Team NYCU at Defactify4: Robust Detection and Source Identification of AI-Generated Images Using CNN and CLIP-Based Models

Tsan-Tsung Yang, Kuan-Ting Chen,

I-Wei Chen, Shang-Hsuan Chiang and Wen-Chih Peng

National Yang Ming Chaio Tung University

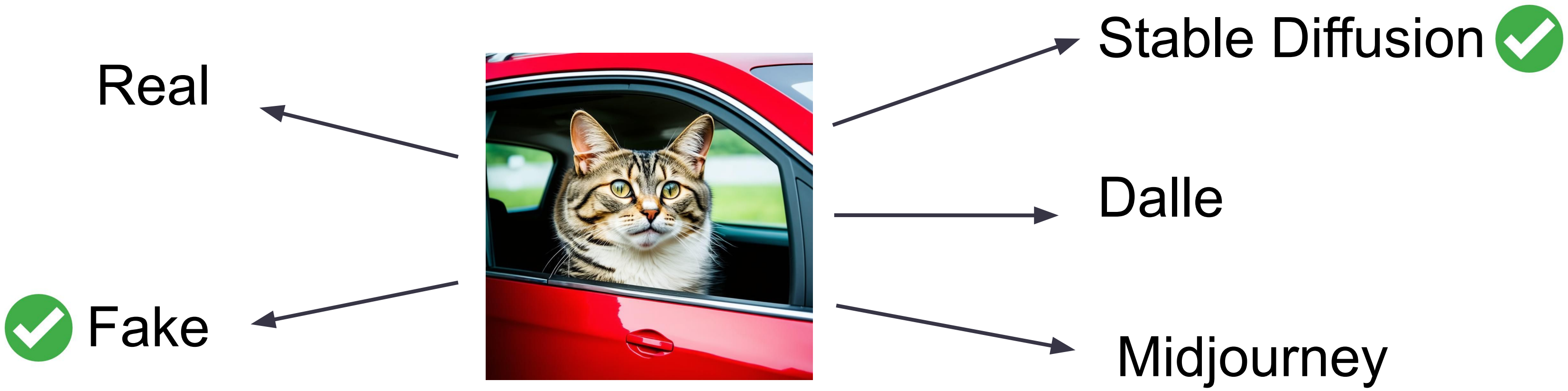
Outline

- Introduction
- Problem
- Solution
- Experiment
- Conclusion

Introduction

Introduction

- Text-to-image generation models are able to produce high-quality images from simple prompts
- To understand and regulate AI-generated images, it is crucial to detect whether the content is real or fake and also to identify the source model
- In real-world scenario, images are often with different perturbations, such as JPEG compression, noise, blurring, and so on



Dataset

- Defactify dataset
  - 4 splits
  - real image and 5 generative models
  - final testing: consist of unknown perturbation

Split Name	Number of Samples
Training	42000
Validation	9000
Testing	9000
Final Testing	45000

Real image from MS-COCO dataset

Stable Diffusion 2.1

Stable Diffusion xl

Stable Diffusion 3

DALL·E

MidJourney

R. Roy, N. Imanpour, A. Aziz, S. Bajpai, G. Singh, S. Biswas, K. Wanaskar, P. Patwa, S. Ghosh, S. Dixit, N. R. Pal, V. Rawte, R. Garimella, A. Das, A. Sheth, V. Sharma, A. N. Reganti, V. Jain, A. Chadha, Overview of image counter turing test: Ai generated image detection, in: proceedings of DeFactify 4: Fourth workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2025.

Problem

Problem

- In the real-world scenario, images are often with different perturbations
  - Noise
  - JPEG compression
  - Blurring
  - Brightness transformation
  - …

Solution

Solution

- We compare two main methods
  - CLIP
  - CNN-based
- We also add different kinds of perturbations for training
  - JPEG compression
  - Gaussian blurring
  - Gaussian noise
  - Brightness transformation

Solution - CLIP

- Backbone: openai/clip-vit-base-patch16
- We trained a SVM classifier based on the pretrained image features
- Perform the grid search to find best parameters set

Solution - CNN

- Backbone: EfficientNet-B0
- We construct more image features, such as VAE reconstruction error and FFT
- Train a CNN classifier from the (512, 512, 5) features

