Optimizing ConvNeXt for TinyImageNet Classification using Self-Attention and Aggressive Regularization

1ST NADHIF ADITA FERNANDA
Department of Computer Science
Universitas Brawijaya
Malang, Indonesia
nadhifadita@student.ub.ac.id

2ND BISMA ADIKA KUSUMA PUTRA
Department of Computer Science
Universitas Brawijaya
Malang, Indonesia
bisma1002@student.ub.ac.id

3RD FARHAN VIER SYARIF HILMI
Department of Computer Science
Universitas Brawijaya
Malang, Indonesia
farhanvier@student.ub.ac.id

4TH KHAIRUMAM FIKRI
Department of Computer Science
Universitas Brawijaya
Malang, Indonesia
khairumamfikri@student.ub.ac.id

5TH REYNO BENEDICT
Department of Computer Science
Universitas Brawijaya
Malang, Indonesia
benedictreyno@student.ub.ac.id

ABSTRACT

Convolutional Neural Networks (CNNs) have long served as a fundamental backbone for image classification due to their strong inductive biases and effective hierarchical feature extraction. Recently, ConvNeXt has emerged as a modernized convolutional architecture that incorporates design principles inspired by Vision Transformers while retaining pure convolutional operations. Although ConvNeXt achieves strong performance on large-scale benchmarks, its application to small-scale datasets remains challenging due to overfitting and limited generalization. This study investigates the adaptation of a ConvNeXt-Tiny model for image classification on the TinyImageNet dataset. A baseline transfer learning configuration using a frozen pretrained backbone is evaluated alongside an improved training strategy that integrates full fine-tuning, aggressive regularization, RandAugment-based data augmentation, and a lightweight self-attention mechanism. Experimental results demonstrate that the improved configuration increases validation accuracy from 82.61% to 86.30% and reduces the generalization gap observed in the baseline model. These results indicate that combining attention mechanisms with aggressive regularization is an effective approach for enhancing the generalization performance of ConvNeXt on small-scale image classification tasks.

Index Terms—CNN, ConvNeXt, image classification, fine-tuning

I. INTRODUCTION

Designing models for computer vision tasks has attracted significant interest as Convolutional Neural Networks (CNNs) have become the state-of-the-art approach in visual recognition [1]. CNNs are specifically designed to extract hierarchical features from images of varying sizes through convolution and downsampling operations, enabling models to classify images in a manner that closely resembles human visual perception. The adoption of CNN-based models plays a crucial role in various smart-X technologies, such as smart cities, which heavily rely on image recognition systems for traffic management, public safety, and surveillance applications.

CNNs were originally proposed as a more automatic pattern recognition model, aiming to reduce dependence on hand-crafted features and heuristics by enabling end-to-end feature learning directly from raw image data. This paradigm shift significantly improved robustness and adaptability across diverse visual tasks, establishing CNNs as the dominant backbone for image classification, object detection, and semantic segmentation over the past decade.

However, as a computer vision tasks continue to grow in complexity and scale, limitations of tradition CNN architectures have become more apparent. In recent years, Vision Transformer-based models have demonstrated strong performance by leveraging self-attention mechanisms and large-scale pretraining [2], leading to a growing perception that convolution-based models may be insufficient for modern vision challenges. This shift has motivated renewed interest in revisiting and modernizing CNN architectures to bridge the performance gap between convolutional and transformer-based approaches.

ConvNeXt was introduced as a response to this challenge, aiming to re-examine classical CNN designs by incorporating architectural and training principles inspired by Vision Transformers, while retaining pure convolutional operations [3]. Through modifications such as large kernel convolutions, inverted bottleneck structures, Layer Normalization, and modern optimization strategies, ConvNeXt demonstrates that convolution-based models can achieve competitive performance compared to transformer-based architectures on large-scale benchmarks.

Despite these advancements, deploying ConvNeXt on

smaller-scale datasets presents additional challenges. High-capacity models are prone to overfitting when trained or fine-tuned on limited data, resulting in strong training performance but reduced generalization to unseen samples. This issue is particularly relevant for datasets such as TinyImageNet, where the number of training samples per class is significantly smaller than that of ImageNet-1K.

