Implicit Generation and Representation Learning with Energy-Based Models

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Two fundamental problems with deep learning are data efficiency and out of distribution generalization. Generative models are able to capture world knowledge and enable faster learning. At the same time, this model prevents catastrophical failure in out of distribution cases.

Generative modeling has seen a flux of interest. Many approaches reply on directly maximizing the likelihood. Modeling a correlated high dimensional data distribution is difficult. Auto-regressive models [Van Oord et al., 2016, Graves, 2013] solve this by completely factorizing the underlying distribution, but such an approach leads to compounding error and a loss of underlying structural information. Other approaches such as the variational auto-encoder [Kingma and Welling, 2014] or flow based models [Dinh et al., 2014, Kingma and Dhariwal, 2018] rely on a factorized prior distribution to simplify likelihood estimation. Flow models require invertible Jacobian transformations, which can limit model capacity and have difficulty fitting discontinuous data. Such approximations have prevented likelihood models from generating high quality images in diverse domains. In contrast, approaches based off generative adversarial networks [Goodfellow et al., 2014] put no constraints on latent space and have generated high quality images but do not cover the entire data distribution.

Energy based models(EBMs) are flexible likelihood model with no latent constraints [LeCun et al., 2006]. EBMs received attention in the past [Hinton et al., 2012, Dayan et al., 1995] but have not seen wide adoption due to expensive negative sampling phase and training instability. We propose a sampling method for training energy models called Generative Energy Optimization (GEO). We use a sampling procedure based on Stochastic Gradient Langevin Dynamics [Welling and Teh, 2011] allowing scaling to high dimensional distributions such as images. We find GEO images are competitive to GANs on CIFAR10 (inception score) in image quality. We further find that EBMs learn useful representations for supervised learning. Finally, we show that GEO trained EBMs generalize well, achieving similar likelihoods on both CIFAR10 train/test sets while showing good image generation and denoising/inpainting ability on test images, and also significantly outperforming feed-forward models in time series prediction domain. We find that EBMs also generalize out of distribution, assigning lower probability to SVHN, a curious failure in GLOW, PixelCNN, and VAE models [Anonymous, 2019a].

2 Generative Energy Optimization

Method Define the data distribution as $p_d(x)$, and model distribution as $p(x)$, for which we use Boltzmann distribution $p(x) = \exp\{-E(x,\theta)\}/Z(\theta)$ where $Z(\theta)$ is the normalization constant, $E(x,\theta) \in \mathbb{R}$ is a neural network [*]. Given training data $x_1 \sim p_d(x)$, the objective is to minimize

$$\min_{\theta} \mathbb{E}_{x\sim p_d(x)} \left[E_\theta(x) - \log \sum_{\tilde{x}\sim q(x)} \exp(-E_\theta(\tilde{x})) \right] \tag{1}$$

This loss is described in more detail in the appendix. The gradient is $\nabla_\theta E(x) - \nabla_\theta Z(\theta)$, which we approximate with respect to the partition function by sampling from a proposal distribution $q(x)$

[*]We use residual networks [He et al., 2015] with zero initialization [Anonymous, 2019b] for images. For time series prediction we use a combination of fully connection, self-attention and 1D convolutions (see appendix)

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

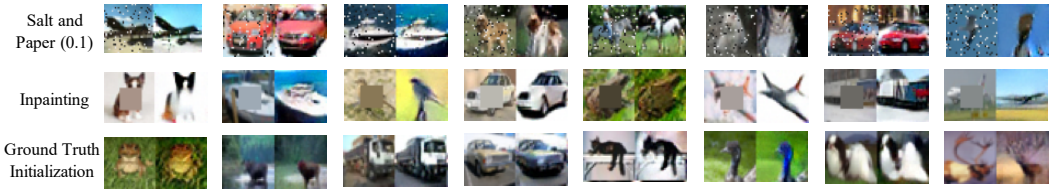


Figure 1: Conditional EBM image restoration on images in the **test** set through GEO. The right column shows failure (approx. 10% objects change with ground truth initialization and 30% of objects change in salt/pepper corruption or in-painting. Right column shows max amount of change.)

which is finite step MCMC approximation of $p(x)$ to obtain the object. We note that in the infinite limit of MCMC steps, of proposal distribution $q(x)$ is exactly $p(x)$.

