Improving Deep Network Robustness to Unknown Inputs with Objectosphere

Akshay Raj Dhamija, Manuel Günther & Terrance E. Boult
Vision and Security Technology Lab, University of Colorado Colorado Springs
{adhamija | mgunther | tboult} @ vast.uccs.edu

Abstract

Deep Neural Networks trained on academic datasets often fail when applied to the real world. These failures generally arise from unknown inputs that are not of interest to the system. The mis-classification of these unknown inputs as one of the known classes highlights the need for more robust deep networks. The problem of identifying samples that are not of interest to the system has previously been tackled by either thresholding softmax, which by construction cannot return none of the known classes *itself, or by learning new features for the unknown inputs using an additional background or garbage class. As demonstrated, both of these approaches help but are generally insufficient when previously unseen classes are encountered. This paper overviews our recent publication Reducing Network Agnostophobia, NeurIPS 2018. The paper presented two novel loss functions that effectively handle unseen classes while providing a new measure for uncertainty. The ability to identify unknown samples plays a crucial role in developing robust networks that may be used in open-world problems. The paper also introduced an evaluation metric that focused on comparing performance of multiple approaches in an open-set setting.*

1. Introduction

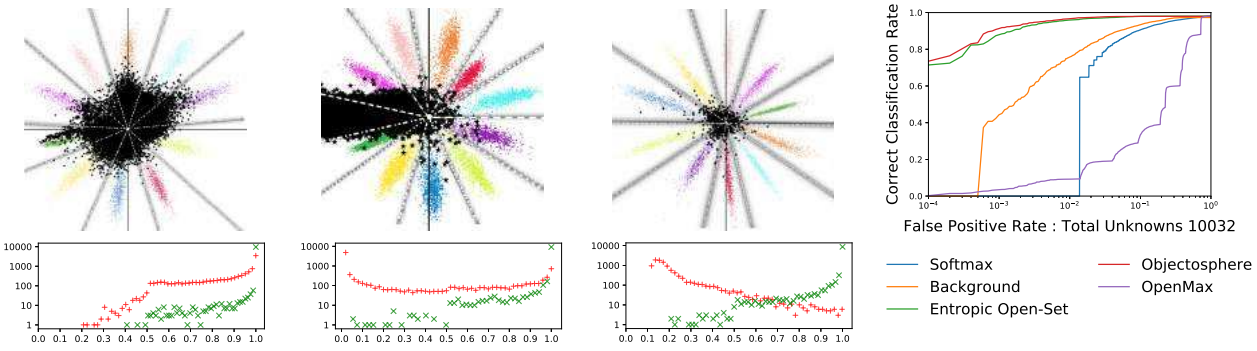
When deployed, detection systems are usually exposed to many more instances of classes that are not of interest than the classes of interest. These detection systems include object detectors [5, 4, 15, 10, 14], face detectors [7], pedestrian detectors [18], etc. Interestingly, though each year new state-of-the-art-algorithms emerge from each of these domains, a crucial component of their architecture remains unchanged – handling unwanted or unknown inputs. The majority of detection approaches can be divided into two modules, a localization module and a classification module. While the localization module provides the bounding box for a detection, the classification module decides what object has been detected. Almost all state-of-the-art object detectors use deep networks for these modules. Most commonly, during training the classification network includes a background class to identify a region as not having an object of interest. In [2], we provide a visual understanding of why the background class approach is insufficient when handling unknown samples and propose novel loss functions that constitute a viable substitution.

In order to better understand the problem, let us assume $\mathcal{Y} \subset \mathbb{N}$ be the infinite label space of all classes. This label space contains the *known classes of interest* ($\mathcal{C}$) and the *unknown classes* ($\mathcal{U} = \mathcal{Y} \setminus \mathcal{C}$). While the *known classes of interest* are the classes the network is trained to identify, the *unknown classes* are the classes the network is supposed to reject as none of the known classes. Since $\mathcal{Y}$ is infinite and $\mathcal{C}$ is finite, $\mathcal{U}$ is also infinite. The set $\mathcal{U}$ is further divided as:

1. $\mathcal{B} \subset \mathcal{U}$: The *background*, *garbage*, or *known unknown* classes. Since $\mathcal{U}$ is infinitely large, during training only a small subset $\mathcal{B}$ can be used.
2. $\mathcal{A} = \mathcal{U} \setminus \mathcal{B} = \mathcal{Y} \setminus (\mathcal{C} \cup \mathcal{B})$: The *unknown unknown* classes, which represent the rest of the infinite space for $\mathcal{U}$. Samples of these classes are not available during training, but only occur at test time.

