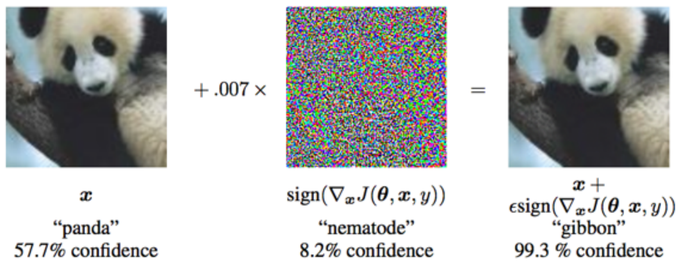
Applying adversarial examples to neural language modeling

Onkur Sen (with Isaac Caswell and Allen Nie)

Overview of adversarial learning

- Idea: add noise to correctly-classified example to create slightly-perturbed, incorrectly-classified example with high confidence



- Illustrates the sensitivity of classifiers
- Training with adversarial examples makes classifiers more robust

Our approach: apply to language models

- Model: mean-pooling RNN with LSTM
- CNN implemented but untested
- Modify objective to simulate training with adversarial examples
- Generate adversarial examples by perturbing training examples

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1 - \alpha) J(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon\text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right))$$

Summary of results

- Sometimes it works better, sometimes it doesn't
- Best successes: large perturbations! (ϵ = 0.5 performs 1-2% better)
- Massive hyperparameter search needed: is our **best** model better than **best** vanilla model?

Nearest neighbor visualization of adversarial example

Original:

```
this <UNK> guy is a real genius ! the movie is of excellent quality and both entertaining and <UNK> . &lt; br / &gt; &lt; br / &gt; i didn &apos;t know what a - girl was before i learned it here . - - - - - - - - - - - - - - - - - - - - - - - -
```

Adversarial:

```
speech <UNK> pretentious is a horse genius ! the age is of excellent quality and both entertaining and <UNK> . &lt; br / &gt; &lt; br / &gt; i horse &apos;t know what a horse girl was before i learned it here . horse horse horse horse horse horse horse horse horse horse horse horse horse horse horse horse …
```

Perturbation at word level

```
"the"     => "the"
"world"   => "world"
"is"      => "is"
"so"      => "movie"
"full"    => "full"
"of"      => "movie"
"a"       => "a"
"number"  => "movie"
"of"      => "movie"
"things"  => "things"
```

NN of perturbation to pairwise word distances

```
<UNK>-role <UNK>-role <UNK>-role <UNK>-role <UNK>-role ...
```

Next steps

- Better visualization of adversarial sentences, incorporating language model
- Create adversarial examples directly?

Special thanks: Jon Gauthier