# Learning with Knowledge Distillation for Fine Grained Image Classification

Muhammed Uzair Khattak Adnan Khan Khaled Dawoud
Mohmmad Bin Zayed University of Artificial Intelligence
{muhammad.uzair, adnan.khan, khaled.dawoud}@mbzuai.ac.ae

## Abstract

*Fine-grained Image Classification (FGIC) is one of the challenging tasks in Computer Vision. Many recent methodologies including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have tried to solve this problem. In this study we show the effectiveness of using both CNNs and ViTs hand in hand to produce state of the art results on challenging FGIC datasets. We show that by using DeiT as student model and ConvNext as teacher model in knowledge distillation settings, we achieve top1 and top5 accuracies of 92.52% and 99.15% respectively on combined CUB + Stanford Dogs datasets. On a more challenging dataset named FoodX-251 we achieved top1 and top5 accuracies of 74.71% and 92.99% respectively.*

## 1. Introduction

Humans have the natural ability to classify fine-grained objects in the scenes. A human can inherently differentiate between the dog, the bird and in addition, also differentiate a sparrow from an ostrich. The motivation of Fine-Grained Image Classification (FGIC) is to make computers able to understand and recognize the sub-ordinate categories (Sparrow and Ostrich) of a basic level category (Birds).

In the past decade, computer vision research took a leap from traditional feature engineering to Convolutional Neural Networks (CNNs) architectures. Deep learning [1] methods proved to be the de-facto research line for classification tasks since AlexNet [2] won the ILSRVC [3] in 2012. Deep learning is also proved to be a robust tool for learning discriminative features and there is significant progress made in the field of FGIC [4, 5, 6]. Remarkable progress is being made using (CNNs) to learn weakly-supervised models for FGIC [7, 8] . However back in 2020, the introduction of Vision Transformers (ViT)[9] outperformed ResNets on an image classification task. After that many such attention-based architectures [10, 11, 12, 13] are introduced for computer vision tasks which are based on global attention design.

In this study, we combine the best of the best architectures from both the CNNs and Transformers families to apply them to the problem of FGIC. We experimented with different models and chose Data-Efficient Image Transformers (DeiT) [14] and ConvNext [15] which works best hand in hand on CUB200[16] , Stanford Dogs[17], and Food251[18] datasets. We use DeiT as student model and ConvNext as teacher model in knowledge distillation settings. fine tuning the teacher model on CUB+Stanford Dog Dataset we could achieve top1 and top5 accuracies of 92.52% and 99.15%. Our main contributions of this study are as follows:

- We propose a hybrid model based on DeiT and ConvNext for FGIC.
- We conduct comprehensive experiments on three difficult datasets namely CUB200, Stanford Dogs, and Food251 to achieve state-of-the-art results.

## 2. Methodology

### 2.1. Background

In this section, we briefly discuss the selected classification models from vision transformers and CNNs family.

**Data-Efficient Image Transformers (DeiT):** Vision Transformers, despite giving good performance on vision tasks, they were still considered difficult to train mainly because of the large pretraining dataset requirement. The maximum performance of ViTs was achieved when they were pretrained on a large scale private dataset, JFT-300M [19]which is the superset of ImageNet-1k and ImageNet-21k. In practice, the availability of such large-scale datasets are very difficult to obtain as well as it requires a lot of compute to train the models.

To introduce improvements for such constraints, DeiT [14] was introduced which is similar to ViTs in design but it performs better than ViTs, when both are trained solely on only ImageNet dataset. DeiT used two strategies for their improvement against standard ViTs, (i) long training schedules along with strong data-augmentations (such as cutmix, mixup, randaugment etc), (ii) knowledge distillation (KD) teacher-student model training, where teacher model,

| Dataset | Train Size | Test Size | # classes |
|---|---|---|---|
| FoodX | 118K | 28K | 251 |
| CUB | 5.9K | 5.8K | 200 |
| Stanford Dogs | 12K | 8.6K | 120 |
| CUB + DOG | 18K | 14K | 320 |

Table 1. Number of train and test images along with classes for different datasets. FoodX dataset contains the highest number of train-test images.

preferably taken from CNN family helps DeiT model to train and learn from it. For KD, the modify a original ViT by adding another learnable token, known as the distillation token. This KD token interacts with the final logits of teacher model and tries to model teacher model capabilities in the DeiT student model, thus improving the overall performance of DeiT.

**ConvNext:** Motivated from the architectural and training strategies of vision transformers, a new family of CNNs called ConvNexts are recently introduced which has shown very competitive results on major computer vision tasks. Just like the vision transformers used some of the design principles (such as windowing attention etc) from CNNs, ConvNexts has done the same exact counterpart, i-e implementing new design and training ideas from the vision transformer models, specifically SWIN transformers. ConvNext model is built on vanilla ResNet model and it has modernized it with several macros (block design changes such as inverted bottleneck design) and micro modifications, totally inspired from SWIN transformers.

**ResNet BiT:** To add more choice in the selection of teacher model, we also experiment with ResNet Big Transfer models [20]. ResNet BiT models are improved versions of standard ResNet which incorporates group normalization and standardized convolution instead of batch normalization and normal convolution respectively. These changes show improved performance on top of the vanilla ResNet model.