Therefore, this study focuses on optimizing ConvNeXt for image classification on the TinyImageNet dataset by addressing overfitting and improving generalization. We explore the effectiveness of fine-tuning strategies combined with aggressive regularization techniques and lightweight attention mechanisms. By comparing a baseline transfer learning setup with an improved training configuration, this work aims to evaluate how ConvNeXt can be adapted to small-scale datasets while maintaining strong classification performance.

II. Related Work

A. CNN & ConvNeXt

Convolutional Neural Networks (CNNs) have long been a central paradigm in the field of computer vision, owing to their inherent ability to effectively model spatial hierarchies in visual data [4]. By leveraging convolutional operations, CNNs embed strong inductive biases such as spatial locality and translation invariance, which make them particularly well-suited for tasks including image classification, object detection, and semantic segmentation [5]. Early and influential architectures such as AlexNet, VGG, and ResNet played a pivotal role in advancing deep learning for visual understanding [6], [7], [1]. These models employed stacked convolutional layers combined with nonlinear activation functions and pooling mechanisms to progressively extract high-level features from raw pixel inputs. Moreover, architectural innovations such as residual connections in ResNet demonstrated that significantly deeper networks could be trained effectively, leading to substantial improvements in accuracy when appropriate optimization strategies and regularization techniques were applied [8].

More recently, ConvNeXt has emerged as an effort to re-examine and modernize conventional CNN architectures by incorporating design principles that were popularized by Vision Transformer (ViT) models. Rather than abandoning convolution entirely, ConvNeXt revisits classical CNN components and refines them through systematic architectural updates [9]. Key modifications include the adoption of larger convolutional kernel sizes to enhance the receptive field, the use of inverted bottleneck structures to improve computational efficiency, the replacement of batch normalization with layer normalization, and a simplification of network stage designs. Through these carefully designed changes, ConvNeXt is able to close the performance gap between convolution-based and transformer-based models. As a result, it achieves competitive or superior accuracy on benchmark vision tasks while preserving the computational efficiency, scalability, and strong inductive biases that characterize convolutional neural networks [9], [10].

B. Transfer Learning

Transfer learning has become an essential and widely adopted strategy in deep learning research, especially in scenarios where the availability of labeled training data is limited or costly to obtain [11]. The core idea of transfer learning lies in initializing a model using parameters that have been pre-trained on large-scale benchmark datasets, such as ImageNet, which contain millions of labeled images across diverse categories [12]. By doing so, the model is able to reuse low-level and mid-level visual features, such as edges, textures, and object parts that have already been learned during the pre-training phase. This reuse of learned representations not only accelerates the training process by enabling faster convergence but also enhances the model's ability to generalize to unseen data [13]. Consequently, transfer learning has become a standard practice in CNN-based architectures and has consistently demonstrated significant performance improvements across a wide range of downstream computer vision tasks [11], [13].

Within the context of more recent and modern architectures, such as ConvNeXt, transfer learning continues to play a crucial role in achieving strong performance. Pre-trained ConvNeXt models provide rich and expressive feature representations that can be effectively adapted to domain-specific applications through fine-tuning [9], [10]. Various fine-tuning strategies can be employed depending on factors such as dataset size, task complexity, and similarity to the pre-training domain. These strategies include freezing early layers to preserve general-purpose features while updating only the higher-level layers, as well as fully fine-tuning the entire network to maximize task-specific adaptation [13], [14]. The flexibility of these approaches allows practitioners to balance computational cost and performance, making transfer learning with ConvNeXt a practical and powerful solution for a wide range of real-world applications.

C. Attention

Attention mechanisms have been widely introduced in deep neural network architectures as a means to enhance feature representation by enabling models to selectively focus on the most informative components of the input data. Instead of treating all features equally, attention allows the network to assign different levels of importance to specific features based on their relevance to the task at hand [14]. Within CNN-based architectures, attention mechanisms are commonly implemented in the form of channel attention, spatial attention, or hybrid approaches that combine both types. Channel attention focuses on identifying the most discriminative feature maps, while spatial attention emphasizes important regions within the spatial dimensions of the feature maps. By selectively enhancing relevant features and suppressing less informative ones, these mechanisms contribute to more

effective and robust feature extraction. This information defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)v \quad (1)$$

where Q is query, K is key, V is value, and d_k is the dimension of key vectors.

