Comparative Analysis of Feature Extraction Techniques for Robust Deepfake Image Detection

Vittorio Pisapia - 1918590

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents

Introduction

Introduction to the problem

Introduction

- Deepfake image detection has become a critical challenge in computer vision, due to the development of generative AI.
- Deepfakes raise critical ethical, legal and security issues.
- Robust techniques are essential for identifying manipulated or generated content.
- The goal of the project is to compare various feature extraction techniques for deepfake image detection: LBP, HOG, CNNS...

Human Performance

Introduction

- 24 partecipants were tested and 132 runs were collected.
- Each partecipant was presented with 10 images and asked to classify them as fake or real. The required time was recorded for each run.
- **Mean accuracy**: 65%
- **Avarage time per test**: 49 seconds

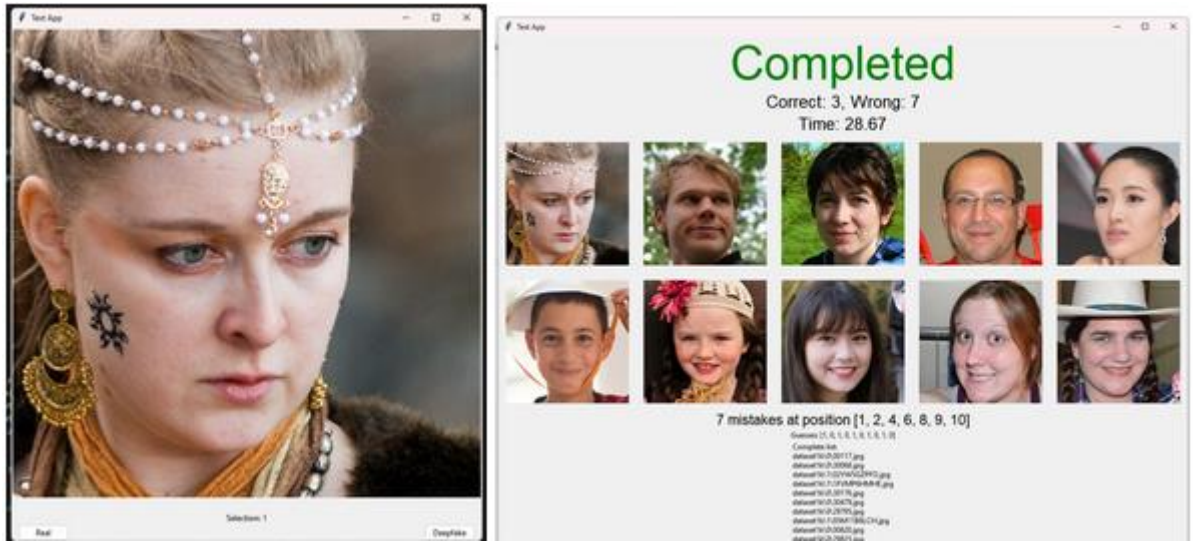


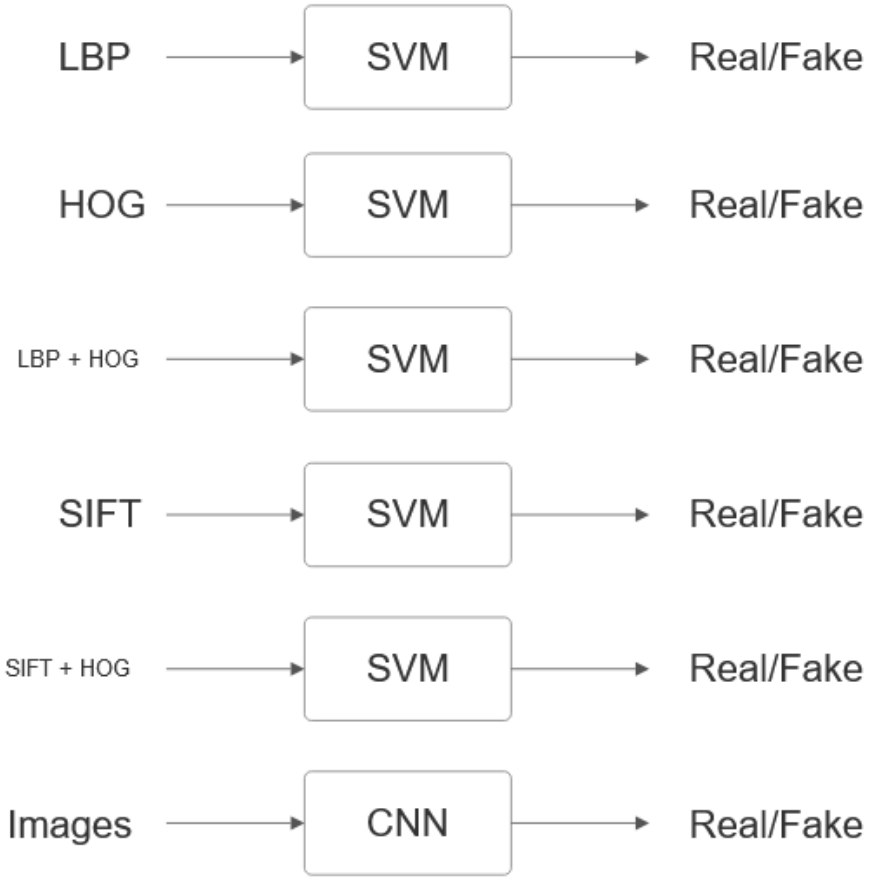
Table of Contents

Introduction

Traditional Feature Extraction

Feature Extractors

- **LBP**: Texture descriptor, encodes relationship between a pixel and its neighboring pixels.
- **HOG**: Focuses on gradient orientation distributions, making it effective in detecting shapes and edges.
- **SIFT**: identifies and describes local keypoints of a image
  - **Dense SIFT**: keypoints are predetermined points arranged on a grid.
  - **Mean SIFT**: Computes the mean of all SIFT descriptor found in the image.



Deep Learning-based feature extraction

Feature Extractors

- **CNN**: Designed to perform feature extraction and classification in a single pipeline. Directly trained on raw image data.

