Better Reconstruction Loss for VQ-VAE

TTIC 31230 - Fundamentals of Deep Learning - Final Project

Divij Sinha

2023-12-08

1 The Paper and the Idea

The paper I have chosen to extend is the VQ-VAE: Neural Discrete Representation Learning.

Compared to VAEs, VQ-VAEs trade off a more thorough exploration of the latent space for better images, however, we see that image quality especially with small datasets like CIFAR-10 remains an issue. A prevalent problem that occurs with VQ-VAE image outputs is image blurriness. The goal of the extension is to address this blurriness issue. The original approach utilizes Mean Squared Error (MSE) loss to quantify the difference between the original and reconstructed images.

Here in this extension, we aim to explore and evaluate various loss functions to check for potential improvements that might exist beyond the traditional MSE loss.

The focus is on utilizing image reconstruction metrics that focus on more perceptual losses.

There is also the idea of using a VQGAN to better discriminate the image loss. However, our approach is centered on computationally efficient methods. We therefore only use methods that do not require extra training.

We use the CIFAR-10 dataset as directed.

2 The Loss Functions

We look at the following loss functions -

Loss Function	Implementation Without MSE	Weighted Combinations with MSE
MSE loss	original baseline	-
PSNR loss	No MSE	1; 0.5; 0.25
SSIM loss	No MSE	1; 0.5; 0.25
MS-SSIM loss	No MSE	1; 0.5; 0.25
HPSI loss	No MSE	1; 0.5; 0.25
GMSD loss	No MSE	1; 0.5; 0.25
MS-GMSD loss	No MSE	1; 0.5; 0.25

That is, each loss functions subjected to 4 experimental runs - 1 without MSE; 3 with MSE with the final reconstruction loss equal to `MSE_loss` + (`w*loss_func`), with `w` being one of the specified above.

2.1 Descriptions of Loss Functions

1. Peak Signal-to-Noise Ratio (PSNR) - PSNR expands on MSE, by taking into account brightness of the image. Since the problem at hand is blurriness, we do not expect it to perform the best, but as a standard metric, it seems useful to add.
2. Structural Similarity Index Measure (SSIM) - SSIM is one of the standards of image comparisons, trying to approximate human perception focussing.
3. Multi-Scale Structural Similarity (MS-SSIM) - MS-SSIM is an expansion of SSIM, utilising the same principles at Multiple Scaling levels.[1]
4. Haar Perceptual Similarity Index (HPSI) - HPSI employs the Haar wavelet transform, known for its sensitivity to subtle changes in texture and structure. [2]
5. Gradient Magnitude Similarity Deviation (GMSD) - GMSD is a measure that focuses on the sharpness and clarity of images. It seems like it would be particularly useful in helping the blurriness issue we face here.
6. Multi-Scale Gradient Magnitude Similarity Deviation (MS-GMSD) - MS-GMSD is an expansion of GMSD, utilising the same principles at Multiple Scaling levels.

3 The Results

Empirically, we take a look at two metrics.

1. Image reconstructions (human perceptions)
2. Fréchet inception distances

We calculate FIDs on a sample of 2,000 images and their reconstructions[3]. Given the small sample we use, the raw FID scores are quite high. To facilitate comparisons, we scale our FID scores to a multiple of the lowest one.

3.1 Image reconstructions (human perceptions)

Following are 10 image reconstruction examples. Each block refers to one of the 4 experiments discussed above (no MSE + weighted with MSE). Each column refers to a different Loss Function.
The first two columns are always the original image and the MSE loss function (baseline) image.

Post that, the order in every block is as follows - PSNR, SSIM, MS-SSIM, HPSI, GMSD, MS-GMSD

We can observe that while no one method is a standout, most of them do perform well. Also, it is evident that different methods prioritise different things. While it is clear that HPSI, GMSD, MS-GMSD without using HPSI do not do a great job of reconstructing the image (and their kernel is clearly visible), they do find different areas of the image to to "focus" on, especially GMSD, which really outlines the subject well.

