing a pre-trained ResNet-18 model and fine-tuning it using transfer learning techniques, tailoring its knowledge to our specific waste classification task.
- Custom Transformer Model: Building a Transformer architecture from the ground up, unraveling the dynamics of self-attention mechanisms and positional encodings in image classification.
- Pre-trained Transformer Model: Employing a pre-trained Transformer architecture and adapting it to our dataset, exploring its adaptability and performance.

This parallel analysis across four distinct models enables us to glean comprehensive insights into the efficacy, adaptability, and optimal use cases of both CNN and Transformer architectures in waste image classification scenarios.

III. Methodology

To achieve a comprehensive comparative analysis between Convolutional Neural Networks (CNNs) and Transformers in the context of image classification, the following methodology was employed:

A. Dataset Preparation

The dataset used in this project consists of images categorized into two classes: Organic (O) and Recyclable (R) waste. The dataset was obtained from https://www.kaggle.com/datasets/techsash/waste-classification-data/data.

The dataset characteristics are as follows:

- Number of classes: 2
- Class names: ['O', 'R']
- Train data size: 22634 images
- Test data size: 2513 images

The original images in the dataset varied in size and resolution. To ensure uniformity and compatibility for model training, all images were transformed to a standard size of 224x224 pixels. Additionally, the pixel values were normalized to enhance convergence during model training. The normalization process involved adjusting the mean pixel values to [0.485, 0.456, 0.406] and the standard deviation to [0.229, 0.224, 0.225].

These transformation steps were critical in standardizing the dataset and enabling consistent feature extraction and classification of waste images across the different models evaluated in this project.

B. Model Construction and Training

Two different types of deep learning models were constructed for this project:

1) ResNet-18: ResNet-18, is a convolutional neural network architecture designed to address the vanishing gradient problem in very deep networks. It consists of 18 layers, featuring residual or skip connections that facilitate the training of deeper networks. Here's an overview of its architecture:

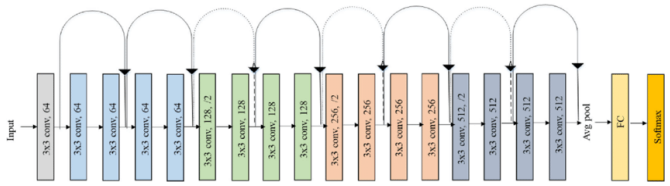

Fig. 1: ResNet-18 Model Architecture

- High-Level Structure: ResNet-18 comprises several layers grouped into stages. The architecture follows a pattern of down-sampling the spatial resolution while increasing the number of filters (or channels) in the feature maps.
- Components:
  - Input Layer: Accepts the input image data.
  - Conv1: Initial convolutional layer with a kernel size of 7×7, followed by a stride of 2, and 64 output channels.
  - Max Pooling: Down-samples the spatial dimensions of the feature maps.
  - Residual Blocks (Conv2x to Conv5x): These blocks form the backbone of the network and are composed of several residual units.
- Residual Unit:
  A residual unit in ResNet-18 comprises two convolutional layers, each followed by batch normalization (BN) and Rectified Linear Unit (ReLU) activation. The key characteristic is the identity shortcut connection that skips one or more layers:
  - Basic Block: The basic building block consists of two convolutional layers with small filters (typically 3×3 kernels).
  - Shortcut Connection: This connection sums the output of a convolutional block with the original input, allowing the network to learn residual functions.
- Stage Details:
  - Conv2x to Conv5x: These stages are composed of a varying number of residual units, each block increasing the number of channels while reducing the spatial dimensions.
  - Global Average Pooling: This pooling layer aggregates spatial information across each feature map, generating a fixed-size output for the fully connected layer.
  - Fully Connected Layer: The final layer acts as a classifier, providing output predictions for the classes.
- Training: During training, the network parameters, especially the weights of convolutional layers, are learned using gradient descent-based optimization techniques like stochastic gradient descent (SGD) or its variants.

