Explaining an image classifier with a generative model conditioned by uncertainty

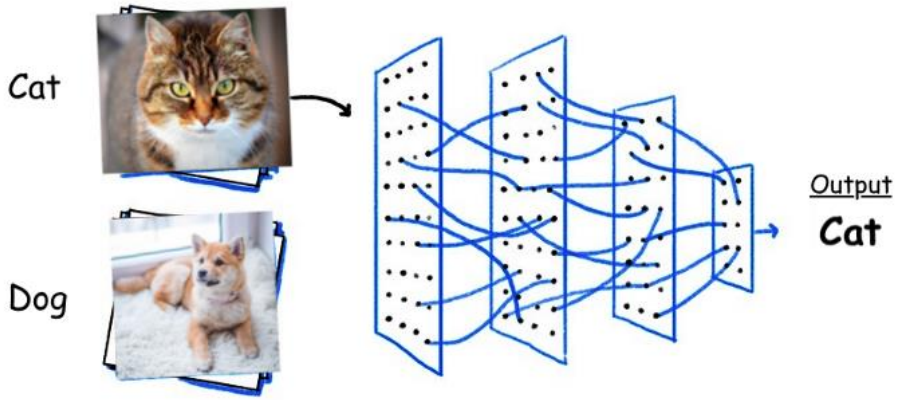
Adrien Le Coz[1,2], Stéphane Herbin[2], Faouzi Adjed[1]

[1]IRT SystemX, Palaiseau, France

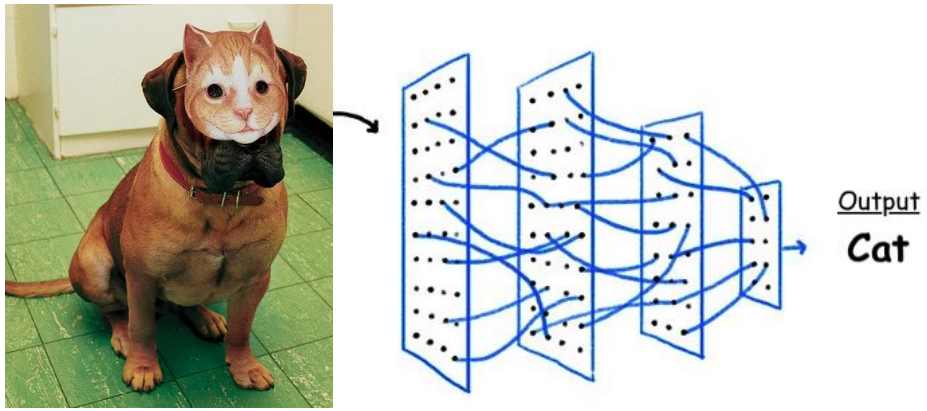
[2]DTIS, ONERA, Université Paris Saclay F-91123 Palaiseau - France

