Noise Experiments on a Denoising Variational Autoencoder

Sahil Jayaram - 6 December 2018
Math 151B - Professor Shay Deutsch - UCLA
Original paper: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14213
Source code available at: https://github.com/SahilJ97/DVAE

Abstract - **Denoising Variational Autoencoders (DVAEs) are a class of autoencoder neural networks well-suited for the tasks of de-noising and dimensionality reduction. While Denoising Autoencoders (DAEs) are trained by injecting noise into only the input layer, DVAEs are also subjected to stochastic noise in their latent space, drawing from the design of Variational Autoencoders (VAEs). In this paper, I investigate the effects of input noise levels on DVAE convergence under Mini-Batch Gradient Descent (MBGD), using both Fashion-MNIST and a toy dataset. The results of my experiment shed light on the relationship between input noise and DVAE performance, and suggest the existence of a practical noise ceiling—that is, a level of input perturbation above which MBGD fails to converge to any reasonable local minimum.**

Keywords - DVAE, Noise, Stability, Mini-Batch Gradient Descent, VAE

Introduction

In a basic autoencoder architecture, m-dimensional inputs are mapped to a lower-dimensional space via an "encoder network," whereupon the resulting "latent representation" is used to approximately reconstruct the original input through a "decoder network." The VAE, first proposed by Kingma and Welling, enhances this model with a family of statistical techniques known as Variational Bayesian methods. On a high level, VAEs differ from standard autoencoders in that they enforce strong continuity on the learned manifold. This is achieved by means of stochastic noise in the latent space, as well as a regularization term penalizing divergence between the learned distribution of latent variables and a "true posterior" distribution of choice (often the standard multivariate normal distribution). Given an input, the encoder network produces a distribution of latent representations rather than a latent vector, and the output is constructed using a vector randomly drawn from this distribution (Kingma, Welling).

DVAEs can be described as VAEs with noise injected at the input level during training, inducing noise-robustness. In the original paper on DVAEs, Im et al. demonstrate that their proposed model outperforms both VAEs and Importance Weighted Autoencoders (IWAEs) on the Frey Face and MNIST datasets. Yet, the question remains as to how to determine an appropriate level of training noise for learning DVAEs (Im et al., 2017). Although time constraints prevented me from providing a rigorous or general answer to this question, the empirical results mentioned in this paper offer some potentially useful insight into the problem.

Background on Mini-Batch Gradient Descent

MBGD is an iterative optimization method based on Gradient Descent. At each iteration, the model parameter θ is updated as follows:

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\alpha}{p}\sum_{i=1}^{p} J_{\theta^{(k)}}(x_i) \qquad (1)$$

where α is the learning rate, J is the Jacobian of the cost function with respect to θ, and $x_1, x_2, \ldots, x_p$ constitute a "batch" selected from the training data set.

The *stability* of an iterative optimization algorithm measures the algorithm's resilience to generalization. Informally, an algorithm that achieves similar solutions with different data is stable, while an algorithm whose solution is highly sensitive to changes in the data is unstable. If the optimization problem is well-posed—i.e. "its unique solution attracts all approximate solutions corresponding to small perturbations of the given problem"—stability can be reformulated as the algorithm's ability to converge to the optimal solution given different data (Zolezzi).

One measure of algorithm stability is ε-uniform stability. Largely quoting from Hardt et al., we say that an algorithm A is ε-uniformly stable if for all data sets S, S' such that S and S' differ in at most one sample, we have

$$\sup_z \mathbb{E}_A[f(A(S)) - f(A(S'))] \leq \varepsilon. \qquad (2)$$

Due to time constraints, I was unable to find or derive any stability conditions for MBGD. However, for Stochastic Gradient Descent (SGD)—equivalent to MBGD with a batch size of 1—we have the following approximate inequality:

$$\varepsilon_{stab} \lessapprox \frac{T^{1-1/(\beta c+1)}}{n} \qquad (3)$$

where the loss function f has an image contained in the interval [0, 1], β is such that f is β-smooth ($\bigtriangledown f$ is Lipschitz-continuous with Lipschitz constant β), $c \geq \alpha_t t$ (where α_t is the step size at step t), and T steps are run in total (Hardt et al.). Because MBGD is similar in nature to SGD with a more stable loss gradient, I apply (3) to my model design later in this paper as part of an informal theoretical analysis.

