Building robust classifiers through generation of confident out of distribution examples

Kumar Sricharan, Ashok Srivastava
Central Data Science Organization, Intuit Inc.
{sricharan_kumar, ashok_srivastava}@intuit.com

Abstract

Deep learning models are known to be overconfident in their predictions on out of distribution inputs. There have been several pieces of work to address this issue, including a number of approaches for building Bayesian neural networks, as well as closely related work on detection of out of distribution samples. Recently, there has been work on building classifiers that are robust to out of distribution samples by adding a regularization term that maximizes the entropy of the classifier output on out of distribution data. To approximate out of distribution samples (which are not known apriori), a GAN was used for generation of samples at the edges of the training distribution. In this paper, we introduce an alternative GAN based approach for building a robust classifier, where the idea is to use the GAN to explicitly generate out of distribution samples that the classifier is confident on (low entropy), and have the classifier maximize the entropy for these samples. We showcase the effectiveness of our approach relative to state-of-the-art on hand-written characters as well as on a variety of natural image datasets.

1 Introduction

Deep learning approaches have been shown to be susceptible to being overly confident on unseen instances that are outside of the training distribution. To address this problem, there have been a number of recently proposed techniques. This body of work can be divided into two broad categories: (i) methods that analyze the output of networks trained in a standard fashion to detect out of sample distributions, and (ii) methods that build robust neural networks through modified loss functions including work on Bayesian networks as well as regularized networks. We note that techniques from (i) and (ii) are independent of each other in that techniques from (i) can be combined with techniques from (ii) to produce more effective results.

The work on using the maximum probability of the classifier [4], its modification to use temperature based scaling [12], and the work on using the underlying feature representations in combination with the Mahalanobis distance [10] belong to category (i). Methods under category (ii) include Bayesian neural networks based on placing prior over weights of the network [5], using dropout as a Bayesian approximation [11], ensemble based approaches [7] and adding a regularization term to explicitly maximize entropy of out of distribution samples [9].

Our focus in this work is on the later category. In particular, we build on the recent, state-of-the-art work by Lee *et.al.* [9] on generating robust classifiers that don't overfit on out of distribution samples by explicitly maximizing the uncertainty of a classifier over out of distribution samples. This work in [9] uses a GAN to generate out-of-distribution samples that are close to training distribution but also simultaneously have high entropy in terms of classifier output over these samples. Their key intuition is that by maximizing the uncertainty of the classifier over these out of distribution samples close to the training distribution, the same effect will be propagated to all samples outside of the training distribution.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

Contribution In this paper, we use the same idea of maximizing the classifier uncertainty over out of distribution samples, but take an alternative approach to generating the out of sample distribution. In particular, we use a GAN to generate samples which are *not* from the training distribution that the classifier *is* confident about (i.e., has low entropy), and then have the classifier maximize the uncertainty of the predicted labels over these samples in an iterative fashion. This is in direct contrast to [9], where the GAN is used to generate samples close to the training distribution that the classifier is not confident about. This difference can be seen clearly by studying the equations (1) and (2) below.

This proposed approach leads to improved robust classifiers for out of sample detection because the GAN is explicitly optimized to generate samples away from the training data that the classifier is confident about. These samples can span a much larger space beyond the edges of the training distribution, unlike in [9], where the classifier is restricted to maximizing the uncertainty of samples close to the edge of the training distribution only. Henceforth, we will refer to the method in [9] as BoundaryGAN, and our proposed method as ConfGAN.

2 Training robust classifiers using GANs

We start with the confidence loss proposed by Lee *et.al.*[9]: $L_c(\theta) = \mathbb{E}_{P_{in}(\hat{x},\hat{y})}[-\log \mathbb{P}_\theta(y = \hat{y} \mid \hat{x})] + \beta \mathbb{E}_{P_{out}(x)}[KL(\mathbb{U}(y) \mid\mid \mathbb{P}_\theta(y \mid x))]$ where the first term is the standard cross-entropy loss, and the second term forces the network (parameterized by θ) to generate close to uniform distributions for out of distribution samples (represented by P_{out}). We note that the proposed loss function corresponds to a Bayesian classifier under a prior on the weights that corresponds to the classifier producing a uniform distribution on the network output of class probabilities $P_\theta(y \mid x)$.

