Leveraging generative models to characterize the failure conditions of image classifiers

Adrien LE COZ\textsuperscript{1,2}, Stéphane HERBIN\textsuperscript{2}, and Faouzi ADJED\textsuperscript{1}

\textsuperscript{1}IRT SystemX, Palaiseau, France
\textsuperscript{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
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])

We choose gaussian noise and blur to corrupt our dataset
Generative Adversarial Networks (GANs): from noise to data

- High-quality generation, rich image representations

[Karras et al. 2018]
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
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)
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]

- [Karras et al. 2018]

- [Karras et al. 2019]

- [Wu et al. 2020]