Improved Contrastive Divergence Training of Energy-Based Models Supplementary

Yilun Du [1] **Shuang Li** [1] **Joshua Tenenbaum** [1] **Igor Mordatch** [2]

In this supplement, we present additional image generation results in Section A. Next we detail experimental settings in Section B. We provide derivations of gradients of $\mathcal{L}_{\text{CD}}$ and $\mathcal{L}_{\text{KL}}$ and show there equivalence to the original contrastive divergence objective in Section C. Finally we provide additional analysis of our method in Section D.

A More Image Results

A.1 Nearest Neighbor Generations

We present L2 nearest neighbors in CelebA-HQ training dataset of unconditional image samples from our trained EBM in Figure 2. We find that our approach generates images distinct from the training set.

A.2 Additional Quantitative Results

We further quantitatively compare our generations with those of SNGAN on LSUN 128x128 bedroom scenes. We find that an SNGAN model trained on LSUN 128x128 bedroom scenes obtains an FID of 64.05 compared to our approach, which obtains an FID of 33.46. To report SNGAN scores, we re-implemented the SNGAN model using the default hyper parameters to train models on ImageNet 128x128.

A.3 Additional Qualitative Images

We present qualitative visualizations of unconditional samples generated from an EBM. Figure 3 shows unconditional image generations from LSUN bedroom scenes. Figure 4 shows unconditional image generations on the CIFAR-10 dataset. Finally, Figure 5 shows unconditional image generations on the ImageNet 32x32 dataset. In all three different settings, we find that our generated unconditional images are relatively globally coherent.

[1]MIT CSAIL [2]Google Brain. Correspondence to: Yilun Du <yilundu@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

B Training Details

B.1 Model Architectures

In this part, we provide the model architectures used in our experiments. When training multiscale energy functions, our final output energy function is the sum of energy functions applied to the full resolution image, the half resolution image, and the quarter resolution image. We use the architecture reported in Table 1 for the full resolution image on CIFAR-10 and ImageNet 32x32 (used in the main paper Section 3.2 and 3.3). The model architecture used on the CelebA-HQ and LSUN datasets are reported in Table 2 (used in the main paper Section 3.2 and 3.4). The half-resolution models share the architecture listed in Table 1, but with the first down-sampled residual block removed. Similarly, the quarter resolution models share the architectures listed, but with the first two down-sampled residual blocks removed. We utilize group normalization (Wu & He, 2018) inside each residual block and utilize the Swish nonlinearity (Ramachandran et al., 2018).

B.2 Experiment Configurations For Different Datasets

CIFAR-10/ImageNet 32x32. For CIFAR-10 and ImageNet 32x32, we use 40 steps of Langevin sampling to generate a negative sample. The Langevin sampling step size is set to be 500, with Gaussian noise of magnitude 0.001 at each iteration. The data augmentation transform consists of color augmentation of strength 1.0 from (Chen et al., 2020), a random horizontal flip, and a image resize between 0.02 and 1.0. This is used in the main paper Section 3.2 and 3.3.

CelebA/LSUN Bedroom. For the CelebA-HQ and LSUN bed datasets, we use 40 steps of Langevin sampling to generate negative samples. The Langevin sampling step size is set to be 1000, with Gaussian noise of magnitude 0.001 applied at each iteration. The data augmentation transform consists of color augmentation of strength 1.0 from (Chen et al., 2020), a random horizontal flip, and a image resize between 0.08 and 1.0. This is used in the main paper Section 3.2 and 3.4.

