LEARNING TO GENERATE IMAGES WITH PERCEPTUAL SIMILARITY METRICS

University of Toronto
Vector Institute

Jake Snell[1,2], Karl Ridgeway[3], Renjie Liao[1,2], Brett D. Roads[3], Michael C. Mozer[3], Richard S. Zemel[1,2]

[1]University of Toronto, [2]Vector Institute, [3]University of Colorado, Boulder

University of Colorado Boulder

GOAL AND MOTIVATION

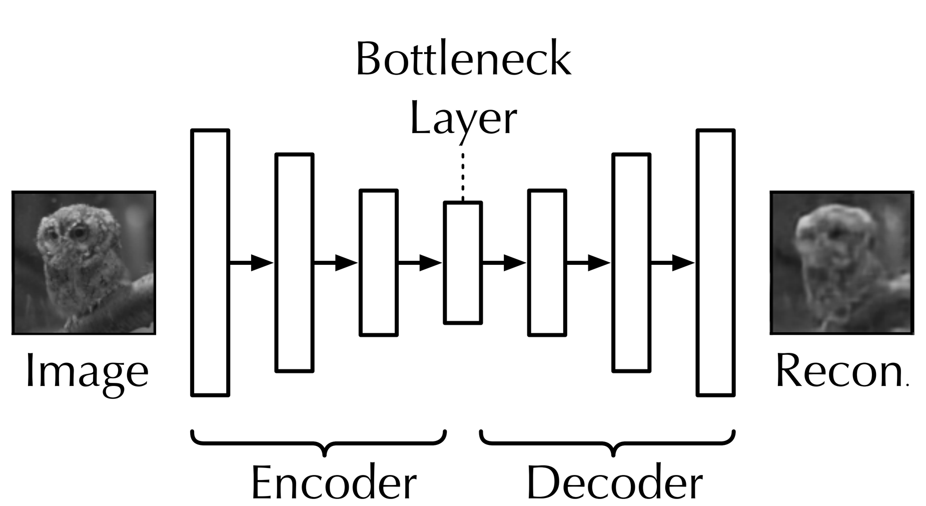
- Deep neural networks are increasingly being applied to image synthesis tasks.
- Supervised training typically uses a pixelwise-loss (PL) to indicate the mismatch between a generated image and its corresponding target.
- We propose to use a loss function better calibrated to human perceptual judgments of image quality: the multiscale structural-similarity score (MS-SSIM) [1].
  - Differentiable, compatible with SGD
- Human observers tend to prefer images synthesized by MS-SSIM-optimized models over PL-optimized models.
  - We found MS-SSIM improves image super-resolution and can also lead to better representations for image classification.
- **Takeaway**: training objectives should be aligned to characteristics of human perception.

MSE Original MS-SSIM

Image reconstructions by a standard approach (left) and ours (right). The compression factor is high to emphasize the differences.

BACKGROUND

- An autoencoder is a common image synthesis network with two components.
  - **Encoder**: compresses an image into a feature vector (typically low dimension).
  - **Decoder**: reconstructs the original image from the bottleneck representation.



- Bottleneck representation may be useful for auxiliary tasks, including classification.
- Loss function quantifies mismatch between reconstruction and target.
  - Mean-squared error: $\mathcal{L}^{MSE}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2$
  - Mean-absolute error: $\mathcal{L}^{MAE}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i|$

PERCEPTION-BASED ERROR METRICS

- We propose to use the multiscale structural-similarity score (MS-SSIM) [1] as a loss function for training image synthesis networks.
- MS-SSIM compares luminance (I), contrast (C), and structure (S) of local neighborhoods of pixels:

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

- Luminance is applied at the coarsest scale, while contrast and structure are computed at multiple scales resulting from iteratively downsampling:

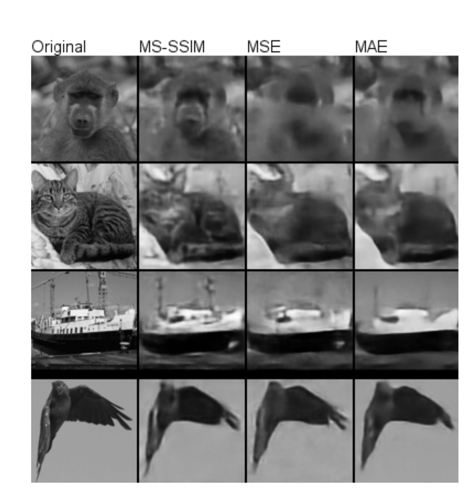
$$\text{MS-SSIM}(x,y) = I_M(x,y)^{\alpha_M}\prod_{j=1}^{M} C_j(x,y)^{\beta_j} S_j(x,y)^{\gamma_j}$$

- Image synthesis networks are trained to minimize negative MS-SSIM over all image pixels:

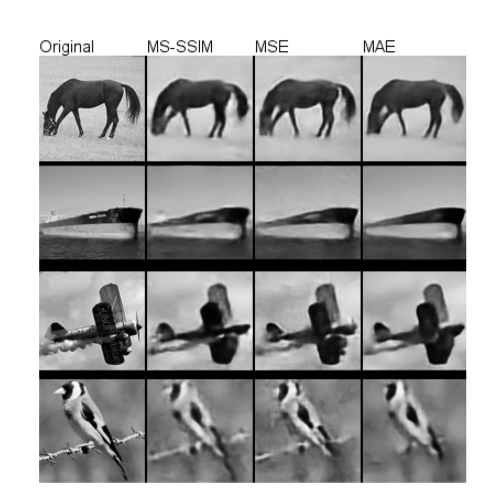
$$\mathcal{L}^{MS-SSIM}(X,Y) = -\sum_i \text{MS-SSIM}(X_i, Y_i)$$

