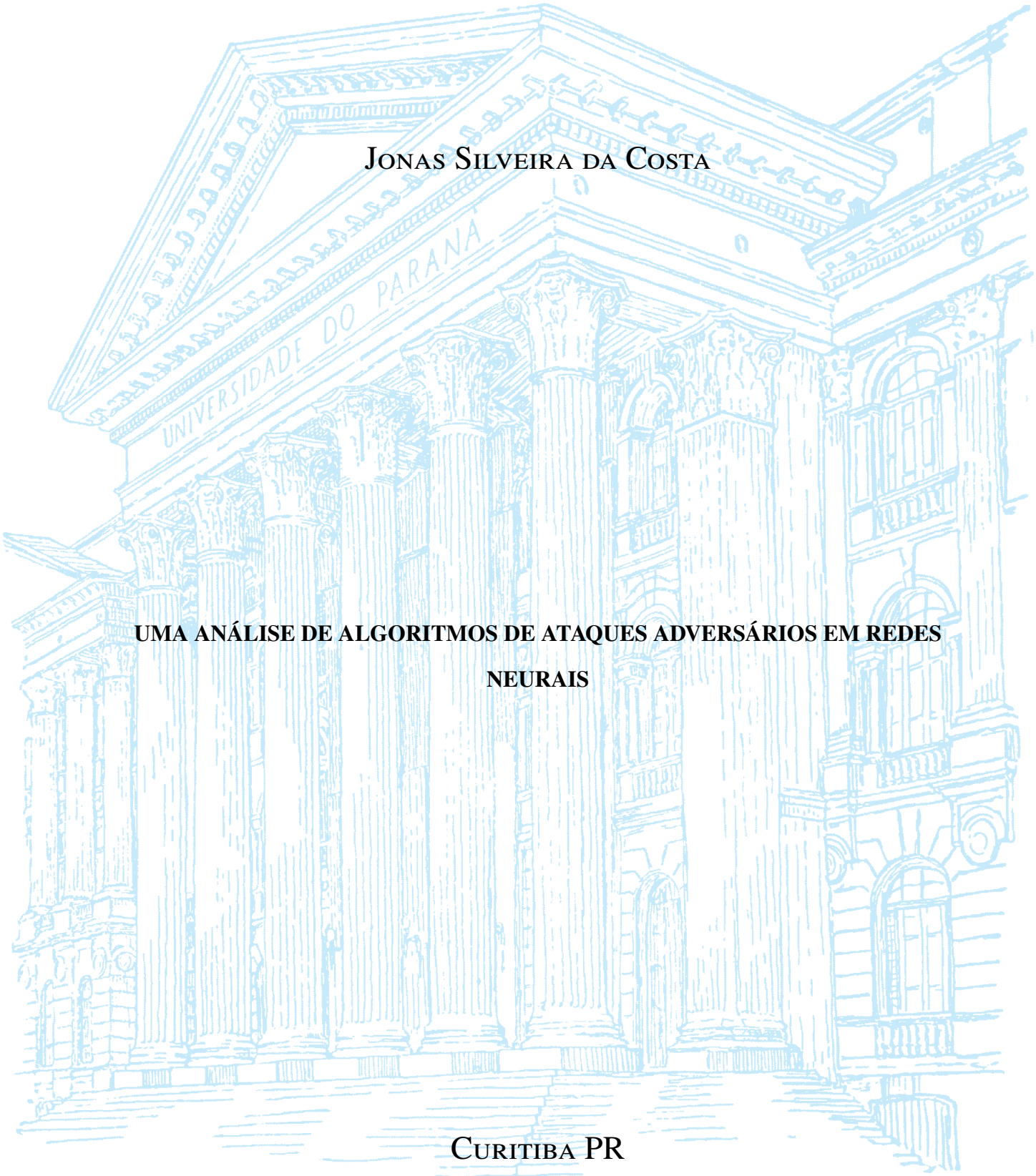
Universidade Federal do Paraná

Jonas Silveira da Costa



UMA ANÁLISE DE ALGORITMOS DE ATAQUES ADVERSÁRIOS EM REDES NEURAIS

Curitiba PR
2018

Jonas Silveira da Costa

UMA ANÁLISE DE ALGORITMOS DE ATAQUES ADVERSÁRIOS EM REDES NEURAIS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Eduardo Jaques Spinosa.

Curitiba PR
2018

Resumo

Redes neurais podem ser implementadas como método de reconhecimento e classificação de imagens, obtendo bastante sucesso em diversas aplicações. Apesar disso, estudos recentes demonstram que esses classificadores são vulneráveis a ataques adversários. Este trabalho tem como objetivo apresentar uma visão geral do avanço dos ataques e defesas nos últimos anos, assim como realizar experimentos para testar a eficácia dos algoritmos de ataque e os efeitos de utilizar algoritmos de defesas.

Palavras-chave: Ataques adverários.

Abstract

Neural networks can be implemented as a method of image recognition and classification, achieving a certain success in several applications, but the latest works demonstrate the vulnerability of these networks to adversarial attacks. This work aims to present an overview of the adversarial attacks and defenses in the last few years, as well as to make experiments to analyze how effective adversarial attacks algorithms can be, and what effects using a defense algorithm can bring.