Proposal Distributions To construct our proposal distribution q(x), we use stochastic gradient descent with Langevin dynamics (SGLD) [Welling and Teh, 2011] which allows for efficient sampling, particularly in high-dimensional spaces. We generate samples with

$$\tilde{x}^q = \tilde{x}^k, \quad \tilde{x}^k = \tilde{x}^{k-1} - \lambda(\nabla_{x^{k-1}} E_\theta(\tilde{x}^{k-1}) + \omega^k), \quad \omega \sim N(0, \sigma) \tag{2}$$

where we clip gradients above a certain magnitude for stability. One interpretation of generating samples with SGLD for convolutional networks is that generator and discriminator are derived from a single function. An energy function can serve as a discriminator while the generator is derived implicitly via a backwards gradient pass. The backward pass of a convolution is a transposed convolution, which mirrors the discriminator, much like GAN architectures where generator is the mirror of discriminator [Radford et al., 2016, Miyato et al., 2018]. Use of a single energy function no long requires explicitly training the generator. However, GEO is not limited to any particular MCMC proposal distribution $q(x)$, but rather any that can approximate $p(x)$. To demonstrate this we also use MPPI based MCMC sampling [Williams et al., 2017] on time series domain (Equation 6).

Efficiently Sampling Negative Samples An issue when estimating the partition function is generating negative samples X^- from probability modes. Even using gradient-based sampling methods, the energy landscape of an arbitrary network can be difficult to sample from.

One solution is to increase smoothness by constraining the Lipschitz constant of an energy model. We follow the method of [Miyato et al., 2018] and add spectral normalization to all layers of the model. Constraining the Lipschitz constant reduces capacity/expressiveness of functions.

An orthogonal solution is to add a loss to minimize $D_{KL}(q(x)||p(x))$. Specifically, we add an additional KL loss $\mathbb{E}[q(x)\log(p(x))]$ which we approximate by sampling $x_1, \ldots, x_k \sim q(x)$ and minimize $L_{kl} = \sum_i E_{stop\_gradient(\theta)}(x_i)$ where we backpropogate through the sampling procedure, changing a model's landscape to be samplable. This approach requires backpropagating through SGLD procedure, which is a more expensive procedure. We use the first solution for images and the second for time series, but both are complementary.

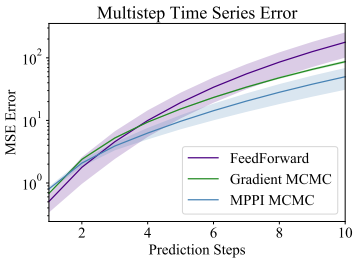
Improving Mode Exploration MCMC methods are known to explore slowly, following random walk behavior [Neal, 2011]. Tieleman [2008] increases diversity by starting chains from past samples. We further propose to use a replay buffer of past samples, where for each new batch of replay samples, we initialize a small fraction from random noise. This allows us to store diverse chains and prevent cyclical behavior. This simultaneously helps our generation procedure increase diversity and prevents our likelihood model from having spurious modes.

3 Evaluation

3.1 Images

We measure EBM's ability to model complex distribution on both class-conditional and unconditional CIFAR10 dataset. Our model is based on the ResNet architecture (using conditional gains and biases per class) with details in Section 4.5. Our models have around 7 million parameters, comparatively much smaller than hundred of millions to billions of parameters found in PixelCNN and Glow models respectively.

Image Quality We evaluate image quality of EBMs with Inception score [Salimans et al., 2016], where we obtain unconditional and conditional Inception scores of 6.43 ± 0.073 and 8.52 ± 0.13 respectively. Our unconditional scores are higher then other likelihood models such as PixelCNN (4.60) [Van Oord et al., 2016] and is similar to DCGAN (6.4) [Radford et al., 2016]. Our conditional score is comparable to 8.59 in SNGAN, [Miyato et al., 2018] and is better than ImprovedGAN (8.09), [Salimans et al., 2016]. During sampling, we found that unconditional GEO was unable to generate



(a) Multistep time series prediction MSE errors (**log** scale). EBMs show out of distribution generalization by reduced long term rollout error.

