T2FNorm: Train-time Feature Normalization for OOD Detection in Image Classification

Sudarshan Regmi[1,2]*, Bibek Panthi[2], Sakar Dotel[3],
Prashnna K Gyawali[4], Danail Stoyanov[1], Binod Bhattarai[1,5]

[1] University College London, UK
[2] NepAl Applied Mathematics and Informatics Institute for research, Nepal
[3] Tribhuvan University, Nepal
[4] West Virginia University, USA
[5] University of Aberdeen, UK

Abstract

*Neural networks are notorious for being overconfident predictors, posing a significant challenge to their safe deployment in real-world applications. While feature normalization has garnered considerable attention within the deep learning literature, current train-time regularization methods for Out-of-Distribution(OOD) detection are yet to fully exploit this potential. Indeed, the naive incorporation of feature normalization within neural networks does not guarantee substantial improvement in OOD detection performance. In this work, we introduce **T2FNorm**, a novel approach to transforming features to hyperspherical space during training, while employing non-transformed space for OOD-scoring purposes. This method yields a surprising enhancement in OOD detection capabilities without compromising model accuracy in in-distribution(ID). Our investigation demonstrates that the proposed technique substantially diminishes the norm of the features of all samples, more so in the case of out-of-distribution samples, thereby addressing the prevalent concern of overconfidence in neural networks. The proposed method also significantly improves various post-hoc OOD detection methods.*

1. Introduction

The efficacy of deep learning models is contingent upon the consistency between training and testing data distributions; however, the practical application of this requirement presents challenges when deploying models in real-world scenarios, as they are inevitably exposed to OOD samples. Consequently, a model's ability to articulate its limitations and uncertainties becomes a critical aspect of its performance. While certain robust methodologies exist that endeavor to achieve generalizability despite domain shifts, these approaches do not always guarantee satisfactory performance. Hence, detecting OOD samples is of paramount importance.

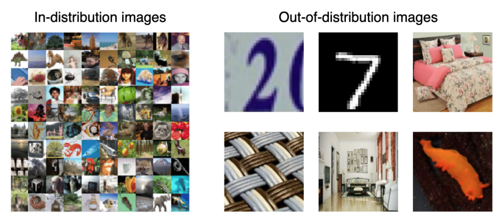


Figure 1. Objective of OOD detection is to differentiate between samples from in-distribution and out-of-distribution categories.

OOD detection (Figure 1) approaches can be broadly grouped into three approaches: post-hoc methods, outlier exposure, and training time regularization. Post-hoc methods, deriving OOD likelihood from pre-trained models, have significantly improved while outlier exposure, despite the challenges in predefining OOD samples ideally, is prevalently adopted in industrial contexts. Another approach involves training time regularization. This line of work due to its capacity to directly impose favorable constraints during training, potentially offers the most promising path to superior performance. The training-time regularization method, LogitNorm [41], employs L2 normalization at the logit level to mitigate overconfidence, leading to an increased ratio of ID norm to OOD norm compared to the results from simple cross-entropy baseline or Logit Penalty [41]. Nonetheless, importance of feature norm in achieving ID/OOD separability has been underscored in recent works [10, 33, 35, 41].

*Correspondence: {sudarshan.regmi}@ucl.ac.uk

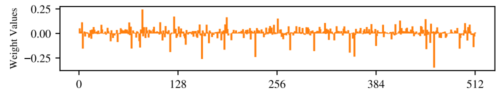


Figure 2. Smooth FC weights of Airplane class in ResNet18 induced by LogitNorm [41] optimization.

Furthermore, in LogitNorm [41], suppressing the norm of logit through logit normalization can leave a short-cut for the model to learn predominantly near-zero FC weights. Indeed, empirical observation (Figure 2) suggests that the optimization process induces smoother uniform weight values closer to zero in the FC layer. However, a recent work DICE [31] has shown that non-trivial dependence on unimportant weights potentially contributes to failure in detecting OOD samples. The presence of smoother weights implies irrelevant features contributing non-trivially to the classification for some predictions resulting in higher output variance for OOD samples. Moreover, though DICE explores the importance of sparsity in a posthoc manner, it operates on overconfident features and there still exists potential room for obtaining sparse features through training time regularization.

