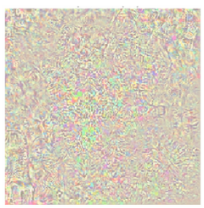
Deep Learning

Adversarial Examples

"golden retriever"
80.6 % confidence



added perturbation
(zoomed in by a factor of 10)

"plane"
99 % confidence

Learning goals

- Adversarial robustness
- Adversarial examples
- Targeted attacks

ADVERSARIAL ROBUSTNESS

- It is critical to examine if a trained neural net is robust and reliable.
- **Adversarial robustness** of a model means that a model is robust to (test time) perturbations of its inputs.
- **Adversarial machine learning** studies techniques which attempt to fool machine learning models through malicious input.
- To make a model more robust, we can train our model on adversarially perturbed examples, called **adversarial examples**, derived from the training set.
- This chapter summarizes high-level ideas in adversarial robustness with a particular emphasis on adversarial examples.
- For a deeper dive, ▸ click here.

Adversarial Examples

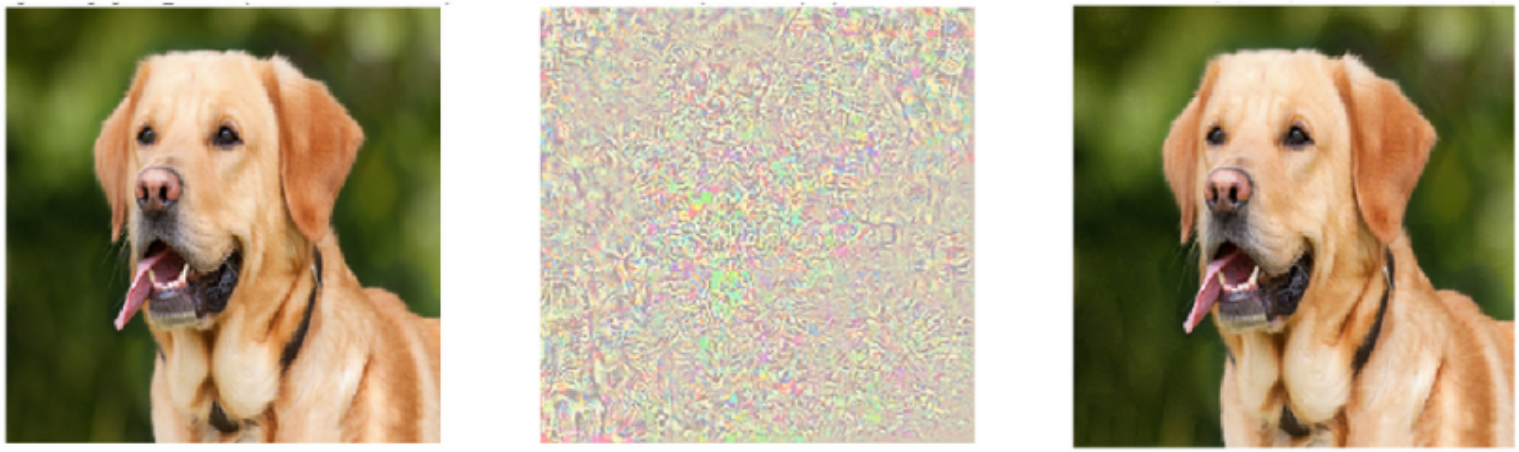
ADVERSARIAL EXAMPLES

- An adversarial example is an input to a model that is deliberately designed to "fool" the model into misclassifying it.
- The test error of a model is only an indicator of how well the model performs with respect to samples from the data-generating distribution.
- The performance of the same model can be drastically different on samples from a completely different distribution (on the same input space).
- It is possible to make changes to an image that makes a pretrained CNN (for example) output a completely different predicted class even though the change is imperceptible to the human eye.
- These examples suggest that even models that have very good test set performance do not have a deep understanding of the underlying concepts that determine the correct output label.

