ENSEMBLES OF GANS FOR SYNTHETIC TRAINING DATA GENERATION

Gabriel Eilertsen[1,2], **Apostolia Tsirikoglou**[1], **Claes Lundström**[1,2,3], **Jonas Unger**[1,2]

[1] Department of Science and Technology, Linköping University, Sweden
[2] Center for Medical Image Science and Visualization, Linköping University, Sweden
[3] Sectra AB, Sweden

ABSTRACT

Insufficient training data is a major bottleneck for most deep learning practices, not least in medical imaging where data is difficult to collect and publicly available datasets are scarce due to ethics and privacy. This work investigates the use of synthetic images, created by generative adversarial networks (GANs), as the only source of training data. We demonstrate that for this application, it is of great importance to make use of multiple GANs to improve the diversity of the generated data, i.e. to sufficiently cover the data distribution. While a single GAN can generate seemingly diverse image content, training on this data in most cases lead to severe over-fitting. We test the impact of ensembled GANs on synthetic 2D data as well as common image datasets (SVHN and CIFAR-10), and using both DCGANs and progressively growing GANs. As a specific use case, we focus on synthesizing digital pathology patches to provide anonymized training data.

1 INTRODUCTION

A deficiency of training data is often limiting the performance of deep learning models, especially in areas such as medical imaging, where acquisition and annotation is highly time-consuming and reliant on busy experts. One potential solution is to use generative techniques to synthesize training data. The high quality of images from generative adversarial networks (GANs) (Goodfellow et al., 2014; Karras et al., 2019; Brock et al., 2019; Miyato et al., 2018; Zhang et al., 2019) have proven effective for image augmentation in deep learning for medical imaging applications (Frid-Adar et al., 2018; Madani et al., 2018; Bowles et al., 2018). Moreover, the possibility of creating purely synthetic collections representative of non-shareable patient data would be an effective anonymization approach (Guibas et al., 2017; Shin et al., 2018; Triastcyn & Faltings, 2019; Yoon et al., 2020).

We focus on the impact of using GANs in ensembles for the purpose of synthesizing training data for downstream deep learning applications. In particular, we are interested in the scenario where releasing the real data is not an option, e.g. if this is of private/sensitive nature or in other ways protected. This problem formulation differs significantly from considering training on real and synthetic data in combination, i.e. using GANs as a tool for data augmentation, where the real data can carry the bulk of information while the synthetic data provide some variations to improve generalization. Requiring the synthetic data to carry all information makes diversity a critical aspect, and missing modes of the underlying data distribution can have a large influence. Previous work has focused on techniques for increasing the mode coverage of GANs (Lin et al., 2018; Hoang et al., 2018; Liu et al., 2020). A powerful solution is to train ensembles of multiple GANs (Wang et al., 2016; Tolstikhin et al., 2017; Grover & Ermon, 2018), and boosting strategies have been considered for forcing the GANs to focus on different parts of the data distribution. While boosting makes intuitive sense, we show that for the purpose of generating training data the most important aspect is to combine multiple GANs; these could even be trained completely independently, meaning that the stochasticity of the optimization is powerful enough to allow for good mode coverage.

Our contributions include: **1)** we perform an evaluation of the behavior of deep classifiers when trained on purely GAN generated training data, **2)** we highlight the importance of using ensembles of GANs for synthesizing datasets diverse enough for training deep models, and show that independently trained GANs could be advantageous compared to boosting strategies, with as good or better performance and less prone to over-fitting, and **3)** we test ensemble performance on 2D data, on SVHN and CIFAR-10, as well as in a more realistic scenario of anonymization in digital pathology.

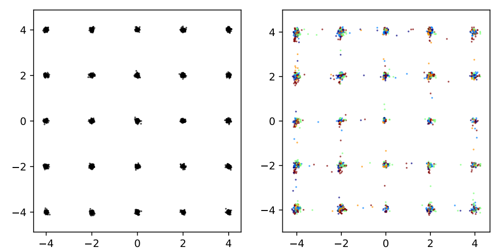


Figure 1: Ground truth (left) and generated by EnsGAN-5 (right). Colors correspond to different GANs in the ensemble.

