CV703 - Assignment 1

Umair Nawaz, Tooba Tehreem Sheikh, Ufaq Khan

A. Introduction

This report articulates the development and formulation of an innovative methodology that integrates Convolutional Neural Networks (ConvNets) with Vision Transformers (ViTs) to enhance the accuracy of fine-grained image classification. The research primarily concentrates on distinct datasets, namely the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [1], the Fine-Grained Visual Classification of Aircraft (FGVC-Aircraft) [5], and the FoodX [3]. The objective was to first train and then augment a baseline model on these datasets by incorporating a novelty in it. We used a ConvNext [4] model as a baseline and then introduced a module of vision transformer (ViT) architecture [2], a leading-edge model recognized for its excellence in image classification. This integration fosters a hybrid model that adeptly captures both the minute details and global patterns within images, a critical factor in the precise classification of fine-grained datasets. The enhanced model embodies a significant advancement in the field, demonstrating the potential of combining ConvNet and ViT architectures for superior image classification performance. Data of this assignment is available at Google Drive

Figure 1. Samples from all three datasets

B. Problem Statement

B.1. Fine-Grained Image Classification

Fine-grained image classification presents a more complex challenge than conventional classification techniques, as it requires distinguishing between subcategories within the same class (intra-class), such as different species of birds, in contrast to identifying distinctions between separate classes (inter-class), for example, differentiating an airplane from a bird. This task is fraught with multiple difficulties, including but not limited to, issues of high dimensionality, the risk of overfitting, and optimization challenges, particularly when employing lighter network architectures. Moreover, when undertaking the task of combined classification, one encounters a notable degree of data imbalance, especially when comparing the scope and diversity of classes across datasets such as the CUB-200-2011 and FGVC-Aircraft. Such imbalance requires meticulous strategy formulation to ensure the classification methodology is both efficient and effective.

In this assignment, we aim to explore and surmount these challenges, aiming to develop a sophisticated multi-class image classification framework that is adept at accommodating the varied characteristics and detailed distinctions of the CUB, FGVC-Aircraft, and FoodX datasets.

B.2. CNN Vs Transformers

For the past decade, ConvNets have been the predominant model for vision-related tasks, attributed to their spatial inductive biases which facilitate learning representations efficiently, utilizing fewer parameters and necessitating reduced training durations. Nonetheless, these networks inherently focus on spatial locality and do not inherently capture global representations. Conversely, ViTs, which are predicated on self-attention mechanisms, excel in modeling long-distance dependencies within data. The remarkable achievements of transformers within the Natural Language Processing (NLP) field have catalyzed academic inquiries into their applicability for vision tasks, where their application has recently shown promising outcomes.

C. Datasets

The proposed architecture was tested on three different datasets for fine-grained image classification. A visual representation of samples from all three datasets is presented in Figure 1.

C.1. CUB-200 [1]

The Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset is an image dataset of North American bird species published by Caltech UCSD and stands as a preeminent resource for the fine-grained visual categorization task, extensively employed in research and development. Comprising a total of 11,788 images, the dataset encompasses 200 subcategories of birds, with 5,994 images designated for training and 5,794 for testing.

C.2. FGVC [5]

FGVS is a second dataset also called as Fine-Grained Visual Classification of Aircraft (**FGVC-Aircraft**) dataset, which consists of 10,200 images. 102 different aircraft model variants are provided with 100 samples each.

C.3. FoodX-251 [3]

The third dataset, denoted as **FoodX**, encompasses 158k images (118k for training, 12k for validation, and 28k for testing) featuring 251 diverse food categories, including cakes, soups, puddings, and more. Originally employed in organizing the iFood-2019 challenge at the Conference on Computer Vision and Pattern Recognition (CVPR) in 2019.

D. Background Study

D.1. ConvNext

ConvNeXt, developed by Facebook AI Research, has been a state-of-the-art advancement in Convolutional Neural Networks (ConvNets), built from ConvNet modules while integrating key design elements from the Swin Transformer architecture [4]. The evolution began with a standard ResNet-50 model, which was modified in several key ways inspired by transformers. Notably, its block structure was altered from (3, 4, 6, 3) to (3, 3, 9, 3), reflecting the Swin Transformer's design.