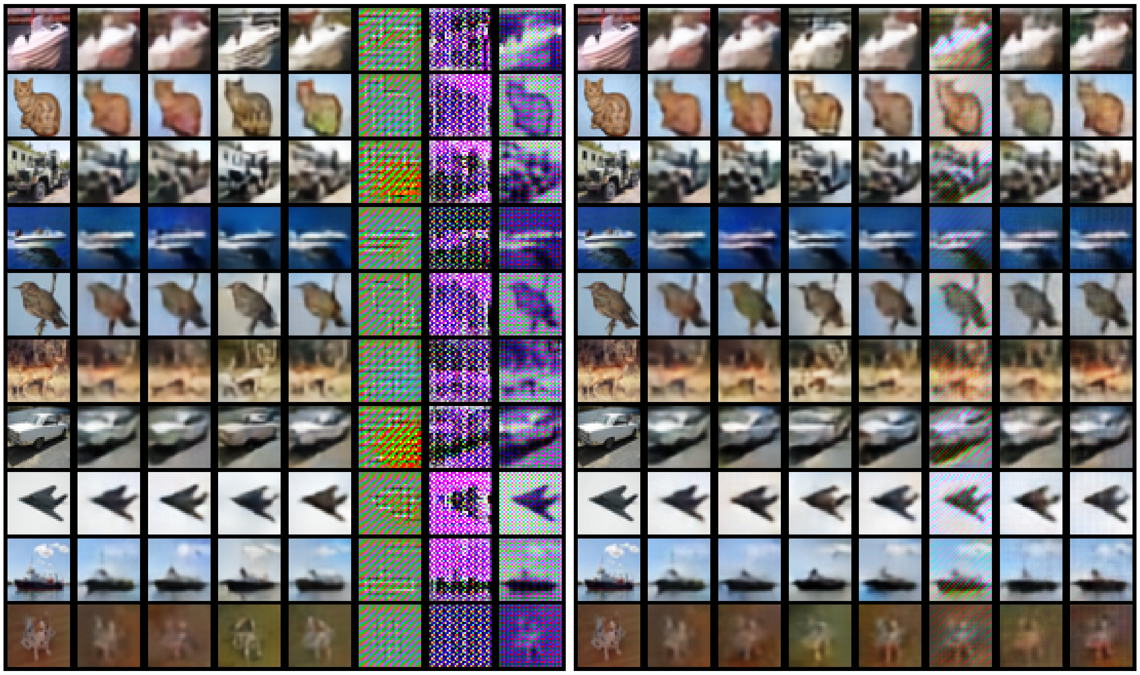
SSIM is an interesting metric as it sacrifices color accuracy, but is reasonably clear.

From this, we can prioritise methods depending on what is important in the outcome image.

[1] Given the 32x32 size of CIFAR-10 images, we don't expect huge differences between SSIM and MS-SSIM, however,fir larger, sharper, images, we expect MS-SSIM to perform better than SSIM

[2] Again, given the small size of CIFAR-10 images, this might not perform as well as it would on larger datasets.

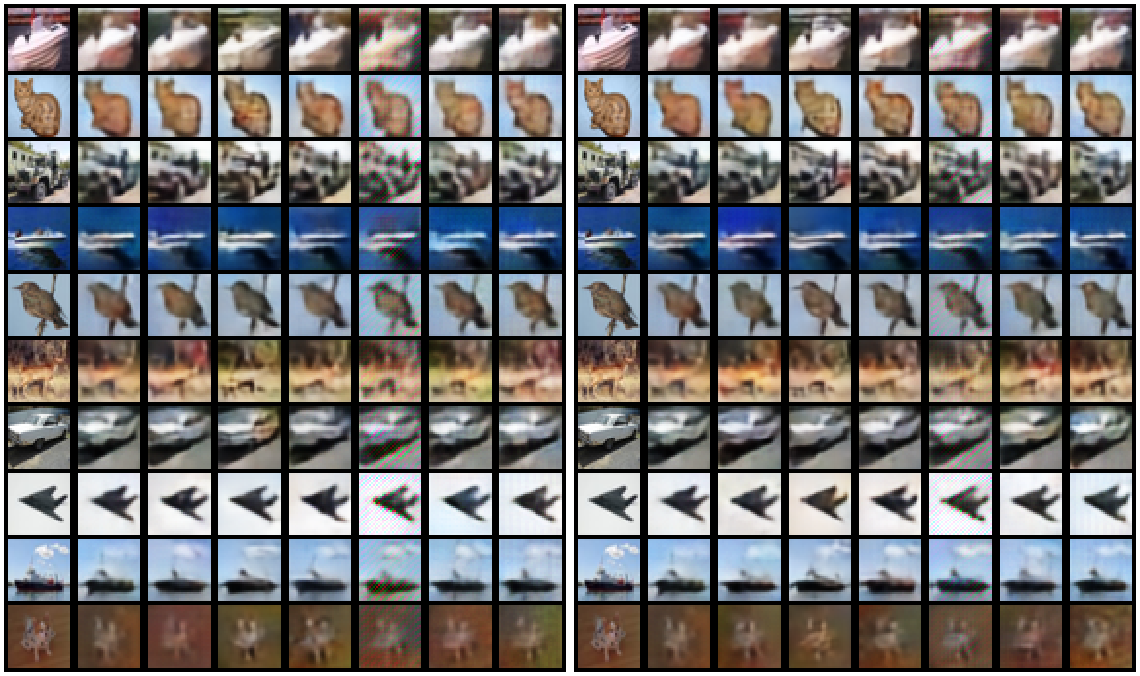
[3] This is an extremely memory intensive process, and due to lack of memory, we only run 2,000 samples.



