LEARNING TO GENERATE IMAGES WITH PERCEPTUAL SIMILARITY METRICS

Jake Snell
University of Toronto

Karl Ridgeway
University of Colorado, Boulder

Renjie Liao
University of Toronto

Brett D. Roads
University of Colorado, Boulder

Michael C. Mozer
University of Colorado, Boulder

Richard S. Zemel
University of Toronto

ABSTRACT

Deep networks are increasingly being applied to problems involving image synthesis, e.g., generating images from textual descriptions and reconstructing an input image from a compact representation. Supervised training of image-synthesis networks typically uses a pixel-wise loss (PL) to indicate the mismatch between a generated image and its corresponding target image. We propose instead to use a loss function that is better calibrated to human perceptual judgments of image quality: the multiscale structural-similarity score (MS-SSIM) [1]. Because MS-SSIM is differentiable, it is easily incorporated into gradient-descent learning. We compare the consequences of using MS-SSIM versus PL loss on training autoencoders. Human observers reliably prefer images synthesized by MS-SSIM-optimized models over those synthesized by PL-optimized models, for two distinct PL measures (L_1 and L_2 distances). We also explore the effect of training objective on image encoding and analyze conditions under which perceptually-optimized representations yield better performance on image classification. Finally, we demonstrate the superiority of perceptually-optimized networks for super-resolution imaging. We argue that significant advances can be made in modeling images through the use of training objectives that are well aligned to characteristics of human perception.

Index Terms— Perceptual Losses, Deep Learning

1. INTRODUCTION

There has been a recent explosion of interest in developing methods for image representation learning, focused in particular on training neural networks to synthesize images. Surprisingly little work has been done to study loss functions that are appropriate for image generation. A basic method for learning generative image models is the *autoencoder* architecture. Autoencoders are made up of two functions, an encoder and a decoder. The encoder compresses an image into a feature vector, typically of low dimension, and the decoder takes that vector as input and reconstructs the original image as output. The standard loss function is the squared Euclidean (L_2) distance between the original and reconstructed images, also referred to as the *mean squared error* or *MSE*. A city-block (L_1) distance is sometimes used as well, referred to as the *mean absolute error* or *MAE*. As we will show, both loss functions yield blurry results–synthesized images that appear to have been low-pass filtered.

In this paper, we explore loss functions that, unlike MSE and MAE, are grounded in human perceptual judgments. We show that these perceptual losses lead to representations are superior to other methods, both with respect to reconstructing given images and image classification. This superiority is demonstrated both in quantitative studies and human judgements. Beyond achieving perceptually superior synthesized images, we also demonstrate that perceptual losses yield a convincing win when applied to a state-of-the-art architecture for single image super-resolution.

2. BACKGROUND AND RELATED WORK

2.1. Neural Networks for Image Synthesis

The standard neural network for image synthesis is the autoencoder, in which the input is mapped directly through hidden layers to output a reconstruction of the original image. The autoencoder is trained to reproduce an image that is similar to the input, where similarity is evaluated using a pixel-wise loss between the image and its reconstruction. A second approach to building generative models for image synthesis uses variants of Boltzmann Machines [2, 3] and Deep Belief Networks [4]. While these models are very powerful, each iteration of training requires a computationally costly step of MCMC to approximate derivatives of an intractable partition function (normalization constant), making it difficult to scale them to large datasets. A third approach to learning generative image models, which we refer to as the *direct-generation* approach, involves training a generator that maps random samples drawn from a uniform distribution through a deep neural network that outputs images. Generative Adversarial Networks (GANs) [5, 6, 7] is a paradigm that involves training a discriminator that attempts to distinguish real from generated images, along with a generator that attempts to trick the

discriminator. Drawbacks of the GAN include the need to train a second network, a deep and complicated adversary, and the fact that the training of the two networks are interdependent and lack a single common objective. An alternative direct-generation approach, moment-matching networks [8], directly trains the generator to make the statistics of these two distributions match.

