Leveraging generative models to characterize the failure conditions of image classifiers

Adrien LE COZ[1,2], Stéphane HERBIN[2], and Faouzi ADJED[1]

[1]IRT SystemX, Palaiseau, France
[2]DTIS, ONERA, Université Paris Saclay F-91123 Palaiseau - France

Introduction

Introduction

Context:

- Behavior of Machine Learning (ML) models is opaque
- Many applications need guaranteed performance
- Accuracy on test dataset is not enough to characterize a ML model

Big Goal:

- Development of methods to guarantee the ML model behavior

Here:

- Use the power of recent generative models to find the visual attributes that impact a classifier prediction

Approach

Classifier failure conditions

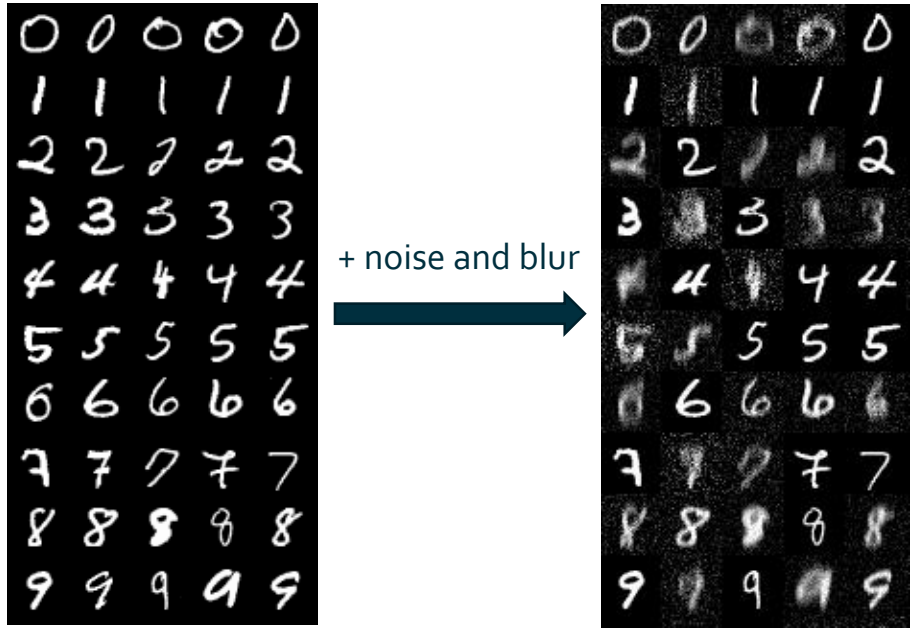
- We want to characterize a classifier on the task: digit classification on **corrupted MNIST** dataset.
- In particular, understand its failure conditions more precisely than looking at global indices.

A classifier trained on clean data is not robust to corruptions

(corruptions from [Hendrycks *et al*. 2019])

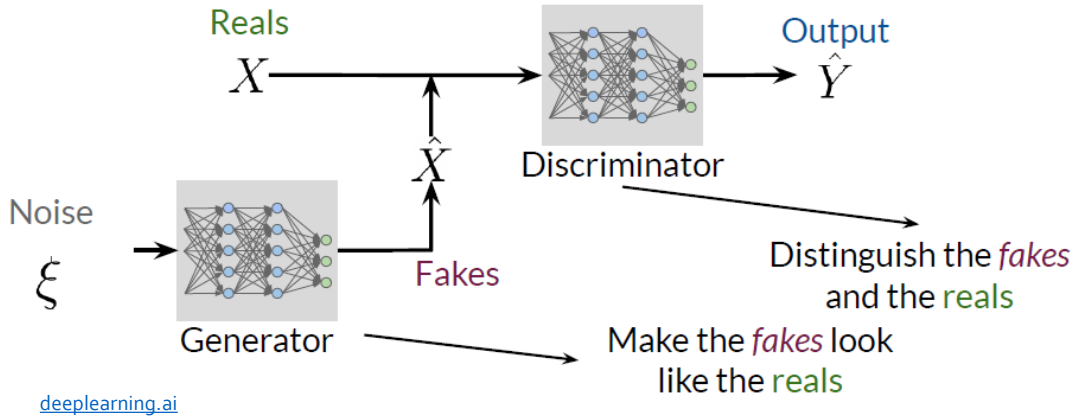
	severity=1	severity=2	severity=3	severity=4	severity=5
uncorrupted	0.9775	0.9775	0.9775	0.9775	0.9775
gaussian_noise	0.9701	0.9588	0.9227	0.8137	0.5975
shot_noise	0.978	0.9772	0.9755	0.9706	0.9695
impulse_noise	0.9693	0.9542	0.9334	0.8097	0.6183
contrast	0.7282	0.4414	0.2401	0.0883	0.0891
jpeg_compression	0.9763	0.9753	0.9757	0.9748	0.9734
speckle_noise	0.977	0.9767	0.9745	0.9747	0.9699
gaussian_blur	0.9737	0.915	0.7841	0.6147	0.3601

