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 image annotation (Xu et al., 2016; Venkatesh N. Murthy et al., 2015). Inspired by "Adversarial Examples" proposed by Szegedy et al., which are images with small perturbations that is able to throw off classification models, but are unnoticeable to humans. Szegedy et al. also found that these adversarial examples generalize across models. I want to investigate if image annotation networks suffer from the same vulnerabilities. Hence, the following research questions:

- Are Image Annotation Networks vulnerable to adversarial examples?
  - Can the addition of small (crafted) perturbations significantly affect the output of the network?
  - Can the addition of small (crafted) perturbations significantly affect the BLUE(Papineni, Roukos, Ward, & Zhu, 2001) score of the network?
- Is the output annotation controllable?

- Can a perturbation be crafted in such a way that a word gets repeated continuously.
- Can a perturbation be crafted in such a way that it forces a certain annotation?

The priority will be on the first research question, and if time permits I would like to also answer the second research question.

To further minimize the initial scope, I will focus on a single image annotation model, most likely Show, Attend and Tell (S.A.T.) by Xu et al. as it is a fully deep learning based approach.

2 Pre Study

Relevant research has been done in the area of image classification, most notably by Goodfellow, Shlens, and Szegedy, who propose a faster way of finding adversarial examples for image classification. Also, the findings by Venkatesh N. Murthy et al. show that adversarial examples have a cross-model generalization. The result of this project could thus also be used to strengthen current datasets. Furthermore, there has been a lot of research into image annotation. One of the more basic full deep learning image annotation models, S.A.T. (Xu et al., 2016), will be my first focus, as it is closely related to the image classification structure used in aforementioned research.

3 Methodology

First reproducing some relevant papers as to gain experience with the field and project. Starting with (re)producing an adversarial attack on a MNIST classification model (Szegedy et al., 2014) to get familiar with handling images and producing adversarial examples. Then applying that on a deep learning image annotation model, like the one proposed by Xu et al.. After which I will combine the two methods to see if S.A.T. is susceptible to adversarial examples. For training and testing I will make use of the publicly available datasets MS COCO (Lin et al., 2015) and Flickr8K (Hodosh, Young, & Hockenmaier, n.d.). I will be mainly looking at BLUE (Papineni et al., 2001) score.

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 MNIST classification
22 March	Have a working Image Annotation model
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[1]. Which will also contain the working version of the paper and other resources, such as this plan.

References

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples.*

Hodosh, M., Young, P., & Hockenmaier, J. (n.d.). *Flickr8k dataset.*

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., . . . Dollár, P. (2015). *Microsoft coco: Common objects in context.*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). Bleu. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. doi: 10.3115/1073083.1073135

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks.*

[1] `https://github.com/dikvangenuchten/bep-adversarial-image-annotation`

Venkatesh N. Murthy, V. N., Amherst, U. o. M., of Massachusetts Amherst, S. M. U., Maji, S., of Massachusetts Amherst, R. M. U., Manmatha, R., ... et al. (2015, Jun). *Automatic image annotation using deep learning representations: Proceedings of the 5th acm on international conference on multimedia retrieval.* Retrieved from `https://dl.acm.org/doi/pdf/10.1145/2671188.2749391`

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2016). *Show, attend and tell: Neural image caption generation with visual attention.*