HEM: a margin-based loss for visual categorisation tasks

Michael W. Spratling[1] and Heiko H. Schütt[1]

Department of Behavioural and Cognitive Sciences, University of Luxembourg,
L-4366 Esch-sur-Alzette, Luxembourg
{michael.spratling,heiko.schutt}@uni.lu

Abstract. Training deep neural networks (DNNs) on classification tasks can be performed with a number of different losses, but cross-entropy (CE) loss is the de-facto standard. Here, we propose an alternative loss, high error margin (HEM), which is a margin based loss modified to improve the training dynamics of neural networks. HEM loss is evaluated extensively using a wide range of DNN architectures and benchmark datasets with all experimental settings and training hyper-parameters taken from the literature, and hence, optimised for CE loss. HEM is found to be more effective than CE loss across a range of image-based tasks: unknown class rejection, adversarial robustness, learning with imbalanced data, continual learning, and semantic segmentation (a pixel-wise classification task). HEM is inferior to CE only in terms of clean and corrupt image classification with balanced training data, and this difference is small. We also compare HEM to specialised losses that have previously been proposed to improve performance for specific vision tasks. Logit-Norm, a loss achieving state-of-the-art performance on unknown class rejection, produces similar performance to HEM for this task, but is much poorer for continual learning and semantic segmentation. Logit-adjusted loss, designed for imbalanced data, has superior results to HEM for that task, but performs worse on unknown class rejection and semantic segmentation. DICE, a popular loss for semantic segmentation, is inferior to HEM loss on all tasks, including semantic segmentation. Overall, HEM is competitive with the best alternative loss for all the tasks we have used and performs better than all other tested losses in terms of rejecting out-of-distribution examples, for continual learning, and by a substantial margin for semantic segmentation[1].

Keywords: deep learning · loss functions · robustness · generalisation · image classification · imbalanced data · continual learning · semantic segmentation

1 Introduction

Deep neural networks (DNNs) are generally trained using variants of stochastic gradient descent. These optimisers require the loss function to have a gradient.

[1] Code at https://codeberg.org/mwspratling/HEMLoss

This means that it is not possible to maximise the classification accuracy directly as this function is piece-wise constant, and therefore, does not define a usable gradient. As a result, it is necessary to use a *surrogate* loss function that has a gradient, but still encourages few classification errors. The design of the loss function is important as different losses will lead to different training speeds and cause convergence to different parameters. While many possible surrogate loss functions have been proposed [56, 59], cross-entropy (CE) loss is by far the most common choice for classification tasks.

In some specific domains better performance can be obtained by using other losses. For example, alternative losses and regularisation terms have been proposed to improve robustness to adversarial attack [3, 13, 29, 30, 37, 44, 46, 55, 63, 66]. In the domain of open-set recognition, where the aim is to better detect and reject images from "unknown" classes, alternative losses (such as contrastive losses) have been employed together with architectural modifications to improve performance beyond that achieved by CE [41, 69]. Alternatively, LogitNorm loss [60], can be employed to improve unknown class rejection without the need for modifications to the network architecture or training procedure. To deal with situations where the training data contains drastically different numbers of samples for different classes ("class imbalance"), state-of-the-art approaches use a logit-adjusted loss which weights minority classes more heavily [14, 39, 48]. For semantic segmentation, where the aim is to assign a class label to each image pixel, the two largest groups of losses centre around CE and its variants, and DICE and its variants [4, 36]. DICE loss [40] is designed to be particularly effective when there is class imbalance, a common situation in segmentation tasks. While these specialised losses can outperform CE on the specific tasks for which they were developed, they tend to perform poorly outside of their specialised domain, as is confirmed by our results which are summarised in Fig. 1. The fact that specialised losses out-perform CE on certain tasks motivates our search for a better classification loss function, that performs well on a range of tasks.

For simpler statistical models, margin based losses [11] are known to be general-purpose and show better generalisation behaviour than CE loss. We hypothesised that similar advantages could be achieved for DNNs by using a margin based loss. Particularly, we expected a margin-based loss to train networks that were less susceptible to making over-confident predictions, and hence, that would be better able to distinguish known from unknown classes. Furthermore, a margin-based loss should be less prone to over-write previously learnt weights, which could reduce catastrophic forgetting in continuous learning and over-writing weights necessary to classify minority classes when training with imbalanced data. Hence, a margin-based loss is a promising candidate for a general-purpose classification loss function. However, the existing multi-class margin-based loss (MM) results in performance that is frequently much worse than that achieved with CE loss (Fig. 1).

Here, we propose high error margin (HEM) loss, a new margin-based loss function that can be used as a general-purpose loss. To motivate our new loss we first describe issues with CE loss (Sec. 2.1) and MM loss (Sec. 2.2). Sec. 3

