Fine-Grained Image Classification with ConvNeXt and SCR-Attention

Rusiru Thushara, Mohamed Insaf Ismithdeen, Chaimaa Abi

Mohamed bin Zayed University of Artificial Intelligence

{rusiru.thushara, mohamed.ismithdeen, chaimaa.abi}@mbzuai.ac.ae

Abstract

This study presents a comprehensive approach to fine-grained image classification on three distinct datasets: CUB birds, a combined dataset of CUB birds and FGVC Aircraft, and FoodX dataset. The primary objective was to enhance classification performance without significantly exceeding the computational FLOPs of the chosen baseline model, ConvNeXt-V2-Large, by more than 5%. The baseline ConvNeXt-V2-Large model achieved an accuracy of 89.49% on the combined dataset surpassing the threshold of 88.00%. Through strategic adjustments and incorporation of Selective Channel Recalibration Attention (SCRA), we improved this accuracy to 89.99% while staying within the set computational constraints.

1. Introduction

Fine-grained image classification is a specialized process in image analysis that aims to identify and categorize objects within narrowly defined, specific subgroups of larger categories. This method presents significant challenges due to high intra-class variance, where objects within the same category exhibit considerable differences, and low inter-class variance, where objects across different categories show minimal distinctions.

We begin by evaluating the performance of a pre-trained baseline model, which we fine-tune for the specific tasks at hand. To assess the model's adaptability and accuracy in fine-grained image classification, we evaluate on the following benchmark datasets: the CUB birds [5](Task 1), combined dataset of CUB birds and FGVC Aircraft [8](Task 2), and FoodX dataset [3](Task 3). Furthermore, we explore the impact of various modifications, ranging from data-level changes through data augmentation to adjustments in loss computations using regularization, and architectural changes to the original model.

2. Related Work & Contributions

We analyzed different backbone architectures and their performance on the fine-grained datasets. We noticed that the recent network architecture, ConvNeXt V2[7] pre-trained on ImageNet 22k, performs much better with simple transfer learning. ConvNeXt V2 demonstrates the efficiency of convolutional neural networks(CNNs) in the backdrop of the increasing popularity of Vision Transformers[2]. Further, they argue the state-of-the-art performance of ViTs is not solely due to the backbone design choice of transformers over CNNs but is an aggregate of improved training recipes and macro and micro architecture choices. They show that applying similar techniques on a CNN-based network provides comparable and often surpassing the performance of ViTs. We use the ConvNeXT V2 Large model pre-trained on ImageNet 22k as our baseline model and perform extensive experiments.

We summarize our contributions as follows:

- We assess the performance of transfer learning in fine-grained image classification, and adaptation of models to specific tasks.
- We explore the impact of diverse modifications on the ConvNeXt V2 Large baseline model's performance, providing a comprehensive understanding of their effects.
- We explore a diverse range of modifications at various levels encompassing adjustments to image data, refinement of the loss function, and enhancements to the baseline architecture.

3. Methodology

3.1. Proposed Modifications

In the initial phase of our study, we fine-tuned a pre-existing baseline model, ConvNeXt V2 Large [7]. The original final classification layer, designed for a standard 1000-class scenario, was substituted with a novel linear layer customized to align with the distinct class counts in our experimental datasets: 200 for Task 1, 300 for Task 2, and 251 for Task 3. Following this modification, the entire model underwent fine-tuning, with the newly integrated classification layer being trained from scratch.

