Latent Space Energy-Based Prior Model for Images, Texts, and Molecules

Ying Nian Wu

Department of Statistics
University of California, Los Angeles

Statistics-Imaging Data Workshop
December 4, 2020

Bo Pang, Erik Nijkamp, Tian Han, S.-C. Zhu
Papers can be downloaded from
`http://www.stat.ucla.edu/~ywu/research.html`

Generative modeling

T Han*, E Nijkamp*, X Fang, M Hill, SC Zhu, YN Wu, CVPR, 2019

Images generated by the learned generator model:

Interpolation in latent space:

Generative modeling

R Gao, Y Song, B Poole, YN Wu, and DP Kingma (2020)

Images generated by the learned energy-based models:

Generative modeling

Images generated by the learned energy-based models:

Generator model

x: observed example. z: latent vector.

$$p_\theta(x, z) = p_\alpha(z) p_\beta(x|z)$$

Non-informative prior model: uniform or isotropic Gaussian

$$z \sim p_0(z)$$

Generator model for image:

$$x = g_\beta(z) + \epsilon$$

Generator model for sequence:

$$p_\beta(x|z) = \prod_{t=1}^{T} p_\beta(x^{(t)}|x^{(1)}, ..., x^{(t-1)}, z)$$

Mapping unimodal prior to multimodal data distribution

Energy-based prior model in latent space

B Pang*, T Han*, E Nijkamp*, SC Zhu, and YN Wu, NeurIPS, 2020

x: observed example. z: latent vector.

$$p_\theta(x, z) = p_\alpha(z) p_\beta(x|z)$$

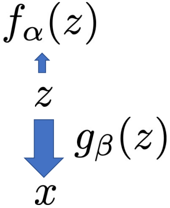
Energy-based prior model: informative, learnable, empirical Bayes

$$p_\alpha(z) = \frac{1}{Z(\alpha)} \exp(f_\alpha(z)) p_0(z)$$

$-f_\alpha(z)$: energy function, exponential tilting
$Z(\alpha)$: normalizing constant.
Standing on generator model

$$x = g_\beta(z) + \epsilon$$

Energy-based prior model in latent space

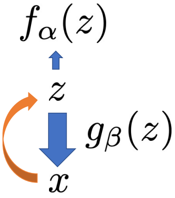


Energy-based prior model:

$$p_\alpha(z) = \frac{1}{Z(\alpha)} \exp(f_\alpha(z)) p_0(z)$$

$f_\alpha(z)$: scalar valued, value, cost or objective, regularities and rules
z: low-dimensional, small network, less multimodal, easy to sample
Origin: statistical physics, Gibbs distribution, random field

Energy-based prior model in latent space



Marginal:

$$p_\theta(x) = \int p_\theta(x, z)dz = \int p_\alpha(z)p_\beta(x|z)dz$$

Posterior:

$$p_\theta(z|x) = p_\theta(x, z)/p_\theta(x) = p_\alpha(z)p_\beta(x|z)/p_\theta(x)$$

Maximum likelihood

Training examples $(x_i, i = 1, ..., n)$.

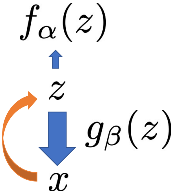
$$L(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

Learning gradient:

$$\begin{aligned}\nabla_\theta \log p_\theta(x) &= \mathrm{E}_{p_\theta(z|x)} \left[\nabla_\theta \log p_\theta(x, z)\right] \\ &= \mathrm{E}_{p_\theta(z|x)} \left[\nabla_\theta (\log p_\alpha(z) + \log p_\beta(x|z))\right]\end{aligned}$$

$p_\theta(z|x)$: inference, posterior, imputation
Similar to EM algorithm

Maximum likelihood

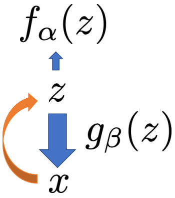


Learning gradient for prior model:

$$\delta_\alpha(x) = \nabla_\alpha \log p_\theta(x) = \mathrm{E}_{p_\theta(z|x)}[\nabla_\alpha f_\alpha(z)] - \mathrm{E}_{p_\alpha(z)}[\nabla_\alpha f_\alpha(z)]$$