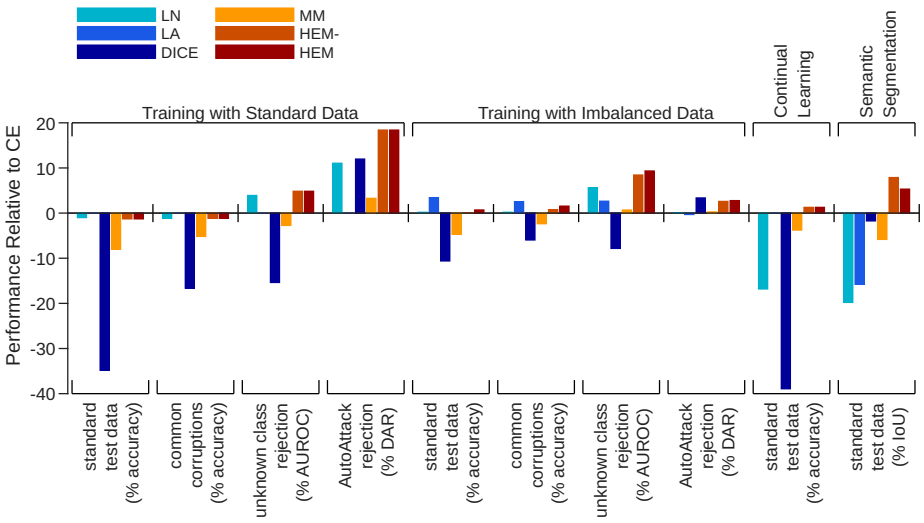

Fig. 1: Summary results for all the different tasks considered, comparing the average performance of cross-entropy (CE) loss to each of the alternative losses that have been evaluated: LogitNorm (LN), Logit-adjusted (LA), DICE, multi-class margin (MM), high error margin with shared margin (HEM-), and high error margin with adjusted margins (HEM). Results are averaged using the arithmetic mean over all other factors that were varied in the experiments. Specifically, for all experiments on "Training with Standard data" (the first four segments of the figure), each bar is an average of 71 experiments (5 data-sets x 3 network architectures x 5 repeats, except for one combination of data-set and architecture where only one trial was performed). For all experiments on "Training with Imbalance Data" (the fifth to eighth segments) each bar is an average of 75 experiments (5 data-sets x 3 network architectures x 5 repeats). For the experiments on continual learning (the ninth segment), each bar is an average of 80 experiments (4 data-sets x 4 continual learning techniques x 5 repeats). For experiments on semantic segmentation (the tenth segment), each bar is an average of 76 experiments (3 data-sets x 4 network architectures x 5 repeats + 1 data-set x 4 network architectures x 4 repeats). For all the evaluation metrics used, higher values indicate better performance. The relative performance is calculated by subtracting the performance produced by CE loss from the corresponding metric for each of the other losses. Hence, positive values indicate average performance better than that of CE loss. Note that the results for LA are equal to those of CE, and the results for HEM are equal to those of HEM- when the training data is balanced: *i.e.* when using standard training data (results in segments 1 to 4 of the figure) and when performing continual learning (segment 9).

Table 1: Example outputs (logits), $\mathbf{y}$, from the last layer of an imaginary neural network being trained to perform a 4-way classification task, and the corresponding losses associated with each of these predictions when the first output represents the correct class. Hyper-parameters for each loss were set to $\tau = 1$ for cross-entropy (CE) loss, $\tau = 0.04$ for LogitNorm (LN) loss, and $\mu = 0.5$ for multi-class margin (MM) and high error margin (HEM).

$\mathbf{y}$	CE	LN	MM	HEM
[1.0 −1.0 −1.0 −1.0]	0.34	0.00	0.00	0.00
[0.6 0.1 0.1 0.1]	1.04	0.00	0.00	0.00
[0.6 0.3 0.0 −0.1]	1.02	0.00	0.11	0.45
[0.6 0.5 0.0 −0.7]	1.00	0.09	0.16	0.63
[0.6 0.7 0.0 −3.5]	0.98	1.10	0.19	0.77
[0.0 1.0 0.0 0.0]	1.74	25.00	0.66	0.88

describes HEM loss which fixes the shortcomings of the existing losses that we have identified in our analysis in Sec. 2. Finally, we present extensive evaluations performed using nineteen different neural network architectures (ranging in size from LeNet to Vit-B/16) trained on many different data-sets (ranging in size from MNIST to ImageNet1k). We find that HEM is competitive with or outperforms CE loss, and several specialised losses, across a range of classification tasks (Sec. 4). Full details of CE and all the other existing loss functions that we consider are given in Appendix A.

2 Analysis and Motivation

2.1 Issues with CE Loss

Even when a classifier produces the correct classification with high confidence, the CE loss is far from zero (Tab. 1, row 1). Consequently the gradients will be non-zero and each presentation of a correctly classified training sample will cause the weights to be modified so that the outputs become ever more extreme (highly positive for the logit representing the correct class and highly negative for the logits representing the incorrect classes). CE loss therefore encourages a DNN to map every training exemplar to an output where confidence in the predicted classification is extremely high. It is to be expected that such a DNN will produce high confidence for all samples, including ones not seen during training. It is unsurprising, therefore, that CE-trained DNNs have issues using prediction confidence to distinguish known from unknown classes.

CE loss continuing to update the weights even when the correct classification has been learnt successfully, could cause weights to be over-written. This behaviour may underlie some of the issues when CE loss is used for continual learning and learning with imbalanced training data. For good performance in these tasks it is necessary to maintain weights that represent classes learnt earlier or that have few training samples. Even after the classifier has achieved

perfect performance, CE loss will update the weights which may cause further forgetting of the previously learnt classes or the minority classes. This intuition is consistent with the observation that models learn fewer features relevant to the minority classes than the majority classes [15].

Another issue with CE loss is that it can be lowered by reducing the logit for an already clearly rejected class without increasing the difference to the closest competitor class. As a result, CE loss can behave quite counter-intuitively: decreasing even though the prediction is becoming poorer (Tab. 1, second to third rows). Most concerningly, CE loss can even be lower for incorrect classification (Tab. 1, penultimate row), than for correct classification.

LogitNorm (LN), multi-class margin (MM) and the proposed high error margin (HEM) losses, do not suffer from these issues. These losses produce a loss of zero for samples where the output of the neuron representing the true class is sufficiently larger than the outputs of other neurons (Tab. 1 top rows) and non-zero losses only for predictions that are worse at distinguishing the true class from the alternatives (Tab. 1 bottom rows). In the case of MM and HEM, this is due to these losses using a margin. In the case of LN loss this is due to it behaving like a margin loss (Appendix A.2).

2.2 Issues with MM Loss

The analysis in Sec. 2.1 suggests that a margin-based loss could have advantages over CE loss. A margin loss, including our proposed High Error Margin (HEM) loss, defines an error, e_i, associated with each logit, y_i, as follows:

$$e_i = \begin{cases} \max(0, y_i - y_l + \mu_i) & \text{if } i \neq l \\ 0 & \text{if } i = l \end{cases} \tag{1}$$

where μ_i is the margin (a non-negative hyper-parameter) for logit i, and l is the index corresponding to the correct class (*i.e.* the ground-truth class label). The error is zero when the output from the neuron representing the correct class exceeds the outputs of the other logit by at least the margin.

