Weather Image Recognition using Vision Transformers (ViTs)

Franklin Xavier Anthony Albin Anbaiya
antho264@umn.edu

1 Introduction

The automated categorization of weather conditions from images is crucial for various applications, such as environmental monitoring, agriculture, aviation, and intelligent transportation systems. Accurate weather predictions and recognitions directly affect decision-making strategies, from redirecting flights to avoid intense storms and modifying farming methods in expectation of droughts or rain, to providing safety warnings for autonomous vehicles in low visibility situations.

Traditional methods for weather detection typically depend on classic computer vision strategies or convolutional neural networks (CNNs). Although CNNs have yielded promising results in deriving hierarchical features from original images, they may find it challenging to entirely grasp global contexts and long-range dependencies essential for nuanced weather events. Differentiating between light haze and heavy fog, for instance, may require an understanding of the entire image's uniformity and texture gradients, rather than localized patterns alone.

Recent advancements in Vision Transformers (ViTs)(Dosovitskiy et al., 2021) offer a compelling option, as these models utilize self-attention directly on image patches, approaching images similarly to sequences of visual tokens. This method enables the modeling of local and global dependencies right from the beginning, making it particularly effective for weather classification tasks that rely on understanding wide spatial patterns, like cloud formations or the spread of fog. Moreover, ViT-Hybrid architectures that merge convolutional backbones and transformer layers can harness the advantages of local feature extraction alongside the global reasoning abilities of transformers. These hybrid models could be especially proficient at managing nuanced weather variations, effortlessly combining detailed textures obtained through CNNs with the comprehensive scene perception enabled by self-attention strategies.

In this project, we assess and analyze two advanced transformer models—one entirely reliant on ViTs and the other utilizing a hybrid approach—within a multi-class weather classification task. Through comparing these methods, we seek to establish if pure or hybrid transformer architectures more effectively satisfy the strict demands of weather classification, thereby guiding more precise and resilient solutions for various visually intricate, globally reliant recognition tasks.

2 Related Work

The classification of weather conditions using images has been explored extensively with varying methodologies ranging from traditional feature-based approaches to deep learning models. Traditional methods often rely on hand-crafted features combined with classifiers like Support Vector Machines (SVMs)(Ship et al., 2024).

On the other hand, deep learning approaches, particularly those leveraging Convolutional Neural Networks (CNNs), have significantly advanced the field. (Xiao et al., 2021) introduced a CNN-based approach tailored for weather classification. Their method achieved a normalized classification accuracy of 92%, a substantial improvement over traditional feature-engineering methods.

Recent advancements in transformer architectures, such as Vision Transformers (ViTs), have further propelled the field by enabling the modeling of global and local dependencies. Unlike CNNs, ViTs treat images as sequences of patches, allowing them to capture both coarse and fine-grained patterns essential for nuanced tasks like weather classification. Additionally, hybrid models combining convolutional backbones with transformer layers have shown promise in leveraging the strengths of both architectures.

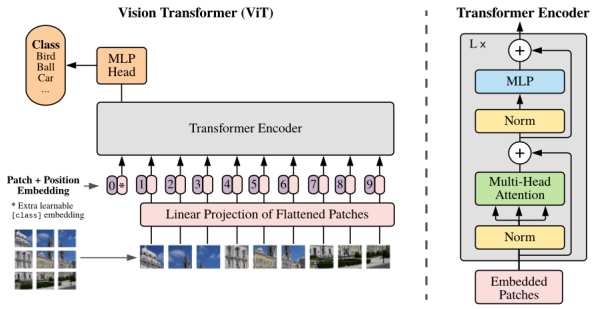

Figure 1: ViT Architecture

3 Architectural Details

3.1 ViT Base Model

The base ViT model is google/vit-base-patch16-224-in21k which is a Vision Transformer pre-trained on the ImageNet-21k (Ridnik et al., 2021) dataset. It divides the input image into patches of size 16×16. For a 224×224 resolution input, the image is split into $14 \times 14 = 196$ patches. Each patch is linearly projected into a 768-dimensional embedding, forming a sequence of embeddings to which standard Transformer encoders (consisting of multi-head self-attention and feed-forward layers) are applied.