$p_\alpha(z)$: prior. $p_\theta(z|x)$: posterior. Match prior to aggregated posterior
$f_\alpha(z)$: value or critic, self-critical (adversarial)

Maximum likelihood

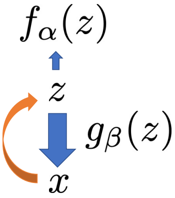


Learning gradient for generation model:

$$\delta_\beta(x) = \nabla_\beta \log p_\theta(x) = \mathrm{E}_{p_\theta(z|x)}[\nabla_\beta \log p_\beta(x|z)]$$

Reconstructing x by $g_\beta(z)$

Prior and posterior sampling

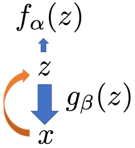


Short run MCMC (Langevin dynamics):

$$z_0 \sim p_0(z)$$
$$z_{k+1} = z_k + s\nabla_z \log \pi(z_k) + \sqrt{2s}\epsilon_k, \; k = 0, ..., K-1$$

e.g., $K = 20$
Can be amortized by learned networks for inference and synthesis sampling

Learning and sampling algorithm

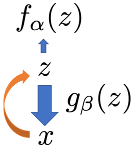


for $t = 0 : T-1$ **do**

1. **Mini-batch**: Sample observed examples $\{x_i\}_{i=1}^m$
2. **Prior sampling**: For each x_i, sample $z_i^- \sim p_{\alpha_t}(z)$
3. **Posterior sampling**: For each x_i, sample $z_i^+ \sim p_{\theta_t}(z|x_i)$
4. **Learning prior model**:
   $\alpha_{t+1} = \alpha_t + \eta_0 \frac{1}{m} \sum_{i=1}^m [\nabla_\alpha f_{\alpha_t}(z_i^+) - \nabla_\alpha f_{\alpha_t}(z_i^-)]$
5. **Learning generation model**:
   $\beta_{t+1} = \beta_t + \eta_1 \frac{1}{m} \sum_{i=1}^m \nabla_\beta \log p_{\beta_t}(x_i|z_i^+)$

end

Amortized sampling networks



Learned prior sampling: $q_\psi(z)$ (flow-based model)
Learned posterior sampling: $q_\phi(z|x)$ (encoder or inference model in VAE)
Perturbation of maximum likelihood:

$$\Delta(\theta, \phi, \psi) = D_{KL}(p_{\text{data}}(x) \| p_\theta(x)) + D_{KL}(q_\phi(z|x) \| p_\theta(z|x)) - D_{KL}(q_\psi(z) \| p_\alpha(z))$$

$$\min_\theta \min_\phi \max_\psi \Delta(\theta, \phi, \psi)$$

Positive phase for posterior sampling and negative phase for prior sampling
Variational learning and adversarial learning, contrastive divergence
Short run MCMC (or only MCMC for prior sampling)