METHODS

I implemented a DVAE as described by Im et al., using one hidden layer in the encoder network and a standard multivariate normal distribution as the true posterior. My implementation is different from theirs in one important way: in order to avoid "mode collapse," whereby a model learns to map all inputs to roughly the same output, I had to decrease the amount of stochastic noise applied at the latent level, a common trick for VAEs exhibiting underfitting. More specifically, my model draws each latent variable from a distribution with a standard error one-tenth of what it "should be" according to the encoder output.

To test my algorithm on a dataset actually generated from the true posterior distribution, I constructed a toy dataset. In accordance with Fashion-MNIST, each datapoint is a 28x28 image. To determine the pixel values for each image, my script generated a 50-dimensional tensor from a standard multivariate normal distribution and mapped it to a 784-dimensional tensor via an affine transformation.

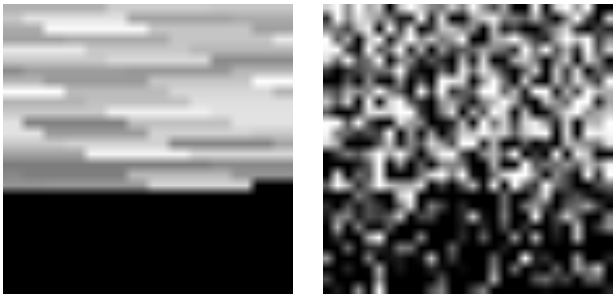

Figure 1
Left: A sample from the toy dataset; Right: The same sample, corrupted with the *maximum* level of noise used in this experiment (.50)

Like that of Im et al., my algorithm applies Gaussian noise at the input level. In my implementation, "noise level" refers to the standard deviation of the zero-centered Gaussian distribution from which the perturbation applied to each pixel is generated.

EXPERIMENT

For both datasets, I trained the DVAE using noise levels ranging from 0 to .50 in increments of .02. The training algorithm was run three times for each input noise level, and the best performance data was kept. 25 evaluations were made per model: one for each level of test set input noise. Figures 2 and 3 illustrate the performance data I collected. Each curve corresponds to a group of models trained with the specified level of input noise, and each datapoint along the curve represents the minimum reconstruction error achieved by a model from that group with the specified level of test set noise.

INFORMAL ANALYSIS

I loosely define the term *reasonable* within the context of my empirical framework as follows: a learned model is *reasonable* if its test performance is best when input noise is low and worst when input noise is high. In particular, models whose performance was constantly poor across test noise levels—e.g. models trained on Fashion-MNIST at noise level .14 (Figure 2 in grey)—were *unreasonable*, as their large reconstruction errors and absolute robustness to noise suggest that their learned manifolds do not accurately reflect the sample space. Upon closer inspection, it becomes clear that all of these models suffered from mode collapse (Figure 4).

With respect to the DVAE performance results for Fashion-MNIST (Figure 2), I present the following observation:

Using a training noise level greater than .12 always resulted in an unreasonable model

In order to determine whether this trend is due to the nature of the data itself, consider the performance results for the toy dataset (Figure 3). Here, the observation remains true, suggesting that the observed ceiling results from properties of the optimization problem that are mainly independent of the dataset used.

It is important to note that this learning problem is not a well-posed optimization problem; if any given solution is a global minimum, the symmetry of my true posterior distribution (the standard 50-variable normal distribution) implies the existence of (50!-1) other global minima, thus eliminating the possibility of a unique global minimum. Because the problem is not well-posed, we may not treat 'stability' as a measure of the attracting power of some optimal solution. My analysis instead takes the term 'stability' to refer to the attracting power of reasonable solutions. Thus, assuming the existence of at least one reasonable solution, I say that an MBGD optimization problem is highly unstable if all models converged to unreasonable solutions.

Because the toy dataset was actually generated from the enforced posterior distribution, we may assume that for any of the toy data optimization problems, there exists a local minimum corresponding to a reasonable solution, regardless of input noise during training. I therefore conjecture that for this DVAE implementation in general, using a training noise greater than or equal to .14 results in conditions under which MBGD is highly unstable, i.e. under which reasonable solutions have little attracting power.