We choose gaussian noise and blur to corrupt our dataset



Generative model

- Generative Adversarial Networks (**GANs**): from noise to data

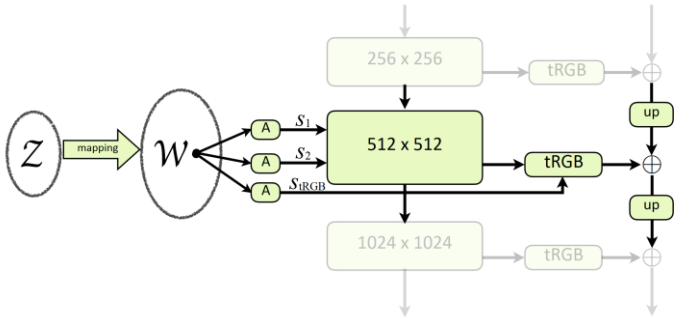


[Karras *et al.* 2018]

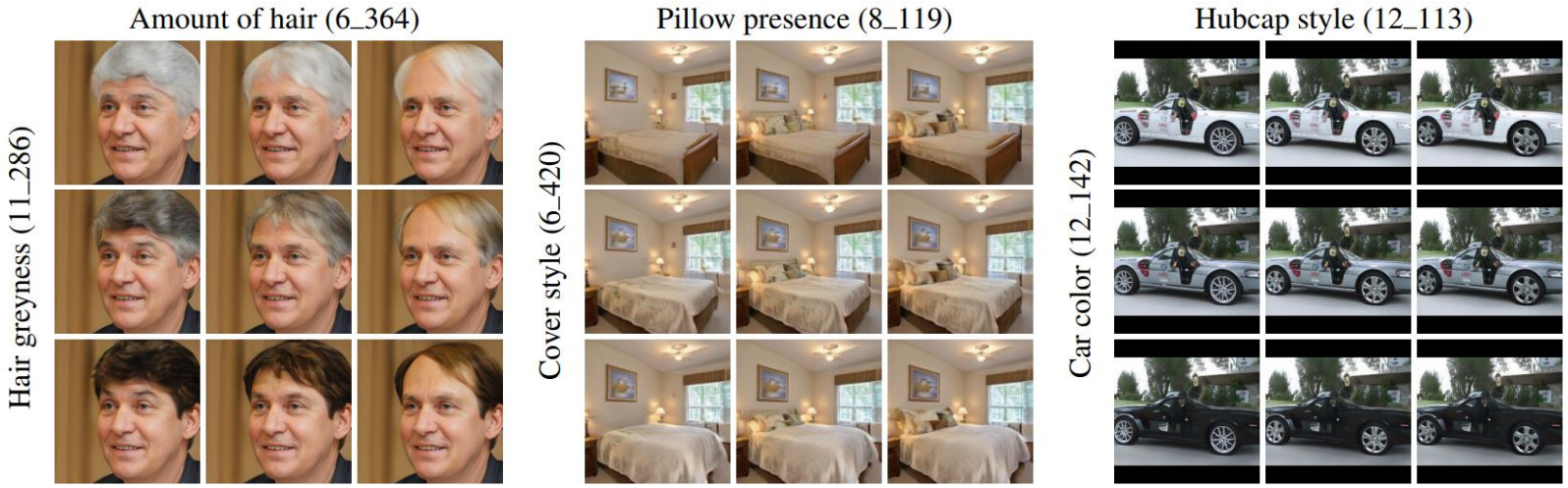
- High-quality generation, rich image representations

Latent space and image editing

- The model **StyleGAN2** [Karras *et al.* 2019] has a **disantangled** latent space, **StyleSpace** [Wu *et al.* 2020]

