OUT-OF-DISTRIBUTION AS A TARGET CLASS IN SEMI-SUPERVISED LEARNING

Antoine Tadros *Sébastien Drouyer* *Rafael Grompone von Gioi*

Centre Borelli, ENS Paris-Saclay, CNRS, Université Paris-Saclay, France

ABSTRACT

A key limitation of supervised learning is the ability to handle data from unknown distributions. Often, such methods fail when presented with samples from a source not represented in the training data. This work proposes an effective way of controlling the behavior of a neural network in the presence of out-of-distribution examples. For this, the training dataset is supplemented with extraneous data assigned to an additional out-of-distribution class. The extraneous data may come from a different dataset or be even noise. By applying a Gaussian mixture model on the latent representation, and by taking advantage of the ability of these models to generalize well, the method described thereafter performs well. Training the model on a segregated dataset helps the model to distinguish out-of-distribution data, including the ones the model were never confronted to during training.

Index Terms— out-of-distribution, neural network, latent space, Gaussian mixture, classification

1. INTRODUCTION

Neural networks produced a radical change of the paradigm in several fields, including image processing and computer vision [1]. Neural network models are now routinely used in critical activities such as medical image analysis [2], finances [3] or even agriculture [4]. Nevertheless, a drawback that came with this family of methods is the lack of information about the certainty that the model has on its prediction. Recently, several methods have been proposed to overcome this limitation.

In [5], the authors propose to use dropout to approximate a Bayesian neural network, which allows to compute an uncertainty on the prediction made by the model. Similar approaches have been presented with ensemble methods [6]. Other approaches consist in evaluating directly the lack of information, or "evidence", using the Dempster-Shafer theory [7] and the subjective logic [8] derived from the latter. [9] proposed to train the neural network to produce a set of evidence, that allows to compute a measure of uncertainty based on subjective logic theory. Instead of computing uncertainty, an alternative approach consists in directly detecting out-of-distribution samples, i.e. samples that seems to have been drawn from a different distribution than the one which generated the training set that was used to train the model [10].

This work proposes a method to detect out-of-distribution samples without computing uncertainties. It consists in training the neural network on the training dataset supplemented with extraneous data assigned to an additional out-of-distribution class. While neural networks for classification widely use the softmax activation to predict a class, the method proposed performs instead a distance-based prediction. The Mahalanobis distance between the input and the classes representative vectors is computed in the feature space delivered by the penultimate layer. The extraneous data may be composed of samples from a different dataset or even noise. We show that while only confronted to a single set of external data, the neural network manages to generalize well and to eliminate most of out-of-distribution samples, regardless of their origin.

This work is organised as follows. Section 2 describes briefly the main state-of-the-art methods, one based on the Dempster-Shafer theory [9] and an other based on Gaussian mixture models [10]. Then, Section 3 introduces the proposed method. An experimental comparison and discussion of these methods is presented in Section 4 where the model is evaluated on different MNIST-like datasets. Finally, Section 5 concludes the paper.

2. RELATED WORK

2.1. Subjective logic in neural network

The Dempster-Shafer Theory [7] was proposed as a generalization of the Bayesian theory of subjective probability. This theory is at the base of subjective logic, in which each event is associated with a belief mass. In the case of a classification problem with C classes, we note $b_c \in [0, 1]$ the belief mass of the class $c \in \{1, \dots, C\}$. An uncertainty measure $u \in [0, 1]$ is defined such that

$$\sum_{c=1}^{C} b_c = 1 - u. \tag{1}$$

Then, the subjective logic theory defines the evidence e_c linked to b_c as $b_c = \frac{e_c}{A}$, with $A = \sum_{c=1}^{C} (e_c + 1)$. It follows

Work partly financed by Office of Naval Research grant N00014-17-1-2552 and N00014-20-S-B001, DGA Astrid project "Filmer la Terre" n° ANR-17-ASTR-0013-01 and Kayrros Inc.

that $u = \frac{C}{A}$.

In [9], the authors consider the evidence e_c as a quantity that indicates how much a sample is likely to be classified as c. In the subjective logic framework, instead of proposing a definitive probability distribution, it is used as a way to develop a distribution over the possible probability distribution on the C classes. Let $\alpha_c = e_c + 1$. Using $(\alpha_c)_{c=1\dots C}$ as the parameters of a Dirichlet distribution over the C-dimensional unit simplex, the authors provide a new way to characterize the probability distribution over the C classes. To adapt the subjective logic theory for neural networks, the classification layer is considered to output an estimate of the evidence e_c for every class c, and the mean of the Dirichlet distribution provides an estimate of the probability distribution as $\hat{p}_c = \frac{a_c}{A}$, for $c = 1, \dots, C$.