Keywords: Adversarial attacks.

Lista de Figuras

Lista de Acrônimos

FGSM	*Fast Gradient Sign Method*
JSMA	*Jacobian-based Saliency Map Attack*
C&W	*Carlini and Wagner*
NIPS	*Neural Information Processing Systems*
MNIST	*Modified National Institute of Standards and Technology*

Sumário

Capítulo 1
Introdução

Velocidade e taxa de acerto são aspectos de grande importância para classificadores de imagens, pontos em que redes neurais são capazes de obter bons resultados comparados com outras técnicas de aprendizado de máquina (Russakovsky et al. [2014]).

A segurança é outro aspecto que pode ser analisado, pesquisas recentes demonstram a vulnerabilidade das redes neurais a imagens adversárias (Szegedy et al. [2013]), as quais são capazes de enganar classificadores com alta taxa de acerto em situações normais. Devido a esse problema, seu uso pode ser limitado a aplicações em que a segurança é crítica, por exemplo uma placa de trânsito ser interpretada incorretamente por um carro autônomo ou conteúdo ilegal ser modificado de forma a não ser detectado pelas autoridades.

Para tentar melhorar a robustez dessas redes neurais, existem técnicas de defesas, entretanto muitas acabam sendo apenas soluções temporárias, ou que afetam a velocidade e taxa de acerto do classificador, existindo ainda a própria incerteza em ser possível construir um modelo realmente robusto (Madry et al. [2017]).

Este trabalho tem como objetivo apresentar o que são os ataques adversários, suas propriedades e exemplos. Para analisar redes neurais como classificadores de imagens ao confrontado com imagens adversárias, realizando experimentos para verificar taxas de acerto, assim como aplicar estratégias de defesa para entender seus efeitos.

1.1 Estrutura do Documento

Esta monografia está organizada em 6 capítulos. O capítulo 2 apresenta os conceitos básicos sobre redes neurais, os neurônios que as formam e o gradiente descendente. O capítulo 3 apresenta alguns ataques adversários, os contextualizando, quais são suas áreas de ataque e objetivos. O capítulo 4 apresenta os experimentos realizados. Inicia-se com uma explicação das ferramentas utilizadas e metodologia. Em seguida são analisados os dados de taxa de acerto encontrados ao executar os ataques. O capítulo 5 apresenta possíveis estratégias de defesa, experimentos e comparações.

Capítulo 2
Fundamentos

2.1 Redes Neurais

O cérebro humano é capaz de organizar seus neurônios a fim de realizar ações como reconhecimento de padrões e controle motor, de forma muito rápida, mais do que os computadores existentes. Grande parte dessa habilidade vem através da experiência e aprendizado durante os anos.

Uma rede neural artificial é modelada de forma semelhante ao cérebro humano, construída com diversos neurônios interconectados, sendo capaz de aprender, armazenar esse conhecimento e utilizá-lo.

O processo de treinamento é feito através de um algoritmo de leitura, o qual basicamente, é uma função que modifica os pesos internos da rede de forma ordenada. O aprendizado consiste em encontrar pesos que exibem o comportamento desejado.

Um grande benefício é a capacidade de generalização, referindo a habilidade de uma rede neral produzir saídas para dados inicialmente desconhecidos, utilizando como base seu conhecimento prévio acumulado com outros dados semelhantes. O que torna as redes neurais ótimas alternativas para problemas complexos e de grande escala, sendo capazes de superar humanos em diversas situações (Silver et al. [2017]).

O neurônio, é uma unidade de processamento fundamental para o funcionamento e operação da rede neural, podendo ser dividido nas seguintes partes:

- **Entrada**: São os valores de entrada da rede
- **Pesos**: é um valor que é calculado durante o treinamento, ao fim sendo responsável para classificar as novas entradas
- ***Bias***: é uma função aditiva para somar os pesos, serve para calibrar o valor final
- **Função de ativação**: limita a amplitude de saída do neurônio, geralmente aproximando para valores entre 0 e 1. Funções comuns são *Sigmoid*, *Tanh*, *ReLu* e *Softmax*.

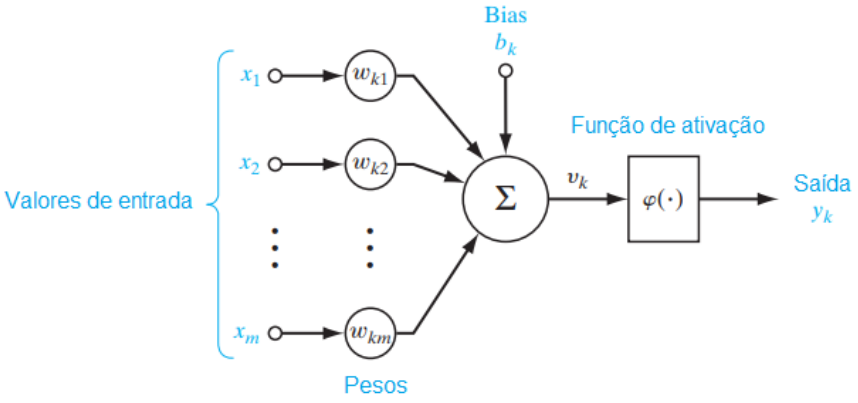


Figura 2.1: Estrutura do neurônio. Fonte: Haykin [2008]

2.2 Classificadores

Reconhecimento de padrões é algo que humanos são aptos a fazer com facilidade, imediatamente reconhecendo faces familiares, vozes e outros tipos de informações através dos sentidos, habilidade desenvolvida ao longo dos anos.

