Research Report: Enhancing Classification on Noisy CIFAR-100 Dataset

Author: Anfimov Vladimir

---

1. Introduction

The CIFAR-100 dataset, a benchmark for image classification, introduces challenges with label noise. Tackling this problem requires robust model architectures and training techniques. This report documents the use of Vision Transformer (ViT) models, comparing a baseline and an improved approach to manage noisy labels effectively.

1.1 Network Architecture: ViT_Small_Patch16_224 The Vision Transformer (ViT) `vit_small_patch16_224` processes images by dividing them into non-overlapping patches of size 16×16 pixels, which are flattened and embedded into a sequence of tokens. These tokens are then fed into a transformer encoder consisting of:

- **Multi-Head Self-Attention:** Captures relationships between image patches by attending to different regions of the input.
- **Feedforward Layers:** Fully connected layers that process token embeddings.
- **Positional Encodings:** Adds spatial information to tokens, enabling the model to understand positional context.

The final output is passed through a classification head. This classification head includes:

- **Classification Token:** A learnable token prepended to the patch sequence, aggregating global information during the self-attention process.
- **Fully Connected Layer:** A single linear layer maps the classification token to the desired number of output classes, 100 in the case of the CIFAR-100 dataset.

This structure ensures that the ViT effectively captures both local and global features of the image, making it robust for classification tasks, even in noisy label environments.

2. Baseline Implementation

2.1 Dataset and Common Techniques: The dataset comprises CIFAR-100 images with noisy labels. Both baseline and improved models used the following methods:

- **Data Preprocessing:** Images resized to 224×224 pixels and normalized using CIFAR-100 mean and standard deviation.
- **Optimization Strategy:** Adam optimizer with a learning rate of 1e-4 and CosineAnnealingLR scheduler for decay.
- **Batch Size and Augmentation:** Batch size of 64 with standard resizing.

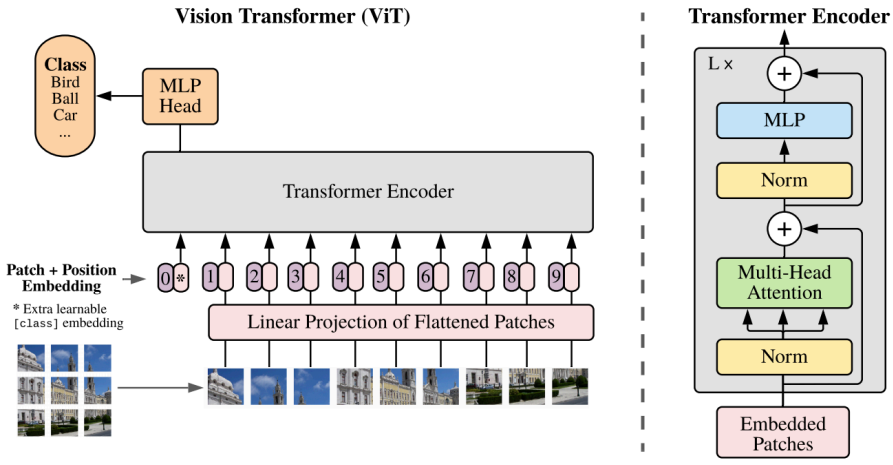


Figure 1: ViT Model Architecture

- **Early Stopping:** Training stopped after 3 epochs without improvement.
- **Hardware Optimization:** Mixed precision training with GradScaler for memory and speed efficiency.

2.2 Baseline Model: The baseline model employed the `vit_small_patch16_224` architecture from TIMM, pre-trained on ImageNet. Excluding advanced augmentations like MixUp and CutMix, the model was trained for 5 epochs, achieving a validation accuracy of 73.66%. Training stalled for 3 iterations without progress, triggering early stopping as part of the optimization strategy.

---

3. Enhancements Over Baseline

3.1 Advanced Augmentation Techniques:

- **RandAugment:** Applied randomized augmentations with 6 operations and a magnitude of 9 for improved data diversity.
- **MixUp:** Blended image-label pairs with an alpha value of 1.0, ensuring equal weighting between the two images and their labels to reduce overfitting and encourage generalization.
- **CutMix:** Replaced image patches and labels with an alpha value of 1.0, ensuring balanced patch sizes that enhance the model's ability to generalize across diverse data contexts.

