DADA: DEEP ADVERSARIAL DATA AUGMENTATION FOR EXTREMELY LOW DATA REGIME CLASSIFICATION

Xiaofeng Zhang[⋆] *Zhangyang Wang*[†] *Dong Liu*[⋆] *Qing Ling*[‡]

[⋆]University of Science and Technology of China
[†]Texas A&M University
[‡]Sun Yat-Sen University

ABSTRACT

Deep learning has revolutionized the performance of classification, but meanwhile demands sufficient labeled data for training. Given insufficient data, while many techniques have been developed to help combat overfitting, the challenge remains if one tries to train deep networks, especially in the ill-posed *extremely low data regimes*: only a small set of labeled data are available, and nothing – including unlabeled data – else. Such regimes arise from practical situations where not only data labeling but also data collection itself is expensive. We propose a deep adversarial data augmentation (DADA) technique to address the problem, in which we elaborately formulate data augmentation as a problem of training a class-conditional and supervised generative adversarial network (GAN). Specifically, a new discriminator loss is proposed to fit the goal of data augmentation, through which both real and augmented samples are enforced to contribute to and be consistent in finding the decision boundaries. Tailored training techniques are developed accordingly. Source code is available at `https://github.com/SchafferZhang/DADA`

***Index Terms*—** classification, extremely low data regime, GAN, data augmentation

1. INTRODUCTION

The performance of classification and recognition has been tremendously revolutionized by the prosperity of deep learning [1]. Deep learning-based classifiers can reach unprecedented accuracy given that there are sufficient labeled data for training. Meanwhile, such a blessing can turn into a curse: in many realistic settings where either massively annotating labels is a labor-intensive task, or only limited datasets are available, a deep learning model will easily overfit and generalizes poorly. Many techniques have been developed to help combat overfitting with insufficient data, ranging from classical data augmentation [2], to dropout [1] and other structural regularizations [3], to pre-training [4], transfer learning [5] and semi-supervised learning [6]. However in low data regimes, even these techniques will fall short, and the resulting models usually cannot capture all possible input data variances and distinguish them from nuisance variances. The high-variance gradients also cause popular training algorithms, e.g., stochastic gradient descent, to be extremely unstable.

To resolve the challenges, we have made multi-fold technical contributions in this paper: 1) For learning deep classifiers in extremely low data regimes, we focus on boosting the effectiveness of data augmentation, and introduce learning-based data augmentation, that can be optimized for classifying general data without relying on any domain-specific prior or unlabeled data. We call the proposed framework *Deep Adversarial Data Augmentation* (**DADA**). 2) We propose a new loss function for the GAN discriminator, that not only learns to classify real images, but also enforces fine-grained classification over multiple "fake classes". That is referred to as the $\mathbf{2}k$ **loss**, in contrast to the k*+1 loss* used by several existing GANs (to be compared in the context later). The novel loss function is motivated by our need of data augmentation: *the generated augmented ("fake") samples need to be discriminative among classes too, and the decision boundaries learned on augmented samples shall align consistently with those learned on real samples*. 3) We conduct simulations on CIFAR-10, to train deep classifiers in the extremely low data regimes, demonstrating significant performance improvements through DADA compared to using traditional data augmentation. To further validate the practical effectiveness of DADA, we train deep classifiers on real-world small dataset: the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) dataset for the tumor classification task. Numerical experiments demonstrate that DADA leads to highly competitive generalization performance.

2. RELATED WORK

GANs [7] have gathered a significant amount of attention due to their ability to learn generative models of multiple natural image datasets. Conditional GAN [8] generates data conditioned on class labels via label embeddings in both discriminator and generator. Conditioning generated samples on labels sheds light the option of semi-supervised classification using GANs. In [9], the semi-supervised GAN has the dis-

criminator network to output class labels, leading to a $k+1$ class loss function consisting of k class labels if the sample is decided to be real, and a single extra class if the sample is decided to be fake. Such a structured $k+1$ loss has been re-emphasized in [10] to provide more informed training that leads to generated samples capturing class-specific variances better. Even with the proven success of GANs for producing realistic-looking images, tailoring GANs for classification is not as straightforward as it looks like [11].