C Loss Gradient Derivation

We show that the gradient of the contrastive divergence objective, $\mathcal{L}_{\text{CD\_Full}}$ is equivalent to that of the $\mathcal{L}_{\text{Full}} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{CD}}$. Recall that the contrastive divergence objective is given by

$$\mathcal{L}_{\text{CD\_Full}} = \text{KL}(p_D(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x})) - \text{KL}(q_\theta(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x})). \tag{1}$$

The gradient of the first KL term with respect to θ, $\frac{\partial \text{KL}(p_D(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x}))}{\partial \theta}$ is

$$-\mathbb{E}_{p_D(\boldsymbol{x})}\left[\frac{\partial E_\theta(\boldsymbol{x})}{\partial \theta}\right] \tag{2}$$

while the gradient of the second KL term with respect to θ, $\frac{\text{KL}(q_\theta(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x}))}{\partial \theta}$

$$\frac{\partial q(\boldsymbol{x}')}{\partial \theta}\frac{\partial \text{KL}(q_\theta(\boldsymbol{x}') \mid\mid p_\theta(\boldsymbol{x}'))}{\partial q_\theta(\boldsymbol{x}')}) - \mathbb{E}_{q_\theta(\boldsymbol{x}')}[\frac{\partial E_\theta(\boldsymbol{x}')}{\partial \theta}] \tag{3}$$

with the overall gradient being

$$\frac{\mathcal{L}_{\text{CD\_Full}}}{\partial \theta} = -(\mathbb{E}_{p_D(\boldsymbol{x})}\left[\frac{\partial E_\theta(\boldsymbol{x})}{\partial \theta}\right] - \mathbb{E}_{q_\theta(\boldsymbol{x}')}[\frac{\partial E_\theta(\boldsymbol{x}')}{\partial \theta}] + \frac{\partial q(\boldsymbol{x}')}{\partial \theta}\frac{\partial \text{KL}(q_\theta(\boldsymbol{x}') \mid\mid p_\theta(\boldsymbol{x}'))}{\partial q_\theta(\boldsymbol{x}')}). \tag{4}$$

We have that

$$\mathcal{L}_{\text{CD}} = \mathbb{E}_{p_D(\boldsymbol{x})}[E_\theta(\boldsymbol{x})] - \mathbb{E}_{\text{stop\_grad}(q_\theta(\boldsymbol{x}'))}[E_\theta(\boldsymbol{x}')], \tag{5}$$

with corresponding gradients

$$\frac{\partial \mathcal{L}_{\text{CD}}}{\partial \theta} = \mathbb{E}_{p_D(\boldsymbol{x})}\left[\frac{\partial E_\theta(\boldsymbol{x})}{\partial \theta}\right] - \mathbb{E}_{q_\theta(\boldsymbol{x}')}[\frac{\partial E_\theta(\boldsymbol{x}')}{\partial \theta}]. \tag{6}$$

Furthermore, we have that

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{q_\theta(\boldsymbol{x})}[E_{\text{stop\_grad}(\theta)}(\boldsymbol{x})] + \mathbb{E}_{q_\theta(\boldsymbol{x})}[\log(q_\theta(\boldsymbol{x}))], \tag{7}$$

can be rewritten as

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{q_\theta(\boldsymbol{x})}[-\log(p_\theta(x))] + \mathbb{E}_{q_\theta(\boldsymbol{x})}[\log(q_\theta(\boldsymbol{x}))] \tag{8}$$

$$= \text{KL}(q_\theta(\boldsymbol{x}) \mid\mid p_{\text{stop\_gradient}(\theta)}(\boldsymbol{x})). \tag{9}$$

The corresponding gradient of the objective is

$$\frac{\partial \mathcal{L}_{\text{KL}}}{\partial \theta} = \frac{\partial q(\boldsymbol{x})}{\partial \theta}\frac{\partial \text{KL}(q_\theta(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x}))}{\partial q_\theta(\boldsymbol{x})}. \tag{10}$$

Thus the sum of the gradients in $\frac{\partial \mathcal{L}_{\text{CD}}}{\partial \theta}$ (Equation 6) and $\frac{\partial \mathcal{L}_{\text{KL}}}{\partial \theta}$ (Equation 10) is equal to the full contrastive divergence gradient $\frac{\mathcal{L}_{\text{CD\_Full}}}{\partial \theta}$ (Equation 4).

