Reliable Attribution of GAN and Diffusion-based Image Generato

Barbara Hammer

Viktor.Bengs@lmu.de

LMU Munich

Alireza Javanmardi

Alireza.Javanmardi@lmu.de

LMU Munich

Introduction

- advancements of generative AI *show images*
- cant rely on human eye anymore
- potential misuse of AI
- goal: attribute images to sources, point the companies at the misuse of their product, hold them responsible
- capture open set problem with bayesian approach, analyze the effects of post-processing

Background

- GAN vs Diffusion based
- metadata tags -> can be removed
- fingerprints in frequency domain *show images*

Open Set Problem

- models shooting up left and right
- likely to encounter images from models not included in training
- classical approaches make oddly confident predictions since they cant express uncertainty
- approaches to solve this issue e.g. with multiple classifiers

Proposed Approach

- Bayesian Neural Networks
- key components, distribution over weights, updated according to bayes theorem
- measure uncertainty with variance of weights distribution, distribution over classes
- desired outcome with out-of-distribution data: low certainty
- desired outcome with post-processed images: robustness, or at least lost certainty with wrong predictions

Timeline

- development of a baseline model with "classical" components
- development of a Bayesian Neural Network
- compare their predictive reliability
- compare both results on the open world problem
- compare the predictive reliability on post-processed images

References

Appendix

Appendix