Redes neurais artificiais podem possuir essa mesma capacidade através do processo de treinamento. Para o reconhecimento de imagens, é necessário uma série de imagens rotuladas com o padrão a que pertencem, sendo então utilizadas como entrada da rede neural, as quais são processadas e os pesos internos da rede são definidos.

O próximo passo consiste em apresentar a rede uma imagem com um padrão desconhecido, mas que pertença a uma mesma população dos padrões usados durante o treinamento. A rede é capaz de identificar a classe dessa nova dado porque extraiu informações dos dados de treinamento.

Para esse trabalho será explorado a propriedade das redes neurais de servirem como classificadores, analisando as fragilidades quando confrontadas com exemplos adversários.

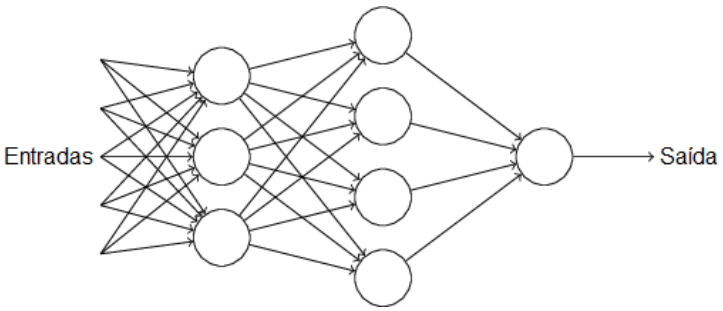


Figura 2.2: Exemplo de estrutura de uma rede neural. Fonte: adaptado de Haykin [2008]

2.3 Gradiente descendente

É comum tarefas de aprendizado de máquina em geral serem resumidos à problemas de otimização, para isso serve o algoritmo de descida do gradiente, o qual tem o objetivo de otimizar funções complexas iterativamente, como as funções de ativação das redes neurais, que é realizado ao encontrar um mínimo da função.

Normalmente uma função possui vários valores mínimos não ótimos, o ideal é encontrar o melhor mínimo dentre eles, o que pode ser uma tarefa complicada e é um problema comum durante a fase de treinamento.

Para redes neurais, as funções podem envolver uma grande quantidade de pesos e bias de forma complexa. A estratégia utilizada para encontrar o mínimo da função nesses casos é iniciar com um ponto aleatório e através de cálculos de derivadas mover esse ponto em direção a um menor.

De forma geral, a descida do gradiente é um algoritmo de otimização usado para encontrar valores de coeficientes, pesos e bias, da função de ativação. Existem ainda variações como o *Batch Gradient Descent*, onde os dados de treinamento são divididos em lotes e o *Stochastic Gradient Descent* que é uma variante mais rápida, podendo acelerar o processo de treinamento quando a quantidade de dados é muito grande.

A relevância do algoritmo de descida do gradiente para ataques adversários é o fato que é alvo de diversos modelos, também sendo responsável em certos casos por um ataque não conseguir encontrar imagens adversárias (Carlini and Wagner [2016]).

Capítulo 3
Ataques adversários

3.1 Introdução

O objetivo de um ataque é construir exemplos adversários, que são dados capazes de enganar o classificador, como nesse trabalho serão utilizados classificadores de imagens, os ataques devem gerar imagens adversárias. Segundo Yuan et al. [2017] o modelo de ataque pode ser decomposto em quatro aspectos:

- **Falsificação adversária**: relacionado ao tipo do exemplo adversário gerado. Sendo classificado como falso positivo quando o exemplo adversário gerado é reconhecível por um humano ou falso negativo quando não é reconhecível por um humano.
- **Conhecimento do adversário**: relacionado ao conhecimento que um adversário têm do classificador alvo, seus parâmetros internos, com as funções de ativação utilizadas e quantidade de camadas. Um ataque pode ser classificado como do tipo "caixa preta" quando não possui nenhum conhecimento ou "caixa branca" quando sabe tudo sobre a rede alvo.
- **Especificação do adversário**: relacionado ao objetivo do adversário, sendo dividido em "ataques direcionados" quando o objetivo é gerar um exemplo adversário com uma classe especifica ou "ataques não direcionados" quando a classe gerada não importa, nesse caso o objetivo é apenas enganar o classificador.
- **Frequência de ataque**: relacionado a como o ataque é gerado, sendo dividido em "iterativo" quando o exemplo adversário é gerado diversas vezes até estar otimizado ou "passo único"quando é gerado apenas uma vez.

  É importante complementar que um ataque não está fixo nessas categorias, por exemplo, um mesmo ataque pode ser capaz de gerar exemplos adversários direcionados e não-direcionados, inclusive alguns ataques possuem variantes para aspectos diferentes.

