Project Proposal: Enhancing Deepfake Detection with ID-Miner Extensions

Team Members Details

[VAIBHAV KUMAR, 2021mcb1219@iitrpr.ac.in, **2021MCB1219**]

[HITESH SINGLA, 2021csb1094@iitrpr.ac.in, **2021CSB1094**]

[CHINMAY UDAYBHANU THAKUR, 2021mcb1233@iitrpr.ac.in, **2021MCB1233**]

[KANHAIYA KUMAR SAHU, 2021mcb1235@iitrpr.ac.in, **2021MCB1235**]

[KARAN DEEP DAS, 2021mcb1236@iitrpr.ac.in, **2021MCB1236**]

Problem Statement

With the rapid advancement of deepfake generation techniques, traditional detection models face significant challenges in identifying high-quality, artifact-free deepfakes. ID-Miner, a recently proposed deepfake detection model, focuses on identity-anchored detection rather than artifact-based methods. However, its effectiveness can be further improved by incorporating **transformer-based sequence modeling, expanded benchmarking, and evaluation on diverse datasets**.

This project aims to enhance ID-Miner's capabilities and evaluate its performance under a broader set of conditions to develop a more robust deepfake detection framework.

Background and Summary of the Problem

Deepfake detection has become increasingly difficult as generative models evolve, producing highly realistic fake videos. Many existing detection methods rely on identifying digital artifacts, which may no longer be present in the most advanced deepfake techniques.

ID-Miner attempts to address this challenge by leveraging **motion-based identity detection** rather than artifact detection. However, its current design may not fully exploit advanced sequence modeling techniques, such as transformers, which have demonstrated superior performance in capturing long-range dependencies. Additionally, the evaluation of ID-Miner has been conducted on a limited set of deepfake detection scenarios and datasets, which may not fully reflect real-world conditions.

To enhance deepfake detection and improve model generalization, this project proposes the following:

1. **Extending ID-Miner with Transformer-Based Sequence Modeling** to improve identity-anchored feature extraction.
2. **Benchmarking multiple deepfake detection models** under the RDDP evaluation framework to assess their robustness.
3. **Validating performance on diverse datasets** to test real-world applicability and generalization.

Proposed Solution

Enhancement 1: Extending ID-Miner with Transformer-Based Sequence Modeling

- Replace LSTMs or GRUs in ID-Miner with self-attention-based models such as **TimeSformer or Perceiver IO**, which have demonstrated improved performance in temporal feature extraction.
- Incorporate **spatiotemporal transformer encoders** to better capture frame-level motion patterns while preserving identity consistency across video sequences.
- Train models using the **RDDP-WHITEHAT and RDDP-SURROGATE** datasets and experiment with different transformer depths to identify the optimal architecture.

Enhancement 2: Applying the RDDP Evaluation Framework to Other Deepfake Detection Models

- Evaluate the performance of multiple state-of-the-art deepfake detection models, including **XceptionNet, EfficientNet, LSTM-based models, and Transformer-based architectures**, under both **standard deepfake detection protocols and the RDDP framework**.
- Compare model performance using key metrics, such as **precision, recall, F1-score, and robustness against unseen deepfakes**.
- Analyze whether certain models exhibit a greater dependence on digital artifacts and propose potential training strategies, such as **adversarial training or self-supervised learning**, to improve their effectiveness.

Enhancement 3: Evaluating on Diverse Datasets for Improved Generalization

- Expand testing to datasets beyond those originally explored, incorporating **WildDeepfake, DeepFake-TIMIT, ForgeryNet, and FaceShifter** to evaluate model generalization across different deepfake creation techniques.
- Conduct **cross-dataset learning** by training models on one dataset and testing them on another to assess their adaptability.

- If significant performance degradation is observed, investigate **domain adaptation techniques**, such as **few-shot learning and self-supervised learning**, to improve model transferability.

Objectives

- Enhance ID-Miner by integrating transformer-based sequence modeling for improved deepfake detection.
- Benchmark multiple deepfake detection models under the RDDP evaluation framework to assess their robustness.
- Evaluate model performance across diverse datasets to test real-world applicability.
- Identify and propose architectural enhancements to improve deepfake detection methods.

Pipeline Flowchart

1. **Data Collection and Preprocessing** – Obtain and preprocess deepfake datasets.
2. **Enhancement of ID-Miner** – Integrate transformer-based sequence modeling.
3. **Training and Evaluation** – Train models under both standard protocols and the RDDP framework.
4. **Benchmarking Multiple Models** – Assess CNNs, LSTMs, and Transformer-based models for deepfake detection.
5. **Cross-Dataset Generalization Testing** – Evaluate model adaptability across different datasets.
6. **Result Analysis and Recommendations** – Identify performance gaps and propose improvements.