Recent research has demonstrated that the integration of attention modules into convolutional neural networks leads to notable performance improvements across a variety of computer vision tasks. One of the key advantages of attention-based enhancements is their ability to capture global contextual information and model long-range dependencies that are not easily handled by standard convolution operations with limited receptive fields [14]. This capability is particularly important for tasks that require fine-grained recognition, precise localization, or complex spatial reasoning, where subtle differences in visual patterns must be distinguished [12]. As a result, attention mechanisms have become an increasingly popular component in modern CNN designs, complementing traditional convolutional operations and further improving the expressive capacity of deep learning models.

D. Data Augmentation

Data augmentation is a widely adopted technique in deep learning, particularly in computer vision, to enhance model robustness and mitigate the risk of overfitting. By artificially increasing the diversity of training samples through various transformations, data augmentation enables models to learn more invariant and generalizable feature representations [13]. Conventional augmentation techniques, such as random cropping, horizontal and vertical flipping, rotation, scaling, and color jittering, have been extensively studied and shown to effectively improve generalization performance in image-based recognition tasks. These transformations expose the model to variations that commonly occur in real-world data, thereby reducing sensitivity to minor changes in input appearance.

Beyond traditional approaches, more advanced data augmentation methods have been proposed to further optimize the training process. One notable example is RandAugment, which simplifies and automates the augmentation strategy by randomly selecting a fixed number of augmentation operations with predefined magnitudes during training. Unlike earlier automated augmentation methods that rely on computationally expensive search procedures or extensive manual hyperparameter tuning, RandAugment provides a more efficient and practical alternative. Empirical studies have demonstrated that RandAugment consistently leads to performance improvements across a wide range of neural network architectures. These improvements are particularly pronounced when training deep models on relatively small or limited datasets, where effective regularization and data diversity play a critical role in achieving strong generalization.

III. Methodology

A. Dataset

This study utilizes Tiny-ImageNet, a standard benchmark dataset commonly used for evaluating image classification models under constrained computational settings [13]. Tiny-ImageNet is an official subset of the ImageNet dataset and retains its semantic diversity while significantly reducing dataset scale.

The dataset contains 200 distinct object classes, with each class comprising 500 training images, 50 validation images, and 50 test images, resulting in a total of 100,000 images. All images are provided at a fixed spatial resolution of 64 × 64 pixels and are evenly distributed across classes, ensuring balanced class representation [13].

Tiny-ImageNet was selected for several reasons. First, its moderate dataset size enables efficient training and experimentation without the extensive computational requirements of ImageNet-1K. Second, as a derivative of ImageNet, it remains highly compatible with models pretrained on ImageNet, including ConvNeXt, which is employed in this study. Third, the images can be resized to 224 × 224 pixels, allowing seamless integration with standard training pipelines and pretrained architectures.

Due to these characteristics, Tiny-ImageNet serves as an effective benchmark for assessing model performance, architectural modifications, and training strategies in multi-class image classification tasks, particularly in the context of transfer learning and fine-tuning.

TABLE I

Summary of The Tiny-ImageNet Dataset

Attribute	Description
Number of classes	200
Training images	100,000 (500 per class)
Validation images	10,000 (50 per class)
Test images	10,000 (50 per class)
Image resolution	64 × 64 pixels
Task type	Multi-class image classification

Figure 1 presents representative samples from the Tiny-ImageNet dataset, illustrating the diversity and complexity of images across different object categories. Although Tiny-ImageNet is a reduced subset of the original ImageNet dataset, it preserves substantial visual variability in terms of object appearance, background clutter, illumination, and viewpoint.