ResNet-18's design enables the training of deeper networks by mitigating the degradation problem with increased depth. The residual connections aid in smoother gradient flow, allowing for the successful training of networks with many layers, improving accuracy and ease of optimization. This architecture laid the groundwork for subsequent ResNet variants and remains influential in modern CNN designs.

2) Transformer: Vision Transformer architecture is based on the idea of dividing the image and looking into specific part using Attention. Below is the architecture from the original paper[3]

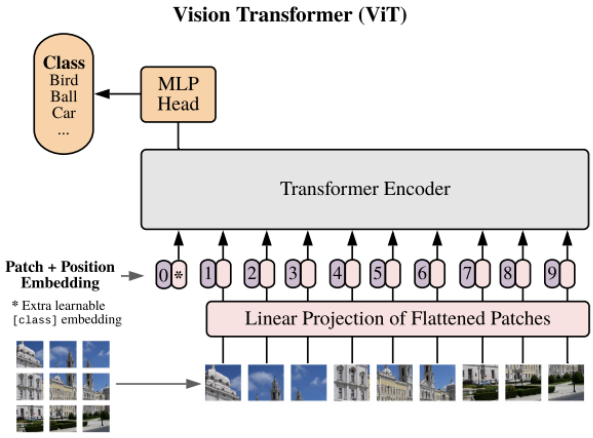

Fig. 2: Vision Transformer Model Architecture

According to this architecture we will break down the image into smaller parts called patches. Each patch is typically a square region of the image. In the original paper[3] author has divide the image into 16x16, stride used in the paper is also 16.

After getting the patches we flatten the 2D vector to a 1D vector, each patch is treated as separate input token. This will then go through linear projection component which preserving the relationships and importance features this involves Weight matrix multiplication and bias addition.

In the next step, position embedding is added to each patch with its corresponding location. This transformer vector of all patches will then fed to the encoder block to learn the self attention and finally going through Multi-layer perceptron head to predict the class.

IV. Experiment Setup

The experiments were conducted using Google Colab, a cloud-based Jupyter notebook environment, which provided access to a T4 GPU accelerator. The utilization of the T4 GPU significantly expedited the model training process, allowing for

efficient computation of complex neural network architectures such as ResNet-18 and Vision Transformer.

The hardware configuration on Google Colab was instrumental in facilitating the training of deep learning models, leveraging the GPU's parallel processing capabilities to expedite matrix operations and backpropagation during training.

Moreover, the experiments were performed using the PyTorch deep learning framework, which offers extensive support for neural network operations and optimization on GPU architectures.

The code snippets were executed in separate cells within the Jupyter notebook environment, ensuring modularity and ease of experimentation. The training and evaluation procedures were meticulously designed and executed to ensure reproducibility and accuracy in the model assessments. The following configuration was used :

- T4GPU
- Google Colab
- Python - 3.10.12
- PyTorch - 2.1.0+cu118

V. Evaluation

A. Model Performance Metrics

In evaluating the performance of our models for waste image classification, several crucial metrics were employed, including accuracy, precision, recall, F1-score, and confusion matrices.

1) ResNet-18 Model Evaluation: The ResNet-18, a Convolutional Neural Network (CNN), served as the backbone architecture for our classification task. Testing this model on a dataset comprising 2513 images balanced between Organic (O) and Recyclable (R) waste categories revealed an approximate accuracy of 88.3%.

Breaking down the classification performance, precision, recall, and F1-score metrics were calculated separately for each waste category ('O' and 'R'). These metrics provided deeper insights into the model's ability to correctly identify images within specific classes.

The confusion matrices, depicted in Figures 3 and 4, offer a comprehensive view of the model's classification performance. These matrices detail the true positive, true negative, false positive, and false negative predictions across the waste categories, giving a nuanced understanding of classification errors.

