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

Sven Elflein, Bertrand Charpentier, Daniel Zügner, Stephan Günnemann

Project page: https://www.daml.in.tum.de/ood-ebm/

TL;DR

- EBMs do **not** strictly outperform Normalizing Flows across multiple training methods.
- Semantic features induced by **supervision improves OOD detection** in recent discriminative EBMs [2].
- **Architectural modifications** can also be used to improve OOD detection with EBMs.

What is an energy-based model?

EBM defines a probability distribution over the data $\mathbf{x} \in \mathbb{R}^D$ through the energy function E_θ as

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

where $Z(\theta) = \int \exp(-E_\theta(\mathbf{x}))d\mathbf{x}$ is the normalizing constant.

Properties:

✓ Flexible transformations	✗ Exact density evaluation
✓ Dimensionality reduction	✗ No direct maximum likelihood training

Why study EBMs for OOD detection?

- Recent research on generative models for OOD detection focuses on exact likelihood methods, e.g., Normalizing Flows
  → We compare behavior for EBMs vs. Normalizing Flows
- EBMs show good OOD detection capabilities [2], however, no detailed analysis has been done
  → We consider EBMs trained with different approaches
  → We investigate influences like supervision, dimensionality reduction, and architecture

How to train an EBM?

We consider three approaches to train EBMs.

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

$$\mathbb{E}_{p_\mathbf{v}}\mathbb{E}_{p(\mathbf{x})}\left[\mathbf{v}^T\nabla_\mathbf{x} s_\theta(\mathbf{x})\mathbf{v} + \frac{1}{2}\|s_\theta(\mathbf{x})\|_2^2\right]$$

where $s_\theta(\mathbf{x}) = \nabla_\mathbf{x} p_\theta(\mathbf{x})$ and $\mathbf{v} \sim p_\mathbf{v}$ is a simple distribution of random vectors.

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

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

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

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

which can be plugged into Eq. (1) to obtain an alternative method for training EBMs with a variational approximation to estimate the entropy term $H(q_\phi)$.

Setup

Natural and non-natural datasets. *Natural datasets*, e.g., images of other classes, require learning semantic features to differentiate. *Non-natural datasets*, e.g., noise, require detection farther from the training data manifold.

OOD detection eval. We compute the density $p_\theta(\mathbf{x})$ and treat ID data as class 1 and OOD data as class 0 to compute AUC-PR.

Experiments & Results

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%)

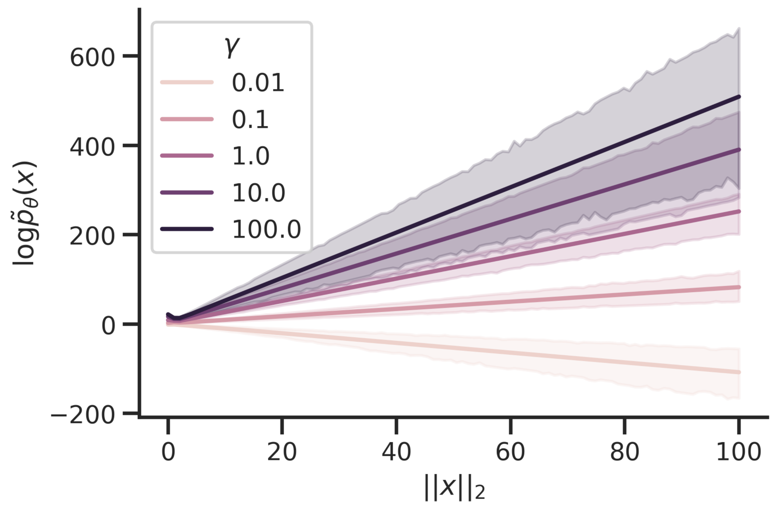
Does supervision improve OOD detection?

Use Joint Energy model (JEM) which incorporates supervision through cross-entropy objective.