Context

Image classifiers work well most of the time…

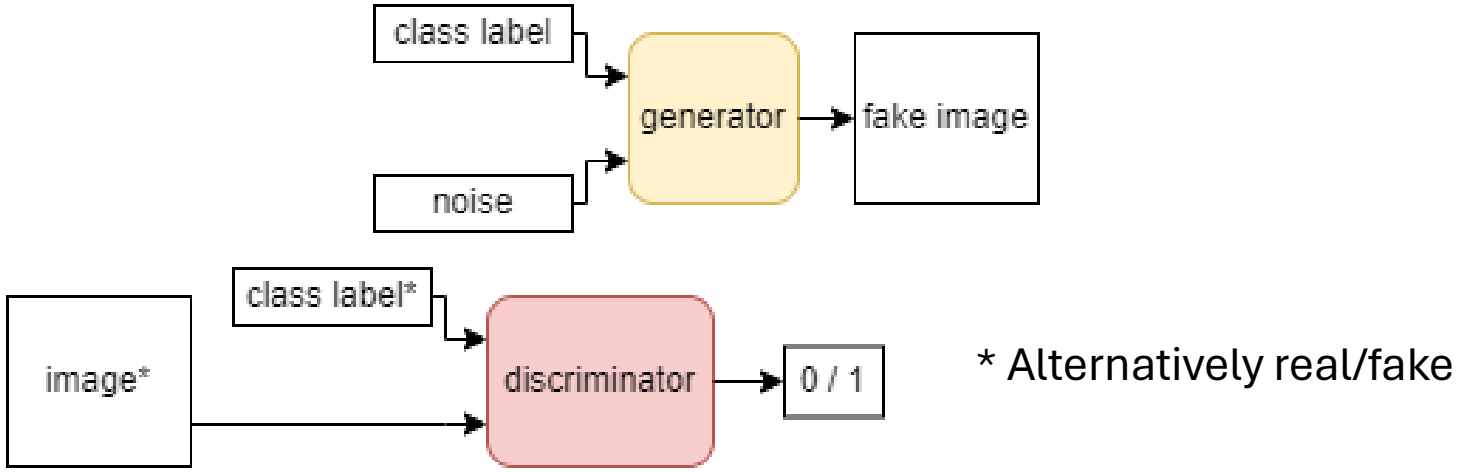


… but not always!



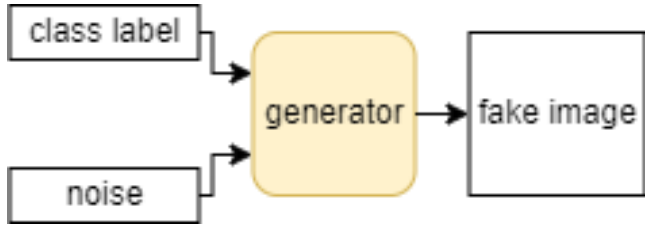
What makes them fail?

Generative adversarial networks



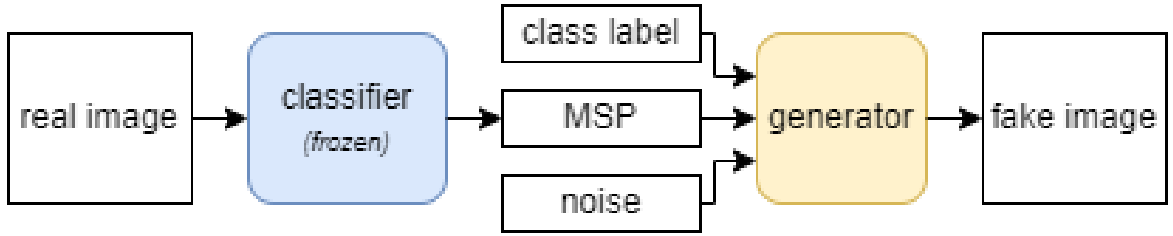
* Alternatively real/fake

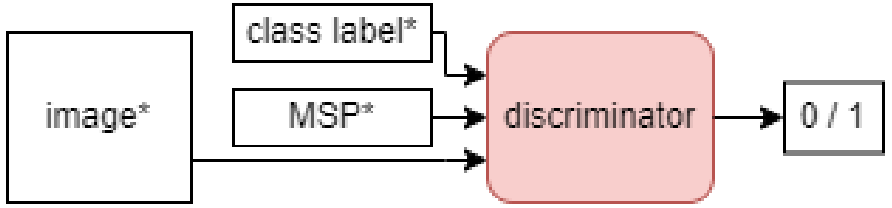
Can be used to generate data of a chosen class (vary label input)



Conditioning with uncertainty

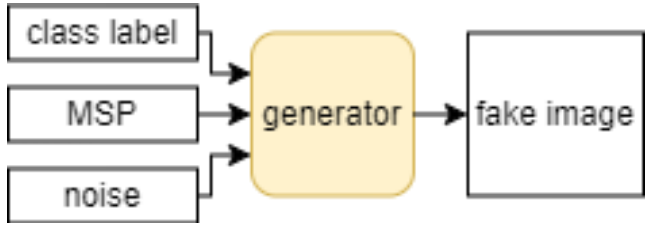
Here, Maximum Softmax Probability (MSP) = classifier confidence