AUTOENCODER RECONSTRUCTIONS

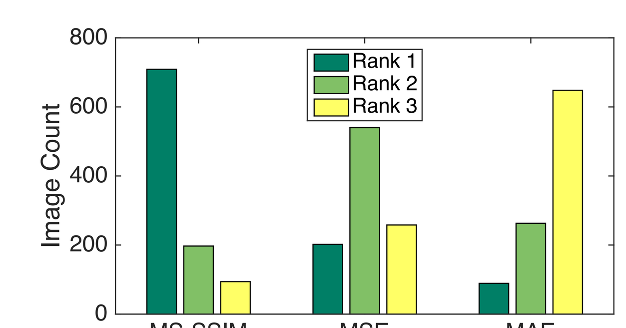
- We trained convolutional autoencoders on grayscale images from the STL-10 dataset (96 x 96 pixels).
- After training, we collected judgments of perceptual quality on Amazon Mechanical Turk to assess whether human observers prefer reconstructions from pixelwise-loss or perceptually-optimized networks.
- We collected 1,000 rankings (20 participants each ranked 50 images).
- MS-SSIM appears to better capture fine details than MSE or MAE.



Images where MS-SSIM reconstruction ranked first.



Images where MS-SSIM reconstruction ranked second or third.



Distribution of image quality rankings on 1,000 held-out STL-10 images.

LEARNED REPRESENTATIONS

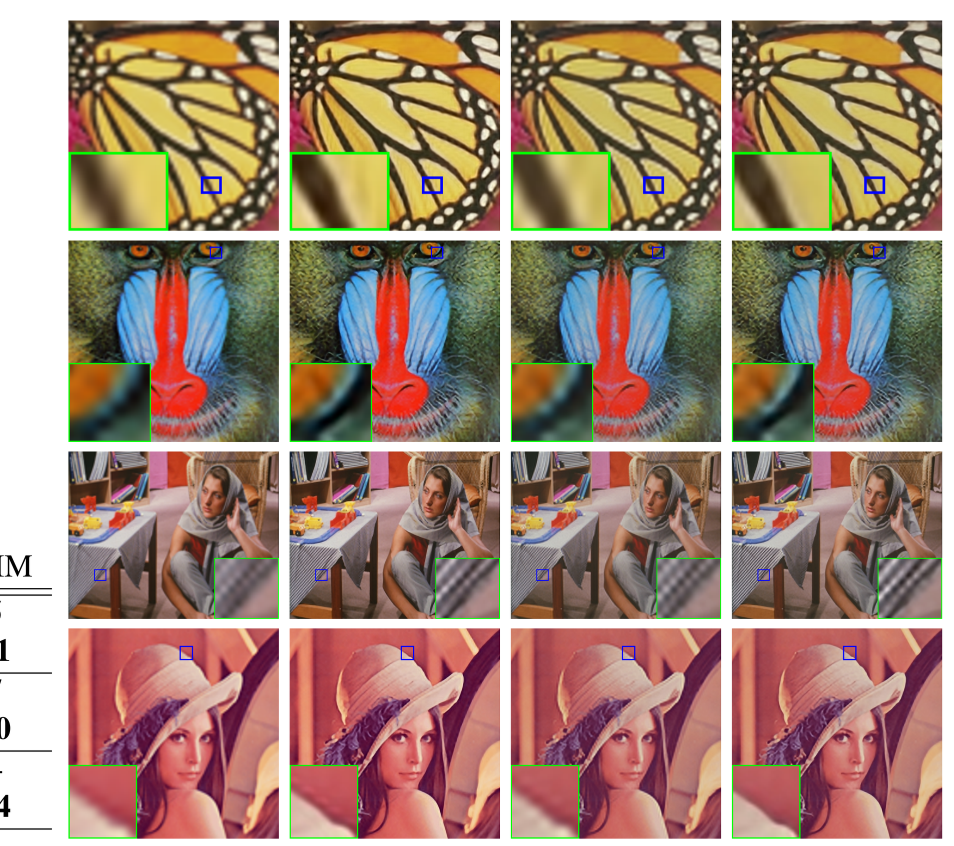
- We compared the learned representations by training conv. autoencoders on grayscale images from the Yale B face dataset (48 x 48 pixels).
- SVMs were trained on top of bottleneck representations to predict identity, azimuth, and elevation.
- Results suggest that MS-SSIM yields better encodings of low- and mid-level visual features such as edges and contours.

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**

IMAGE SUPER-RESOLUTION

- We used our perceptual loss to perform image super-resolution using the architecture of the SRCNN [2], a state-of-the-art SR method.
- Architecture consists of 3 conv. layers and 2 fully-connected layers of ReLUs with 64, 32, and 1 filters in conv. layers and filter sizes of 9, 5 and 5.
- Trained on 5 million patches randomly cropped from a subset of the ImageNet dataset.
- Performed 4x SR with all measures are computed on the Y channel of YCbCr color space.
- MS-SSIM achieves comparable PSNR to MSE and outperforms other losses significantly in the SSIM measure.

	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**



Bicubic MSE MAE MS-SSIM

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] 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.