WEAK ADVERSARIAL BOOSTING

Sreekalyan Deepakreddy and Raghav Kulkarni
LinkedIn (Bangalore) India

ABSTRACT

The *adversarial training* methods have recently been emerging as a promising avenue of research. Broadly speaking these methods achieve efficient training as well as boosted performance via an adversarial choice of data, features, or models. However, since the inception of the Generative Adversarial Nets (GAN), much of the attention is focussed on adversarial *models*, i.e., machines learning by pursuing competing goals. In this note we investigate the effectiveness of several (weak) sources of adversarial *data* and *features*. In particular we demonstrate: (a) low precision classifiers can be used as a source of adversarial data-sample closer to the decision boundary (b) training on these adversarial data-sample can give significant boost to the precision and recall compared to the non-adversarial sample. We also document the use of these methods for improving the performance of classifiers when only limited (and sometimes no) labeled data is available.

1 INTRODUCTION

Over the last few decades, machine learning has witnessed several beautiful techniques for obtaining efficient training and boosting the performance of the models. Many of these methods, for example boosting, work well in supervised settings which require abundance of labeled data. Obtaining a large amount of labeled data is expensive and time consuming. However, unstructured and unlabeled data is only exploding day by day in the internet-era and it is easily available. This has motivated several innovative research ideas to obtain an insight from the unlabeled data or to use as few labeled examples as possible to obtain high performance. The area of *active learning*, for instance, focusses on some of these methods.

Another area has recently seen a growing interest and has been a topic active research, which is *adversarial training*. In traditional machine learning methods, much of the importance is given to selecting *good* data, *good* features, and a *good* model. However, in adversarial training methods, one chooses *adversarial* data, *adversarial* features, or *adversarial* models in order to boost the performance. This is a counter-intuitive approach that can sometimes give efficient performance with as few training examples as possible. The adversarial training methods have especially been popular since the emergence of Generative Adversarial Nets (GAN). The GAN focus on *adversarial models* approach, and is a rich and active research area by itself.

We investigate more basic adversarial choices, namely that of *data* and *features*. In essence, we attempt to formulate an abstraction of the notion of *adversarial training* starting from the basic choices of data, features, and models supporting with experiments the growing belief that adversarial choices can drastically boost the performance with significantly fewer labeled examples. Moreover one can use low-precision classifiers as a weak source of adversarial data and one can use low-precision auto-encoders as a weak source of adversarial features. Our motivation to explore such methods came out of practical need for improving the machine learning classifiers for text classification at LinkedIn. We have documented the experiments and results in later sections.[1]

[1]We can not release the full data for privacy reason. However we can make the methods and results available on partial and similar data-sets.

2 ADVERSARIAL DATA CAN BOOST PRECISION AND RECALL

2.1 WHAT IS ADVERSARIAL DATA?

The adversarial data can mean different things based on the context. For instance, people have explored adversarial images that change say a picture of cat in an adversarial way on few pixels so as to fool a classifier to identity it as a dog. The generator in GAN generates such adversarial examples starting from a random noise. However we will focus on weaker notion of adversarial data. There have also been notions of adversarial examples in the context of security attacks. For simplicity let us assume that our classification problem is binary, i.e., we only have two labels one positive and other negative. Then by adversarial data-sample we mean a sample in which it is difficult to tell apart positively labeled data point from a negatively labeled data point. In other words the data-sample is spread closer to the decision boundary. Of course we want a sample that also represents the decision boundary well enough to have a better generalization ability.

2.2 LOW PRECISION CLASSIFIERS CAN BE A WEAK SOURCE OF ADVERSARIAL DATA

One thing for sure is that low-precision classifiers are bad at classifying. However our experiments demonstrate that one can still re-use them in a constructive way. One of the things low-precision classifiers might be good at is generating adversarial data. Say after some heuristic attempts someone obtained a classifier with say only 50 percent precision, it is likely that the positives of the classifier are producing examples close to the decision boundary and they are harder to distinguish.