ADVERSARIAL EXAMPLES

- Adversarial examples are *not* unique to deep neural nets. Many other models (such as logistic regression) are also susceptible to them.
- They pose serious security concerns in many areas.
- Example: Fooling autonomous cars into thinking that a stop sign is a 45 km/h sign.
- Example: Evading law enforcement by fooling facial recognition systems into misidentifying individuals.

ADVERSARIAL EXAMPLES



"golden retriever"
80.6 % confidence

added perturbation
(zoomed in by a factor of 10)

"plane"
99 % confidence

Figure: The difference between the left and the right golden retriever is imperceptible to humans. The last image was classified as a plane by ResNet50 with more than 99% confidence. The only difference between the left and right image are small pixel perturbations, shown in the second picture (Prevost, 2018).

ADVERSARIAL EXAMPLES

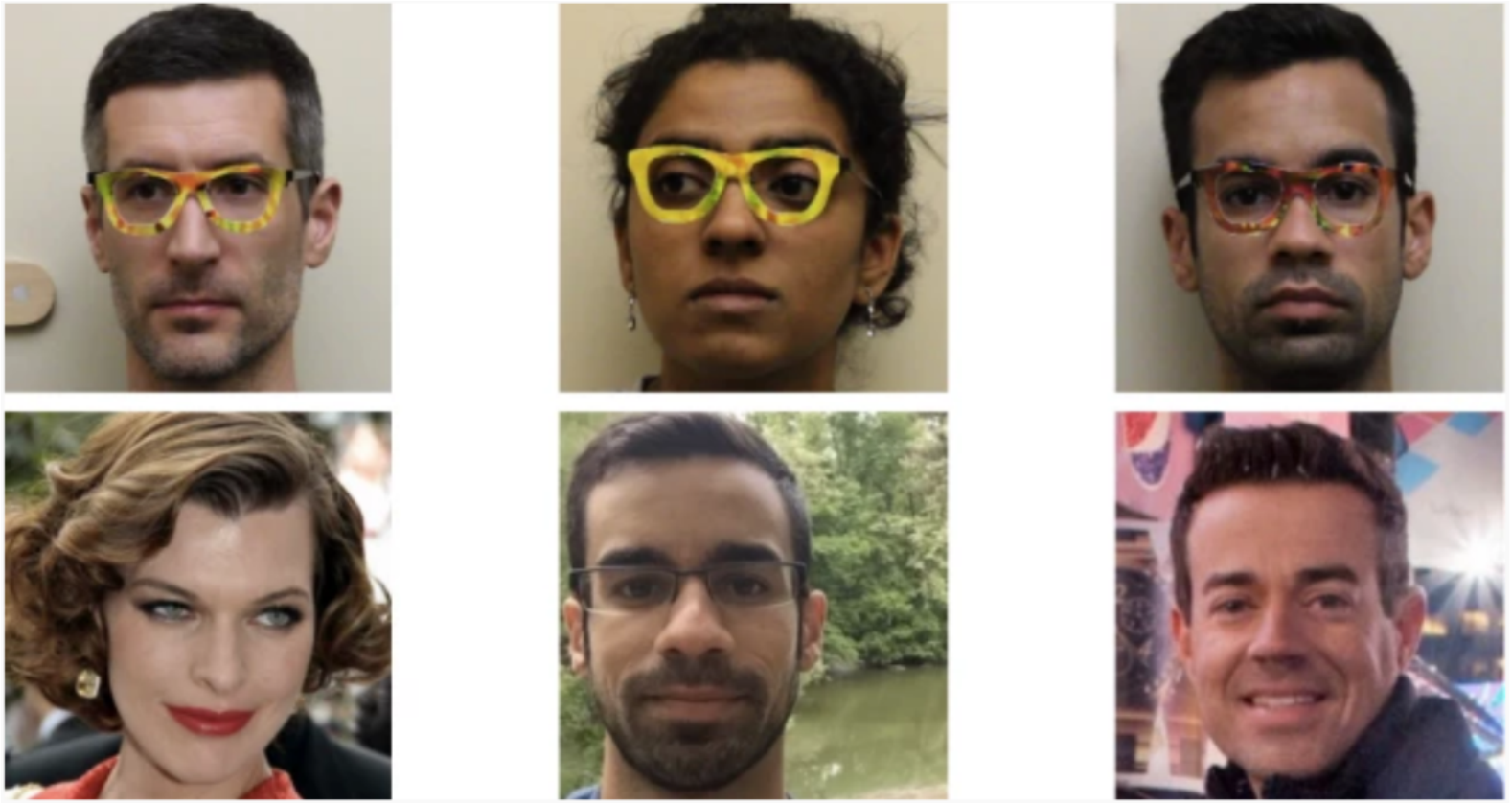


Figure: A CNN misidentified each person in the top row (with the funky looking "adversarial" glasses) as the one in the corresponding position in the bottom row. The generated images do often contain some features of the target class (Sharif et al., 2016).

ADVERSARIAL EXAMPLES

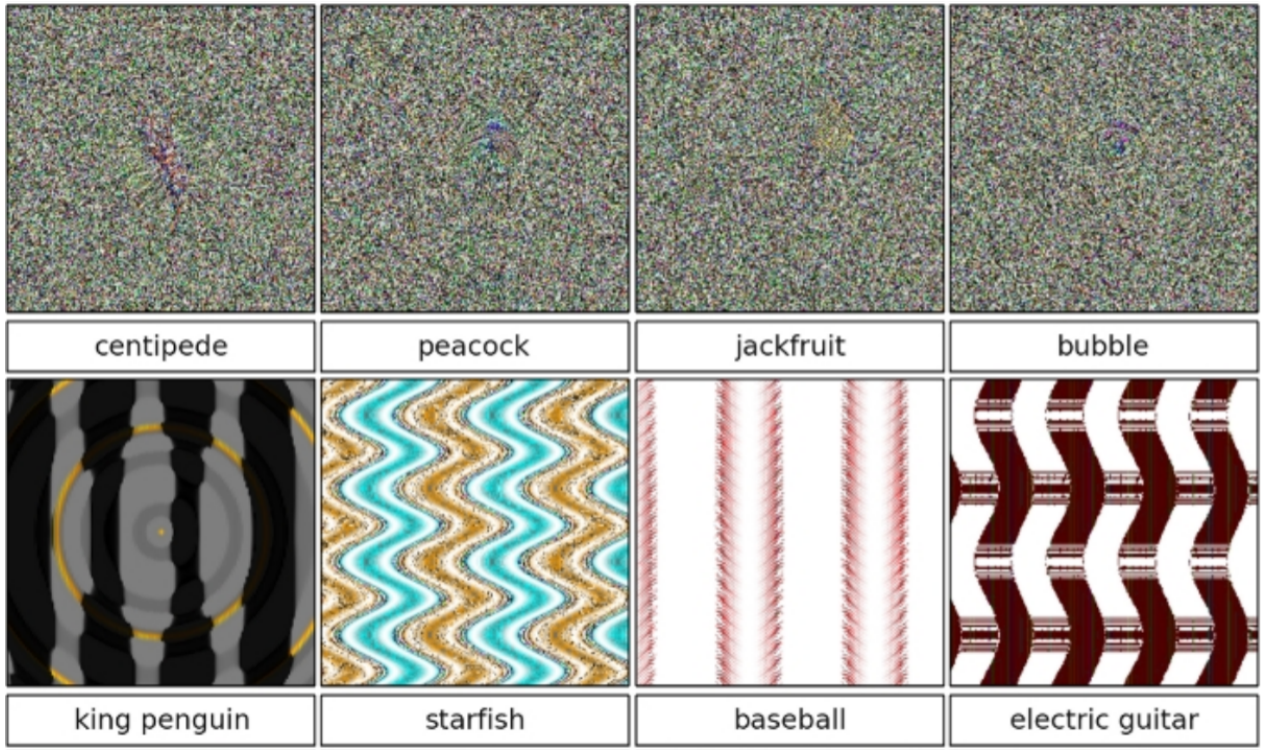


Figure: All 8 images above are unrecognizable to humans but are misclassified by a CNN with higher than 99% confidence. The CNN was trained by Krizhevsky et al. on the ImageNet dataset and consists of five convolutional layers, some of which are followed by max-pooling layers,and three fully-connected layers with a final softmax (Nguyen et al., 2015).

