Reliable Attribution of GAN and Diffusion-based Image Generato

Lorenz
Lorenz.Gartmeier@campus.lmu.de
LMU Munich

Anatol Maier
anatol.maier@neuraforge.de
supervisor at Neuraforge AI Solutions GmbH

Introduction

source: https://www.theguardian.com/commentisfree/2023/mar/27/pope-coat-ai-image-baby-boomers

Introduction

Convenience Store

Midjourney v3, 2022

Midjourney v6, 2024

source: https://medium.com/@junehao/comparing-ai-generated-images-two-years-apart

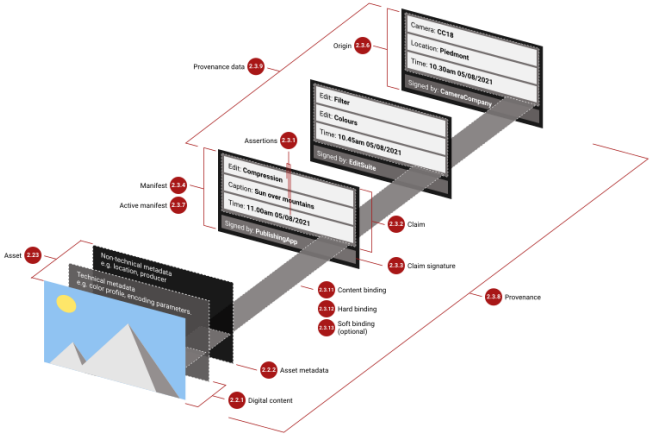
Introduction

Attribute images to source models, so that

- image can be identified as fake
- companies can be pointed at the misuse of their product by users
- companies can be hold responsible
- images from unknown sources don't get falsely attributed

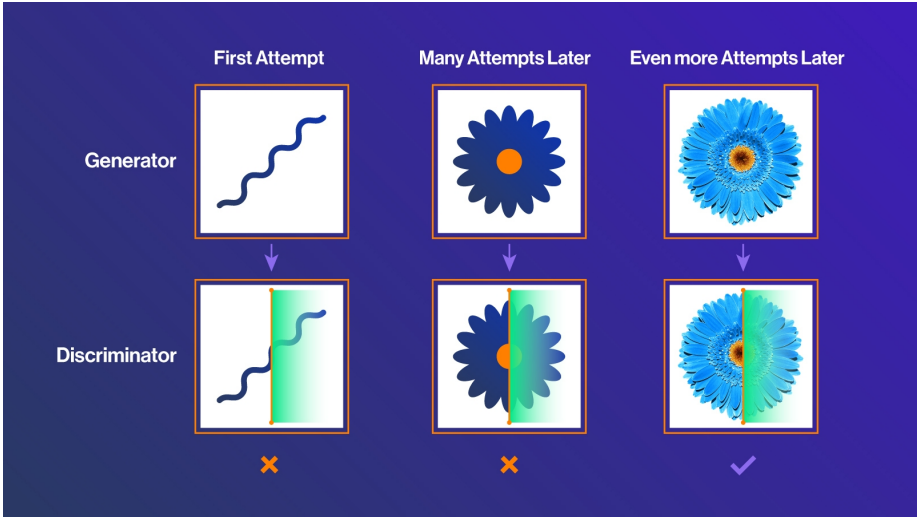
by using a Bayesian Deep Learning approach

Background



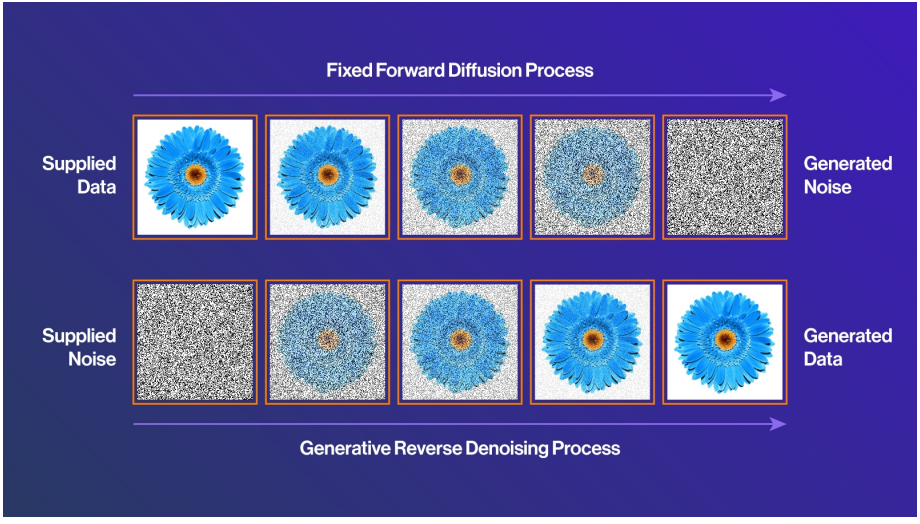
source: https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html