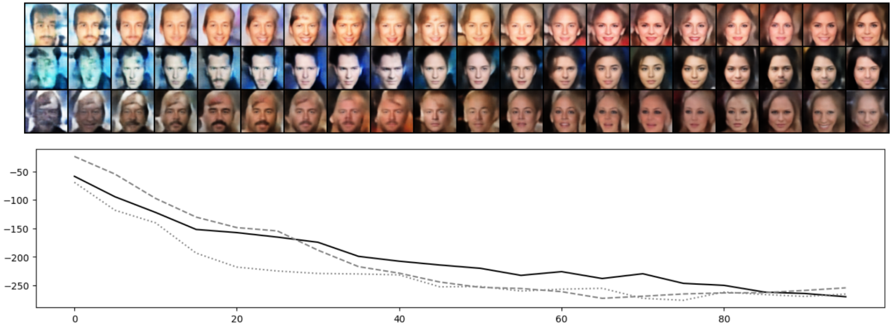
Image generation

Image generation

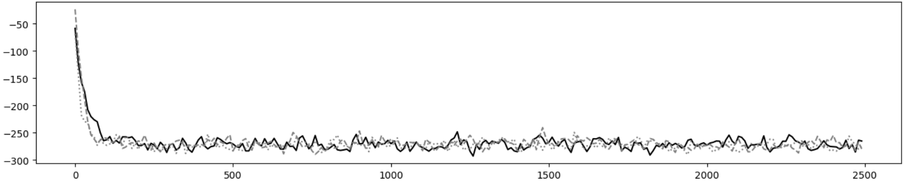
Models		VAE	2sVAE	RAE	SRI	SRI (L=5)	Ours
SVHN	MSE	0.019	0.019	0.014	0.018	0.011	**0.008**
	FID	46.78	42.81	40.02	44.86	35.23	**29.44**
CIFAR-10	MSE	0.057	0.056	0.027	-	-	**0.020**
	FID	106.37	109.77	74.16	-	-	**70.15**
CelebA	MSE	0.021	0.021	0.018	0.020	0.015	**0.013**
	FID	65.75	49.70	40.95	61.03	47.95	**37.87**

Table 1: MSE of testing reconstructions and FID of generated samples for SVHN $(32 \times 32 \times 3)$, CIFAR-10 $(32 \times 32 \times 3)$, and CelebA $(64 \times 64 \times 3)$ datasets.

Short run MCMC



Long run MCMC



Text generation

judge in <unk> was not
west virginia bank <unk> which has been under N law took effect of october N
mr. peterson N years old could return to work with his clients to pay
iras must be
anticipating bonds tied to the imperial company 's revenue of $ N million today
many of these N funds in the industrial average rose to N N from N N N
fund obtaining the the
ford 's latest move is expected to reach an agreement in principle for the sale of its loan operations
wall street has been shocked over by the merger of new york co. a world-wide financial board of the companies said it wo n't seek strategic alternatives to the brokerage industry 's directors

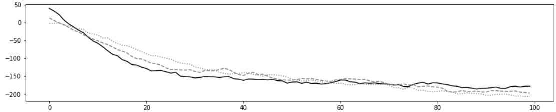


Table 3: Transition of a Markov chain initialized from $p_0(z)$ towards $\tilde{p}_\alpha(z)$. *Top:* Trajectory in the PTB data-space. Each panel contains a sample for $K'_0 \in \{0, 40, 100\}$. *Bottom:* Energy profile.

Text generation

Models	SNLI FPPL	SNLI RPPL	SNLI NLL	PTB FPPL	PTB RPPL	PTB NLL	Yahoo FPPL	Yahoo RPPL	Yahoo NLL
Real Data	23.53	-	-	100.36	-	-	60.04	-	-
SA-VAE	39.03	46.43	33.56	147.92	210.02	101.28	128.19	148.57	326.70
FB-VAE	39.19	43.47	28.82	145.32	204.11	92.89	123.22	141.14	319.96
ARAE	44.30	82.20	28.14	165.23	232.93	91.31	158.37	216.77	320.09
Ours	**27.81**	**31.96**	28.90	**107.45**	**181.54**	91.35	**80.91**	**118.08**	321.18

Table 2: Forward Perplexity (FPPL), Reverse Perplexity (RPPL), and Negative Log-Likelihood (NLL) for our model and baselines on SNLI, PTB, and Yahoo datasets.

