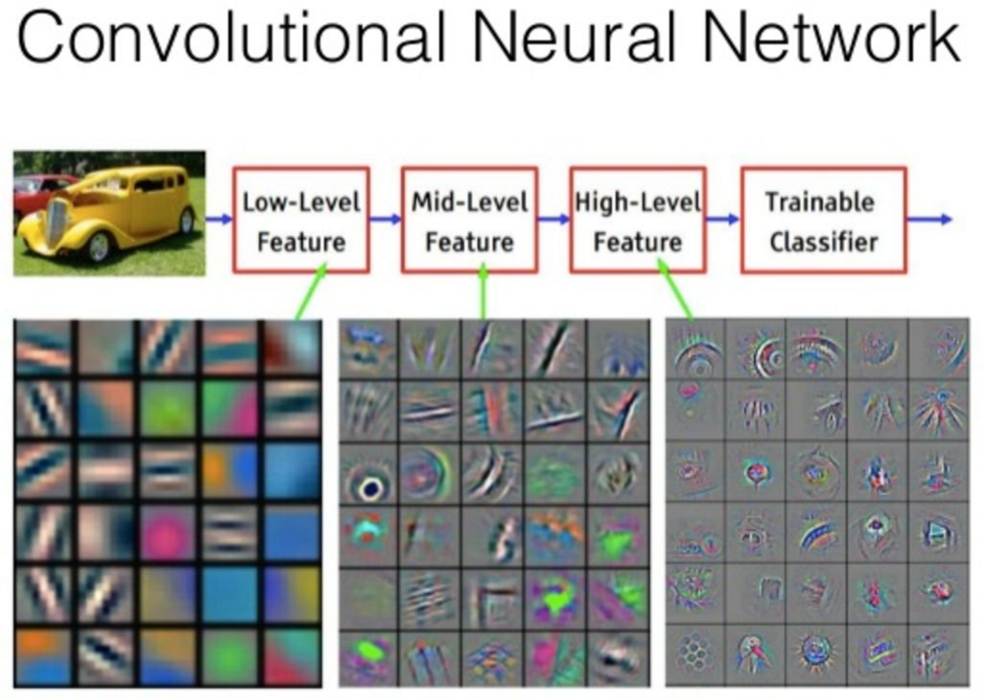
Adversarial Examples: a Generalization Failure?

Gauthier Gidel,
Mila, UdeM, CCAI Chair

Chapter 1: The Dream

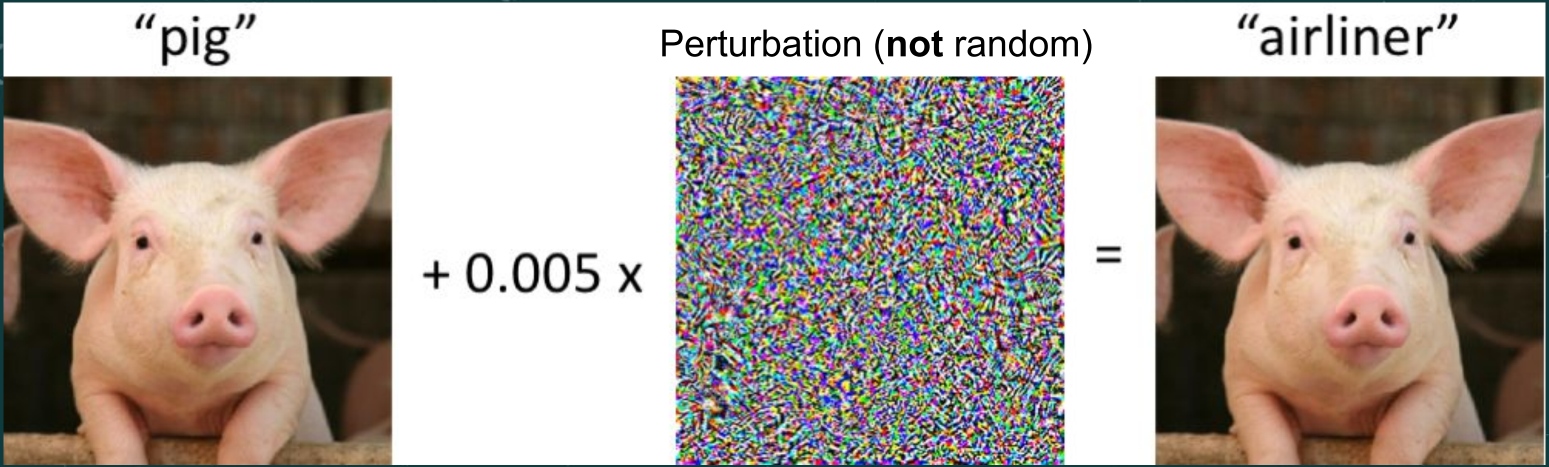


A Yellow car

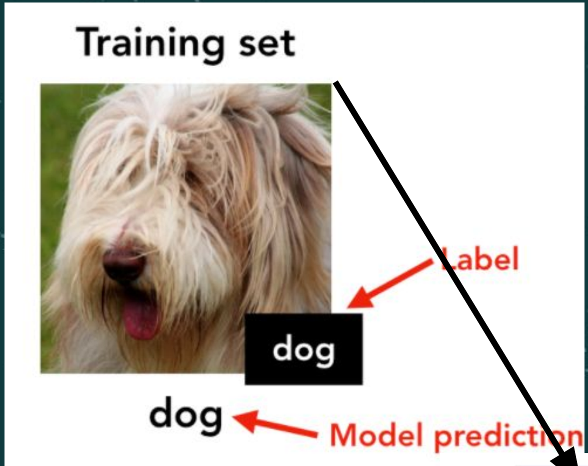
Source: https://medium.com/abraia and [Zeiler & Fergus 2013]