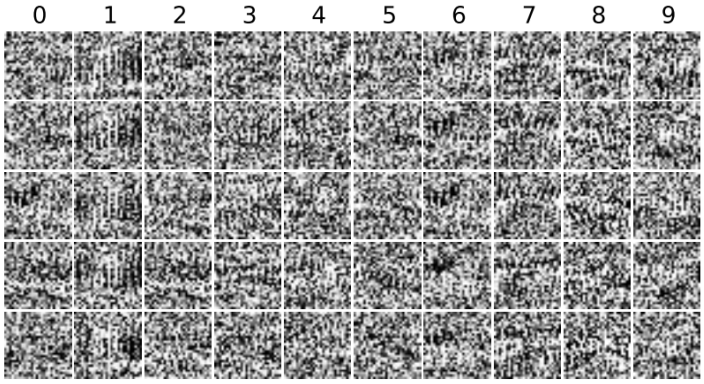


Figura 3.1: Imagem adversária do com falsificação adversária do tipo falso positivo. Fonte: Nguyen et al. [2014]

3.2 Ataque L-BFGS

Em 2014, Szegedy et al. [2013] fez a descoberta que redes neurais são vulneráveis à ataques adversários, demonstrando ser possível gerar imagens adversárias capazes de enganar redes neurais. Por ter sido o primeiro tipo de ataque adversário, acabou impulsionando a área de pesquisa, proporcionando espaço para diversos outros modelos de ataques e discussões sobre robustez e segurança em redes neurais.

Esse é um ataque que se encaixa nas categorias iterativa, direcionado e caixa branca. Apesar de ser caixa branca, os autores exploram que as imagens adversárias geradas podem ser generalizadas a diferentes modelos.

O seu funcionamento se resume a um problema de otimização que utiliza o algoritmo L-BFGS(*Limited memory-Broyden–Fletcher–Goldfarb–Shanno*) para aproximação de valores. O problema identificado a ser resolvido é o seguinte:

$$\text{minimizar } c \cdot \|x - x'\|_2^2 + \text{loss}_{F,l}(x') \text{ tal que } x' \in [0,1]^n$$

Onde dado uma imagem x, o modelo tem por objetivo encontrar uma imagem x' que é similar, porém que seja classificada de forma diferente, e consequentemente de forma incorreta. Sendo $\text{loss}_{F,l}$ a função de ativação do classificador alvo. O problema se resume a encontrar a constante c, ou múltiplas constantes c

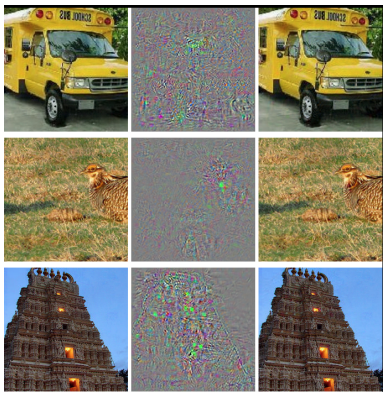


Figura 3.2: Imagem adversária gerada com o uso do FGSM. Fonte: Goodfellow et al. [2014]

3.3 FGSM (*Fast Gradient Sign Method)*

Esse modelo proposto por Goodfellow et al. [2014], se encaixa nas categorias "passo único", direcionado e caixa branca. Esse método foi projetado para ser rápido, e não necessariamente ótimo, portanto perturbações nas imagens não são as mínimas possíveis. Seu funcionamento se resume em aplicar uma função de perturbação em cada pixel, afetando sua intensidade.

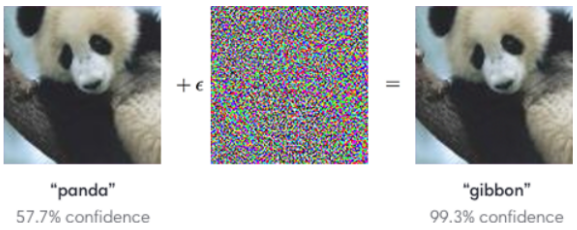


Figura 3.3: Imagem adversária gerada com o uso do FGSM. Fonte: Goodfellow et al. [2014]

Ainda serviu como base para diversas outras variações, por exemplo Kurakin et al. [2016], que tem como objetivo utilizar câmeras reais e outros sensores para capturar as imagens antes de passar pelo classificador.

Figura 3.4: Imagens utilizando câmera de celular usadas como entrada de um classificador. Fonte: Kurakin et al. [2016]

3.4 JSMA (*Jacobian-based Saliency Map Attack)*

É um modelo de ataque proposto por Papernot et al. [2015] do tipo direcionado e iterativo, de forma geral é um algoritmo guloso que escolhe um pixel por vez para modificar, a cada iteração tentando aproximar mais do rótulo alvo.

De forma mais específica, o JSMA faz uso do gradiente para computar um mapa de saliências, o qual representa o impacto que cada pixel possui na classificação resultante, ou seja, no rótulo da imagem. Valores altos na matriz indicam que alterar aquele pixel tem boa chance de modificar a classificação que o modelo dará para a imagem.

