On Out-of-distribution Detection with Energy-based Models

Sven Elflein, Bertrand Charpentier, Daniel Zügner, Stephan Günnemann
Technical University of Munich

Problem Definition & Contribution

Goal: Investigate superior OOD detection performance of EBMs vs. other generative models.

Motivation:

- Recent research on density estimation focuses on exact likelihood methods.
- Findings of superior OOD detection performance of EBMs without analysis.

Key Contributions:

- Find that EBMs do not strictly outperform Normalizing Flows across multiple training methods.
- Identify that learning semantic features induced by supervision improves OOD detection in recent discriminative EBMs.
- Show that one can use architectural modifications to improve OOD detection with EBMs.

Method

Energy-based model (EBM). Energy-function E_θ defines a density over the data x as

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)} \tag{1}$$

where $Z(\theta) = \int \exp(-E_\theta(x))dx$.

Joint Energy model (JEM) [2]. Given a classifier $f : \mathbb{R}^D \mapsto \mathbb{R}^C$ assigning logits for C classes for a datapoint $x \in \mathbb{R}^D$

$$p_\theta(y \mid x) = \frac{\exp(f_\theta(x)[y])}{\sum_{y'} \exp(f_\theta(x)[y'])} \tag{2}$$

where $f_\theta(x)[y]$ denotes the y-th logit. The logits $f_\theta(x)[y]$ can be interpret as unnormalized probabilities of the joint distribution $p_\theta(x, y)$ which yields the marginal distribution over x as

$$p_\theta(x) = \sum_y p_\theta(x, y) = \sum_y \frac{\exp(f(x)[y])}{Z(\theta)} \tag{3}$$

Training. We follow [2] and optimize the factorization

$$\log p_\theta(x, y) = \log p_\theta(x) + \log p_\theta(y \mid x) \tag{4}$$

using 2 and 3. In particular, we use a Cross Entropy objective to optimize $p_\theta(y \mid x)$ weighted with hyperparameter γ.

For optimizing $p_\theta(x)$, we consider different approaches which have shown to scale to high-dimensional data.

Sliced score matching (SSM) [4]. Efficient update formula based on random projection

$$\mathbb{E}_{p_v}\mathbb{E}_{p(x)}\left[v^T \nabla_x s_\theta(x) v + \frac{1}{2}\|s_\theta(x)\|_2^2\right] \tag{5}$$

where $v \sim p_v$ is a simple distribution of random vectors.

Contrastive divergence (CD) [3]. Approximation of the gradient of the maximum likelihood objective by

$$\nabla_\theta p_\theta(x) = \mathbb{E}_{p_\theta(x')}\left[\nabla_\theta E_\theta(x')\right] - \nabla_\theta E_\theta(x) \tag{6}$$

VERA [1]. Learn the parameters ϕ of a auxiliary distribution q_ϕ as the optimum of

$$\log Z(\theta) = \max_{q_\phi} \mathbb{E}_{q_\phi(x)}\left[f_\theta(x)\right] + H(q_\phi) \tag{7}$$

which can be plugged into 1 to obtain an alternative method for training EBMs with a variational approximation to estimate the entropy term H_{q_ϕ}.

OOD Detection.
For OOD detection, we compute the density $p_\theta(x)$ at the considered datapoint x. We treat ID data as class 1 and OOD data as class 0 and compute AUC-PR.

Experiments & Results

Differentiation of natural and non-natural dataset.

<u>*Natural OOD*</u>: Requires learning semantic features to differentiate, e.g., images of classes not in training set.
<u>*Non-natural OOD*</u>: Requires detection farther from the training data manifold, e.g., noise, OODomain

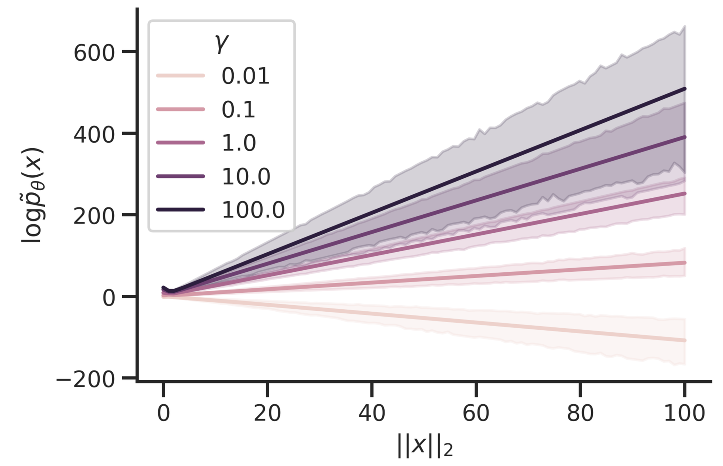
Are EBMs better than Normalizing Flows?