The ViT architecture includes a learnable [CLS] token prepended to the patch embeddings, Layer normalization, multi-head self-attention, MLP blocks repeated several times (12 transformer encoder layers for the base model), and Positional embeddings to preserve spatial information.

After passing through the transformer layers, the [CLS] token's final representation is used for classification via a fully connected MLP head.

3.2 ViT-Hybrid Model

The ViT-Hybrid model is google/vit-hybrid-base-bit-384 which is a hybrid architecture that uses a ResNet (BiT variant) (Kolesnikov et al., 2020) backbone as a convolutional stem. The CNN stem extracts low-level features and outputs feature maps, which are then flattened into patches for the transformer layers. This approach can provide richer local representations compared to pure patch embeddings, potentially benefiting tasks requiring subtle discriminations.

Key aspects of this architecture includes a ResNet-based stem pre-trained on large datasets providing strong initial local feature extraction, A similar transformer encoder stack to the pure ViT model, applied after the CNN stem's feature maps are patchified and embedded, and the final classification head that operates on the [CLS] token as in the pure ViT.

This larger input resolution and CNN stem can help capture more fine-grained details before the transformer layers aggregate global context.

4 Approach

4.1 Dataset and Preprocessing

We use a publicly available weather image dataset from Kaggle jehanbhathena/weather-dataset, comprising 6,862 images across 11 classes: dew, fog/smog, frost, glaze, hail, lightning, rain, rainbow, rime,sandstorm, and snow. The dataset, derived from the Weather Phenomenon Database (WEAPD)(Xiao et al., 2021), is ideal for weather classification tasks due to its diversity and real-world complexity. We split the dataset into training, validation, and test subsets using stratified sampling and preprocess images to resolutions of 224×224 and 384×384 for compatibility with ViT and ViT-Hybrid models.

4.2 Data Augmentation and Transformations

We apply random resized crops, horizontal flips, and color jitter to the training set for both ViT and ViT-Hybrid models, though at their respective resolutions (224 and 384). Validation and test sets are only resized and normalized, ensuring a fair performance assessment.

4.3 Handling Class Imbalance

To address class imbalance, we employ WeightedRandomSampler to sample underrepresented classes more frequently. We also use mixup with $\alpha = 0.4$, blending pairs of images and labels to encourage the model to learn smoother decision boundaries and reduce overfitting.

4.4 Training and Optimization

Both the ViT and ViT-Hybrid models were fine-tuned using the AdamW optimizer with an initial learning rate of 2×10^{-5} and a batch size of 32. Training was conducted for 50 epochs to thoroughly evaluate the convergence behavior of both architectures.To stabilize training and prevent overfitting, a weight decay of 0.01 was applied to the optimizer.

Both models were trained on a single NVIDIA A100 GPU on Google Colab for around 5 hours.