O processo se repete até que um máximo de iterações seja alcançado ou a classificação seja alterada. Esse modelo tem como objetivo causar uma pequena perturbação na imagem, de forma a modificar uma pequena porção da imagem (em torno de 4%) e ainda assim enganar a rede, entretanto utilizar um valor máximo de iterações muito baixo pode resultar na falha de gerar imagens adversárias, ou no caso de um valor muito alto, pode fazer com que a modificação gerada seja muito grande.

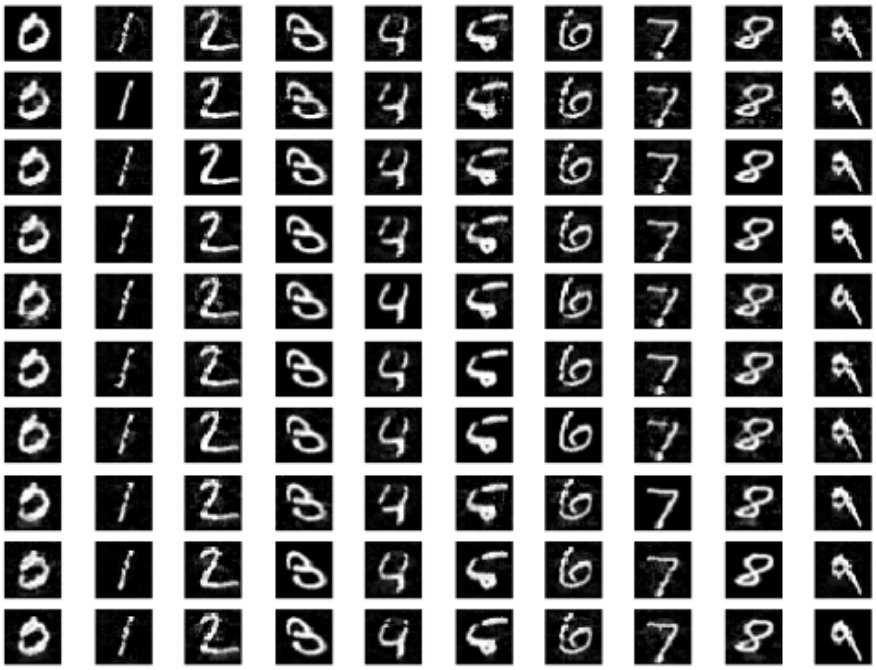


Figura 3.5: Imagens adversárias com o uso do JSMA. Fonte: Papernot et al. [2015]

3.5 C&W (Carlini and Wagner)

Modelo de ataque do tipo direcionado proposto por Carlini and Wagner [2016], surgiu com o objetivo de derrotar um tipo de defesa (*defensive distillation*), e se provou eficiente para esse propósito, obtendo também sucesso para grande parte das defesas existentes.

O método proposto surgiu a partir de uma adaptação do L-BFGS e acabou se dividindo em três variantes, todas obtendo sucesso no objetivo inicial. E assim como no JSMA, as modificações feitas nas imagens adversárias geradas são pequenas.

Os autores propuseram que novas defesas usassem esse e outros ataques eficientes para seus testes, para que seja possível desenvolver defesas que realmente tornem o classificador mais robusto.

3.6 *Provably minimally-distorted*

Carlini et al. [2017] mostrou que dos artigos aceitos em 2018 na ICLR (*International Conference on Learning Representations*), mais da metade das defesas propostas já haviam sido quebradas. Foi proposto uma técnica de como construir imagens adversárias com distorção miníma, utilizando métricas formais para avaliar as imagens geradas, ao mesmo tempo, essas métricas se provaram eficazes em provar propriedades sobre defesas.

Figura 3.6: Imagens adversárias geradas com o uso do C&W L2. Fonte: Carlini and Wagner [2016]

Utilizando essas métricas os autores analisaram as defesas propostas na ICLR, identificando que uma delas Madry et al. [2017] teve sucesso em aumentar a distorção necessária nas imagens adversárias para que fossem capazes de enganar o classificador.

3.7 NIPS *(Neural Information Processing Systems)*

O NIPS é uma conferência sobre aprendizado de máquina realizada todo ano em dezembro. Em 2017, foi proposta uma competição de ataques adversários e defesas pelo *Google Brain*, com a intenção de atrair atenção ao risco de segurança causado pela vulnerabilidade dos classificadores às imagens adversárias.

A competição foi dividida em categorias, sendo duas delas para ataques, direcionado e não direcionado, a outra específica para defesas. Nas categorias de ataque, o atacante recebia 1 ponto cada vez que fosse capaz de enganar uma defesa. Para as defesas, recebia 1 ponto para cada imagem classificada corretamente e os pontos foram então normalizados.