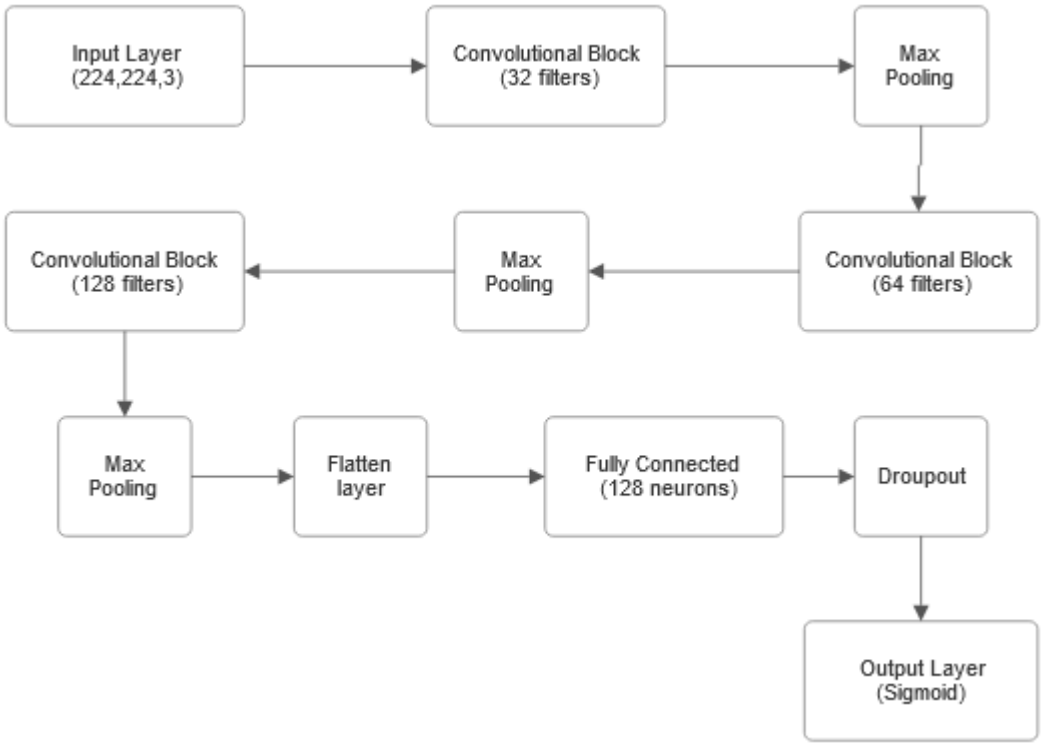


Table of Contents

Introduction

Dataset and Training Data

Dataset

- The dataset includes both real and fake images:
  - **Real images**: Flickr-Faces-HQ (FFHQ) provided by Nvidia.
  - **Fake images**: ‘’1 Million Fake Faces’’ created with StyleGand and provided by Bojan Tunguz.
- A total of 3000 samples were used for training (1500 real images and 1500 fake images).
- For simpler feature extracton methods (LBP), 20.000 total samples were tested for training, but it led to similar performance while significantly increasing training time.

Table of Contents

Introduction

Accuracy, Efficiency and Robustness

Evaluation Metrics

- Performance will be evaluated using the following criteria:

1. **Accuracy**: evaluated alongside F1-score to measure balance between precision and recall.

2. **Computational Efficiency**: Extraction times and training times will be compared to assess computational performance.

3. **Robustness to Adversarial Attacks**:Using adversarial attack generation techniques (Adversarial Robustness Toolbox), the resilience of each model against perturbed inputs will be analyzed.

Adversarial Robustness

Evaluation Metrics

- Adversarial robustness was evaluated using two attacks from the Adversarial Robustness Toolbox (ART):

1. **Fast Gradient Method (FGM):**
   FGM computes the gradient of the loss function with respect to the input features. Adversarial examples are generated by perturbing the input features in the direction that maximizes the loss function. This is a straightforward and computationally efficient attack method.

2. **Carlini and Wagner L2 Attack (CL2):**
   A state-of-the-art iterative method, CL2 formulates adversarial example generation as an optimization problem. It seeks the minimal perturbation required to misclassify the model, offering a more precise and effective attack compared to FGM.

