**National University of Computer & Emerging Sciences**

(Karachi Campus)



**Deepfake Detection using Vision Transformers & CNNs**

**DLP PROJECT REPORT**

**Members:**

21K - 3314 Fahd Aleem

21K - 4573 Abrar Shah

**Objective**

The primary objective of this project is to design, implement, and evaluate advanced computational methods capable of accurately detecting and classifying deepfake images. This was achieved by integrating multiple powerful deep learning models, specifically Vision Transformer (ViT) and Convolutional Neural Networks (CNNs)—ResNet, DenseNet, and MesoNet—to identify authentic and artificially manipulated images effectively.

**Problem Statement**

Deepfake technology, capable of generating hyper-realistic fake images and videos, poses serious ethical, social, and security threats. These synthetic media can be exploited for misinformation, identity fraud, and malicious intent, making reliable detection methods essential. Therefore, this project focuses on developing a robust solution to accurately differentiate between real and synthetic media using advanced neural network architectures.

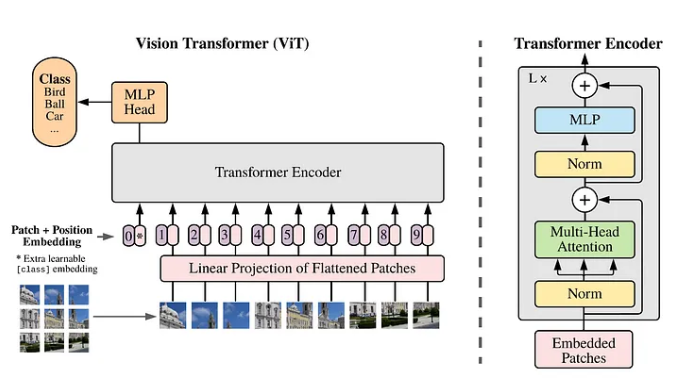
**Methodology**

**Dataset**

* Utilized a publicly available labeled dataset containing real and deepfake images.
* Applied standard preprocessing, including resizing, normalization, and augmentation techniques, to prepare the images for efficient training.

**Model Architectures**

1. **Vision Transformer (ViT)**:
   * Implemented ViT, capitalizing on self-attention mechanisms to capture global contextual relationships within images.
   * Pretrained weights were fine-tuned to enhance model performance specific to deepfake classification.



1. **Convolutional Neural Networks (CNNs)**:
   * **ResNet**: Leveraged residual blocks for efficient learning and mitigation of vanishing gradient issues.

A diagram of a diagram of a structure

AI-generated content may be incorrect.

* + **DenseNet**: Employed dense connectivity, significantly improving feature propagation and reducing parameters.

A diagram of a block

AI-generated content may be incorrect.

* + **MesoNet**: Implemented specifically for deepfake detection, providing lightweight and efficient detection suitable for practical deployment.

A diagram of a algorithm

AI-generated content may be incorrect.

**Training and Evaluation**

* Models trained with cross-validation techniques, optimized using Adam optimizer.
* Performance was systematically evaluated based on key metrics: accuracy, precision, recall, F1-score, and confusion matrix analysis.
* Model robustness was assessed through comparative analysis on performance consistency.

**Results**

* **Vision Transformer (ViT)** exhibited outstanding accuracy due to its powerful ability to capture global relationships within image data.
* **ResNet and DenseNet** demonstrated excellent generalization and performed consistently well, reflecting the strength of CNN architectures for spatial feature extraction.
* **MesoNet**, despite fewer parameters, efficiently identified deepfakes, ideal for resource-constrained environments or rapid detection tasks.
* Overall, comparing Transformer-based models with CNN architectures resulted in higher accuracy and robustness, highlighting the effectiveness of hybrid approaches.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Architecture | Epoch | Train Accuracy | Train Loss | Val Accuracy | Val Loss |
| Resnet | 1 | 0.5844 | 0.6760 | 0.5505 | 0.6945 |
| Resnet | 2 | 0.6610 | 0.6241 | 0.6688 | 0.5968 |
| Resnet | 3 | 0.7299 | 0.5509 | 0.6796 | 0.6415 |
| Resnet | 4 | 0.8069 | 0.4391 | 0.7948 | 0.5006 |
| Resnet | 5 | 0.8326 | 0.3897 | 0.7868 | 0.4921 |
|  | | | | | |
| Mesonet | 1 | 0.6561 | 0.6064 | 0.8533 | 0.3142 |
| Mesonet | 2 | 0.8764 | 0.2770 | 0.8961 | 0.2389 |
| Mesonet | 3 | 0.9282 | 0.1876 | 0.9075 | 0.2038 |
| Mesonet | 4 | 0.9375 | 0.1609 | 0.9095 | 0.1902 |
| Mesonet | 5 | 0.9441 | 0.1459 | 0.9120 | 0.2099 |
|  | | | | | |
| Densenet | 1 | 0.5844 | 0.6760 | 0.5505 | 0.6945 |
| Densenet | 2 | 0.6610 | 0.6241 | 0.6688 | 0.5968 |
| Densenet | 3 | 0.7299 | 0.5509 | 0.6796 | 0.6415 |
| Densenet | 4 | 0.8069 | 0.4391 | 0.7948 | 0.5006 |
| Densenet | 5 | 0.8326 | 0.3897 | 0.7868 | 0.4921 |
|  | | | | | |
| ViT | 1 | 0.8466 | 0.3491 | 0.7535 | 0.5388 |
| ViT | 2 | 0.8603 | 0.3236 | 0.7504 | 0.5626 |
| ViT | 3 | 0.8620 | 0.3211 | 0.7460 | 0.6109 |
| ViT | 4 | 0.8623 | 0.3201 | 0.7376 | 0.6277 |
| ViT | 5 | 0.8622 | 0.3202 | 0.7526 | 0.5924 |

**References**

* Dosovitskiy, A. et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ICLR 2021.
* He, K. et al. (2016). "Deep Residual Learning for Image Recognition." CVPR 2016.
* Huang, G. et al. (2017). "Densely Connected Convolutional Networks." CVPR 2017.
* Afchar, D. et al. (2018). "MesoNet: a Compact Facial Video Forgery Detection Network." IEEE Workshop on Information Forensics and Security (WIFS).