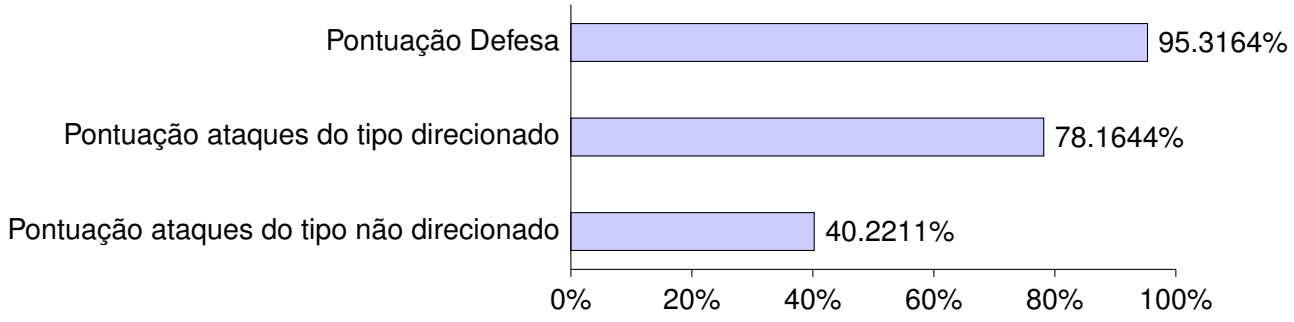


Figura 3.7: Melhores resultados de cada categoria na competição NIPS 2017. Fonte: Kurakin et al. [2018]

Como demonstrado por Kurakin et al. [2018], a competição serviu para aumentar a atenção ao problema dos ataques adversários, com mais de 100 times participando. A competição levou os competidores a melhorar métodos já conhecidos, assim como explorar novas técnicas.

Em todas as categorias da competição existiu alguma melhora em comparação com o que existia anteriormente. Em especial a categoria de defesa apresentou bons resultados, com exceção do pior caso que ficou abaixo da média.

Capítulo 4
Experimentos

4.1 *Cleverhans* e *TensorFlow*

Cleverhans (Papernot et al. [2018]) é uma biblioteca que oferece referências de implementação de ataques adversários, que possui os objetivos de construir imagens adversárias para treinamento e servir como *benchmark* para testar a robustez de classificadores. Faz uso do *TensorFlow*, ferramenta do *Google* utilizada para criação e treinamento de redes neurais.

4.2 Metodologia

Foi utilizado o *dataset* do *MNIST* para os experimentos. Esse *dataset* inclui imagens de dígitos de zero à nove manuscritos. Foi utilizado um classificador de imagens pré-treinado para o MNIST, oferecido na biblioteca *cleverhans*, com taxa de acerto de 99.3%.

Possuindo então um classificador com o treinamento realizado, e sendo capaz de classificar novas entradas, foram geradas imagens adversárias com o uso da biblioteca *cleverhans* para os ataques FGSM, JSMA e C&W L2, com o objetivo de analisar e comparar a taxa de acerto que o modelo consegue atingir ao tentar classificar as imagens geradas por cada um desses ataques.

4.3 Resultados

Os experimentos realizados com os ataques adversários utilizaram os dados de taxa de acerto alcançados pelo classificador para as imagens originais e as imagens adversárias geradas pelos ataques, comparando seus resultados e mostrando exemplos das imagens geradas para exemplificar as modificações feitas por cada algoritmo de ataque na imagem original.

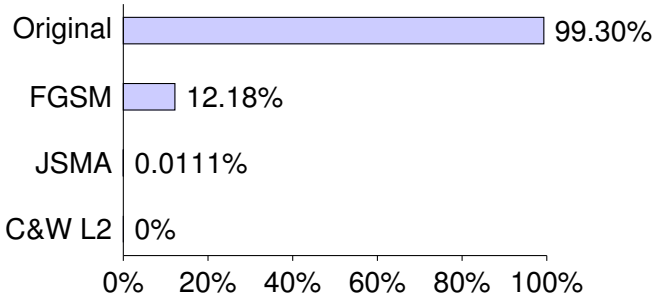


Figura 4.1: Porcentagem de taxa de acerto para o modelo original, e os ataques FGSM, JSMA e C&W L2.

Foi identificado que os ataques são muito eficazes (figura 4.1). O algoritmo FGSM teve o pior resultado dentre os ataques, entretanto foi capaz de gerar as imagens adversárias com maior velocidade.

Os algoritmos JSMA e C&W, apesar de mais lentos para gerar as imagens adversárias, enganaram o classificador com pequenas modificações nas imagens.

Devido a taxa de acerto ser reduzida a 0%, será também analisado casos com aplicação de defesa no capítulo 5 para verificar os comportamentos de uma rede mais robusta.

Figura 4.2: Exemplo de imagem adversária gerada com os ataques experimentados.

A figura 4.2 mostra imagens geradas para o dígito "1"com cada um dos ataques, neste caso o classificador foi capaz de classificar corretamente a imagem gerada pelo modelo FGSM e incorretamente os outros dois.

Em todos os casos é possível para um humano identificar o dígito sem muita dificuldade, é possível ainda perceber que as perturbações geradas nas imagens para o modelo FGSM afetam todo o "bloco"da imagem, enquanto os outros dois modelos aplicaram modificações mais diretas ao dígito, possuindo pouca diferença ao comparar com a imagem original. Em média, entre os três ataques, FGSM realiza a maior modificação e C&W a menor.