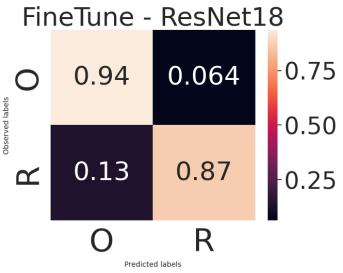

Fig. 3: Confusion Matrix: Custom-trained ResNet-18 Model

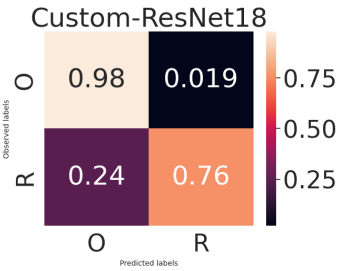

Fig. 4: Confusion Matrix: Pre-trained ResNet-18 Model

Furthermore, we conducted experiments with a Fine-Tuned and a Custom model, both trained across 10 epochs using a batch size of 64, a learning rate of 0.001, and the Adam optimizer.

The fine-tuned model, trained for approximately 2 hours and utilizing around 14.6 GB on T4GPU, exhibited an average accuracy of 95.6%. In contrast, the custom-built model, trained for 3.5 hours and consuming about 13.7 GB on T4GPU, achieved an average accuracy of 88%. This discrepancy suggests that the fine-tuned model, likely due to its larger dataset during training, outperformed the custom model.

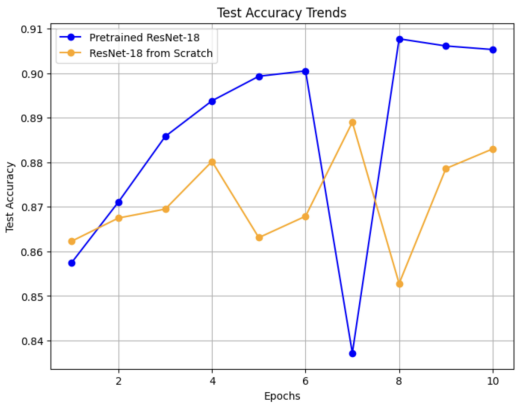

Fig. 5: Accuracy Comparison: Custom vs. Pre-trained ResNet-18

Moreover, a detailed bar chart comparison (Figure 5) illustrates the precision, recall, F1-score, and accuracy for both models. These metrics provide a visual overview, highlighting the varying performance aspects of the models across different evaluation parameters.

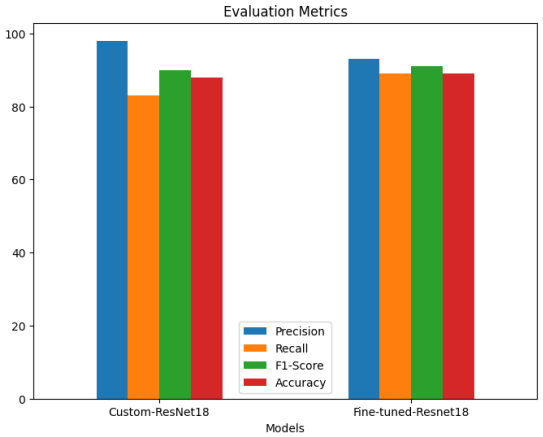

Fig. 6: Comparison of Precision, Recall, F1, and Accuracy

2) Transformer Model Evaluation: The Transformer architecture, specifically the Vision Transformer, was also utilized for waste image classification. This model underwent evaluation on the same test dataset, allowing a comparative analysis with the ResNet-18 model. We performed both the custom training of the model(from scratch) and fine tuning using the pre-trained model.

We generated confusion matrices for both models, showcasing their classification performance and F1 score.