ADVERSARIAL EXAMPLES

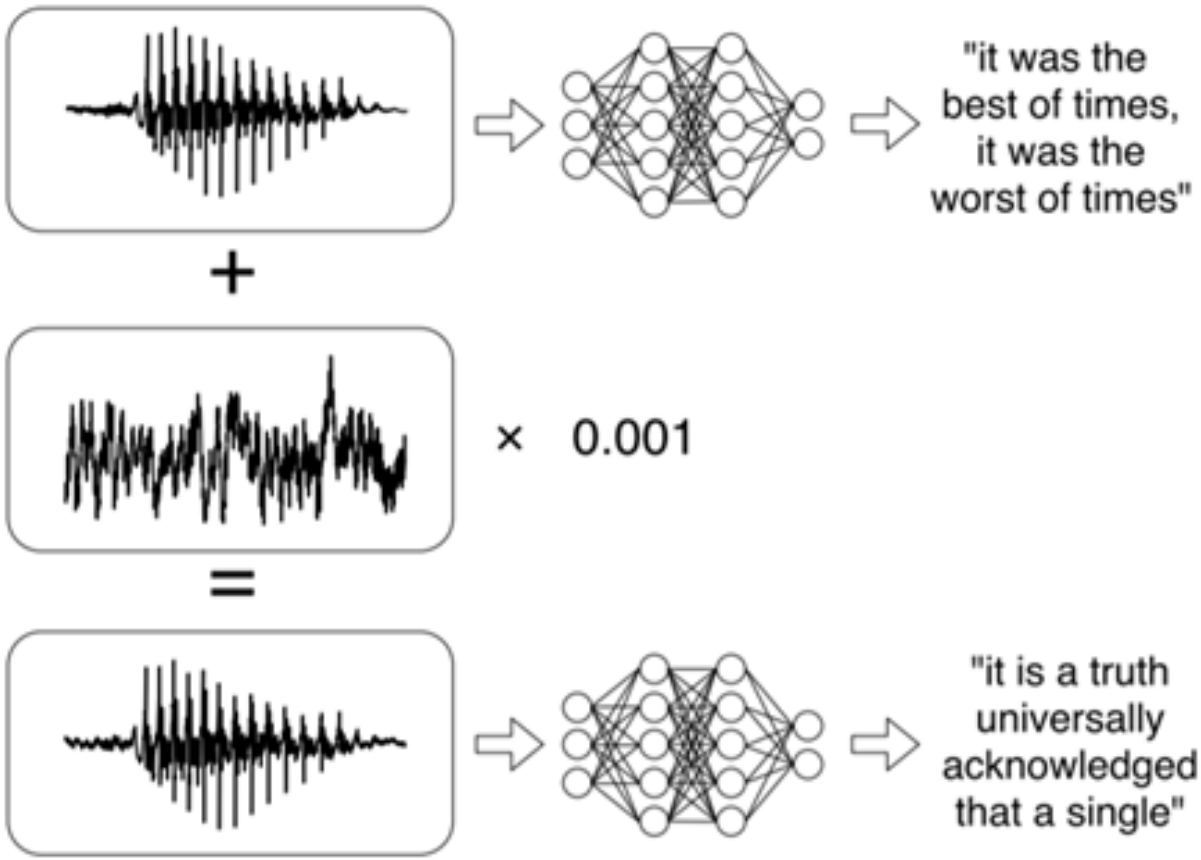


Figure: It is possible to add a small perturbation to any waveform in order to fool a speech-to-text neural network into transcribing it as any desired target phrase (Carlini et al.).