Background



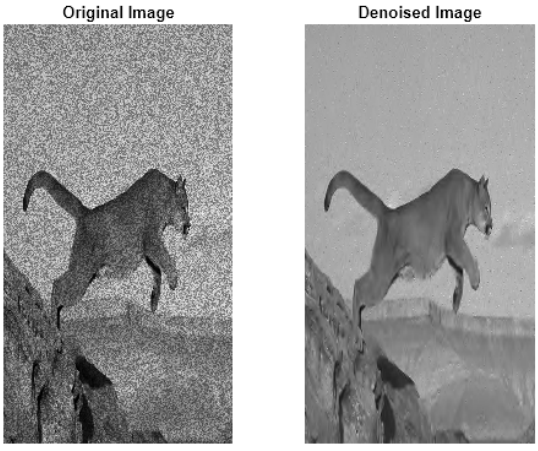
source: https://www.sabrepc.com/blog/Deep-Learning-and-AI/gans-vs-diffusion-models

Background



source: https://www.sabrepc.com/blog/Deep-Learning-and-AI/gans-vs-diffusion-models

Background



source: https://www.mathworks.com/help/wavelet/ug/wavelet-denoising.html

Background

Figure 1: fingerprint estimate with 2 residuals

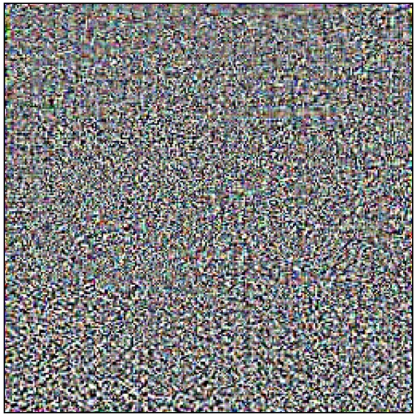


Figure 2: fingerprint estimate with 8 residuals

source: Marra et al, "Do GANs leave artificial fingerprints?", 2018

Background

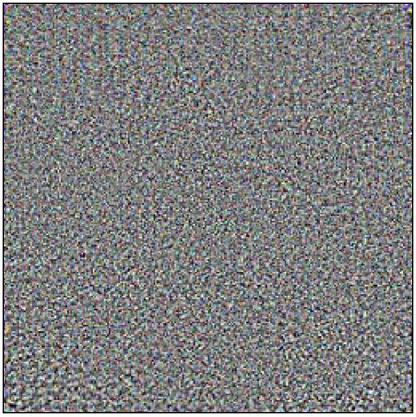


Figure 3: fingerprint estimate with 32 residuals

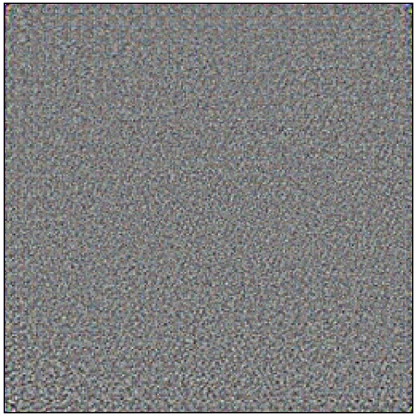


Figure 4: fingerprint estimate with 128 residuals

ɜ: Marra et al, "Do GANs leave artificial fingerprints?", 2018

Background

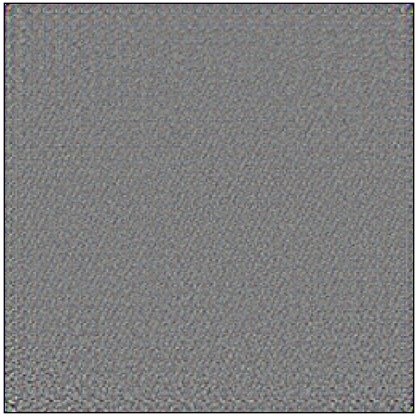


Figure 5: fingerprint estimate with 512 residuals

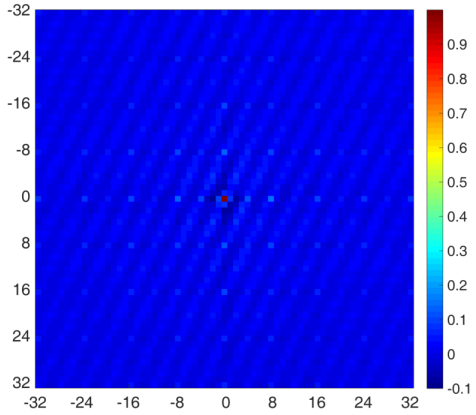


Figure 6: frequency domain

: Marra et al, "Do GANs leave artificial fingerprints?", 2018

Background

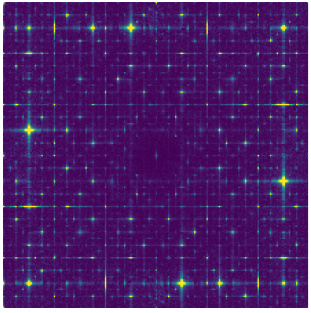


Figure 7: Big Gan

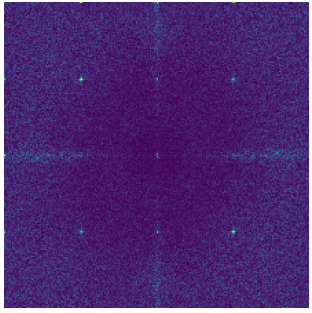


Figure 8: Glide

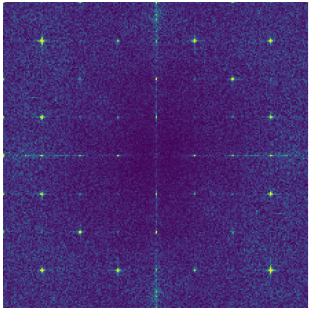


Figure 9: Stable Diffusion

source: Corvi et al, "ON THE DETECTION OF SYNTHETIC IMAGES GENERATED BY DIFFUSION MODELS"

