Sampling in low confidence area of a targeted model

Adrien Chan-Hon-Tong

Sampling in low confidence area of a targeted model.

Adrien CHAN-HON-TONG

January 22, 2020

Abstract

Generative adversarial networks are traditionally used to generate data for itself (super resolution, procedural texture) or for helping training tasks (semi supervised training, zero shot learning, adversarial loss). In this paper, we aim to generate testing data.

At first glance, this may seem useless because these data will require human annotation to ensure their relevancy and correct classes. But this human annotation is the main lock in many domains where there is a profusion of data (remote sensing, youtube video ...).

Yet, the main idea of the offered framework is to sample especially in the low confidence area of a targeted model. This way, the human annotation cost is focused on hard samples.

Despite this framework does not allow to mitigate the issue of false output with high confidence, it may be an interesting to tackle the rare event probability problem of estimating error probability in low confidence area of a targeted network.

1 Introduction

Deep learning [5] raising after [4] is due to the performance reached by these models on classical tasks like image classification. But, it is also due to the facility to adapt such model for some not classical tasks e.g. deep network allows a rise in semantic segmentation [12] just by tricking last layers of classical networks.

An other task which has been more or less successfully tackled is data generation. Indeed, using generative adversarial networks [3], one can generate *somehow* realistic image.

Such data generation ability is mainly considered for the data itself e.g. super resolution [6] or procedural texture [13]. More recently, GAN framework has been used for adversarial loss [8], zero shot learning [1], or, semi supervised learning [14].

But, today GAN has never been used to generate testing images. This is not surprising, because, data itself can not be used in testing dataset: data

has to be annotated (both from semantic point of view and to be sure that the data deserves to be processed). Yet, the cost to annotate the data is the main issue ! Indeed, there is profusion of data in many domain such as low resolution earth observation (4To of sentinel2 data are released by ESA every day). So, generating data is not relevant as it.

Yet, in this paper, we argue that generating hard data is both feasible and relevant. It is feasible by forcing the GAN to generated both realistic and targeted-model-low-confidence data. And, it is relevant because evaluating a *good deep network* is somehow dealing with a rare event probability estimation [9].

Indeed, in context of certification of perception software for critical application, probability error[1] will be required to be very low - typically 10^{-8}. Yet, performing a statistical test to measure such low probability straightforwardly required a very large set of annotated data. In this context, biasing the sampling toward data which may contain the error is relevant. Unfortunately, this is not trivial in machine learning setting. An approximation is to bias the sampling toward data on with the targeted model has low confidence. This may obviously forget samples which may be wrongly classified with high confidence. Yet, empirically, samples with high confidence (except under adversarial attack setting [11, 10] and/or data poisoning [2]) have lower probability to be wrongly classified. Currently, in detection setting, the standard academic metric [7] (mean average precision) even measures the agreement between confidence and correctness instead of probability of failures. So, biasing the sampling toward data on with the targeted model has low confidence may already be an advance into generating cheaper testing datasets.

2 GAN for sampling in low confidence area

The global pipeline of our framework aiming to generate hard and annotated testing samples is presented in figure 1.

The global idea is to have a GAN module generating data on which the targeted network has low confidence. This can be tested on the fly, so it is possible to reject any data for which targeted network has not a low confidence. It is even possible to update the seed of the generated data by gradient descent to fasten the generation of hard (i.e. low confidence for the targeted network) data.

On the other hand, the GAN module is trained under adversarial network setting, so offered data can be checked by the discriminator (and rejected if discriminator considers the image as fake).

This way, only hard image considered as true by the discriminator are selected. Then, human should review the image, and, provides two information[2]:

[1] Today, probability error of software is required to be 0, but, this is unrealistic for ambiguous objective software like perception software. So, perception software may at some point just be required to have a low empirical error rate.

[2] We assume an image without class is not realistic. Also, it is either possible to generate

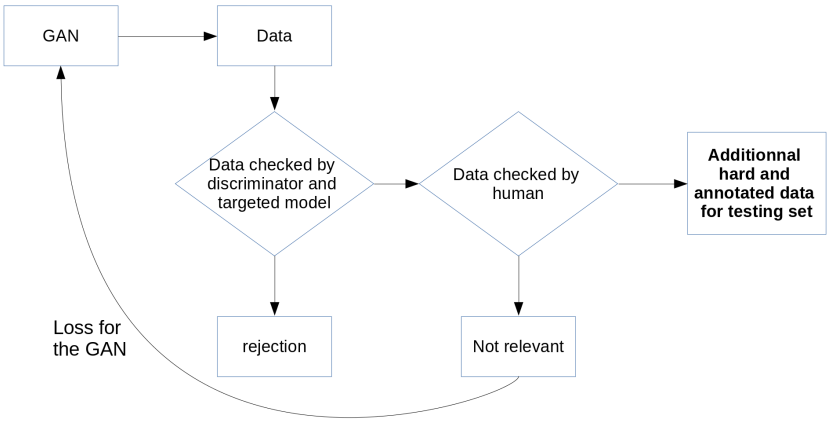

Figure 1: Overview of our pipeline to generate additional hard and annotated testing samples.

is the image is realistic ?, and, *what is the real class of the image ?*. If the human rejects the image, it means that the discriminator has failed, and, thus that the discriminator should be updated.

Now, when the human accepts the image, it means that the framework has generated additional hard data for the testing set (with low human effort is discriminator is efficient).

Experimental experiments should now be conducted to assess if such process is able to make the empirical error estimator converging faster than naive estimator. This can be done using a very large testing dataset by comparing estimator measured on $k\%$ of the testing data sampled uniformly, or, the same amount of data generated by the framework.

References

[1] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2666–2673, 2017.

[2] Adrien CHAN-HON-TONG. An algorithm for generating invisible data poisoning using adversarial noise that breaks image classification deep learning. *Machine Learning and Knowledge Extraction*, 1(1):192–204, Nov 2018.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative

prelabeled image using conditional GAN, this way the human just has to confirm/reject the label

adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[6] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[8] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[9] Jérôme Morio and Mathieu Balesdent. *Estimation of rare event probabilities in complex aerospace and other systems: a practical approach*. Woodhead Publishing, 2015.

[10] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[11] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[13] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[14] Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.