Deepfake Detection using Attention-Driven CNN and TTA

Robust Deepfake Detection via ResNet-18 with Explainable Grad-CAM and Test-Time Augmentation

Vasu Tandon

February 16, 2026

Abstract

The proliferation of hyper-realistic deepfake content poses a significant threat to digital media integrity. This paper presents a lightweight yet robust deepfake detection system based on the ResNet-18 architecture, optimized for resource-constrained environments. Our approach integrates Grad-CAM (Gradient-weighted Class Activation Mapping) for model explainability and employs a Test-Time Augmentation (TTA) strategy involving horizontal flips and center crops to enhance prediction reliability. We introduce an 'Aggressive Detection' thresholding mechanism to minimize false negatives in high-stakes scenarios. Trained on a balanced subset of 40,000 images from the '140k Real and Fake Faces' dataset, the model demonstrates high efficiency and reliable classification performance.

1. Introduction

Deep learning advancements have democratized the creation of synthetic media, necessitating equally advanced detection mechanisms. While large transformer-based models offer high accuracy, they often require significant computational resources. This research focuses on optimizing a convolutional neural network (CNN) for efficiency without compromising detection capabilities. We utilize a modified ResNet-18 backbone and introduce a resource-aware training loop that includes active cooling pauses to maintain hardware stability during training on standard laptops. Furthermore, we address the 'black box' nature of CNNs by embedding Grad-CAM visualization directly into the inference pipeline, allowing users to see which facial features triggered a 'fake' classification.

2. Methodology

2.1 Dataset and Preprocessing

We utilized the '140k Real and Fake Faces' dataset, curating a balanced subset of 20,000 real and 20,000 fake images for training. Images were resized to 224x224 pixels and normalized using ImageNet standards (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). A custom 'FastDeepfakeDataset' class was implemented to handle efficient data loading with error resilience.

2.2 Model Architecture

The core architecture is a ResNet-18 model, chosen for its balance between depth and computational cost. The final fully connected layer was adapted for binary classification (Real vs. Fake). We utilized Binary Cross

Entropy with Logits Loss (BCEWithLogitsLoss) and the AdamW optimizer (lr=5e-5) for stable convergence.

2.3 Test-Time Augmentation (TTA)

To improve inference robustness, we implemented TTA. For every input image, the model predicts on three variations: (1) the original image, (2) a horizontally flipped version, and (3) a 90% center-cropped zoom. The final confidence score is the average of these three probabilities. This ensemble-like approach reduces susceptibility to noise and pose variations.

2.4 Explainability via Grad-CAM

We integrated Grad-CAM by hooking into the final convolutional layer (layer4[-1]) of the ResNet backbone. This generates a heatmap highlighting the regions most influential in the model's decision, providing interpretability for end-users.

3. Experimental Results & Discussion

The model was trained for 4 epochs with a batch size of 16. To mitigate overfitting and ensure high detection rates for manipulated content, we implemented an 'Aggressive Mode' during inference. Any image with a fake probability score greater than 0.30 via TTA is flagged as 'DEEPFAKE'. This lowers the threshold for detection, prioritizing recall for security-critical applications. During validation, the model achieved consistent accuracy improvements, demonstrating the effectiveness of the TTA strategy in refining borderline predictions.

4. Conclusion

We successfully developed a deepfake detection system that balances performance with interpretability. By combining a lightweight ResNet-18 backbone with Test-Time Augmentation and Grad-CAM, the system provides reliable and explainable predictions suitable for deployment on consumer hardware. Future work will explore temporal analysis for video-based deepfake detection.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition,' CVPR, 2016.
[2] R. R. Selvaraju et al., 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,' ICCV, 2017.
[3] Xhlulu, '140k Real and Fake Faces Dataset,' Kaggle, 2020.