Let the samples that belong to $\mathcal{B}$ and were seen during training be depicted as $\mathcal{D}'_b$ and the ones seen during testing depicted as $\mathcal{D}_b$. Similarly, the samples seen during testing belonging to $\mathcal{A}$ are represented as $\mathcal{D}_a$. The samples belonging to the known classes of interest $\mathcal{C}$, seen during training and testing are represented as $\mathcal{D}'_c$ and $\mathcal{D}_c$, respectively. Finally, the unknown test samples are $\mathcal{D}_u = \mathcal{D}_b \cup \mathcal{D}_a$.

Rather than rejecting unknown samples $x \in \mathcal{D}_u$, the two novel loss functions develop deep feature representations that are more robust to unknown inputs. When training the models with background samples $\mathcal{D}'_b$, in contrast to common approaches, our loss functions do not rely on an additional softmax output for the background class. Instead, they yield networks where *thresholding softmax scores* is more effective at rejecting unknown samples $x \in \mathcal{D}_u$ than using a dedicated background class. The novel uncertainty measure combines properties of the learned deep features and the softmax scores to provide robust networks. Inspired from real-world requirements of detectors, we also propose a new evaluation metric for comparing performances of different approaches under the presence of unknown samples.



(a) Softmax (b) Background (c) Objectosphere (d) Open-Set Classification Rate Curve

Figure 1: 2D LeNet++ Responses To Known and Unknown Samples. *The network in (a) was only trained to classify the 10 MNIST classes ($\mathcal{D}'_c$) using softmax, while the networks in (b) and (c) added NIST letters [6] as known unknown samples ($\mathcal{D}'_b$) trained with an additional background class or the novel Objectosphere loss, respectively. In the feature representation plots on top, colored dots represent test samples from the ten MNIST classes ($\mathcal{D}_c$), while black dots represent samples from the previously unseen Devanagari [11] dataset ($\mathcal{D}_a$), and the dashed gray-white lines indicate class borders where softmax scores for neighboring classes are equal. [2] addresses how to improve recognition by reducing the overlap of network features from known samples $\mathcal{D}_c$ with features from unknown samples $\mathcal{D}_u$. The figures in the bottom are histograms of softmax probability values for samples of $\mathcal{D}_c$ and $\mathcal{D}_a$ with a logarithmic vertical axis. For known samples $\mathcal{D}_c$ the probability of the correct class is used, while for samples of $\mathcal{D}_a$ the maximum probability of any known class is taken. In an application, a score threshold θ should be chosen to optimally separate unknown from known samples. Unfortunately, such a threshold is difficult to find for either (a) or (b), a better separation is achieved with the Objectosphere loss (c). The proposed Open-Set Classification Rate (OSCR) curve in (d) depicts the high accuracy of the Objectosphere approach even at a low false positive rate, where competitive algorithms such as OpenMax [1] fail.*

2. Approach

One of the limitations of training with a separate background class is that the features of all unknown samples are required to gather in one region of the feature space. This restriction is independent of the similarity that an input might have to one of the known classes. In Fig. 1(a), from the depiction of the test set of MNIST [8] and samples from the Devanagari dataset [11], we observe that magnitudes for unknown samples in deep feature space are often lower than those of known samples. This observation leads us to believe that the magnitude of the deep feature vector captures information about a sample being unknown. We exploit and exaggerate this property to develop a network where for $x \in \mathcal{D}'_b$ we reduce the deep feature magnitude ($\|F(x)\|$) at the network layer prior to the logit layer and, therewith, maximize entropy of the softmax scores in order to separate them from known samples. Unlike the background class approach, this approach allows the network to have unknown samples that share features with known classes as long as they have a small feature magnitude. It may also allow the network to focus learning capacity to respond to the known classes instead of spending effort in learning specific features for unknown samples. This in achieved in two stages.