For a very rough theoretical assessment of the stability of these optimization problems, I apply the inequality (3) to my

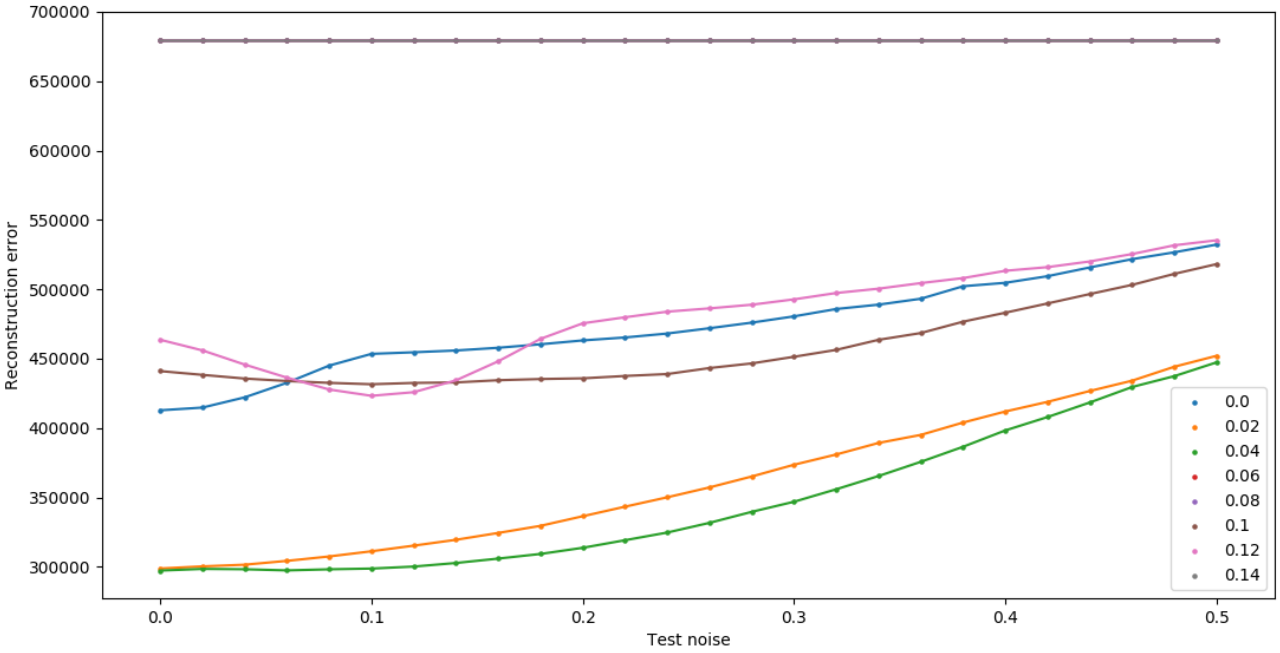

Figure 2

DVAE performance on FASHION-MNIST. The group of models trained at noise level .04 outperformed all other groups. For noise levels .06, .08, .14, …, .5, no reasonable model was learned (all of those models achieved a relatively constant reconstruction error of about 690,000).