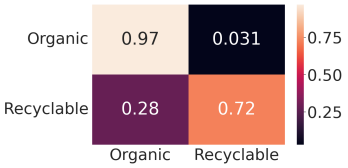

Fig. 7: Confusion Matrix for Custom-trained ViT Model

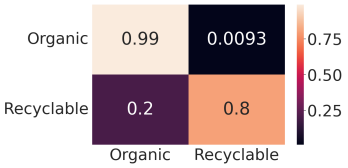

Fig. 8: Confusion Matrix for Pre-trained ViT Model

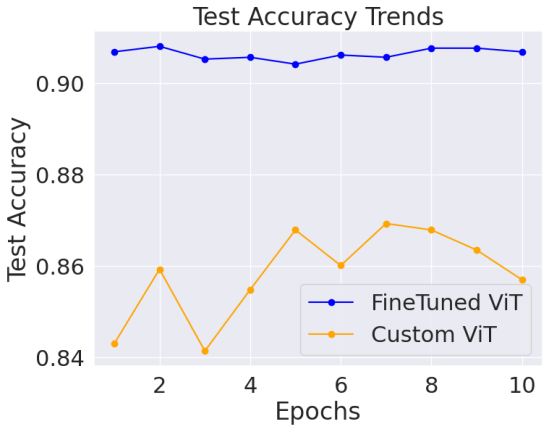

Fig. 9: Comparison of Accuracy for Custom and Pre-trained ViT Models

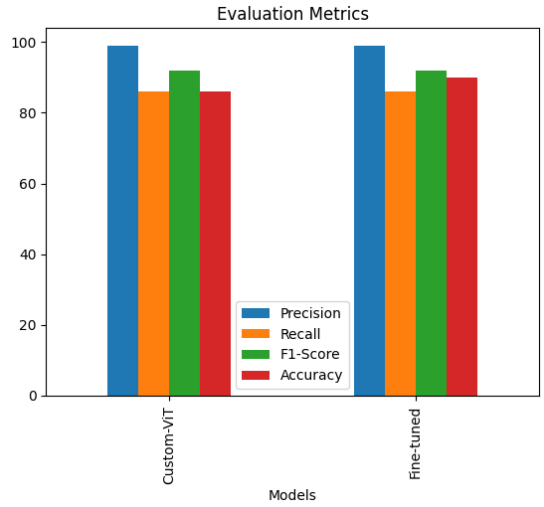

Fig. 10: Bar Chart Comparison of Precision, Recall, F1, and Accuracy ViT

The evaluation metrics, including accuracy, precision, recall, and F1-score, were computed for the Vision Transformer. The confusion matrix was generated to illustrate the model's performance in classifying Organic (O) and Recyclable (R) waste images.

B. Comparative Analysis

Comparing the performance metrics between the ResNet-18 and Vision Transformer models provides valuable insights into the strengths and weaknesses of each architecture in handling waste image classification tasks.

A comparative analysis of the ResNet-18 and Vision Transformer models elucidated the distinctive characteristics and efficacy of these architectures in waste classification scenarios. The analysis focused on examining the trade-offs, advantages, and suitability of each model in accurately classifying Organic and Recyclable waste images.

VI. Conclusion

Our comprehensive experimentation emphasized the critical role of model selection, particularly when working with constrained datasets. Contrary to opting for the highest-performing model outright, our analysis underscores the pragmatic advantage of leveraging Convolutional Neural Network (CNN) architectures. The ResNet18-FineTuned model emerged as the top performer with the highest accuracy of 95.7%, showcasing its robustness in accurately classifying waste images.

Moreover, the ResNet18-FineTuned model exhibited not just superior accuracy but also proved to be more time-efficient, requiring only 1 unit of training time. This efficiency highlights its viability in resource-constrained environments or scenarios demanding swift model deployment.

Models	**Accuracy**	**Train Time(GPU)**	**Precision**	**Recall**
ResNet18-Custom	88%	2.5	98	83
ResNet18-FineTuned	95.7%	1	93	89
ViT-Custom	86%	3	93	83
ViT-FineTuned	90%	2	99	86

While other models such as ResNet18 Custom, ViT Fine-Tuned, & ViT Custom demonstrated respectable performance, they fell slightly short in either accuracy, training time, or both when compared to the ResNet18-FineTuned model.