Figure 2: Structure of the Dataset

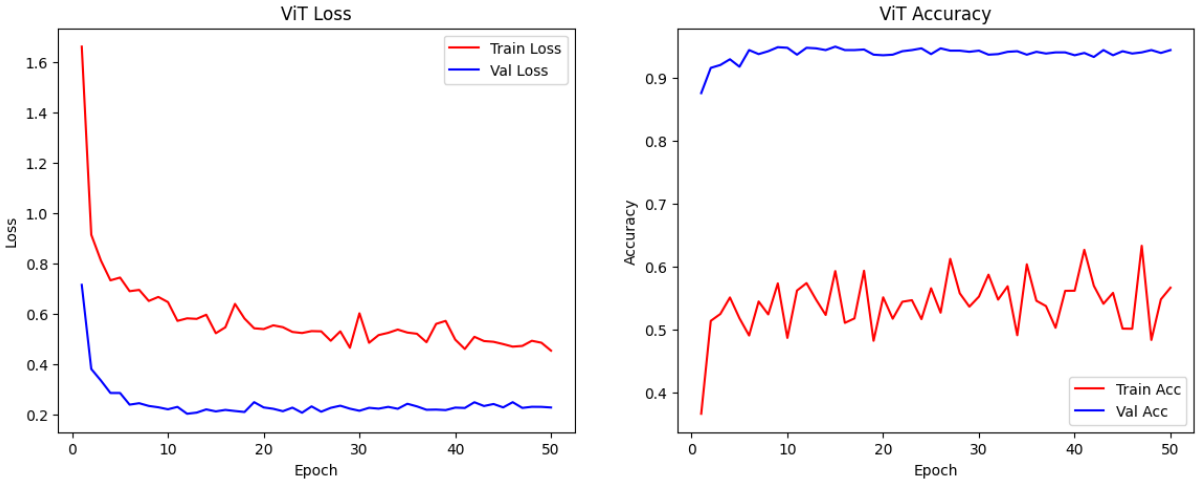

Figure 3: ViT Training and Validation Loss and Accuracy

4.5 Training and Validation Performance

The training process demonstrated distinct learning patterns for the two models:

- **ViT:** The pure transformer architecture exhibited steady improvements in training and validation accuracy, with minimal fluctuations. Over the 50 epochs, the model achieved a best validation accuracy of **94.90%**, indicating its capability to capture global dependencies effectively. The training loss steadily decreased, suggesting smooth convergence.
- **ViT-Hybrid:** The hybrid model achieved a best validation accuracy of **95.45%**, slightly outperforming the ViT model. However, the validation loss exhibited mild oscillations, particularly in later epochs, indicating potential sensitivity to the learning rate or data augmentations. This behavior suggests that while the hybrid model benefits from the convolutional backbone for local feature extraction, its performance might stabilize further with a learning rate scheduler or longer training.

The use of mixup augmentation and WeightedRandomSampler was particularly beneficial, as evidenced by the balanced class-wise performance across both models. These techniques ensured that even minority classes, such as *frost* and *glaze*, were represented adequately during training, preventing overfitting to dominant classes.

4.6 Evaluation Metrics

The evaluation metrics for both models included:

- **Top-1 Accuracy:** Measures the percentage of correctly classified images in the test set.
- **Macro-F1 Score:** Provides a balanced metric by equally weighting all classes, ensuring fair