MM loss, the existing margin-based loss (see Appendix A.5 for full details), produces classifiers that have significantly lower accuracy on the standard test data compared to equivalent CE-trained classifiers (as shown in Fig. 1). We suspect that this may be due to the method used to combine the errors. In MM loss this is achieved by averaging (or summing) the errors (across logits and samples in the batch). We believe this method of combining the error causes the gradient magnitude to change too much over the course of training. The loss is much higher at the start of training than near its end, because most errors become zero. This effect could either prevent the suppression of the last non-zero errors towards the end of training or lead to instability at the start. If our intuition is correct, we predict that the severity of this problem will increase for tasks with more classes.

3 High Error Margin Loss

We propose to solve the training issues of MM loss (Sec. 2.2) by combining the errors (Eq. (1)) in a more adaptive way to reduce the change in gradient magnitude during training. For each sample, all error values below the mean are set to zero, and the mean of above-zero values is calculated:

$$\mathcal{L}_{HEM} = \frac{\sum_{i=1}^{n} \mathbb{1}[e_i \geq \frac{1}{n}\sum e_j] \times e_i}{\sum_{i=1}^{n} \mathbb{1}[e_i \geq \frac{1}{n}\sum e_j]} \tag{2}$$

where n is the number of classes, and $\mathbb{1}[\cdot]$ is the indicator function, which equals 1 if the argument is true and 0 otherwise. We call this loss the high error margin (HEM) loss, as it takes the average only of high errors. Losses for different samples in the batch are also combined by finding the mean of above-zero values. Note that the computation of the mean error, used by the indicator function in Eq. (2), is detached from the computational graph so that it does not affect the gradients.

At the start of learning, when a large number of logits produce errors (especially when n is large), the mean error for each sample will be significant and, by only considering those losses above the mean, HEM concentrates on reducing the largest errors. Later in learning, there will be many zero errors and as a result, the mean error will be small (likely smaller than the few non-zero errors that remain). Hence, at this stage in training thresholding the errors by the mean will have little effect. However, by taking the mean of only the above-zero values, the loss will remain large even when there are few non-zero errors in each sample, and/or few incorrectly classified samples in a batch. As a result, HEM loss concentrates on the logits that produce the highest errors throughout learning. The effectiveness of the proposed method of combining errors, compared to that used in MM loss, was confirmed in an ablation study presented in Sec. 4.5.

We present results for two variants of the proposed loss:

High Error Margin with adjusted margins (HEM) which uses per-logit margins as defined in Eq. (1). Each margin was set to be inversely proportional to the number of training samples associated with that class. Specifically, $\mu_i = \sqrt{M/(ns_i)}$, and s_i is the number of samples in class i. The separate, class specific, margins help deal with class imbalance.

High Error Margin with shared margin (HEM-) which (like MM loss) sets all margins to the same value (*i.e.* $\mu_i = \mu \; \forall i$). The shared margin was made equal to $\sqrt{M/\sum_{i=1}^{n} s_i}$. HEM- is an ablated version of the proposed loss that we use to provide a fairer comparison with CE and MM losses which do not employ mechanisms to deal with class imbalance.

Both versions employ a single hyper-parameter, M. Based on preliminary experiments (see Appendix D.1) M was set to a value of 2000 for all other experiments described in this paper. Appendix D.1 demonstrates empirically that the results are not sensitive to the size of the margin, and also provides an intuitive argument for this lack of sensitivity to the choice of the hyper-parameter value. Note

that when the training data contains the same number of samples per class, all the margins used by HEM are equal, and HEM is identical to HEM-.

4 Results

We compared the performance of networks trained with our proposed loss, HEM, to the performance of identical networks trained with the alternative losses described in Appendix A. We have sought to produce a fair and representative evaluation of the different losses by using many different tasks, data-sets, and network architectures. Performance was tested using a number of different metrics to assess accuracy, generalisation, and robustness. For all metrics larger values correspond to better performance. Full details of the tasks, data-sets, evaluation metrics, DNN architectures and training set-ups are provided in Appendix B.

The tasks were chosen to include essential and important applications in computer vision (image classification and semantic segmentation), and tasks where we expected our margin loss to perform well, due to it not learning to make predictions with very high confidence (unknown class rejection) and stopping weight updates once adequate performance is achieved (continual learning and learning with imbalanced data).

For each experimental condition (combination of data-set, network architecture, and loss function) a network was trained and evaluated multiple times (typically five), each time with a different random weight initialisation and random presentation order of training samples. In the main text, summary results are presented by showing the average performance relative to CE loss. Detailed results showing absolute performance together with error-bars are reported in Appendix C.

Each experiment was performed using a single NVIDIA Tesla V100 GPU with 16GB of memory, except for experiments with ImageNet1k which were executed in parallel on four such GPUs. Performance differences can be interpreted independently of computational cost because the time taken to compute any of the losses is negligible compared to the overall execution time.

4.1 Learning with Standard Data-sets

Performance on standard image classification tasks was assessed using five benchmark data-sets: MNIST, CIFAR10, CIFAR100, TinyImageNet and ImageNet1k. For each training data-set experiments were performed using three different neural network architectures (see Tab. 3 in Appendix B.1). Performance was evaluated in terms of the following criteria: accuracy on standard test data, accuracy on common corruptions test data, ability to identify and reject samples from unknown classes, and the proportion of adversarial samples correctly classified or rejected (details in Appendix B.1).