Anomaly detection

Based on $\log p(x, z)$. Out of distribution examples

Heldout Digit	1	4	5	7	9
VAE	0.063	0.337	0.325	0.148	0.104
MEG	0.281 ± 0.035	0.401 ± 0.061	0.402 ± 0.062	0.290 ± 0.040	0.342 ± 0.034
BiGAN-σ	0.287 ± 0.023	0.443 ± 0.029	0.514 ± 0.029	0.347 ± 0.017	0.307 ± 0.028
Ours	**0.336 ± 0.008**	**0.630 ± 0.017**	**0.619 ± 0.013**	**0.463 ± 0.009**	**0.413 ± 0.010**

Table 3: AUPRC scores for unsupervised anomaly detection on MNIST.

Trajectory prediction

B Pang, T Zhao, X Xie, and YN Wu (2020)

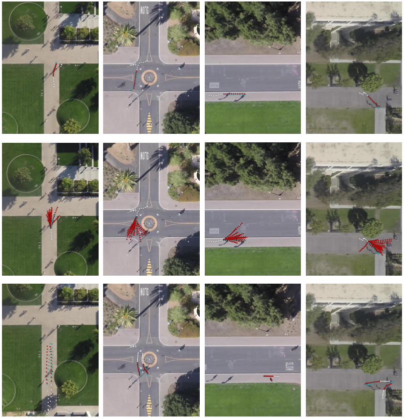


Figure 2. Qualitative results of our proposed method across 4 different scenarios in the Stanford Drone. First row: The best prediction result sampled from 20 trials from LB-EBM. Second row: The 20 predicted trajectories sampled from LB-EBM. Third row: prediction results of agent pairs that has social interactions. The observed trajectories, ground truth predictions and our model's predictions are displayed in terms of white, blue and red dots respectively.

$f_\alpha(z|c)$: value or cost of trajectory given condition, inverse control

Trajectory prediction

	ADE	FDE
S-LSTM [1]	31.19	56.97
S-GAN-P [13]	27.23	41.44
MATF [52]	22.59	33.53
Desire [21]	19.25	34.05
SoPhie [42]	16.27	29.38
CF-VAE [3]	12.60	22.30
P2TIRL [7]	12.58	22.07
SimAug [24]	10.27	19.71
PECNet [28]	9.96	15.88
Ours	**8.87**	**15.61**

Table 1. ADE / FDE metrics on Stanford Drone for several methods compared to ours are shown. The lower th

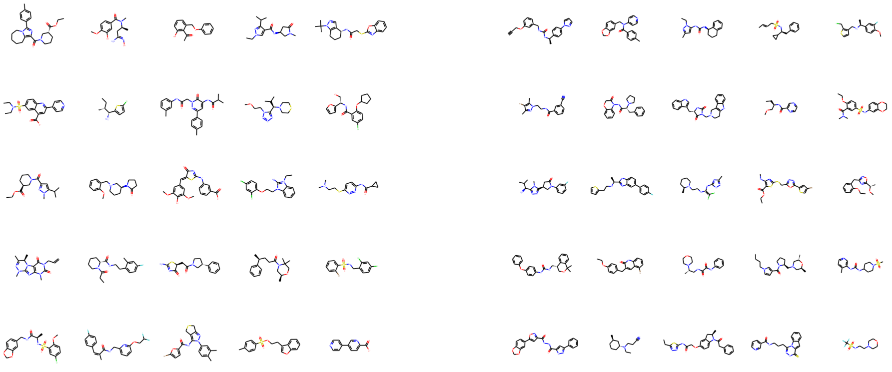
	ETH	HOTEL	UNIV	ZARA1	ZARA2	A
Linear * [1]	1.33 / 2.94	0.39 / 0.72	0.82 / 1.59	0.62 / 1.21	0.77 / 1.48	0.79
SR-LSTM-2 * [51]	0.63 / 1.25	0.37 / 0.74	0.51 / 1.10	0.41 / 0.90	0.32 / 0.70	0.45
S-LSTM [1]	1.09 / 2.35	0.79 / 1.76	0.67 / 1.40	0.47 / 1.00	0.56 / 1.17	0.72
S-GAN-P [13]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61
SoPhie [42]	0.70 / 1.43	0.76 / 1.67	0.54 / 1.24	0.30 / 0.63	0.38 / 0.78	0.54
MATF [52]	0.81 / 1.52	0.67 / 1.37	0.60 / 1.26	0.34 / 0.68	0.42 / 0.84	0.57
CGNS [22]	0.62 / 1.40	0.70 / 0.93	0.48 / 1.22	0.32 / 0.59	0.35 / 0.71	0.49
PIF [26]	0.73 / 1.65	0.30 / 0.59	0.60 / 1.27	0.38 / 0.81	0.31 / 0.68	0.46
STSGN [50]	0.75 / 1.63	0.63 / 1.01	0.48 / 1.08	0.30 / 0.65	0.26 / 0.57	0.48
GAT [19]	0.68 / 1.29	0.68 / 1.40	0.57 / 1.29	0.29 / 0.60	0.37 / 0.75	0.52
Social-BiGAT [19]	0.69 / 1.29	0.49 / 1.01	0.55 / 1.32	0.30 / 0.62	0.36 / 0.75	0.48
Social-STGCNN [30]	0.64 / 1.11	0.49 / 0.85	0.44 / 0.79	0.34 / 0.53	0.30 / 0.48	0.44
PECNet [28]	0.54 / 0.87	0.18 / 0.24	0.35 / 0.60	0.22 / 0.39	0.17 / 0.30	0.29
Ours	**0.30 / 0.52**	**0.13 / 0.20**	**0.27 / 0.52**	**0.20 / 0.37**	**0.15 / 0.29**	**0.21**