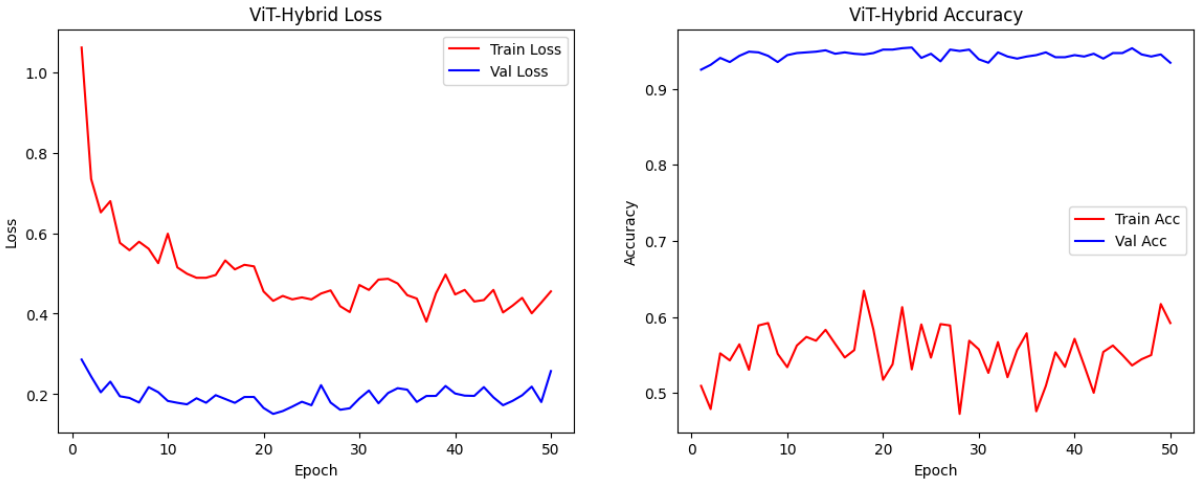

Figure 4: ViT Hybrid Training and Validation Loss and Accuracy

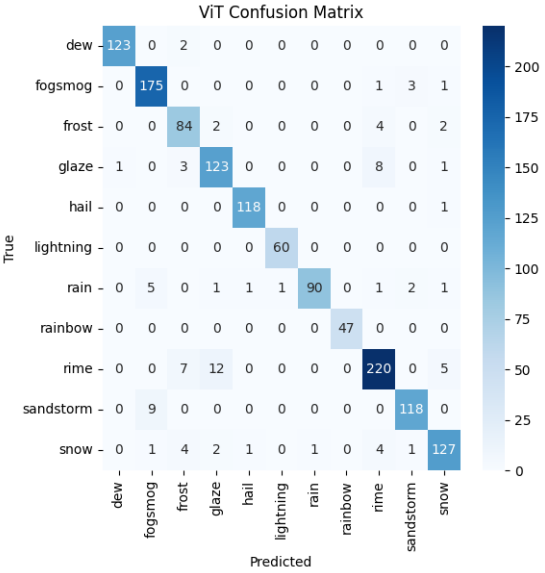

Figure 5: ViT Confusion Matrix

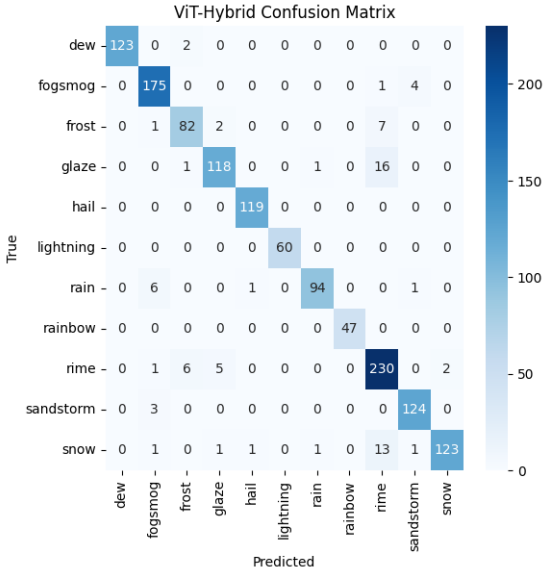

Figure 6: ViT Hybrid Confusion Matrix

performance evaluation across minority and majority classes.

- **Micro-F1 Score:** Averages performance across all samples, favoring the majority class.
- **Confusion Matrix and Per-Class Accuracy:** Offers insights into class-wise performance and identifies specific categories where models might struggle.

4.7 Test Set Results

Table 1 summarizes the test set performance for both models. Notably, the ViT model achieved a test accuracy of **93.59%**, while the ViT-Hybrid model slightly outperformed with **94.32%** accuracy. Both models demonstrated strong generalization capabilities, with high Macro-F1 and Micro-F1 scores exceeding 0.94. These results highlight the effectiveness of Vision Transformers for weather image classification.

Model	Test Acc	Macro-F1	Micro-F1
ViT	0.9359	0.9432	0.9359
ViT-Hybrid	0.9432	0.9522	0.9432

Table 1: Test set performance metrics for both models.

4.8 Per-Class Performance

Figure 7 and 8 illustrates the per-class accuracy for both models. The analysis reveals:

- Both models excelled in categories with distinct visual features, such as *rainbow*, *dew*,