Table of Contents

Introduction

Accuracy, f1-score and training time

Results

FEATURE EXTRACTOR	Training time (seconds)	accuracy	f1-score
LBP	86,17	0,64	0,64
HOG	67,97	0,74	0,74
LPB+HOG	1908,29	0,82	0,82
Mean SIFT	47,84	0,62	0,62
Dense SIFT	1848,04	0,82	0,82
Mean SIFT + HOG	2125,73	0,81	0,81
CNN (20ep)	919,38	0,69	0,69
CNN (10ep)	456,57	0,7	0,7
CNN (5ep)	232,03	0,67	0,66

- **Training time**:
  - **Mean SIFT** (47.84 seconds) is the fastest, followed by **HOG** (67.97 seconds) and **LBP** (86.17 seconds), highlighting their efficiency.
  - **CNN** (5 epochs) has a reasonable training time (232.03 seconds).
  - **LBP + HOG** (1908.29 seconds), **Mean SIFT + HOG** (2125.73 seconds), and **Dense SIFT** require significantly more time due to high-dimensional input space.
  - **CNN** (20 epochs) is a middle ground at 919.38 seconds.

Accuracy, f1-score and training time - 2

Results

FEATURE EXTRACTOR	Training time (seconds)	accuracy	f1-score
LBP	86,17	0,64	0,64
HOG	67,97	0,74	0,74
LPB+HOG	1908,29	0,82	0,82
Mean SIFT	47,84	0,62	0,62
Dense SIFT	1848,04	0,82	0,82
Mean SIFT + HOG	2125,73	0,81	0,81
CNN (20ep)	919,38	0,69	0,69
CNN (10ep)	456,57	0,7	0,7
CNN (5ep)	232,03	0,67	0,66

- **Accuracy and F1-score**:
  - **Dense SIFT** (0.82), **LBP + HOG** (0.82) and **Mean SIFT + HOG** (0.81) are the best performers, demonstrating the benefit of combining complementary features.
  - **HOG** (0.74) achieves a strong performance despite its semplicity and fast training time, outperforming **LBP** (0.64)
  - **CNN** (10 epochs) achieves an accuracy and f1-score of 0.70, outperforming its counterpart with fewer epochs and approaching **HOG**'s performance.