2.1 GAN for generating boundary samples

The confidence loss function is very intuitive, but the key difficulty is in determining P_{out}. In [9], they take the approach of approximating P_{out} by learning to generate 'boundary' samples at the edges of the training distribution P_{in} using a GAN [3], with the intuition that if the network learns to generate labels distributions with high entropy for these boundary samples, then this will propagate to all out of distribution samples. To generate the boundary out of distribution samples, the generator in BoundaryGAN is asked to learn samples $\bar{x}$ close to the training distribution (via standard GAN loss) which also have low entropy wrt classifier output (via KL divergence loss wrt uniform distribution). In particular, the overall loss function is given by:

$$\min_\theta \min_G \max_D \quad \underbrace{\mathbb{E}_{P_{in}(\hat{x},\hat{y})}[-\log \mathbb{P}_\theta(y = \hat{y} \mid \hat{x})]}_{\text{A}} + \underbrace{\beta \mathbb{E}_{P_{prior}(z)}[KL(\mathbb{P}_\theta(y \mid G(z)) \mid\mid \mathbb{U}(y))]}_{\text{B}} \\ \underbrace{\mathbb{E}_{P_{in}(\hat{x})}[\log D(\hat{x})] + \mathbb{E}_{P_{prior}(z)}[\log(1 - D(G(z))]}_{\text{C}} \tag{1}$$

Here, (A) is the standard cross-entropy loss, and (C) is the standard GAN loss. The term B is introduced in [9] to both force the generator to generate samples at the boundary of the training distribution, and to regularize the classifier via these out of distribution samples.

2.2 Proposed method: GAN for generating confident samples

In contrast to [9], we suggest an alternative approach where instead of forcing the generator to learn a distribution where the classifier output is close to uniform, we instead ask the generator to identify samples the classifier is confident about (i.e., low entropy) that are away from the training distribution P_{in}. We simultaneously force the classifier to update its parameters θ to produce probability outputs with low entropy for these samples from the GAN. The overall loss function for our proposed ConfGAN method is given by:

$$\min_\theta \max_G \max_D \quad \underbrace{\mathbb{E}_{P_{in}(\hat{x},\hat{y})}[-\log \mathbb{P}_\theta(y = \hat{y} \mid \hat{x})]}_{\text{a}} + \underbrace{\beta \mathbb{E}_{P_{prior}(z)}[KL(\mathbb{P}_\theta(y \mid G(z)) \mid\mid \mathbb{U}(y))]}_{\text{b}} \\ \underbrace{\mathbb{E}_{P_{in}(\hat{x})}[\log D(\hat{x})] + \mathbb{E}_{P_{prior}(z)}[\log(1 - D(G(z))]}_{\text{c}} \tag{2}$$

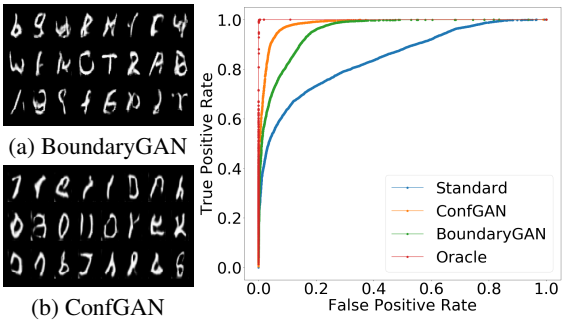
This is similar to the loss function (Eq 1) for BoundaryGAN with the only difference being that in BoundaryGAN, the overall loss is minimized with respect to the generator G instead of being maximized as in Eq.2. For this reason, in term (b), we use the reverse KL divergence instead of the forward version used in [9]. In practice, for term (b), we use the reverse KL term for maximization wrt the generator, and the original forward KL term from [9] for minimization wrt the classifier. This is because the forward KL terms supplies stronger gradients that push the predicted distribution $\mathbb{P}_\theta(y = \hat{y} \mid \hat{x})$ towards the uniform distribution.

Asking the generator to maximize the loss function has the effect of forcing the generator to generate samples the classifier is confident about through term (b), and at the same time, are not from the training distribution through term (c).

