Energy Based Out-Of-Distribution Detection
CS726 Project Report

Aziz Shameem, 20d070020
Shikhar Mundra 200050131
Shivam Patel, 200070077
Vedang Asgaonkar, 200050154

May 2023

Task Description

The Problem

With the advent of large machine learning models for image and text recognition, classification, generation etc., it is imperative to understand their boundaries with respect to the data inputs that they can correctly work on. Determining the 'out-of-distribution' (OOD) samples is crucial for ascertaining the accuracy and validity of machine learning models for any input data stream. Many machine learning models are trained on typical data which might not be representative of real life usage data. In such cases, it is essential that the model does not take deleterious decisions which can cause harm. For example, we might want an autonomous driving system to flag an alert or raise and alarm, seeking human intervention when it encounters unexpected situations. This not only makes machine learning models safer, but also increases their utility and usability. These concerns gave rise to the OOD detection literature in the machine learning diaspora, which detect distribution shift, semantic outliers and anomalies.

Background

Energy based models (EBMs) utilize non-probabilistic scalar 'energy' score for training and testing of machine learning models. The mapping of the input to a scalar output through the energy function $E(x) : \mathbb{R}^D \rightarrow \mathbb{R}$ can be converted to a probability measure by using the Gibbs distribution :

$$p(y|x) = \frac{exp(-E(x,y)/T)}{\int_{y'} exp(-E(x,y')/T)} = \frac{exp(-E(x,y)/T)}{Z}$$

Here, Z represents the partition function which acts as the normalising factor in the Gibbs distribution. T is the temperature parameter of the Gibbs distribution. The partition function marginalises over the set of outputs y' in the continuous case, and sums over outputs in the discrete space framework. Our implementation of energy function does not necessarily require a generative model, and extends to purely discriminative models as well, both for training and evaluation purposes.

Energy Function: For delineating the energy function for EBMs, we first define *Helmholtz Free Energy* $E(x)$ of any datapoint $x$as

$$E(x) = -T \cdot log \int_{y'} e^{-E(x,y')/T}$$

Now let us consider any discriminator neural network $f(x) : R^D \to R^K$, which takes $x \in R^D$ as input and outputs softmax scores for k output classes. This can be represented by the logit scoring

$$p(y|x) = \frac{exp(f_y(x)/T)}{\Sigma_i^K exp(f_i(x)/T)}$$

Where $f_i(x)$ is the score for i^{th} class. Connecting the Gibbs Distribution and Helmholt Free Energy, we can express the free energy function as

$$E(x; f) = -T \cdot log \Sigma_i^k e^{f_i(x)/T}$$

LSTM

Long Short-Term Memory(LSTM) Networks is a deep learning, sequential neural network that allows information to persist. It is a special type of Recurrent Neural Network which is capable of handling the vanishing gradient problem faced by RNN. LSTM was designed by Hochreiter and Schmidhuber that resolves the problem caused by traditional rnns and machine learning algorithms.

This type of an architecture is used for text/NLP tasks, due to its ability to learn inter-dependencies between arrays of inputs. It outperforms other architectures in its ability to understand context in the inputs.

Outline of Method

Energy based OOD detection is compatible the softmax trained discriminative neural networks, with the newer probability density function being

$$p(x) = \frac{exp(-E(x;f)/T)}{\int_x exp(-E(x;f)/T)}$$

Taking the logarithm gives us the log likelihood maximization problem -

$$log\, p(x) = -E(x; f)/T - log\, Z$$

Where the logarithm of partition function $log\, Z$ is independent of any input x. This directly implies the linear alignment of $-E(x; f)$ with the log likelihood function. Thus, implying a threshold on $-E(x; f)$ gives us an OOD detection setup, with higher values of energy implying out of distribution samples.

Comparison of Softmax Score and Energy Score

Energy score based training mathematically turns out to be an bias compensated softmax classifier. For a softmax confidence score (for the general case of $T = 1$):

$$\begin{aligned}
\max_y p(y|x) &= \max_y \frac{exp(f_y(x)}{\Sigma_i exp(f_i(x))} \\
&= \frac{exp(f^{max}(x))}{\Sigma_i exp(f_i(x))} \\
&= \frac{1}{\Sigma_i exp(f_i(x) - f^{max}(x))} \\
\implies log \max_y p(y|x) &= E(x; f(x) - f^{max}(x)) = E(x; f) + f^{max}(x)
\end{aligned}$$

Hence, in the softmax confidence score, the log likelihood function is shifted by the maximum logit function value. For OOD samples, $f^{max}(x)$ is lower and $E(x; f)$ is lower, which destroys the linearity between the probability density and the energy function. Consequently, the softmax score is not well aligned with $p(x)$ unlike the energy score. Thus, actual energy score values are more reliable and informative than the softmax score for each input sample.

