CNN face à l'adversité autour d'un cas d'étude

J.E Campagne (IJCLab/Orsay)

Action Dark Energy 2020

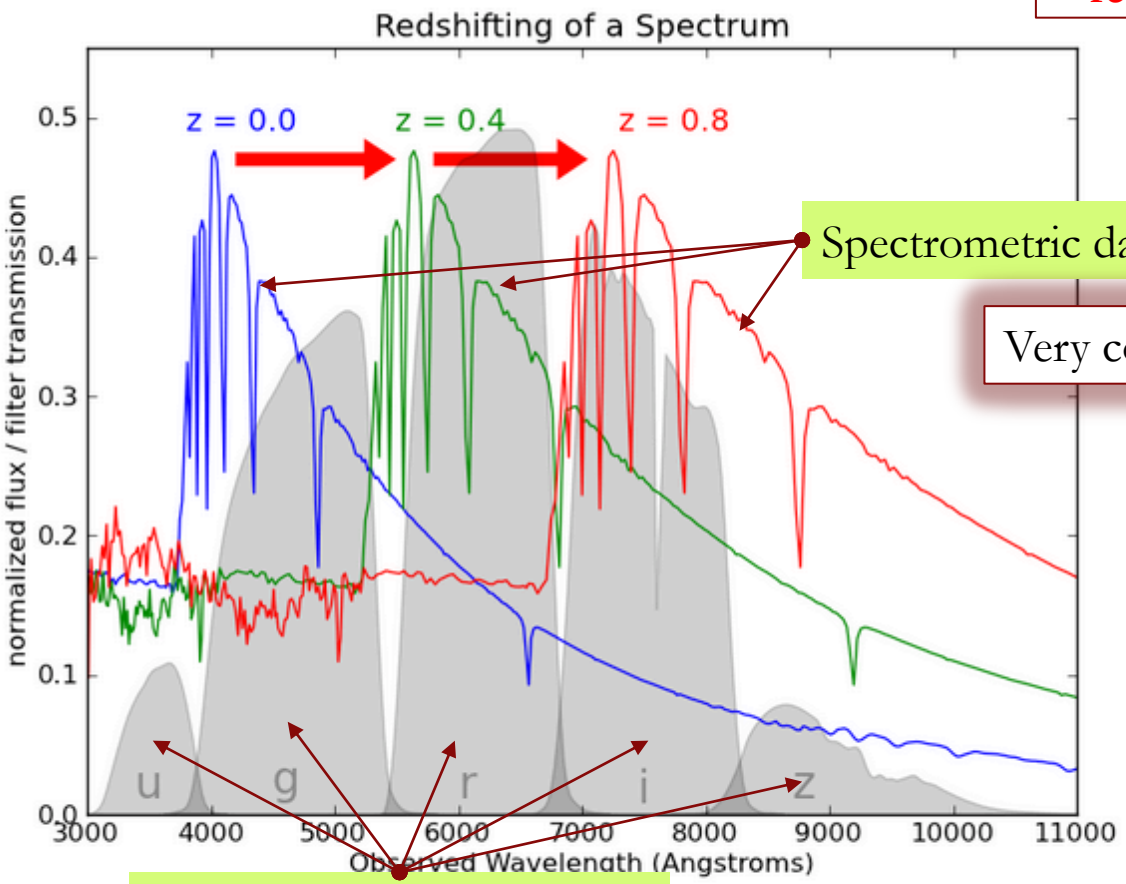
Outline

- Photometric redshift determination : the use case
- Case of *Inception* model first developed by J. Pasquet et al.
- Adversarial Samples
- Some results towards robustness
- Summary/Outlooks

Photo-z: photometric redshift

$\lambda_{rec} = \lambda_{src}(1+z)$

the redshift



Spectrometric data

Very costly

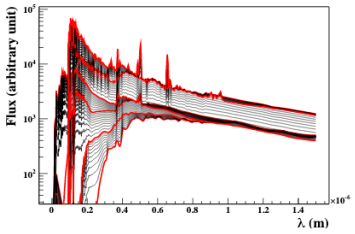
Efficiency curves of filters

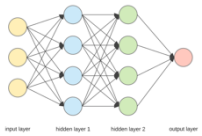
The spectrum of the star Vega (α-Lyr) at three different redshifts. The SDSS ugriz filters are shown in gray for reference.

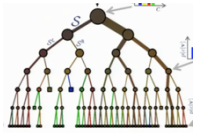
Vega 25ly, z=0

Methods for Photo-z

Since the pioneering work in the 60's, several methods have been developed to estimate the redshift from the multi-bands photometric measurements, basically:







Possible combination

- *Template* fitting
  - eg. uses the SEDs of known galaxies and a fitting method
- *Feature* based Machine Learning
  - Uses a certain number of precooked features extracted from the measurements and feed to an engine as k-NN, NN/MLP, Decision Tree, BDT or Random Forest

The use-case

- *Image* based Deep Learning

Nb. Absolutely non exhaustive list of methods.