2.1. Entropic Open-Set Loss

First, the *Entropic Open-Set loss* aims to maximize entropy of unknown samples by making their softmax response uniform across classes. This means, while we keep the softmax loss calculation untouched for samples of $\mathcal{D}'_c$, we modify it for training with the samples from $\mathcal{D}'_b$ seeking to equalize their logit values l_c, which will result in equal softmax scores S_c. The intuition here is that if an input is unknown, we know nothing about what classes it relates to or what features we want it to have and, hence, we want the maximum entropy distribution of uniform probabilities over the known classes. Let S_c be the softmax score as above, the Entropic Open-Set loss J_E is defined as:

$$J_E(x) = \begin{cases} -\log S_c(x) & \text{if } x \in \mathcal{D}'_c \text{ is from class } c \\ -\frac{1}{C}\sum_{c=1}^{C} \log S_c(x) & \text{if } x \in \mathcal{D}'_b \end{cases}$$

2.2. Objectosphere Loss

While the Entropic Open-Set loss produces a network that provides a higher softmax entropy for the unknown samples, there is often a modest overlap between the feature magnitudes of known and unknown samples. This should not be surprising as nothing is forcing known samples to have a large feature magnitude or always force features of unknown samples to be short. Thus, in the second stage we attempt to put a distance margin (ξ) between them. In particular, we seek to push known samples into what we call the *Objectosphere* where they have large feature magnitude and low entropy, i.e., we train the network to have a large response to only known classes $x \in \mathcal{D}'_c$ while penalizing

$\|F(x)\|$ for $x \in \mathcal{D}'_b$ to minimize feature length for unknown samples. Formally, the Objectosphere loss is calculated as:

$$J_R = J_E + \lambda \begin{cases} \max(\xi - \|F(x)\|, 0)^2 & \text{if } x \in \mathcal{D}'_c \\ \|F(x)\|^2 & \text{if } x \in \mathcal{D}'_b \end{cases}$$

The above formalization both penalizes samples of the known classes if their feature magnitude is inside the boundary of the Objectosphere, and samples of unknown classes if their magnitude is not vanishing. Note that larger ξ values will generally scale up deep features, including the unknown samples, but what matters is the overall separation.

Finally, instead of thresholding just on the final softmax score $S_c(x)$ of the network, one can exploit the fact that the known and unknown samples have different deep feature magnitudes which may be multiplied with the softmax scores to obtain the *Scaled Objectosphere* score: $S_c(x) \cdot \|F(x)\|$. Thresholding this product seems to be more reasonable and justifiable. Theorems and related proofs for the Entropic Open-Set and the Objectosphere loss may be found in [2].

3. Evaluating Open-Set Systems

An open-set system has a two-fold goal, it needs to correctly classify samples from known classes $\mathcal{D}_c$ while rejecting samples belonging to unknown classes $\mathcal{D}_u$. This makes evaluating open-set problems more complex. We briefly discuss a few important traits for such an evaluation. The system needs to be evaluated separately for unknwons $\mathcal{D}_u$ and knowns $\mathcal{D}_c$. Considering all unknown samples as a separate class and reporting accuracy on all the combined classes may not provide a clear indication of the system's performance in both closed and open-set conditions. Moreover, systems are generally deployed at a particular operating point, therefore rather than comparing algorithms with overall accuracy, performance needs to be evaluated at separate operating points. For such an evaluation, the choice of the operating point plays a key role. For example, softmax confidence scores cannot serve as operating points because they are incomparable across different systems. In order to provide a unified measure of performance, the Area Under the Curve (AUC) is an acceptable measure only as long as it is calculated for a monotonic curve. Sadly, the popularly reported precision-recall (PR) curve is non-monotonic and, hence, its AUC is not dependable. The shortcomings of various evaluation metric such as the AUC of the PR curve, Recall@K and the Accuracy vs. Confidence curve are discussed in [2].