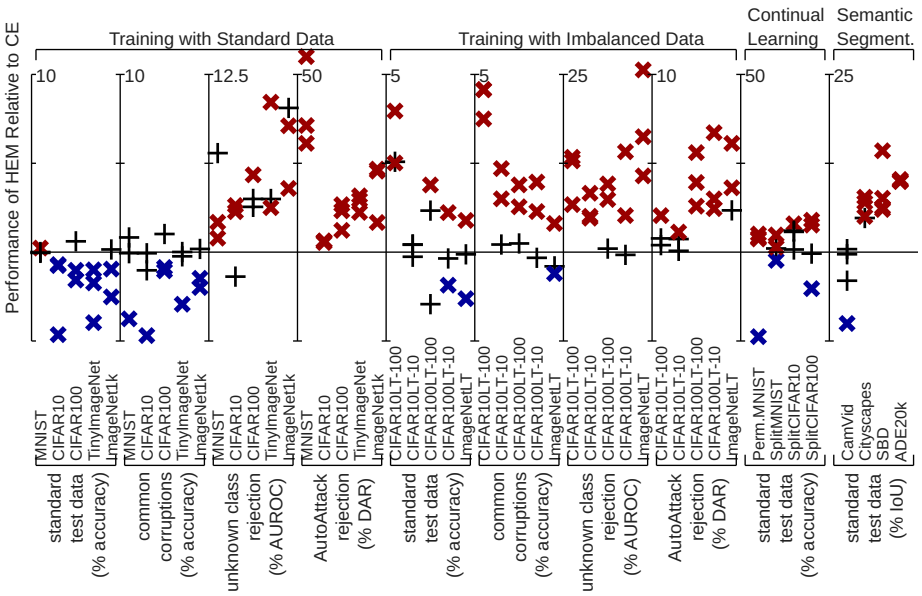

Fig. 2: Summary results for all the different tasks considered, showing the relative performance of high error margin (HEM) loss compared to Cross-Entropy (CE) loss. Relative performance is calculated as described in Fig. 1. Hence, points above the horizontal line indicate HEM performance better than that of CE. Please note the separate y-axis scales used in each segment of this figure. In contrast to Fig. 1, here separate results are shown for each training data-set. When there are multiple results for the same data-set, these were obtained using different DNN architectures, or in the case of continual learning different techniques for preventing catastrophic forgetting. Each marker shows the difference in the mean performance achieved across multiple trials. The 'x' and '+' markers indicate experiments where there was or was not a significant statistical difference in the performance produced by the two losses, as evaluated using the two-sample t-test (with $p<0.05$). The blue/red markers indicate conditions where CE/HEM loss had the significantly better performance. For information about the variability of performance across trials please see the more detailed results in Appendix C.

Performance on Standard Test Data and Common Corruptions Data Networks trained using CE loss and HEM loss (here, because the training data is balanced, HEM- = HEM) have comparable performance on classifying the standard test data and the common corruptions data (Figs. 1 and 2, 1^{st} and 2^{nd} segments), although CE loss has a small advantage. On average across the fifteen conditions (five data-sets with three network architectures per data-set) this advantage is 1.19% for clean accuracy and 1.07% for common corruptions. This difference is small compared to the changes in clean accuracy that can be produced by small changes to the training setup [22, 45, 61]. Of the specialized losses, LN loss achieves similar accuracy on clean and corrupt test data as HEM,

while DICE and MM losses perform much worse (Fig. 1, 1^{st} and 2^{nd} segments). Detailed results are given in Appendix C.1.

Performance on Unknown Class Rejection Classification accuracy measured on the standard test set, which contains samples from a similar input distribution to the training data, has been the main pre-occupation of most research in the history of machine learning so far. However, more recently, there has been growing concern that accuracy on the standard test data is insufficient to ensure that classifiers are safe, reliable, and trustworthy in more realistic scenarios [1, 2, 7, 18, 19, 23, 27, 38, 43, 47, 50–53, 64]. In particular, it is well known that CE-trained DNNs are susceptible to making over-confident predictions. For example, when shown samples that do not belong to any of classes in the training data a DNN may predict with high confidence that these samples belong to one of the known categories [2, 25, 33, 43]. Such over-confidence for unknown classes may cause errors in real-world scenarios where such samples might be commonly encountered and in situations where dishonest actors deliberately attempt to fool the classifier into making erroneous predictions.

To evaluate how susceptible a network is to this kind of overconfidence error, we can test whether we can detect and reject out-of-distribution samples based on the confidence of the network (see Appendix B.1 for details). This evaluation method is called open-set recognition [58, 62], out-of-distribution (OOD) detection/rejection [6, 25, 26, 42, 67], or unknown class rejection [53]. For the results reported here, Maximum Softmax Probability (MSP) [25] was used as the confidence score, but similar results were obtained using Maximum Logit Score (MLS) [24, 58] (see Fig. 4(c) in Appendix C.1).

For unknown class rejection, HEM provides significantly better performance on average than all other tested losses (Fig. 1, 3^{rd} segment). In all but one of the fifteen conditions we tested (five data-sets with three network architectures per data-set) HEM outperforms CE and its average performance is 4.52% better than CE (Fig. 2, 3^{rd} segment). DICE loss and MM loss perform very badly (Fig. 1, 3^{rd} segment). HEM loss even beats LN loss despite LN being a specialised loss that produces state-of-the-art performance on unknown class rejection. Detailed results are given in Appendix C.1, and an analysis of the prediction confidence scores, and the calibration of these scores, is given in Appendices D.2 and D.3.

Performance on AutoAttack Rejection Another robustness problem faced by DNNs is their susceptibility to adverserial attacks: being fooled into making the wrong prediction by small perturbations that do not change the class of the perturbed sample to a human observer [5, 16, 20, 34, 54]. Here we test susceptibility to this problem using the DAR score [53], which is the proportion of adverserially perturbed samples that are rejected as out-of-distribution or are not rejected but still classified correctly. The rejection/acceptance threshold is set so that 95% of correctly classified clean examples are accepted. Adverserial samples were generated using AutoAttack (AA) [12] (details in Appendix B.1).