« *Inception* » for photo-z

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26 arXiv:1806.06607v2

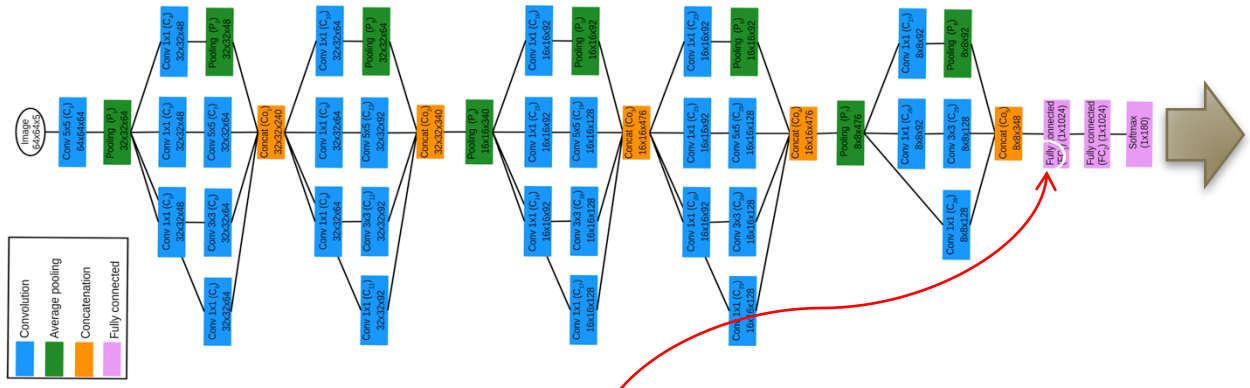
Inspired from GoogleLeNet with multi-levels of conv-layers (Szegedy et al. 2014)

27.5M parameters

~30Layers

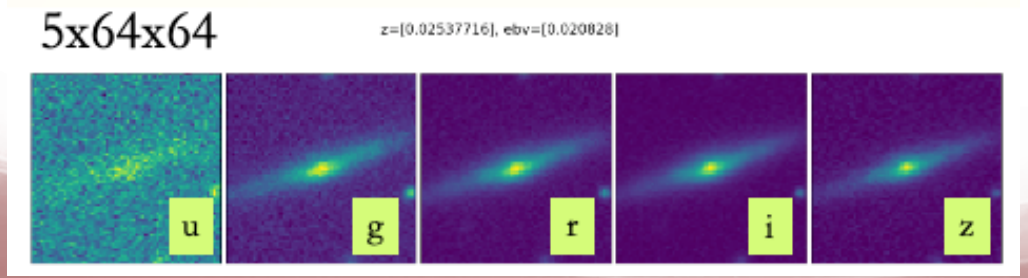
Convolution (6.5% param.)

FC (93.5% param.)

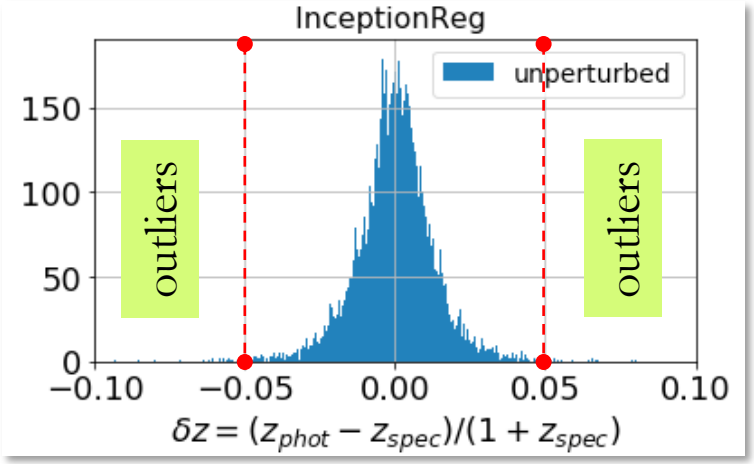


Output Photo-z

SDSS DR12 images + redenning

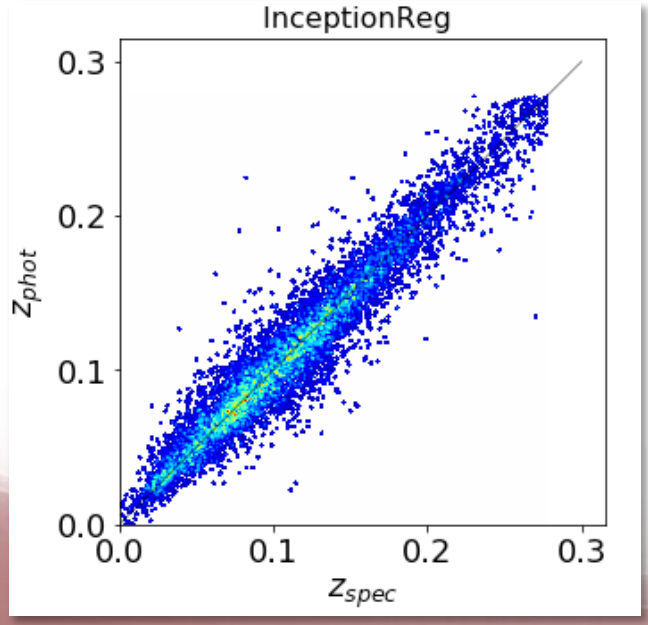


Results from my *Inception*



Training/Test samples/plot :100k/100k/10k
from a total of ~ 600k input dataset

Output: z-photo (regression, MSE)*



Bias ×10−4	σmad ×10−3	η (%)
+1.98	10.31	0.47

Results in agreement with J. Pasquet et al.

(*: arXiv:2002.10154 I was using the output of Pasquet et al.)

Common work

Teams spend some times to:

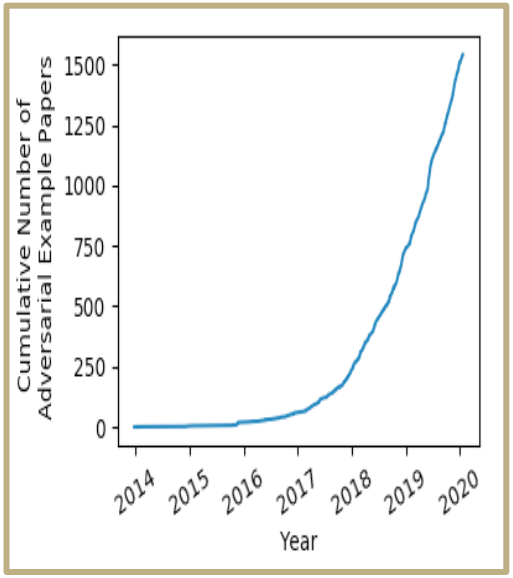
- Elaborate ML/DL architectures
- Apply some ML paradigm to tune the hyper-parameters using for instance: the triptych Training/Testing/Validation sets, Under/Over fitting aspects
- Compare their results against "State-of-the-art" competitors
- Perform systematics studies* on the Input Data: eg. are they representative of the use-case, what about their quality…

But, haven't we forgotten something ?