The introduction of a patchify strategy using a 4×4 convolution layer with a stride of 4 aimed to reduce input redundancy. Adopting depthwise convolution, akin to the approach in ResNeXt, served as an alternative to the self-attention mechanism, allowing for efficient channel-wise processing. Innovations such as an inverted bottleneck design and a 7×7 depthwise convolution kernel were implemented to improve the accuracy-efficiency trade-off, boosting accuracy from 78.7% to 80.6%. The ConvNeXt model's architecture, depicted in Figure 2, illustrates these modifications.

Further micro-adjustments enhanced the model's performance, including replacing ReLU with GELU as the primary activation function, simplifying batch normalization to a single-layer norm before the point-wise convolution layer, and introducing a 2×2 downsampling layer at each network stage, mirroring the Swin Transformer's structure. These improvements collectively raised the model's accuracy to 82.0%.

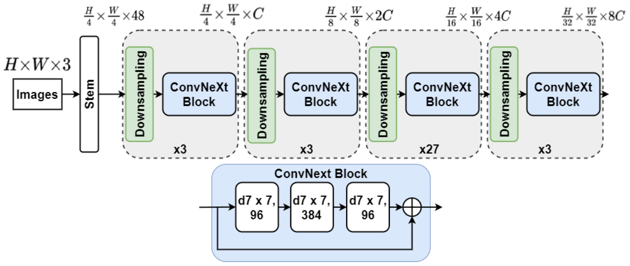
$\frac{H}{4}\times\frac{W}{4}\times 48$
$\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
$H\times W\times 3$
Images
Stem
Downsampling
ConvNeXt Block
x3
Downsampling
ConvNeXt Block
x3
Downsampling
ConvNeXt Block
x27
Downsampling
ConvNeXt Block
x3
ConvNext Block
d7 x 7, 96
d7 x 7, 384
d7 x 7, 96

Figure 2. High-level architecture of ConvNeXt-L

D.2. Vision Transformer

The Vision Transformer (ViT) introduces a significant shift in computer vision, moving beyond the traditional convolutional neural networks (CNNs) to offer a versatile solution for tasks like image classification, detection, and segmentation, achieving top-tier performance. ViT starts by partitioning an input image into uniform patches, converting these into linear embeddings, and adding positional embeddings to maintain spatial context. This transforms the image into a sequence of tokens, akin to text processing in NLP transformers.

The sequence undergoes processing by the transformer encoder, which alternates between multi-headed self-attention and multilayer perceptrons (MLPs), incorporating layer normalization and residual connections in line with the original NLP transformer design. The self-attention mechanism enables ViT to assess the relevance of different patches for comprehensive image understanding, capturing wide-ranging dependencies within the image. The processed output is then funneled through a classification head, usually a linear layer, for the final prediction. This approach allows ViT to learn global visual representations, setting it apart from the localized processing of CNNs.

E. Proposed Architecture

In this assignment, we introduce a novel architecture designed to enhance the capabilities of ConvNeXt by integrating it with the structural elements of the Vision Transformer (ViT). This integration results in a hybrid model that not only meets but surpasses the baseline performance of ConvNeXt in terms of accuracy while maintaining the original model's parameter count and computational efficiency (FLOPs) i.e., under the 5% limit as shown in Table 1. Notably, this development was achieved without the necessity for pre-training the composite model on the ImageNet1K dataset, relying instead on the separate pre-trained components.

The ConvNeXt architecture, as described, includes four primary blocks following the initial stem. Our proposed configuration merges these four ConvNeXt-L stages with the encoder block of the ViT, which is characterized by its multi-head attention module. This strategic addition of attention to the network's conclusion significantly enhances the feature representation extracted by the ConvNeXt-L model. It presupposes that the ConvNeXt-L has already identified the most pertinent features for classification, which are subsequently refined by the ViT's main block to discern even finer details, thereby improving learning accuracy. A linear layer is appended following the attention block to map to the desired number of output classes. Among various configurations explored, this particular architecture demonstrated superior accuracy outcomes. Alternative combinations, such as merging the initial two ConvNeXt-L blocks with the final two Swin blocks, were considered but found lacking during the training phase due to insufficient learning, underscoring the necessity for training from scratch on ImageNet. Consequently, the decision to incorporate the attention module at the terminal stage of the network was validated.

