Critique Report on the Paper: "Multi-Class Data Description for Out-of-distribution Detection [2]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Predictions given by deep learning classifiers are often considered unreliable in practical settings as even input samples that the network was not trained on are classified with high confidence [6, 5]. It had been noted in this paper that the reason behind this unreliable prediction scores was the nature of the softmax function typically used for finding class conditional probabilities in deep learning models. As the softmax function finds linear decision boundary between different classes in the latent space, there is significant overlap between in-distribution (ID) and out-of-distribution (OOD) samples. Hence it this paper, an alternative to the softmax based classifiers was proposed that found spherical decision boundary for the classes in the latent space instead of linear.

Methods Proposed

The following objective function called Deep-MCDD was proposed in the paper for training the classifier instead of the traditional cross entropy loss.

$$\min_{W,\mu,\sigma,b} \frac{1}{N}\sum_{i=1}^{N}[D_{y_i}(x_i) - \frac{1}{\nu}log\frac{exp(D_{y_i}(x_i)+b_{y_i})}{\sum_{k=1}^{K}exp(D_k(x_i)+b_k)}] \tag{1}$$

where,

$$\begin{aligned} D_k(x) &= -logP(x|y=k) \\ &= -logN(f(x;W)|\mu_k, \sigma_k^2 I) \\ &\approx \frac{||f(x;W)-\mu_k||^2}{2{\sigma_k}^2} + log{\sigma_k}^d \end{aligned} \tag{2}$$

The first part of the proposed objective function, as could be intuitively understood from the equations, bound samples from each class into an isotropic Gaussian distribution in the latent space, hence leading to spherical distribution and the second part maximized the separation between spheres representing different classes and thus aiding classification. These properties of the proposed objective function was established using theoretical formulation of the objective using Gaussian Discriminant Analysis (GDA) also. After training the classifier using this proposed Deep-MCDD objective, during inference, ID samples were classified and OOD samples were detected using confidence scores defined using measure of class conditional distance obtained using equation 2 written above.

Strength of the Work

- Novel approach for OOD detection as well as ID classification proposed that formulated classification using Deep Neural Network (DNN) into a GDA problem, without using any OOD sample for fine-tuning that could make the detector biased towards those samples. Also the proposed approach was backed with strong theoretical justifications.
- Unlike some other related works [3], the proposed approach did not make any random assumption. Every detail was taken into account and rigorously proved and established. For example, it was justified that the

proposed objective function conformed with the assumption of the class conditional Gaussian distribution being isotropic. Also the plotted visualizations qualitatively established that the obtained distributions for the classes were indeed spherical and hence isotropic, unlike linear boundaries visualized in case of softmax function.

- An interesting side observation was that the proposed approach sometimes resulted in better ID accuracy compared to softmax classifier.
- A Deep-SVDD based objective function was also proposed initially that would have required two-phase and hence slightly complicated training. The proposed Deep-MCDD objective overcame that limitation.
- Details like applying non-negative constraint on $log\sigma_k$, so that the proposed distance metric in section 3.1 satisfied triangular inequality, were handled in the paper.

Limitations of the Work

- The proposed approach for OOD detection would not work on top of any readily available off-the-shelf softmax classifiers trained with cross-entropy loss function. It would require training the network from scratch. OOD detectors that can be externally incorporated into off-the-shelf pretrained models are valuable from the perspective of ready deployment, something that the proposed approach did not offer.
- The parameter d that was used in the equations written in Sections 3.1 and 3.2 was never defined anywhere in the paper. I had to google the GitHub repository for the paper that contained the source codes [1] to understand what this parameter stood for.
- Ablation studies regarding the effect of the regularization parameter ν on the image datasets were reported only for the tabular datasets, not for the image datasets.
- OOD detection results were only reported for 85% TPR, whereas in the literature its typically reported for 95% TPR. Because rejecting as high as 15% of the ID samples as OOD samples is not desirable from practical standpoint. Results for 95% TPR should have been reported to validate that the proposed approach is valuable for deploying in real-life applications.
- Technically, evaluation of the proposed approach on image datasets was not compared with any of the State-of-the-Art OOD detectors. It was argued that since the proposed approach did not fine-tune using any OOD sample, it was not compared with the results reported by detectors that required fine-tuning using OOD samples. Comparison was done only with a simplified version of the approach proposed by [3], that worked without fine-tuning. Even if this argument is accepted, there are other OOD detectors proposed in literature that did not involve fine-tuning either as this is by no means the first work that got rid of fine-tuning. Results on some of those approaches were even reported in the paper on tabular datasets.

So the experimental results on the image datasets should have been compared with those approaches. To me, approaches like ODIN [4] perform too good to be ignored even if they require fine-tuning. Hence works that don't fine-tune also compare the results with ODIN. This should have been done in this work also.

- The proposed approach detected input samples as OOD at the very last layer of the classifier. Hence the approach was inefficient in terms of throughput in case of batch inference.

Questions Unanswered

- It was not clear that how the value of the confidence score described in section 3.3, to distinguish between ID and OOD samples, was decided without using any OOD sample as had been claimed. In other literature these confidence scores were fine-tuned using some OOD samples. It should have been described how this was done in this paper without fine-tuning.

- In Figure-6, as ν was increased, it was expected that TNR should have increased. It increased upto $\nu = 10$, after that there was a sharp drop. The reason for this drop was not explained in the paper.

Suggested Future Studies

- The proposed objective function was very theoretically solid. An OOD detector could be trained using feature maps obtained from pretrained models on this objective function for fast and readily implementable OOD detection.

- Instead of two-term objective function proposed in the paper, it would be interesting to see if a single term objective function could be formulated for faster and easier convergence, that performs similar optimization.

References

[1] GitHub repository for Deep-MCDD Paper. https://github.com/donalee/DeepMCDD. Accessed: 10-05-2020.

[2] Dongha Lee, Sehun Yu, and Hwanjo Yu. Multi-class data description for out-of-distribution detection. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1362–1370, 2020.

[3] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[4] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[5] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.