VIRTUAL ADVERSARIAL TRAINING SYSTEM FOR DCASE 2019 TASK 4

Technical Report

Anthony Agnone, Umair Altaf

817 West Peachtree Street NW #770
Atlanta, GA 30308, USA
{aagnone,ualtaf} at pindrop.com

ABSTRACT

This paper describes the approach used for Task 4 of the DCASE 2019 Challenge. This tasks challenges systems to learn from a combination of labeled and unlabeled data. Furthermore, the labeled data is itself a combination of strongly-informed, coarse time-based data and weakly-informed, fine time-based synthetic data. The baseline system builds off of the winning solution from last year, and adds the synthetic data, which was not provided in that iteration of the challenge. Our solution uses the semi-supervised virtual adversarial training method, in addition to the Mean Teacher consistency loss, to encourage generalization from weakly-labeled and unlabeled data. The chosen system parametrization achieves a 59.57% macro F1 score.

***Index Terms*—** Semi-supervised learning, Virtual adversarial training, Data augmentation, Sound event detection

1. INTRODUCTION

The multi-offering DCASE competition is now in its fifth year of execution. In the past few years, Task 4 has been reserved for large-scale and weakly supervised applications, which are two critical steps for machine learning algorithms as they are applied into real-world, unstructured environments. The winning solution [1] last year used a convolutional and recurrent system trained with the Mean Teacher approach [2], which assesses a consistency loss between a teacher network and student network. Although these two networks are indeed different, their architecture is always identical; in order to encourage temporal consistency, the weights of the student network are set as an exponential moving average of the respective weights of the teacher network. This method was initially demonstrated on image data, and was shown to also be very effective on audio, achieving the highest F1 score on the DCASE 2018 evaluation data. In this work, we introduce Virtual Adversarial Training [4] to the DCASE community, and encourage its application and study in future applications.

2. PROPOSED METHOD

Since its introduction, the Mean Teacher method has become quite popular for semi-supervised and weakly supervised deep learning applications. By penalizing deviation of a network from the exponential moving average of its weights over time, the network learns to be *consistent with respect to weight change*. Separately, it has also become a common heuristic to apply either isotropic Gaussian noise or realistic augmentation to the network inputs, in order to achieve *consistency with respect to function input*. While these methods have shown notable improvements in well-established benchmarks, they still only create new samples which are highly correlated with the original samples, and thus are quite limited in the extent to which new information may be presented to the learner during training.

In [3] and [4], Miyato et al propose the method of Virtual Adversarial Training (VAT) through Local Distribution Smoothing (LDS). In this method, one computes an estimate of the adversarial sample for a sample.This sample is that which, when backpropagating the error gradient, results in the largest change in the predictive distribution of the learner. As opposed to simple isotropic and otherwise highly-correlated perturbations, the VAT perturbation seeks the direction from a vector in which the learner is most affected. By penalizing the KL divergence of the predictive distributions of the network after seeing each of the normal and adversarially augmented samples, the network is encouraged to exhibit smooth response behavior, even in (estimated) adversarial cases. Since the derivation of the estimate does not require class labels, and thus makes no explicit claims of class assignment, the term ”virtual” is used to describe the adversarial vector.

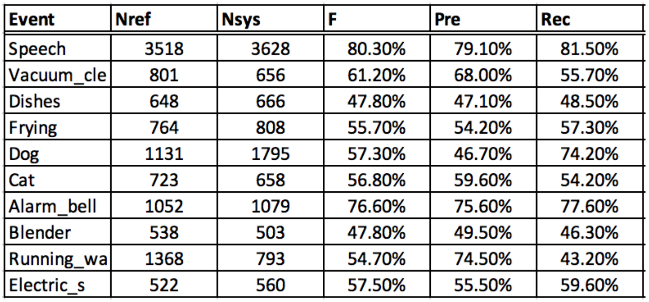
Tarvainen et al. note in [2] that the Mean Teacher method is likely to be mostly complementary to that of the VAT method, although a full study has not been performed as of yet. This paper also makes no formal attempt at the comparison, and simply uses the methods in conjunction, to evaluate its effect on the baseline system. We hope that our simple demonstration of improvement over the baseline, without any further modifications or optimizations, encourages other researchers to apply the VAT loss to their implementations in the future, in order to control the variance of the learner's predictive distribution as its complexity grows.

3. RESULTS

The macro F1 scores are shown in Table 1. Parameters varied include the number of power iterations performed (np), the magnitude movement in the adversarial direction (eps), and the regularization coefficient for the computed VAT loss (alpha). For details regarding further explanation of the function of each parameter, the reader is referred to [4]. A value of 10 for the regularization coefficient was determined empirically, such that the range of the regularization loss term was of a similar magnitude as the supervised loss term. Figure 3 shows the class-wise performance of the chosen VAT system parametrization on the validation data set.

System	**Macro F1 Score (%)**
baseline	54.23
np1_eps05_alpha10	57.38
np1_eps10_alpha10	57.63
np2_eps10_alpha10	58.06
np1_eps2.5_alpha10	**59.57**

Table 1: Macro F1 validation set performance.



Event	Nref	Nsys	F	Pre	Rec
Speech	3518	3628	80.30%	79.10%	81.50%
Vacuum_cle	801	656	61.20%	68.00%	55.70%
Dishes	648	666	47.80%	47.10%	48.50%
Frying	764	808	55.70%	54.20%	57.30%
Dog	1131	1795	57.30%	46.70%	74.20%
Cat	723	658	56.80%	59.60%	54.20%
Alarm_bell	1052	1079	76.60%	75.60%	77.60%
Blender	538	503	47.80%	49.50%	46.30%
Running_wa	1368	793	54.70%	74.50%	43.20%
Electric_s	522	560	57.50%	55.50%	59.60%

Figure 1: Class-wise validation set performance.

4. CONCLUSION AND FUTURE WORK

In this work, we performed some rudimentary experiments of the efficacy of virtual adversarial training for deep neural networks on weakly-supervised audio event detection. The results suggest that VAT indeed does provide generalization gains for a DNN, and encourages further experimentation moving forward. Future work will explicitly compare the effects of Mean Teacher and VAT on generalization, as well as how each aids learning on real and synthetic audio.

5. REFERENCES

[1] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," *Detection and Classification of Acoustic Scenes and Events*, September 2018.

[2] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," pp. 1195–1204, 2017.

[3] T. Miyato, S. ichi Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," 2015.

[4] T. Miyato, S. ichi Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," 2017.