Report on "Your Classifier is secretly an energy based model and you should treat it like one" by Grathwol et al

Ronak Dedhiya Dinesh (ronakdedhiya@iisc.ac.in; SR No: 06-18-05-19-52-21-1-20116)
Prashanth N Bhat(prashanthbn@iisc.ac.in; SR No: 04-03-06-19-52-21-1-20082)

September 28, 2024

1 Introduction

The paper [1] proposes an interesting interpretation of a classifier. The standard discriminative classifier $p(y|x)$ is interpreted as an energy based model for the joint distribution $p(x, y)$. This framework allows standard classification as well as calculation of $p(x)$ and $p(x|y)$. It can also be used to train on unlabelled data.

Generative models generate data that can be used for downstream tasks such as semi-supervised learning, generation of missing data, and calibration of uncertainty etc. However, the performance of generative models are not as good as purpose built solutions to such tasks. This could be due to the fact that downstream tasks are discriminative in nature. Energy-based models (EBMs) fit more naturally in a disriminative framework.

2 Energy-based models

EBMs use the observation that any probability density $p(x)$ where $x \in \mathbb{R}^D$ is expressed as

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

where $E_\theta(x) : \mathbb{R}^D \to \mathbb{R}$ is the *energy function* and $Z(\theta) = \int_x exp(-E_\theta(x))$ is the *partition function*.

It is not possible to compute $Z(\theta)$ for most choices of E_θ.

The gradient of log probability is calculated to be

$$\nabla_x = -\nabla_x E_\theta(x)$$

We can then sample from the distribution using

$$x \sim p_\theta(x)$$

$$x_{i+1} = x_i - \frac{\alpha}{2}\frac{\partial E_\theta(x|i)}{\partial x_i} + \epsilon$$

$$\epsilon \sim \mathbb{N}(0, \alpha)$$

3 Classifier

A classifier with K classes is defined as $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$. f_θ outputs K real-valued numbers called logits which can then be used to parameterize a categorical distribution using the softmax transfer function:

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

where $f_\theta(x)[y]$ indicates the y^{th} index of $f_\theta(x)$ i.e the logit corresponding to the y^{th} class label.

The logits obtained from f_θ can be re-interpreted to define $p(x, y)$ and $p(y)$ as well. Defining an energy based model of joint distribution of x and y is

$$p_\theta(x, y) = \frac{exp(f_\theta(x)[y]}{Z(\theta)}$$

where $Z(\theta)$ is an unknown normalizing constant and $E_\theta(x, y) = -f_\theta(x)[y]$.

Marginalizing on y,

$$p_\theta(x) = \sum_y p_\theta(x, y) = \frac{\sum_y exp(f_\theta(x)}{Z(\theta)}$$

Now, the LogSumExp(.) of the logits of any classifier can be re-used as the energy function at a datapoint x as

$$E_\theta(x) = -LogSumExp_y(f_\theta(x)[y]) = -log\sum_y exp(f_\theta(x)[y])$$

In a typical classifier, shifting the logits $f_\theta(x)$ by a scalar does not affect the model, but here, shifting the logits for a data point x will affect $logp_\theta(x)$. There is an extra degree of freedom hidden within the logits which can be used to define the density function over input examples as well as the joint density among examples and labels.

Also, $p_\theta(y|x) = p_\theta(x, y)/p)_\theta(x)$ so that the normalizing constant cancels out, yielding the standard softmax parameterization. A generative model hidden in the standard discriminative model is found. This method is called Joint Energy-based Model (JEM) whose general architecture is presented in Figure 1

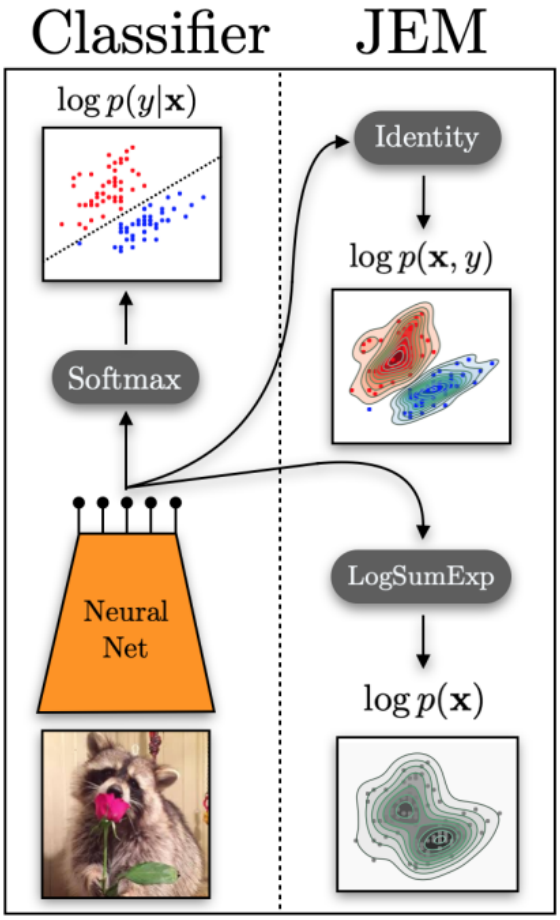

Figure 1: JEM architecture. Image copied from [1]

4 Applications

4.1 Calibration

A classifier is said to be calibrated if its predictive confidence, $max_y p(y|x)$ (ex. 0.9) aligns with its accuracy(ex.: 90% chance of being correct). Classifiers can be highly accurate but have low calibration. JEM on the other hand has near perfect calibration. So, in applications where calibration is more important than accuracy, JEM is a very useful model.

4.2 Robustness

Classifiers that are adversarially robust can be used to generate convincing images, do in-painting, and translate examples from one class to another. This is done through iterative refinement procedure. A classifier derived from EBM, of which JEM is an example is exceptionally robust to adversarial inputs.

5 Limitations

Normalized likelihood cannot be computed for an EBM, so it is challenging to accurately judge the learning rate. One way to check would be to generate samples to assess learning, but this method cannot be measured or assessed.

Gradient estimators to train JEMs are unstable and are prone to diverging if the sampling and estimation parameters are not tuned correctly. Diverged models will need to restart the learning process with a lower learning rate and increased regularization.

6 Conclusion

The paper presents a novel architecture that retains the performance of discriminative models while adding the benefits of generative modeling approaches.

References

[1] 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, 2019.