(*: The study of J. Pasquet et al. is very detailed)

Adversarial samples: brief history

- After "**AlexNet**" the winner of ImageNet competition **2012**
- Topic rising since Szegedy et al. (**2013**): "**Intriguing properties of neural networks**"
- 1st explanation Goodfellow et al. (**2014**) : "**Explaining and Harnessing Adversarial Examples**"
- Part of the NIPS ' **2017** Competition
- Kurakin et al @ ICRL **2017**: "**Adversarial Machine Learning at Scale**"
- Ilyas et al (**2019**): "**Adversarial Examples Are Not Bugs, They Are Features**"
- Madry et al (**2017-19**) @ ICLR 18: "**Towards Deep Learning Models Resistant to Adversarial Attacks**"
- **… Towards a deeper understanding of what is going on and how to overtake this intrinsic problem.**



NSIP: Neural Information Processing Systems
ICLR: International Conference on Learning Representations

Empirical risk/adversarial sample

$\{x_i, z_i\}_{i \leq N} \in D_{train}$ Eg. x_i: images, z_i: spectro-z

Classical Empirical risk

$$R_e(f_\theta, D_{train}) = \frac{1}{|D_{train}|} \sum_{(x,z)\sim D_{train}} \ell(f_\theta(x), z)$$

Adversarial Empirical risk

Adversarial perturbation

$$R_{adv,e}(f_\theta, D_{train}) = \frac{1}{|D_{train}|} \sum_{(x,z)\sim D_{train}} \left[\max_{\|\delta\| \leq \varepsilon} \ell(f_\theta(x+\delta), z) \right]$$

1) Min-max/saddle point problem: no general solution in non-convex problem
2) Which norm $|\delta|$, which value of ε ?

Simple perturbation mechanism

Goodfellow et al. 2014

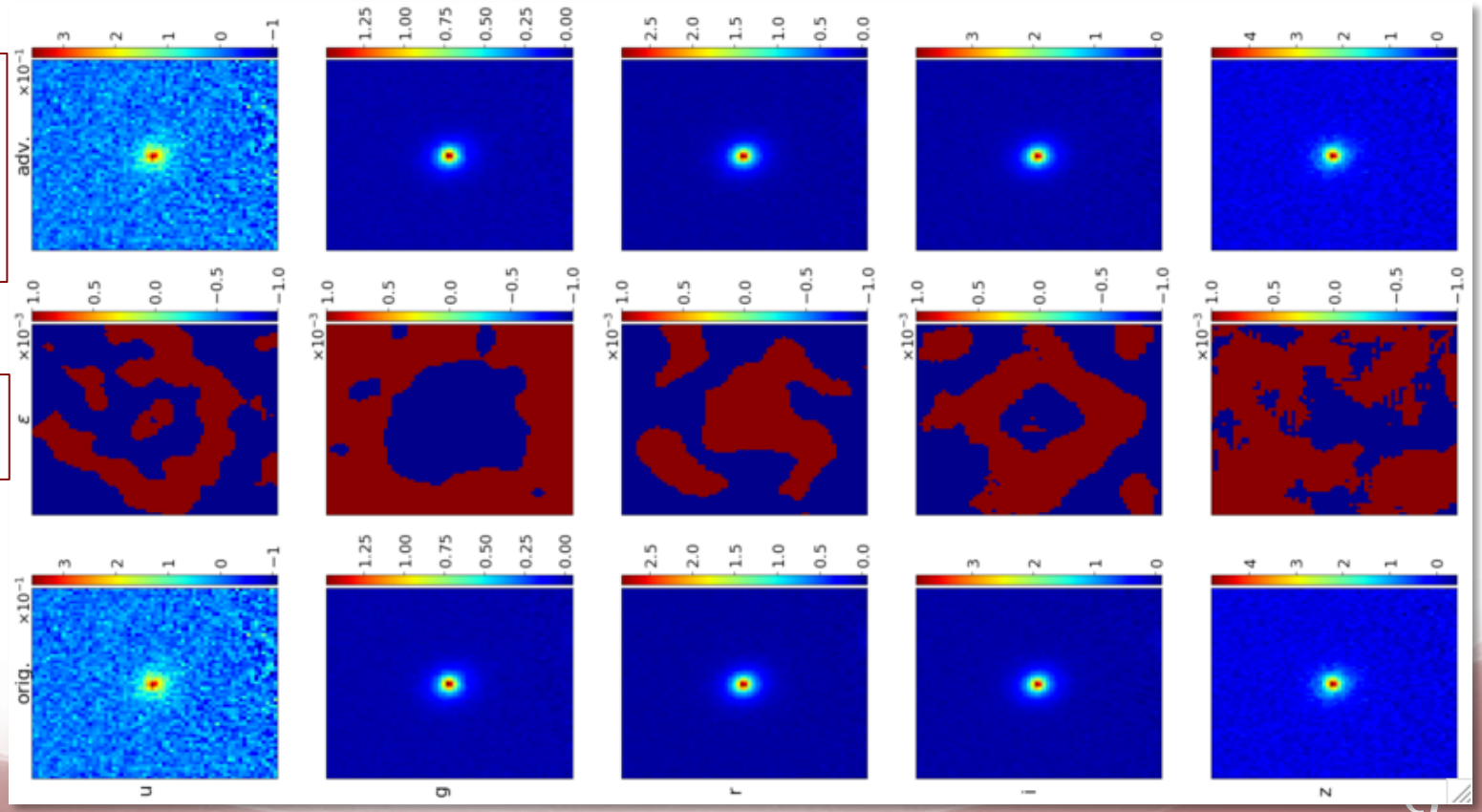
Fast Sign Gradient Method

$$f_\theta \text{ "linear"} + \|\delta\|_\infty \leq \varepsilon \Rightarrow \delta^*(x) = \varepsilon \times \text{sign}\left(\nabla_\delta \ell(f_\theta(x+\delta), z)\right)$$