HEM-trained networks have a large advantage over identical CE-trained architectures in terms of correctly dealing with adversarial attacks (Fig. 1, 4^{th} segment). In this case, HEM outperforms CE in all fifteen conditions (Fig. 2, 4^{th} segment), and its average performance is 17.17% better. Compared to CE, HEM loss has a much larger advantage in terms of its ability to enable accurate unknown class rejection, and also to detect adversarial attacks, than the small disadvantage it has in terms of clean and corrupt accuracy. HEM also outperforms, by a large margin, the other tested losses on adverserial robustness (Fig. 1,4^{th} segment). Detailed results are given in Appendix C.1.

4.2 Learning with Imbalanced Data-sets

The ability to learn when the training data contains a very different number of samples for different classes (*i.e.* with long-tailed data) was tested using the CIFAR10, CIFAR100 and ImageNet training data. This training data was modified to produce long-tailed data, using standard methods used in previous literature, by removing different numbers of samples from each class. Performance was evaluated on multiple network architectures using all the performance criteria used in the Sec. 4.1 to evaluate networks trained with balanced data-sets. Full details of the experimental methods are provided in Appendix B.2.

A comparison of the performance of networks trained on imbalanced data using CE and HEM losses reveals a similar pattern of results as where obtained with standard training data. Specifically, similar performance for the two losses on standard and corrupt test data (on average HEM is better by 0.70% and 1.54% respectively, see Figs. 1 and 2, 5^{th} and 6^{th} segments), better performance with HEM loss on unknown class rejection and adversarial attacks (on average HEM is better by 9.33% and 2.77% respectively, see Figs. 1 and 2, 7^{th} and 8^{th} segments).

Comparing HEM and LA loss (a version of CE designed to improve performance on imbalanced data) shows that LA has an advantage in terms of clean and corrupt accuracy, but that networks trained with HEM are better at identifying, and rejecting, unknown and adversarial samples. The high performance of LA loss on the clean data raises the prospect that there may be more optimal methods of adjusting the HEM margins to take account of the class imbalance.

HEM (and HEM-) perform as well as, or better than LN on all of the four evaluation metrics. DICE loss performs significantly worse than HEM on all evaluation criteria. However, the performance of DICE loss on adversarial attacks can, surprisingly, be improved beyond all other losses by using the MLS score as the rejection criterion (see Fig. 6(e) in Appendix C.2). A full set of more detailed results are provided in Appendix C.2.

4.3 Continual Learning

The performance of HEM loss when applied to continual learning was assessed using standard benchmark tasks: PermutedMNIST, SplitMNIST, SplitCIFAR10,

and SplitCIFAR100. Due to catastrophic forgetting [17], the over-writing of previously learned weights when training on a new task, all loss functions perform very poorly at continual learning unless a strategy is used to reduce forgetting. Many such strategies have been proposed. Here five were used: Replay [9, 49], Synaptic Intelligence (SI) [65], Elastic Weight Consolidation (EWC) [32], Less-Forgetful Learning (LFL) [28], and Learning without Forgetting (LwF) [35]. Each loss function was used in combination with each of these continual learning strategies, and performance was evaluated at the end of a sequence of five training episodes using unseen test data for all the five sub-tasks that were learnt during training. Full details of the experimental methods are provided in Appendix B.3

HEM performed 1.27% better on average at continual learning than CE loss (Fig. 1, 9^{th} segment), consistent with our expectations (Sec. 2.1). In 11 of the 16 conditions tested, better performance was obtained using HEM loss rather than CE loss (Fig. 2, 9^{th} segment). Furthermore, if only the best performing combination of loss and strategy of reducing catastrophic forgetting is considered for each loss, then in three of the four tasks HEM loss produces better performance than CE loss (Figs. 7(a) to 7(d) in Appendix C.3). This is remarkable as the training recipes used were designed to produce the best performance for each method of reducing catastrophic forgetting when paired with CE loss.

HEM has even greater advantages over the other applicable losses that were tested: LN, DICE and MM. LN and DICE losses perform very poorly on continual learning. This suggests that, unlike HEM and CE losses, they do not generalise to tasks outside of the specialised domain for which they were developed. As there is no class imbalance in the training data used here, LA is equivalent to CE, and HEM is equivalent to HEM-. Detailed results are given in Appendix C.3.

4.4 Semantic Segmentation

Performance on semantic segmentation was assessed using four standard data-sets: CamVid [8], Cityscapes [10], SBD [21], and ADE20k [68]. For each data-set multiple experiments were performed using the FPN architecture [31] with four different backbones. Full details can be found in Appendix B.4.

HEM- (the ablated version of the proposed loss with a single, shared, margin) performs better than HEM (the proposed loss that uses class specific margins to help deal with imbalanced data), suggesting that our heuristic for setting class-specific margins does not generalise from image classification to semantic segmentation (see Fig. 1, 10^{th} segment). However, even with sub-optimal margins, HEM produces 5.31% better average performance than CE, and exceeds by an even greater margin the performance of all the other existing losses considered. For some backbone architectures CE loss produced superior image segmentation performance to HEM loss on the CamVid data-set (Fig. 2, 10^{th} segment and Fig. 8(a) in Appendix C.4). However, for the three larger data-sets, HEM loss out-performed CE loss with all four backbone architectures (Fig. 2, 10^{th} segment and Figs. 8(b) to 8(d) in Appendix C.4).

Table 2: Ablation study on the effects of the proposed changes to MM loss on classification accuracy. Results are for ResNet18 networks trained on CIFAR10 and CIFAR100 using a margin of μ =0.2. Results are averaged over five trials and the standard deviation is given after the $\pm$ symbol. The best result in each column is highlighted in bold. The changes made to standard MM loss are denoted as "maz" for taking the mean of above-zero errors, and "thres" for setting errors below the mean to zero.

