1. Introduction

Out-of-distribution (OOD) detection is a fundamental challenge in the field of computer science and engineering, as it plays a crucial role in ensuring the reliability and safety of machine learning models in real-world applications. The ability to identify instances that lie outside the distribution of the training data is essential for preventing unexpected model behavior and erroneous predictions. Traditional methods for OOD detection, such as uncertainty estimation and confidence scores, have demonstrated some success but often suffer from limitations when facing complex, high-dimensional data and distribution shifts[1].

In recent years, energy-based models have emerged as a promising approach to tackle the OOD detection problem. These models leverage the concept of energy functions, which assign low energy values to in-distribution data and high energy values to out-of-distribution samples. The key idea behind energy-based OOD detection is to train the model on a dataset representing the in-distribution data while making it generalize to identify novel and unseen samples during inference.

In this work, various methodologies and techniques using energy-based models has been explored to improve the accuracy and robustness of OOD detection systems. This research aims to advance the state-of-the-art in OOD detection.

2. Energy based models

The history of EBMs is long and dates back to 80 of the previous century when models dubbed Boltzmann Machines which is taken from statistical physics were proposed. In contrast to traditional methods, energy-based approaches do not require explicit probabilistic assumptions and can better handle complex data distributions. By formulating the detection task as an energy minimization problem, the model can learn to assign low energy to familiar data patterns and high energy to unfamiliar ones. This makes energy-based methods more flexible, interpretable, and capable of capturing subtle differences in the data distribution.

The idea behind EBMs is that we may create an energy function, E(x), that gives a value (energy) to a certain state rather than suggesting a particular distribution, such as Gaussian or Bernoulli. Then, the probability distribution could be obtained by transforming the energy to the unnormalized probability $e^{-E(x)}$ and normalizing it by Z = $\sum_x e^{-E(x)}$ that yields the Boltzmann (also called Gibbs) distribution[2]:

$$\mathrm{P(x)} = \frac{e^{-E(x)}}{Z} \tag{2.1}$$

To compute the probability for each point, we divide all exponentiated energy by their sum, much as we do when calculating softmax. When dealing with continuous random variables, we must normalise by computing an integral. Therefore, the Gaussian distribution, for example, can be described as the Boltzmann distribution with an analytically tractable partition function and the energy function, that yields:

$$\mathrm{P(x)} = \frac{e^{-E(x)}}{\int e^{-E(x)}dx} = \frac{e^{\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi\sigma^2}} \tag{2.2}$$

Now we formulate the learning model based on energy. We define an energy function with certain observable and decision random variables, E(x, y; θ), that assigns a value (an energy) to a pair (x, y) where $x \in R_D$ and $y \in \{0, 1, . . . , K - 1\}$.

$$E(x,y;\theta) = -NN_{\theta}(x)[y] \tag{2.3}$$

where we indicate by [y] the specific output of the neural net NNθ (x). Then, the joint probability distribution is defined as the Boltzmann distribution:

$$p_{\theta}(x,y)= \frac{\exp\{NN_{\theta}(x)[y]\}}{\sum_{x,y}\exp\{NN_{\theta}(x)[y]\}} \tag{2.4}$$

we define the partition function as $Z_{\theta} = \sum_{x,y}\exp\{NN_{\theta}(x)[y]\}$. [2]

The marginal distribution and conditional distribution can then be calculated easily. Here, we can calculate the conditional distribution pθ (y|x). We know that pθ (x, y) = pθ (y|x) pθ (x) thus:

$$p_{\theta}(y|x) = \frac{p_{\theta}(x,y)}{p_{\theta}(x)} = \frac{\exp\{NN_{\theta}(x)[y]\}}{\sum_{y}\exp\{NN_{\theta}(x)[y]\}} \tag{2.5}$$

The the equition (2.6) indicates that the energy-based model could be used either as a classifier or as a marginal distribution. In fact the EBMs has two functionality. They can be used either as a classifier or as a generator.

Next, we write the logarithm of the joint distribution to calculate the loss function as following:

$$\ln p_{\theta}(y,x) = \ln p_{\theta}(y|x) + \ln p_{\theta}(x) \tag{2.6}$$

$$= \ln \text{softmax}\{ NN_{\theta}(x)[y]\} + (\text{LogSumExp}_{x}\{NN_{\theta}(x)[y]\} - \ln Z_{\theta}) \tag{2.7}$$

It is evident that the model needs a common neural network that computes both distributions. We choose the final activation function to get a certain distribution.

