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

Soledad Villar

Center for Data Science
Courant Institute of Mathematical Sciences

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?
    Do not use the test set to select your models.

---

Recht, Roelofs, Schmidt, Shankar, 2018
Goodfellow, Shlens, Szegedy, 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, Szegedy, 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.

![Image of adversarial example]

hog (91%) + 0.005 x noise (not random) = airliner (99%)

---

Recht, Roelofs, Schmidt, Shankar, 2018
Goodfellow, Shlens, Szegedy, 2014
https://adversarial-ml-tutorial.org/ Kolter, Madry, 2019
What is so bad about this example?

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

hog (91%) + 0.005 x noise (not random) = airliner (99%)
Why should an astrophysicist care?
They reveal non-robustness of models

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

Physical model: $x = G_P(y)$ \textit{(generative)}

The Cannon: $x = G_C(y)$ \textit{(generative)}

Linear model: $y = F_L(x)$ \textit{(discriminative)}

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

Physical model: $x = G_P(y)$ (generative)

The Cannon: $x = G_C(y)$ (generative)

Linear model: $y = F_L(x)$ (discriminative)

AstroNN: $y = F_A(x)$ (discriminative)

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
What would an adversarial attack look like in this setting?

Attacks to regression

Let $x$ be a stellar spectrum,

- Very small perturbation $\delta_x$.
- 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 \( h_\theta : \text{data} \rightarrow \text{labels} \)

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

Adversarial example

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

\[
\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) \quad (1)
\]

Adversarial example

\[
\max_{\delta \in \Delta} \ell(h_\theta(x + \delta), y), \quad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2 \quad (2)
\]

Fast Gradient Sign Method (FGSM).  

\[
g := \nabla_x \ell(h_\theta(x), y) \quad (3)
\]

\[
\delta := \epsilon \cdot \text{sign}(g) \quad (\text{for} \quad \|\delta\|_\infty \leq \epsilon) \quad \delta := \epsilon \cdot \frac{g}{\|g\|_2} \quad (\text{for} \quad \|\delta\|_2 \leq \epsilon). \quad (4)
\]
How to produce an adversarial example?

Adversarial example

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

Projected Gradient Descent (PGD)

\[
x^{t+1} := \Pi_{x+\Delta}(x^t + \alpha \nabla_x \ell(h_\theta(x), y))
\]
How to produce an adversarial example?

Adversarial example

$$\text{maximize}_{\delta \in \Delta} \ell(h_\theta(x + \delta), y), \quad \Delta = \{\delta : \|\delta\|_s \leq \epsilon\}, \quad s = \infty, 2$$ \hspace{1cm} (5)

Projected Gradient Descent (PGD)

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

Targeted attacks

$$\text{maximize}_{\delta \in \Delta} (\ell(h_\theta(x + \delta), y) - \ell(h_\theta(x + \delta), y_{\text{target}}))$$ \hspace{1cm} (7)
How to produce an adversarial example?

Adversarial example

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

Projected Gradient Descent (PGD)

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

Targeted attacks

$$\max_{\delta \in \Delta} \left(\ell(h_\theta(x + \delta), y) - \ell(h_\theta(x + \delta), y_{\text{target}})\right)$$ (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

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

(8)

- Comparison to relevant physical perturbation.

Observed flux \( x \), physical flux \( \hat{x} \), label \( y \).

Nearest labels \( y_1, ..., y_k \), physical fluxes \( \hat{x}_1, ..., \hat{x}_k \).

\[ A_{\text{physics}}(\delta_x, x; h_\theta) = \max_{t=1, ..., k} \|y_t - h_\theta(x)\|_2 \|\hat{x}_t - \hat{x}\|_2 \|\delta_x\|_2 - \|h_\theta(x) - h_\theta(x + \delta_x)\|_2. \]  

(9)

- Bayesian setting.

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

\[ \text{Var}(\theta; \mu) = \mathbb{E}_{x \sim \mu} \sigma_\theta(x) + \mathbb{V}_{x \sim \mu} h_\theta(x) \]  

(10)

\[ \text{Inv}(\theta; \mu) = \frac{\text{R}(\theta; \mu)}{\sqrt{\text{Var}(\theta; \mu)}}. \]
How to evaluate the success of an attack to a regression?

- Comparison to random perturbations

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

- Comparison to relevant physical perturbation.

  Observed flux \( x \), physical flux \( \hat{x} \), label \( y \).

  Nearest labels \( y_1, \ldots y_k \), physical fluxes \( \hat{x}_1, \ldots \hat{x}_k \).

\[ A_{\text{physics}}^k(\delta_x, x; h_\theta) = \max_{t=1,\ldots,k} \frac{\|y_t - h_\theta(x)\|_2}{\|\hat{x}_t - \hat{x}\|_2} \frac{\|\delta_x\|_2}{\|\hat{x}_t - \hat{x}\|_2} - \|h_\theta(x) - h_\theta(x + \delta_x)\|_2}. \] (9)
How to evaluate the success of an attack to a regression?

▶ Comparison to random perturbations

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

(8)

▶ Comparison to relevant physical perturbation.

Observed flux \( x \), physical flux \( \hat{x} \), label \( y \).

Nearest labels \( y_1, \ldots y_k \), physical fluxes \( \hat{x}_1, \ldots \hat{x}_k \).

\[ A^k_{\text{physics}}(\delta_x, x; h_\theta) = \max_{t=1,\ldots,k} \frac{\| y_t - h_\theta(x) \|_2}{\| \hat{x}_t - \hat{x} \|_2} \| \delta_x \|_2 - \| h_\theta(x) - h_\theta(x + \delta_x) \|_2. \]  

(9)

▶ Bayesian setting.

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

\[ \text{Var}(\theta; \mu) = \mathbb{E}_{x \sim \mu} \sigma_\theta(x) + \nabla_{x \sim \mu} h_\theta(x) \]  

\[ \text{Inv}(\theta; \mu) = \mathcal{R}(\theta; \mu)/\sqrt{\text{Var}(\theta; \mu)}. \]  

(10)
Preliminary results
Attacks for stellar spectroscopy

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 reasonable change.
▶ It may be useful to train models to be adversarial-attack-robust.

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