Because the goal of image generation is to synthesize images that humans would judge as high quality and natural, current approaches seem inadequate by failing to incorporate measures of human perception. In this paper, we describe an alternative approach using the autoencoder architecture. We focus on autoencoders over direct-generation approaches because autoencoders *interpret* images in addition to generating images. That is, an input image can be mapped to a compact representation that encodes the underlying properties of the world responsible for the observed image features. The encoder can thus be used as the initial image mapping that can be utilized for many different applications. Although adversarial training can be combined with autoencoding, here we explore autoencoding in isolation, to study the effects of optimizing with perceptually-based metrics.

Because autoencoders reconstruct training images, training the network requires evaluating the quality of the reconstruction with respect to the original. This evaluation is based on a pixel-to-pixel comparison of the images—a so-called *full-reference metric*. Autoencoders typically use *mean-squared error* (*MSE*), the average square of the pixel intensity differences, or *mean-absolute* error (*MAE*), the average of the absolute difference in pixel intensity. In many instances, these standard measures fail to capture human judgments of quality. For example, distorting an image with salt-and-pepper impulse noise obtains a small perturbation by standard measures but is judged by people as having low visual quality relative to the original image.

2.2. Perception-Based Error Metrics

As digitization of photos and videos became commonplace in the 1990s, the need for digital compression also became apparent. Lossy compression schemes distorted image data, and it was important to quantify the drop in quality resulting from compression in order to optimize the compression scheme. Researchers attempted to develop full-reference image quality metrics that take into account features to which the human visual system is sensitive and that ignore features to which it is insensitive. Some are built on complex models of the human visual system, such as the Sarnoff JND model [9], the visual differences predictor [10], the moving picture quality metric [11], the perceptual distortion metric [12], and the metric of [13].

Others take more of an engineering approach, and are based on the extraction and analysis of specific features of an image to which human perception is sensitive. The most popular of these metrics is the structural similarity metric (SSIM) [14], which aims to match the luminance, contrast, and structure information in an image. Alternative engineering-based metrics are the visual information fidelity metric [15] and the visual signal-to-noise ratio [16]. Some of these metrics have been used for optimization in traditional image reconstruction paradigms [17, 18], but not in the context of deep learning.

2.3. Structural Similarity

In this paper, we train neural nets with MS-SSIM [1], a multiscale extension of the structural-similarity metric (SSIM) [14]. We chose the SSIM family of metrics because it is well accepted and frequently utilized in the literature. Further, its pixelwise gradient has a simple analytical form and is inexpensive to compute. In this work, we focus on the original grayscale MS-SSIM, although there are interesting variations such as colorized versions [19, 20].

The SSIM family of metrics compares corresponding pixels and their neighborhoods in two images, denoted x and y, with three comparison functions—luminance (I), contrast (C), and structure (S):

$$I(x,y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1} \quad C(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}$$
$$S(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}$$

The variables μ_x, μ_y, σ_x, and σ_y denote mean pixel intensity and the standard deviations of pixel intensity in a local image patch centered at either x or y. Following [14], we chose a square neighborhood of 5 pixels on either side of x or y, resulting in 11×11 patches. The variable σ_{xy} denotes the sample correlation coefficient between corresponding pixels in the patches centered at x and y. The constants C_1, C_2, and C_3 are small values added for numerical stability. The three comparison functions are combined to form the SSIM score:

$$\text{SSIM}(x,y) = I(x,y)^\alpha C(x,y)^\beta S(x,y)^\gamma$$

This single-scale measure assumes a fixed image sampling density and viewing distance, and may only be appropriate for certain range of image scales. This issue is addressed in [1] with a variant of SSIM that operates at multiple scales simultaneously. The input images x and y are iteratively downsampled by a factor of 2 with a low-pass filter, with scale j denoting the original images downsampled by a factor of 2^{j-1}. The contrast $C(x,y)$ and structure $S(x,y)$ components are applied at all scales. The luminance component is applied only at the coarsest scale, denoted M. Additionally, a weighting is allowed for the contrast and structure components at each scale, leading to the definition:

$$\text{MS-SSIM}(x,y) = I_M(x,y)^{\alpha_M} \prod_{j=1}^{M} C_j(x,y)^{\beta_j} S_j(X,y)^{\gamma_j}$$

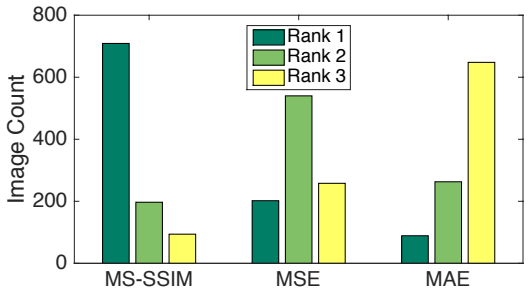

Fig. 1. Distribution of image quality ranking for MS-SSIM, MSE, and MAE on 1000 held-out STL-10 images.