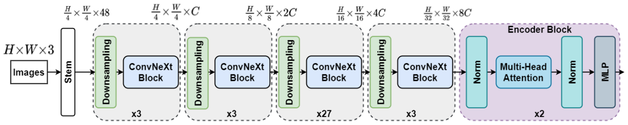
$H \times W \times 3$
Images
Stem
$\frac{H}{4} \times \frac{W}{4} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Downsampling
ConvNeXt Block
x3
Downsampling
ConvNeXt Block
x3
Downsampling
ConvNeXt Block
x27
Downsampling
ConvNeXt Block
x3
Encoder Block
Norm
Multi-Head Attention
Norm
MLP
x2

Figure 3. High-level architecture of ConvNeXt-ViT.

F. Experimental Setup

For implementation, we developed a hybrid model named ConvNeXt-ViT that merges the spatial efficiency of ConvNeXt-L with the Vision Transformer's (ViT) attention mechanisms, harnessing the strengths of both to process information effectively. We structured the training in two phases, starting with a linear learning rate scheduler for the first 20 epochs, followed by a Cosine scheduler, optimizing the model's performance gradually and effectively.

Using the AdamW optimizer for its weight decay management, we fine-tuned hyperparameters to prevent overfitting, supported by a robust data augmentation strategy to enhance generalizability. The model's efficacy was evaluated through training and testing accuracy.

Experiments were performed on NVIDIA Quadro RTX 6000 with a batch size of 32. We relied on PyTorch for its flexibility, complemented by libraries like pandas and NumPy for data manipulation, tqdm for monitoring training progress through intuitive progress bars, and timm for streamlined model loading and management.

G. Results

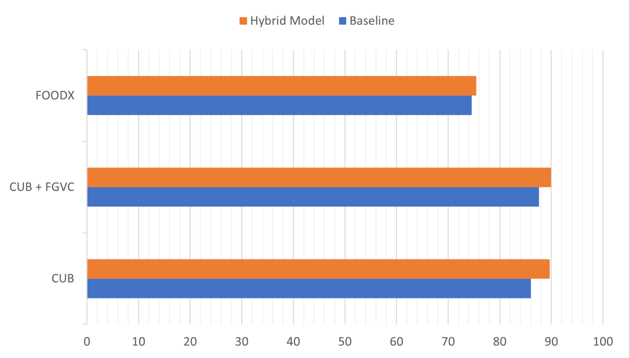

Figure 4. Accuracy results on Testing split for each dataset

In this section, the results obtained upon fine-tuning process for both the ConvNeXt and ConvNext-ViT model across the selected datasets (CUB, a combination of CUB and FGVC, and FoodX) are explained as illustrated in Figure 4. A notable enhancement in performance was observed on the CUB dataset, where accuracy escalated from 86.05% to 89.69%. In the case of the combined CUB and FGVC dataset, an increase of 2.36%, from 87.58% to 89.94%, was recorded. The FoodX dataset, characterized by its extensive diversity and volume of images, presented a more challenging scenario. Here, the baseline model achieved an accuracy of 74.60%, while the hybrid model showed a slight improvement, reaching an accuracy of 75.42%. These results underscore the efficacy of the hybrid model, particularly in handling datasets with intricate and varied visual content.

Table 1. Comparison of Baseline and Hybrid Model Characteristics

Architecture	GFLOPS	Params (M)
Baseline	34.3602816	196.450376
ConvNext-ViT Model	34.3603062	196.45652

H. Conclusion

In conclusion, our results led us to understand that early convolution makes transformers (such as ViT) see better. Our hybrid model's accuracy exceeds the traditional ConvNext-L model for all the 3 datasets also referred to as CUB, CUB + FGVC, and FoodX datasets, only by finetuning. We are confident that if our hybrid model is pre-trained on ImageNet, the accuracy will be boosted by a good margin. However, this was not available in our study, because of the limited computational resources we were provided with. Our contributions in this work are: (1) Fine-tune the ConvNext model on three datasets and compare their results; (2) Propose a hybrid model, ConvNeXtViT achieving better results than the individual architecture; (3) Compare the proposed model's performance with the baselines.

References

[1] Cub-200-2011 dataset — papers with code. https://paperswithcode.com/dataset/cub-200-2011. (Accessed on 02/01/2024).

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: A dataset for fine-grained food classification, 2019.

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[5] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.