Adversarial Robustness - LBP

Results

LBP	Mean of original L2 norms	Mean of L2 norm of differences	% wrt mean of original L2 norms	Successful attacks (out of 100)
FGM eps = 0.01	4,7768	0,051	1,067	5
FGM eps = 0.03	4,7768	0,153	3,2	15
FGM eps = 0.08	4,7768	0,4079	8,53	49
CL2 c = 0.05	4,7768	0,5024	10,51	95
CL2 c = 0.08	4,7768	0,5491	11,49	95
CL2 c = 0.1	4,7768	0,5695	11,92	95

- **FGM Attack**:
  - **LBP** shows robustness against FGM, with only 49% of successful attacks even at high epsilon value (0.08)
  - For epsilon = 0.08, the perturbation induces a 8.53% variation in the L2 norms.
- **CL2 Attack**:
  - **LBP** is significantly more vulnerable to the advanced CL2 attack, with a 95% success rate even at lower confidence thresholds.

Adversarial Robustness - HOG

Results

HOG	Mean of original L2 norms	Mean of L2 norm of differences	% wrt mean of original L2 norms	Successful attacks (out of 100)
FGM eps = 0.01	89,7785	0,9	1,002	22
FGM eps = 0.03	89,7785	2,7	3,007	73
FGM eps = 0.08	89,7785	7,2	8,019	100
CL2 c = 0.05	89,7785	0,3745	0,4171	41
CL2 c = 0.08	89,7785	0,3963	0,4414	40
CL2 c = 0.1	89,7785	0,3464	0,3858	36

- **FGM Attack**:
  - **HOG** features shows vulnerability to FGM attacks: at epsilon = 0.01, 22% of the attacks succeed , despite only a 1% difference in the L2 norms.
  - At higher epsilon values, the success rate reaches 100%.
- **CL2 Attack**:
  - **HOG** performs well against CL2 attacks, keeping the success rate under 50% across all confidence level.

Adversarial Robustness - Mean SIFT

Results

Mean SIFT	Mean of original L2 norms	Mean of L2 norm of differences	% wrt mean of original L2 norms	Successful attacks (out of 100)
FGM eps = 0.01	11,0991	0,1131	1,019	16
FGM eps = 0.03	11,0991	0,3394	3,057	45
FGM eps = 0.08	11,0991	0,9051	8,1547	89
CL2 c = 0.05	11,0991	0,5006	4,5102	95
CL2 c = 0.08	11,0991	0,5656	5,0959	95
CL2 c = 0.1	11,0991	0,6084	5,4815	95

- **FGM Attack**:
  - **Mean SIFT** shows sligtly better performance than **HOG**, especially for epsilon = 0.01 and 0.03
- **CL2 Attack**:
  - **Mean SIFT** demonstrates limited robustness to Carlini and Wagner L2 attacks. For all tested confidence levels, the success rate remains consistently 95%.

Adversarial Robustness – Concatenation and Dense SIFT

Results

LBP + HOG	Mean of original L2 norms	Mean of L2 norm of differences	% wrt mean of original L2 norms	Successful attacks (out of 100)
FGM eps = 0.01	162,3967	1,6208	0,998	24
FGM eps = 0.03	162,3967	4,8624	2,9941	77
FGM eps = 0.08	162,3967	12,9664	7,9843	99

Mean SIFT + HOG	Mean of original L2 norms	Mean of L2 norm of differences	% wrt mean of original L2 norms	Successful attacks (out of 100)
FGM eps = 0.01	162,7266	1,6239	0,9979	20
FGM eps = 0.03	162,7266	4,8718	2,9938	76
FGM eps = 0.08	162,7266	12,9916	7,9836	98

Dense SIFT	Mean of original L2 norms	Mean of L2 norm of differences	% wrt mean of original L2 norms	Successful attacks (out of 100)
FGM eps = 0.01	313,5703	3,1397	1	30
FGM eps = 0.03	313,5703	9,4192	3	81
FGM eps = 0.08	313,5703	25,1178	8.01	99