(b) Illustration of cross class mapping using GEO on a conditional EBM. The EBM is conditioned on a particular class but is initialized with an image from a separate class(left). Additional images in Figure 5

#	Baseline	FT	Baseline + DA	FT + DA
Accuracy	83.6	86.5	89.7	90.2

Table 1: Test Accuracy on CIFAR10 with or without finetuning (FT) with or without data augmentation (DA). We using horizontal flip and random crop data augmentations. Energy based fine-tuning allow better generalization.

many of the images we observed in the replay buffer without a prohibitively large number of steps. Therefore, for unconditional GEO, we also tested SGLD on the last 10 snapshots of the trained modwl, allowing better initial mode exploration from random noise. Under this scheme, we obtain 6.79 ± 0.053 for the unconditional model. We urge readers to see qualitative generations in Figure 3. We note definite objects even in unconditional models and show comparisons against GLOW models.

Quantitative Evaluation To measure overfitting, we found unconditional model had average energies of -0.00169 ± 0.0196 on the train dataset and 0.001454 ± 0.0176 on the test dataset. For conditional model, we found energies of 0.00198 ± 0.0369 on the train dataset and 0.00751 ± 0.0374 on the test dataset. The small mean difference relative to individual standard deviation indicates EBMs assigns close likelihoods on train and test sets. This is supported by results in Figure 1. [†] On different images in SVHN, unconditional GEO has energies of 0.00713 ± 0.01 and conditional 0.0453 ± 0.0459 showing that energy models do not suffer from problems in VAE, GLOW, and PixelCNN models of assigning higher probability on SVHN despite training on CIFAR10 [Anonymous, 2019a]

Image Restoration To further evaluate the likelihood model and generalization of conditional models, we evaluate EBM's ability to restore **test set** images after corruption. We show GEO decorruption results on images corrupted with 10% salt and pepper noise or with the central portion removed Figure 1 and find good restoration. To ensure results are not due to mode collapse, we also initialize and images to random real test image samples and find limited image change, indicating presence of high probability modes at test images. A unique property of GEO's optimization-based sampling is its ability to restore errors in images without explicitly specifying corrupted pixels. We also investigate ability to restore images from one class to images from another class in Figure 2b.

Representation Learning We further investigate representation learning in EBMs, which we measure by fine-tuning energy models to a supervised classification task. We remove the last linear layer of our model and replace it with a classification layer. During training, we backpropogate through all weights and get results found in Table 1. We find EBMs learns representations that allow better generalization on CIFAR10. We believe even larger gains may be achieved by large pre-training or joint training.

3.2 Time Series Prediction

To demonstrate the generality of our technique, we also explore the ability of energy functions to model future predictions on particle-based dynamics. We simulate one moving ball with wall collisions, drag and friction. We train models to predict the next ball state given the past 3 states using 4500 training trajectories with 500 time-steps and evaluate MSE of future state predictions on 500 test trajectories. We preserve the same architecture for feed-forward model but modify last layer to predict the next state. Results are shown in Figure 2a. When using model rollouts, we find that energy functions have significantly lower error for multistep prediction despite higher initial error, indicating that energy functions have stronger out of distribution generalization.

[†] We found exact log likelihood of energy-based models difficult to estimate as calculating the log partition function using AIS [Neal, 2001] with HMC transitions took too long to explore modes.

References

Anonymous. Do deep generative models know what they don't know? In *Submitted to International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=H1xwNhCcYm. under review. 1, 3

Anonymous. The unreasonable effectiveness of (zero) initialization in deep residual learning. In *Submitted to International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=H1gsz30cKX. under review. 1, 8

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural Comput.*, 7(5):889–904, 1995. 1

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 1

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

Alex Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850*, 2013. 1

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2015. 1

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012. 1

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018. 1

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

Yann LeCun, Sumit Chopra, and Raia Hadsell. A tutorial on energy-based learning. 2006. 1

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2, 8

Radford M Neal. Annealed importance sampling. *Stat. Comput.*, 11(2):125–139, 2001. 3

Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011. 2

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 2

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. ACM, 2008. 2

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 1, 2

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011. 1, 2

Grady Williams, Andrew Aldrich, and Evangelos A Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 40(2):344–357, 2017. 2, 7

4 Appendix

4.1 Qualitative Evaluation

(a) GLOW unconditional samples

(b) GEO unconditional EBM samples

(c) GEO Historical ensemble(10) EBM unconditional samples

Figure 3: Illustrations of image generation from GLOW as compared to our EBM models. Our models are able to more accurately generate objects.