Bounded Energy Learning

So far we looked at adopting energy based OOD detection on an already trained model. Now we try to include energy based OOD detection paradigms during training of model itself, which optimizes the gap between the out and in-distribution data for and effective and more distinguishable OOD detection. This results in a greater contrast in the energy surface for the in and out-distribution samples. Our newer training objective for the energy-based classifier is

$$\min_\theta E_{(x,y)\sim \mathcal{D}_{in}^{train}}[-logF_y(x)] + \lambda \cdot L_{energy}$$

where $\mathcal{D}_{in}^{train}$ is the in-distribution training data and $F(x)$ is the softmax output of the classification model. Thus, the newer objective blends the cross-entropy loss with the regularization loss for maximising energy difference between in and out of distribution samples. The regularizer is

$$\begin{aligned}
L_{energy} = &E_{(x_{in},y)\sim \mathcal{D}_{in}^{train}}(\max(0, E(x_{in}) - m_{in}))^2 \\
&+ E_{x_{out}\sim \mathcal{D}_{out}^{train}}(\max(0, m_{out} - E(x_{out})))^2
\end{aligned}$$

where $\mathcal{D}_{out}^{train}$ is the additional out of distribution training data. Here we regularized the the energy with squared hinge loss using the margin hyperparameters m_{in} and m_{out}. Once the model is fine tuned, we can proceed with normal OOD detection using energy score as described previously. In our implementation, we have replaced this regularizer with a ranking loss on the energies of ID and OOD samples, to create a margin between them. This improves the FPR metric of the model, while maintaining comparable scores on other metrics.

$$L_{energy} = \sum_{x_{in}} \sum_{x_{out}} ReLU(m + E(x_{in}) - E(x_{out}))$$

Summary of Implementation

In this project, we build upon energy based scoring. Our main contributions are summarised as follows :

- We reproduced the results stated in the paper, using energy based scoring for OOD detection
- We replaced the margin loss implemented with **ranking loss**, as described in the previous section.
- We built an **LSTM** based model, with the discussed margin losses (ranking and energy), and tried it on *20news* dataset(text data with 20 classes), with the first 10 classes taken as **In-Distribution**, and the rest as **Out-Of-Distributions**. The results are summarised below.

Although the obtained results are not great, we observe an improvement in the error when ranking loss is used, as opposed to energy-based scoring (as implemented by the source).

Experiment Details and Main Results

We summarize the results of our experiments with the CIFAR-10 dataset:

Table 1: Texture dataset as OOD

	Ranking(our)	Energy Ft(fine tuned by us)	Energy baseline
FPR95	0.24	0.34	0.52
AUROC	99.81	99.81	85.27
AUPR	99.96	99.96	95.38

Related Work

OOD detection is relatively nascent field which emerged out of the pressing need for ascertaining validity and reliability of outputs for any input data to the model. Listed below are some findings and methologies related to energy-based OOD detection.

Table 2: Speckle Noised images as OOD

	Ranking(our)	Energy Ft(fine tuned by us)	Energy baseline
FPR95	0.31	0.24	0.66
AUROC	99.91	99.93	89.37
AUPR	99.97	99.97	97.96

Table 3: LSTMs for OOD on 20NewsGroup dataset

Number of Epochs	Method	Test Error
1	Energy	89.78
1	Ranking	89.70

- '**On Out-of-distribution Detection with Energy-based Models**' [1] hypothesizes that EBMs fail to learn semantic features for OOD detection. Supervisional and architectural restrictions are imposed and improved OOD detection is observed, ascertaining the hypothesis.

- '**Unsupervised Energy-based Out-of-distribution Detection using Stiefel-Restricted Kernel Machine**' [2] uses the Stiefel-Restricted Kernel Machine (St-RKM) instead of the softmax scoring that we focused on so far. Newer energy function discriptions are also delineated based on the RKM framework.

- '**Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem**' [3] provides an insight into the the failure of ReLU based networks in OOD detection and how an adversarial training like technique overcomes the problem.

- '**Semantic Driven Energy based Out-of-Distribution Detection**' [4] uses representation score (for respective class labels) along with energy score as a training objective for OOD detection.

Conclusions

Work split up amongst the team members

- **Vedang, Aziz**: Training and fine-tuning models, literature review and solution roadmap, exploring and setting up codebase, setting up the dataloader and implementation and training of LSTM model for text data.

- **Shivam, Shikhar**: Formulating solution roadmap, drafting report, literature review

References

[1] S. Elflein, B. Charpentier, D. Zügner, and S. Günnemann, "On out-of-distribution detection with energy-based models," *CoRR*, vol. abs/2107.08785, 2021.

[2] F. Tonin, A. Pandey, P. Patrinos, and J. A. K. Suykens, "Unsupervised energy-based out-of-distribution detection using stiefel-restricted kernel machine," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2021.

[3] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," *CoRR*, vol. abs/1812.05720, 2018.

[4] A. Joshi, S. Chalasani, and K. N. Iyer, "Semantic driven energy based out-of-distribution detection," in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 01–08, 2022.

[5] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *CoRR*, vol. abs/2110.11334, 2021.

[6] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, 2020.

[7] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," *CoRR*, vol. abs/2111.05826, 2021.