Adversarial examples (and robustness) of machine learning methods for stellar spectroscopy

Soledad Villar

Center for Data Science
Courant Institute of Mathematical Sciences

NEW YORK UNIVERSITY

Machine Learning Tools for Research in Astronomy
Ringberg, December 10 2019

Adversarial examples

- Deep neural networks are the state of the art for image classification.
- There's a lot of "overfitting".
  - Evidence 1: Do CIFAR10 Classifiers Generalize to CIFAR-10?
    *D*o not use the test set to select your models.

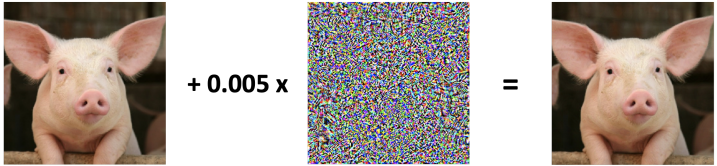
Recht, Roelofs, Schmidt, Shankar, 2018

Goodfellow, Shlens, Szegedyr, 2014

https://adversarial-ml-tutorial.org/ Kolter, Madry, 2019

Adversarial examples

- Deep neural networks are the state of the art for image classification.
- There's a lot of "overfitting".
  - Evidence 1: Do CIFAR10 Classifiers Generalize to CIFAR-10?
    Do not use the test set to select your models.
  - Evidence 2: Adversarial examples.



Recht, Roelofs, Schmidt, Shankar, 2018

Goodfellow, Shlens, Szegedyr, 2014

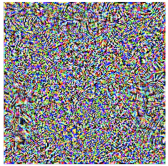
https://adversarial-ml-tutorial.org/ Kolter, Madry, 2019

Adversarial examples

- Deep neural networks are the state of the art for image classification.
- There's a lot of "overfitting".
  - Evidence 1: Do CIFAR10 Classifiers Generalize to CIFAR-10?
    Do not use the test set to select your models.
  - Evidence 2: Adversarial examples.

hog (91%)

+ 0.005 x



noise (not random)

=

airliner (99%)

Recht, Roelofs, Schmidt, Shankar, 2018

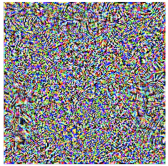
Goodfellow, Shlens, Szegedyr, 2014

https://adversarial-ml-tutorial.org/ Kolter, Madry, 2019

What is so bad about this example?

hog (91%)

+ 0.005 x



noise (not random)

=

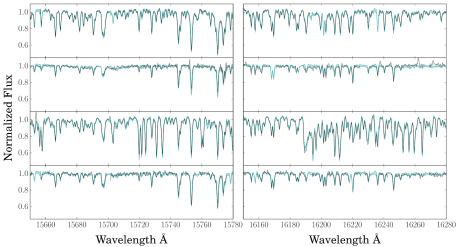
airliner (99%)

- Very small perturbation.
- Uninformative (looks like noise).
- Changes the output immensely.

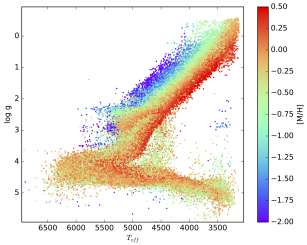
Why should an astrophysicist care?

They reveal non-robustness of models

Example: stellar spectroscopy



$\longleftrightarrow$



Example

Let x be the spectra (data), y physical parameters ($\log g$, T_{eff}, Fe/H) (labels):

$$\begin{aligned} \text{Physical model:} \quad & x = G_P(y) \quad (\textit{generative}) \\ \text{The Cannon:} \quad & x = G_C(y) \quad (\textit{generative}) \\ \text{Linear model:} \quad & y = F_L(x) \quad (\textit{discriminative}) \\ \text{AstroNN:} \quad & y = F_A(x) \quad (\textit{discriminative}) \end{aligned}$$

ASPCAP, Pérez et at. 2016

Ness, Hogg, Rix, Ho, Zasowski, 2015

Hogg, Eilers, Rix, 2019

Leung, Bovy, 2018

Example

Let x be the spectra (data), y physical parameters ($\log g$, T_{eff}, Fe/H) (labels):

$$\begin{aligned} \text{Physical model:} \quad & x = G_P(y) \quad (\textit{generative}) \\ \text{The Cannon:} \quad & x = G_C(y) \quad (\textit{generative}) \\ \text{Linear model:} \quad & y = F_L(x) \quad (\textit{discriminative}) \\ \text{AstroNN:} \quad & y = F_A(x) \quad (\textit{discriminative}) \end{aligned}$$

In generative models, given data $\hat{x}$ find $\hat{y} = F_G(\hat{x}) = \arg\min_y \|G(y) - \hat{x}\|$.
In discriminative models, given label $\hat{y}$ find $\hat{x}$ such that $F(\hat{x}) = \hat{y}$ is extremely ill-posed.