3.2 Overall Impact: RandAugment, MixUp, and CutMix collectively enabled faster convergence and higher accuracy, significantly improving performance over the baseline.

4. Results

4.1 Performance Metrics: The improved Vision Transformer model achieved a validation accuracy of 78.42% after 7 epochs, compared to the baseline's 73.66% after 5 epochs. Both models adhered to the 20-epoch training limit.

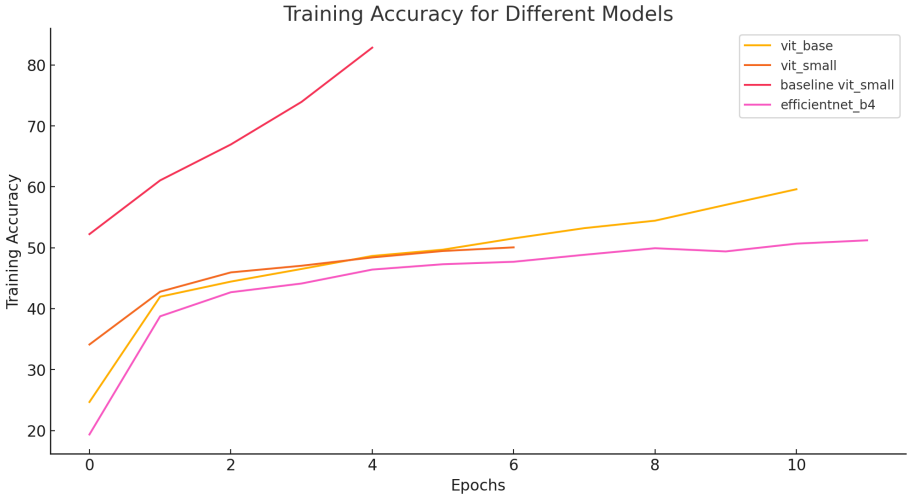
4.2 Contribution Analysis:

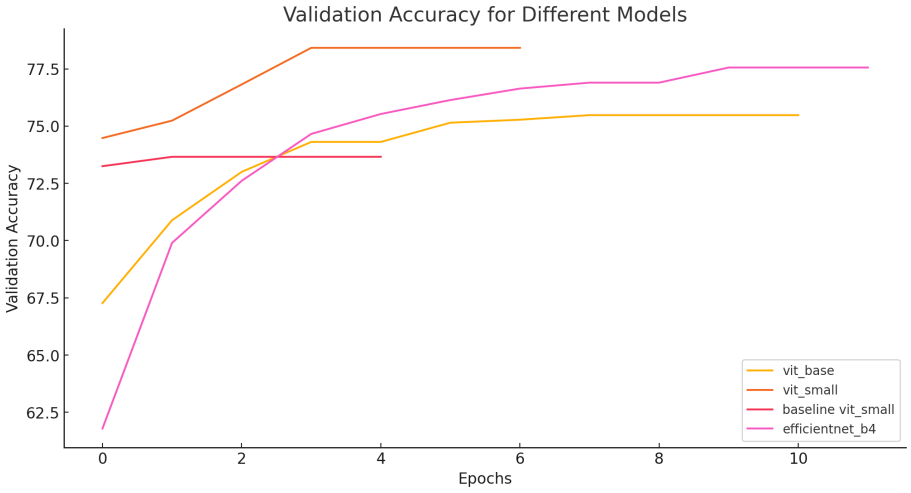
- **Augmentation Benefits:** Adding RandAugment (6 operations, magnitude 9), MixUp, and CutMix contributed to a 4.76% accuracy improvement.
- **Efficiency Gains:** Advanced techniques ensured faster convergence and robustness against label noise.

4.3 Other Experiments:

Additional experiments were conducted to explore the potential of other architectures:

- **ViT_Base_Patch16_224**: This model achieved a maximum validation accuracy of 75.4%, but training required 2 hours and 44 minutes, significantly increasing computational costs without accuracy gains.
- **EfficientNet_B4**: This model achieved a maximum validation accuracy of 77.56%, slightly lower than the improved ViT model, indicating that while EfficientNet offers competitive accuracy.





5. Conclusion

This study confirms that augmentations like RandAugment, MixUp, and CutMix, paired with the ViT architecture, excel in noisy label environments. Future work could explore semi-supervised approaches, noise-cleaning techniques, and scaling to larger datasets.