IMPROVING THE VISUAL QUALITY OF GENERATIVE ADVERSARIAL NETWORK (GAN)-GENERATED IMAGES USING THE MULTI-SCALE STRUCTURAL SIMILARITY INDEX

Parimala Kancharla, Sumohana S. Channappayya

Department of Electrical Engineering, IIT Hyderabad, Kandi - 502285, Telangana, India.
Email: {ee15mtech11024, sumohana}@iith.ac.in

ABSTRACT

This paper presents a simple yet effective method to improve the visual quality of Generative Adversarial Network (GAN) generated images. In typical GAN architectures, the discriminator block is designed mainly to capture the class-specific content from images without explicitly imposing constraints on the visual quality of the generated images. A key insight from the image quality assessment literature is that natural scenes possess a very unique local structural and (hence) statistical signature, and that distortions affect this signature. We translate this insight into a constraint on the loss function of the discriminator in the GAN architecture with the goal of improving the visual quality of the generated images. Specifically, this constraint is based on the Multi-scale Structural Similarity (MS-SSIM) index to guarantee local structural and statistical integrity. We train GANs (Boundary Equilibrium GANs, to be precise,) using the proposed approach on popular face and car image databases and demonstrate the improvement relative to standard training approaches both visually and quantitatively.

Index Terms— Generative Adversarial Networks (GANs), Multi-scale Structural Similarity, Natural Scene Statistics.

1. INTRODUCTION

Generative modeling is an unsupervised learning method, whose goal is to generate data that resembles the large corpus of unlabeled training data it learns from. Goodfellow *et al.* [1] proposed Generative Adversarial Networks (GANs) that are generative models designed to learn the probability distribution of data that is aided by adversarial learning. Various GAN architectures have since been proposed in the literature and we briefly introduce them in the following in order to place the proposed approach in context.

Deep Convolutional GAN (DCGAN) [2] is the first GAN architecture designed using convolutional neural networks (CNNs). The generator objective function in the DCGAN and earlier GAN architectures is inspired by the Jensen-Shannon (JS) divergence. This leads to instability in training the GAN due to the vanishing gradient problem. Wasserstein GAN (WGAN) [3] addressed the problem of instability by introducing the Wassertein-1 distance into GAN objective as an alternative to JS-divergence. WGAN solved the problem of instability and it showed a better mode coverage but at the expense of slow training speed. Boundary Equilibrium GAN (BEGAN) [4] is an autoencoder based GAN whose loss is derived from the Wassertein distance. It provides a stable training procedure in addition to good visual quality images (without explicitly using a visual quality constraint in the optimization).

Vertolli *et al.* [5] proposed a training approach for BEGAN by modifying the distance metric a combination of l_1 score, Gradient Magnitude Similarity Mean (GMSM) score and chrominance score. They have reported the results by varying the weights for individual distance metrics. They showed their improvement by visually comparing gradient maps and chrominance maps of the input and the decoded images by their proposed discriminator. Zeng *et al.* [6] analyzed the statistics of DCGAN and WGAN images. The statistics include mean power spectrum, number of connected components in a given image area, distribution of random filter response and contrast distribution. They observed significant differences in the statistics of natural images and deep generated images. Further, they have addressed this problem by replacing the deconvolution layers with sub-pixel convolution in the GAN architecture. Their work concludes that there is a need for designing new loss functions for GANs to improve visual quality.

Snell *et al.* [7] demonstrated that autoencoders give superior results when trained using the MS-SSIM index [8] instead of the standard L_1 or L_2 distances. They showed improved performance using their optimized networks in case of image super resolution. Dosovitskiy and Brox [9] have proposed a class of loss functions which they call deep perceptual similarity metrics (DeepSIM). Instead of computing the distances in the image space they compute the distances between image features extracted by deep neural networks. They have showed that their proposed metrics reflect the perceptual similarity well and they demonstrated their proposed metrics by optimizing the variational autoecoder (VAE) [10] to gener-

ate realistic images. The VAE is also an image generation model that is trained by maximizing the lower bound of the log likelihood for data distribution. Though the VAE offers a tractable recognition model compared to GAN, the images generated from VAE are blurry and have poor visual quality. In summary, existing GAN architectures do not include visual quality constraints in their design. While Dosovitskiy and Brox [9] have demonstrated improved visual quality using a local image statistics constraint in the loss function, an explicit use of image quality assessment metrics in the design of GANs has not been made yet (to the best of our knowledge). In this work, we propose a simple yet effective training approach for BEGANs to improve the visual quality of images by introducing the MS-SSIM index as a constraint in the discriminator's objective function. We demonstrate the effectiveness of this approach both qualitatively and quantitatively. We describe the standard GAN and BEGAN theory next and present our proposed approach subsequently.