2.2. Gaussian mixture in the latent space

The method described in the previous section requires to retrain the full neural network from scratch in order to provide the uncertainty measure. Instead, Lee et al. [10] proposed to estimate the uncertainty by building a logistic regression (LR) on top of the neural network. The method is proposed, among other applications, to detect out-of-distribution samples.

By assuming that the distribution of each class is Gaussian in the intermediate feature spaces of a trained neural network, a Gaussian mixture model (GMM) is estimated on several layers using the maximum a posteriori (MAP) on a validation set. For a layer l, the method proposes to assign the same global covariance matrix to the Gaussian of every class. Therefore, the GMM takes the form:

$$\mathcal{GMM}_l(x) = \sum_{c=1}^{C} \mathcal{N}\left(f_l(x), \hat{\mu}_{l,c}, \hat{\Sigma}_l\right), \tag{2}$$

where $f_l(x)$ is the output of the l^{th} layer, and $\hat{\mu}_{l,c}, \hat{\Sigma}_l$ are the MAP estimate of the Gaussians' parameters.

To classify a sample x, a set of L layers is selected and a GMM is fitted for each one of them. The minimum of the Mahalanobis distance to each Gaussian is then selected as the estimate provided by the layer,

$$\hat{c}_l = \arg\min_c \left(f_l(x) - \hat{\mu}_{l,c}\right)^T \hat{\Sigma}_l^{-1} \left(f_l(x) - \hat{\mu}_{l,c}\right). \tag{3}$$

Noise is then added to the sample based on the value of $\hat{c}_l$ to see how well the x lies in the Gaussian of the class $\hat{c}_l$. This creates a new sample $\hat{x}$. The Mahalanobis distance is computed again with this $\hat{x}$ to provide a confidence score M_l

$$M_l = \min_c \left(f_l(\hat{x}) - \hat{\mu}_{l,c}\right)^T \hat{\Sigma}_l^{-1} \left(f_l(\hat{x}) - \hat{\mu}_{l,c}\right). \tag{4}$$

The scores M_l of several layers are then combined to perform a logistic regression trained on a validation set made of the same amount of in-distribution and out-distribution samples. This is an important difference with the previous method, since it explicitly uses the out-of distribution samples to train another classifier on top of the neural network.

3. PROPOSED APPROACH

In [10], the authors use out-of-distribution samples to train a classifier. We propose to use out-of-distribution sample only during the training steps of the neural-network itself. Let A be the in-distribution training set and B the out-of-distribution one. We train the neural network on A and on only a small amount of data of B. The overall neural network, as well as the training process, remains the same. The only difference between this approach and the training of usual neural-network-based clasifier, is that the data from the out-of-distribution (OOD) dataset are considered as a class in itself, the "OOD class". Therefore, the output vector of the neural network is of size $C + 1$. The assumption is that by showing to the neural network a small amount of different data, it should be able to differentiate samples similar to the ones in the training set A, on which it has been specialized, from the one that are too different.

The method described in [10] assumed Gaussian mixture distributions on the data of several layers. Fig. 1 shows that this is a fair supposition for the last layer. Nevertheless, the first layers of a neural network does not necessary have this property as their feature space is less likely to be refined enough by the learning process to display such a feature.

Notice in Fig. 1 that "Letters" is the only class that does not form a proper separated ellipsoid. This motivates furthermore the need for out-of-distribution samples in the training phase to fine-tune the embedding.

To this end, only the GMM estimated on the last layer will be considered,

$$\mathcal{GMM}_l(x) = \sum_{c=1}^{C} \mathcal{N}\left(f_l(x), \hat{\mu}_{l,c}, \hat{\Sigma}_{l,c}\right). \tag{5}$$

Here, the estimate of the covariance matrix is class-specific ($\hat{\Sigma}_{l,c}$) while in [10] the covariance matrix is assumed to be the same for all Gaussians ($\hat{\Sigma}_l$) and is computed on the whole dataset in the feature space. To counter-weight this simplification, [10] trained a logistic regression classifier to perform the OOD, with the added perturbation to the input to control the robustness of the prediction. Our approach allows us to simplify the overall process while improving the overall classification performances.