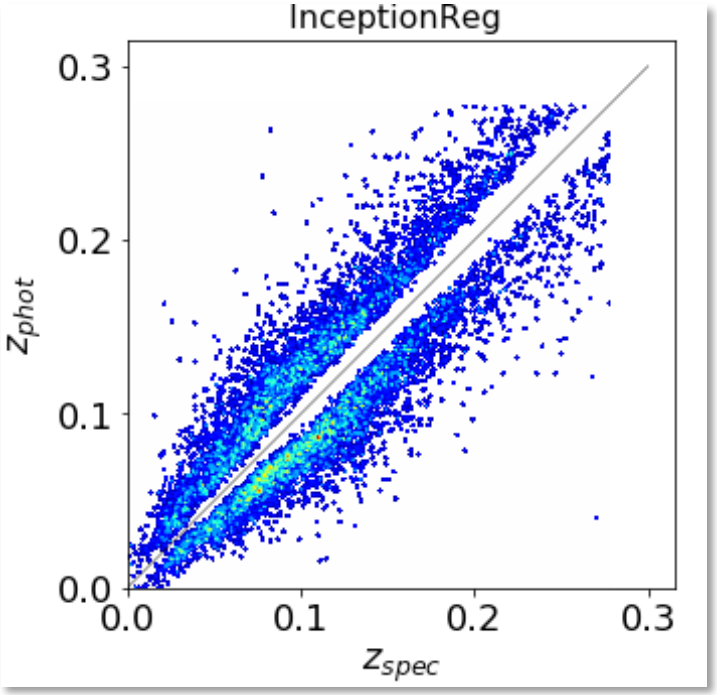
x
z-phot: 0.10
z-spec: 0.07

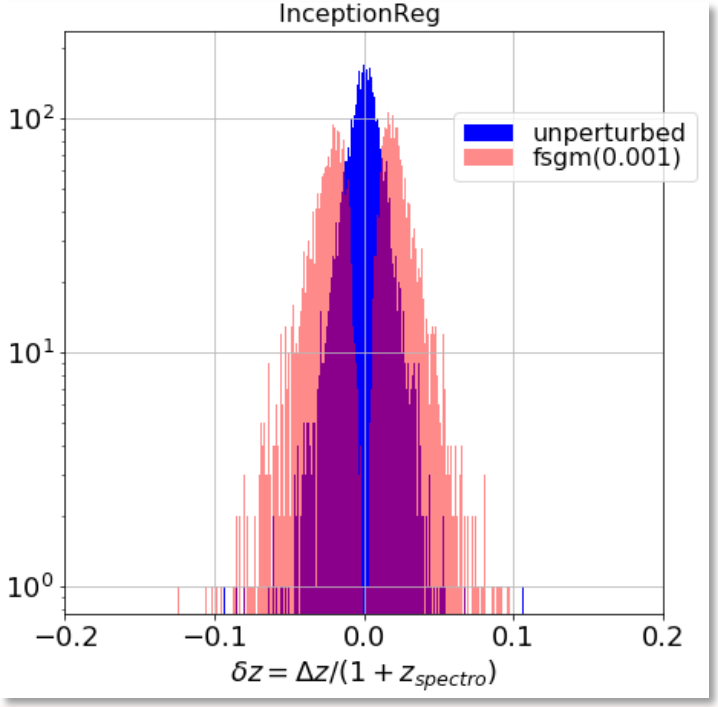
δ^*, $\varepsilon=10^{-3}$

x+δ^*
z-phot: **0.13**



FSGM perturbations impact





Images	Bias ×10−4	σmad ×10−3	η (%)
Perturbed	*+66.40*	33.63	5.31
Non-perturbed	+1.98	10.31	0.47

(nb. arXiv:2002.10154 I was using $\varepsilon=10^{-2}$ which produce of course even more dramatic effects, but perturbation mode visible)

