IN5400 Week 12: create attacks yourself

Alex

March 25, 2021

These homeworks are smaller in scope as I understand that you may want to let things go slower during Easter time. Besides having had a mandatory exercise just recently. You will get a bigger GAN homework in April :).

1 Task – do a simple attack by yourself

This is Mr. Shout.

- Implement the Algorithmically Improved Fooling with solutions for A, B and at least the simple way to deal with discretization
- make him a Strawberry or any other class you like.
- you can call .backward() on any 1-dimensional tensor
- what flags you need to set so that you can compute gradients with respect to the image?
- synset_words.txt tells you which index in imagenet is which class

- you need a pretrained neural net, a way to load the image into a pytorch tensor. Note that involves loading images into PIL/numpy, converting them to a pytorch tensor in shape $(1, 3, h, w)$. That will include swapping color and spatial axes $(h, w, c) \rightarrow (1, c, h, w)$, preprocessing it by the neural network preprocessing rules (mean/standard deviation). You do not need datasets or data loaders. You can apply pytorch transforms for that, for example.

- you need a way to clip the tensor into integer value and to save it back as an image, this involves undoing of: swapping color and spatial axes , preprocessing it by the neural network preprocessing rules (mean/standard deviation) . `PIL.Image` is one option for the actual numpy to image step.

2 Task2 Boundary attack from Foolbox

In order to get a feeling for speed differences between gradient-guided whitebox and blackbox attacks, run from foolbox a boundary attack. Of course, while the latter are slower, they can be applied to models where you only know the predicted labels, and lack of speed is a tradeoff for the lack of information.
Thats a 10 minutes issue to do.

Happy Easter holidays