(a) GEO Samples from conditional CIFAR10 EBM model

(b) GEO samples from unconditional ImageNet 32x32 EBM model

We present qualitative images of unconditional image generation using the recent GLOW model compared to using GEO in Figure 3. We find that our unconditional model is able to construct many more object like shapes. For historical ensemble generation, we use the last 10 model snapshots to sampling over. We find that although EBMs tend to put high likelihood in many image modes, gradients of an individual EBM tend to point to specific images. By combining 10 models, we get more diverse gradients and samples. We present qualitative images of images from a conditional model in Figure 4a. Our conditional image model is able to generate reasonable looking images in all of the classes in CIFAR10. We further present images from a unconditional generation on ImageNet in Figure 4b, which we generate using the last 10 model snapshots of energy models. We find the presence of objects and scenes in some of the generated image with occasional hybrids (such as a presence of a toaster cat in middle bottom row).

We provide further images of cross class conversions using a conditional EBM model in Figure 5. Our model is able to convert images from different classes into reasonable looking images of the target class while sometimes preserving attributes of the original class.

Finally, we analyze nearest neighbors of images we generate in Figure 6.

4.2 Sampling Process

We provide illustration of image generation from conditional and nonconditional EBM model starting from random noise in Figure 7 with small amounts of random noise added. Dependent on the image generated there is slight drift from some start image to a final generated image. We typically observe that as sampling continues, much of the background is lost and a single central object remains.

We find that if small amounts of random noise are added, all sampling procedures generate a large initial set of diverse, reduced sample quality images before converging into a small set of high

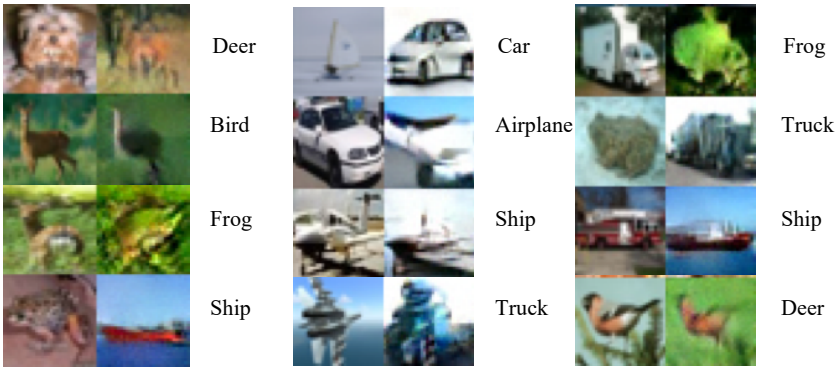


Figure 5: Illustration of more cross class conversion applying GEO on a conditional EBM. We condition on a particular class but is initialized with an image from a another class(left). We are able to preserve certain aspects of the image while altering others.

probability/quality image modes that are modes of images in CIFAR10. However, we find that if sufficient noise is added during sampling, we are able to slowly cycle between different images with larger diversity between images (indicating successful distribution sampling) but with reduced sample quality.

Due to this tradeoff, we use a replay buffer to sample images at test time, with slightly high noise then used during training time. For conditional energy models, to increase sample diversity, during initial image generation, we flip labels of images early on in sampling.

4.3 Loss Functions

When training EBMs, our overall loss function is given by

$$L_{\text{total}} = L_{\text{ML}} + L_{\text{KL}} + L_{\text{Reg}}$$

where L_{ML} is negative log likelihood loss and is given by

$$L_{\text{ML}} = \frac{1}{N}\sum_{i=0}^{N} E_\theta(x_i^+) + \log(\sum_{i=0}^{N} e^{-E_\theta(x_i^-)})$$

In practice, we found that the negative term to be numerically unstable due to an exploding term in the denominator. We therefore note that an expression with equivalent gradient to the negative term is

$$\sum_{i=0}^{N} -\text{stop\_gradient}\left(\frac{e^{-E_\theta(x_i^-)}}{\sum_i e^{-E_\theta(x_i^-)}}\right) E(x_i^-)$$

where x^+ are real images and x^- are negative images sampled from $q(x)$. Therefore, we add $\epsilon = 1e-4$ to the denominator of the above expression for an overall expression of L_{ML} of

$$L_{\text{ML}} = \frac{1}{N}\sum_{i=0}^{N} E_\theta(x_i^+) - \sum_{i=0}^{N} -\text{stop\_gradient}\left(\frac{e^{-E_\theta(x_i^-)}}{\sum_i e^{-E_\theta(x_i^-)} + \epsilon}\right) E(x_i^-) \tag{3}$$

