Title: Energy-based model

URL: https://en.wikipedia.org/wiki/Energy-based_model

PageID: 63087895

Categories: Category:Hamiltonian mechanics, Category:Machine learning, Category:Statistical mechanics, Category:Statistical models

-----

An energy-based model (EBM) (also called Canonical Ensemble Learning or Learning via Canonical Ensemble – CEL and LCE , respectively) is an application of canonical ensemble formulation from statistical physics for learning from data . The approach prominently appears in generative artificial intelligence .

EBMs provide a unified framework for many probabilistic and non-probabilistic approaches to such learning, particularly for training graphical and other structured models. [citation needed]

An EBM learns the characteristics of a target dataset and generates a similar but larger dataset. EBMs detect the latent variables of a dataset and generate new datasets with a similar distribution.

Energy-based generative neural networks is a class of generative models , which aim to learn explicit probability distributions of data in the form of energy-based models, the energy functions of which are parameterized by modern deep neural networks .

Boltzmann machines are a special form of energy-based models with a specific parametrization of the energy.

Description

For a given input x {\displaystyle x} , the model describes an energy E θ (x) {\displaystyle E_{\theta }(x)} such that the Boltzmann distribution P θ (x) = exp ■ (– β E θ (x)) / Z (θ) {\displaystyle P_{\theta }(x)=\exp(-\beta E_{\theta }(x))/Z(\theta)} is a probability (density), and typically β = 1 {\displaystyle \beta =1} .

Since the normalization constant:

Z (θ) := ∫ x ∈ X exp ■ (– β E θ (x)) d x {\displaystyle Z(\theta):=\int _{x\in X}\exp(-\beta E_{\theta }(x))dx}

(also known as the partition function) depends on all the Boltzmann factors of all possible inputs x {\displaystyle x} , it cannot be easily computed or reliably estimated during training simply using standard maximum likelihood estimation .

However, for maximizing the likelihood during training, the gradient of the log-likelihood of a single training example x {\displaystyle x} is given by using the chain rule :

∂ θ log ■ (P θ (x)) = E x ′ ~ P θ [∂ θ E θ (x ′)] – ∂ θ E θ (x) (∗) {\displaystyle \partial _{\theta }\log \left(P_{\theta }(x)\right)=\mathbb {E} _{x'\sim P_{\theta }}[\partial _{\theta }E_{\theta }(x')]-\partial _{\theta }E_{\theta }(x)\,(*)}

The expectation in the above formula for the gradient can be approximately estimated by drawing samples x ′ {\displaystyle x'} from the distribution P θ {\displaystyle P_{\theta }} using Markov chain Monte Carlo (MCMC).

Early energy-based models, such as the 2003 Boltzmann machine by Hinton , estimated this expectation via blocked Gibbs sampling . Newer approaches make use of more efficient Stochastic Gradient Langevin Dynamics (LD), drawing samples using:

x 0 ′ ~ P 0 , x i + 1 ′ = x i ′ – α 2 ∂ E θ (x i ′) ∂ x i ′ + ■ {\displaystyle x_{0}'\sim P_{0},x_{i+1}'=x_{i}'-{\frac {\alpha }{2}}{\frac {\partial E_{\theta }(x_{i}')}{\partial x_{i}'}}+\epsilon } , where ■ ~ N (0 , α) {\displaystyle \epsilon \sim {\mathcal {N}}(0,\alpha)} . A replay buffer of past values x i ′ {\displaystyle x_{i}'} is used with LD to initialize the optimization module.

The parameters θ {\displaystyle \theta } of the neural network are therefore trained in a generative manner via MCMC-based maximum likelihood estimation: the learning process follows an "analysis by synthesis" scheme, where within each learning iteration, the algorithm samples the synthesized examples from the current model by a gradient-based MCMC method (e.g., Langevin dynamics or Hybrid Monte Carlo), and then updates the parameters θ {\displaystyle \theta } based on the difference between the training examples and the synthesized ones – see equation (∗) {\displaystyle (*)} . This process can be interpreted as an alternating mode seeking and mode shifting process, and also has an adversarial interpretation.

Essentially, the model learns a function E θ {\displaystyle E_{\theta }} that associates low energies to correct values, and higher energies to incorrect values.

After training, given a converged energy model E θ {\displaystyle E_{\theta }} , the Metropolis–Hastings algorithm can be used to draw new samples. The acceptance probability is given by:

P a c c (x i → x ∗) = min (1 , P θ (x ∗) P θ (x i)) . {\displaystyle P_{acc}(x_{i}\to x^{*})=\min \left(1,{\frac {P_{\theta }(x^{*})}{P_{\theta }(x_{i})}}\right).}

History

The term "energy-based models" was first coined in a 2003 JMLR paper where the authors defined a generalisation of independent components analysis to the overcomplete setting using EBMs.