Data augmentation is an alternative strategy to bypass the unavailability of labeled training data, by artificially synthesizing new labeled samples from existing ones. A latest work [12] presented a novel direction to select and compose pre-specified base data transformations (such as rotations, shears, central swirls for images) into a more sophisticated "tool chain" for data augmentation, using generative adversarial training. They achieve highly promising results on both image and text datasets, but need the aid of unlabeled data in training (the same setting as in [10]). *We experimentally compare the method [12] and DADA and analyze their more differences in Section 5.*

A number of works [13, 14, 15, 16, 17, 18] explored the "free" generation of labeled synthetic examples to assist training. However, they either relied on extra information, e.g., 3D models of the subject, or were tailored for special object classes, e..g, face or license plates. The synthesis could be viewed as a special type of data augmentation that hinges on stronger forms of priori invariance know,ledge.

3. TECHNICAL APPROACH

3.1. Problem Formulation a,nd Solution Overview

Data augmentation approaches seek an **augmenter** A, to synthesize a new set $\mathcal{D}'$ of augmented labeled data $(\bar{x}^i, y^i)$ from (x^i, y^i), constituting the new augmented training set of size $|\mathcal{D}| + |\mathcal{D}'|$. Traditional choices of A, being mostly ad-hoc minor perturbations, are usually class-independent, i.e., constructing a sample-wise mapping from x^i to $\bar{x}^i$ without taking into account the class distribution. Such mappings are usually limited to a small number of priori known, hand-crafted perturbations. They are not learned from data, and are not optimized towards finding classification boundaries. To further improve A, one may consider the inter-sample relationships [19], as well as inter-class relationships in $\mathcal{D}$, where training a generative model A over (x^i, y^i) becomes a viable option.

The conceptual framework of DADA is depicted in Fig. 1. If taking a GAN point of view towards this, A naturally resembles a generator: its inputs can be latent variables z^i conditioned on y^i, and outputs $\bar{x}^i$ belonging to the same class y^i but being **sufficiently diverse** from x^i. C can act as the discriminator, if it will incorporate typical GAN's real-fake classification in addition to the target k-class classification. Ideally, the classifier C should: (1) **be able to** correctly classify both real samples x^i and augmented samples $\bar{x}^i$ into the correct class y^i; (2) **be unable to** distinguish x^i and $\bar{x}^i$. The entire DADA framework of A and C can be jointly trained on (x^i, y^i), whose procedure will bear similarities to training a class-conditional GAN. However, existing GANs may not fit the task well, due to the often low diversity of generated samples. We are hence motivated to introduce a novel loss function towards generating more diverse and class-specific samples.

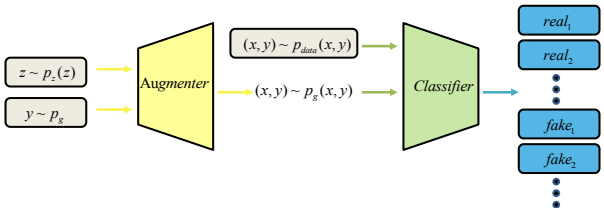

Fig. 1. An illustration of DADA.

3.2. Going More Discriminative: From $k+1$ Loss to $2k$ Loss

The discriminator of a vanilla GAN [7] has only one output to indicate the probability of its input being a real sample. In [10, 9], the discriminator is extended with a semi-supervised fashion $k+1$ loss, whose output is a $(k+1)$-dimensional probabilistic vector: the first k elements denote the probabilities of the input coming from the class 1, 2, ..., k of real data; the $(k+1)$-th denotes its probability of belonging to the generated fake data. In that way, the generator simply has the semi-supervised classifier learned on additional unlabeled examples and supplied as a new "generated" class. In contrast, when in extremely low data regimes, we tend to be more "economical" on consuming data. We recognize that the unlabeled data provides weaker guidance than labeled data to learn the classification decision boundary. Therefore, if there is no real unlabeled data available and we can only generate from given limited labeled data, generating labeled data (if with quality) should benefit classifier learning more, compared to generating the same amount of unlabeled data. Further, the generated labeled samples should join force with the real labeled samples, and their decisions on the classification boundary should be well aligned. Motivated by the above philosophy, we build a new $2k$ loss function, whose first group of k outputs represent the probabilities of the input data from the class 1, 2, ..., k of real data; its second group of k outputs represent the probabilities of the input from the class 1, 2, ..., k of fake data.

