POST-HOC OUT-OF-DISTRIBUTION DETECTION

Harshit Varma
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai, India
harshitvarma@cse.iitb.ac.in

Aaron Jerry Ninan & Eeshaan Jain & Ipsit Mantri
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
{190100001,19D070022,180070032}@iitb.ac.in

NOTATION

Acronyms

ID	In-Distribution
OOD	Out-Of-Distribution
DNN	Deep Neural Network

1 PROBLEM STATEMENT

We focus on OOD detection, in a classification setting, after a classifier has already been trained (i.e., a *post-hoc/post-training* setting). Many popular baselines propose score functions to detect OOD data. These score functions should assign low scores to OOD data and high scores to ID data. We aim to provide better scoring functions that are effective, computationally cheap, generalize across different OOD settings and have a theoretical backing. In cases when the performance is not good enough, we aim to provide ways to further improve the separability between the ID and OOD data without using a large and diverse auxiliary OOD dataset (like 80 Million Tiny Images and ImageNet-22K) for fine-tuning, unlike many other popular approaches (Hendrycks et al., 2018; Liu et al., 2020).

2 RELATED WORK

LeCun et al. (2006) is a tutorial on general energy-based approaches.
Yang et al. (2021) is a survey paper on OOD detection.
Hendrycks & Gimpel (2016) propose a simple and commonly used baseline for OOD detection based on the maximum (over classes) softmax probability.
Hendrycks et al. (2018) propose the use of an auxiliary OOD dataset to enforce the softmax probabilities for the OOD data to be close to a uniform distribution over the classes.
Lee et al. (2018) fit class-conditional multivariate Gaussian distributions (with shared covariance matrix) on the penultimate layer output of a DNN softmax classifier via maximum likelihood, and then use the Mahalanobis distance as a score.
Wang et al. (2021) propose a a Wasserstein distance based score.
Our base paper is Liu et al. (2020), which interprets softmax classifiers as energy-based models and proposes a simple new score based on this which outperforms many popular baselines and can also be theoretically shown to be better than Hendrycks & Gimpel (2016). It also provides a way to further improve the performance using by fine-tuning on an auxiliary dataset using the proposed score.

3 DATASETS AND CODE

Datasets: (considered till now)

- MNIST
- Fashion MNIST
- CIFAR-10
- MNIST-35689: MNIST with classes 3, 5, 6, 8, and 9 (as subset of the entire MNIST). Similarly MNIST-01247.

We plan to use more datasets, like the Describable Textures Dataset (DTD), notMNIST, etc.

Reference Code:

- `https://github.com/wetliu/energy_ood`
- `https://github.com/tayden/ood-metrics`

Libraries: Our code is written in PyTorch and uses standard libraries like NumPy, Matplotlib, etc. Most experiments were run locally. Google Colab was used for fine-tuning experiments and training models which required more compute.

4 PROPOSED APPROACH

We provide a simple and novel (to the best of our knowledge) scoring function based on the Dirichlet distribution. We assume a Dirichlet distribution over the softmax probability outputs generated by the the classifier and estimate the distribution's concentration parameters via maximum likelihood; using the softmax probabilities outputted by the classifier after training as the dataset for maximum likelihood estimation. This is computationally cheap as only a forward pass through the classifier is required and the estimation converges quickly (≈ 5 epochs) in practice. We also provide theoretical reasoning and show the superiority of our method in comparison to the method by Liu et al. (2020) by analyzing the asymptotic behaviour of our approach. **Please have a look at the appendix for all the details.** Empirically as well, our method consistently outperforms Liu et al. (2020)'s method[1] across multiple datasets and metrics.

Our method also naturally leads to a loss function for enforcing further separation of the ID and the OOD data by the model. We are yet to test this.

From the results (e.g. MNIST-35689 vs MNIST-01247 compared to MNIST-35689 vs CIFAR-10), we also observe that the all scores perform poorly when the ID and the OOD data share low-level features (like edges, low-order image statistics, etc). We aim to enforce further separation between the ID and the OOD data particularly in this setting. We propose a way that doesn't require the usage of another large and diverse auxiliary dataset unlike Hendrycks et al. (2018); Liu et al. (2020). We can achieve this by creating auxiliary OOD data by augmenting the ID data itself. Particularly, we plan to use elastic distortions and random patching (patching the image and shuffling the patches randomly) as data augmentation strategies. Both of these methods preserve low-level image features but remove the higher level structural information. We think this will help improve the performance by pushing the model to give more importance to high level features like shape and structure.