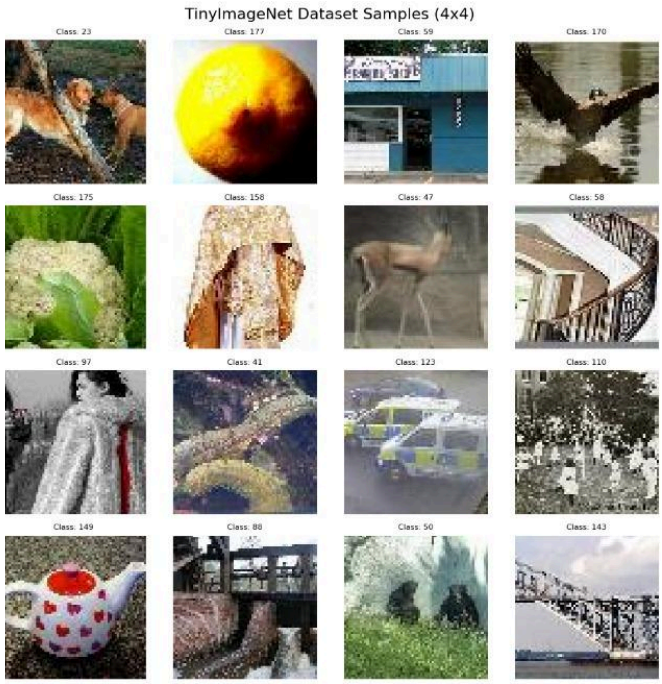


Fig. 1. Sample images from the Tiny-ImageNet dataset.

As shown in the figure, objects often occupy a limited portion of the image due to the low spatial resolution of 64 × 64 pixels. This characteristic makes discriminative feature extraction more challenging, as fine-grained details are easily lost during downsampling. Additionally, several classes exhibit high visual similarity, increasing the difficulty of distinguishing between semantically related categories.

The presence of complex backgrounds and varying object scales further emphasizes the need for models that can effectively capture both local texture information and global contextual relationships. These characteristics make Tiny-ImageNet a suitable benchmark for evaluating the generalization capability and robustness of convolutional neural networks, particularly when adapted from large-scale pretraining such as ImageNet-1K.

B. Proposed Improvement

This study proposes an enhanced ConvNeXt architecture by integrating a lightweight self-attention mechanism at the final stage of the network, as illustrated in Fig. 2. While the original ConvNeXt architecture relies exclusively on convolutional operations for feature extraction, the proposed modification introduces explicit global context modeling to complement the strong local spatial representations learned by convolutional layers. ConvNeXt has demonstrated that modernized convolutional networks can achieve competitive performance with Transformer-based models when trained using appropriate architectural designs and optimization strategies [3].

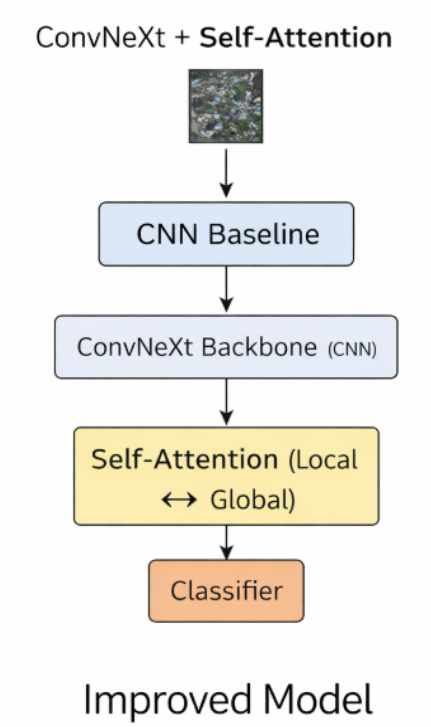


Fig. 2. Architecture of the proposed ConvNeXt model with a self-attention module at the final stage.

As shown in Fig. 2, the input image is first processed through a standard convolutional pipeline, where hierarchical features are extracted using a ConvNeXt-Tiny backbone. The backbone operates as the primary feature extractor and preserves the convolutional inductive bias that is effective for capturing local patterns such as edges, textures, and object parts. The output feature map from the final ConvNeXt stage is then forwarded to a self-attention module for further contextual refinement.