2. PROPOSED APPROACH

We revisit the GAN and BEGAN models to help setup the necessary background and notation for the proposed approach.

2.1. Generative Adversarial Network (GAN) [1]

A GAN is composed of two models: the generator model (G) and the discriminator model (D). The generator model $G(z;\theta_G)$ with parameters θ_G maps samples (z) from the noise distribution $p(z)$ to the true data distribution $p_{model}(x)$. The discriminator model $D(x;\theta_D)$ represents the probability that the sample (x) belongs to the true data distribution $p_{data}(x)$ rather than $p(z)$. The generative model $G(z;\theta_G)$ and discriminative model $D(x;\theta_D)$ both are learned jointly by alternating the training of D and G described by a two player minimax game. D is trained to maximize the probability of assigning the correct label to both training examples and samples from G. G is simultaneously trained to minimize $\log(1 - D(G(z)))$. Thus, $G(z;\theta_G) : z \to x$ and $D(x;\theta_D) : x \to [0, 1]$ play the following two-player minimax game with value function $V(G; D)$:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log(D(x))] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where $p_{data}(x)$ is the true data distribution and $p_z(z)$ is a simple noise distribution to draw samples for e.g. Uniform or Gaussian distribution.

2.2. Boundary Equilibrium GAN (BEGAN) [4]

BEGAN is an extension of GAN (and a variant of WGAN [3]) where the discriminator block is replaced with an autoencoder. If x is a real sample of dimension N_x (i.e., $x \in R^{N_x}$), $D(x) : R^{N_x} \to R^{N_x}$ the discriminative model, is an autoencoder. BEGAN aims to match the autoencoder loss distributions using the Wassertstein distance measure which helps with stable convergence.

The loss function $L : R^{N_x} \to R^{N_x}$ used to train a pixel-wise autoencoder is defined as

$$L(x) = |x - D(x)|^n; n = 1, 2, \quad (2)$$

where $L(x)$ is the loss for real sample x and $L(G(z))$ is the loss for generated sample $G(z)$. Let m_1 and m_2 be the representative means of distributions of loss functions $L(x)$ and $L(G(z))$ respectively. In order to match the distributions of loss functions, the difference of means of two loss functions is minimized. The main goal of the discriminator is to differentiate the real samples from generated samples well. Therefore, the discriminator should be trained to mismatch the distributions of $L(x)$ and $L(G(z))$. This turns to be maximizing the difference between representative means m_1 and m_2. One possible way is minimizing m_1 and maximizing m_2.

The objective function of BEGAN then becomes

$$\begin{aligned} L_D &= L(x) - k_t L(G(z)) \text{ for } \theta_D, \\ L_G &= L(G(z)) \text{ for } \theta_G, \\ k_{t+1} &= k_t + \lambda_k(\gamma L(x) - L(G(z))) \text{ for training step } t, \end{aligned} \quad (3)$$

where θ_G and θ_D are the parameters of generator and discriminator respectively, and are updated by minimizing the loss functions L_D and L_G respectively. k_t is the variable to control how much emphasis should be put on $L(G(z))$ during gradient descent. λ_k is the proportional gain for k.

In the standard BEGAN setting, the autoencoder loss in the discriminator is defined by (2), i.e., either Mean Absolute Difference (MAD) or Mean Squared Error (MSE). It is well known in the image quality assessment literature that MAD/MSE are not ideally suited for perceptual quality estimation [11]. Thus, the BEGAN cost functions as defined in (2), (3) provide the motivation for our proposed approach and is described next.

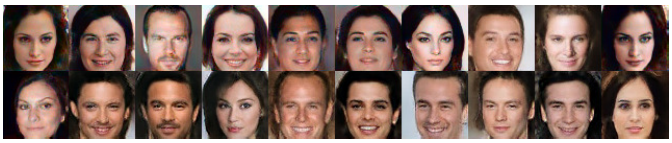
2.3. Proposed MS-SSIM index Constrained BEGAN

As widely noted in the image quality assessment literature, natural scenes possess a very unique local structural and statistical signature [14]. To ensure that the structural information is preserved well in the BEGAN discriminator's autoencoder, we propose to modify the loss function in (2) to include a structural integrity constraint term. Specifically, the MS-SSIM index [8] is introduced into the loss function given its proven strength as a robust image quality assessment algorithm. The training aims to maximize the MS-SSIM between input and reconstructed image. Therefore, the loss function of the BEGAN's discriminator is modified to be a weighted average of MAD and 1-(MS-SSIM) and is defined as

$$L(x) = \lambda_1 L_1(x) + \lambda_2 L_2(x), \quad (4)$$

(a) BEGAN-MAD ($\lambda_1 = 1$, $\lambda_2 = 0$) based approach.

(b) BEGAN-MAD ($\lambda_1 = 1$, $\lambda_2 = 0$) based approach.