Conjecture: Generative models are more robust to adversarial attacks

---

ASPCAP, Pérez et at. 2016
Ness, Hogg, Rix, Ho, Zasowski, 2015
Hogg, Eilers, Rix, 2019
Leung, Bovy, 2018

What would an adversarial attack look like in this setting?

Attacks to regression

Let x be a stellar spectrum,

- Very small perturbation δ_x.
- Uninformative $\langle \delta_x, \nabla_x F_P(x) \rangle \approx 0$ (the noise has no physical information).
- Changes the output immensely: $F(x + \delta_x)$ very different from $F(x)$
  - or makes no sense physically

How to produce an adversarial example?

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^m$ data, $y \in \mathbb{R}^t$ labels. Consider h_θ : data $\rightarrow$ labels

$$\text{minimize}_\theta \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) \tag{1}$$

Adversarial example

$$\text{maximize}_{\delta \in \Delta} \ell(h_\theta(x + \delta), y), \qquad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2 \tag{2}$$

How to produce an adversarial example?

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^m$ data, $y \in \mathbb{R}^t$ labels. Consider h_θ : data $\to$ labels

$$\text{minimize}_\theta \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) \tag{1}$$

Adversarial example

$$\text{maximize}_{\delta \in \Delta}\, \ell(h_\theta(x+\delta), y), \qquad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2 \tag{2}$$

Fast Gradient Sign Method (FGSM) .

$$g := \nabla_x \ell(h_\theta(x), y) \tag{3}$$

$$\delta := \epsilon \cdot \mathrm{sign}(g) \text{ (for } \|\delta\|_\infty \leq \epsilon) \qquad \delta := \epsilon \cdot \frac{g}{\|g\|_2} \text{ (for } \|\delta\|_2 \leq \epsilon). \tag{4}$$

How to produce an adversarial example?

Adversarial example

$$\text{maximize}_{\delta \in \Delta}\, \ell(h_\theta(x+\delta), y), \qquad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2 \tag{5}$$

Projected Gradient Descent (PGD)

$$x^{t+1} := \Pi_{x+\Delta}(x^t + \alpha \nabla_x \ell(h_\theta(x), y)) \tag{6}$$

How to produce an adversarial example?

Adversarial example

$$\text{maximize}_{\delta \in \Delta}\ \ell(h_\theta(x+\delta), y), \qquad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2 \tag{5}$$

Projected Gradient Descent (PGD)

$$x^{t+1} := \Pi_{x+\Delta}(x^t + \alpha \nabla_x \ell(h_\theta(x), y)) \tag{6}$$

Targeted attacks

$$\text{maximize}_{\delta \in \Delta}\ (\ell(h_\theta(x+\delta), y) - \ell(h_\theta(x+\delta), y_{\text{target}})) \tag{7}$$

How to produce an adversarial example?

Adversarial example

$$\text{maximize}_{\delta \in \Delta}\, \ell(h_\theta(x+\delta), y), \qquad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2 \tag{5}$$

Projected Gradient Descent (PGD)

$$x^{t+1} := \Pi_{x+\Delta}(x^t + \alpha \nabla_x \ell(h_\theta(x), y)) \tag{6}$$

Targeted attacks

$$\text{maximize}_{\delta \in \Delta}\, (\ell(h_\theta(x+\delta), y) - \ell(h_\theta(x+\delta), y_{\text{target}})) \tag{7}$$

All these attacks require to access gradients. What if we don't?

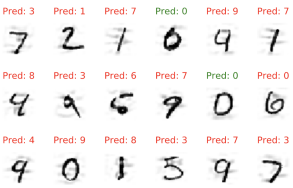
Examples

FGSM (fully connected)

FGSM (convNET)

PGD (convNET)

targeted PGD (convNET)



L2-PGD (convNET)

https://adversarial-ml-tutorial.org/ Kolter, Madry, 2019

How to evaluate the success of an attack to a regression?

- Comparison to random perturbations

$$A(\delta_x, x; h_\theta) = \frac{\ell(h_\theta(x + \delta_x), y = h_\theta(x))}{\mathbb{E}_{s \in \Delta} \ell(h_\theta(x + s), h_\theta(x))}. \tag{8}$$

How to evaluate the success of an attack to a regression?