In our work, we weight each component and each scale equally ($\alpha = \beta_{1..M} = \gamma_{1..M} = 1$), a common simplification of MS-SSIM. Following [1], we use $M = 5$ downsampling steps. Our objective is to minimize the loss related to the sum of structural-similarity scores across all image pixels,

$$\mathcal{L}(X, Y) = -\sum_i \text{MS-SSIM}(X_i, Y_i),$$

where X and Y are the original and reconstructed images, and i is an index over image pixels.

3. AUTOENCODER RECONSTRUCTIONS

We now turn experiments that compare autoencoders trained with a pixelwise loss (MSE and MAE) to those trained with a perceptually optimized loss (MS-SSIM). We trained networks on 96×96 images with a convolutional autoencoder architecture [21]: convolutional layers encode the input and deconvolutional layers decode the feature representation in the bottleneck layer. For training and testing, we use the STL-10 dataset [22], which consists of RGB color images from 10 categories. The images were converted to grayscale using the ITU-R 601-2 luma transform. For our experiments, we train our models on the 100,000 images in STL-10 referred to as the "unlabeled" set, and of the remaining data, we formed a hold-out set of 10,400 images. More details of the dataset and architecture are provided in the supplementary materials[1].

After training, we collected judgments of perceptual quality on Amazon Mechanical Turk to assess whether human observers prefer reconstructions produced by perceptually-optimized networks or by the pixelwise-loss optimized networks. Images were chosen randomly from the STL-10 hold-out set. Participants were presented with a sequence of screens showing the original (reference) image on the left and a set of of three reconstructions on the right. Participants were instructed to drag and drop the images vertically into the correct order, so that the best reconstruction is on top and the worst on the bottom. The initial vertical ordering of reconstructions was randomized. We asked 20 participants to each rank 50 images, for a total of 1000 rankings. Figure 1 shows the distribution over rankings for each of the three training objectives. If participants chose randomly, one would expect to see the same number of high rankings for each model. However, MS-SSIM is ranked highest for a majority of images (709 out of 1000).

[1]Supplementary materials are available at `http://www.cs.toronto.edu/~jsnell/perceptual_supplementary.pdf`.

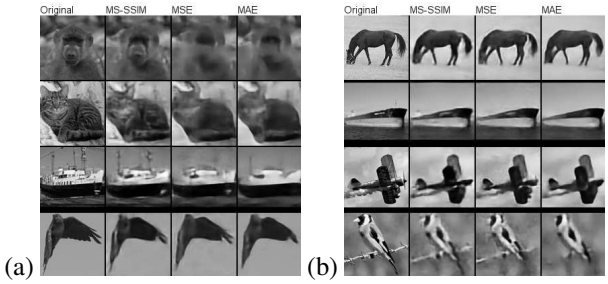

Fig. 2. (a) Four randomly selected, held-out STL-10 images and their reconstructions for the 128-hidden-unit networks. For these images, the MS-SSIM reconstruction was ranked as best by humans. (b) Four randomly selected test images where the MS-SSIM reconstruction was ranked second or third.

Figure 2a shows examples of images whose MS-SSIM reconstruction was ranked as best by human judges. Figure 2b shows examples of images whose MSE or MAE reconstruction was ranked as the best. The strong preference for MS-SSIM appears to be due to its superiority in capturing fine detail such as the monkey and cat faces and background detail such as the construction cranes. MS-SSIM seems to have less of an advantage on simpler, more homogeneous, less textured images. Note that even when MSE or MAE beats MS-SSIM, the MS-SSIM reconstructions have no obvious defects relative to the other reconstructions.