To properly address the evaluation of an open-set system, we introduce the *Open-Set Classification Rate* (OSCR) curve as shown in Fig. 1(d), which is an adaptation of the *Detection and Identification Rate* (DIR) curve used in open-set face recognition [12]. For evaluation, we split the test samples into samples from known classes $\mathcal{D}_c$ and samples from unknown classes $\mathcal{D}_u$. Let θ be a score threshold. For samples from $\mathcal{D}_c$, we calculate the *Correct Classification Rate* (CCR) as the fraction of the samples where the correct class $\hat{c}$ has maximum probability, which needs to be greater than θ. We compute the *False Positive Rate* (FPR) as the fraction of samples from $\mathcal{D}_u$ that are classified as *any* known class $c \in \mathcal{C}$ with a probability greater than θ:

$$\mathrm{FPR}(\theta) = \frac{\left|\{x \in \mathcal{D}_a \mid \max_c P(c \mid x) \geq \theta\}\right|}{|\mathcal{D}_a|}$$

$$\mathrm{CCR}(\theta) = \frac{\left|\{x \in \mathcal{D}_c \mid \arg\max_c P(c \mid x) = \hat{c} \wedge P(\hat{c} \mid x) > \theta\}\right|}{|\mathcal{D}_c|}$$

Finally, for the OSCR curve we plot CCR versus FPR, varying the probability threshold θ values from largest to smallest from left to right. For the smallest θ, the CCR is identical to the closed-set classification accuracy on $\mathcal{D}_c$. OSCR is not prone to any of the issues discussed in [2] and, hence, is ideal for evaluating open-set systems.

4. Conclusion

Our approaches exploit the default response of a network to unknown samples and provides a better separation between $\mathcal{D}_u$ and $\mathcal{D}_c$. As supported by the evaluation in Fig. 1(d), this separation results in networks that are much more robust to $\mathcal{D}_u$. Further experiments on datasets such as CIFAR and architectures such as ResNet-18 demonstrate [2] that the approach may also be successfully applied to other datasets and higher dimensional feature spaces. While there was considerable prior work on open set, rejection, out-of-distribution detection, or uncertainty estimation, our work [2] presents the first significant and theoretically grounded steps to an improved network representation for addressing samples of unknown classes. Though traditionally softmax has been used as a uncertainty score, our ability to differentiate unknown ($\mathcal{D}_u$) from known samples ($\mathcal{D}_c$) based on the deep feature magnitude provides a better measure of uncertainty. The benefits of this new measure can be witnessed by the success of the *Scaled Objectosphere* approach as seen in the additional experiments [2].

The improved uncertainty estimation provides more robust deep networks that are not drastically affected in performance when ported from lab environments into the real world where the identification of unknown samples is of utmost importance. While there has been significant progress in zero-shot, one-shot, few-shot, and incremental learning [9, 17, 3, 13, 16], if systems incorrectly but confidently classify unknown samples as one of the known classes, there is no reason for a system to consider learning these samples as new classes. This means that open-world systems need to have the ability to know that they do not know. Systems that are able to identify unknown are more robust in the real world and also have the ability to learn novel classes using concepts such as incremental learning.

References

[1] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[2] Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 3

[3] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2018. 3

[4] Ross Girshick. Fast R-CNN. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015. 1

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. 1

[6] Patrick Grother and Kayee Hanaoka. NIST special database 19 handprinted forms and characters 2nd edition. Technical report, National Institute of Standards and Technology (NIST), 2016. 2

[7] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 1

[8] Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits, 1998. 2

[9] Xin Li, Yuhong Guo, and Dale Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015. 3

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*. Springer, 2016. 1

[11] Ashok Kumar Pant, Sanjeeb Prasad Panday, and Shashidhar Ram Joshi. Off-line Nepali handwritten character recognition using multilayer perceptron and radial basis function neural networks. In *Asian Himalayas International Conference on Internet (AH-ICI)*. IEEE, 2012. 2

[12] P. Jonathon Phillips, Patrick Grother, and Ross Micheals. *Handbook of Face Recognition*, chapter Evaluation Methods in Face Recognition. Springer, 2nd edition, 2011. 3

[13] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 3

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 1

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1

[16] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *International Conference on Computer Vision (ICCV)*. IEEE, 2017. 3

[17] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 3

[18] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 1