Generating and defending against adversarial examples in vision-optimized neural architectures

Daniel Donoghue
Email: ddonogh1@jhu.edu

Nicholas Lines
Email: nicholasalines@gmail.com

Arnaldo Pereira
Email: aepereira@gmail.com

***Abstract*—As automated decision-making becomes more popular and more dependent upon artificial intelligence, securing sensitive models from adversarial behavior has become essential. Neural networks are particularly vulnerable to so-called adversarial examples [3], and various attacks and defences have been explored in the literature.**

Our intention in this paper is to demonstrate and confirm the results of such attacks at an informative but modest scale. We apply two common attacks to both the wide ResNet and GoogLeNet neural models, and test two defences, in a reproducible computational environment. We show that significant improvements in network robustness are available with minimal defence measures.

The authors are listed alphabetically, and all made equal contributions. This work is performed in association with the Johns Hopkins Engineering for Professionals Program, as a project for EN.625.638.8VL2.FA20 Neural Networks.

All code and further reference materials are available online at https://github.com/linesn/adversarial_examples.

CONTENTS

I Executive Summary 1

II Project overview 1
- II-A Why are adversarial examples effective? 1
- II-B Attacks 2
- II-C Defences 2

III Computational results 2
- III-A Creating Adversarial Examples 2
- III-B Defences 2

IV Analysis 2

V Conclusions 2

References 2

I. EXECUTIVE SUMMARY

Since 2014 when Szegedy et al [3] published the first observation on the subject, adversarial examples have gained much attention in both the study of adversarial machine learning research and the more results-oriented world of practical neural archetecture, due to the alarming weaknesses they expose and the interesting robustness that can be introduced via defence efforts. The term "adversarial example" is used to describe "an input to a machine learning model that is intentionally designed to cause the model to make a mistake in its predictions, despite resembling a valid input to a human" [5]. As such these examples are classed as evasion techniques by adversarial machine learning theory, since their goal is to evade detection while producing misinterpretations [4].

Most of the literature on the subject (in keeping with traditions in the neural network community) uses image recognition tasks to demonstrate the efficacy of attacks and defences, and we will do the same. In this paper we will demonstrate successful us of the Fast Gradient Sign Method (FGSM) [1] and Directed Gradient Sign (DGSM) [2] attacks against convolutional neural networks trained with Imagenette data. We will then examine the results of applying two common defences: first, perturbed prediciton averaging, and second, training using adversarial examples. We confirm the observations of Goodfellow et al [1] and show that, while the Wide ResNet and GoogLeNet architectures are very susceptible to the above attacks, the named defences also produce significant improvement to the robustness of the classifiers.

II. PROJECT OVERVIEW

A. Why are adversarial examples effective?

In practice neural architectures based on linear components are prefered (over, for example, radial basis components) because of their speed in training and inference. However, it is this property that makes them particularly vulnerable to the most common form of adversarial example [1]. Neural classifier inputs or features naturally have some precision limit, such as a color range or pixel count, below which perturbations are ignored. Consider an input vector $\mathbf{X}$, to which we add a noise vector $\boldsymbol{\eta}$, where every $\eta \in \boldsymbol{\eta}$ is smaller than ϵ, the precision limit. To humans and a first pass review by machines, $\mathbf{X}$ and $\mathbf{X} + \boldsymbol{\eta}$ are identical. However, when the linear activity function is computed, the network's weights are dotted with the input, yielding approximately

$$\mathbf{W}^\intercal(\mathbf{X} + \boldsymbol{\eta}) = \mathbf{W}^\intercal\mathbf{X} + \mathbf{W}^\intercal\boldsymbol{\eta},$$

where the noise term $\mathbf{W}^\intercal\boldsymbol{\eta}$ can grow very large if $\mathbf{W}$ is ill-conditioned. This means, in practice, that networks reliant on linear activity functions can produce extremely different outputs when given only minimally altered inputs.

B. Attacks

The FGSM attack [1] takes advantage of this weakness in a straightforward manner. The attacker forms the perturbation vector $\boldsymbol{\eta}$ to match the cost function gradient sign for a given input, computing

$$\boldsymbol{\eta} = \epsilon_* \text{sign}(\nabla_{\mathbf{X}} J(\boldsymbol{\theta}, \mathbf{X}, y))$$

where ϵ_* is the allowable level of perturbation, J is the network cost funtion, $\boldsymbol{\theta}$ is the vector of model parameters, and y is the true label. Thus, if the attacker is in posession of the model and labeled training data, it is easy to train the network to behave badly using simple backpropagation. The result is that the network will lose certainty in the true label classification, and often misclassify the data at random. The attack parameter ϵ_* may be scaled, of course, but making ϵ_* much larger than the network precision level ϵ may produce examples whose alteration is visible to human reviewers, so smaller ϵ_* are desireable from the adversarial perspective.