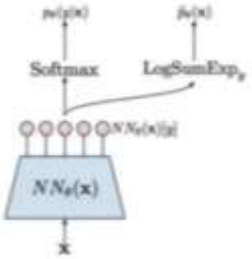


Fig 2.1 - The energy-based model could be used either as a classifier or as a marginal distribution.[2]

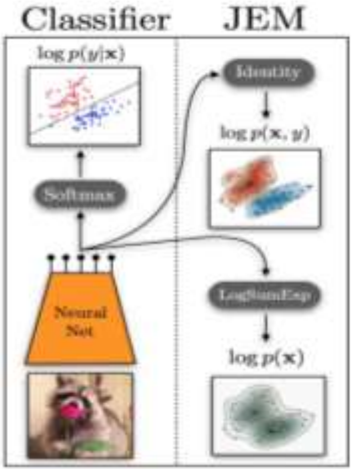


Fig 2.2 – Visualization of EBMs function as a classifier and generator [4]

2.1. Pros and cons of EBMs:

As mentioned before EBMs offer a promising alternative to address many challenges in both classification and generation area. One of the benefits of EBMs is that the energy function is unconstrained which means it could be any function such as a NN. The energy function can be multimodal without being defined as such (opposing to a mixture distribution). Also, there is no difference if we define it over discrete or continues variables.

Despite the benefits of EBMs, these models may cause some drawbacks as well. One of the negative points of EBMs is that the partition function which is the key element in EBMs, is problematic in learning energy-based models, since energy functions do not result in a nicely computable partition function.

In general, it is hard to sample from such models. Because, we know the probability for each point but there is not generative process like ARMs, flows, or VAEs. It is vague what is the graphical model for an EBM. Another obstacle of EBMs is that they do not distinguish variables in any way (like a black box that receivs an X as input and I gives an value as output.

3. Energy-based Out-of-Distribution Detection

Out-of-distribution detection is a binary classification issue that uses a score to distinguish between cases that are in- and out-of-distribution. A scoring function should generate numbers that can be differentiated between in- and out-of-distribution. In other words, Out-of-distribution (OOD) detection is a binary classification problem in which the model must generate a score that the scores for in-distribution examples are higher than that out-of-distribution examples [4].

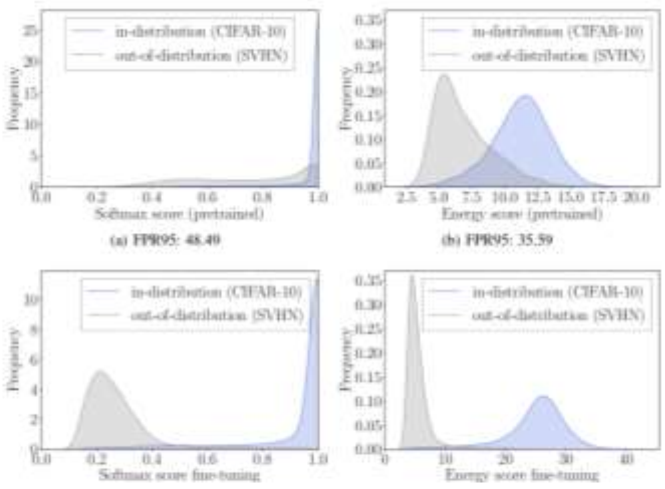
Traditional OOD detection techniques rely on measures of uncertainty, such as softmax to estimate the model's confidence. Energy-based models offer a promising alternative to address these challenges. Energy-based OOD maps each input to a single scalar that is lower for observed data (In-Distribution) and higher for unobserved ones (Out-of-Distribution). In contrast to traditional methods, energy-based approaches do not require explicit probabilistic assumptions and can better handle complex data distributions.[1]

Mathematically speaking, we use the EBM density function (eq. 2.2) for model discrimination. The eq. 2.2 is linearly aligned with the log likelihood function, which is desirable for OOD detection. Samples with higher energies (lower likelihood) are considered as OOD inputs. We define G(x; τ, f) as following:

$$\mathrm{G}(\mathrm{x};\tau,\mathrm{f}) = \begin{cases} 0 & if - E(x;f) \leq \tau \\ 1 & if - E(x;f) \geq \tau \end{cases} \tag{3.1}$$

We pick the threshold based on in-distribution data so that the OOD detector G(x) accurately classifies a large proportion of inputs.

Figure 3.1 shows how the in-distribution and out-of-distribution are classified energy-based OOD compared to softmax. The threshold τ can be choosed based on the following plots.



(a) FPR95: 48.40

(b) FPR95: 35.59

Fig 3.1 - Distribution of softmax scores vs. energy scores from pre-trained WideResNet [1]

To find the optimal diffrentiation of in- and out-of-distribution data, an energ-bounded learning objective is defined to fine-tune the network.

$$\underset{\theta}{min} E_{(x,y)\sim} D_{in}^{train} [\text{-log}\mathrm{F}_{\mathrm{y}}(\mathrm{x})] + \lambda.\mathrm{L}_{\text{energy}} \tag{3.2}$$

It assigns low energy values to in-distribution data and higher energy values to OOD training data, resulting in data that is more distinct between in- and out-of-distribution. The regularisation term consists of two terms that penalise in-distribution for being more than a particular margin and penalise out-of-distribution for being less than the margin (with an auxiliary dataset).