	Modes (max 25)	High quality samples (%)
GAN	18.8 ± 2.6	75.7 ± 4.6
PacGAN4	24.8 ± 0.2	93.6 ± 0.6
MGAN	22.5 ± 0.6	51.1 ± 1.6
ScGAN	25.0 ± 0.0	99.5 ± 0.1
EnsGAN-2	24.6 ± 0.8	75.1 ± 3.4
EnsGAN-3	24.8 ± 0.7	75.3 ± 2.9
EnsGAN-4	25.0 ± 0.1	75.2 ± 2.6
EnsGAN-5	25.0 ± 0.0	75.3 ± 2.3

Table 1: Covered modes and fraction of high quality points, comparing different ensemble sizes to a selection of previous methods.

2 ENSEMBLE GANS

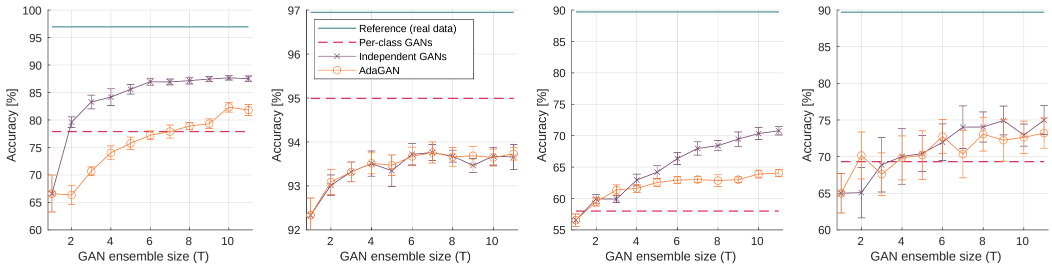
A GAN generator $G(z)$ produces synthetic image samples by drawing from the latent distribution $z \sim \mathcal{Z}$. We define an ensemble of T GANs as the mixture $\hat{G}_T = \sum_{t=1}^{T} p_t G_t$, i.e. a weighted combination of individually trained GANs G_t, where $\sum_t p_t = 1$. In practice, generating images from $\hat{G}_T$ entails drawing individual samples from G_t with probability p_t. This means that a generated dataset of N_g samples will contain $p_t N_g$ samples from G_t. In practice, we consider classification as the downstream application for the synthetic data. This means that we have access to class labels $k \in [1, K]$. To make use of this information, we use bootstrap aggregation (bagging), by training GANs separately for each class k in each ensemble iteration. Given that we strive for maximal diversity and mode coverage, it is a sensible choice to reflect the different parts of the data distribution covered by the individual classes of the dataset. This means that in each ensemble iteration we train K models $G_{t,k}$, and the total mixture is $\hat{G}_T = \frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T} p_t G_{t,k}$. Training a separate GAN for each class means, in practice, that we will have an ensemble/mixture of KT GANs. However, for better clarity we will only refer to the iterations T.

We are mainly interested in independent ensembles, where G_t are trained in isolation from each other. This means that the benefit of combining multiple GANs comes entirely from the stochastic nature of the optimization, which will make each G_t focus on slightly different parts of the data distribution. We compare the naive approach to a sophisticated boosting scheme, AdaGAN (Tolstikhin et al., 2017), where the training samples are re-weighted for each G_t based on discriminator score. The AdaGAN re-weighting puts more focus on the images that the discriminator can easily single out as real, meaning that the current ensemble at a certain ensemble iteration is unable to cover the distribution around these images. For both ensemble approaches in consideration we use $p_t = 1/T$ for equal contribution from each G_t. Moreover, we always keep the number of generated training images fixed to the same number as in the real dataset. That is, if we use N real images, the synthetic dataset from a GAN ensemble with T generators will use N/T images from each G_t.

3 EVALUATION

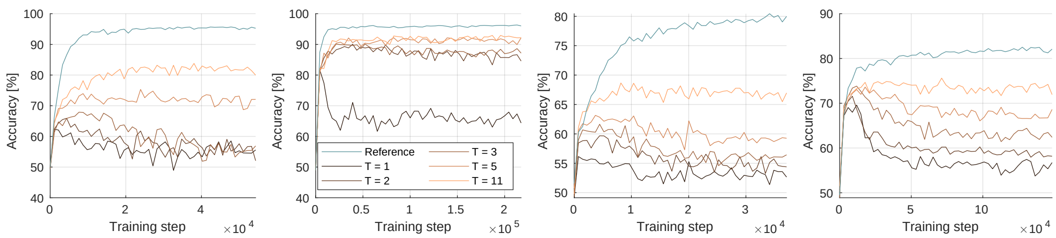
In order to get a sense for the impact of GANs in ensembles on the quality of synthetic training data, we perform a series of experiments on data of increasing difficulty.