(c) BEGAN-MAD+MS-SSIM ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$) based approach.

(d) BEGAN-MAD+MS-SSIM ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$) based approach.

Fig. 1: Randomly selected BEGAN generated images trained on the CelebA dataset [12] (left image set) and the Stanford Cars dataset [13] (right image set). The top row images correspond to the standard BEGAN approach while the bottom row images correspond to the proposed MS-SSIM index constrained approach.

where λ_1 and λ_2 are normalized weights given to each of the metrics. Specifically, we choose $0 \leq \lambda_1, \lambda_2 \leq 1$ and $\lambda_2 = 1 - \lambda_1$. Further,

$$\begin{aligned} L_1(x) &= |x - D(x)|^n, \\ L_2(x) &= 1 - (\text{MS-SSIM}(x, D(x))). \end{aligned} \tag{5}$$

We have chosen $n = 1$ in this work (i.e., MAD). Also, we have considered the gray scale MS-SSIM index. The MS-SSIM index is computed over $M = 5$ downscaling steps using same β, γ parameters proposed in [8]. The loss function in (4) does not lend itself to a closed form solution given the non-convex nature of the MS-SSIM index. This in turn forces us to choose λ_1 and λ_2 empirically. This and other implementation details are discussed next along with a presentation of the results of the proposed method.

3. RESULTS AND DISCUSSION

We implemented our proposed approach using the standard BEGAN architecture proposed in [4]. The discriminator is an autoencoder with both deep encoder and decoder. The generator has the same architecture of the discriminator's decoder with different weights. Convolutions are 3×3 in size with exponential linear units applied at their outputs. In our experiments, we have used λ_k to be 0.001 and k_0 to be 0. As noted in the previous section, the non-convex nature of our loss function forces us to report our performance by manually varying the weights λ_1 and λ_2.

The proposed method has been evaluated qualitatively and quantitatively on two databases. CelebA dataset [12] is the standard database for evaluating GAN frameworks in the literature. This dataset has more than 202,000 face images. We have also used Stanford Cars [13] dataset. This car dataset is composed of 16185 images. We trained our models with a batch size of 64 for 633122 and 50500 iterations for the CelebA and Cars datasets respectively.

3.1. Performance Evaluation

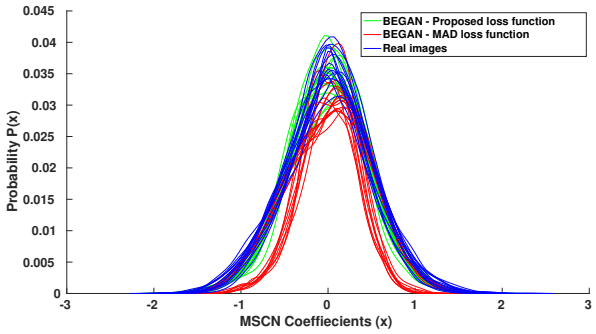


Fig. 2: Normalized histograms of MSCN coefficients.

For a fair comparison of our approach with the standard BEGAN approach, we have used a recently proposed and more consistent metric called the Frechet Inception Distance (FID) [15] for evaluating GANS. The statistically inspired FID score is also appropriate to our setting given our claim that the local structural constraint imposed in our method also leads to improved local image statistics. Futher, the FID score is shown to be more consistent than the popularly used inception score [16] .

To measure the quality of the generated samples using FID, they are passed through the Inception net to compute the feature space from the specific layer. Then, treating the embedding layer as a continuous multivariate Gaussian, the mean and covariance is estimated for both the generated data and the real data. The Frechet distance between these two Gaussians is then used to quantify the quality of the samples.

$$FID(x, g) = ||\mu_x - \mu_g||_2^2 + \text{Tr}\big(\Sigma_x + \Sigma_g - 2\big(\Sigma_x \Sigma_g\big)^{\frac{1}{2}}\big) \tag{6}$$

where (μ_x, Σ_x) and (μ_g, Σ_g) are the mean vector and the covariance matrix of the sample embeddings from the real data distribution and the generated data distribution, respectively.

FID scores are negatively correlated with visual quality. We have used more than 3000 generated images from the proposed BEGAN and more than 3000 real images to compute the FID scores.

We also used the Natural Image Quality Evaluator (NIQE) [17] score for evaluating the proposed approach. NIQE is a popular no reference image quality assessment technique, which computes image quality by measuring the deviation in the statistics of mean subtracted contrast normalization (MSCN) from pristince natural image statistics. The standard NIQE implementation is used for computing NIQE score of the 3000 generated images. The mean value of all the NIQE scores is reported for both the databases for the proposed model. As with the FID score, the NIQE score is negatively correlated with visual quality.