Towards this, in this work, we propose a novel strategy of adopting scaled feature normalization during training while excluding it during OOD detection. The feature normalization, during training, prevents the network from being incentivized to learn dense features thereby addressing overconfidence. The avoidance of normalization during OOD detection is necessary to preserve the difference in response of the network towards OOD and ID samples. We demonstrate that with the proposed feature normalization, we achieve a clear distinction between the norm of ID and OOD data samples, eventually contributing toward a substantial performance improvement without compromising the model's accuracy. We show a boost in OOD detection in a number of OOD benchmark datasets (Table 1). For instance, our method reduces the FPR@95 score by **34%** with respect to baseline and by **7%** with respect to LogitNorm on average across a variety of 9 OOD datasets with DICE scoring on ResNet-18 architecture. In addition, our method works well in conjunction with many post hoc methods. Our key results and contributions are:

- We propose T2FNorm – a surprisingly trivial yet powerful plug to regularize the model for OOD detection. We quantitatively show that train time normalization approximately projects the features of ID samples to the surface of a hypersphere differentiating it from OOD samples thereby achieving significantly higher ***separability ratio***.
- We show T2FNorm is equally effective across multiple deep learning architectures and multiple datasets. It also works well in conjunction with multiple post-hoc methods.
- We perform both qualitative and quantitative analysis showing our method's ability to reduce overconfidence and also perform a sensitivity study to show the robustness of our model to the temperature parameter τ.
- We show that **skipping feature normalization during OOD scoring time** is a key contributor to our method thus paving the way for exploring the effectiveness of other forms of normalization discrepancies during OOD scoring.

2. Related Works

OOD Detection Numerous studies have emerged in recent years focusing on OOD detection. A straightforward method for OOD detection is a simple maximum softmax probability [7]. However, it remains an unreliable scoring metric for OOD detection because of inherent overconfidence imposed by training with one-hot labels [25]. OOD detection has been primarily tackled with three lines of approach in the literature (a) post-hoc methods, (b) outlier exposure and (c) train-time regularization. Post-hoc methods [5, 7, 9, 19, 22, 29, 31, 32, 39] aim to improve the ID/OOD separability with pretrained models trained only with the aim for accuracy. Outlier exposure is another less studied line in academic research, as the assumption of the nature of OOD limits the ideal applications. However, it is found to be commonly used for industrial purposes. Training time regularization [2, 8, 13, 18, 23, 28, 40] employs some form of regularizer in the training scheme, and this line of work due to its capacity to directly impose favorable constraints during training potentially offers the most promising path to superior performance for OOD detection. For instance, LogitNorm [41] employs logit normalization as training time regularization to address the overconfidence issue and, thereby, improve OOD detection. Furthermore, LogitNorm shows overconfidence can somewhat be addressed sub-optimally with logit penalty too. Different from LogitNorm, our work pertains to addressing overconfidence in the feature space thereby automatically addressing overconfidence in the logit space. Our work deals with high-dimensional normalization. For the first time, we delve into the feature normalization discrepancy between the training and OOD evaluation phases.

Normalization The utility of normalization in ensuring consistent input distribution and reducing covariate shift has proven beneficial in various subareas of deep learning [27, 30, 43, 47]. Normalization consisting of learnable parameters such as Batch Normalization [12], Layer Normalization [1], and Group Normalization [42], have been effective in mitigating training issues of neural networks. On the other hand, the strategic placement of L2 normalization has also been a popular recipe for training more effective deep learning models. Similar to our work, Ranjan et al. [27] constrains the features to lie on the hypersphere of fixed