$$\mathrm{L}_{\text{energy}} = E_{(x_{in},y)\sim D_{in}^{train}}(\max(0,\mathrm{E}(\mathrm{x}_{\text{in}}) - \mathrm{m}_{\text{in}}))^2 + E_{x_{out}\sim D_{out}^{train}}(\max(0,\mathrm{m}_{\text{out}} - \mathrm{E}(\mathrm{x}_{\text{out}})))^2 \tag{3.3}$$

4. Results

4.1. EBM as a Classifier

We train our energy-based OOD model on CIFAR10 and CIFAR100, and tested on SVHN and texure. We used 80 Million Tiny Images dataset as auxilary outlier dataset. Table 4.1 indicates the classification of data using energy-based OOD compared to softmax.

D_{in}	Fine-tune	OOD Dataset	FPR95	AUROC	AUPR
			Softmax[2],[5] / Energy Score (ours)		
WideResNet CIFAR10	×	Texure	59.28/52.43	88.50/85.21	97.16/95.40
		SVHN	48.49/34.90	91.89/91.14	98.27/97.71
WideResNet CIFAR10	✓	Texure	12.94/**6.08**	97.73/96.72	99.52/98.92
		SVHN	4.36/**1.21**	98.63/**99.14**	99.74/**99.43**

Table 4.1 - Comparison of OOD detection performance utilising softmax vs. energy-based

To evaluate our solution, we use two evaluation metrics. we employ the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR). The ROC curve is a graph that compares the true positive rate ($t_{pr} = t_p/(t_p + f_n)$) with the false positive rate ($f_{pr} = f_p/(f_p + t_n)$). Furthermore, the AUROC may be viewed as the likelihood that a positive example will have a higher detector score/value than a negative example. While with AUPR, the PR curve plots the precision (tp/(tp+fp)) and recall (tp/(tp + fn)) against each other. The baseline detector has an AUPR approximately equal to the precision.

It is worth noting that comparing detectors is not as simple as using accuracy. We have two classes for detection, and the detector returns a score for both the positive and negative classes. If the negative class is significantly more probable than the positive class, a model may always anticipate the negative class and achieve high accuracy, which can be misleading. We must next define a score threshold to ensure that some positive cases are correctly identified, but this is determined by the trade-off between false negatives (fn) and false positives (fp)[3].

4.2. EBM as a Generator

As shown in eq.2.9, EBM can act as a generator since the second term in the eq.2.9 is a marginal distribution. By utilizing EBMs neural nets we were able to synthesize digits after 70 epoches as shown in fig 4.2.

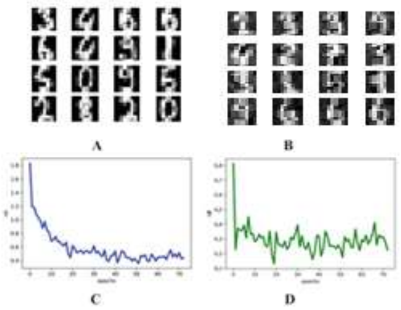


Fig 4.2 - (a) Randomly selected real images. (b) Unconditional generations from the EBM

5. Discussion and Future work

This article presents a novel approach to out-of-distribution (OOD) detection using energy scores instead of traditional softmax confidence scores. This approach`s strengths is solving the problem of being bias in softmax. Since EBMs use neural network to calculate the enegry instead of a constant maximum value in softmax, it is more dianamic and therefore more accurate than softmax[1]. Alao, existing techniques based on pre-trained models may require numerous hyperparameters to be tweaked and, in certain cases, new data. In comparison, the energy score is a parameter-free measure that is simple to use and implement and delivers equivalent or even higher performance in many circumstances[1].

One weakness of the article is that the partition function in the energy distribution fomula is difficult to compute[2].

Potential areas for future research include exploring the use of energy-based OOD detection in other machine learning tasks beyond image classification, such as natural language processing or speech recognition. Additionally, further investigation into the impact of hyperparameter tuning on the method's performance could help optimize the framework for specific applications. Another potential avenue for future research is the development of methods for online or incremental learning with energy-based OOD detection, which could be useful in scenarios where new data is constantly being generated.

Overall, the article presents a promising approach to OOD detection using energy scores, with potential applications in a variety of fields. However, further research is needed to fully understand the theoretical foundations of energy-based learning and to optimize the method for specific applications.

Reference

[1] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 1802, 21464–21475.

[2] Jakub M. Tomczak, Deep Generative Modeling.

[3] Dan Hendrycks and Kevin Gimpel 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136.

[4] Will Grathwohl, Kuan-Chieh Wang & Jorn-Henrik Jacobsen 2020. Your classifier is secretly an energy based model and you should treat it like one.

[5] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich 2019. Deep anomaly detection with outlier exposure. In International Conference on Learning Representations.