3.3. Training Algorithm Details

The training procedure of DADA is divided into two different phases. In training phase I, which we call **Generation training**, the classifier and the augmenter compete with each other

Table 1. The comparison of loss functions among GAN discriminators

Model	Class Number	Classes	Training Data
Vanilla GAN [7]	2	real, fake	unlabeled only
Improved GAN [10]	$k+1$	$real_1$, ..., $real_k$; $fake$	labeled + unlabeled
Proposed	$2k$	$real_1$, ..., $real_k$; $fake_1$, ..., $fake_k$	labeled only

within a specific class. The game between the two players will have its optimum only if $p_{data}(x|y) = p_g(x|y)$. Thus, the optimal classifier has $C(x|y) = p_{data}(x|y)/(p_g(x|y) + p_{data}(x|y)) = 1/2$, indicating that the augmenter is trained well enough so that the classifier can not discriminate them.

Similar to the vanilla GAN formulation, the loss functions of the augmenter and the classifier in training phase I are:

$$\begin{aligned} L_C^I = &- \mathbf{E}_{x,y\sim p_{data}(x,y)} \log[p(y|x, y < k+1)] \\ &- \mathbf{E}_{x,y\sim p_g(x,y)} \log[p(y|x, k < y < 2k+1)] \end{aligned} \quad (1)$$

$$\mathcal{L}_G^I = -\mathbf{E}_{x,y\sim p_g(x,y)} \log[p(y-k|x, k < y < 2k+1)]. \quad (2)$$

Based on the observation of the Improved-GAN that the feature matching technique can help improve the classification performance of the generated samples, we make some modifications on this training strategy. The conditioned feature matching is formulated as:

$$\begin{aligned} \mathcal{L}_{fm} = &\|\mathbf{E}_{x,y\sim p_{data}(x,y)} f(x|y) - \\ &\mathbf{E}_{z\sim p_z(z), y\sim p_c} f(G(z,y)|y)\|. \end{aligned} \quad (3)$$

Here $f(x)$ denotes activations on an intermediate layer of the classifier. We keep p_c the same to the true data label. With the regularization of feature matching, the objective function of generator in training phase I is hence:

$$L_G^I = \mathcal{L}_G^I + \lambda \mathcal{L}_{fm}. \quad (4)$$

Once the training phase I is finished, assuming that the generator can capture the class-wise data distribution, then it comes to the training phase II called **Classification training**. In this phase, the generator is fixed just as a data provider. We only train the classifier on the generated data and the real training data. The loss function of the classifier in training phase II can be written as:

$$L_C^{II} = \mathcal{L}_{data} + \mathcal{L}_{gen} \quad (5)$$

where

$$\begin{aligned} \mathcal{L}_{data} = &- \mathbf{E}_{x,y\sim p_{data}(x,y)} \log[p(y|x, y < k+1) + \\ &p(y+k|x, y < k+1)] \end{aligned} \quad (6)$$

and

$$\begin{aligned} \mathcal{L}_{gen} = &- \mathbf{E}_{x,y\sim p_g(x,y)} \log[p(y|x, k < y < 2k+1) + \\ &p(y-k|x, k < y < 2k+1)]. \end{aligned} \quad (7)$$

The entire training procedure is summarized in Algorithm 1

Algorithm 1 Minibatch stochastic gradient descent training of DADA

Require: The training epochs $\mathcal{K}_G$, $\mathcal{K}_C$ in phase I and phase II, the training set $\mathcal{D}$, the test set $\mathcal{T}$, the batch size $\mathcal{B}$

for number of epochs $\mathcal{K}_G$ **do**
  Sample a batch of pairs (z, y), $z \sim p_z(z), y \sim p_g$, a batch of pairs $(x, y) \sim p_{data}(x, y)$.
  Update the classifier by performing stochastic gradient descent on L_C^I
  for number of epochs e **do**
    Sample a batch of pairs $(x, y) \sim p_{data}(x, y)$, a batch of pairs (z, y), $z \sim p_z(z)$, keep y the same with the true data
    Update the generator by performing stochastic gradient descent on L_G^I
  end for
end for
for number of epochs $\mathcal{K}_C$ **do**
  Sample a batch of pairs (z, y), $z \sim p_z(z), y \sim p_g$, a batch of pairs $(x, y) \sim p_{data}(x, y)$.
  Update the classifier by performing stochastic gradient descent on L_C^{II}
end for