In the proposed design, the final-stage feature map is reshaped into a sequence of spatial tokens, where each token represents a distinct spatial location enriched with high-level semantic information. A multi-head self-attention mechanism is applied to this token sequence, in which each token simultaneously serves as the query, key, and value. Through this formulation, each spatial location is able to attend to all other locations in the feature map, enabling the model to capture long-range dependencies and global spatial relationships that are difficult to model using convolutional operations alone due to their inherently local receptive fields [14].

The self-attention module is incorporated without altering the internal structure of the ConvNeXt backbone, ensuring that the original convolutional architecture remains intact. This design choice preserves computational efficiency and architectural simplicity while enhancing the model's ability to integrate global contextual information with local feature representations. Finally, the attention-enhanced features are aggregated and passed to a classification head to produce the final class prediction. Overall, this improvement aims to strengthen feature expressiveness and generalization

performance, particularly in small-scale image classification settings such as Tiny-ImageNet.

C. Training Strategy

In this study, two fine-tuning strategies are investigated to evaluate the effectiveness of ConvNeXt-Tiny on the Tiny-ImageNet dataset. The first strategy serves as a baseline, in which a ConvNeXt-Tiny model pretrained on ImageNet-1K is adapted using a transfer learning approach. In this configuration, the backbone network is kept frozen, and only the final classification head is replaced and trained to accommodate the 200 target classes. This setting evaluates the discriminative capability of pretrained convolutional features under a feature extraction regime.

The second strategy, referred to as the improved configuration, applies full fine-tuning by unfreezing all layers of the ConvNeXt-Tiny model. This allows the network to fully adapt its representations to the Tiny-ImageNet domain. To improve computational efficiency during training, automatic mixed precision (AMP) is employed. Full fine-tuning enables more effective task-specific feature learning by leveraging the full capacity of the model, particularly when combined with the proposed architectural modifications and regularization techniques.

For both baseline and improved strategies, an early stopping mechanism is utilized to select the optimal model checkpoint based on the highest validation accuracy. This approach prevents unnecessary computation and mitigates overfitting by terminating training once performance on the validation set no longer improves.

IV. Experiment

A. Baseline Performance

The baseline experiment, utilizing the backbone freezing strategy, achieved a validation accuracy of 82.61% with a validation loss of 0.8057. The training process was governed by an Early Stopping mechanism (patience=5) to prevent redundant computation. As shown in Fig. 3, the training accuracy continued to improve significantly, reaching approximately 91%, while the validation accuracy plateaued around 82%. This resulted in a generalization gap of ~9%, indicating that the model began to overfit the training data despite the frozen backbone. Furthermore, the loss curve in Fig. 1 demonstrates a clear divergence starting from Epoch 2, where validation loss increased while training loss decreased. Consequently, training was automatically halted at Epoch 6.

To qualitatively assess the model's limitations, we visualized random predictions in Fig. 4. We define the word “Sample” as randomly sampling 8 correct predictions and 8 wrong predictions. The model demonstrates robust performance on objects with distinct shapes and clear backgrounds. However, misclassifications tend to occur in images with complex backgrounds or visual ambiguity (e.g., Class 145 and Class 87), suggesting that the frozen feature extractor struggles to separate foreground objects from background noise in challenging samples. To address the smaller scope, we would also like to see the first 16 classes (0-15) shown in Fig. 5.