(a) Implementation without MSE

(b) Implementation with weight=1

Figure 1: Image reconstructions



(a) Implementation with weight=0.5

(b) Implementation with weight=0.25

Figure 2: Image reconstructions

3.2 Fréchet inception distances

Since the original paper and implementation used different samples for FID, we recalculate the FID for the original MSE loss as well.

Loss Function	Implementation Without MSE	Weight=1	Weight=0.5	Weight=0.25
MSE loss	86.3	-	-	-
PSNR loss	**81.9**	**84.3**	**81.6**	**83.7**
SSIM loss	**69.0**	**84.2**	**74.5**	**72.3**
MS-SSIM loss	**78.4**	90.0	87.8	86.3
HPSI loss	384.	251.	205.	161.
GMSD loss	361.	109.	87.9	88.1
MS-GMSD loss	278.	97.1	90.4	**83.8**

While the numbers are all quite high for FID distances, we put that down to small sample size. These were stable over multiple runs. In bold, we have highlighted all the ones that are better than the baseline (the first row)

3.3 Discussion of results

While it's evident that no single method significantly surpasses Mean Squared Error (MSE) in performance, several methods do offer good improvements. Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are more effective than MSE on a consistent basis. Additionally, it's noteworthy that certain methods yield better results when used in combination with MSE, Multi-Scale Gradient Magnitude Similarity Deviation (MS-GMSD) is a good example of this: while its standalone performance is quite bad, integrating it with MSE using appropriate weightings leads to an improvement.

4 Conclusion & Future Steps

Given that these are low-to-no-cost improvements over MSE computationally, with the added computational time being consistently <10% over a few dozen runs, we can absolutely hope to improve our VQ-VAE models using these. We only tested them on CIFAR-10, and used pre-defined levels of hyperparameters (the weights = 1, 0.5, 0.25). Some hyperparameter tuning, and we can expect to see reasonable improvements at no cost. We also expect these methods to perform better when paired with larger images, as perceptual losses can be better captured in higher fidelity images.

Here are all the loss functions that performed better than the baseline, along with another set of reconstructions. The first row is the original image. The second row onwards are the loss functions sorted by FID from the table above (SSIM, SSIM w=0.25, SSIM w=0.5, MS-SSIM, . . . , MS-GMSD w=0.25, SSIM w=0.5, PSNR w=05). The last row is the baseline MSE loss function.