- Comparison to random perturbations

$$A(\delta_x, x; h_\theta) = \frac{\ell(h_\theta(x+\delta_x), y = h_\theta(x))}{\mathbb{E}_{s\in\Delta}\ell(h_\theta(x+s), h_\theta(x))}. \tag{8}$$

- Comparison to relevant physical perturbation.
  Observed flux x, physical flux $\hat{x}$, label y.

  Nearest labels $y_1, \dots y_k$, physical fluxes $\hat{x}_1, \dots \hat{x}_k$.

$$A^k_{physics}(\delta_x, x; h_\theta) = \max_{t=1,\dots,k} \frac{\|y_t - h_\theta(x)\|_2}{\|\hat{x}_t - \hat{x}\|_2} \|\delta_x\|_2 - \|h_\theta(x) - h_\theta(x+\delta_x)\|_2. \tag{9}$$

How to evaluate the success of an attack to a regression?

- Comparison to random perturbations

$$A(\delta_x, x; h_\theta) = \frac{\ell(h_\theta(x+\delta_x), y = h_\theta(x))}{\mathbb{E}_{s\in\Delta}\ell(h_\theta(x+s), h_\theta(x))}. \tag{8}$$

- Comparison to relevant physical perturbation.
  Observed flux x, physical flux $\hat{x}$, label y.

  Nearest labels $y_1, \dots y_k$, physical fluxes $\hat{x}_1, \dots \hat{x}_k$.

$$A^k_{physics}(\delta_x, x; h_\theta) = \max_{t=1,\dots,k} \frac{\|y_t - h_\theta(x)\|_2}{\|\hat{x}_t - \hat{x}\|_2}\|\delta_x\|_2 - \|h_\theta(x) - h_\theta(x+\delta_x)\|_2. \tag{9}$$

- Bayesian setting.

  Network predicts labels $y = h_\theta(x)$ and uncertainties $\sigma = \sigma_\theta(x)$

$$\mathrm{Var}(\theta; \mu) = \mathbb{E}_{x\sim\mu}\sigma_\theta(x) + \mathbb{V}_{x\sim\mu}h_\theta(x) \tag{10}$$

$$\mathrm{Inv}(\theta; \mu) = \mathcal{R}(\theta; \mu)/\sqrt{\mathrm{Var}(\theta; \mu)}.$$

Preliminary results

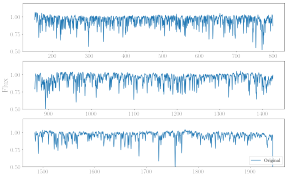
Attacks for stellar spectroscopy

+0.01* =

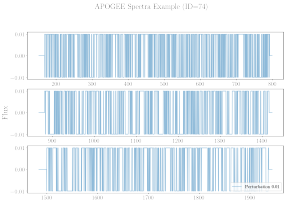
Huang, Martin, Scanlon, Wang, Hogg, V., in preparation

Preliminary results

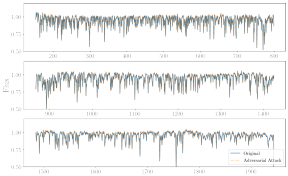
Attacks for stellar spectroscopy

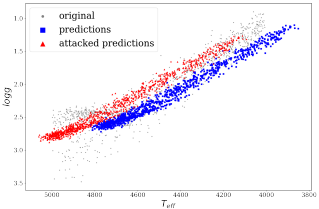


+0.01*

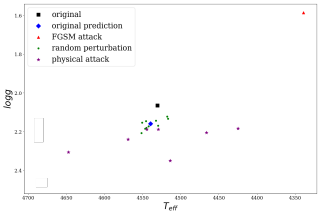


=

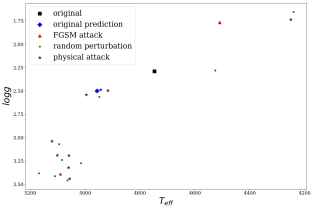




Attack to linear model



Attack to AstroNN



Attack to Cannon

Huang, Martin, Scanlon, Wang, Hogg, V., in preparation

Conclusions

- Discriminative models in astronomy are likely to be susceptible to attacks.
  - Single pixel attacks
  - L_2 attacks
- The notion of an attack cannot be disentangled from the notion of a <u>reasonable change</u>.
- It may be useful to train models to be adversarial-attack-robust.

$$\text{minimize}_\theta \, \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\delta\in\Delta} \ell(h_\theta(x+\delta), y).$$

Thanks

Adversarial Attacks Against Linear and Deep-Learning Regressions in Astronomy
Teresa Huang, Zacharie Martin, Greg Scanlon, Eva Wang, David W. Hogg and Soledad Villar
in preparation