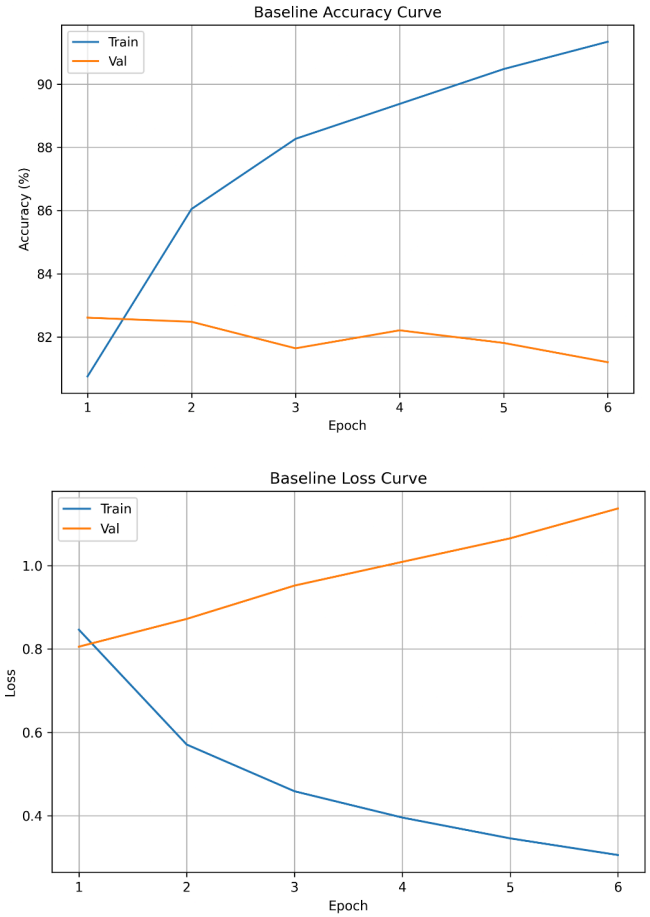


Fig .3. Plots of training-validation accuracy and loss of baseline

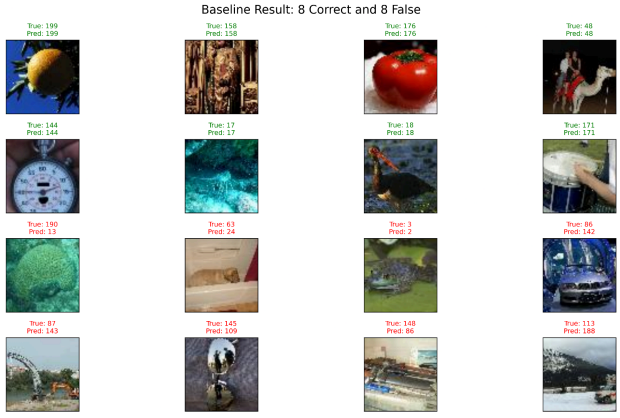


Fig. 4. Baseline Sample Prediction Result

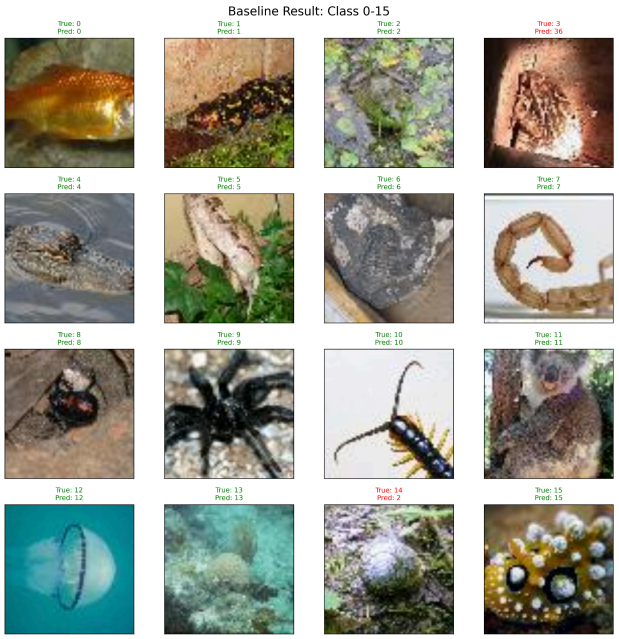
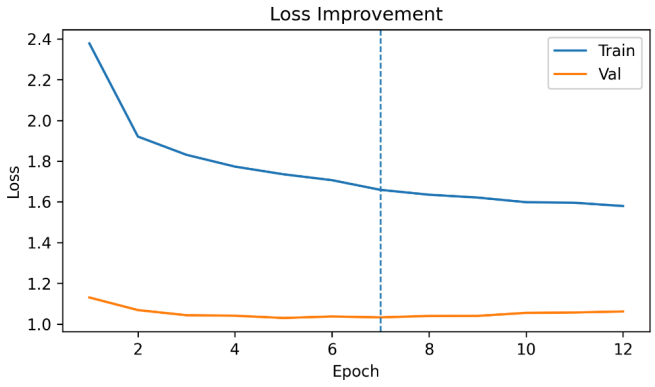