5 WORK DONE AND REMAINING WORK

Experiments using various scores across the datasets using multiple metrics are done. Results are available on this link. Notation used: m: negative of the minimum logit, M: maximum logit, E: energy-score as proposed by Liu et al. (2020), S: softmax-score as proposed by Hendrycks & Gimpel (2016), D: dirichlet-score (our method). Density plots of various scores on different datasets are available on this link. Please use IITB LDAP for viewing all results.

[1] We skip the comparison of our methods with other methods like Lee et al. (2018) for brevity since Liu et al.'s method outperforms them.

Fine-tuning experiments using different kinds of losses, scores, augmentations are remaining. We plan to focus on this from now on. In the end if time permits, we also plan to extend the experiments to include adversarial examples (including degraded images) and analyze the results.

REFERENCES

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Yinan Wang, Wenbo Sun, Jionghua Jin, Zhenyu Kong, Xiaowei Yue, et al. Wood: Wasserstein-based out-of-distribution detection. *arXiv preprint arXiv:2112.06384*, 2021.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

APPENDIX

More details about the proposed approach and background material are provided here.

1 INTRODUCTION

Below sections closely follow Liu et al. (2020) and LeCun et al. (2006).

1.1 DISCRIMINATIVE MODELS

Let $\mathcal{X}$ denote the input space. Let K be the number of classes. Let $\mathcal{Y} = \{\text{oh}(k)\}_{k=1}^{K}$ be the output space, where $\text{oh}(k)$ is the one-hot encoding of k. We only consider deep neural networks (DNNs) as discriminative models. Let $F(x;\theta_F) : \mathcal{X} \to \mathbb{R}^K$ denote a DNN with parameters θ_F that maps an input to un-normalized logits. To get $p(y|x)$ we (usually) pass these through a softmax function:

$$p(y|x) = \frac{\exp\left(\langle\, y, F(x;\theta^F)\rangle\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\langle\, y', F(x;\theta^F)\rangle\right)}$$

1.2 ENERGY-BASED MODELS (EBMS)

Energy-based approaches aim to build a function $E(x, y; \theta^E) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, parametrized by θ^E, that maps an input to a single scalar called the "energy", which can be seen as measure of compatibility between x and y. In a discriminative setting, we can define $p(y|x)$ via the Gibbs distribution.

$$\begin{aligned} p(y|x) &= \frac{\exp\left(-E(x, y; \theta^E)/T\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(-E(x, y'; \theta^E)/T\right)} \\ &= \frac{\exp\left(-E(x, y; \theta^E)/T\right)}{\exp\left(-E(x; \theta^E)/T\right)} \\ E(x;\theta^E) &= -T \log \left(\sum_{y \in \mathcal{Y}} \exp\left(-E(x, y; \theta^E)/T\right) \right) \end{aligned}$$

Where $E(x;\theta^E)$ is the Helmholtz Free Energy.

1.3 DISCRIMINATIVE MODELS AS EBMS

Discriminative models can thus be thought of as an EBM implicitly parametrized by θ^F

$$\begin{aligned} E(x, y; \theta^E) = E(x, y; \theta^F) &= -T \langle\, y, F(x;\theta^F)\rangle \\ E(x;\theta^E) = E(x;\theta^F) &= -T \log \left(\sum_{y \in \mathcal{Y}} \exp\left(\langle\, y, F(x;\theta^F)\rangle\right) \right) \end{aligned}$$

1.4 OOD DETECTION

OOD Detection can be thought of as a binary classification problem which aims to use a scoring function to assign high scores to ID data and low scores to OOD data.

A simple candidate for the scoring function is the maximum of the predicted softmax probabilities as explored by Hendrycks & Gimpel (2016).

$$\begin{aligned} \text{softmax\_score}(x) &= \max_{y \in \mathcal{Y}} p(y|x) \\ &= \max_{y \in \mathcal{Y}} \left(\frac{\exp\left(\langle\, y, F(x;\theta^F)\rangle\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\langle\, y', F(x;\theta^F)\rangle\right)} \right) \\ &= \frac{\max_{y \in \mathcal{Y}} \left(\exp\left(\langle\, y, F(x;\theta^F)\rangle\right) \right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\langle\, y', F(x;\theta^F)\rangle\right)} \end{aligned}$$

Based on this, we define the following two scores (the motivation for defining these will be clear soon)

$$\text{max\_logit\_score}(x) = \max_{y \in \mathcal{Y}} \left(\langle y, F(x; \theta^F) \rangle \right)$$
$$\text{avg\_logit\_score}(x) = -\frac{1}{K} \sum_{y \in \mathcal{Y}} \langle y, F(x; \theta^F) \rangle$$

Note that the average of the logits for a "good" classifier is expected to be < 0 for ID data as roughly all but one logits are expected to be < 0. This behaviour is observed in practice as well. Thus the score is defined to be negative of the average.