To decide which class is the most appropriate, the closest one in the sense of the Mahalanobis distance will be selected. The Gaussian distributions are fitted with the same training set A as the rest of the parameters of the neural network, using the MAP estimator.

4. EXPERIMENTS

The proposed method will be compared to the approaches described in Section 2. We will refer to the Subjective Logic Learning method as SLL [9] and to the approach based

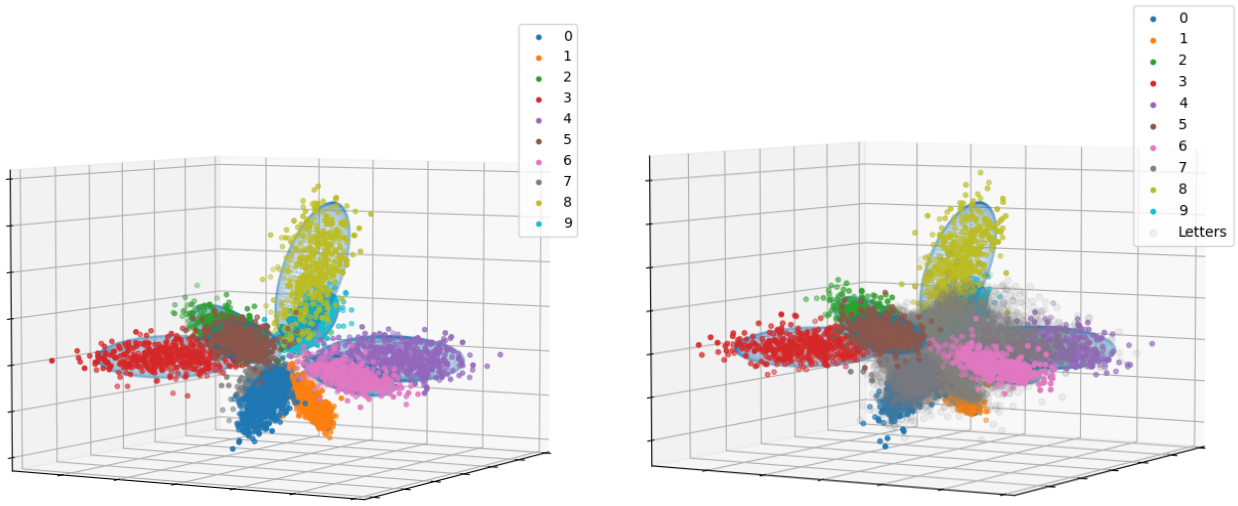

Fig. 1. MNIST representation on the feature space of the last layer of a vanilla LeNet architecture. A 3D representation is used for better visualisation without the need for dimension reduction. The ellipsoid shape of each class indicates that a Gaussian model is appropriate. The letters are data points from the EMNIST dataset.

on Gaussian Mixture Model with Logistic Regression as GMM+LR [10]. The three models have been trained on MNIST [11] as the main dataset. The out-of-distribution datasets used are EMNIST [12], KMNIST [13], Fashion-MNIST [14] and Uniform Noise.

A five-layers LeNet architecture [15] was trained on 1500 epochs for each method with a batch-size of 128. For GMM+LR and the proposed approach, the models were trained on one of the out-of-distribution dataset and evaluated on all of them to validate that the method can generalize well. Recall that SLL does not use OOD during training.

During the experiments, our model used without any Gaussian modeling performed better than used with the Mahalanobis distance. The result in Table 1 and Table 2 use the classic one. Using the Mahalanobis distance gave results closer to what can be found with GMM+LR. For our model, the out-of-distribution samples made 10% of the training set, which is, in proportion, less than the 50% used to train the logistic regression of [10].

From Table 1 and Table 2, it appears, on the one hand, that the two supervised methods performs the best in out-of-distribution detection. On the other hand, SLL [9], outperform the other models when it comes to make predictions on in-distribution samples.

It can also be noticed that our method, in its simpler setup, performs as well as [10], which needs an extra supervised classifier in order to work. In both cases, it is interesting to see that by training the model on one out-of-distribution dataset, the model is able to also reject samples that are neither from the in-distribution nor the out-of-distribution dataset that it encounters during its training.