4. SIMULATIONS

To evaluate our approach, we first conduct simulations CIFAR-10 image classification benchmark. We intentionally sample the given training data to simulate the extremely low data regimes, and compare the following training options. 1) C: directly train a classifier using the limited training data; 2) C_augmented: perform traditional data augmentation (including rotation, translation and flipping), and then train a classifier; 3) DADA: the proposed data augmentation; 4) DADA_augmented: first apply the same traditional augmentation as C_augmented on the real samples, then perform DADA. We use absolutely **no unlabeled data or any pre-trained initialization** in training, different from the setting of most previous works. We use the original full test sets for evaluation. The network architectures that we used have been exhaustively tuned to ensure the best possible performance of all baselines in those unusually small training sets.

To illustrate the advantage of our proposed $2k$ loss, we also use the vanilla GAN [7] (which adopt the 2-class loss), as well as the Improved GAN [10] (which adopt the $(k+1)$-class loss), as two additional baselines to augment samples. For the

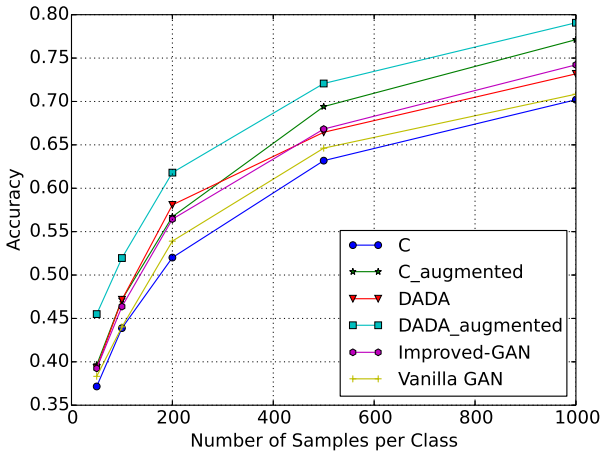

Fig. 2. Results on CIFAR-10, the test accuracy in different training settings with respect to the number of training images per class.

vanilla GAN, we train a separate generator *for each class*. For Improved GAN, we provide only the labeled training data without using any unlabeled data: a different and more challenging setting than evaluated in [10]. They work with traditional data augmentation too, similarly to the DADA_augmented pipeline. For all compared methods, we generate samples so that the augmented dataset has 10 times the size of the given real labeled dataset.

Fig. 2 summarizes the performance of the compared methods. The vanilla GAN augmentation performs slightly better than the no-augmentation baseline, but the worst in all other data augmentation settings. It concurs with [20] that, though GAN can generate visually pleasing images, it does not naturally come with increased data diversity from a classification viewpoint. While improved GAN achieves superior performance, DADA (without using traditional augmentation) is able to outperform it at the smaller end of sample numbers (less than 400 per class). Comparing with vanilla GAN, Improved GAN and DADA_augmented reveal that as the discriminator loss goes "more discriminative", the data augmentation becomes more effective along the way.

Furthermore, DADA_augmented is the best performer among all, and consistently surpass all other methods for the full range of [50, 1000] samples per class. It leads to around 8 percent top-1 accuracy improvement in the 500 labeled sample, 10 class subset, without relying on any unlabeled data. It also raises the top-1 performance to nearly 80%, using only 10% of the original training set (i.e. 1000 samples per class), again with neither pre-training nor unlabeled data.

5. EXPERIMENTS WITH REAL-WORLD SMALL DATA

In this section, we discuss real-data experiments which fall into extremely low data regimes. The data, not just labels, are difficult to collect and subject to high variability. We show that in this cases, the effects of transfer learning are limited, and/or no ad-hoc data augmentation approach might be available to alleviate the difficulty to train deep networks. In comparison, DADA can be easily plugged in and boost the classification performance in all experiments.

Table 2. Comparison between DADA and Tanda.