Experiment

Result

- Different models with different noises

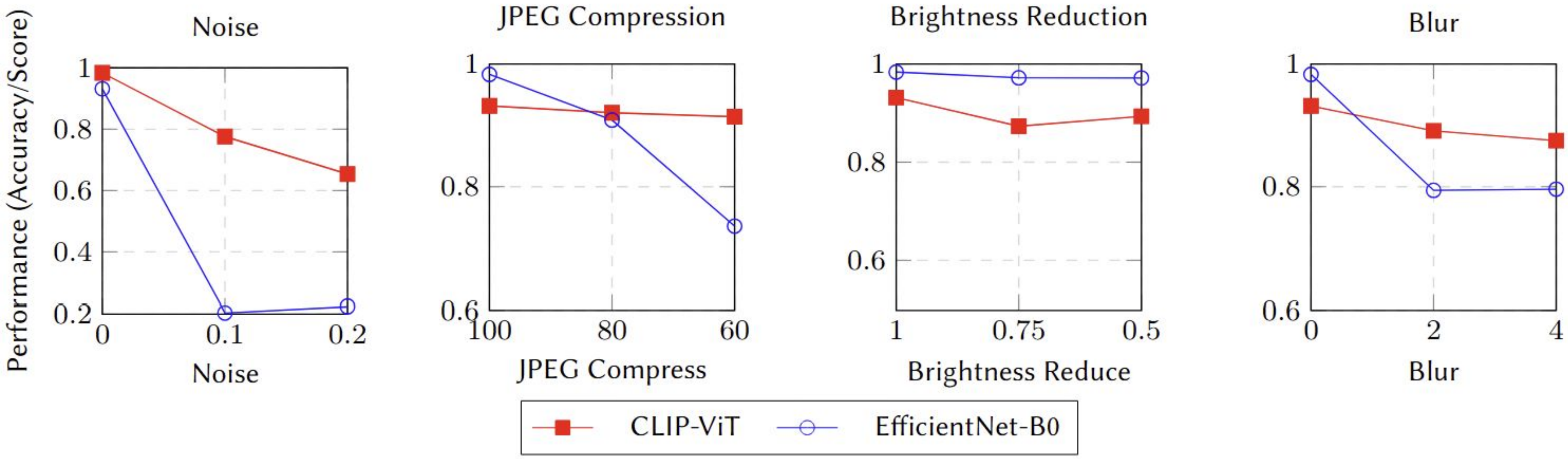


Figure 3: The generalization of CLIP-ViT and EfficientNet on different perturbations. The red square line indicates the CLIP-ViT and the blue circle one indicates EfficientNet. The result shows that CLIP-ViT's performance is better while EfficientNet's performance drops dramatically.

Comparison to Baseline

- We also compare our methods with some SOTA methods
  - AEROBLADE (CVPR 2024)
  - OCC-CLIP (ECCV 2024)
- Our methods achieve competitive results on Task A and outperform all baseline methods on Task B. (This experiment is done on validation set)

Method	Task A		Task B	
	Acc.	F1	Acc.	F1
AEROBLADE	0.8149	0.6986	-	-
OCC-CLIP	**0.9934**	**0.9881**	0.8693	0.8721
Ours: EfficientNet-B0	0.9849	0.9833	**0.9951**	**0.9951**
Ours: CLIP-ViT	0.9421	0.9421	0.9377	0.9317

- For final testing set, we get **0.8329** on Task A and **0.491** on Task B

Ablation Study

- The importance of data augmentation

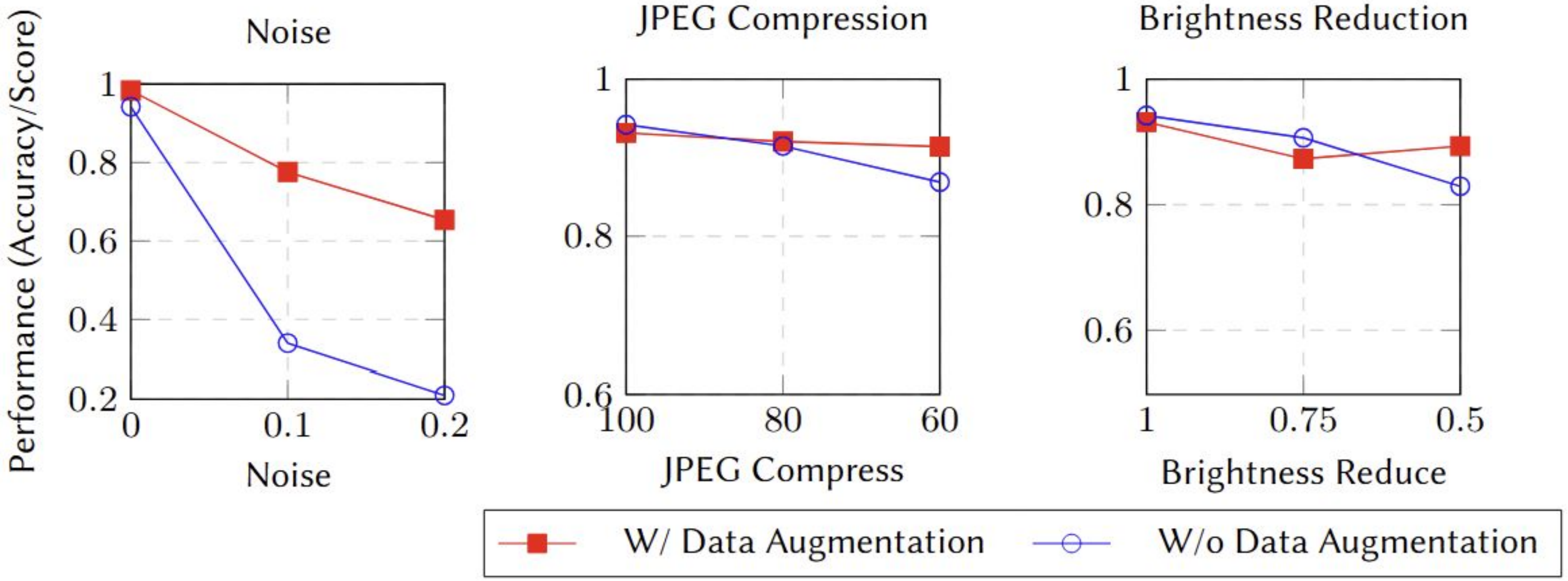


Figure 4: The importance of data augmentation. The red square line indicates training with data augmentation and the blue circle one indicates training without data augmentation.

Conclusion

Conclusion

- Both EfficientNet-B0 and CLIP-ViT models perform well in task A and task B, with CLIP-ViT showing greater robustness against real-world image degradations.
- Our methods achieve competitive or superior results compared to baselines like AEROBLADE and OCC-CLIP, especially in source model identification.
- Data augmentation with perturbations (e.g., Gaussian noise, JPEG compression) significantly improves model generalization and robustness.

Thanks