D Additional Analysis

D.1 Alternative Sampling Distributions

Instead of utilizing $q_\theta(\boldsymbol{x})$ as $\Pi_\theta^t(p_D(\boldsymbol{x}))$, as noted in the method section, our approach can further maximize likelihood as long as $\text{KL}(p_D(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x}))$ is greater $\text{KL}(q_\theta(\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{x}))$. We test an alternative sampler $q_\theta(\boldsymbol{x})$ consisting of initializing Langevin dynamics from random noise in Figure 1. We find again that our approach improves the training stability.

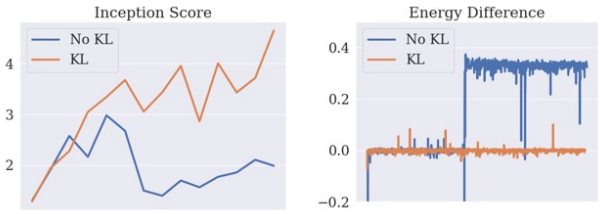

Figure 1: Inception Score and energy difference plots when $\mathcal{L}_{\text{KL}}$ is applied to MCMC initialized from random noise.

D.2 Analysis of Truncated Langevin Backpropagation

To better understand the training effect of $\mathcal{L}_{\text{KL}}$, we analyze the effect of truncating backpropogation through Langevin sampling. We train two separate models on MNIST, one with backpropogation through all Langevin steps, and one with backpropogation through only the last Langevin step. We obtain an FID of 90.54 with backpropogation through only 1 step of Langevin sampling and an FID of 94.85 with backpropogation through all steps of Langevin sampling. We present illustrations of samples generated with one step in Figure 6 and with all steps in Figure 7. Overall, we find little degradation in performance with the truncation of backpropogation, but note that backpropogation through all steps of sampling is over 3 times slower to train.

D.3 Analysis of Effect of KL Loss on Mode Sampling

We illustrate the effect of $\mathcal{L}_{\text{KL}}$ as a regularizer to prevent EBM sampling collapse. When training an EBM, $\mathcal{L}_{\text{KL}}$ serves as a repelling term encouraging MCMC samples from an EBM to both have low energy and exhibit diversity. In the absence of $\mathcal{L}_{\text{KL}}$, we find that EBM sampling always collapses and eventually always generates samples illustrated in Figure 8. These samples are significantly less diverse than those generated when training with $\mathcal{L}_{\text{KL}}$ (Figure 1), which never suffers from sampling collapse.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2018. URL https://openreview.net/forum?id=SkBYYyZRZ.

Yuxin Wu and Kaiming He. Group normalization. *arXiv:1803.08494*, 2018.

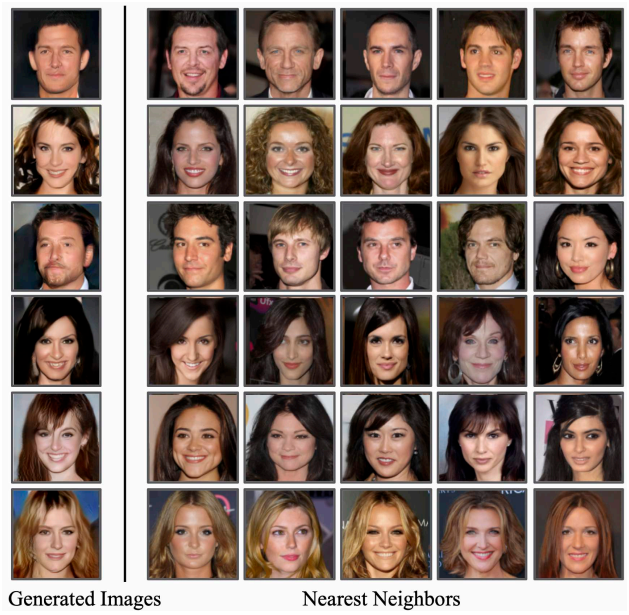

Figure 2: Nearest neighbors in the L2 space of generated images in CelebA-HQ 128x128.