Therefore, our study underscores the importance of considering not just raw performance metrics but also training efficiency when selecting models, especially in scenarios with limited data availability. This pragmatic approach ensures a balanced consideration of accuracy and resource utilization, leading to more effective and efficient model deployment strategies

VII. Future Work

In future endeavors, expanding the scope of this study involves exploring and evaluating various convolutional neural network (CNN) architectures beyond ResNet-18. Models such as Inception, AlexNet, and VGG present intriguing alternatives, and their performance will be thoroughly examined and compared in the context of waste image classification. This comparative analysis aims to discern the strengths, weaknesses, and overall suitability of these architectures for the specific task, potentially uncovering more efficient or accurate models for waste classification scenarios.

Furthermore, scaling up the dataset size and diversity could offer valuable insights into the scalability and robustness of these models. Training CNNs on larger datasets allows for a more comprehensive understanding of their generalization capabilities and performance under varying conditions.

Additionally, in the realm of transformer architectures, the investigation will extend to explore newer variants like the Swin Transformer. Assessing the Swin Transformer's applicability and efficacy in waste image classification tasks provides an opportunity to delve into evolving transformer architectures and their advantages compared to conventional CNNs.

Moreover, experimentation with a broader range of hyperparameters, including learning rates, batch sizes, and optimization strategies, remains a key aspect of future investigations. Fine-tuning these parameters could significantly impact model performance, and comprehensive hyperparameter tuning aims to extract the optimal configuration for improved accuracy and efficiency.

Overall, the future trajectory of this study involves a systematic exploration of diverse CNN architectures, scaling experiments to larger datasets, and embracing newer transformer models. These endeavors aim to refine waste image classification methodologies, fostering the development of more accurate, scalable, and adaptable models for sustainable waste management solutions.

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Appendix

Repository

Code for all the experiments, data preprocessing, and data analysis can be found in the following Github repositories:

github : https://github.com/Shreyjaradi/ImageClassification-CNN-vs-Transformer The code has been written in Python and uses popular libraries such as PyTorch, MatPlot, Pandas, Numpy, and Scikit-learn.The repository is private right now and the access is only to authors. Please request access on sjaradi@hawk.iit.edu/ kkumarkaiplody@hawk.iit.edu

Individual Contributions

Shrey Jaradi: Shrey Jaradi primarily focused on the design, implementation, and experimentation with Convolutional Neural Network (CNN) models for waste image classification. He meticulously curated the dataset, developed and fine-tuned the CNN-based ResNet-18 model, and executed extensive experiments to assess its performance. Shrey was responsible for training, evaluating, and refining the CNN architecture, optimizing hyperparameters, and generating model evaluation metrics, including accuracy, precision, recall, and F1-score. Additionally, he prepared visualizations, such as confusion matrices and performance graphs, to illustrate the results effectively.

Karthik Kumar Kaiplody: Karthik Kumar Kaiplody spearheaded the exploration, implementation, and evaluation of Transformer-based architectures, specifically focusing on Vision Transformer models for waste image classification. He extensively researched and developed the custom Transformer architecture, orchestrated the model's training, fine-tuning, and optimization strategies. Karthik conducted experiments, assessed the Transformer's adaptability to the waste classification task, and generated insights into its performance metrics. He also contributed to the report's methodology, architectural descriptions, and evaluation sections concerning the Transformer-based models.

Both authors actively collaborated, brainstormed ideas, shared insights, and collectively contributed to the conceptualization and preparation of the report. Their joint efforts ensured a comprehensive exploration of diverse deep learning architectures, fostering an inclusive and holistic understanding of Convolutional Neural Networks and Transformers in the context of waste image classification.