EBMs do not consistently outperform Normalizing Flows across different training methods (Improvements: CD 11.9%, VERA 4.3%, SSM −4.3%)

Does supervision improve OOD detection?

Model	ID dataset	Natural	Non-natural
CD	CIFAR-10	-10.82	-9.11
	FMNIST	47.17	3.24
	Segment	1.85	0.89
	Sensorless	29.72	-0.02
SSM	CIFAR-10	7.33	-27.94
	FMNIST	50.61	-20.26
	Segment	25.89	-21.94
	Sensorless	22.13	-40.73
VERA	CIFAR-10	-1.16	-3.00
	FMNIST	33.66	-15.53
	Segment	4.98	-0.57
	Sensorless	97.93	0.07

- Leveraging supervision with JEMs improves some results significantly on *natural* OOD datasets
- Results do not improve or even degrade on *non-natural* OOD datasets
- Investigation shows: Weighting parameter of cross-entropy loss γ affects the density estimates far from training data
  $\rightarrow$ *Non-natural* data becomes harder to detect



Sidestepping tuning of γ

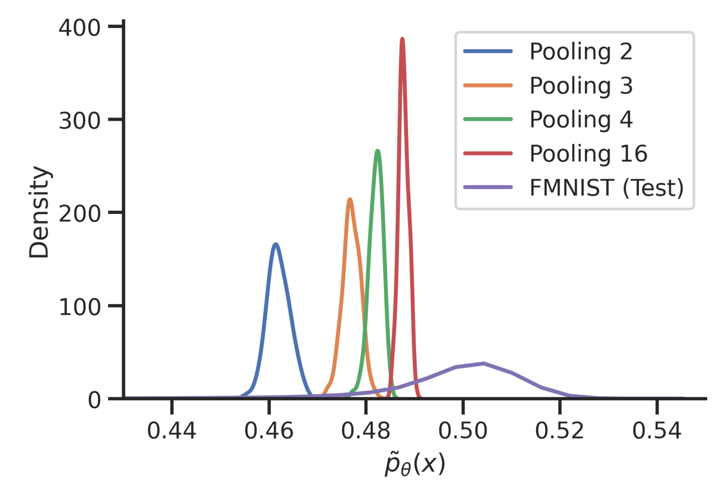
Model	ID dataset	Natural	Non-natural
CD	CIFAR-10	48.60	3.37
	FMNIST	95.79	-13.52
SSM	CIFAR-10	53.84	-2.31
	FMNIST	58.40	59.59
VERA	CIFAR-10	50.16	16.97
	FMNIST	15.12	1.80

- Introduce supervision by training on embeddings obtained from classification model
- OOD detection significantly improves on *natural* and in cases on *non-natural* datasets
- Shows that vanilla EBMs struggle to extract high-level, semantic features

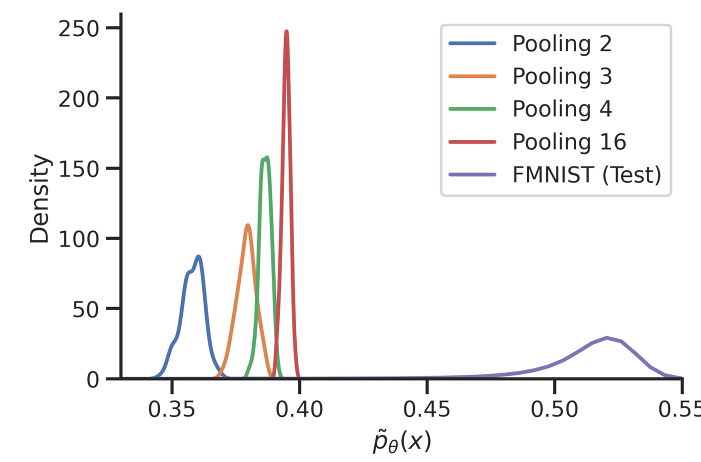
Can we encourage semantic features?

Model	ID dataset	Natural	Non-natural
CD	CIFAR-10	20.18	20.38
	FMNIST	67.95	10.88
SSM	CIFAR-10	14.76	33.34
	FMNIST	1.75	-5.92
VERA	CIFAR-10	19.66	33.22
	FMNIST	26.84	32.94

- Introduce bottlenecks through 1×1 convolutions
- OOD detection improves consistently upon the baseline EBMs by learning higher-level features
- The difference in density assigned for low-level features to images increases significantly



No Bottleneck.



With Bottleneck.

References

[1] Will Grathwohl, Jacob Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No MCMC for me: Amortized sampling for fast and stable training of energy-based models. *arXiv:2010.04230 [cs]*, October 2020.

[2] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. *arXiv:1912.03263 [cs, stat]*, September 2020.

[3] Geoffrey E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, August 2002.

[4] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced Score Matching: A Scalable Approach to Density and Score Estimation. *arXiv:1905.07088 [cs, stat]*, June 2019.