2.3 WHY ADVERSARIAL DATA MIGHT BOOST PRECISION?

The adversarial data acts as a hack for support vectors. The small number of support vectors are enough to define the decision boundary. The support vectors are in fact closest to the decision boundary. In similar way, a small amount of adversarial data can define the decision boundary highly accurately as the adversarial data is closer to the decision boundary. These examples are more informative than non-adversarial examples. Therefore if we have a constraint on number of examples to get labeled, it might be a good idea to choose those from adversarial data for better precision.

3 COOPERATIVE ADVERSARIAL BOOSTING

Suppose we have two correlated classifiers A and B and we have already trained A with high (say 95% precision) and B has relatively lower precision (say 75%). One can use adversarial data choice here to lift the precision and recall of B as follows:

3.1 BOOSTING PRECISION IN ABSENCE OF LABELED DATA

Take $A \cap B$ as the positive labels for B and random sample outside B as negative label for B. This will improve the precision of B (to say 85%). Iterating this process few times can lift precision of B close to that of A.

3.2 BOOSTING RECALL USING ADVERSARIAL DATA

Furthermore if we want to improve the recall of B then one can take $A \cap B$ as the positive labels for B and $A - B$ as negative. This will force B to output more positives outside of A. Thus increasings the total number of positives by A and B and improving the overall recall.

4 EXPERIMENTS AND RESULTS

We had a classifier with precision 30% attempting to catch spam articles at LinkedIn. The classifier was catching a lot of genuine articles which contained some bad words or non-common language. One needed human intervention to read these articles careful before calling them spam or genuine. Thus the classifier was outputting examples close to boundary. We had a limited labeling bandwidth

(say 1000 labels). We experimented training first with a random sample, which gave a precision of 75% only. This was not good as we needed a precision of at least 95% to run automatic take-down of the positives. Therefore instead of choosing a random 1000 samples, we chose 1000 positives of the low-precision classifier and trained our classifier on that. This gave 80% precision on the training data and surprisingly 95% precision for real online-data. The lift in precision was because the real-data was non-adversarial whereas our training data was adversarial. Notably all these lifts were obtained using only small number of labeled examples.

After training a high precision classifier to catch the spam, we wanted to improve the recall. There were also other correlated classifiers whose precision and recall needed a boost. We used co-operative adversarial boosting described above to lift the precision of one correlated classifier from 75% to 95%. The recall of the combined classifier also saw a drastic improvement (15% to 65%) due to adversarial data choice. Throughout this process surprisingly we did not use any additional labeled data.

5 Open ends

5.1 Boosting via adversarial features

In an ongoing experiment, we are exploring usefulness of adversarial features for boosting precision and recall. Suppose for instance, we are interested in high precision classifier to distinguish between images of horses from those of cars. We start with a (badly tuned) low precision auto-encoder trained on horse images. It barely maintains the essential features and decodes a horse as a horse-shape. Moreover if we run this auto-encoder on an image of a car, it will try to identity horse-like features within the car image and will decode part of the shadow for instance as a horse tail. In other words, original pictures of horse and cars were far from the boundary whereas the decoded horse and decoded car are harder to distinguish and therefore come closer to decision boundary. Hence the encoded features, which are maintaining only barely minimal features to be able to distinguish between horse and a car are acting like an adversarial choice of features. With this choice we observed that a small number of training examples suffice to give a boosted precision using the adversarial features as opposed to the original image features.

5.2 Do iterations help?

Iteratively modifying the adversarial data or adversarial features seems to be a simple yet powerful trick. We believe that such a careful choice of iterations can further boost the performance. It is intriguing to compare these simple iterative adversarial boosting methods to more sophisticated adversarial models such as GANs where the adversarial nature is enforced by choosing a common loss functions combining opposite goals.

References

J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio: Generative Adversarial Nets. NIPS 2014: 2672-2680

Recent workshop on adversarial training: https://sites.google.com/site/nips2016adversarial/

Resources on active learning: https://en.wikipedia.org/wiki/Active-learning