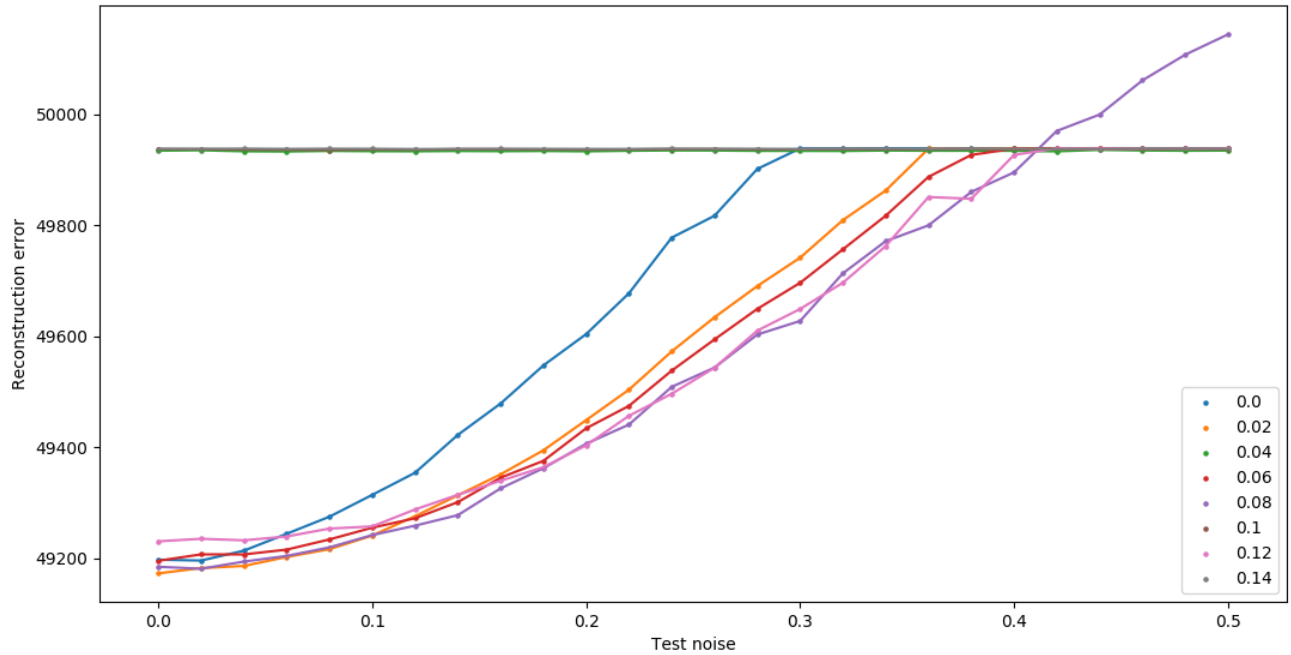

Figure 3

DVAE performance on the toy dataset. For noise levels .04, .10, .14, .16, …, .5, no reasonable model was learned.

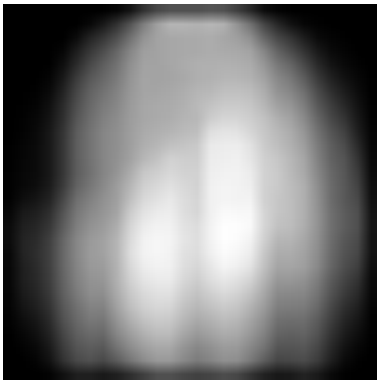

Figure 4
"Mode collapse" for Fashion-MNIST. The learned manifold consists entirely of images resembling this one, which approximates the dataset mean ("mean collapse" would be a more suitable term).

model. To find a suitable β, I analyzed the Hessian matrix of my training loss function for one thousand training steps. The largest encountered second-order partial derivative was 731.22, so I assume that my cost function is 800-smooth. Disregarding the KL-Divergence portion of the training loss function (which always converged to a number << 1 within one epoch), and scaling the function such that its values fall in [0, 1] rather than [0, 7840], I thus selected $\beta = 0.1$. Using T = 60,000 and a step size of 0.0001 (true to my model design), I found $c = 6$ to be an appropriate choice. Using these values and n = 60,000 (the size of the Fashion-MNIST training set), we get 0.001 as an approximate least upper bound for ε_{stab}. Because (3) was only derived for well-posed optimization using SGD, and not for ill-posed problems using MBGD, this epsilon bears little relevance to my experiment. If, however, this least upper bound turned out to be very large (>> 1), it would be sensible to conjecture that MBGD is inherently very unstable with respect to this DVAE.

Conclusions

Much of the *Informal Analysis* section is too speculative to be of any practical use. Had I managed to derive a stability condition more specific to the optimization problem addressed by Im et al. and in this paper, my theoretical analysis would have been rigorous. Nonetheless, my empirical results do seem to provide some interesting insight into DVAE learning.

Particularly, this specific DVAE exhibits what I will refer to as a *practical noise ceiling* around Gaussian input noise level .12: a level of input perturbation above which MBGD is highly unstable, failing to converge to any reasonable solution. Although a broader range of evidence would be needed to substantiate this observation, my results may indicate that all DVAE implementations like the one proposed by Im et al. are associated with a firm, data-independent ceiling of this nature. This pattern could be useful in developing a numerical method for estimating an optimal training noise level, among other potential applications.

References

Han Xiao, Kashif Rasul, Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. Available at: <arXiv:1708.07747>

Hardt, M.; Recht, B.; Singer, Y.. Train faster, generalize better: Stability of stochastic gradient descent. Available at <https://arxiv.org/abs/1509.01240>

Im, D.; Ahn, S.; Memisevic, R.; Bengio, Y.. Denoising Criterion for Variational Auto-Encoding Framework. AAAI Conference on Artificial Intelligence, North America, feb. 2017. Available at: <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14213>.

Kingma, D.; Welling, M.. Auto-Encoding Variational Bayes. Available at: <https://arxiv.org/abs/1312.6114>

Zolezzi, T.. Well-Posedness and Conditioning of Optimization Problems. Studia Mathematica Bulgaria, 1998 Available at: <http://www.math.bas.bg/pliska/Pliska-12/Pliska-12-1998-267-280.pdf>