Background

GAN based

- NVIDIA GauGAN/Canvas
- Rosebud.AI
- DeepArt
- StyleGAN
- BigGAN
- ProGAN
- CycleGAN
- StarGAN
- Pix2Pix
- DiscoGAN
- WGAN-GP

Diffusion based

- Stable Diffusion
- Midjourney
- DALL-E 3
- Adobe Firefly
- Runway Gen-2
- Google's Imagen/ImageFX
- Anthropic's Claude 3 Sonnet Vision
- Leonardo.AI
- ControlNet
- Kandinsky
- karlo-diffusion

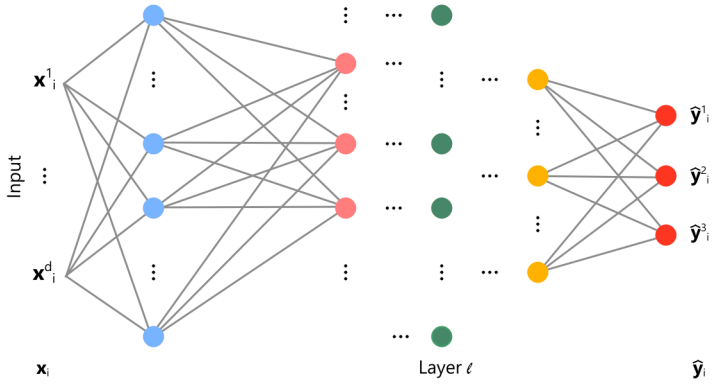
Open Set Problem

- 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

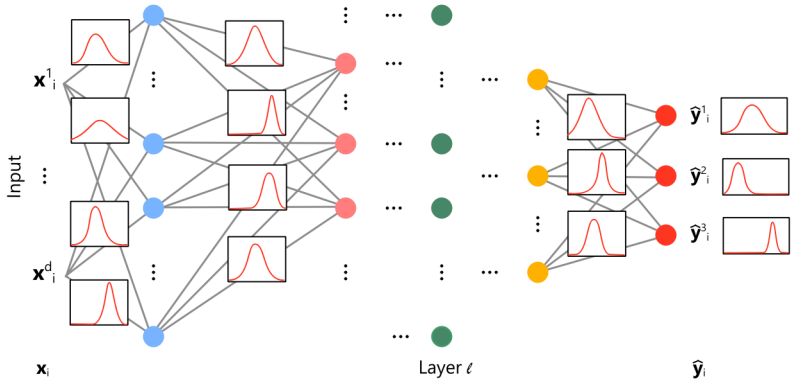
- Bayesian Neural Networks
- key components, distribution over weights, updated according to bayes theorem
- measure uncertainty with variance of weights distribution, distribution over classes
- desired outcome with out-of-distribution data: low certainty
- desired outcome with post-processed images: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

Proposed Approach



source: Magris et al, "Bayesian Learning for Neural Networks: an algorithmic survey", 2022

Proposed Approach



source: Magris et al, "Bayesian Learning for Neural Networks: an algorithmic survey", 2022

Proposed Approach

Update model weights according to Bayes theorem

$$P(w|D) = \frac{P(D|w) \cdot P(w)}{P(D)} \tag{1}$$

Timeline

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