Experiment 1 – Synthetic 2D data As a simple way of analyzing the mode coverage capabilities, we perform an initial experiment on 2D data. We follow the same exact setup as reported in (Lin et al., 2018), which has been used in other evaluations as well (Hoang et al., 2018; Liu et al., 2020). The dataset is generated by drawing from a mixture of 25 Gaussians arranged in a grid, each representing a separate mode. A generated point is said to be of high quality if it falls within 3 standard deviations from the center of a mode, and a mode is recovered if it has at least one high quality point. We compare to vanilla GAN (Goodfellow et al., 2014) and previously reported numbers for PacGAN (Lin et al., 2018), MGAN (Hoang et al., 2018), and self-conditioned GAN (ScGAN) (Liu et al., 2020). We refer to an ensemble as *EnsGAN-T*, where T is the number of GANs.



(a) SVHN-II – DCGAN (b) SVHN-II – PG-GAN (c) CIFAR-II – DCGAN (d) CIFAR-II – PG-GAN

Figure 2: Classification performance on synthetic datasets, comparing different ensemble sizes and approaches. Each datapoint has been estimated from the mean of 10 separate trainings, and standard deviations are reported with error bars.



(a) SVHN-II – DCGAN (b) SVHN-II – PG-GAN (c) CIFAR-II – DCGAN (d) CIFAR-II – PG-GAN

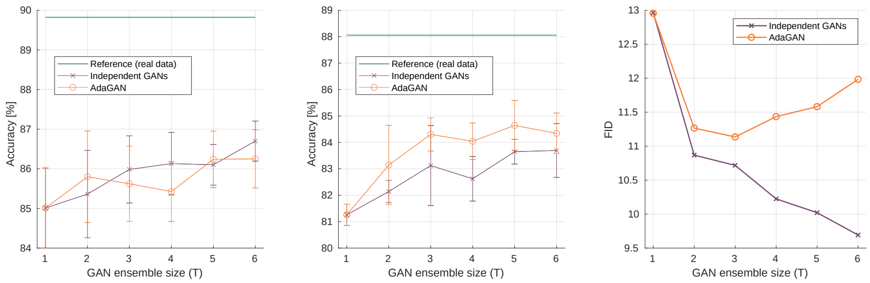
Figure 3: Test accuracy over training steps, for different ensemble sizes. There is a high tendency to overfit with few GANs in an ensemble (small T), and the variance between training steps is high.

The ground truth and generated data are displayed in Figure 1. Mode coverage and the fraction of high quality points are presented in Table 1. For this dataset 2 or 3 GANs is sufficient to effectively cover all 25 modes. However, while the modes are easily recovered the quality is not affected. We believe that this is not a negative thing since a "tightening" around each mode, as can be seen e.g. for ScGAN (Liu et al., 2020), could potentially affect the diversity although all modes are recovered.

Experiment 2 – SVHN/CIFAR-10 For SVHN (Netzer et al., 2011) and CIFAR-10 (Krizhevsky & Hinton, 2009), each consisting of 10 classes, we partition the data into two separate classes. In SVHN, one class contains digits 0-4 and one uses 5-9. In CIFAR-10 one class is $\{$*airplane*, *automobile*, *bird*, *cat*, *deer*$\}$, and the other uses $\{$*dog*, *frog*, *horse*, *ship*, *truck*$\}$. We call these two-class datasets SVHN-II and CIFAR-II, respectively. For reference, we also train GANs separately on each of the original 10 classes, followed by partitioning the generated data into the aforementioned classes. We train ensembles using both DCGAN (Radford et al., 2015) and progressively growing GAN (PG-GAN) (Karras et al., 2018), which allows us to analyze ensemble performance under different GAN complexities. For details on experimental setup, we refer to the supplementary material.

Figure 2 shows the accuracy on the downstream classification task, evaluated using a ResNet-18 (He et al., 2016). Overall there is a significant improvement when utilizing increasingly large ensembles of GANs, effectively reducing the gap to the performance on real data, especially for DCGAN. In none of the experiments the more advanced AdaGAN boosting performed significantly better, and for the considered problem independently trained GANs is the better option at least for DCGAN. Moreover, and perhaps the most interesting finding, is how ensembles of DCGAN can outperform a single PG-GAN on CIFAR-II. Since the quality of individual images is not affected by ensembles, this points to how important the improved diversity is for the quality of the synthetic dataset.