Liu et al. (2020) use $-E(x; \theta^F)$ with $T = 1$ as the scoring function. We shall also assume $T = 1$ throughout. If NLL is used as the loss function, it can be shown that the energy $E(x; \theta^F)$ will be minimized for the ID data.

$$\begin{aligned} \text{energy\_score}(x) &= -E(x; \theta^F) \\ &= \log \left(\sum_{y \in \mathcal{Y}} \exp \left(\langle y, F(x; \theta^F) \rangle \right) \right) \end{aligned}$$

The two scores are related as follows

$$\begin{aligned} \log \text{softmax\_score}(x) &= \log \max_{y \in \mathcal{Y}} p(y|x) \\ &= \log \max_{y \in \mathcal{Y}} \exp \left(\langle y, F(x; \theta^F) \rangle \right) - \log \left(\sum_{y' \in \mathcal{Y}} \exp \left(\langle y', F(x; \theta^F) \rangle \right) \right) \end{aligned}$$
$$\log \text{softmax\_score}(x) = \text{max\_logit\_score}(x) - \text{energy\_score}(x)$$

Thus, the softmax_score is composed of a difference two scores (i.e., both effectively act in opposite directions). Due to this, the softmax_score cannot be reliably used for OOD detection.

2 Asymptotic Analysis of the Energy Score

Let $l_k = \langle y_k, F(x; \theta^F) \rangle$ (the k^{th} logit)
Let $M = \text{max\_logit\_score}(x) = \max_{y \in \mathcal{Y}} \left(\langle y, F(x; \theta^F) \rangle \right)$ and let this be achieved at the m^{th} logit.

$$\begin{aligned} \text{energy\_score}(x) &= \log \left(\sum_{y \in \mathcal{Y}} \exp \left(\langle y, F(x; \theta^F) \rangle \right) \right) \\ &= \log \left(\sum_{k=1}^{K} \exp \left(l_k \right) \right) \\ &= \log \left(\exp \left(M \right) \cdot \sum_{k=1}^{K} \exp \left(l_k - M \right) \right) \\ &= M + \log \left(1 + \sum_{k \neq m} \exp \left(l_k - M \right) \right) \end{aligned}$$

For ID data, and for a good classifier, the second term in the $\log$ is expected to be $\ll 1$.
Thus, $\text{energy\_score}(x) \approx \text{max\_logit\_score}(x)$. This is also observed in practice as shown in the results section, with the energy_score performing only marginally better than the max_logit_score. This is also seen in the softmax_score values, which are concentrated heavily near 1 for the ID data, implying $\log \text{softmax\_score}(x) = \text{max\_logit\_score}(x) - \text{energy\_score}(x) \approx 0$.

3 DIRICHLET-BASED OOD DETECTION

The Dirichlet PDF paramterized by the concentration parameters $\alpha \in \mathbb{R}_+^K$ is given by

$$p(s|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K s_k^{\alpha_k - 1} \text{ for } s \in \mathcal{S}_{K-1}$$

It's a natural choice for a distribution over values $\in \mathcal{S}_{K-1}$, the open standard $K-1$ simplex.

We assume a Dirichlet distribution over the softmax-ed logits of the DNN and estimate α via maximum likelihood and then use $\log p(s|\hat{\alpha})$ as the OOD detection score.

$$\begin{aligned}
D &= \{s^{(i)} = \text{softmax}(F(x^{(i)}; \hat{\theta}^F))\}_{i=1}^N \\
\text{NLL}(\alpha) &= \sum_{i=1}^N \left(\sum_k \log\Gamma(\alpha_k) - \log\Gamma\left(\sum_k \alpha_k\right) - \sum_k \left((\alpha_k - 1)\log s_k^{(i)}\right)\right) \\
&= N\sum_k \log\Gamma(\alpha_k) - N\log\Gamma\left(\sum_k \alpha_k\right) - \sum_k \left((\alpha_k - 1)\sum_i \log s_k^{(i)}\right)
\end{aligned}$$

We get $\hat{\alpha} = \arg\min_{\alpha>0} \text{NLL}(\alpha)$ via gradient descent. Using the Adam optimizer, we converge after a few epochs.