Capítulo 5
Estratégias de defesa

As estratégias de defesa têm por objetivo aumentar a robustez dos modelos classificadores de forma a torná-los mais seguros e capazes de classificar corretamente imagens adversárias, ou ao menos melhorar sua taxa de acerto nesses casos. Como dito por Yuan et al. [2017], as defesas podem ser divididas em duas categorias:

- **Reativa**: detecta as imagens adversárias após a construção da rede neural. Defesas desse tipo tentam identificar que a entrada é uma imagem adversária.
- **Proativa**: durante a fase do treinamento são aplicados métodos para aumentar a robustez. Defesas dessa categoria tentam classificar corretamente as imagens adversárias.

5.1 *Adversarial Training*

Esta forma de defesa é da categoria proativa e se baseia em utilizar imagens adversárias geradas previamente como parte do treinamento de um novo classificador. Idealmente o modelo será capaz de se tornar mais robusto aos ataques, porém esta técnica sofre de possuir um grande custo de treinamento, e sua eficácia também depende muito das imagens adversárias utilizadas, já que as imagens usadas durante o treinamento podem ter sido geradas de forma diferente às imagens que serão utilizadas de fato durante um ataque, e como visto no capitulo 3, existe uma grande variedade de ataques possíveis, e essa quantidade só tem aumentado.

Outro problema é que essa forma de defesa não é adaptável, propriedade que as redes neurais possuem em adaptar seus pesos internos a alterações pequenas no ambiente Haykin [2008]. O que causa essa limitação é o fato de que o treinamento deve ser feito com um grande número de dados para ser significativo, e essa necessidade de atualização pode ser muito constante para que seja viável. Ainda, como mostrado em Papernot et al. [2016], existe uma possibilidade de perda de taxa de acerto para casos normais ao utilizar essa estratégia.

Outro ponto identificado por Nguyen et al. [2014] é que realizar o treinamento utilizando imagens do tipo falso positivos não traz benefícios.

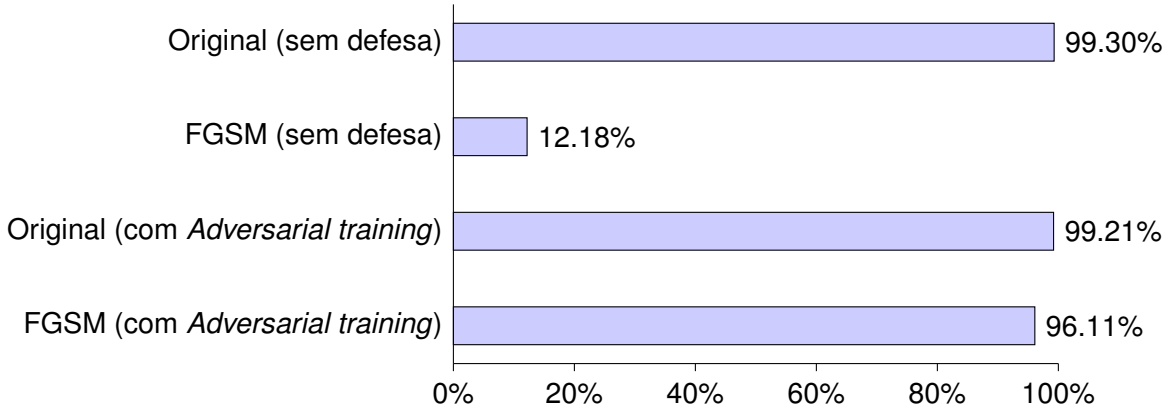


Figura 5.1: Taxa de precisão ao utilizar classificador sem defesa com as imagens originais e imagens adversárias geradas com o FGSM e feito o mesmo experimento para classificador com o uso do *Adversarial training*.

Foram realizados experimentos aplicando essa estratégia de defesa contra imagens geradas com o FGSM, e os resultados (figura 5.1) obtidos demostram que para treinar o modelo com esse método existe um pequeno custo na taxa de acerto comparado com o classificador sem defesa, por outro lado teve um grande aumento na robustez do classificador, novamente comparado ao classificador sem defesa.

5.2 *Adaptive Noise Reduction*

Defesa da categoria reativa que pode é aplicado sobre o classificador, pois opera de forma independente. Tem por objetivo identificar que a entrada é uma imagem adversária e aplicar um filtro na imagem de forma que seja possível classificar corretamente.

O problema com essa defesa, é o fato que aplicar esse filtro pode gerar falsos positivos, e uma de suas vantagens é que não é necessária nenhuma alteração ao classificador, inclusive podendo ser usado em conjunto com uma defesa da categoria proativa.

5.3 *Adversarial Retraining*

Método de defesa por Madry et al. [2017] da categoria proativa. É uma estratégia de defesa baseada no *adversarial training*, entretanto mais específica nas imagens usadas para o novo treinamento da rede, utilizando as imagens adversárias que foram mais eficientes em enganar o classificador.

É bastante dependente de experimentações e investigação dos resultados para que seja realmente eficiente, devido ao comportamento só ser evidente ao fim do treinamento.