### 2.2. Approach

**Task 1 and task 3:** For CUB and FoodX dataset, we finetune ConvNext-B and DeiT-B models which are pretrained on imageNet-1k. For finetuning, we remove the final classification head of the model, and replace it with linear layer having output nodes equal to the number of classes of respective finetuning dataset. On the test sets, we report the top1 and top5 accuracies along with other hyperparameters.

**Task2:** For CUB+DOG dataset, we adopt the approach of knowledge distillation during the fine tuning stage. Particularly, we use ConvNext and ResNet BiT as the teacher models, initialized with imageNet-1k pretraining. They are then finetuned on CUB+DOG dataset. For the student model, we choose DeiT-B distilled model, which is also initialized with imageNet-1k pretraining (with distillation). Further, we finetune that DeiT-B on CUB+DOG dataset. We set this model as our baseline. For improving the results, we finetune again the student model using CNN teacher models with several different configurations. For knowledge distillation settings, we study the effect of hard distillation as well as soft distillation [14]. We carry out extensive experimentation and review the performance capabilities of models in such learning settings.

### 2.3. Training details and Dataset Preprocessing

Pretrained weights for imageNet-1k training for DeiT-B, ConvNext-B and ResNet-Bit are obtained from timm library in Pytorch. These models are finetuned on all respective datasets afterwards. Models on CUB and CUB+DOG dataset are finetuned for 60 epochs while for Food dataset, models are finetuned for 30 epochs. Batch size 16 is used for all experiments.

## 3. Results and Experimentation

**Validating results on CUB and FoodX datasets:** For CUB and FoodX datasets, we report the finetuning results of ConvNext-B, DeiT-B and DeiT-B distilled models in Table 2. For the CUB dataset, Convnext-B models provides the highest top1 accuracy of 84.55% while the transformer based DeiT model achieves 83% top1 accuracy. Interestingly, the DeiT distilled model, performs even worse and achieves top1 accuracy of 70.5%. We also tried training the ConvNext-B model from scratch, but it did not give any good performance. For the FoodX dataset, ConvNext-B model again provides the best accuracy in comparison to Deit-B models, achieving top 1 accuracy of 74.77% on the test dataset. The corresponding DeiT-B and Deit-B Distilled model achieves top1 accuracy of 68.33 and 68.15 respectively.

**Validating results on CUB+DOG dataset:** Results on the test set of CUB+DOG dataset on various models are shown in Table 2-b. The selected two teacher models, imageNet pretrained ConvNext-B and ResNet BiT are finetuned on the CUB+DOG dataset and they provide top1 accuracy of 91.91% and 85.42% respectively. Our baseline model, Deit-B distilled achieves finetuning top1 accuracy of 91.17% alone.

In the second step, we perform knowledge distillation learning of fine-tuned Deit-B Distilled model using ConvNext-B and ResNet BiT respectively. From the results in Table 2, we can see that our baseline model, when finetuned with the help of ConvNext improves the top1 accuracy by about 1.5% from 91.17% to 92.52%. It even surpasses the teacher model accuracy, which is 91.92 (top1). On the other hand finetuning with ResNet-BiT top of (CUB+FOOD) finetuned Deit-B distilled model provides top1 accuracy of 89.22%. The performance from knowl-

| Sub-task (a): CUB | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Experiment # | Model | | | Image size | | Pretraining | Acc@1 | Acc@5 | Test loss | Parameters |
| **1** | **ConvNext-B** | | | **384** | | **Imagenet 1K** | **84.55** | **97.72** | **0.737** | **87M** |
| 2 | ConvNext-L | | | 224 | | From scratch | 2.12 | 7.56 | 5.052 | 87M |
| 3 | Deit-B_384 | | | 384 | | Imagenet 1K | 83 | 93.75 | 1.05 | 86M |
| 4 | Deit-B_384_distilled | | | 384 | | Imagenet 1K | 70.5 | 91.69 | 1.44 | 86M |
| **Sub-task (b): CUB + Dogs** | | | | | | | | | | |
| Experiment # | Model | Distillation | Train recipe | Image size | | Pretraining | Acc@1 | Acc@5 | Test loss | Parameters |
| 1 | ConvNext-B | N/A | N/A | 384 | | Imagenet1k | 91.91 | 99.10 | 0.33 | 88M |
| 2 | Deit-B | No | Deit-B | 384 | | Imagenet1k | 80.96 | 95.9 | 0.90 | 87M |
| 3 | Deit-B | No | ConvNext | 384 | | Imagenet1k | 82.5 | 96.04 | 0.87 | 87M |
| 4 | Deit-B Distilled | No | ConvNext | 384 | | Imagenet1k | 91.17 | 98.65 | 0.614 | 87M |
| 5 | Deit-B Distilled (Imagenet finetuned) | ConvNext (soft) | Deit-B | 384 | | Imagenet1k | 86.03 | 97.74 | 0.6054 | 87M |
| 6 | Deit-B Distilled (Imagenet finetuned) | ConvNext (hard) | Deit-B | 384 | | Imagenet1k | 88.53 | 98.48 | 0.422 | 87M |
| **7** | **Deit-B Distilled (CUB+Dogs finetuned)** | **ConvNext (hard)** | **Deit-B** | **384** | | **Imagenet1k** | **92.52** | **99.15** | **0.36** | **87M** |
| 8 | BiT | No | ConvNext | 384 | | Imagenet1k | 85.42 | 97.57 | 0.677 | 87M |
| 9 | Deit-B Distilled (Imagenet finetuned) | ResNet BiT (hard) | Deit-B | 384 | | Imagenet1k | 87.49 | 98.24 | 0.46 | 87M |
| 10 | Deit-B Distilled (CUB+Dog finetuned) | ResNet BiT (hard) | Deit-B | 384 | | Imagenet1k | 89.224 | 98.77 | 0.39 | 87M |
| **Sub-task (c): FOOD** | | | | | | | | | | |
| Experiment # | Model | | | Image size | | Pretraining | Acc@1 | Acc@5 | Test loss | Parameters |
| **1** | **ConvNext-B** | | | **384** | | **Imagenet1k** | **74.77** | **93.272** | **1.01** | **87M** |
| 2 | DeiT-B 384 | | | 384 | | Imagenet1k | 68.33 | 89.578 | 1.359 | 86M |
| 3 | DeiT-B Distilled 384 | | | 384 | | Imagenet1k | 68.15 | 88.99 | 1.422 | 86M |