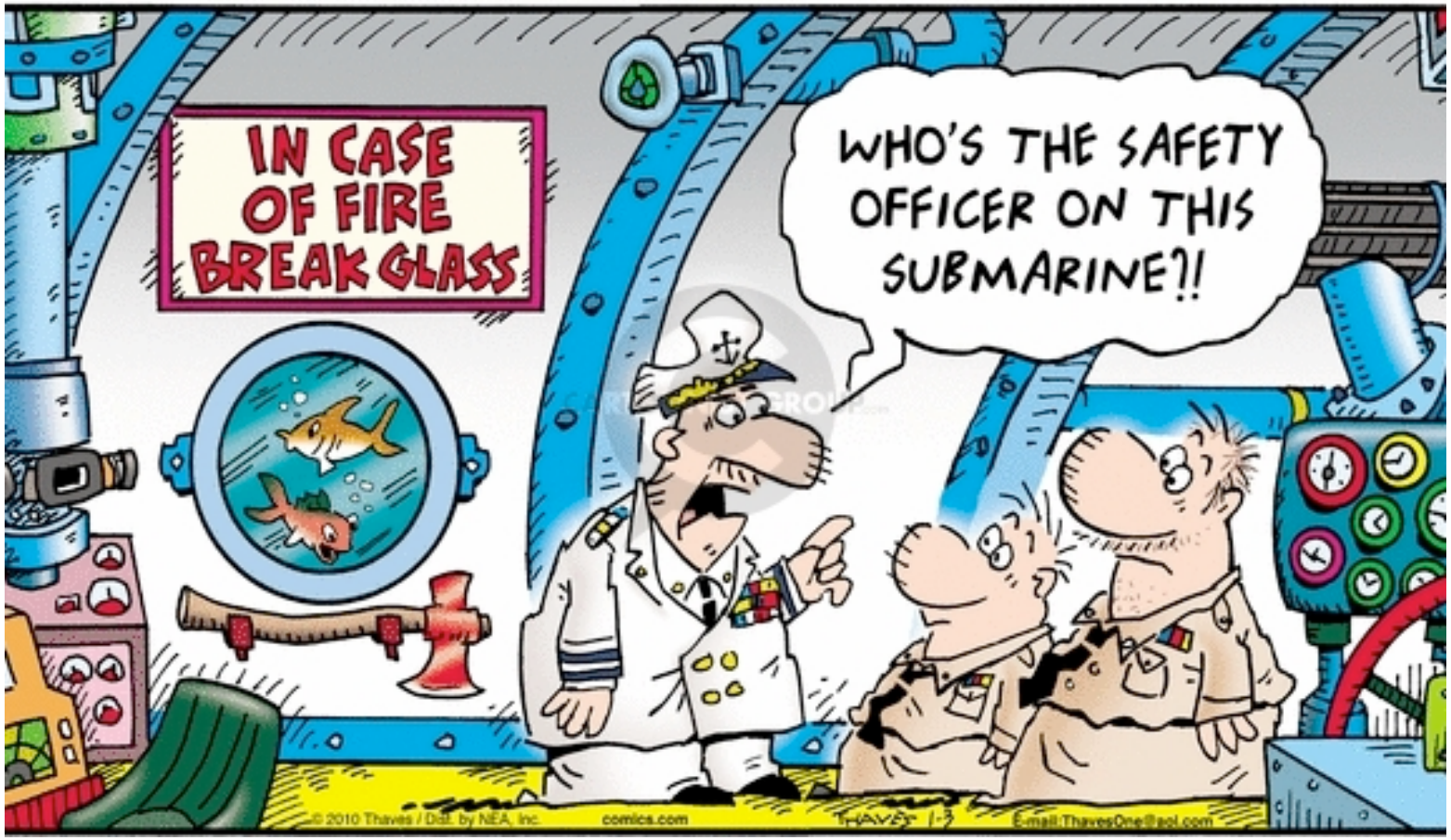
Some other results

- If one train 5 indepedant *Inception* models, and also a simpler CNN model, with each times a different set of training samples:
  - With ***unperturbed*** images: same results for each model (good), one can combine them.
  - With ***perturbed images*** build with one *Inception* model,
    - also perturbes the other *Inception* models: ie. **Combining different models cannot solve the problems**
    - Also **perturbes different architecture models**
- A perturbation δ^* is in principle directly linked to the original x unperturbed image, but it turns out that a single perturbation can impact a large number of images.
- One can use different ways to produce adversarial samples, I use here the FSGM method for the seak of clarity and also that one should first takle such simple perturbation.

What to do?

- A1: What's the problem ?
- A2: Bury one's head in the sand…
- A3: « These kind of perturbations will never append ! »: are you sure ? It is true that trying to generate the perturbations with « real artefact » is not so easy, at least faint objects can mimic small perturbations but not as efficient as the FSGM ones. The question is still open.
- Take it seriously as a sign of a certain (intrinsic) weakness:
  - Training ?
  - Architecture ?
  - Both ?

Countermeasures ?



Countermeasures ?

What about the training?

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{|B_{train}|} \sum_{(x,z)\sim B_{train}} \nabla_\theta \left[\max_{\|\delta\|\leq\varepsilon} \ell(f_\theta(x+\delta), z) \right]_{\theta=\theta_t}$$

1) Solution not known in the general case.