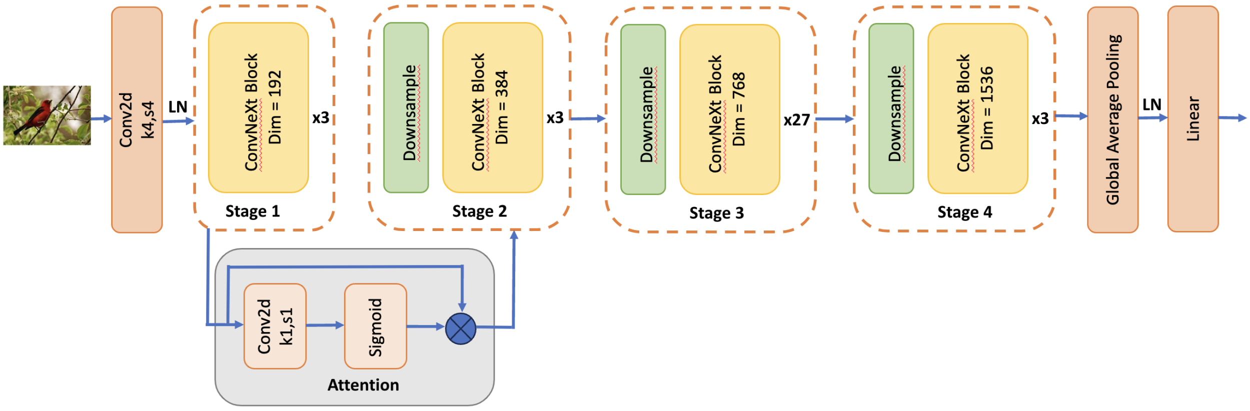

Figure 1: Proposed Architecture: We introduce SCRA block between the stage 1 and stage 2 of the ConvNeXt V2 Large Architecture

After the initial fine-tuning of the ConvNeXt V2 Large model, we conducted a series of experiments aimed at enhancing its performance. These enhancements encompassed a range of modifications as L2SP regularization, Dropout regularization, Data augmentation techniques, and Selective Channel Recalibration Attention (SCRA). The list of these modifications is presented below:

3.1.1. Data Augmentation

We employed horizontal flipping and rotation transformations to the data. By doing so, the model learns to identify and understand objects in images that are not perfectly aligned or are in unusual positions. We also applied color jitter data augmentation to enhance model robustness against image color variations. This technique randomly alters brightness $I' = I \times \beta$, contrast $I' = I \times \alpha$, saturation $I' = \text{sat}(I, \gamma)$, and hue $I' = \text{hue}(I, \delta)$ of the images. The parameters $\beta, \alpha, \gamma, \delta$ are set after extensive experiments, introducing controlled variability and aiding the model's ability to generalize from diverse visual inputs.

3.1.2. Loss level Modifications

L2SP (L2 Soft Penalty) Regularization'[4]: We introduced an additional term to the loss function, applying the L2SP regularization method. Unlike traditional L2 regularization, which penalizes weights based on their magnitude relative to zero, L2SP penalizes the deviation of the weights from a specific 'starting point,' the weights of a pre-trained model. This approach helps in retaining the knowledge from the initial training while allowing necessary adaptations.

The adjusted loss function for the classification task with L2SP is:

$$\text{Total Loss} = \underbrace{-\sum_{i=1}^{m}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic})}_{\text{Cross-Entropy Loss}} + \underbrace{\lambda\sum_{i=1}^{n}(w_i - w_{i,\text{pretrained}})^2}_{\text{L2SP Term}}$$

- λ is the regularization parameter controlling the extent of regularization.
- w_i represents the current weights of the model.
- $w_{i,\text{pretrained}}$ represents the pre-trained weights.
- n is the number of weights in the model.

3.1.3. Architectural Modifications

Dropout: In the ConvNeXt Large V2 architecture, we introduced a dropout regularization layer with a probability of $p = 0.45$ right before the final softmax layer. This regularization technique randomly sets a subset of activations to zero during training, preventing co-adaptation of neurons and enhancing the model's generalization capabilities on the test data.

Selective Channel Recalibration Attention (SCRA): The attention mechanism operates on the principle of feature recalibration by applying a per-channel weighting that is learned in a data-driven manner. The SCRA block (Figure:1) takes in the output feature map of stage 1 and processes it through a 1x1 convolutional layer, maintaining the spatial dimensions while allowing channel interaction. The resultant feature map is then passed through a sigmoid activation function to create an attention map, assigning weights between 0 and 1 to each feature, signifying their importance. This attention map is element-wise multiplied with the original feature map, effectively gating the features by amplifying important ones and diminishing the less important, resulting in an output feature map of the same size, where the network's focus is adjusted towards the most informative features for the task. This idea was inspired by the work [1].