	M	E	F	K	N
SLL [9]	**0.981**	0.923	0.954	0.938	0.906
GMM + LR (E) [10]	0.936	0.936	0.998	0.904	0.999
GMM + LR (F) [10]	0.775	0.894	**0.999**	0.918	0.999
GMM + LR (K) [10]	0.883	0.930	**0.999**	0.927	0.999
GMM + LR (N) [10]	0.893	0.820	0.983	0.866	0.999
Proposed method (E)	0.885	**0.999**	0.993	0.976	0.999
Proposed method (F)	0.885	0.918	**0.999**	0.935	0.999
Proposed method (K)	0.885	0.923	**0.999**	**0.999**	0.999
Proposed method (N)	0.885	0.909	0.910	0.909	**1.00**

Table 1. Accuracy in the out-of-distribution detection. M, E, F, K, N respectively stands for MNIST, EMNIST, Fashion-MNIST, KMNIST and Noise. The accuracy in column M corresponds to the in-distribution accuracy. When a method is followed by a letter in parenthesis, such as (E), it means that the model was trained using the dataset (E) as an out-of-distribution set. The SLL method as no "E","F","N" or "K" line since the method is fully unsupervised. Thus, SLL needs no training on the OOD datasets but its ability to detect OOD samples can still be done by using the uncertainty estimation from Eq. 1.

It is worth highlighting that for both the proposed method and the one designed in [10], using the noise as the out-of-distribution set, while not as good as using a proper dataset, helps the model to accomplish the task of eliminating samples that are not from the in-distribution training set. This

	E	F	K	N
SLL [9]	0.897	0.978	0.954	0.930
GMM + LR (E) [10]	0.936	**0.999**	0.911	**1.00**
GMM + LR (F) [10]	0.864	**0.999**	0.900	**1.00**
GMM + LR (K) [10]	0.921	**0.999**	0.916	**1.00**
GMM + LR (N) [10]	0.756	0.969	0.810	**1.00**
Proposed method (E)	**0.999**	0.993	0.956	0.999
Proposed method (F)	0.835	**0.999**	0.894	**1.00**
Proposed method (K)	0.807	**0.999**	**0.999**	0.999
Proposed method (N)	0.660	0.827	0.642	**1.00**

Table 2. AU-ROC score in the out-of-distribution detection. E, F, K, N respectively stands for EMNIST, Fashion-MNIST, KMNIST and Noise.

makes the method unsupervised for the classification of unknown samples.

5. CONCLUSION

This work proposed a simple approach for handling out-of-distribution samples in semi-supervised learning: the training dataset is supplemented with extraneous data assigned to an additional out-of-distribution class and the Mahalanobis distance in the penultimate feature space is used as a prediction criterion. An advantage of this approach is that it does not require any major change in the model configuration. Our experiments show the proposed approach perform similarly or better than state-of-the-art methods. Interestingly, using noise as the extraneous data, while not optimal, achieves good results in most cases. Further work will aim at using few-shot learning techniques to optimize the learning process.

6. REFERENCES

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] Alexander Selvikvåg Lundervold and Arvid Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019, Special Issue: Deep Learning in Medical Physics.

[3] Ankit Thakkar and Kinjal Chaudhari, "A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions," *Expert Systems with Applications*, vol. 177, pp. 114800, 2021.

[4] Lingjia Gu, Fachuan He, and Shuting Yang, "Crop classification based on deep learning in northeast china using sar and optical imagery," in *2019 SAR in Big Data Era (BIGSARDATA)*, 2019, pp. 1–4.

[5] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[6] Tim Pearce, Felix Leibfried, and Alexandra Brintrup, "Uncertainty in neural networks: Approximately bayesian ensembling," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Silvia Chiappa and Roberto Calandra, Eds. 26–28 Aug 2020, vol. 108 of *Proceedings of Machine Learning Research*, pp. 234–244, PMLR.

[7] Glenn Shafer, *A mathematical theory of evidence*, Princeton university press, 1976.

[8] Audun Jøsang, "Subjective logic: A formalism for reasoning under uncertainty," 2016.

[9] Murat Sensoy, Lance Kaplan, and Melih Kandemir, "Evidential deep learning to quantify classification uncertainty," *arXiv preprint arXiv:1806.01768*, 2018.

[10] 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*, vol. 31, 2018.

[11] Yann LeCun and Corinna Cortes, "MNIST handwritten digit database," 2010.

[12] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik, "Emnist: an extension of mnist to handwritten letters," *arXiv preprint arXiv:1702.05373*, 2017.

[13] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha, "Deep learning for classical japanese literature," 2018, cite arxiv:1812.01718Comment: To appear at Neural Information Processing Systems 2018 Workshop on Machine Learning for Creativity and Design.

[14] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017, cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.