Adversarial Robustness

Results

- **FGM Attack**:
  - All three methods showed similar performance to **HOG**, with **Dense SIFT** slightly underperforming compared to the concatenated feature methods.
  - **LBP + HOG** and **Mean SIFT + HOG** performed almost identically, achieving a 25% attack success rate at epsilon = 0.01 and 75% at epsilon = 0.03.
  - **Dense SIFT** performed worse, with 30% success at epsilon = 0.01 and 81% at epsilon= 0.03.
  - At epsilon = 0.08, almost all attacks were successful for all models.
- **CL2 Attack**:
  - Due to the high dimensionality of inputs and the iterative nature of the Carlini and Wagner L2 attack, computation for these methods was not feasible with the available resources.

Adversarial Robustness - CNN

Results

CNN (10 epochs)	Accuracy on benign samples	Accuracy on adversarial samples	Success rate of attack (%)
FGM eps = 0.01	0,7	0,59	41
FGM eps = 0.03	0,74	0,54	58
FGM eps = 0.08	0,71	0,54	58

- **FGM Attack**:
  - The CNN demonstrates consistent robustness against FGM attacks across different epsilon values (ϵ=0.01,0.03,0.08). The success rate of attacks remains close to 50% regardless of the increasing epsilon.
  - Even at higher values of epsilon, the model does not show a significant accuracy drop compared to lower epsilon values, suggesting resilience to strong perturbations.
- **CL2 Attack**:
  - The Adversarial Robustness Toolbox currently lacks the implementation of the Carlini and Wagner L2 attack specifically for deep learning models.

Table of Contents

Introduction

Conclusions

Conclusions

- Best performance in terms of accuracy and F1-score were achieved by **Dense SIFT** or by concatenating traditional complementary features, such as **LBP** and **HOG**.
- They generate features with a high number of components, increasing significantly computational time (~2000% of LBP's training time) and resulting in models highly vulnerable to adversarial attacks.

- **HOG** features stood out for the balance between performance and computational efficiency.
- It achieved better accuracy than the tested participants while requiring only ~68 seconds of training time (78.8% of **LBP**'s training time).
- In terms of adversarial robustness, it performed comparably to the **LBP + HOG** model for simpler attacks and demonstrated relatively strong resilience against more complex attacks, with a success rate below 45% at the tested confidence levels.

- **CNN** models, have the potential to achieve higher accuracy levels given sufficient computational resources and larger datasets.
- However, under the constraints of this project, **CNN** underperformed compared to more complex traditional features extractor. Despite their lower accuracy, **CNN** demonstrated greater robustness to FGM attacks for higher values of epsilon.

Future Works

Conclusions

Opportunities for Growth:

- **Further analysis on CL2 attacks:**
  - With additional resources, it would be possible to extend the analysis of CL2 attacks to more complex feature extractors, such as **Dense SIFT**, **LBP + HOG**, and **Mean SIFT + HOG**.

- **Dimensionality optimization:**
  - Using Principal Component Analysis (**PCA**), the dimensionality of the best-performing features could be reduced, embedding them in a lower-dimensional space. This would result in shorter training times and improved computational efficiency.

- **Explore other feature extractor:**
  - Investigate additional feature extraction method.

- **Optimization of CNN model:**
  - Further analysis could focus on developing a more complex and optimized neural network architecture.
  - With increased resources, a larger training dataset could be utilized to fully exploit the advantages of deep learning.

- Thanks for the attention!

Pre-processing

Backup slides

For all SVM-based models, images undergo the following processing:

- Image resizing to 224 x 224 pixels;
- Conversion to grayscale;
- Feature extraction, specific to each model;
  - Feature concatanation (applied to **LBP + HOG** and **Mean SIFT + HOG**);
- Transformation using Kera's StandardScaler;

Features dimensions

Backup slides

Feature Dimensions for SVM-based Models:

- **LBP**: Each feature consists of 26 components.
- **HOG**: Each feature consists of 26.244 components.
- **Mean** Sift: Each feature consists 128 components.
- **Dense SIFT**: Each feature consists 100.352 components.