One can alter this attack to cause the network to favor a particular class instead [2]. We will call this the Directed Gradient Sign Method (DGSM). This time we use the loss function to direct the network toward a specific desired label. We iterate using gradient descent for a given number of iterations, and project the gradient onto the l_∞-norm ϵ_*-sphere, which has the effect of insisting that a feature is not altered by more than ϵ_*. For a given input $\mathbf{X}$, the attacker must solve the minimization problem

$$\begin{aligned} \min_{\delta}\{J_{adv}(\mathbf{X}+\boldsymbol{\delta}) &= J(\boldsymbol{\theta}, \mathbf{X}+\boldsymbol{\delta}, y_{desired}) \\ &\quad -J(\boldsymbol{\theta}, \mathbf{X}+\boldsymbol{\delta}, y_{true})\} \\ \text{subject to } ||\boldsymbol{\delta}||_\infty &\leq \epsilon_* \end{aligned}$$

where J is again the loss function, $\boldsymbol{\theta}$ the fixed network parameters, $y_{desired}$ and y_{true} are two different labels, and $\boldsymbol{\delta}$ is the directed perturbation vector. This requires using forward passes and backpropagation within the network over N iterations, applying the update rule

$$\begin{aligned} \boldsymbol{\delta}_t &= \boldsymbol{\delta}_{t-1} - a \text{ sign}(\nabla L_{adv}(\mathbf{X}+\boldsymbol{\delta}_{t-1})), \\ \boldsymbol{\delta}_t &\leftarrow \text{clip}(\boldsymbol{\delta}_t, -\epsilon_*, \epsilon_*) \end{aligned}$$

begining with the zero vector $\boldsymbol{\delta}_0 = \mathbf{0}$. The result of this attack is that the classifier will incorrectly favor the chosen label $y_{desired}$ in adversarial inputs, despite remaining perfectly capable of correctly classifying unaltered inputs.

C. Defences

A theme that has emerged in the literature is that there is a strong correlation between generally robust networks and networks that are not easily swayed by adversarial attacks. Of course, one also expects a cost in effort or accuracy to be associated with increased robustness.

Our first defence we test is simply Perturbed Prediction Averaging. This defence has the advantage that it does not require retraining the network, and the only increased expense is the cost of slower decisions at inference time. We predict the class for each image based on an ensemble prediction for the original image and $N-1$ additional perturbed versions of the image, with the perturbations drawn uniformly from the $\epsilon_{\max}$-ball in the l_∞ sense around the image, choosing $\epsilon_{\max}$ to be larger than any expected adversarial alteration level ϵ_*. For example, $\epsilon_{\max}$ can be set large enough that a uniform random $||\boldsymbol{\delta}||_\infty \leq \epsilon_{\max}$ perturbation would be easily noticed by a human. In that case, we can assume that adversarial attacks will rely on $\epsilon_* << \epsilon_{\max}$. Using a modified softmax function, we can express the probability for class k of classes $\{1, 2, ..., K\}$ predicted using this defense as

$$P(y_k|\mathbf{X}) = \frac{\sum_{i=1}^{N} e^{z_k}(\mathbf{X}+\boldsymbol{\delta}_i)}{\sum_{j=1}^{K}\sum_{i=1}^{N} e^{z_j}(\mathbf{X}+\boldsymbol{\delta}_i)},$$

where $z_j(\cdot)$ is the output of the jth hidden node for a given network input $\mathbf{X}$ and with $\boldsymbol{\delta}_i = \mathbf{0}$, and $||\boldsymbol{\delta}||_\infty \leq \epsilon_{\max}$ for all i.

III. Computational results

A. Creating Adversarial Examples

B. Defences

IV. Analysis

V. Conclusions

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, arXiv preprint arXiv:1412.6572 (2014).

[2] Z Madry, *Solving the inner maximization*, Adversarial Examples, 2020, [Online; accessed 11-November-2020].

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, *Intriguing properties of neural networks*, 2014.

[4] Wikipedia contributors, *Adversarial machine learning — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Adversarial_machine_learning, 2020, [Online; accessed 11-November-2020].

[5] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker, *Adversarial examples in modern machine learning: A review*, arXiv preprint arXiv:1911.05268 (2019).