To further emphasize the importance of using GANs in ensembles when generating synthetic training data, Figure 3 shows the test performance on real data over training iterations with synthetic data. For both DCGAN and PG-GAN it is problematic to train on synthetic data from a single GAN, experiencing severe problems with over-fitting and stability. Including more models in the ensemble effectively stabilizes the training to better reflect how training on real data behaves.



(a) Accuracy, ResNet-18 (b) Accuracy, vanilla CNN (c) FID

Figure 4: Tumor classification performance on synthetic pathology datasets (a-b), averaged over 10 separate trainings, and FID score evaluated using a model pre-trained on Imagenet (c).

Experiment 3 – digital pathology As a more challenging scenario, we look at anonymization of the CAMELYON17 dataset (Litjens et al., 2018). This consists of 50 hematoxylin and eosin (H&E) stained lymph node whole-slide images containing tumor areas, and was sampled to produce 50K tumor and 50K non-tumor patches at a resolution of 128×128. A separate test set with 15K/15K tumor/non-tumor patches was extracted from slides different from those used to produce the training data. For this dataset, we trained ensembles using PG-GAN, and refer to the supplementary material for details on the experimental setup.

Figure 4 shows the downstream classification performance for a ResNet-18, a vanilla CNN, as well as the Fréchet Inception Distance (FID) (Heusel et al., 2017), for different sizes of ensembles. There are consistent improvements in performance on both classifiers, although most significant for the vanilla CNN, and independent ensembles show a similar performance as AdaGAN. While there is a large discrepancy in FID between independent ensembles and AdaGAN, this is not reflected in the classification performance, pointing to the weakness of FID as a measure of training dataset quality.

4 Discussion and Conclusion

We have provided empirical evidence for the importance of using GANs in ensembles to enable synthetic training data that can be used in isolation from the real data, e.g. in order to act as an anonymization method. The experiments point to how independently trained GANs can be a simple yet powerful technique to improve quality, especially for complex data and simpler GAN models. For a simpler dataset and state-of-the-art GAN, such as PG-GAN on SVHN-II, the impact is less pronounced. However, there is a substantial difference in training behavior when comparing different ensemble sizes (Figure 3b). There is also a significant gap to the performance on real data, calling for further work to investigate techniques that can improve GAN synthesized training data.

One benefit of using independently trained GANs is that this technique does not risk focusing on a too narrow part of the data distribution. This could potentially be a problem with boosted GANs, leading to over-fitting to the exact images used to train the GAN (we refer to the supplementary material for an example), which is not acceptable for anonymization applications. We have also pointed to how FID is not very representative for training data quality (see Figure 4 and supplementary material). There has been some studies on training on synthetic data and testing on real data and vice versa (Shmelkov et al., 2018), which is a better indicator of the dataset quality. Nevertheless, there is a need for formulating standardized quality measures tailored to this specific application of generative models, since the aim most likely differs from many of the existing GAN quality metrics (Borji, 2019). Finally, we see a need for a more comprehensive evaluation of existing techniques for diversifying GANs in the context of synthetic training data generation.

Acknowledgments

This project was supported by VINNOVA grant 2019-05144 and grant 2017-02447 (AIDA), the Wallenberg Autonomous Systems and Software Program (WASP-NTU and WASP-AI), and the strategic research environment ELLIIT.

REFERENCES

Ali Borji. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. GAN augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems (NIPS)*, 2014.

Aditya Grover and Stefano Ermon. Boosted generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

John T Guibas, Tejpal S Virdi, and Peter S Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems (NeurIPS)*, pp. 6626–6637, 2017.

Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. MGAN: Training generative adversarial nets with multiple generators. In *International conference on learning representations (ICLR)*, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

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

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. *Advances in neural information processing systems (NeurIPS)*, 2018.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6):giy065, 2018.

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14286–14295, 2020.

Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In *Medical Imaging 2018: Image Processing*, volume 10574, pp. 105741M, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

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*, 2011.

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

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pp. 1–11. Springer, 2018.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my GAN? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2018.

Ilya O Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. AdaGAN: Boosting generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5424–5433, 2017.

Aleksei Triastcyn and Boi Faltings. Generating artificial data for private deep learning. *Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies, AAAI Spring Symposium Series*, 2019.

Yaxing Wang, Lichao Zhang, and Joost Van De Weijer. Ensembles of generative adversarial networks. *arXiv preprint arXiv:1612.00991*, 2016.

J. Yoon, L. N. Drumright, and M. van der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pp. 7354–7363, 2019.