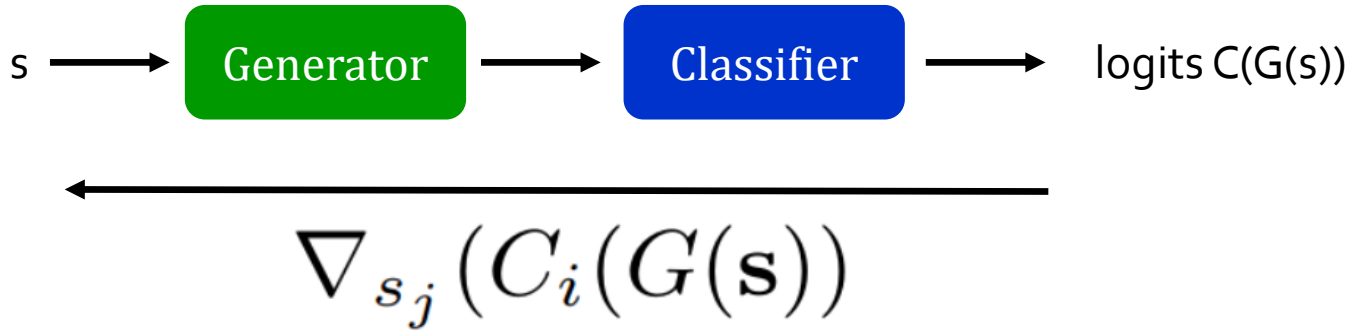


- It allows powerful **image editing**, one attribute at a time



Finding influential dimensions in the latent space

- We find which **dimensions** of StyleSpace have the most impact on predicting the correct class by using the **gradient**



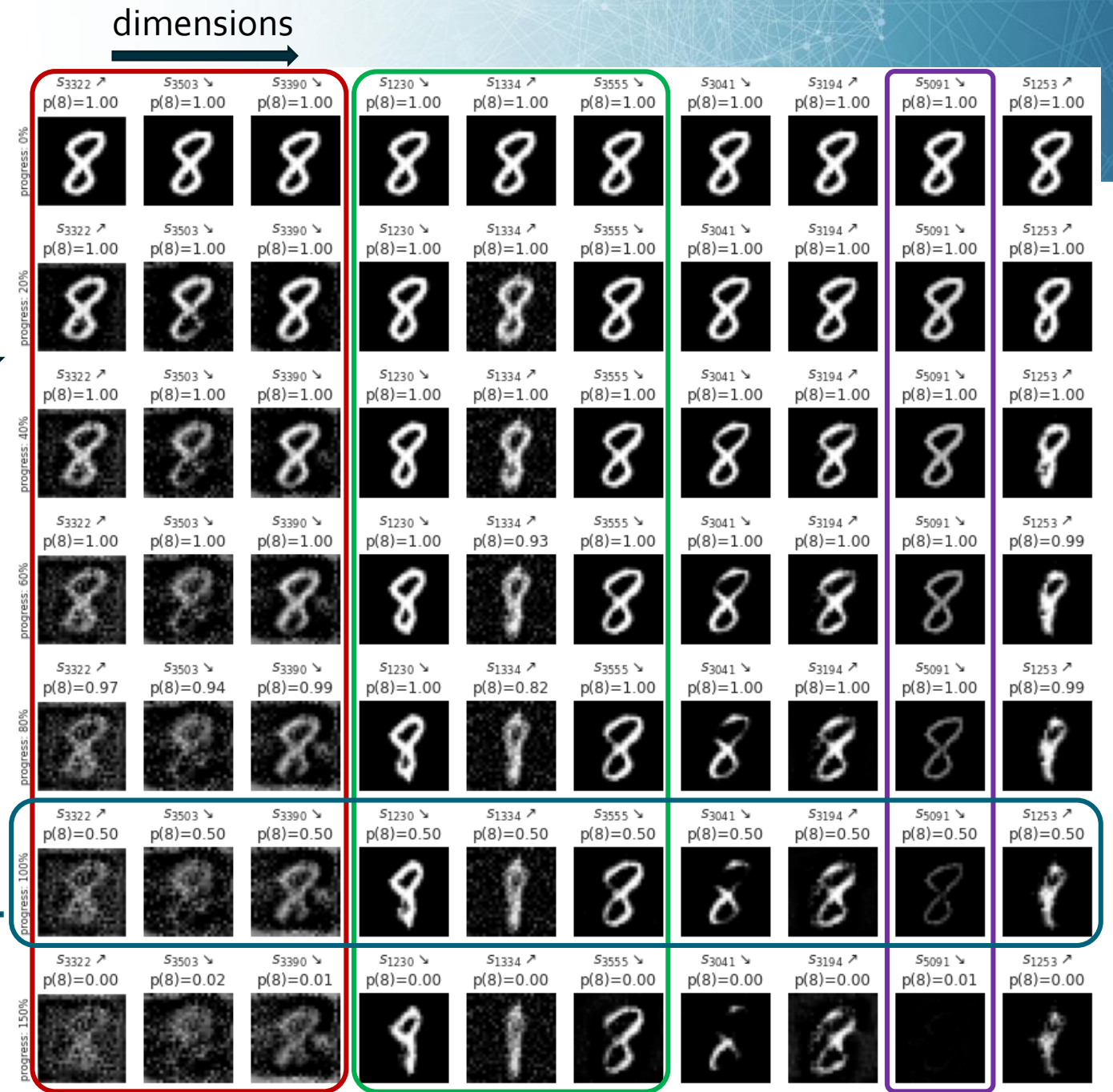
1. Compute average gradients for many inputs sampled from the StyleSpace
2. Select dimensions with the highest average gradients values
3. Change values of selected dimensions and recognize the associated visual attributes
4. Identify corner cases