CREATION OF ADVERSARIAL EXAMPLES

- In the examples earlier, we saw that adversarial examples can seem recognizable to humans or seem like random noise/patterns.
- In the following, given a datapoint $\mathbf{x}$, we want to create an adversarial example $\tilde{\mathbf{x}}$ that is very similar to $\mathbf{x}$.
- Specifically, our goal is to find an input $\tilde{\mathbf{x}}$ close to the datapoint $\mathbf{x}$ such that a pretrained model (which accurately classifies $\mathbf{x}$), ends up misclassifying $\tilde{\mathbf{x}}$.
- When we train a neural network, we typically want to optimize the parameter $\boldsymbol{\theta}$, so that we minimize the loss.
- By contrast, to find an adversarial example $\tilde{\mathbf{x}}$ that is in the vicinity of $\mathbf{x}$, we want to optimize the *input* to *maximize* the loss.

CREATION OF ADVERSARIAL EXAMPLES

- To ensure that $\tilde{\mathbf{x}}$ is close to $\mathbf{x}$, we optimize over the perturbation of $\mathbf{x}$, denoted as $\boldsymbol{\delta}$, and define an feasible set of perturbations Δ.

$$\arg\max_{\boldsymbol{\delta} \in \Delta} L(y, \hat{f}(\mathbf{x} + \boldsymbol{\delta} | \boldsymbol{\theta}))$$

- A common perturbation set is $\mathcal{B}_\epsilon^\infty$ which is the ϵ-ball measured by $\ell_\infty = \| \cdot \|_\infty$

$$\Delta = \mathcal{B}_\epsilon^\infty(\boldsymbol{\delta}) = \{\boldsymbol{\delta} : ||\boldsymbol{\delta}||_\infty \leq \epsilon\} \text{ with } ||\boldsymbol{\delta}||_\infty = \max_i |\delta_i|$$

  It allows each component of the perturbation $\boldsymbol{\delta}$ to lie between $-\epsilon$ and $+\epsilon$.
- In general, Δ can also depend on the input datapoint $\mathbf{x}$, denoted as $\Delta(\mathbf{x})$.