Chapter 2: The Bug

“Deep Learning can make pigs fly’’

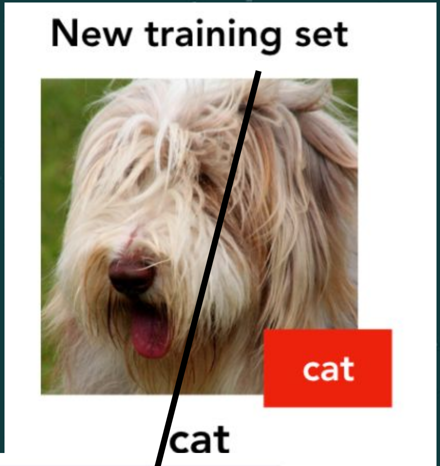


[szegedy et al. 2013]



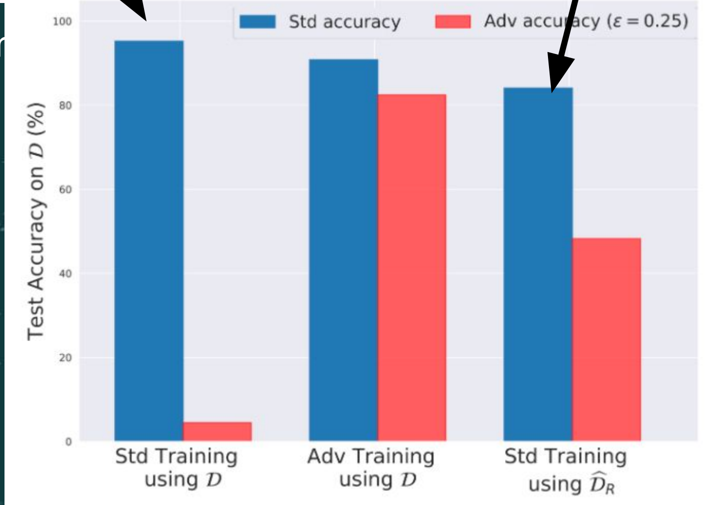
source:https://gradien

Compute adv. Example



Train a Classif on that **bad train set**

??????

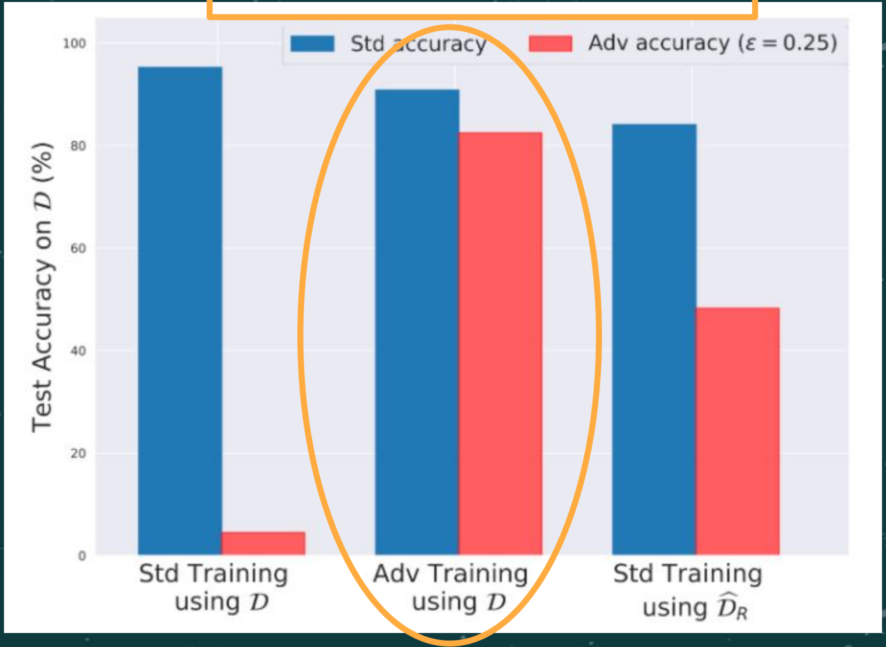


Normal Test Set !!!

Chapter 3: The feature ???

[Ilya et al. 2019]

- These features are **useful** quantities for the prediction task.
- They **GENERALIZE** (in the sense of supervised learning)
- Adversarial Training remove these features.



If we remove these features

Chapter 4: the diagnosis

- Adversarial examples always exists
  [Bubeck, Cherapanamjeri, Gidel, Tachet des Combes 2021] [Daniely and Schacham 2020]
- Adversarial examples can be used for the **in-distribution Task.** [Ilyas et al 2019]
- **My Opinion:** there few hope that these feature will help for OoD generalization. (will learn them with standard supervised learning)
- Something is broken in standard supervised learning. (Adversarial examples are the symptom of that)
- First step of **OoD Generalization**: Robust models generalize to distributions “close” to the data distribution.

Conclusion

- In-distribution Generalization is a somewhat broken task. [Recht et al. 2018]
- Robustness (to adv examples) cannot help to that ‘too easy task’
- Robustness can help is more challenging task (sub-population shift) [Santurkar et al. 2021] (to be proved that it can help for OoD in a broader sense)