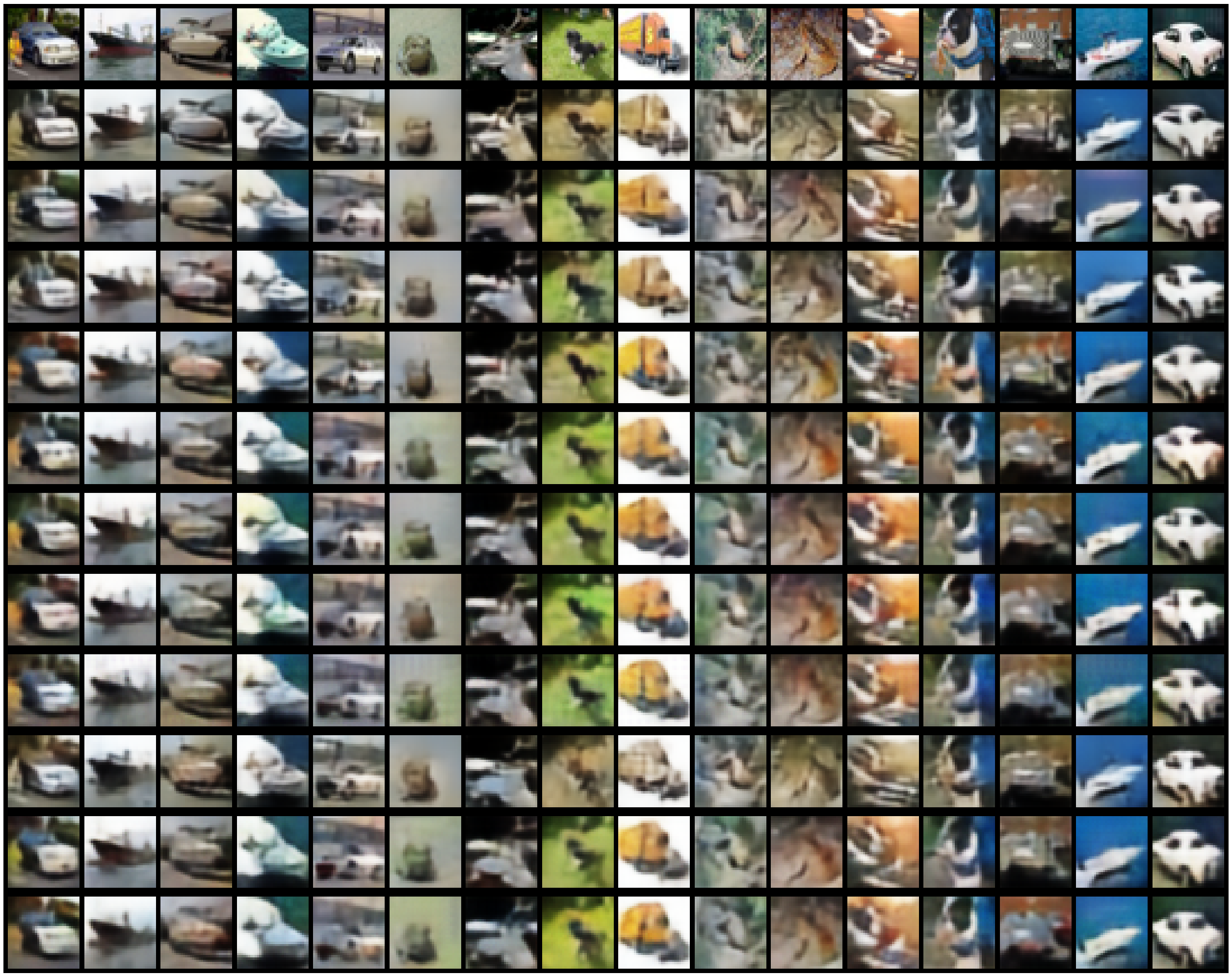


Figure 3: Better than baseline (by FID) image reconstructions

5 Colab notebook

The colab notebook can be found here - https://drive.google.com/file/d/1ruL20tTSLstnE7MWROAzFsf2S7uhFHbh/view?usp=sharing

The google drive folder being mounted in the first step of the notebook is here https://drive.google.com/drive/folders/1W6LW7lZsHF8BogACkjLmm_h_K7UidKV0?usp=sharing

References

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. “GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium.” *CoRR* abs/1706.08500. http://arxiv.org/abs/1706.08500.

Kumar, Rithesh, Tristan Deleu, and Evan Racah. 2018. “Reproducing Neural Discrete Representation Learning.” https://github.com/ritheshkumar95/pytorch-vqvae/.

Oord, Aäron van den, Oriol Vinyals, and Koray Kavukcuoglu. 2017. “Neural Discrete Representation Learning.” *CoRR* abs/1711.00937. http://arxiv.org/abs/1711.00937.

Reisenhofer, Rafael, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. 2016. “A Haar Wavelet-Based Perceptual Similarity Index for Image Quality Assessment.” *CoRR* abs/1607.06140. http://arxiv.org/abs/1607.06140.

Rozet, François. 2020. *PIQA: PyTorch Image Quality Assessement*. https://doi.org/10.5281/zenodo.7821598.

Seitzer, Maximilian. 2020. *pytorch-fid: FID Score for PyTorch*. https://github.com/mseitzer/pytorch-fid.

Wang, Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing* 13 (4): 600–612. https://doi.org/10.1109/TIP.2003.819861.

Wang, Z., E. P. Simoncelli, and A. C. Bovik. 2003. "Multiscale Structural Similarity for Image Quality Assessment." In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2:1398–1402 Vol.2. https://doi.org/10.1109/ACSSC.2003.1292216.

Xue, Wufeng, Lei Zhang, Xuanqin Mou, and A. C. Bovik. 2013. "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index." *CoRR* abs/1308.3052. http://arxiv.org/abs/1308.3052.

Xue, Wufeng, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. 2014. "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index." *IEEE Transactions on Image Processing* 23 (2): 684–95. https://doi.org/10.1109/TIP.2013.2293423.

Zhang, Bo, Pedro V. Sander, and Amine Bermak. 2017. "Gradient Magnitude Similarity Deviation on Multiple Scales for Color Image Quality Assessment." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1253–57. https://doi.org/10.1109/ICASSP.2017.7952357.