EXAMPLE: RESNET50

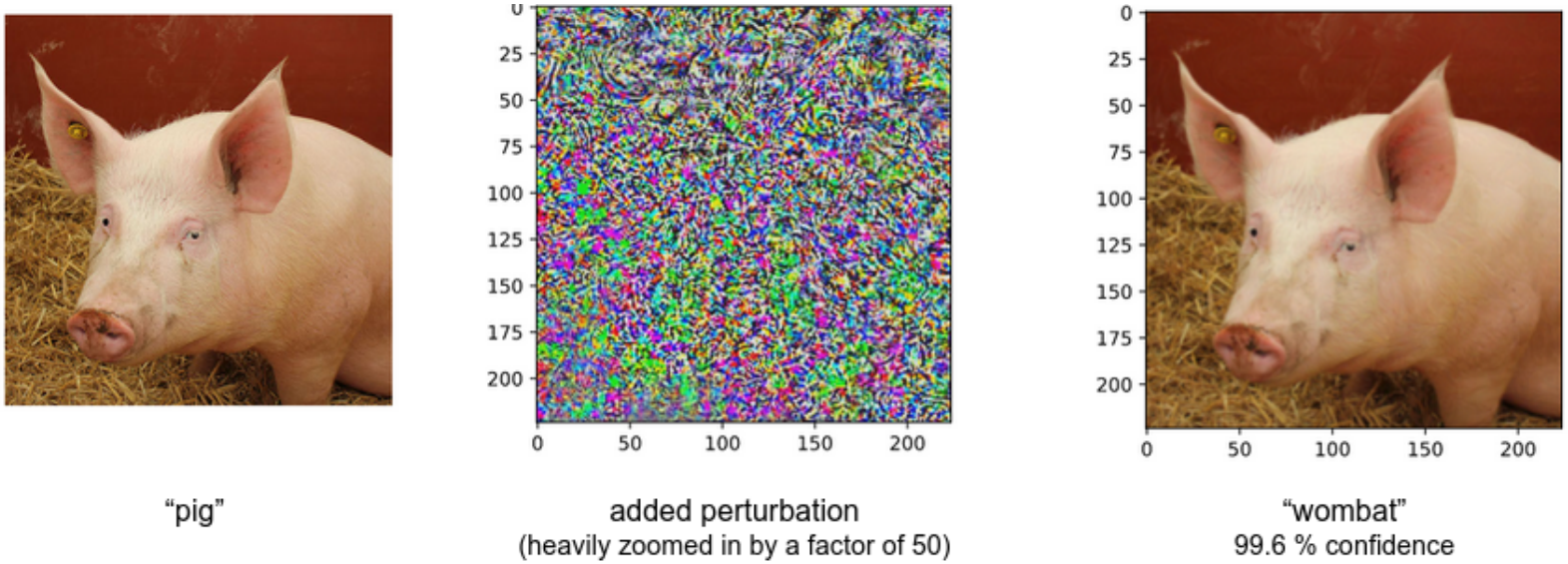


Figure: Adversarial example for one datapoint of the ImageNet dataset and pre-trained ResNet50. By adding an imperceptibly small perturbation to the original image, an image was created that looks identical to our original image, but is misclassified (Kolter & Madry, 2019).

TARGETED ATTACKS

- It is also possible to generate adversarial examples classified virtually as any desired class. This is known as a **targeted attack**.
- The only difference is that, instead of trying to just maximize the loss of the correct class, we maximize the loss of the correct class while also minimizing the loss of a target class y_{target}.

$$\arg\max_{\delta \in \Delta} (L(y, \hat{f}(\mathbf{x} + \delta | \theta)) - L(y_{target}, \hat{f}(\mathbf{x} + \delta | \theta)))$$

REFERENCES

Zico Kolter and Aleksander Madry (2019)

Adversarial Robustness - Theory and Practice

https://adversarial-ml-tutorial.org/

Ian Goodfellow, Yoshua Bengio and Aaron Courville (2016)

Deep Learning

http://www.deeplearningbook.org/

Ian Goodfellow (2017)

Lecture 16 | Adversarial Examples and Adversarial Training

https://www.youtube.com/watch?v=CIfsB_EYsVI

Ian Goodfellow, Nicolas Papernot, Sandy Huang, Rocky Duan, Pieter Abbeel, Jack Clark (2017)

Attacking Machine Learning with Adversarial Examples

https://openai.com/blog/adversarial-example-research/

REFERENCES

Anh Nguyen, Jason Yosinski and Jeff Clune (2015)

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

https://arxiv.org/abs/1412.1897

Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinto (2012)

ImageNet Classification with Deep Convolutional Neural Networks. NIPS.

https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

Maxence Prevost (2018)

Adversarial ResNet50

http://arxiv.org/abs/1207.0580

Mahmood Sharif, Sruti Bhagavatula and Lujo Bauer (2016)

Accessorize to a Crime: Real and Stealthy Attacks onState-of-the-Art Face Recognition. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.

https://dl.acm.org/doi/10.1145/2976749.2978392

REFERENCES

Goodfellow, Shlens (2014)

Explaining and Harnessing Adversarial Examples

https://github.com/maxpv/maxpv.github.io/blob/master/notebooks/Adversarial_ResNet50.ipynb

Papernot , McDaniel, Goodfellow, Jha, Celik, Swamy (2016)

Practical Black-Box Attacks against Machine Learning

https://arxiv.org/abs/1602.02697

Athalye , Engstrom, Ilyas, Kwok (2017)

Synthesizing Robust Adversarial Examples

https://arxiv.org/abs/1707.07397

Tom B. Brown and Catherine Olsson, Research Engineers, Google Brain Team (2018)

Introducing the Unrestricted Adversarial Examples Challenge

https://ai.googleblog.com/2018/09/introducing-unrestricted-adversarial.html