	Clean Accuracy (%)	
Loss	CIFAR10	CIFAR100
MM	93.79 ± 0.11	70.13 ± 0.19
+maz	93.81 ± 0.23	74.94 ± 0.35
+thres	93.78 ± 0.22	73.13 ± 0.26
+maz+thres = HEM	$\mathbf{93.84 \pm 0.19}$	$\mathbf{74.95 \pm 0.46}$

LN and LA losses perform very poorly (Fig. 1, 10^{th} segment) showing that these specialised losses do not work well outside of the specific domain for which they were created. DICE, a specialised loss developed specifically for segmentation tasks, has performance similar to that of CE, but worse than HEM and HEM-. The advantage of HEM is even clearer if for each data-set only the results for the backbone architecture that gives the best results for each loss is considered (Fig. 8(e) in Appendix C.4).

4.5 Ablation Study

HEM, differs from MM loss in terms of 1) using class-specific margins, and 2) how the errors are combined (Sec. 3). The effects of the first modification can be seen from the results for HEM- that have already been presented. The effects of the second modification were tested using ResNet18 networks trained on CIFAR10 and CIFAR100. The training set-up was as described in Appendix B.1.

The first change to how the errors are combined is to include only above-zero error values when calculating the mean error. This modification alone produces an improvement in classification accuracy (Tab. 2). As expected, this improvement is greatest for the data-set with the most class labels, as there will be more zero-valued errors across the larger number of logits that this modification enables the loss to ignore.

The second change to how the errors are combined was to set errors less than the mean to zero. On its own this modification is less effective than the first. This is to be expected, as this modification causes even more zeros to be included in the average, causing the loss to become low and learning to cease prematurely. However, when this modification is combined with the first it provides a small additional boost to performance by encouraging the loss to concentrate on the largest errors, particularly at the start of training when there are many errors. The advantage of HEM over MM in terms of clean accuracy is fairly small for the conditions shown in Tab. 2. However, as shown in Fig. 1, on average, over

many data-sets and network architectures, the advantages of HEM over MM are highly significant.

5 Conclusion

The proposed high error margin (HEM) loss has been shown to perform well across a wide range of tasks, data-set sizes, and network architectures. It was the best loss in our evaluations for unknown class and AutoAttack rejection, for continual learning, and for semantic segmentation, where it outperformed the other losses by a wide margin. Across the ten different types of evaluation we have performed, corresponding to the ten segments of Fig. 1, HEM was the best performing in five situations. In comparison, CE was found to be the best in only two. In a head-to-head comparison with CE loss, HEM was found to be superior in eight of the ten evaluations we performed (Fig. 2). CE was only best for classification of clean and corrupted images when training was performed with balanced training data, and remains the loss of choice to maximise accuracy in such circumstances. However, to improve rejection of out-of-distribution (OOD) samples, for training with long-tailed data, for continual learning, and certainly for semantic segmentation, HEM seems the better choice.

When training with balanced data, the drop in clean accuracy resulting from using HEM compared to CE was about about 1.2%. This reduction is likely to be negligible, as optimising the training hyper-parameters can yielded much bigger improvements in clean accuracy than the difference we observe [22, 45, 61] and in all our experiments, the training and evaluation choices were optimized for CE loss. A simple experiment where only the initial learning rate was modified based on intuition gained from observing the learning dynamics substantiates this claim (see Appendix D.4). Furthermore, clean accuracy is often reduced by techniques that increase robustness [53]. In this context, the reduction in clean accuracy caused by HEM would often be considered acceptable to achieve the large increases in performance on the other metrics.

Our results confirmed that specialised losses performed well on the tasks that they were designed for, but also showed that all these losses had very poor performance on at least one other task. LogitNorm (LN) loss [60], performed almost as well as HEM loss at OOD rejection, but HEM out-performed LN by a considerable margin on continual learning and semantic segmentation. With imbalanced training data, logit-adjusted (LA) loss [39] yielded better performance on standard test data than HEM loss, but HEM was superior at rejecting OOD samples. Furthermore, HEM out-performed LA at semantic segmentation, our only other task with class imbalance. For semantic segmentation the commonly used specialised loss, DICE [40], performed no better than CE loss in the experimental set-ups we used. HEM loss performed semantic segmentation more accurately than DICE, and out-performed DICE by a considerable margin on all other tasks. In contrast to the specialised losses that all failed outside their specific domain, HEM loss performed well on all tasks. This general applicability of HEM might lead to even greater advantages in other applications. For exam-

ple, in an situation where image segmentation needed to be performed with high robustness to unknown classes, a developer might be tempted to use LN loss due to it good performance in training image classifiers to be robust to OOD samples. However, our results show that this would be a poor choice as LN produces poor accuracy in image segmentation. In contrast, HEM loss is likely to do well in such an application. It is not hard to imagine other real-world scenarios where multiple advantages of our loss might combine to yield even greater advantages over CE and the specialised losses. For example, a task where it is necessary to learn continuously with long-tailed data and the resulting classifier needs to be robust to unknown classes.

Training with HEM loss is slightly faster than training with CE loss. HEM loss is zero for any training sample where the activation of the target logit is sufficiently above the value of the other logits. This means that during the later stages of learning many training samples cause no changes to the network weights, and it is possible for autograd to prune the computational graph. For example, this effect reduces training time by approximately 10% for a ResNet18 trained for 200 epochs on TinyImageNet. In dense prediction tasks there will be fewer opportunities to prune the computational graph, however, we still observe a small reduction in training time when using HEM. For example, training the ResNet34 backbone on Cityscapes was approximately 4% quicker using HEM compared to CE.

