Hyperparameter Tuning

Multiple hyperparameters were tuned to study their effect on model performance. The learning rate, weight decay, and number of epochs were varied while keeping other parameters fixed.

Regularization and Preprocessing Techniques

To reduce overfitting and improve generalization, multiple regularization and preprocessing techniques were employed.

Data augmentation techniques such as random horizontal flipping and random cropping were applied to the training images. Since CIFAR-10 images are originally of size 32×32, they were resized to 224×224 to match the input requirements of the Vision Transformer.

Weight decay (L2 regularization) was used through the AdamW optimizer to penalize large weights. Dropout layers present inside the Vision Transformer architecture further helped prevent co-adaptation of neurons.

It was observed that models trained with data augmentation achieved higher test accuracy and reduced overfitting compared to models trained without augmentation.

Vision Transformer Architecture

The Vision Transformer (ViT) adapts the transformer architecture, originally proposed for natural language processing, to image classification tasks.

The input image is divided into fixed-size patches of 16×16 pixels. Each patch is flattened and projected into a lower-dimensional embedding space using a linear layer. Positional embeddings are added to preserve spatial information.

The embedded patches are passed through multiple Transformer encoder blocks. Each block consists of multi-head self-attention, a feed-forward neural network, layer normalization, and residual connections.

A special classification token ([CLS]) is appended to the sequence of patch embeddings. The output corresponding to this token is used for final image classification using a fully connected layer.

Implementation Details

```
model = timm.create_model(
    'vit_tiny_patch16_224',
    pretrained=True,
    num_classes=10
```

The Vision Transformer model was implemented using the timm library. Pretrained weights from ImageNet were used to improve convergence and performance on the CIFAR-10 dataset.

Results

The Vision Transformer achieved competitive performance on the CIFAR-10 dataset. Data augmentation and regularization techniques significantly improved generalization performance.

Conclusion

This experiment demonstrated the applicability of Vision Transformers to image classification tasks. While ViT models are computationally expensive, using smaller variants and data augmentation allows efficient training on limited datasets such as CIFAR-10.