ɔle 2. ADE / FDE metrics on ETH-UCY for several methods compared to ours are shown. The models with * mark are

Molecule generation

B Pang, T Han, and YN Wu (2020)

simplified molecular input line entry systems (SMILES)



(a) ZINC (b) Generated

Figure 1: Sample molecules taken from the ZINC dataset (a) and generated by our model (b).

Latent space EBM captures chemical rules implicitly

Molecule generation

simplified molecular input line entry systems (SMILES)

Model	**Model Family**	**Validity w/ check**	**Validity w/o check**	**Novelty**	**Uniqueness**
GraphVAE (Simonovsky et al., 2018)	Graph	0.140	-	1.000	0.316
CGVAE (Liu et al., 2018)	Graph	1.000	-	1.000	0.998
GCPN (You et al., 2018)	Graph	1.000	0.200	1.000	1.000
NeVAE (Samanta et al., 2019)	Graph	1.000	-	0.999	1.000
MRNN (Popova et al., 2019)	Graph	1.000	0.650	1.000	0.999
GraphNVP (Madhawa et al., 2019)	Graph	0.426	-	1.000	0.948
GraphAF (Shi et al., 2020)	Graph	1.000	0.680	1.000	0.991
ChemVAE (Gomez-Bombarelli et al., 2018)	LM	0.170	-	0.980	0.310
GrammarVAE (Kusner et al., 2017)	LM	0.310	-	1.000	0.108
SDVAE (Dai et al., 2018)	LM	0.435	-	-	-
FragmentVAE (Podda et al., 2020)	LM	**1.000**	-	0.995	0.998
Ours	LM	0.955	-	**1.000**	**1.000**

Table 1: Performance obtained by our model against LM-based and graph-based baselines.

Latent space EBM captures chemical rules implicitly

Molecule generation

simplified molecular input line entry systems (SMILES)

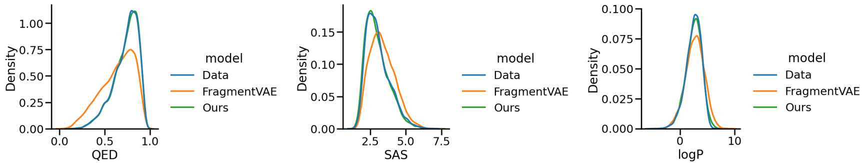
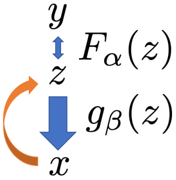


Figure 2: Distributions of molecular properties of data and 10,000 random samples from FragmentVAE and our model.