3 Experimental results

While the modification from the original loss in BoundaryGAN to the proposed loss function is easy to implement, the effect of the modification is significant in terms of the results. We show this through experimental results on two different data sets - handwritten characters, and natural images. For measuring out-of-distribution detection accuracy, as in [9], we use threshold-based detectors [4] that computes the maximum value of predictive distribution on a test sample and classifies it as in-distribution if this value is above some threshold.

3.1 Handwritten characters

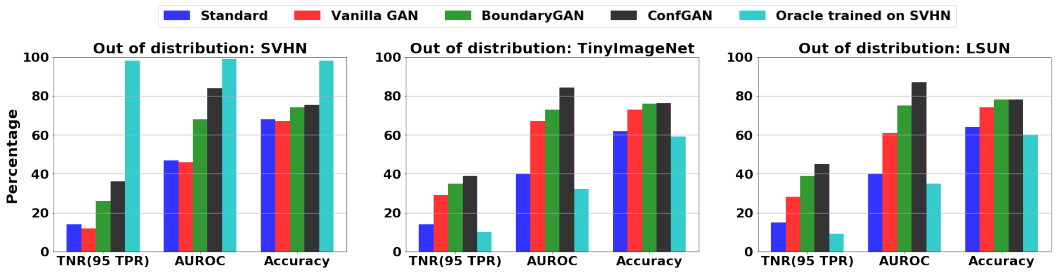


(a) BoundaryGAN

(b) ConfGAN

In the first experiment, we use the handwritten letters from EMNIST [1] as the in-sample distribution, and the FashionMNIST dataset [16] as the out of sample distribution. The neural network for the classifier is LeNet-5 [8] modified to support 47 classes, and the discriminator and generator use the DCGAN architecture [14]. We use $\beta = 1$. We showcase two sets of results: first, we compare the samples of the generator from BoundaryGAN with ConfGAN. As can be seen from the figure, BoundaryGAN produces samples that look like complete characters similar to the actual EMNIST data (similar to Figure 3(d) in [9]), whereas ConfGAN produces images that look like well-defined, but incomplete parts of characters - i.e., samples that are not from the training distribution that a classifier is likely to be overconfident about. Next, we compare the ROC curves measuring detection performance of BoundaryGAN, ConfGAN, the standard classifier trained with no regularization (lower bound), and an oracle classifier that is trained with FashionMNIST as the out of distribution data (upper bound). From Figure 3.2, we can see that ConfGAN comes close to the performance of the oracle, and significantly outperforms BoundaryGAN.

3.2 Image datasets using VGG



The second experiment is on a combination of various datasets: CIFAR [6], SVHN [13], ImageNet [2], and LSUN [17]. We train VGGNet [15] for classifying CIFAR-10, use DCGAN for the generator and discriminator, with $\beta = 0.1$, and use datasets from {SVHN, ImageNet, LSUN} as the out of sample datasets. Under this setup, we measure out of sample detection performance using the same evaluation metrics in [9] (TNR at 95% TPR, AUROC, Detection accuracy). From Figure 3.2, we see that the proposed ConfGAN uniformly outperforms BoundaryGAN and other baselines (except oracle trained and applied on SVHN). The performance improvement of ConfGAN over BoundaryGAN is especially apparent wrt AUROC (with an average improvement of $\sim 12\%$) across the three datasets.

4 Conclusion

We proposed a new method for building robust classifiers to detect out of distribution samples. The method is based on using a GAN to generate samples away from the training data that the classifier is confident about, and have the classifier maximize the uncertainty over these points. Our method outperforms several baselines in this space on handwritten character and natural image datasets.

References

[1] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[6] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[8] Yann LeCun, LD Jackel, Leon Bottou, A Brunot, Corinna Cortes, JS Denker, Harris Drucker, I Guyon, UA Muller, Eduard Sackinger, et al. Comparison of learning algorithms for handwritten digit recognition. In *International conference on artificial neural networks*, volume 60, pages 53–60. Perth, Australia, 1995.

[9] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[10] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *arXiv preprint arXiv:1807.03888*, 2018.

[11] Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. *arXiv preprint arXiv:1703.02914*, 2017.

[12] Shiyu Liang, Yixuan Li, and R Srikant. Principled detection of out-of-distribution examples in neural networks. arxiv preprint. *arXiv preprint arXiv:1706.02690*, 2017.

[13] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011.

[14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[16] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[17] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.