Two applications

Degrade accuracy

We identify the main dimensions of degradation (fast accuracy decrease), and:

- Progressively shift values of these dimensions, e.g. s_{3322} ↗
- Compute classifier output probability $p(8)$
- Interpret the associated visual attributes: noise, shape, contrast…



Corner cases
(data point where classifier prediction gets erroneous)

Corner cases

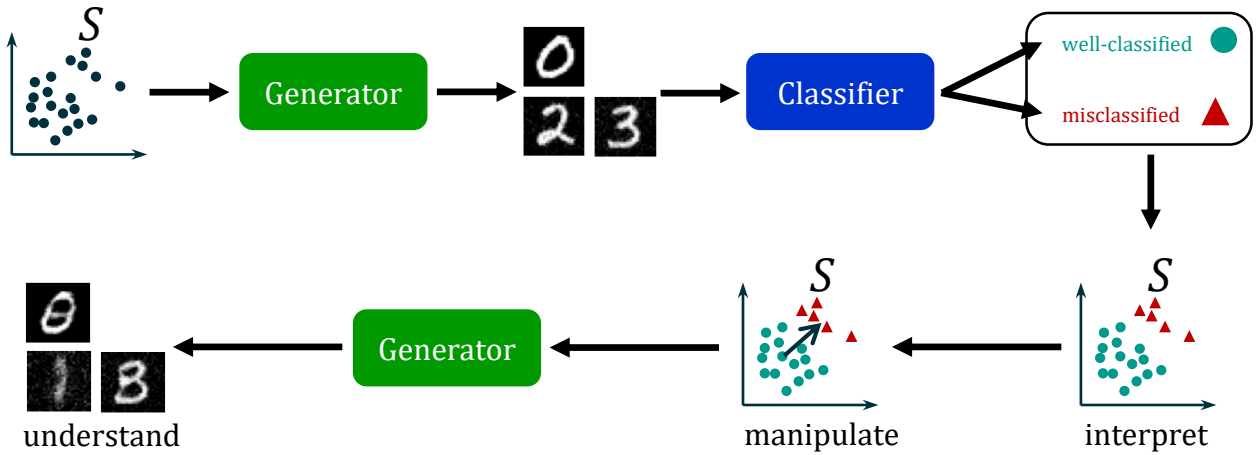
Let's see more examples of corner cases:

- The main dimensions of degradation are computed separately for each class
- Classes are not degraded the same way (e.g. 0 vs. 9)
- *Unlimited* generation of difficult data becomes possible

Conclusion

Conclusion

- We identified the visual attributes which deteriorate the classifier performances and discover corner cases.
- It helps better understand classifier performance.



- Perspectives:

- Scale up. Recent works extend StyleGAN to less structured and more diverse images.
- Incorporate classifier during generator training.
- Define an intelligible operational domain with guaranteed classification performance.

References

- [Hendrycks *et al*. 2019]
  D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL: https://arxiv.org/abs/1903.12261. doi:10.48550/ARXIV.1903.12261.
- [Karras *et al*. 2018]
  T. Karras, S. Laine, T. Aila, A style-based generator
  architecture for generative adversarial networks, 2018. URL: https://arxiv.org/abs/1812.04948.
  doi:10.48550/ARXIV.1812.04948.
- [Karras *et al*. 2019]
  T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, 2019. URL: https://arxiv.org/abs/1912.04958. doi:10.48550/ARXIV.1912.04958.
- [Wu *et al*. 2020]
  Z. Wu, D. Lischinski, E. Shechtman, Stylespace analysis: Disentangled controls for stylegan image generation, 2020. URL: https://arxiv.org/abs/2011.12799.doi:10.48550/ARXIV.2011.12799.