

Real vs AI image Classification (Aadya Arora and Kishan Ved)

1. Approach

The objective of this project is to detect deepfake images by leveraging both RGB content and frequency-based artifacts. Our method enriches input images with edge-based frequency components and utilizes a convolutional vision transformer (LeViT) for binary classification (real vs. fake).

The approach involves:

- Constructing a custom 6-channel dataset by augmenting each image with frequency maps derived from horizontal and vertical gradients.
 - Adapting the LeViT architecture to accept 6-channel inputs.
 - Employing focal loss to address class imbalance.
 - Training with transfer learning and a step learning rate scheduler.
-

2. Model Architecture and Design Decisions

Base Model: LeViT

We use the `LeViT-128` model from the `timm` library, which combines convolutional and transformer components, making it computationally efficient while retaining attention-based modeling power.

Modifications

- **Input Layer:** LeViT expects 3-channel inputs. We replace the first convolution layer to accept 6 channels. The new weights are initialized by duplicating the pretrained weights across the added channels to retain generalization.
- **Input Composition:** The first 3 channels are standard RGB, while the next 3 are edge frequency maps, computed using simple horizontal and vertical pixel differences.
- **Output Layer:** A fully connected linear layer maps the global average pooled output to a single logit for binary classification. In our experimentation, we tested and tuned the

logits and finally set it to 0.6

3. Performance Analysis

Training Setup

- **Optimizer:** Adam with learning rate $1e-4$ and weight decay $1e-5$
- **Loss:** Focal loss ($\alpha=0.75$, $\gamma=2$) to prioritize hard examples
- **Scheduler:** StepLR ($\gamma=0.1$ every 10 epochs)
- **Epochs:** 30
- **Batch Size:** 32
- **Hardware:** Supports multi-GPU with DataParallel

Validation Metrics

- Binary accuracy is reported at the end of each epoch.
- Validation accuracy is printed to monitor overfitting and guide early stopping.

Observations

- Initial experiments show that the frequency-augmented model outperforms the vanilla LeViT on a held-out validation set.
 - The frequency channels aid in distinguishing visually similar but structurally inconsistent fake images.
 - Focal loss significantly improves learning on under-represented fake samples.
-

5. Known Limitations and Potential Improvements

Limitations

- **Explainability is implicit:** While the model uses frequency features, it lacks post-hoc explainability tools (e.g., GradCAM, LIME) to generate visual explanations.
 - **Hardcoded frequency extraction:** The method uses basic gradient-based frequency extraction, which may not generalize well across datasets.
 - **Noisy edge channels:** Simple differencing might introduce noise or amplify compression artifacts.
-

Conclusion

This project demonstrates a lightweight, explainable deepfake detection pipeline that combines handcrafted frequency features with a transformer-based backbone. The approach shows promising results in enhancing detection accuracy with minimal architectural modifications while maintaining interpretability through inductive biases.