Latent space EBM captures chemical rules implicitly

Semi-supervised learning

B Pang, E Nijkamp, J Cui, T Han, and YN Wu (2020)



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

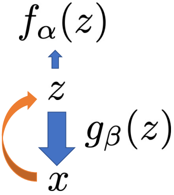
y: one-hot vector, (0, ..., 0, 1, 0, ..., 0). z: continuous dense vector.
Semi-supervised: y given for a small number of x.
Symbol-vector coupling, associative memory:

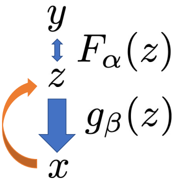
$$p_\alpha(y, z) = \frac{1}{Z(\alpha)} \exp(\langle y, F_\alpha(z)\rangle) p_0(z)$$

$F_\alpha(z) = (F_\alpha^{(1)}(z), ..., F_\alpha^{(c)}(z), ..., F_\alpha^{(C)}(z))$: logit scores for C categories
Soft-max classifier:

$$p_\alpha(y|z) \propto \exp(\langle y, F_\alpha(z)\rangle) = \exp(F_\alpha^{(y)}(z))$$

Symbol-vector coupling



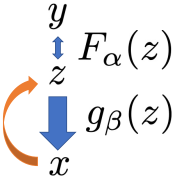


Marginal energy-based prior:

$$p_\alpha(z) = \frac{1}{Z(\alpha)} \exp(f_\alpha(z)) p_0(z)$$

$$f_\alpha(z) = \log \sum_y \exp(\langle y, F_\alpha(z) \rangle)$$

Likelihood-based semi-supervised learning



Only some x are labeled with y:

$$L(\theta) = \sum_{all} \log p_\theta(x) + \lambda \sum_{labeled} \log p_\theta(y|x)$$

Image data

Method	SVHN 1000 Labels	CIFAR-10 4000 Labels
VAE M1+M2	64.0 ± 0.1	-
AAE	82.3 ± 0.3	-
JEM	66.0 ± 0.7	-
FlowGMM	82.4	78.2
Ours	92.0 ± 0.1	78.6 ± 0.3
TripleGAN	94.2 ± 0.2	83.0 ± 0.4
BadGAN	95.8 ± 0.03	85.6 ± 0.03
Π-Model	94.6 ± 0.2	83.6 ± 0.3
VAT	96.3 ± 0.1	88.0 ± 0.1

Table 1: Accuracy on SVHN and CIFAR-10.

Text data

Method	AGNews-Unigram 200 Labels
Self-training	77.3 ± 1.7
Glove (ID)	70.4 ± 1.2
Glove (OD)	68.8 ± 5.7
VAMPIRE	81.9 ± 0.5
Ours	84.5 ± 0.3

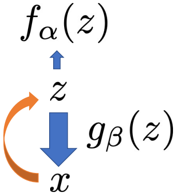
Table 3: Accuracy on AGNews with Unigram.

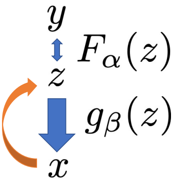
Tabular data

Method	Hepmass 20 Labels	Miniboone 20 Labels	Protein 100 Labels
RBF Label Spreading	84.9	79.3	-
JEM	-	-	19.6
FlowGMM	88.5 ± 0.2	80.5 ± 0.7	-
Ours	89.1 ± 0.1	81.2 ± 0.3	23.1 ± 0.3
Π-Model	87.9 ± 0.2	80.8 ± 0.01	-
VAT	-	-	17.1

Table 4: Accuracy on Hepmass, Miniboone, and Protein.

Discussion





Energy-based model in latent space: simple and expressive
Symbol-vector coupling: hippocampus, entorhinal cortex, visual cortex?
Fast learning f_α and slow learning g_β?