1 Attack Evaluation

333 We compare our targeted attacks to the best results previously reported in prior publications, for each of the three distance metrics.

4444 We re-implement Deepfool, fast gradient sign, and iterative gradient sign.

55555 For fast gradient sign, we search over ϵ to find the smallest distance that generates an adversarial example; failure is returned if no ϵ produces the target class. Our iterative gradient sign method is similar: we search over ϵ (fixing $\alpha = 1$ 256) and return the smallest successful. For JSMA we use the implementation in CleverHans [35] with only slight modification (we improve performance by 50× with no impact on accuracy). JSMA is unable to run on ImageNet due to an inherent significant computational cost:

666666 recall that JSMA performs search for a pair of pixels p, q that can be changed together that make the target class more likely and other classes less likely. ImageNet represents images as 299 × 299 × 3 vectors, so searching over all pairs of pixels would require 236 work on each step of the calculation. If we remove the search over pairs of pixels, the success of JSMA falls off dramatically.

55555 We therefore report it as failing always on ImageNet. We report success if the attack produced an adversarial example with the correct target label, no matter how much change was required. Failure indicates the case where the attack was entirely unable to succeed.

4444 We evaluate on the first 1,000 images in the test set on CIFAR and MNSIT. On ImageNet, we report on 1,000 images that were initially classified correctly by Inception v3. On ImageNet we approximate the best-case and worst-case results by choosing 100 target classes (10%) at random.

The results are found in Table IV for MNIST and CIFAR, and Table V for ImageNet.

4444 For each distance metric, across all three datasets, our attacks find closer adversarial examples than the previous state-of-the-art attacks, and our attacks never fail to find an adversarial example.

55555 Our L_0 and L_2 attacks find adversarial examples with 2× to 10× lower distortion than the best previously published attacks, and succeed with 100% probability. Our L∞ attacks are comparable in quality to prior work, but their success rate is higher. Our L∞ attacks on ImageNet are so successful that we can change the classification of an image to any desired label by only flipping the lowest bit of each pixel, a change that would be impossible to detect visually. As the learning task becomes increasingly more difficult, the previous attacks produce worse results, due to the complexity of the model. In contrast, our attacks perform even better as the task complexity increases. We have found JSMA is unable to find targeted L_0 adversarial examples on ImageNet, whereas ours is able to with 100% success.

4444 It is important to realize that the results between models are not directly comparable.

55555 For example, even though a L_0 adversary must change 10 times as many pixels to switch an ImageNet classification compared to a MNIST classification, ImageNet has 114× as many pixels and so the fraction of pixels that must change is significantly smaller.

22 Generating synthetic digits

333 With our targeted adversary, we can start from any image we want and find adversarial examples of each given target.

Using this, in Figure 6 we show the minimum perturbation to an entirely-black image required to make it classify as each digit, for each of the distance metrics.

333 This experiment was performed for the L_0 task previously [38], however when mounting their attack, "for classes 0, 2, 3 and 5 one can clearly recognize the target digit." With our more powerful attacks, none of the digits are recognizable.

Figure 7 performs the same analysis starting from an all-white image.

333 Notice that the all-black image requires no change to become a digit 1 because it is initially classified as a 1, and the all-white image requires no change to become an 8 because the initial image is already an 8.

22 Runtime Analysis.

333 We believe there are two reasons why one may consider the runtime performance of adversarial example generation algorithms important:

4444 first, to understand if the performance would be prohibitive for an adversary to actually mount the attacks, and second, to be used as an inner loop in adversarial re-training [11]. Comparing the exact runtime of attacks can be misleading.

55555 For example, we have parallelized the implementation of our L_2 adversary allowing it to run hundreds of attacks simultaneously on a GPU, increasing performance from 10× to 100×. However, we did not parallelize our L_0 or L∞ attacks. Similarly, our implementation of fast gradient sign is parallelized, but JSMA is not.

4444 We therefore refrain from giving exact performance numbers because we believe an unfair comparison is worse than no comparison.

333 All of our attacks, and all previous attacks, are plenty efficient to be used by an adversary. No attack takes longer than a few minutes to run on any given instance. When compared to L_0, our attacks are 2 × −10× slower than our optimized JSMA algorithm (and significantly faster than the un-optimized version). Our attacks are typically 10× −100× slower than previous attacks for L_2 and L∞, with exception of iterative gradient sign which we are 10× slower.

commentary

If you haven't already, read my post for part 6 of *Message in Text*. There you will find a general discussion of the hierarchical structure of Message, and you'll also see the wavelength of Section III on display, as above you see the wavelength of Section VII. Because now, briefly, I want to contrast the wavelengths of each section, because the differences tell us a lot about how Message is structured in sections for methods (as in Section III) and in sections for evaluation (as in Section VII).

The wavelength of Section III is tuned to a high frequency. That is to say, the units of content oscillate mostly between 333 and 4444, and there are four crests at 22. The reason for such high-frequency transmission of the content is that there are many methods to cover in a short space. Besides, for the purposes of this paper, the methods require only a cursory treatment. Carlini and Wagner want merely to give a close enough appreciation of each method so that the reader will understand how Carlini and Wagner have adapted it to their own research purposes. Therefore, the methods are just touched upon, each in turn, and only partial detail is provided on any one.

The wavelength of Section VII is very different, and that's because the purpose of Section VII is very different too.

In VII, Carlini and Wagner are not covering many things on the surface, but instead, they are covering few things in depth. This is a section for evaluation. The authors want to explain the entire significance of their attack results, and neither those results nor their interpretations of those results can be found in any other paper. Therefore, Carlini and Wagner cover points of content as deeply as 666666. Moreover, although Sections III and VII have roughly the same word counts, the two sections differ significantly in how that word-material is deployed. In Section III, the content is handled at just seven levels which are located at 4444 or lower. In Section VII, by contrast, the number is a full twelve points of content handled at levels 4444 or lower. That is nearly double the detail in VII, and no wonder: **The purpose of the section is to interpret, exhaustively and convincingly, the significance of Carlini and Wagner's attack results.**

One added complication in Section VII, again in contrast to Section III, is the use by Carlini and Wagner of figures and tables. In my parse above, I mark in red and I also range to the left margin all those clauses which refer the reader to a figure or table. This is my attempt to capture the jump in structure represented by a reader's shifting between the textual modality and the visual modality. Because essentially, in those moments, the flow of the reading gets swiftly redirected outside the discourse. However, the reader will need to find his or her place again; in fact, the reader will need even to be motivated to return to the discourse in the text.

Carlini and Wagner do a fine job at this. Carlini and Wagner guide their readers masterly by (a) generally spacing the visual cross-references so that the reader isn't rushed about the text, here and then there, and they guide by (b) always maintaining the same level of hierarchy both *before* a cross-reference *and after*. For example, the cross-references to Tables IV and V are both preceded and likewise followed by the hierarchical level of 4444. This greatly assists a reader in retrieving his or her thread in the discourse.

Towards Evaluating the Robustness of Neural Networks is a great paper for good reason. In almost every detail of the text, Carlini and Wagner proves themselves to be highly skilled writers of high-impact findings.