Project Report – Privacy-Preserving and Explainable Federated Learning for Robust Digital Forensics

Advanced Research Topics in IT Security (ARTIS)

Omar Abushanab *(mail)*
Ibrahim Selim *(mail)*
Malak Abdelaziz *(mail)*
Peter Schropp *(mail)*
Matthias Högel *(matthias.hoegel@uni-ulm.de)*

January 4, 2026

1. Introduction

Introduction of the paper.

2. Related Work

Related work of the paper.

3. Background

3.1. Digital Forensics

3.2. Federated Learning

3.3. Attack Vectors in FL

3.4. CNNs

3.5. Fundamentals

3.6. Transfer Learning

3.7. Model Explainability using SHAP

4. Methodology

4.1. Centralized Fake Face Detection Model

The authors of [1] developed a method for fake face detection using CNN, which achieves an accuracy of 99.06This approach is to be used as a benchmark for comparing central fake face detection with fake face detection using FL. This work was reimplemented for verification purposes in order to ensure a meaningful comparison.
The dataset used was the well-known 140k real-fake faces dataset, which consists of 70,000 real faces and 70,000 fake faces. The dataset was divided into 100.000 training images, 20.000 test images, and 20.000 validation images. EfficientNetB0 with transfer learning was used as the model. For this purpose, the model was initialized with the pretrained weights of EfficientNetB0 on the ImageNet dataset. A lightweight head was attached to the pre-trained base model, consisting of global average pooling, a 256-dimensional fully connected layer with ReLU activation, batch normalization, and dropout, followed by a 2 dimensional softmax output layer. The output is a 2-dimensional vector with probabilities indicating whether the input image is a fake or

Epoch	**Learning rate**
$epoch \leq 2$	0.01
$2 < epoch \leq 15$	0.001
$epoch > 15$	0.0001

Table 1: Adjustment of the learning rate during training.

real face. The model was optimized with binary cross-entropy loss and the Adam optimizer. Training was performed with a batch size of 32 over 30 epochs with early stopping to reduce training time.
The paper presented a learning rate scheduler that adjusts the learning rate during training based on the epoch, as shown in Table 1. The learning rate scheduler ensures that significant weight adjustments are made early in training. Furthermore, in later iterations, a strong adjustment is prevented by the decreasing learning rate. This leads to faster convergence in early epochs, while weight optimizations can be performed in later itterations.
Due to time constraints, the model was only trained once, as training was very computationally intensive due to the large amount of data.

4.2. Decentralized Fake Face Detection using FL

In this section a decentralized trained fake face detection model is to be developed. The training process will be explained below. In addition, we will discuss how privacy attacks can be prevented during training.

4.2.1 Research Scenario

The following scenario is fictional.
Several research organizations want to work together to train a fake face detection model. Each individual organization has images of fake faces, but also images of faces that belong to their customers. An ML model should be trained together that can distinguish between real and fake images. To do this, a large data set containing all images would have to be created in order to train the model. However, all organizations are interested in protecting the privacy of their customers and therefore do not want to share the real images. The solution is to train the model using FL. This fake face detection model should be trained with all data from all organizations and should be available to everyone without the need to share data between organizations.

4.2.2 Thread Model

All participants in the fake face detection training process are trusted. This means that all model performance attacks that seek to undermine the convergence of the global model can be excluded. After training, participants send their model weights to a trusted third-party server. This ensures the secure aggregation of weights. Weights sent by participants to the trusted third party could be captured during transmission. Capturing the weights of individual organizations represents an attack vector for privacy attacks. This attack vector should be reduced by the proposed encryption.

4.2.3 FL attack mitigations

4.3. Explainability using SHAP

5. Results

This are the results of the paper.

6. Discussion

This is the discussion of the paper.

7. Conclusion

This is the conclusion of the paper.

References

[1] Raidah Salim Khudeyer and Noor Mohammed Almoosawi. “Fake Image Detection Using Deep Learning”. In: *Informatica* 47.7 (2023).