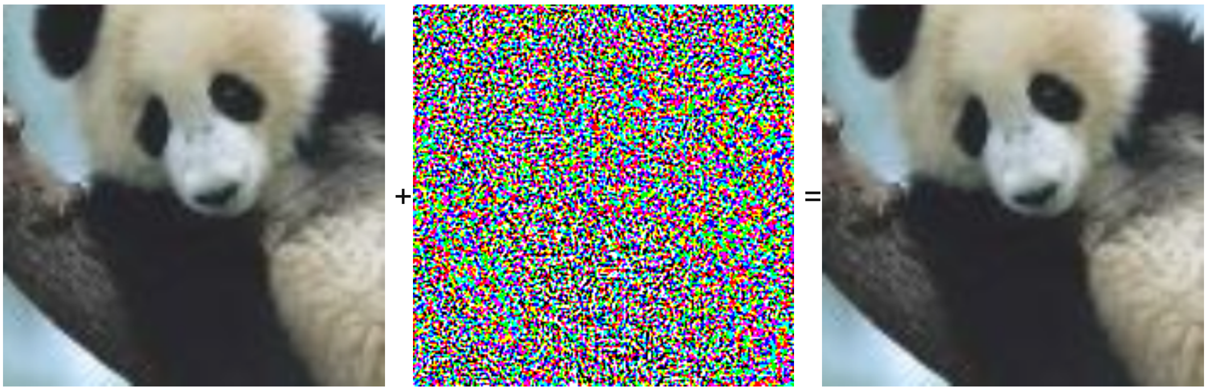
Gregory Piatetsky, Editor: earlier KDnuggets post by Zachary Lipton (Deep Learning's Deep Flaws)'s Deep Flaws led to interesting discussion with Yoshua Bengio (one of leaders of the Deep Learning field), and Ian Goodfellow (Yoshua's student, now a Google Research scientist), but that discussion was buried in the comments. I have asked Ian to expand upon his comments and his work on adversarial examples for KDnuggets readers, and he kindly agreed - here is his post.

Until recently, nearly any input could fool an object recognition model. We were more surprised when object recognition worked than when it didn't. Today, object recognition algorithms have reached human performance as measured by some test set benchmarks, and we are surprised that they fail to perform as well on unnatural inputs. Adversarial examples are synthetic examples constructed by modifying real examples slightly in order to make a classifier believe they belong to the wrong class with high confidence. Rubbish class examples (such as fooling images) are pathological examples that the model assigns to some class with high confidence even though they should not belong to any class.



Original image classified as a panda with 60% confidence.

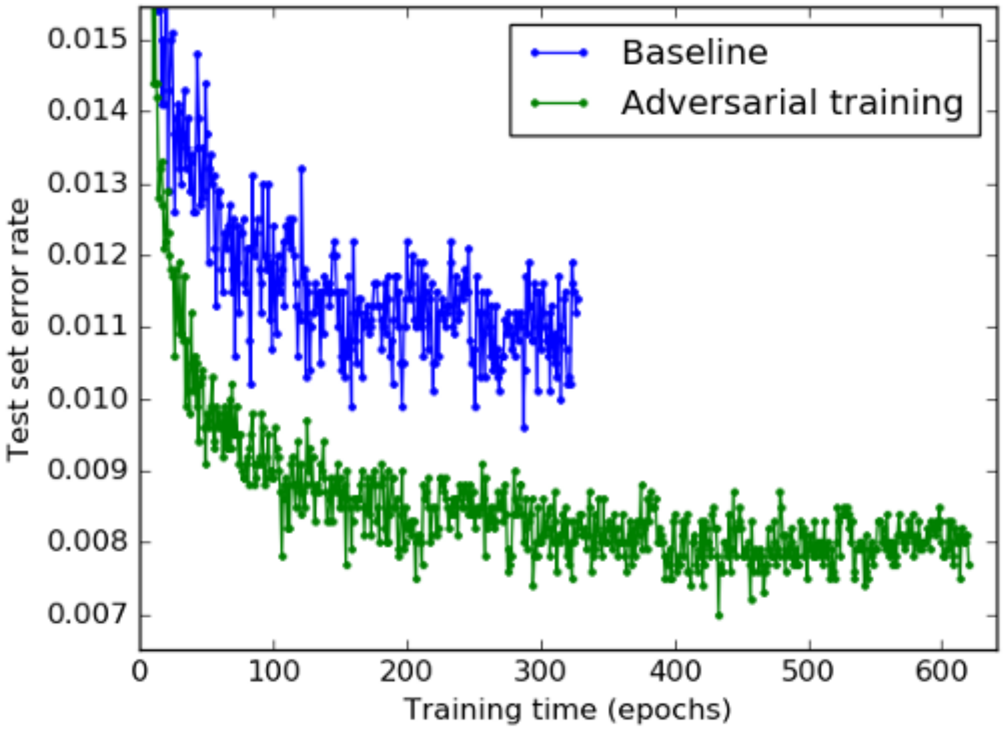
Tiny adversarial perturbation.

Imperceptibly modified image, classified as a gibbon with 99% confidence.

Fig 1. An *adversarial example* constructed by modifying this picture of a panda so that a machine learning model thinks it is a gibbon.
The modification is performed on 32-bit floating point values used as input to the network, and is so small that it does not change the 8-bit representation of the image used for publication. See this paperfor details.

These mistakes have captured the public imagination. In the excitement, some misconceptions about adversarial examples have become widespread. In this blog post, I address some of these misconceptions.