Future work might explore how to exploit the fact that samples can yield strictly zero loss for HEM. To save even more time, one could try presenting such learnt samples less frequently. Rather than saving time, one could focus augmentations or adversarial distortions on such zero-loss training samples to improve generalisation and/or robustness. More generally, future work might explore which augmentations, regularizations and training schedules work best for HEM loss. Some initial experiments showing that HEM is compatible with existing, complementary, methods for improving performance with imbalanced training data, adversarial robustness, and unknown class rejection are shown in Appendix E. Testing HEM with other such methods, and developing new techniques specifically designed to work well with HEM, might also be interesting topics for future research. Additionally, further work might explore alternative heuristics for setting the class-specific margins or ways of learning margins for different tasks. Subsequent research might also test HEM in other domains, such as language, as there is no reason why HEM loss should not also work for non-visual classification tasks.

Acknowledgements

The authors gratefully acknowledge use of the King's Computational Research, Engineering and Technology Environment (CREATE) and the HPC facilities of the University of Luxembourg [57] for carrying out the experiments described in this paper.

References

[1] Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access **6**, 14410–30 (2018). https://doi.org/10.1109/ACCESS.2018.2807385

[2] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety (2016), arXiv:1606.06565

[3] Awasthi, P., Mao, A., Mohri, M., Zhong, Y.: Theoretically grounded loss functions and algorithms for adversarial robustness. In: Ruiz, F., Dy, J., van de Meent, J.W. (eds.) Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). Proceedings of Machine Learning Research, vol. 206, pp. 10077–94 (2023), https://proceedings.mlr.press/v206/awasthi23c.html

[4] Azad, R., Heidary, M., Yilmaz, K., Hüttemann, M., Karimijafarbigloo, S., Wu, Y., Schmeink, A., Merhof, D.: Loss functions in the era of semantic segmentation: A survey and outlook (2023), arXiv:2312.05391

[5] Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition **84**, 317–31 (2018). https://doi.org/10.1016/j.patcog.2018.07.023

[6] Bitterwolf, J., Meinke, A., Augustin, M., Hein, M.: Breaking down out-of-distribution detection: Many methods based on OOD training data estimate a combination of the same core quantities. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2041–74 (2022), arXiv:2206.09880

[7] Bowers, J.S., Malhotra, G., Dujmović, M., Montero, M.L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J.E., Heaton, R.F., Evans, B.D., Mitchell, J., Blything, R.: Deep problems with neural network models of human vision. Behavioral and Brain Sciences **46**, e385 (2023). https://doi.org/10.1017/S0140525X22002813

[8] Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters **30**(2), 88–97 (2009). https://doi.org/10.1016/j.patrec.2008.04.005

[9] Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P.K., Torr, P.H.S., Ranzato, M.: On tiny episodic memories in continual learning (2019), arXiv:1902.10486

[10] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016), arXiv:1604.01685

[11] Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research **2**, 265–92 (2002)

[12] Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 2206–16 (2020), arXiv:2003.01690

[13] Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., Zhang, H.: Decoupled kullback-leibler divergence loss. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Proceedings of the Conference on Advances in Neural Information Processing Systems. pp. 74461–86. Curran Associates, Inc. (2024), https://openreview.net/forum?id=bnZZedw9CM, arXiv:2305.13948

[14] Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 9260–9 (2019). https://doi.org/10.1109/CVPR.2019.00949, arXiv:1901.05555

[15] Dablain, D., Jacobson, K.N., Bellinger, C., Roberts, M., Chawla, N.V.: Understanding CNN fragility when learning with imbalanced data. Machine Learning **113**, 4785–810 (2024). https://doi.org/10.1007/s10994-023-06326-9

[16] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning models. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2018), arXiv:1707.08945

[17] French, R.M.: Catastrophic forgetting in connectionist networks. In: Nadel, L. (ed.) Encyclopedia of Cognitive Science, vol. 1, pp. 431–5. Nature Publishing Group, London, UK (2003)

[18] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–73 (2020). https://doi.org/10.1038/s42256-020-00257-z

[19] Geirhos, R., Temme, C.R.M., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (2018), arXiv:1808.08750

[20] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (2015), arXiv:1412.6572

[21] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of the International Conference on Computer Vision (2011)

[22] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 558–567 (2019). https://doi.org/10.1109/CVPR.2019.00065, arXiv:1812.01187

[23] Heaven, D.: Why deep-learning AIs are so easy to fool. Nature **574**, 163–6 (2019)

[24] Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., , Song, D.: Scaling out-of-distribution detection for real-world settings. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 8759–73 (2022), https://proceedings.mlr.press/v162/hendrycks22a.html, arXiv:1911.11132

[25] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of the International Conference on Learning Representations (2017), arXiv:1610.02136

[26] Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., Steinhardt, J.: Pixmix: Dreamlike pictures comprehensively improve safety measures. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2022), arXiv:2112.05135

[27] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Proceedings of the Conference on Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), arXiv:1905.02175

[28] Jung, H., Ju, J., Jung, M., Kim, J.: Less-forgetting learning in deep neural networks (2016), arXiv:1607.00122

[29] Kanai, S., Yamaguchi, S., Yamada, M., Takahashi, H., Ohno, K., Ida, Y.: One-vs-the-rest loss to focus on important samples in adversarial training. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202 (2023), arXiv:2207.10283

[30] Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing (2018), arXiv:1803.06373

[31] Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 6392–401 (2019). https://doi.org/10.1109/CVPR.2019.00656, arXiv:1901.02446

[32] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences USA **114**(13), 3521–6 (2017). https://doi.org/10.1073/pnas.1611835114

[33] Kumano, S., Kera, H., Yamasaki, T.: Are DNNs fooled by extremely unrecognizable images? (2022), arXiv:2012.03843

[34] Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. In: Proceedings of the International Conference on Learning Representations (2017), arXiv:1607.02533

[35] Li, Z., Hoiem, D.: Learning without forgetting. In: Proceedings of the European Conference on Computer Vision. pp. 614–29 (2016), arXiv:1606.09282

[36] Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021). https://doi.org/10.1016/j.media.2021.102035

[37] Mao, C., Zhong, Z., Yang, J., Vondrick, C., Ray, B.: Metric learning for adversarial robustness. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (2019), arXiv:1909.00900