Fig. 5. Baseline Prediction Result Class 0-15

B. Improving the Model

In the second phase of experiments, we trained the Improved ConvNeXt model incorporating the custom Attention module, RandAugment, and aggressive regularization. The Improved ConvNeXt model achieved a peak Validation Accuracy of 86.30% at Epoch 7, surpassing the baseline by approximately 4%. As visualized in **Fig. 6**, the training dynamics shifted drastically compared to the baseline.



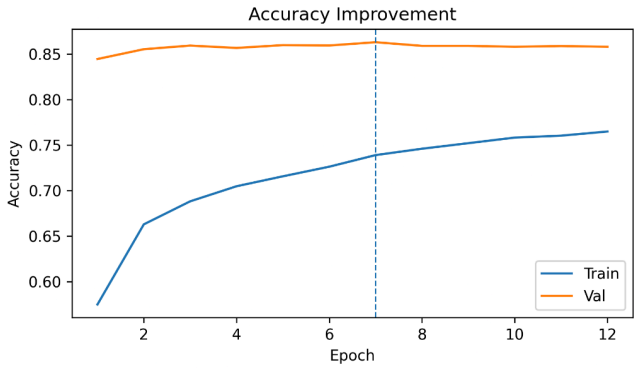


Fig. 6. Plots of training-validation accuracy and loss of improvement

A notable phenomenon observed in Fig. 4 is that the Validation Accuracy consistently exceeds the Training Accuracy. This behavior is a direct consequence of the aggressive regularization pipeline we implemented:

1. RandAugment and Dropout (p=0.2) introduce significant noise and difficulty during the training phase, suppressing the training accuracy to approximately 74%.
2. High Weight Decay (5×10^{-4}) on the Attention mechanism successfully restricts model complexity.

Conversely, during validation, the model is evaluated on clean, unaugmented images with full network capacity (Dropout disabled). The stability of the loss curve (right) and the early plateau indicate that the model has learned generalized features rather than memorizing specific training samples. The Early Stopping mechanism was triggered at Epoch 12, identifying Epoch 7 as the optimal checkpoint with the best trade-off between bias and variance. Prediction through classes could be seen in Fig. 7 and Fig. 8.