4. IMAGE CLASSIFICATION

In the previous section, we showed that using a perceptually-aligned training objective improves the quality of image synthesis, as judged by human observers. In this section, we investigate whether the MS-SSIM objective leads to the discovery of internal representations that are more closely tied to the factors of variation in images. For these experiments we use the Extended Yale B Faces dataset [23]. This dataset contains 2,414 grayscale images of 38 individuals and is labeled with the azimuth ($-130°$ to $+130°$) and elevation ($-40°$ to $+90°$) of the light source in relation to the face. We resized the images to 48×48 and learned convolutional autoencoders using MSE, MAE, and MS-SSIM as loss functions. We then used the bottleneck representations as features for SVMs trained to predicted identity, azimuth, and elevation. Additional details are provided in the supplementary materials.

Loss	Identity	Azimuth	Elevation
MSE	5.60%	277.46	51.46
MAE	5.60%	325.19	50.23
MS-SSIM	**3.53%**	**234.32**	**35.60**

Table 1. Test error for SVMs trained on bottleneck representations of convolutional autoencoders for Yale B. Classification error is the evaluation metric for identity prediction; MSE is the evaluation metric for azimuth and elevation prediction.

We opted to investigate this prediction task as opposed to a more straightforward task (such as STL-10 classification accuracy) because we expect MS-SSIM to obtain superior encodings of low- and mid-level visual features such as edges and contours. Indeed, initial studies showed only modest benefits of MS-SSIM for STL-10 classification accuracy, where coarse classification (e.g., plane versus ship) does not require fine image detail. The resulting test performance (Table 1) demonstrate that MS-SSIM yields more robust representations of relevant image factors than MSE and MAE.

5. IMAGE SUPER-RESOLUTION

We also apply our perceptual loss to the task of super-resolution (SR) imaging. As a baseline model, we use a state-of-the-art SR method, the SRCNN [24]. We used the SRCNN architecture determined to perform best in [24]. It consists of 3 convolutional layers and 2 fully connected layers of ReLUs, with 64, 32, and 1 filters in the convolutional layers, from bottom to top, and filter sizes 9, 5, and 5. All the filters coefficients are initialized with draws from a zero-mean Gaussian with standard deviation 0.001.

We construct a training set in a similar manner as [24] by randomly cropping 5 million patches (size 33×33) from a subset of the ImageNet dataset of [25]. We compare three different loss functions for the SRCNN: MSE, MAE and MS-SSIM. Following [24], we evaluate the alternatives utilizing the standard metrics PSNR and SSIM. We tested $4\times$ SR with three standard test datasets—Set5 [26], Set14 [27] and BSD200 [28]. All measures are computed on the Y channel of YCbCr color space, averaged over the test set. As expected (Table 2), MSE performs best on PSNR because they are equivalent. However, MS-SSIM achieves a PSNR comparable to that of MSE, and outperforms other loss functions significantly in the SSIM measure. Close-up visual illustrations are provided in the supplementary materials.

6. DISCUSSION AND FUTURE WORK

We have investigated the consequences of replacing pixel-wise loss functions, MSE and MAE, with perceptually-grounded loss functions, SSIM and MS-SSIM, in neural networks that synthesize and transform images. Human observers consistently judge SSIM-optimized images to be of higher quality than PL-optimized images. We also found that perceptually-optimized representations are better suited for predicting content-related image attributes. Finally, our promising results on single-image super-resolution highlight one of the key strengths of perceptual losses: they can easily be applied to current state-of-the-art architectures by simply substituting in for a pixel loss such as MSE. Taken together, our results support the hypothesis that the MS-SSIM loss encourages networks to encode relevant low- and mid-level structure in images. We conjecture that the MS-SSIM trained representations may even be useful for fine-grained classification tasks, in which small details are important.

	Bicubic	MSE	MAE	MS-SSIM
SET5 PSNR	28.44	**30.52**	29.57	30.35
SSIM	0.8097	0.8621	0.8350	**0.8681**
SET14 PSNR	26.01	**27.53**	26.82	27.47
SSIM	0.7018	0.7512	0.7310	**0.7610**
BSD200 PSNR	25.92	**26.87**	26.47	26.84
SSIM	0.6952	0.7378	0.7220	**0.7484**

Table 2. Super-resolution imaging results.