Other early work on EBMs proposed models that represented energy as a composition of latent and observable variables.

Characteristics

EBMs demonstrate useful properties:

Simplicity and stability–The EBM is the only object that needs to be designed and trained. Separate networks need not be trained to ensure balance.

Adaptive computation time–An EBM can generate sharp, diverse samples or (more quickly) coarse, less diverse samples. Given infinite time, this procedure produces true samples.

Flexibility–In Variational Autoencoders (VAE) and flow-based models , the generator learns a map from a continuous space to a (possibly) discontinuous space containing different data modes. EBMs can learn to assign low energies to disjoint regions (multiple modes).

Adaptive generation–EBM generators are implicitly defined by the probability distribution, and automatically adapt as the distribution changes (without training), allowing EBMs to address domains where generator training is impractical, as well as minimizing mode collapse and avoiding spurious modes from out-of-distribution samples.

Compositionality–Individual models are unnormalized probability distributions, allowing models to be combined through product of experts or other hierarchical techniques.

Experimental results

On image datasets such as CIFAR-10 and ImageNet 32x32, an EBM model generated high-quality images relatively quickly. It supported combining features learned from one type of image for generating other types of images. It was able to generalize using out-of-distribution datasets, outperforming flow-based and autoregressive models . EBM was relatively resistant to adversarial perturbations, behaving better than models explicitly trained against them with training for classification.

Applications

Target applications include natural language processing , robotics and computer vision .

The first energy-based generative neural network is the generative ConvNet proposed in 2016 for image patterns, where the neural network is a convolutional neural network . The model has been generalized to various domains to learn distributions of videos, and 3D voxels. They are made more effective in their variants. They have proven useful for data generation (e.g., image synthesis, video synthesis, 3D shape synthesis, etc.), data recovery (e.g., recovering videos with missing pixels or image frames, 3D super-resolution, etc), data reconstruction (e.g., image reconstruction and linear interpolation).

Alternatives

EBMs compete with techniques such as variational autoencoders (VAEs), generative adversarial networks (GANs) or normalizing flows .

Extensions

Joint energy-based models

Joint energy-based models (JEM), proposed in 2020 by Grathwohl et al., allow any classifier with softmax output to be interpreted as energy-based model. The key observation is that such a classifier is trained to predict the conditional probability p θ (y | x) = e f → θ (x) [y] ∑ j = 1 K e f → θ (x) [j] for y = 1 , … , K and f → θ = (f 1 , … , f K) ∈ R K , {\displaystyle p_{\theta }(y|x)={\frac {e^{{\vec {f}}_{\theta }(x)[y]}}{\sum _{j=1}^{K}e^{{\vec {f}}_{\theta }(x)[j]}}}\ \ {\text{ for }}y=1,\dotsc ,K{\text{ and }}{\vec {f}}_{\theta }=(f_{1},\dotsc ,f_{K})\in \mathbb {R} ^{K},} where f → θ (x) [y] {\displaystyle {\vec {f}}_{\theta }(x)[y]} is the y-th index of the logits f → {\displaystyle {\vec {f}}} corresponding to class y.

Without any change to the logits it was proposed to reinterpret the logits to describe a joint probability density:

with unknown partition function Z (θ) {\displaystyle Z(\theta)} and energy E θ (x , y) = − f θ (x) [y] {\displaystyle E_{\theta }(x,y)=-f_{\theta }(x)[y]} .

By marginalization, we obtain the unnormalized density

therefore,

so that any classifier can be used to define an energy function E θ (x) {\displaystyle E_{\theta }(x)} .

See also

Empirical likelihood

Posterior predictive distribution

Contrastive learning

Literature

Implicit Generation and Generalization in Energy-Based Models Yilun Du, Igor Mordatch https://arxiv.org/abs/1903.08689

Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One, Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, Kevin Swersky https://arxiv.org/abs/1912.03263

Energy-Based Transformers are Scalable Learners and Thinkers, Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, Tariq https://arxiv.org/abs/2507.02092

References

External links

"CIAR NCAP Summer School" . www.cs.toronto.edu . Retrieved 2019-12-27 .

Dayan, Peter; Hinton, Geoffrey; Neal, Radford; Zemel, Richard S. (1999), "Helmholtz Machine", Unsupervised Learning , The MIT Press, doi : 10.7551/mitpress/7011.003.0017 , hdl :

21.11116/0000-0002-D6D3-E , ISBN 978-0-262-28803-3

Hinton, Geoffrey E. (August 2002). "Training Products of Experts by Minimizing Contrastive Divergence". Neural Computation . 14 (8): 1771– 1800. doi : 10.1162/089976602760128018 . ISSN 0899-7667 . PMID 12180402 . S2CID 207596505 .

Salakhutdinov, Ruslan; Hinton, Geoffrey (2009-04-15). "Deep Boltzmann Machines" . Artificial Intelligence and Statistics : 448– 455.