Adversarial Attack on Image Annotation

H.J.M. van Genuchten, 1297333

March 2020

1 Goal

1.1 General

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

1.2 Specific

I want to create an adversarial attack benchmark targeted at [3, image annotation] . Inspired by [2, 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:

- Are Image Annotation Networks susceptible to adversarial examples?
- 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 research question. I will try my best to answer the second.

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.

1.3 Personal

Although not directly related, I have some personal goals that I would like to mention and rationalize:

- Test Driven Development (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.

2 Pre Study

Relevant research has been done in the area of image classification, most notably [4, 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 [1, surveys] on the difference between them. On of the more basic [3, deep learning image annotation] models will be my first focus, as it is closly related to the image classification structure used in aformentioned research.

3 Methodology

First reproducing some relevant papers as to gain experience with the field and project. Starting with [4, Intriguing properties of neural networks] and then a [3, deep learning image annotation] model. After which I will combine the two methods to see if Image Annotation is susceptible 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).

3.1 Timeline

I have set up the following schedule for myself. Bolded deadlines are from the university, the rest is a rough sketch to keep myself on schedule.

Date	Description
22 February	Hand in draft of Project plan
01 March	**Hand in Project plan**
08 March	Reproduced Adversarial Attack on classification
22 March	Reproduced An Image Annotation network
10 April	Applied Adversarial Attack on Image Annotation
17 April	**Hand in Partial thesis**
06 May	Targeted Output
03 June	Finalized experimentation
19 June	**Hand in Final Thesis**

3.2 Technology

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

References

[1] Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. A survey and analysis on automatic image annotation. *Pattern Recognition*, 79:242–259, 2018.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[3] Venkatesh N. Murthy University of Massachusetts Amherst, Venkatesh N. Murthy, University of Massachusetts Amherst, Subhransu Maji University of Massachusetts Amherst, Subhransu Maji, R. Manmatha University of Massachusetts Amherst, R. Manmatha, Carnegie Mellon University, City University of Hong Kong, Fudan University, and et al. Automatic

image annotation using deep learning representations: Proceedings of the 5th acm on international conference on multimedia retrieval, Jun 2015.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.