[38] Marcus, G.: The next decade in AI: Four steps towards robust artificial intelligence (2020), arXiv:2002.06177

[39] Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: Proceedings of the International Conference on Learning Representations (2021), https://openreview.net/forum?id=37nvvqkCo5, arXiv:2007.07314

[40] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–71 (2016). https://doi.org/10.1109/3DV.2016.79

[41] Ming, Y., Sun, Y., Dia, O., Li, Y.: How to exploit hyperspherical embeddings for out-of-distribution detection? In: Proceedings of the International Conference on Learning Representations (2023), https://openreview.net/forum?id=aEFaE0W5pAd

[42] Mohseni, S., Pitale, M., Yadawa, J., Wang, Z.: Self-supervised learning for generalizable out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5216–23 (2020). https://doi.org/10.1609/aaai.v34i04.5966

[43] Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015), arXiv:1412.1897

[44] Pang, T., Xu, K., Dong, Y., Du, C., Chen, N., Zhu, J.: Rethinking softmax cross-entropy loss for adversarial robustness. In: Proceedings of the International Conference on Learning Representations (2020), arXiv:1905.10626

[45] Pang, T., Yang, X., Dong, Y., Su, H., Zhu, J.: Bag of tricks for adversarial training. In: Proceedings of the International Conference on Learning Representations (2021), arXiv:2010.00467

[46] Panum, T.K., Wang, Z., Kan, P., Fernandes, E., Jha, S.: Exploring adversarial robustness of deep metric learning (2021), arXiv:2102.07265

[47] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: IEEE European Symposium on Security and Privacy (2016), arXiv:1511.07528

[48] Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. In: Larochelle, H., Ranzato, M.,

Hadsell, R., Balcan, M.F., Lin, H. (eds.) Proceedings of the Conference on Advances in Neural Information Processing Systems. vol. 33, pp. 4175–4186. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/2ba61cc3a8f44143e1f2f13b2b729ab3-Paper.pdf, arXiv:2007.10740

[49] Robins, A.: Catastrophic forgetting in neural networks: the role of rehearsal mechanisms. In: Proceedings of the First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems. pp. 65–8. IEEE (1993)

[50] Roy, A., Cobb, A., Bastian, N.D., Jalaian, B., Jha, S.: Runtime monitoring of deep neural networks using top-down context models inspired by predictive processing and dual process theory. In: Proceedings of the AAAI Conference on Artificial Intelligence. Spring Symposium (2022)

[51] Sa-Couto, L., Wichert, A.: Simple Convolutional-Based Models: Are They Learning the Task or the Data? Neural Computation **33**(12), 3334–50 (2021). https://doi.org/10.1162/neco_a_01446

[52] Serre, T.: Deep learning: The good, the bad, and the ugly. Annual Review of Vision Science **5**(1), 399–426 (2019). https://doi.org/10.1146/annurev-vision-091718-014951

[53] Spratling, M.W.: A comprehensive assessment benchmark for rigorously evaluating deep learning image classifiers. Neural Networks **192**(107801) (2025), arXiv:2308.04137

[54] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (2014), arXiv:1312.6199

[55] Tack, J., Yu, S., Jeong, J., Kim, M., Hwang, S.J., Shin, J.: Consistency regularization for adversarial robustness. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022), arXiv:2103.04623

[56] Terven, J., Cordova-Esparza, D.M., Romero-González, J.A., Ramírez-Pedraza, A., Chávez-Urbiola, E.A.: A comprehensive survey of loss functions and metrics in deep learning. Artificial Intelligence Review **58**(195) (2025). https://doi.org/10.1007/s10462-025-11198-7

[57] Varrette, S., Cartiaux, H., Peter, S., Kieffer, E., Valette, T., Olloh, A.: Management of an Academic HPC & Research Computing Facility: The ULHPC Experience 2.0. In: Proc. of the 6th ACM High Performance Computing and Cluster Technologies Conf. (HPCCT 2022). Association for Computing Machinery (ACM), Fuzhou, China (Jul 2022)

[58] Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: a good closed-set classifier is all you need? In: Proceedings of the International Conference on Learning Representations (2022), arXiv:2110.06207

[59] Wang, Q., Ma, Y., Zhao, K., Tian, Y.: A comprehensive survey of loss functions in machine learning. Annals of Data Science **9**, 187–212 (2022). https://doi.org/10.1007/s40745-020-00253-5

[60] Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 23631–44 (2022), arXiv:2205.09310

[61] Wightman, R., Touvron, H., Jégou, H.: ResNet strikes back: An improved training procedure in timm. In: Proceedings of the Conference on Advances in Neural Information Processing Systems. Workshop on ImageNet: Past, Present, and Future (2021), https://openreview.net/forum?id=NG6MJnV16M5, arXiv:2110.00476

[62] Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., Liu, Z.: OpenOOD: Benchmarking generalized out-of-distribution detection. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (2022), arXiv:2210.07242

[63] Yu, Y., Xu, C.Z.: Efficient loss function by minimizing the detrimental effect of floating-point errors on gradient-based attacks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 4056–66 (2023). https://doi.org/10.1109/CVPR52729.2023.00395

[64] Yuille, A.L., Liu, C.: Deep nets: What have they ever done for vision? International Journal of Computer Vision **129**, 781–802 (2021). https://doi.org/10.1007/s11263-020-01405-z

[65] Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3987–95 (2017), arXiv:1703.04200

[66] Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the International Conference on Machine Learning (2019), arXiv:1901.08573

[67] Zhang, L.H., Ranganath, R.: Robustness to spurious correlations improves semantic out-of-distribution detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023), arXiv:2302.04132

[68] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2017)

[69] Zhu, Q., Zheng, G., Yan, Y.: Effective out-of-distribution detection in classifier based on PEDCC-loss. Neural Processing Letters **55**, 1937–49 (2023). https://doi.org/10.1007/s11063-022-10970-y