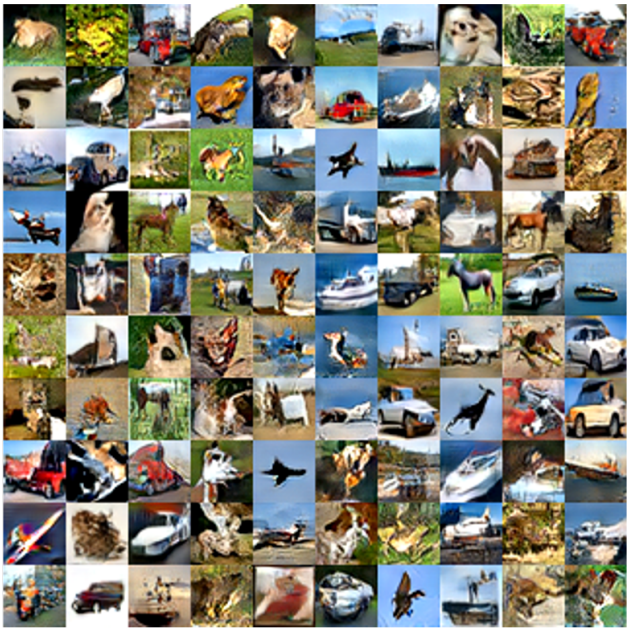

Figure 4: Randomly selected unconditional CIFAR-10 samples from our trained EBM.

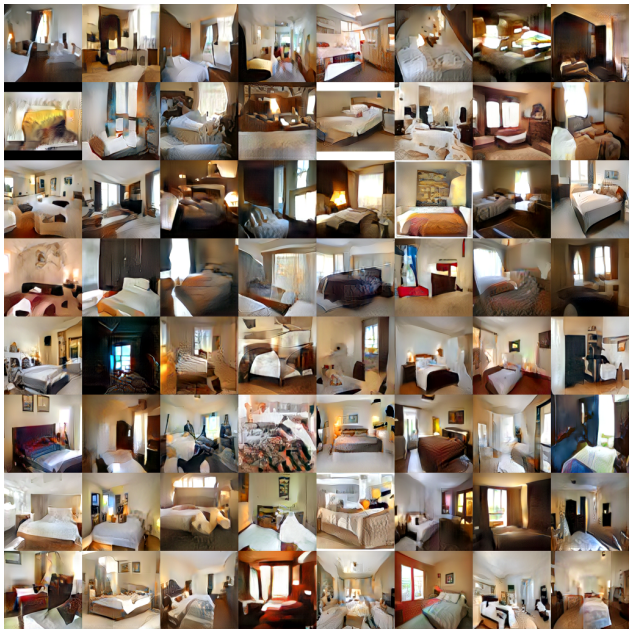

Figure 3: Randomly selected unconditional LSUN bed 128x128 samples from our trained EBM.

Figure 5: Randomly selected unconditional ImageNet 32x32 samples from our trained EBM.

Table 1: The model architecture used for CIFAR-10 and ImageNet-32x32 experiments.

3x3 conv2d, 64
ResBlock 64
ResBlock Down 64
ResBlock 64
ResBlock Down 64
Self Attention 64
ResBlock 128
ResBlock Down 128
ResBlock 256
ResBlock Down 256
Global Mean Pooling
Dense $\rightarrow$ 1

Table 2: The model architecture used for CelebA-HQ/LSUN room experiments.

3x3 conv2d, 64
ResBlock Down 64
ResBlock Down 128
ResBlock Down 128
ResBlock 256
ResBlock Down 256
Self Attention 512
ResBlock 512
ResBlock Down 512
Global mean Pooling
Dense $\rightarrow$ 1

Figure 7: Generations on MNIST with backpropogation through all steps of Langevin sampling.

Figure 6: Generations on MNIST with backpropogation through 1 step of Langevin sampling.

Figure 8: Illustration of collapsed sampling from an EBM. Sampling does not collapse with the addition of the KL loss.