A recent manuscript [29] also proposed using SSIM and MS-SSIM as a training objective for image processing neural networks. In this manuscript, the authors evaluate alternative training objectives based not on human judgments, but on a range of image quality metrics. They find that MAE outperforms MSE, SSIM, and MS-SSIM on their collection of metrics, and not surprisingly, that a loss which combines both PL and SSIM measures does best—on the collection of metrics which include PL and SSIM measures. Our work goes further in demonstrating that perceptually-grounded losses attain better scores on the definitive assessment of image quality: that registered by the human visual cortex.

Given our encouraging results, it seems appropriate to investigate other perceptually-grounded loss functions. SSIM is the low-hanging fruit because it is differentiable. Nonetheless, even black-box loss functions can be cached into a *forward model* neural net [30] that maps image pairs into a quality measure. We can then back propagate through the forward model to transform a loss derivative expressed in perceptual quality into a loss derivative expressed in terms of individual output unit activities. This flexible framework will allow us to combine multiple perceptually-grounded loss functions and additionally refine any perceptually-grounded loss functions with data obtained from human preference judgments, such as those we collected in the present set of experiments.

Acknowledgements

This research was supported by NSF grants SES-1461535, SBE-0542013, and SMA-1041755; the Samsung GRP project; and the Canadian Institute for Advanced Research.

7. REFERENCES

[1] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 9–13, 2003.

[2] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1*, D. E. Rumelhart and J. L. McClelland, Eds., pp. 194–281. MIT Press, Cambridge, MA, 1986.

[3] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in boltzmann machines," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, D. E. Rumelhart and J. L. McClelland, Eds., vol. 1, pp. 283–317. MIT Press, Cambridge, MA, 1986.

[4] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–54, 2006.

[5] I. J. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial networks," *arXiv 1406.266v1 [stat.ML]*, pp. 1–9, 2014.

[6] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks," *arXiv 1506.05751 [stat.ML]*, pp. 1–10, 2015.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[8] Y. Li, K. Swersky, and R. S. Zemel, "Generative Moment Matching Networks," in *Proc. of The 32nd International Conf. on Machine Learning*, 2015, pp. 1718–1727.

[9] J. Lubin, "A human vision system model for objective image fidelity and target detectability measurements," in *Proc. EUSIPCO*, 1998, vol. 98, pp. 1069–1072.

[10] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1992, pp. 2–15.

[11] C. J. Van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1996, pp. 450–461.

[12] S. Winkler, "A perceptual distortion metric for digital color images.," in *ICIP (3)*, 1998, pp. 399–403.

[13] T. Frese, C. A. Bouman, and J. P. Allebach, "Methodology for designing image similarity metrics based on human visual system models," in *Electronic Imaging '97*. International Society for Optics and Photonics, 1997, pp. 472–483.

[14] 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.

[15] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[16] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[17] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate distortion optimization using ssim," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1170–1181, 2013.

[18] V. Jakhetiya, W. Lin, S. P. Jaiswal, S. C. Guntuku, and O. C. Au, "Maximum a posterior and perceptually motivated reconstruction algorithm: A generic framework," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 93–106, 2017.

[19] A. Kolaman and O. Yadid-Pecht, "Quaternion structural similarity: a new quality index for color images," *IEEE Trans. on Image Processing*, vol. 21, no. 4, pp. 1526–1536, 2012.

[20] M. Hassan and C. Bhagvati, "Structural Similarity Measure for Color Images," *International Journal of Computer Applications (0975 8887)*, vol. 43, no. 14, pp. 7–12, 2012.

[21] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 52–59. Springer, 2011.

[22] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[23] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE TPAMI*, vol. 38, no. 2, pp. 295–307, 2016.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[26] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *BMVC*, 2012.

[27] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[28] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.

[29] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Is l2 a good loss function for neural networks for image processing?," *arXiv preprint arXiv:1511.08861*, 2015.

[30] M. I. Jordan and D. E. Rumelhart, "Forward models: Supervised learning with a distal teacher," *Cognitive Science*, vol. 16, no. 3, pp. 307–354, 1992.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.