AI-Driven Distributed Data Management for Deepfake Images in Smart Cities

ANAND M K - 242CS008

1 Theory Plan

Deepfake technology has emerged as a critical challenge in modern smart cities, affecting security, privacy, and public trust. Deepfake images and videos can be used maliciously for misinformation, identity theft, and fraudulent activities. Thus, there is a pressing need for AI-driven approaches to detect and manage deepfakes efficiently within a distributed framework.

This study will begin with an extensive literature review covering various deepfake detection techniques, focusing on deep learning architectures such as Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid models that combine spatial and temporal analysis for better detection accuracy. Additionally, traditional detection techniques based on statistical and frequency domain analysis, such as Discrete Cosine Transform (DCT) and wavelet transformations, will be examined to complement AI-driven methods.

Another essential aspect of this research is the role of distributed AI in managing deepfake datasets. Conventional deepfake detection methods require centralized data storage, which raises privacy concerns. This study will explore federated learning, a decentralized AI training method that allows multiple devices to contribute to model training without sharing raw data. By investigating the advantages and limitations of federated learning, the research aims to highlight its potential in enhancing privacy-preserving deepfake detection in smart city applications.

Challenges such as adversarial deepfakes, where AI-generated modifications deceive detection models, will also be analyzed. Methods such as adversarial training, data augmentation, and explainable AI techniques will be considered to make deepfake detection more robust and interpretable.

Finally, ethical and legal aspects of deepfake management in smart cities will be discussed. The study will provide insights into policy recommendations and future directions, such as integrating AI-driven regulatory frameworks and public awareness initiatives to mitigate deepfake threats effectively.

2 Lab Plan

The practical implementation of this research will focus on developing a deepfake detection system that can efficiently classify real and fake images while ensuring scalability in a distributed environment.

The first step involves dataset selection. Publicly available datasets such as FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge dataset will be used to train and evaluate models. The dataset will undergo preprocessing, including resizing, normalization, and augmentation techniques such as rotation, flipping, and noise injection to enhance model generalization.

Next, a feature extraction pipeline will be designed using OpenCV and deep learning frameworks such as TensorFlow and PyTorch. The model architecture will include a fine-tuned deep learning model, such as EfficientNet or Xception, optimized for detecting deepfake artifacts. Transfer learning will be employed to reduce computational costs and improve performance on limited datasets.

The model will be trained using GPU acceleration, and its performance will be evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. Additionally, Grad-CAM visualizations will be used to interpret the model's decision-making process and highlight regions of manipulated content.

To simulate real-world deployment scenarios, the trained model will be tested on unseen deepfake images and videos to assess its robustness.

Finally, the experimental results will be analyzed to understand the strengths and weaknesses of the approach. Limitations such as dataset bias, computational constraints, and model generalizability will be discussed, along with potential future improvements, including adversarial training and multi-modal deepfake detection techniques.