Models	Acc
Tanda (MF)	0.5990
Tanda (LSTM)	0.6270
DADA	0.6196
DADA_augmented	**0.6549**

In the existing learning-based data augmentation work *Tanda* [12], most training comes with the help of unlabeled data. One exception we noticed is their experiment on the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [21, 22, 23], a medical image classification task whose data is expensive to collect besides labeling. Since both Tanda and DADA use the only available labeled dataset to learn data augmentation, we are able to perform a fair comparison on CBIS-DDSM between the two. We follow the same classifier configuration used for CBIS-DDSM by Tanda: a four-layer all-convolution CNN with leaky ReLUs and batch normalization. We resize all medical images to 224×224. Note that Tanda heavily relied on hand-crafted augmentations: on DDMS, they used many basic heuristics (crop, rotate, zoom, etc.) and several domain-specific transplantations. For DADA_augmented, we apply only rotation, zooming, and contrast as the traditional augmentation pre-processing, to be consistent with the user-specified traditional augmentation modules in Tanda. We compare DADA and DADA_augmented with two versions of Tanda using mean field (MF) and LSTM generators [12], with results in Table 2 showing the clear advantage of our approaches.

6. CONCLUSION

We present DADA, a learning-based data augmentation solution for training deep classifiers in extremely low data regimes. We leverage the power of GAN to generate new training data that both bear class labels and enhance diversity. A new $2k$ loss is elaborated for DADA and verified to boost the performance. We perform extensive simulations as well as real-data experiments, where results all endorse the practical advantage of DADA.

7. ACKNOWLEDGEMENT

Dong Liu is supported by NSF China Grant 61331017. Qing Ling is supported by NSF China grants 61573331 and U1811464.

8. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] Luis Perez and Jason Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[3] Song Han, Jeff Pool, John Tran, and William Dally, "Learning both weights and connections for efficient neural network," in *NIPS*, 2015, pp. 1135–1143.

[4] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 625–660, 2010.

[5] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng, "Self-taught learning: Transfer learning from unlabeled data," in *ICML*, 2007, pp. 759–766.

[6] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-supervised learning with deep generative models," in *NIPS*, 2014, pp. 3581–3589.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[8] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[9] Augustus Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training GANs," in *NIPS*, 2016, pp. 2234–2242.

[11] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan R Salakhutdinov, "Good semi-supervised learning that requires a bad GAN," in *NIPS*, 2017, pp. 6513–6523.

[12] Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré, "Learning to compose domain-specific transformations for data augmentation," in *Advances in Neural Information Processing Systems*, 2017, pp. 3239–3249.

[13] Zhangyang Wang, Jianchao Yang, Hailin Jin, Eli Shechtman, Aseem Agarwala, Jonathan Brandt, and Thomas S Huang, "Deepfont: Identify your font from an image," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 451–459.

[14] Tuan Anh Le, Atilim Giineş Baydin, Robert Zinkov, and Frank Wood, "Using synthetic data to train neural networks is model-based reasoning," in *IJCNN*, 2017, pp. 3514–3521.

[15] Leon Sixt, Benjamin Wild, and Tim Landgraf, "Rendergan: Generating realistic labeled data," *arXiv preprint arXiv:1611.01331*, 2016.

[16] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb, "Learning from simulated and unsupervised images through adversarial training," in *CVPR*, 2017, vol. 3, p. 6.

[17] Xinlong Wang, Zhipeng Man, Mingyu You, and Chunhua Shen, "Adversarial generation of training examples: Applications to moving vehicle license plate recognition," *arXiv preprint arXiv:1707.03124*, 2017.

[18] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.

[19] Søren Hauberg, Oren Freifeld, Anders Boesen Lindbo Larsen, John Fisher, and Lars Hansen, "Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation," in *Artificial Intelligence and Statistics*, 2016, pp. 342–350.

[20] Shibani Santurkar, Ludwig Schmidt, and Aleksander Madry, "A classification-based perspective on GAN distributions," *arXiv preprint arXiv:1711.00970*, 2017.

[21] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[22] M Heath, K Bowyer, D Kopans, R Moore, and P Kegelmeyer, "The digital database for screening mammography," *Digital mammography*, pp. 431–434, 2000.

[23] R Sawyer Lee, F Gimenez, A Hoogi, and D Rubin, "Curated breast imaging subset of DDSM," *The Cancer Imaging Archive*, 2016.