BEP Adversarial Image Annotation

Goal

General

Goal of the BEP is to generate a new benchmark tool for NLP. Preferably one that is adversarial as it is less subseptibal to overfitting, and can find weaknesses in current state of the art methods.

Specific

I want to create an adversarial attack benchmark targeted at image annotation. Inspired by Adversarial Examples, which is able to produce a specifically crafted noise field that is able to throw off classifcation models, but are undetectable to humans. The research questions then boil down to: * *Commit:* Are Image Annotation Networks subseptical to adversarial examples? * *Stretch:* Is the output annotation controlable? (i.e. can we produce an adversarial example which will generate a (random) chosen output)

I want to definetly answer the first (commit) research question. I will try my best to answer the second (stretch).

Although this project will focus on Adversarial Attack on Image Annotation, it should be applicable to the broader scope of image-to-text models. However I do not plan to go into that, due to the limited time of the BEP.

Personal

Although not directly related, I have some personal goals that I would like to mention and rationalize. * TDD: To ensure that the code that is written is correct, modular and easy to change. Also making sure everything is deterministic and therefore reproducable. * A better understanding of how Neural Networks work and train. NNs are usefull but to most still a magic black box that just works. I want to be able to better reason as to why certain things will work and others won't. * Mono-repo: Keeping everything related to the project under a single github repository. Including the code, paper and other resources. This will

Pre Study

Relevant research has been done in the area of image classification, most notably Intriguing properties of neural networks, which proposes a way of finding adversarial examples for image classification. It also discusses the cross-model generalization of the produced samples. The result of this project could thus also be used to strengthen current datasets. There are a broad range of Image Annotation Networks and surveys on the difference between them. On of the more basic deep learning image annotation models will be my first focus, as

it is closly related to the image classification structure used in aformentioned research.

Methodology

First reproducing some relevant papers as to gain experience with the field and project. Starting with Intriguing properties of neural networks and then a deep learning image annotation model. After which I will combine the two methods to see if Image Annotation is subseptical to Adversarial attakcs. Training and Testing will be done using the same sets as used in the aformentioned paper (Coral5k, ESP Game and IAPR-TC12).

Timeline

Date	Description of Goal
22 February	Draft of project plan
01 March	Hand in project plan
08 March	Reproduced Intriguing properties of neural networks
15 March	Reproduced Deep learning image annotation
29 March	Applied Adversarial Attack on Image Annotation Network
17 April	Hand in Partial Thesis
06 May	Targeted output
03 June	Finalized experimentation
19 June	Hand in Final Thesis

Technology

The code will be written in Python, with Pytorch as Machine-Learning backend. Versioning will be done with git, see repo. Which will also contain the working version of the paper and other resources, such as this plan.