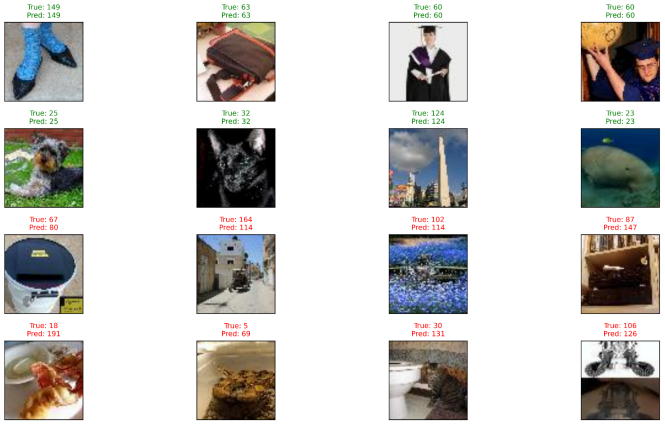


Fig. 7. Improved Sample Prediction Result

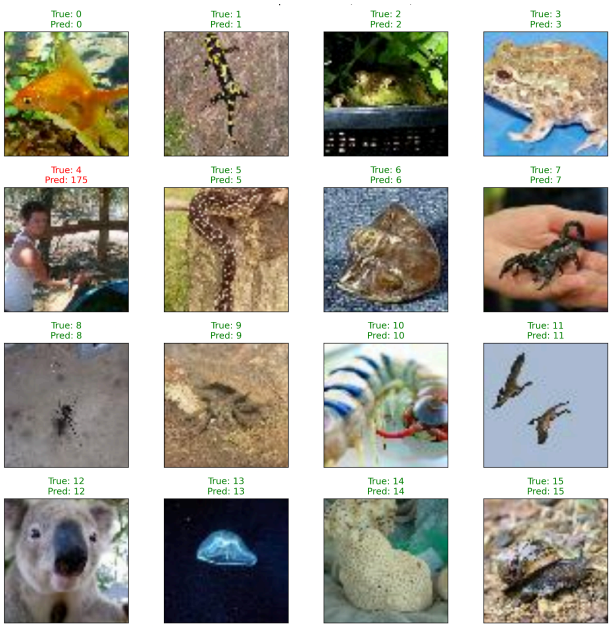


Fig. 8. Improved Prediction Result Class 0-15

V. Discussion

The experimental results demonstrate that the proposed improvements effectively enhance the performance of ConvNeXt on the Tiny-ImageNet dataset. In the baseline configuration, where the pretrained ConvNeXt-Tiny backbone is frozen, the model exhibits a notable generalization gap between training and validation accuracy. This behavior indicates that relying solely on pretrained convolutional features is insufficient for capturing the complexity of Tiny-ImageNet, particularly given its limited resolution and high inter-class visual similarity.

By contrast, the improved configuration achieves a substantial performance gain by combining full fine-tuning, aggressive regularization, and the integration of a lightweight self-attention module. The introduction of self-attention at the final stage enables the model to capture global contextual relationships across spatial locations, complementing the local feature representations learned by convolutional layers. This enhancement is particularly beneficial for images containing cluttered backgrounds or objects occupying small regions of the image, which are common characteristics of Tiny-ImageNet.

Furthermore, the observed training dynamics provide insight into the effectiveness of the proposed regularization strategy. The lower training accuracy compared to validation accuracy in the improved model can be attributed to the use of RandAugment, dropout, and high weight decay, which collectively introduce significant noise during training. This behavior suggests that the model is discouraged from memorizing training samples and is instead guided toward learning more robust and generalizable feature representations.

Overall, the results indicate that augmenting ConvNeXt with self-attention and aggressive regularization is an effective approach for mitigating overfitting and improving generalization on small-scale image classification datasets. These findings highlight the importance of combining architectural enhancements with appropriate training strategies when adapting high-capacity convolutional models to data-limited settings.

VI. Conclusion

In conclusion, this study demonstrates the effectiveness of the improved ConvNeXt-Tiny architecture for low-resolution image classification. By integrating a Self-Attention mechanism and a Differential Learning Rate strategy, we achieved a peak validation accuracy of 86.30%, surpassing the freezing-based baseline (82.61%). Qualitative analysis in Fig. 8 further corroborates this performance, showing minimal misclassification among random samples. Most importantly, the implementation of Self Attention and RandAugment proved critical in solving the overfitting problem observed in the baseline, effectively bridging the generalization gap.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90

[2] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learning Representations (ICLR), 2021, doi: 10.48550/arXiv.2010.11929.

[3] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986, 2022.

[4] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2012, doi: 10.1145/3065386.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learning Representations (ICLR), 2015, doi: 10.48550/arXiv.1409.1556.

[7] S. Woo et al., “ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16133–16142, doi: 10.1109/CVPR52729.2023.01549.

[8] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2009, doi: 10.1109/TKDE.2009.191.

[9] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2014, doi: 10.48550/arXiv.1411.1792.

[10] H. Shin et al., “Deep convolutional neural networks for computer-aided detection,” IEEE Trans. Medical Imaging, vol. 35, no. 5, pp. 1285–1298, 2016, doi: 10.1109/TMI.2016.2528162.

[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Computer Vision (ECCV), 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[12] A. Mumuni and F. Mumuni, “Data augmentation: A comprehensive survey of modern approaches,” Array, vol. 16, p. 100258, 2022, doi: 10.1016/j.array.2022.100258.

[13] Y. Le and X. Yang, “Tiny ImageNet Visual Recognition Challenge,” Stanford University CS231N Course Project, 2015. [Online].

Available: zh-plus/tiny-imagenet · Datasets at Hugging Face
[14] A. Vaswani et al., “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, doi: 10.48550/arXiv.1706.03762.