$$\nabla_\theta \ \ell(f_\theta(x+\delta^*), z)$$

$$\delta^* = \max_{\|\delta\|\leq\varepsilon} \ell(f_\theta(x+\delta), z)$$

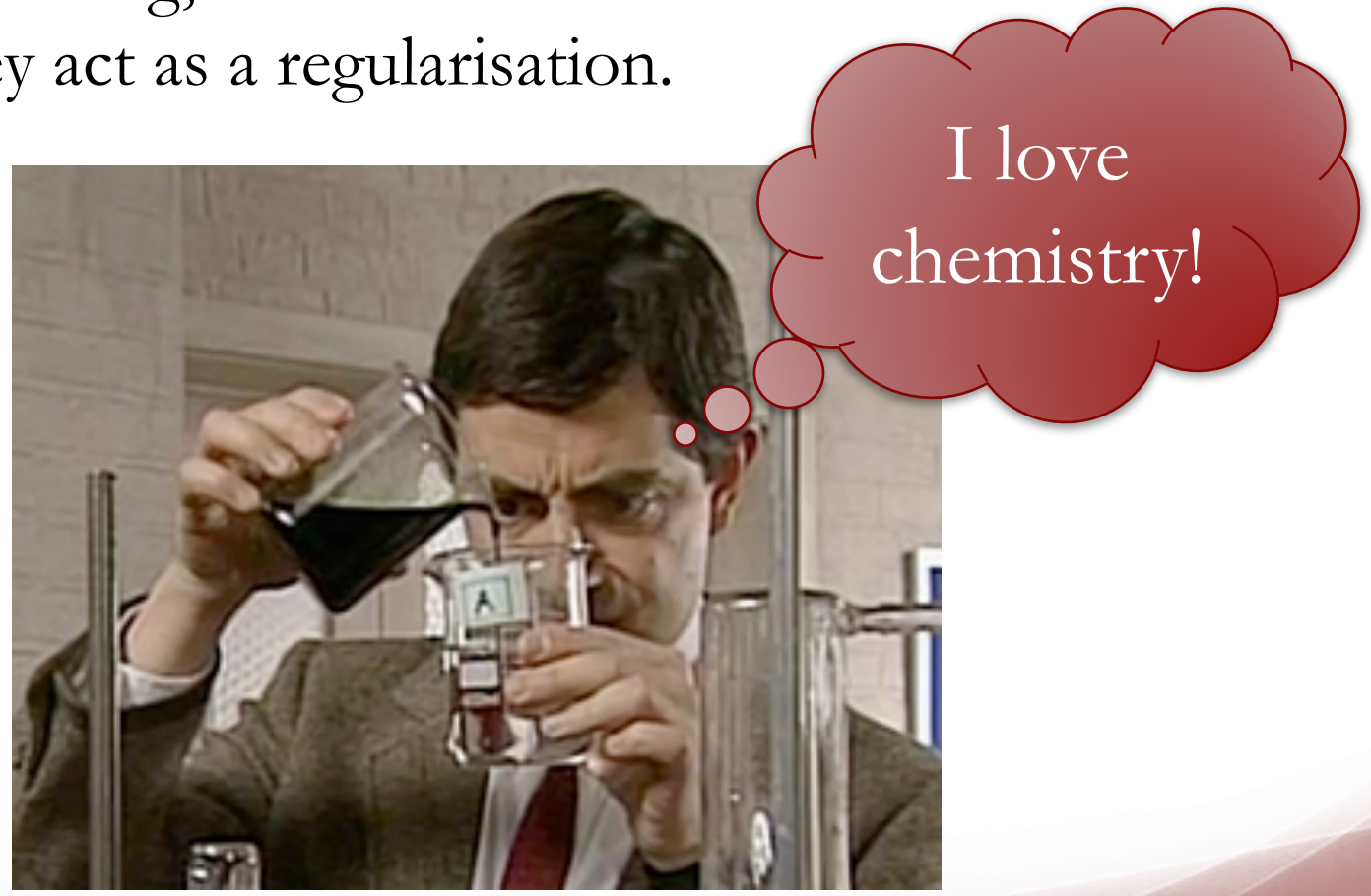
J. Danskin 1966
Convex case

2) Mix up normal images & adversarial ones acts as regularisation terms.

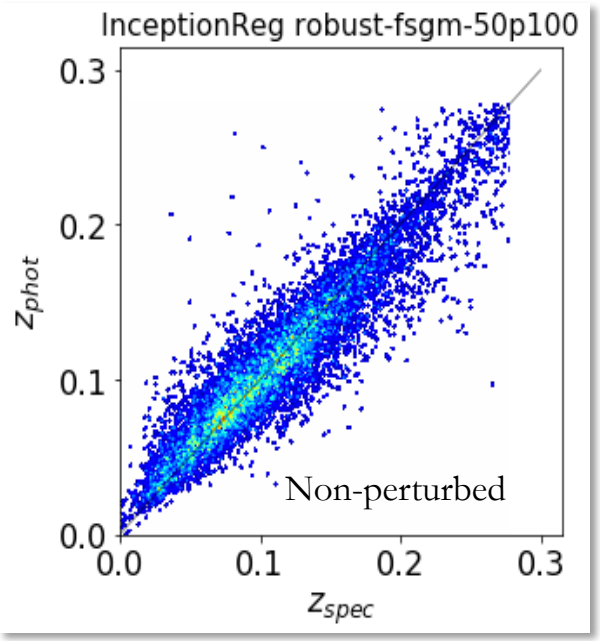
Finlay et al. 2018;
Bietti et al. 2018

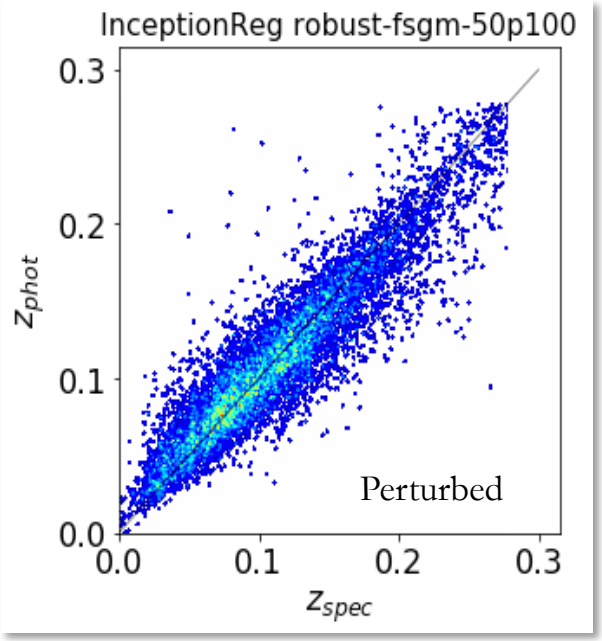
Adversarial training

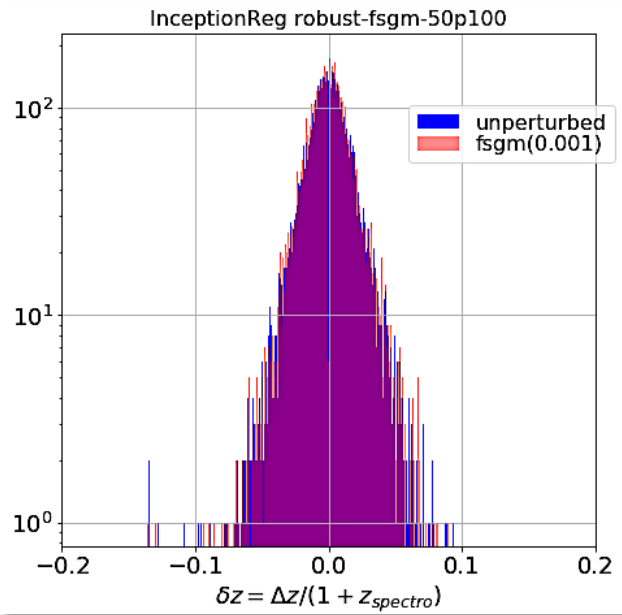
During the training, add a certain fraction of adversarial samples. They act as a regularisation.