Along the lines of NIQE scores, we further demonstrate the improvement offered by the proposed method using the normalized histograms of the generated images. Fig. 2 shows the normalized histograms of the mean subtracted contrast normalized (MSCN) coefficients of real images, and the images generated by the standard BEGAN and the proposed approaches. A well-known result in natural scene modeling is that the MSCN coefficients follow a Gaussian distribution [18]. From the figure, there is a noticeable difference in the statistics of MSCN coefficients of real images and standard BEGAN generated images. The proposed method is able to better match the generated image statistics to real image statistics, thereby highlighting its potential in improving visual image quality.

Tables 1 and 2 show the performance of the proposed approach for various combinations of λ_1 and λ_2. From the results, it is observed that both the FID scores and NIQE scores are the lowest for the $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ case. All other tested combinations of λ_1 and λ_2 including the MAD-only and the MS-SSIM-only case showed inferior performance.

Through our experiments we observed that as the weight assigned to the MS-SSIM index increases, our model starts to become unstable. To address this issue, we had to reduce the learning rate (γ). As the learning rate decreases, image diversity also decreases. As this is a trade-off, equal weightage to MAD and MS-SSIM gave us the best results in terms of both diversity and image quality.

Finally, Fig. 1 provides qualitative evidence for the improvement provided by the proposed approach. Figs. 1a and 1b show a set of face and car images generated by the standard BEGAN approach. Figs. 1c and 1d show a set of face and car images generated by the proposed approach. While the improvement in the face images is subtle (cleaner faces and hair), the improvement in the car images is very obvious. The improved shape of the cars (well-defined shapes and boundaries) and the details in the wheels support our claim of improved visual quality. These structural improvements can be attributed to the MS-SSIM index constraint in our formulation.

Table 1: Proposed BEGAN results on CelebA dataset [12].

Model Parameters		FID	NIQE
λ_1(MAD)	λ_2(1-(MS-SSIM))		
1	0	77.41	8.53
0.9	0.1	72.91	8.93
0.5	0.5	64.96	7.61
0.1	0.9	71.35	8.54
0	1	70.72	8.83

Table 2: Proposed BEGAN results on Cars dataset [13].

Model Parameters		FID	NIQE
λ_1 (MAD)	λ_2 (1-(MS-SSIM))		
1	0	235.89	9.72
0.9	0.1	245.57	8.42
0.5	0.5	205.03	7.33
0.1	0.9	268.52	8.14
0	1	235.00	8.52

Through these quantitative and qualitative comparisons, we have shown that the proposed approach is able to clearly improve the visual quality of generated images.

4. CONCLUSIONS AND FUTURE WORK

We presented a training approach for BEGANs which imposes constraints on the local image structure and statistics in addition to constraints related to class information. Specifically, we used the MS-SSIM index to improve visual quality in combination with the typical class related distance (MAD). Through this preliminary attempt, we demonstrate that explicitly integrating an image quality assessment model into the image generation model is indeed promising. We provided corroborative evidence both quantitatively (in terms of the FID and NIQE scores) and qualitatively on two popular image datasets. We also showed that the local image statistics of the proposed method are more closely matched with natural images compared to the standard BEGAN approach.

As future work, we plan to build on these preliminary results by leveraging the rich literature on natural scene statistics models as well on the insights gained from the image quality assessment literature.

5. ACKNOWLEDGEMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

6. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[2] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 214–223. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html

[4] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: boundary equilibrium generative adversarial networks," *CoRR*, vol. abs/1703.10717, 2017. [Online]. Available: http://arxiv.org/abs/1703.10717

[5] M. O. Vertolli and J. Davies, "Image quality assessment techniques show improved training and evaluation of autoencoder generative adversarial networks," *CoRR*, vol. abs/1708.02237, 2017. [Online]. Available: http://arxiv.org/abs/1708.02237

[6] Y. Zeng, H. Lu, and A. Borji, "Statistics of deep generated images," *CoRR*, vol. abs/1708.02688, 2017. [Online]. Available: http://arxiv.org/abs/1708.02688

[7] K. Ridgeway, J. Snell, B. Roads, R. S. Zemel, and M. C. Mozer, "Learning to generate images with perceptual similarity metrics," *arXiv preprint arXiv:1511.06409*, 2015.

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. Ieee, 2003, pp. 1398–1402.

[9] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 658–666. [Online]. Available: http://papers.nips.cc/paper/6158-generating-images-with-perceptual-similarity-metrics-based-on-deep-networks.pdf

[10] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in neural information processing systems*, 2016, pp. 2352–2360.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[13] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[14] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6629–6640. [Online]. Available: http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium.pdf

[16] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *CoRR*, vol. abs/1606.03498, 2016. [Online]. Available: http://arxiv.org/abs/1606.03498

[17] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[18] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," in *Advances in neural information processing systems*, 1994, pp. 551–558.