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

% improvement in AUC-PR

- Supervision improves some results significantly on *natural* OOD datasets, but degrade on *non-natural* datasets.
- Investigation shows: Weighting parameter of cross-entropy loss γ affects the density estimates far from training data
  → *Non-natural* data becomes harder to detect



Sidestepping tuning of γ

Introduce supervision indirectly by training on embeddings obtained from a classification model.

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

% improvement in AUC-PR

- OOD detection significantly improves on *natural* and in cases on *non-natural* datasets
- Shows that vanilla **EBMs struggle to extract high-level, semantic features**

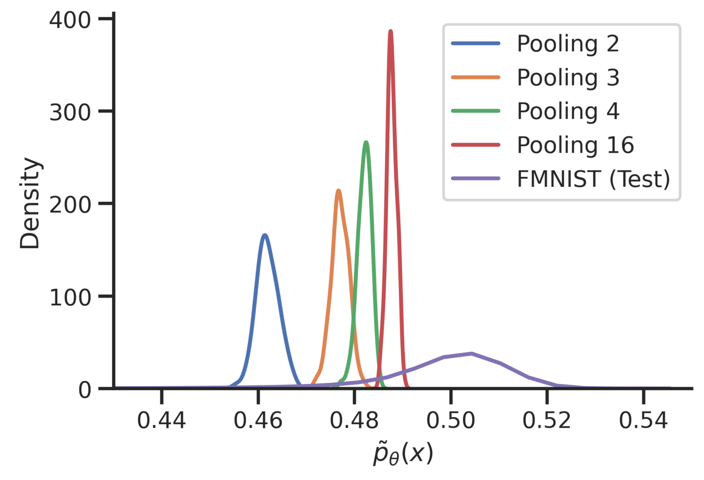
Can we encourage semantic features?

Introduce bottlenecks through 1×1 convolutions into the architecture.

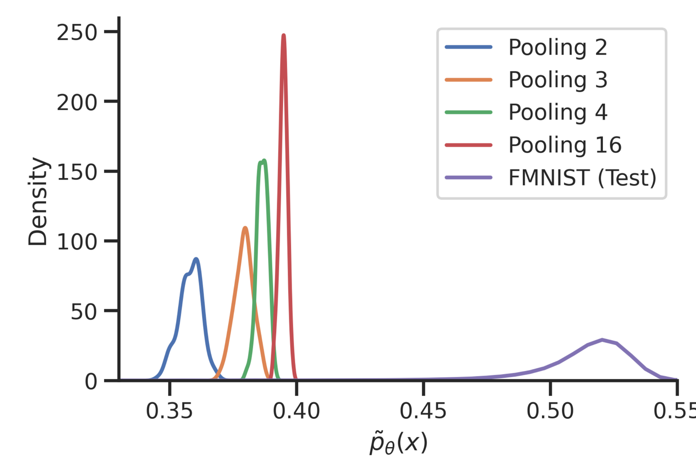
Model	ID dataset	Natural	Non-natural
CD	CIFAR-10	20.18	20.38
	FMNIST	67.95	10.88
SSM	CIFAR-10	14.76	33.34
	FMNIST	1.75	-5.92
VERA	CIFAR-10	19.66	33.22
	FMNIST	26.84	32.94

% improvement in AUC-PR

- OOD detection improves consistently upon the baseline EBMs by learning higher-level features
- The difference in density assigned for low-level features compared to images increases significantly



No Bottleneck.



With Bottleneck.

References

[1] W. Grathwohl, J. Kelly, M. Hashemi, M. Norouzi, K. Swersky, and D. Duvenaud. No MCMC for me: Amortized sampling for fast and stable training of energy-based models. *arXiv:2010.04230 [cs]*, Oct. 2020.

[2] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. *arXiv:1912.03263 [cs, stat]*, Sept. 2020.

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

[4] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced Score Matching: A Scalable Approach to Density and Score Estimation. *arXiv:1905.07088 [cs, stat]*, June 2019.