1 ATTACK ALGORITHMS

22 L-BFGS

333 Szegedy *et al.* [46] generated adversarial examples using box-constrained L-BFGS.

4444 Given an image x, their method finds a different image x' that is similar to x under L_2 distance, yet is labeled differently by the classifier. They model the problem as a constrained minimization problem [mathematical problem] This problem can be very difficult to solve, however, so Szegedy *et al.* instead solve the following problem: [mathematical problem] where $\text{loss}_{F,l}$ is a function mapping an image to a positive real number. One common loss function to use is cross-entropy. Line search is performed to find the constant $c > 0$ that yields an adversarial example of minimum distance: in other words, we repeatedly solve this optimization problem for multiple values of c, adaptively updating c using bisection search or any other method for one-dimensional optimization.

22 Fast Gradient Sign

333 The fast gradient sign [11] method has two key differences from the L-BFGS methods: first, it is optimized for the L_∞ distance metric, and second, it is designed primarily to be fast instead of producing very close adversarial examples.

4444 Given an image x the fast gradient sign method sets [mathematical problem] where ϵ is chosen to be sufficiently small so as to be undetectable, and t is the target label. Intuitively, for each pixel, the fast gradient sign method uses the gradient of the loss function to determine in which direction the pixel's intensity should be changed (whether it should be increased or decreased) to minimize the loss function; then, it shifts all pixels simultaneously.

333 It is important to note that the fast gradient sign attack was designed to be *fast*, rather than optimal. It is not meant to produce the minimal adversarial perturbations. Iterative Gradient Sign: Kurakin *et al.* introduce a simple refinement of the fast gradient sign method [26] where instead of taking a single step of size ϵ in the direction of the gradient-sign, multiple smaller steps α are taken, and the result is clipped by the same ϵ.

4444 Specifically, begin by setting [mathematical problem] and then on each iteration [mathematical problem].

333 Iterative gradient sign was found to produce superior results to fast gradient sign [26].

22 JSMA

333 Papernot *et al.* introduced an attack optimized under L_0 distance [38] known as the Jacobian-based Saliency Map Attack (JSMA). We give a brief summary of their attack algorithm; for a complete description and motivation, we encourage the reader to read their original paper [38].

4444 At a high level, the attack is a greedy algorithm that picks pixels to modify one at a time, increasing the target classification on each iteration. They use the gradient [mathematical definition] to compute a saliency map, which models the impact each pixel has on the resulting classification. A large value indicates that changing it will significantly increase the likelihood of the model labeling the image as the target class 1. Given the saliency map, it picks the most important pixel and modify it to increase the likelihood of class 1. This is repeated until either more than a set threshold of pixels are modified which makes the attack detectable, or it succeeds in changing the classification.

55555 In more detail, we begin by defining the saliency map in terms of a pair of pixels p,q. Define [mathematical problem] so that α_{pq} represents how much changing both pixels p and q will change the target classification, and β_{pq} represents how much changing p and q will change all other outputs. Then the algorithm picks [mathematical problem] so that $\alpha_{pq} > 0$ (the target class is more likely), $\beta_{pq} < 0$ (the other classes become less likely), and $-\alpha_{pq} \cdot \beta_{pq}$ is largest.

4444 Notice that JSMA uses the output of the second-to-last layer Z, the logits, in the calculation of the gradient: the output of the softmax F is *not* used. We refer to this as the JSMA-Z attack. However, when the authors apply this attack to their defensively distilled networks, they modify the attack so it uses F instead of Z. In other words, their computation uses the output of the softmax (F) instead of the logits (Z). We refer to this modification as the JSMA-F attack. When an image has multiple color channels (e.g., RGB), this attack considers the L_0 difference to be 1 for each color channel changed independently (so that if all three color channels of one pixel change change, the L_0 norm would be 3).

333 While we do not believe this is a meaningful threat model, when comparing to this attack, we evaluate under both models.

22 Deepfool

333 Deepfool [34] is an untargeted attack technique optimized for the L_2 distance metric. It is efficient and produces closer adversarial examples than the L-BFGS approach discussed earlier.

4444 The authors construct Deepfool by imagining that the neural networks are totally linear, with a hyperplane separating each class from another. From this, they analytically derive the optimal solution to this simplified problem, and construct the adversarial example. Then, since neural networks are not actually linear, they take a step towards that solution, and repeat the process a second time. The search terminates when a true adversarial example is found.

333 The exact formulation used is rather sophisticated; interested readers should refer to the original work [34].

commentary

Turn the above parse 90° clockwise, and you'll see the wavelike structure of Message. You begin high at the crest of 1, because that is the topic of the entire Section III. From there you descend into troughs at depths five times the heigh of the crest at 1. However, you also ascend back up to 22, back up to 22, and again back up to 22. This undulation is quite regular, rolling most of the way down to 4444 and back up to 333.

Now it is important to note that although this structuring force of the Message is caused to undulate by each and every Theme in the clauses of this text, the waves of the Message here are not strictly determined by the grammar of the clause. The reason is quite simple: The grammar of the clause is limited exclusively to the clause itself. In other words, the grammar of English really stops at the verb, and above that ceiling of the grammatical structuring, other things must take over the function of producing a coherent text of this Section III. One such thing is the hierarchization of Message.

Message above the clause, or if you prefer, above the verb — Message at this height is structured as a hierarchy.

For example, in Section III above, you read about the high-level differences between fast gradient sign and L-BFGS *before* you learn the particulars of how fast gradient sign itself works. Then you are told about the design features of fast gradient sign and about one refinement to those features, *again before* you learn the mathematics of the method. And to conclude, you read about (and are able thus to appreciate) the superiority of the L-BFGS refinement over the L-BFGS original. *That* is the hierarchical structure of a well-laid Message.

Carlini and Wagner have not just dumped content on the page. No, they have a signal that they are sending out, and that signal transmits at varying frequencies, all according to which place in the text the signal is sending from, and as well, all according to which item in the content the signal is sending through.

For example, consider the JSMA of Papernot *et al*. At 333, the signal being sent across the topic of JSMA is that the method will be properly understood only by reading the original paper. Then, at 4444, the signal across the same topic is that a general appreciation of JSMA is indeed

possible, at least for purposes of this paper, *Towards Evaluating the Robustness of Neural Networks*. Next, the choice of the words *in more detail* intensifies the transmission so that at 55555, the output of JSMA becomes the focus. This signal across the output of JSMA lengthens so that, at 4444, Carlini and Wagner can repurpose that output for their own modifications to JSMA. Then finally, back out at 333, the signal reattains the same long reach as the opening lines of the subsection, and Carlini and Wagner simply state why they evaluate under *both* their modified versions of JSMA.

So, there you have a well-tuned Message in one of the best papers written in your community. Examine above how Carlini and Wagner do it, and look out for my parse again of the hierarchy of Message in this paper, in Section VII, *Attack Evaluation*.