* Alternatively real/fake

Can be used to generate uncertain images (vary MSP input)

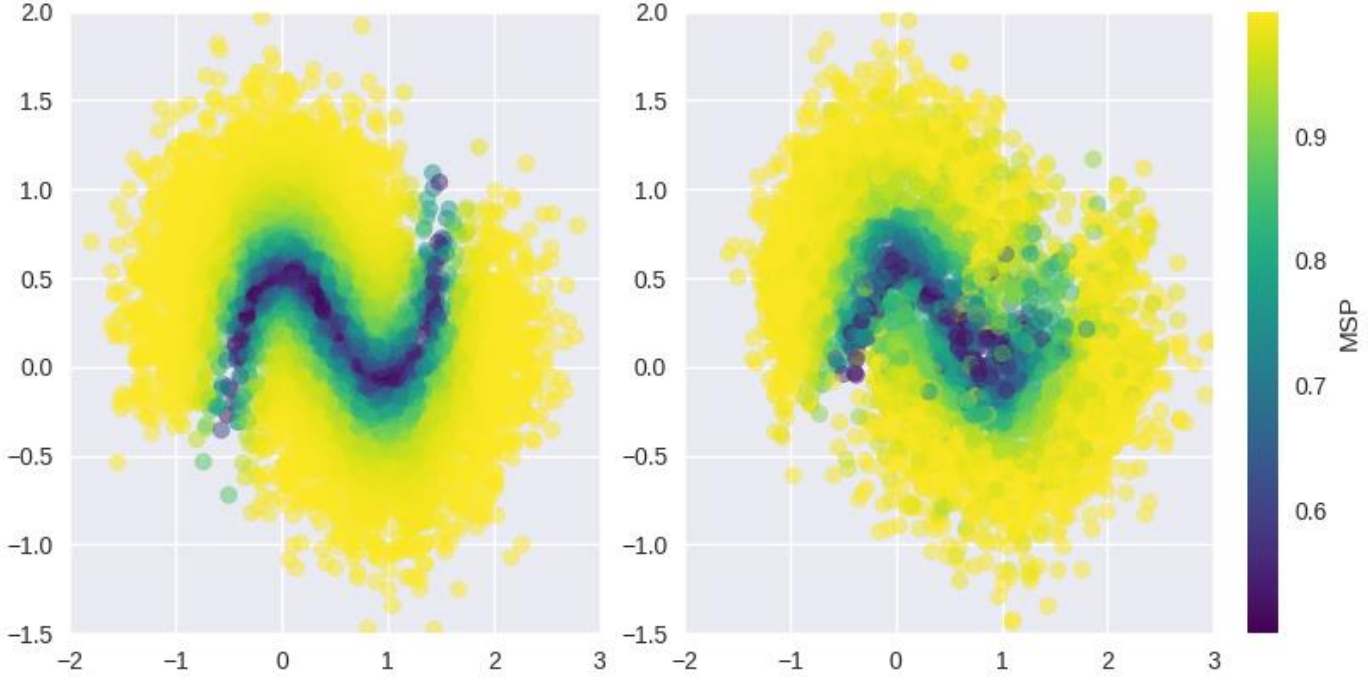


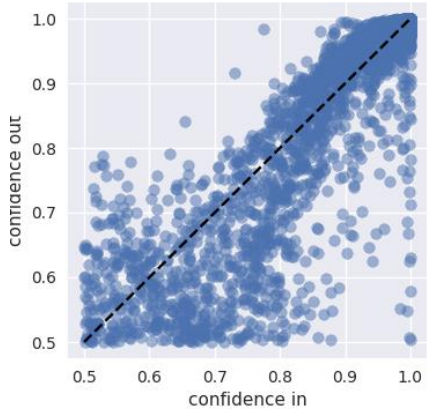
Results

Toy data

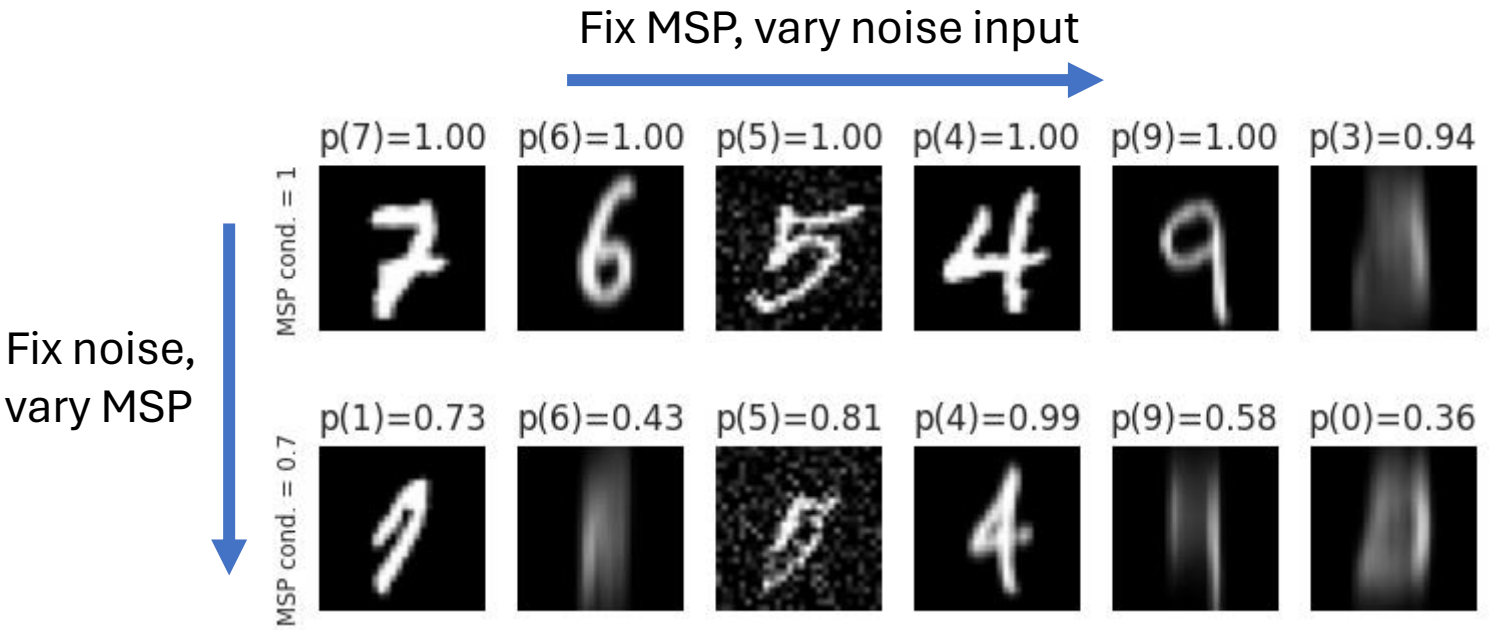
Real data → MSP

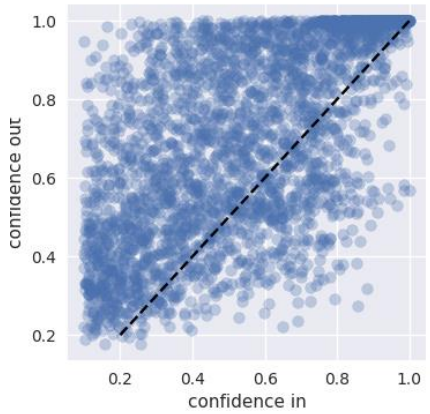
MSP_{cond} → fake data





Corrupted MNIST





Conclusion

- GAN conditioned by uncertainty can:
  - Corrupt images (fix noise, vary confidence)
  - Generate uncertain images (fix confidence, vary noise)
- Identification of uncertainty sources is manual
- Still limited results for images

Sources of images:

https://towardsdatascience.com/10-papers-you-should-read-to-understand-image-classification-in-the-deep-learning-era-4b9d792f45a7

https://www.reddit.com/r/photoshopbattles/comments/3hpcvq/psbattle_dog_with_cat_mask/