3.2. Dataset

For model training and evaluation, the following datasets are used:

– **CUB-200-2011 Dataset:** This dataset comprises 11,678 images spanning 200 bird species. It is widely used for fine-grained image classification tasks, particularly for bird species recognition. The training set consists of 5,884 images, while the test set comprises 5,794 images.

Dataset	CUB Birds	CUB Birds + FGVC Aircraft	FoodX	GFLOPS	Parameters (M)
Baseline	89.49	89.48	78.62	68.72	196.660
Baseline + Augmentation	89.47 (-0.02)	89.82 (+0.34)	-	68.72	196.660
Baseline + Augmentation + L2SP	89.42 (-0.07)	89.78 (+0.30)	-	68.72	196.660
Baseline + Augmentation + L2SP + Dropout	89.6 (+0.11)	89.64 (+0.16)	-	68.72	196.660
Baseline + Augmentation + L2SP + Dropout + SCRA	-	89.96 (+0.48)	-	68.80	196.881
Baseline + Augmentation + Dropout + SCRA	**90.4 (+0.91)**	**89.99 (+0.51)**	**79.44 (+0.82)**	**68.80 (+0.11%)**	**196.881 (+0.11%)**

Table 1: Comparison of top-1 accuracy between the baseline with different modifications

- **Combined Dataset (CUB-200-2011 + FGVC Aircrafts):** This dataset combines 200 bird species from the CUB-200-2011 and 100 aircraft models from the FGVC Aircrafts, totaling 300 classes. It includes 12,661 images in the training set and 9,127 images in the test set.
- **FoodX Dataset:** Comprises 123,841 samples distributed among 251 food categories. The training set contains 111,847 images, while the test set has 11,994 images.

3.3. *Implementation Details*

In the study, the pre-trained ConvNeXt V2 Large model from the timm[6] library is used. For data preprocessing, all images are resized to 224x224 pixels, aligning with ConvNeXt's default input size, and normalized to standardize pixel values. During fine-tuning, the Adam optimizer is employed with a learning rate of 1e-4 and a weight decay of 1e-5, complemented by a learning rate scheduler that adjusts the rate every 5 epochs with a reduction factor of 0.1. The study introduces specific modifications for enhanced performance: L2SP regularization with a weight (λ) of 0.01, a dropout layer with a 0.45 probability to improve generalization, and data augmentation techniques including random rotations up to 15 degrees and horizontal flipping, each with a 25% probability. We set the values for Color Jitter as brightness=[0.75, 1.25], contrast=[0.75, 1.25], saturation=[0.75, 1.25], hue=[-0.05, 0.05]. We used 30 epochs and a batch size of 128 to train all the models on an NVIDIA 80G A100 GPU.

4. Results

Table 1 presents the top-1 accuracy achieved through various experimental modifications applied to our models. A key observation is that the highest performance on the combined dataset (CUB + FGVC datasets) was achieved with a combination of data augmentation(horizontal flipping, rotation and color jitter), dropout, and SCRA yielding a notable accuracy increase of 0.51% to reach 89.99% with respect to the baseline performance. Moreover, this increase is more apparent for CUB Birds (Task1) and FoodX datasets (Task3).

The observed decline in performance when employing L2SP regularization may be attributed to its inherent design, which promotes the retention of features from the pre-trained model. While this approach is beneficial in certain contexts, it presents a potential drawback for fine-grained classification tasks that necessitate the acquisition of substantially new or distinct features.
We also observed that the use of data augmentations and dropout increased accuracy in general. Incorporation of SCRA to the basline, ConvNeXt V2 Large model showed an accuracy increase across all datasets while adding minimal computational overhead in terms of GFLOPS and Model Parameters.

5. Conclusion

In this study focused on fine-grained image classification, we initially fine-tuned a baseline ConvNeXt V2 Large model. Subsequently, we conducted extensive experiments aimed at enhancing model performance through various configurations. Our findings reveal that the use of data augmentations, dropout, and SCRA gives the best performance improvement across all datasets.

References

[1] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns, 2020.

[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. *arXiv preprint arXiv:1907.06167*, 2019.

[4] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[5] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Jul 2011.

[6] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[7] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023.

[8] Yabin Zhang, Hui Tang, and Kui Jia. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In *Proceedings of the european conference on computer vision (ECCV)*, pages 233–248, 2018.