Table 2. (a) Model comparision on CUB dataset. ConvNext outperformes the DeiT-base 384 model with +1.55%. DeiT Distilled (during pretraining) model performes lower as compared to ConvNext and DeiT.(b) Experimental results on CUB+DOG dataset. Here knowledge distillation during finetuning is performed as well using DeiT Distilled as student and ConvNext as teacher model. (c) For the FOOD dataset, the ConvNext shows highest results by achieving more than 6% span between DeiT-B 384 and DeiT-B distilled 384 models.

edge distillation results on top of raw Deit-B distilled model (without any CUB+DOG dataset) are not so high. In this setting, Deit-B distilled using convnext model reaches up to 88.53% and with ResNet BiT, it reaches to top1 accuracy of 87.49% respectively. It is worth to mentioned, hard knowledge distillation strategy gave better results than soft distillation. Lastly, for observing the effectiveness of knowledge distillation, we also evaluated results of stand along Deit-B models without using any knowledge distillation and it provides up to 82.5% top1 accuracy.

## 4. Discussions

**CUB and FoodX datasets:** From the results, ConvNext model provided the best accuracy on CUB and FoodX dataset, as compared to the transformer based DeiT model. The CNNs has by default a number of inherited inductive biases, which makes it to have an edge over transformers. Also, we tried training on scratch using CUB but the performance was very low (as shown in Table 2 a), this emphasizes that small datasets cannot generalize the large-scale models when trained from scratch. For FoodX dataset, we observed better performance from ConvNext model, but all in all, the overall results from both datasets are not so promising ($<$75%). We believe that, due to very large dataset size, these models should be trained for longer schedule with adaptive learning rates scheduler in order to further improve their performance.

**CUB+DOG Dataset:** Using CUB+DOG dataset, we tried to improve the baseline DeiT distilled model using the hard distillation strategy, where the DeiT model was finetuned not only based on the actual ground truth labels, but also on the predictions of the teacher model. For the case of using ConvNext as teacher model, the baseline model improved its performance by about 1.5%. This model also surpassed the performance of the teacher model itself, which also validates the official experiments performed on DeiT models. This improvement tends to propose that, using the double strategy and taking help from both CNNs and ViTs can provide results better than their individual performances. The final DeiT model not only achieves the best performance via learning from CNN, but also it enjoys other capabilities for being a transformer based model which makes it a more robust model, as proposed in [12].

## 5. Conclusion

In this work, we introduced knowledge distillation strategy in the FGIC problem. We validated the idea, that even for such difficult datasets, training vision transformer models, via supervision from CNNs helps in improving the overall performance of the vision transformer model, particularly DeiT in our study. The results also emphasize on the fact that CNNs, if properly designed and trained can still be considered the best models in image classification tasks and are on par with transformer based vision models.

## References


[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[4] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *European conference on computer vision*, pages 316–332. Springer, 2020. 1

[5] Sultan Daud Khan and Habib Ullah. A survey of advances in vision-based vehicle re-identification. *Computer Vision and Image Understanding*, 182:50–63, 2019. 1

[6] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. Fine-grained person re-identification. *International journal of computer vision*, 128(6):1654–1672, 2020. 1

[7] Jianlong Fu, Yue Wu, Tao Mei, Jinqiao Wang, Hanqing Lu, and Yong Rui. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *Proceedings of the IEEE international conference on computer vision*, pages 1985–1993, 2015. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[11] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *arXiv preprint arXiv:2012.12556*, 2020. 1

[12] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 3

[13] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 1

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 2

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 1

[16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[17] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011. 1

[18] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 1

[19] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1

[20] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020. 2