After estimating α, the Dirichlet score is defined as follows and can be used for OOD detection

$$\text{dirichlet\_score}(x) = -\sum_k \left((\alpha_k - 1)\sum_i \log s_k^{(i)}\right)$$

4 ASYMPTOTIC ANALYSIS OF THE DIRICHLET SCORE

For a good classifier $F(x; \hat{\theta}^F)$ we are expected to have $\alpha_k \approx \alpha_0 \; \forall \; k \in \{1, \ldots, K\}$ with $\alpha_0 \ll 1$. This corresponds to a Dirichlet distribution having the density concentrated at the corners of the simplex $\mathcal{S}_{K-1}$.
Thus, we analyse the behaviour of $\log p(s|\alpha)$ when $\alpha_k = \alpha_0 \; \forall \; k \; \alpha_0 \to 0^+$

$$\begin{aligned}
\lim_{\alpha_0 \to 0^+} \log p(s|\alpha) &= \lim_{\alpha_0 \to 0^+} \left(\log\Gamma(K\alpha_0) - \sum_k \log\Gamma(\alpha_0)\right) - \sum_k \log s_k \\
&= \lim_{\alpha_0 \to 0^+} (\log\Gamma(K\alpha_0) - K\log\Gamma(\alpha_0)) - \sum_k \langle\, y_k, F(x; \hat{\theta}^F)\rangle + \sum_k \log\left(\sum_{y' \in \mathcal{Y}} \exp\langle\, y', F(x; \theta^F)\rangle\right) \\
&= \lim_{\alpha_0 \to 0^+} (\log\Gamma(K\alpha_0) - K\log\Gamma(\alpha_0)) - K\left(\frac{1}{K}\sum_k \langle\, y_k, F(x; \hat{\theta}^F)\rangle + E(x; \hat{\theta}^F)\right) \\
&= \lim_{\alpha_0 \to 0^+} (\log\Gamma(K\alpha_0) - K\log\Gamma(\alpha_0)) + K\,(\text{energy_score}(x) + \text{avg_logit_score}(x)) \\
&\propto K\,(\text{energy_score}(x) + \text{avg_logit_score}(x))
\end{aligned}$$

Thus, the asymptotic behaviour of the proposed Dirichlet score acts as an ensemble of two different score functions. This behaviour can be reason behind the consistent improvements over the individual scores as observed in the results.

5 FINE-TUNING WITH THE DIRICHLET SCORE

The NLL loss defined leads to a natural auxiliary loss function which can be used to fine-tune the model when auxiliary OOD data is available. For this we can keep α's fixed to the values obtained

after fitting to the ID data. The below loss aims to calibrate the softmax probabilities of the ID data towards the learnt probability distribution and the OOD data anywhere away from it. X_{in}, X_{out} are batches of ID and OOD data respectively. $t_k^{(j)}$ is the softmax probability of the k^{th} class for the j^{th} sample in the OOD batch. $s_k^{(i)}$ defined in a similar way for the ID batch.

$$\begin{aligned} L_{ft}(X_{in}, X_{out}) &= \sum_k \left((\alpha_k - 1) \sum_i \log t_k^{(i)} \right) - \sum_k \left((\alpha_k - 1) \sum_i \log s_k^{(i)} \right) \\ &= \sum_k (\alpha_k - 1) \left(\sum_j \log t_k^{(j)} - \sum_i \log s_k^{(i)} \right) \end{aligned}$$

The below loss can then be used for fine-tuning

$$L(X_{in}, Y_{in}, X_{out}) = L_{ce}(X_{in}, Y_{in}) + \lambda L_{ft}(X_{in}, X_{out})$$

6 Fine-tuning with the Energy Score

The objective described in Liu et al. (2020) is

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[-\log\langle y, F(x;\theta_F)\rangle] + \lambda L_{\text{energy}}$$

where

$$L_{\text{energy}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}\big(\max(0, E(x;\theta_F) - m_{\text{in}})\big)^2 + \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{out}}^{\text{train}}}\big(\max(0, m_{\text{out}} - E(x;\theta_F))\big)^2$$

$\mathcal{D}_{\text{in}}^{\text{train}}$ is the in-distribution training data and $\mathcal{D}_{\text{out}}^{\text{train}}$ is the unlabeled auxiliary out-of-distribution training data. The main issue with this formulation is that there are two margin hyperparameters m_{in} and m_{out} that need to be tuned carefully. Our plan moving forward is to also improve on this and devise a single margin loss function, which does not harm the performance, as the authors of Liu et al. (2020) claimed that dual-margin loss functions performed better empirically.

7 Experimental Setup

We use the VGG16 (with batch normalization) as our model for all the experiments. Inputs are normalized channel-wise on a per-image basis.