Os autores realizaram diversos testes, encontrando resultados com boa taxa de acerto, por exemplo, em um desses testes foi utilizado um classificador que possuía baixa taxa de acerto contra imagens adversárias geradas utilizando o modelo de ataque PGD, essas imagens foram então utilizadas durante uma nova fase de treinamento. Em seguida esse novo classificador foi colocado novamente contra imagens geradas pelo ataque PGD, dessa vez o ataque foi muito menos eficiente, também foram utilizadas imagens adversárias geradas com outros modelos de ataques que também não obtiveram o mesmo sucesso que tinham com o classificador antigo.

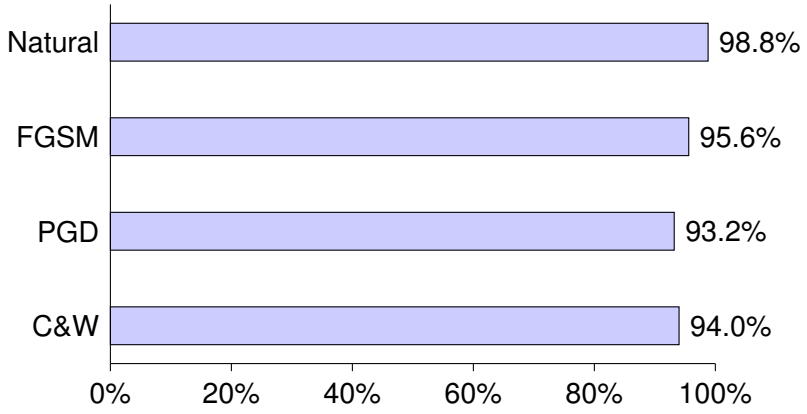


Figura 5.2: Taxa de acerto do *adversarial retraining* utilizando imagens adversárias geradas pelo PGD (*projected gradient descent*) durante a fase de treinamento para o modelo natural, e os ataques FGSM, PGD e C&W L2. Fonte: Madry et al. [2017]

Utilizando o ataque adversário PGD para gerar imagens adversárias e usando-as durante o treinamento do classificador, os autores foram capazes de minimizar os efeitos causados por diversos ataques, mantendo uma taxa de acerto de mais de 90% para todos os casos testados. Outro ponto importante foi que a taxa de acerto para as imagens naturais foi pouco impactada.

Capítulo 6
Conclusão

A existência de exemplos adversários limita as áreas em que *deep learning* pode ser aplicado. Ainda é um problema em aberto se é possível um modelo ser robusto o suficiente para ser considerado seguro, pois defesas fortes atualmente podem ser suscetíveis e vulneráveis a ataques futuros.

Esse trabalho buscou inicialmente apresentar o que são ataques adversários, os categorizando e exibindo alguns exemplos, para em seguida analisar de forma experimental modelos de ataques e defesas, onde foi encontrando diversos exemplos da vulnerabilidade de redes neurais aos exemplos adversários. Por outro lado também foi identificado que existem estratégias de defesas que amenizam o problema, entretanto suas implementações podem afetar a eficácia dos modelos classificadores.

Esses modelos de ataque, estratégias de defesa, a possível troca de eficiência por segurança e a vulnerabilidade dos classificadores às imagens adversárias ainda são descobertas recentes, existindo muito ainda para ser analisado sobre a segurança e robustez das redes neurais.

6.1 Trabalhos Futuros

Observou-se durante os experimentos que os ataques atuais são muito eficazes para modelos sem defesa, e que apenas dados sobre a taxa de acerto não são suficientes para comparar adequadamente os tipos diferentes de ataques. Portanto para trabalhos futuros seria feita uma comparação utilizando outros dados, como quantificar a perturbação realizada na imagem ao comparar com a original.

Seria interessante ainda adicionar mais variações de ataques, realizando comparações entre as diferentes categorias, assim como utilizar mais estratégias de defesas, inclusive em conjunto, para analisar suas eficácias.

Referências

[1] N. Carlini and D. Wagner. Towards Evaluating the Robustness of Neural Networks. *ArXiv e-prints*, August 2016.

[2] N. Carlini, G. Katz, C. Barrett, and D. L. Dill. Provably Minimally-Distorted Adversarial Examples. *ArXiv e-prints*, September 2017.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *ArXiv e-prints*, December 2014.

[4] Simon O. Haykin. *Neural Networks and Learning Machines, 3rd Edition*. Prentice Hall, 2008.

[5] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ArXiv e-prints*, July 2016.

[6] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, S. Huang, Y. Zhao, Y. Zhao, Z. Han, J. Long, Y. Berdibekov, T. Akiba, S. Tokui, and M. Abe. Adversarial Attacks and Defences Competition. *ArXiv e-prints*, March 2018.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv e-prints*, June 2017.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv e-prints*, June 2017.

[9] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *arXiv e-prints*, art. arXiv:1412.1897, December 2014.

[10] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik, and A. Swami. The Limitations of Deep Learning in Adversarial Settings. *ArXiv e-prints*, November 2015.

[11] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami. Practical Black-Box Attacks against Machine Learning. *ArXiv e-prints*, February 2016.

[12] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *ArXiv e-prints*, September 2014.

[14] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv e-prints*, December 2017.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ArXiv e-prints*, December 2013.

[16] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial Examples: Attacks and Defenses for Deep Learning. *ArXiv e-prints*, December 2017.