1. **Myth: Adversarial examples do not matter because they do not occur in practice.**
   Fact: It's true that adversarial examples are very unlikely to occur naturally. However, adversarial examples matter because training a model to resist them can improve its accuracy on non-adversarial examples. Adversarial examples also *can* occur in practice if there really is an adversary - for example, a spammer trying to fool a spam detection system.



Training a network to correctly classify adversarial examples reduces its error rate on the test set - even though the test set examples are not perturbed. This technique improved the state of the art on the MNIST dataset.

2. **Myth: Deep learning is more vulnerable to adversarial examples than other kind of machine learning.**
   Fact: So far we have been able to generate adversarial examples for every model we have tested, including simple traditional machine learning models like nearest neighbor. Deep learning with adversarial training is the *most resistant* technique we have studied so far.
3. **Myth: Adversarial examples are due to the extreme non-linearity of deep models.**
   Fact: Our latest experiments suggest that deep models behave too linearly. Linear models become excessively confident when asked to extrapolate far from the training data. This explains many of the mistakes made on adversarial and rubbish class examples.

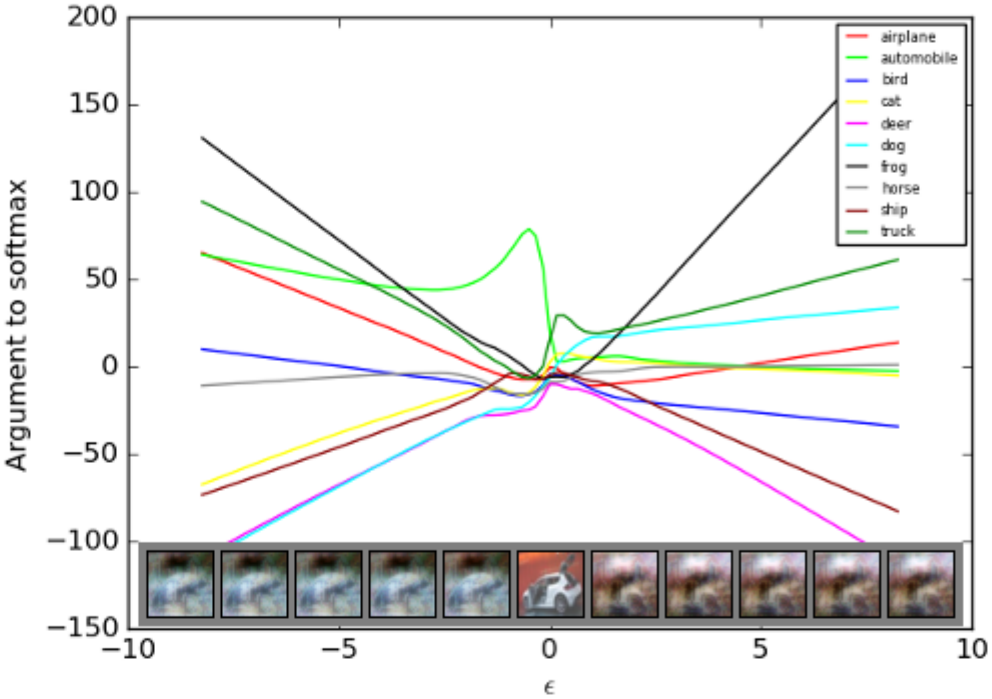


Fig 2. We can trace out a linear path in input space by adding an adversarial perturbation scaled by differing amounts to a clean image of a car. Here we follow the linear path from a scaling factor of negative 10 to positive 10. We see that the logits output by the network behave linearly far from the data. This causes the network's predictions to become extreme, resulting in rubbish class inputs being classified as real classes with high confidence.

4. **Myth: Adversarial examples are hard to find, occurring in small pockets.**
   Fact: *Most* arbitrary points in space are misclassified. For example, one network we tested classified roughly 70% of random noise samples as being horses with high confidence.
5. **Myth: The best we can do is identify and refuse to process adversarial examples.**
   Fact: Refusing to process an adversarial example is better than misclassifying it, but not a satisfying solution. When there truly is an adversary, such as a spammer, the adversary would still gain an advantage by producing examples our system refused to classify. We know it is possible correctly classify adversarial examples because people are not confused by them, and that should be our goal.
6. **Myth: An attacker must have access to the model to generate adversarial examples.**
   Fact: Adversarial examples generalize across models trained to perform the same task, even if those models have different architectures and were trained on a different training set. This means an attacker can train their own model, generate adversarial examples against it, and then deploy those adversarial examples against a model they do not have access to.
7. **Myth: Adversarial examples could easily be solved with standard regularization techniques.**
   Fact: We have unsuccessfully tested several traditional regularization strategies, including averaging across multiple models, averaging across multiple glimpses of an image, training with weight decay or noise, and classifying via inference in a generative model.

8. **Myth: No one knows whether the human brain makes similar mistakes.**
   Fact: Neuroscientists and psychologists routinely study illusions and cognitive biases. Even though we do not have access to our brains' "weights," we can tell we are not affected by the same kind of adversarial examples as modern machine learning. If our brains made the same kind of mistakes as machine learning models, then adversarial examples for machine learning models would be optical illusions for us, due to the cross-model generalization property.

In conclusion, adversarial examples are a recalcitrant problem, and studying how to overcome them could help us to avoid potential security problems and to give our machine learning algorithms a more accurate understanding of the tasks they solve.

Bio: Ian Goodfellow is a Research Scientist at Google. He received Ph.D. in machine learning in 2014 from U. Montreal where he was in Yoshua Bengio group, and BS/MS from Stanford.