Training with 50% adv. images **FSGM** using **ε=0.01** (10x the attack)

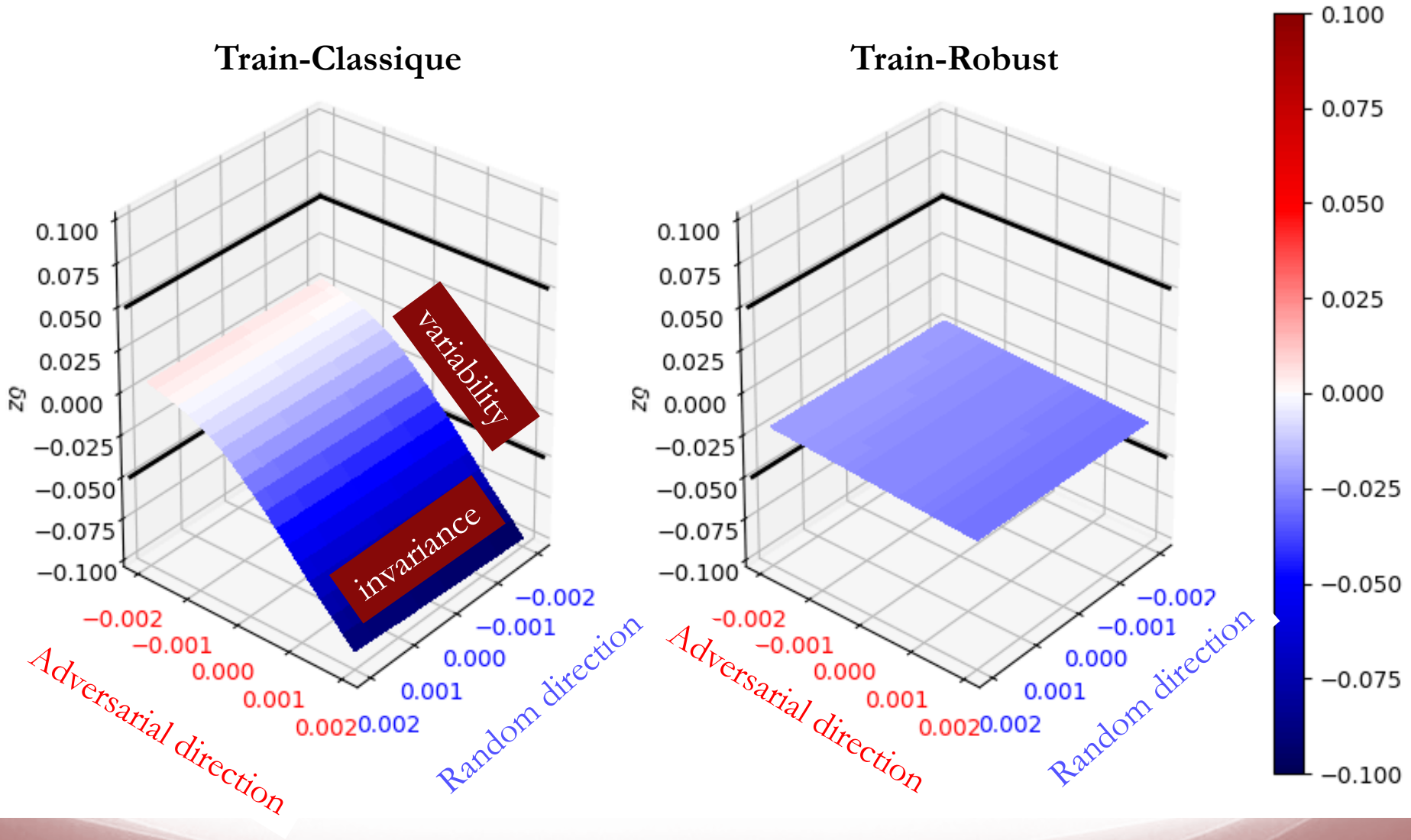






Training	Images	Bias ×10−4	*σmad* ×10−3	*η* (%)
Train-Robust (Min *ℓ test* ~4.4 10^{-4})	**Non-perturbed**	+0.67	14.27	1.95
	Perturbed	+5.54	15.56	2.14
Train-Classique (Min *ℓ test* ~2.2 10^{-4})	**Non-perturbed**	+1.98	10.31	0.47
	Perturbed	*+66.40*	33.63	5.31

Model response (δz) arround a reference image: (x_{ref}+δ)



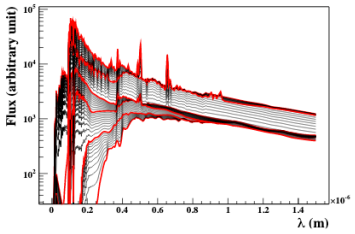
Summary/outlooks

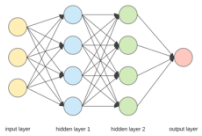
- The **classical** training/testing/validation triptych is **not enough to guarantee the generalisation power** of a network. Notice that **the problem in more general than CNN** (ie. DT, Gradient Boosted DT, R may also be affected as described in reference (Chen et al. 2019)).
- I've shown that mix up normal images with FSGM perturbed images gives some good results for ***Inception* robustness, but for normal images the results are worse than those obtained with classical training.**
- But this is **not the end of the story**: ***Inception* is not immune** against more aggressive methods. Some countermeasures have been elaborated but still it is a very active research domain as no satisfactory solution exists yet
- **So, what next?**
  - **Change the architecture** ? Well, I have tried several "classical" architecture but w/o real success, especially considering the robustness
  - **Go back to the origin** of the ML paradigm and modify the architecture with **operators** which **do not need to be trained** and are **more stable against perturbations**. Active research, new results will come, stay tuned!

