Universidad Nacional Autónoma de México

Instituto de Matemáticas

Interpretabilidad de los Ataques Adversarios

Proyecto Final Reporte

- Redes Neuronales -

Rodrigo Fritz Hernández

Aaron Kelley

18 de junio, 2021

Abstract

Abstract Goes here

Keywords:

Índice

Development Notes/Ideas

- maybe we should take out Projected Gradient Descent
- with cleverhans we can do targeted attacks with Fast Gradient Descent

1. Introducción

Las redes neuronales profundas (DNNs) han ganado una alta reputación con respecto a sus capacidades de clasificar imágenes igual (o hasta superior) que los humanos. Pero en el pasado reciente, ha quedado cada vez más claro que las redes aprenden clasificar de manera muy distinta que los humanos. Una de las propiedades que realmente demostró eso fue el descubrimiento de su susceptibilidad a los ataques adversarias [8]. Esos ataques se construyen agregándoles a las imágenes una pequeña perturbación imperceptibles para los humanos que engañan a la red. Es decir, aunque una imagen adversaria se parezca igual a la original, la red se equivoca con la clasificación de la adversaria con alta probabilidad. Como los ataques pueden no ser detectables por los humanos, se plantea la preocupación que puedan ser usados maliciosamente; por ejemplo, en la tecnología de reconocimiento de imágenes que se utiliza en los automóviles autónomos. Por eso se requiere maá profundizamiento del conocimiento asociado.

A grandes rasgos, los ataques pueden dividirse entre dos categorías: dirigidos y no-dirigidos. Primero hablemos de los ataques dirigidos; muchos usan el gradiente de manera directa. La idea es que se toma el gradiente de la función de pérdida con respecto a la imagen, y eso se usa para diseñar una perturbación que maximiza el error de la clasificación cuando se le agrega a la imagen. Dos ejemplos de ataques no-dirigidos que utilizan el gradiente directamente son el projected gradient decent (PGD) [6] y el fast gradient sign method (FGSM) [3]. Este último saca el signo de cada elemento del gradiente en lugar de usar el gradiente verdadero; eso hace que sea más rápido con grandes cantidades de datos. En este artículo se usan los dos de PDG y FGSM para explorar los ataques no-dirigidos. El objetivo de los ataques no dirigidos es que cambien la clasificación correcta a cualquier otra ataque, mientras que los ataques dirigidos tienen una deseada a la que cambian la clasificación correcta. Se examina el ataque Carlini & Wager (CW), el cual puede actuar como dirigido o no-dirigido.

Cabe mencionar la diferencia entre los ataques caja blanca y caja negra. En el primero todos los detalles de la red (pesos, arquitectura, etc) son conocidos por el atacante. Al contrario, en los caja blanca los detalles son escondidos, y solo pueden saberse las entradas y salidas. Aunque parezca muy difícil, los ataques caja negra son bastante exitosos por la facilidad de aproximar el gradiente solo por las entradas y las salidas. Los ataques que se mencionan en este artículo son de los caja blanca.

Desde el descubrimiento de ese vulnerabilidad que tienen las DNN, se han hecho defensas para tratar de combatir los ataques. Aunque ninguna defensa funciona para completamente resistir los ataques, muchas sí tienen efecto, y vale la pena explorar cuales propiedades contribuyen a su éxito. Igual que los ataques, las defensas también se pueden categorizar en dos modalidades. Las del primer tipo modifican el entrenamiento de la red para que la función de pérdida se vuelva más suave. Entre más suave esa función, más difícil será para que las perturbaciones pequeñas hagan cambios grandes. Un ejemplo conocido de este tipo es el entrenamiento adversario [3, 7, 8]. Las del otro tipo de defensa, y las que vamos a estudiar en este paper, no modifican el entrenamiento ni la arquitectura, sino le ponen un preprocesamiento a las imágenes de entrada. La defensa que se va a utilizar en ese reporte es compresión JPEG [2].

2. Métodos

2.1. Datos

2.1.1. MNIST

[5]

2.1.2. CIFAR-10

[4]

2.2. Ataques

2.2.1. Fast Gradient Method

[**maybe more**, 3]

Sean θ los parametros de un modelo, x la entrada, y las salidas asociadas, y $J(\theta, x, y)$ la función de costo. La función de costo se lineariza alrededor del valor actual de θ. Sea $\epsilon \in \mathbb{R}^+$. Definamos la imagen adversaria

$$\tilde{x} = x + \epsilon\eta_{\text{opt}}$$

Se puede definir η_{opt} por el problema de optimización

$$\eta_{\text{opt}} = \underset{\eta}{\text{argmax}} \left\{ \text{grad}^\top \eta : \|\eta\|_p < \epsilon \right\}$$

Donde $p \in \mathbb{N} \cup \{\infty\}$ y $\text{grad} = \nabla_x J(\theta, x, y)$. Experimentamos con tres valores de p:

a) $p = 1$, no lo sé, pero se encuentra en el codigo

b) $p = 2$,

$$\eta_{\text{opt}} = \frac{\text{grad}}{\|\text{grad}\|}$$

c) $p = \infty$,

$$\eta_{\text{opt}} = \text{sign}(\nabla_x J(\theta, x, y))$$

2.2.2. Carlini & Wagner

[1]

2.3. Defensas

2.3.1. Compresión JPEG

[2]

3. Resultados

- Show graphs of jpeg defense vs epsilon of each attack

Referencias

[1] Nicholas Carlini y David Wagner. *Towards Evaluating the Robustness of Neural Networks*. 2017. arXiv: 1608.04644 [cs.CR].

[2] Nilaksh Das y col. *Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression*. 2017. arXiv: 1705.02900 [cs.CV].

[3] Ian J. Goodfellow, Jonathon Shlens y Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[4] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Inf. téc. 2009.

[5] Yann LeCun, Corinna Cortes y CJ Burges. «MNIST handwritten digit database». En: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010).

[6] Aleksander Madry y col. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. arXiv: 1706.06083 [stat.ML].

[7] Uri Shaham, Yutaro Yamada y Sahand Negahban. «Understanding adversarial training: Increasing local stability of supervised models through robust optimization». En: *Neurocomputing* 307 (sep. de 2018), págs. 195-204. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.04.027. URL: http://dx.doi.org/10.1016/j.neucom.2018.04.027.

[8] Christian Szegedy y col. *Intriguing properties of neural networks*. 2014. arXiv: 1312.6199 [cs.CV].