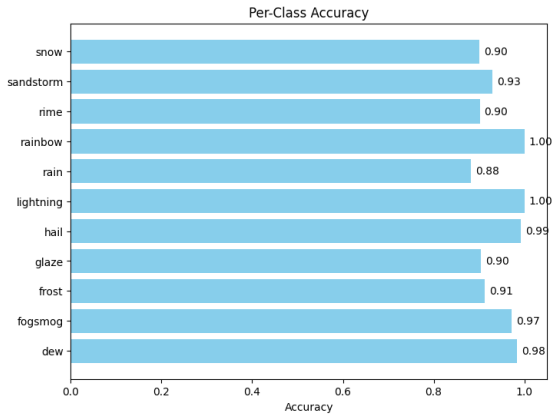

Figure 7: ViT Per Class Accuracy

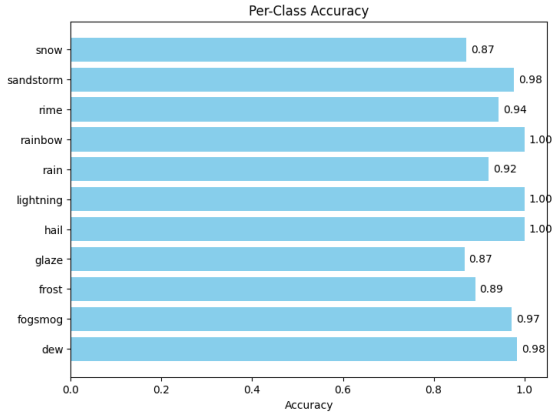

Figure 8: ViT Hybrid Per Class Accuracy

and *lightning*, achieving near-perfect accuracy.

- More nuanced categories, such as *frost*, *glaze*, and *rime*, posed greater challenges due to subtle visual differences. The ViT-Hybrid model outperformed the pure ViT model for *rime*, indicating the hybrid architecture's advantage in capturing local textures.

- For *glaze*, the ViT model achieved higher recall, suggesting its stronger ability to identify such conditions consistently.

5 Conclusion

This study demonstrates the exceptional capabilities of Vision Transformers (ViT) and hybrid architectures (ViT-Hybrid) in addressing the challenges of weather image classification. By leveraging advanced architectural designs, both models excelled in capturing the intricate nuances of diverse weather conditions, achieving test accuracies

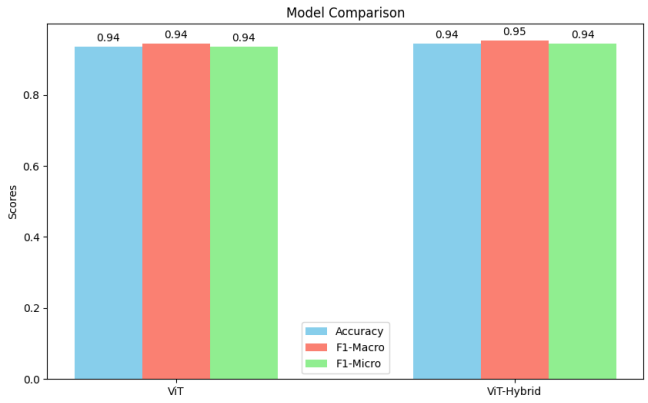

Figure 9: ViT vs ViT Hybrid Comparsion

of **93.59%** and **94.32%** for ViT and ViT-Hybrid, respectively.

The ViT model's superior recall for certain complex classes, such as *glaze*, and the ViT-Hybrid's advantages in capturing local textures, such as *rime*, highlight the complementary strengths of these architectures. While both models demonstrated robust generalization, the ViT-Hybrid's slight edge in stability suggests that hybrid designs could be particularly effective in scenarios requiring both local and global feature extraction.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Xiao, Haixia, Feng Zhang, Zhongping Shen, Kun Wu, and Jinglin Zhang. Classification of weather phenomenon from images by using deep convolutional neural network. In *Earth and Space Science 8, no. 5 (2021): e2020EA001604*, 2021.

Ship, Eden, Eitan Spivak, Shubham Agarwal, Raz Birman, and Ofer Hadar. Real-Time Weather Image Classification with SVM. In *arXiv preprint arXiv:2409.00821*, 2024.

Ridnik, Tal, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *arXiv preprint arXiv:2104.10972*, 2021.

Kolesnikov, Alexander, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp. 491-507. Springer International Publishing*, 2020.