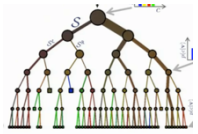
Back-up

Methods for Photo-z

Since the pioneering work in the 60's, several methods have been developed to estimate the redshift from the multi-bands photometric measurements, basically:







Possible combinaison

- template-fitting
  - Uses the SED and a method of fit
  - since Loh & Spillar 1986 ~30 galaxies in cluster 0024+1654,…, Beck et al 2016...
  - for LSST eg. Gorecki et al 2014 and Ansari et al 2019
- feature based Machine Learning
  - Uses a certain number of predefined features extracted from the measurements and feed to an engine as k-NN, NN/MLP, Decision Tree, BDT or Random Forest
  - Eg. Csabai et al. 2007 (k-NN) used by Beck et al 2016, Gorecki et al 2014 (NN) ,Ansari et al 2019 (BDT)...
- image based Deep Learning

Nb. Absolutely non exhaustive list of contributions.

Multi-steps perturbations

$$\delta^*(x) \equiv \underset{\|\delta\| \leq \varepsilon}{\text{argmax}}\ \ell(f_\theta(x+\delta), z)$$

Non-linear case

$$\delta \leftarrow \delta + \underset{\|u\| \leq \alpha}{\text{argmax}} \left[u^T . \nabla_\delta \ell(f_\theta(x+\delta), z) \right]$$

Kurakin et al 2016

$$\|\delta\|_\infty \leq \varepsilon$$

Projected Gradient Descend

$$\delta \leftarrow \mathcal{P}_\varepsilon^\infty \left[\delta + \alpha\ \text{sign}(\nabla_\delta \ell(f_\theta(x+\delta), z)) \right]$$

clip learning rate

One can also use GAN
Jang et al. 2019
Ensemble Adv. Training
Tramer et al. 2018

Different Inception networks and CNN architectures trained independently and tested with the same adversarial samples.

Models (images)	bias ($\times 10^{-4}$)	σ_{mad} ($\times 10^{-3}$)	η (%)
Iref (non perturbed)	0.3	11	1
$\varepsilon = 10^{-2}$, Single FSGM			
Iref	–	66	42
I0-I3 (*Iref* adv)	–	[63, 68]	[40, 43]
CNN (*Iref* adv)	–	76	49
$\varepsilon = 10^{-2}$, PGD, $\alpha = 10^{-3}$, $n_{iter} = 10$			
Iref	–	82	59
I0-I3 (*Iref* adv)	–	[76, 78]	[52, 55]

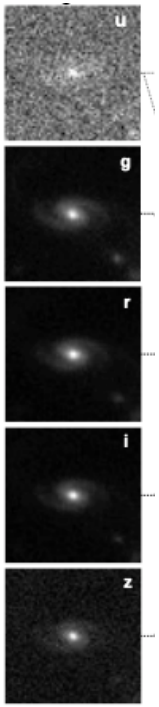
Some syst. studies

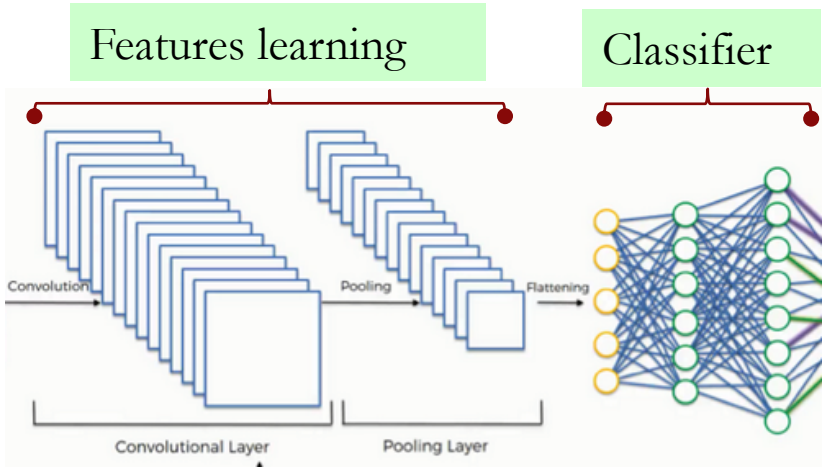
These are a very short summary of the J. Pasquet et al. thorough study.

Item	Comments
Galactic reddening (extra features added at the level of the FC part)	a strong reddening-dependent bias is observed If the information is not provided
Galaxy inclination	the CNN is very robust: large sample & data augmentation
Neighboring galaxies	The CNN learn how to improve redshift with neightboors at z>0.1
Variations throughout the surveyed area	Deviations in the SZ and Strip 82 of the SDSS dataset
PSF	induce a small but measurable amount of systematics on the estimated redshifts. Info can be added at the FC input (not done).

DL: what is promised?

Input images





Cascade of Convolutional layers

Output Photo-z

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26
arXiv:1806.06607v2

1) Variation: D'Isanto & Polsterer (2018) with a Gaussian Mixture Model as output
2) CNN architecture is used in other context: eg. g-g lens finding algo (Lanusse et al 2018), deblending (Burke et al 2019), objects classification (Gonzales et al 2018),…
… Non exhaustive list !