In cases in which we wish to backpropogate through the sampling procedure, we add an additional L_{KL} term that enforces that decreases the distance $\text{KL}(q(x)||p(x))$. Expanding this term and ignoring entropy, we find that

$$L_{\text{KL}} = \frac{1}{N}\sum_{x_i \sim q(x)} E_{stop\_gradient(\theta)}(q(x_i)) \tag{4}$$

During training, we found that energy values can sometimes diverge. This because the above losses only enforce the relative difference between energies of positive and negative samples, but not the actual values. However, large energy values can lead to instability, thus to make the training well-behaved we add a regularization penalty to prevent very large energies

$$L_{\text{Reg}} = \frac{1}{N}\sum_i E_\theta(x_i^+)^2 + E_\theta(x_i^-)^2 \tag{5}$$

(a) Nearest neighbor images in CIFAR10 for conditional energy models (leftmost generated, seperate class per row).

(b) Nearest neighbor images in CIFAR10 for unconditional energy model (leftmost generated)

Figure 6: Nearest neighbor images for images generated with GEO

(a) Illustration of GEO on conditional model of CIFAR10

(b) Illustration of GEO on unconditional model on CIFAR10

Figure 7: Generation of images from random noise.

4.4 MPPI

To use MCMC sampling with MPPI, we use the following sampling procedure. For additional details, we refer the reader to [Williams et al., 2017].

$$\tilde{x}^k = \sum_i w_i x_i^k, \quad x_i^k \sim N(0, \sigma) + x^{k-1}, \quad w_i = \left(\frac{e^{-E_\theta(x_i^k)}}{\sum_j e^{-E_\theta(x_i^j)}} \right) \tag{6}$$

4.5 Model

We use the residual model in Figure 8a for conditional CIFAR10 images generation and the residual model in Figure 8b for unconditional CIFAR10 and Imagenet images. We found unconditional

3x3 conv2d, 128
ResBlock down 128
ResBlock 128
ResBlock down 256
ResBlock 256
ResBlock down 256
ResBlock 256
Global Sum Pooling
dense $\rightarrow$ 1

(a) Architecture used for conditional CIFAR10 experiments

3x3 conv2d, 128
ResBlock down 128
ResBlock 128
ResBlock 128
ResBlock down 256
ResBlock 256
ResBlock 256
ResBlock down 256
ResBlock 256
ResBlock 256
Global Sum Pooling
dense $\rightarrow$ 1

(b) Architecture used for unconditional CIFAR10 experiments

FC 32
Self Attention Block
Conv1D 128
Conv1D 128
dense $\rightarrow$ 1

(c) Architecture used for Time Series Experiments

models need additional capacity. Our conditional and unconditional architectures are similar to architectures in [Miyato et al., 2018].

We found definite gains with additional residual blocks. We further found that replacing global sum pooling with a fully connected network also worked but did not lead to substantial benefits. We use the zero init in [Anonymous, 2019b] and spectral normalization on all weights. We use conditional bias and gains in each residual layer for a conditional model. We found it important when down-sampling to do average pooling as opposed to strided convolutions. We use leaky ReLUs throughout the architecture.

We use the architecture in Figure 8c for particle time series regression.

4.6 Training Hyperparameters

For CIFAR10 experiments, we use 60 steps of SGLD to generate negative samples. We use a replay buffer of size of 10000 image. We scale images to be between 0 and 1. We clip gradients to have magnitude of 0.01 and use a step size of 10 for each gradient step of SGLD. We use random noise with standard deviation of 0.005. We train our model on 1 GPU for 2 days. We use the Adam Optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.999$ with a training learning rate of 1e-4. We use a batch size during training of 128 positive and negative samples. For both experiments, we clip all training gradients that are more than 3 standard deviations from the 2nd order Adam parameters. We use spectral normalization on networks without backpropagating through the sampling procedure. We use the identical setup for ImageNet 32x32 images, but train for 3 days on 1 GPU.

For trajectories, we use 20 steps of SGLD to generate negative samples. We use a noise standard deviation 0.005. We use a batch size of 256 positive and negative samples. We found that a replay buffer was not necessary. We use the Adam Optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.999$. For MPPI, we use 30 steps of simulation with 5 noise simulations per step. We found spectral normalization to be overly restrictive on trajectories so we instead backpropogate through the sampling procedur e.