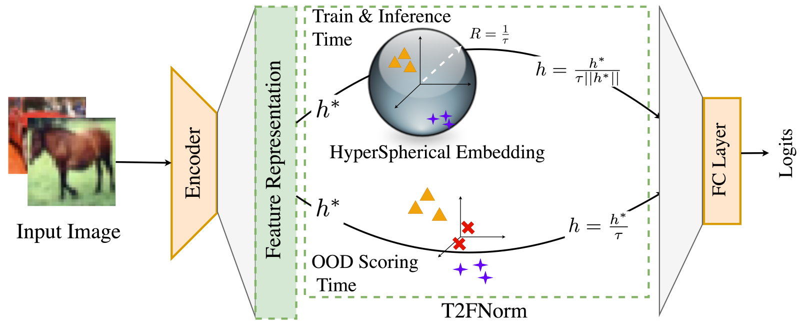


Figure 3. Schematic diagram of our method: T2FNorm. Features are L_2 normalized and scaled during training and inference time, while normalization is avoided for OOD Scoring.

radius for face verification purposes but does so in both the training and testing phase without scaling. Further works in deep metric learning such as ArcFace [3], CosFace [38], SphereFace [21], etc realize the effectiveness of normalization. Specifically, Techapanurak et al. [36] shows the hyperparameter-free OOD detection method introducing cosine loss by taking inspiration from norm face [37] where both the penultimate feature and fully connected layer are normalized. Our approach differs from cosine loss in three different ways. a) The temperature parameter is learned in the cosine loss method whereas we set a fixed temperature across all 6 settings. While it may seem extra hyperparameter is being added, we find the value of τ to be architecture agnostic as well as dataset agnostic. b) Unlike cosine loss, we avoid normalizing the classification layer freeing it to learn non-smooth weight values which, in turn, boost compatibility with various downstream OOD scoring methods as they rely on ID-OOD separability based on magnitudes. c) Importantly, we remove the constraint of hyperspherical embeddings in the OOD scoring phase while Techapanurak et al. [36] uses cosine similarity and is not compatible with other OOD scoring functions. Guo et al. [6] provided a study showing modern neural networks' poor calibration and proposed to use temperature scaling as posthoc method to improve calibration. Platt scaling [26] is another simple postprocessing calibrating technique. Label smoothing [34] helps to avoid overconfident calibration by adding uncertainty to the one-hot encoding of labels.

3. Method

3.1. Preliminaries: Out of Distribution Detection

Setup Let $\mathcal{X}$ be input space, $\mathcal{Y}$ be output space and $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Let the $\mathcal{P}_{in}$ be the marginal distribution of $\mathcal{X}$ which represents the distribution of input we want our classifier to be able to handle. This is the in-distribution (ID) of the input labels x_i.

Supervised Classification In supervised classification, the goal is to minimize the empirical loss $\mathcal{L}$ function formulated as: $\min_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i)$ over the input dataset which is sampled *i.i.d.* from the in-distribution $\mathcal{P}_{in}$ using model f_θ. Here, θ is the model parameters, $f_\theta(x_i)$ is the classification predicted for input x_i by the model with parameters θ. The model f_θ is composed of encoder ϕ and fully connected layer FC.

OOD Detection During test time, the environment can present samples from a different distribution $\mathcal{P}_{out}$ instead of from $\mathcal{P}_{in}$. The goal of Out of Distribution Detection is to differentiate between samples from in-distribution $\mathcal{P}_{in}$ and out-of-distribution $\mathcal{P}_{out}$. In this work, we treat OOD detection as a binary classification where a scoring function $SC(\mathbf{x})$ and a corresponding threshold λ provide a decision function $g(\mathbf{x})$ that performs OOD detection:

$$g(\mathbf{x}) = \begin{cases} \text{In-distribution,} & \text{if } SC(\mathbf{x}) \geq \lambda \\ \text{Out-of-distribution,} & \text{if } SC(\mathbf{x}) < \lambda \end{cases} \quad (1)$$

The simplest of the scoring function $SC(\mathbf{x})$ is the Maximum Softmax Probability (MSP) obtained by passing the logits from the final layer of the network to the softmax function and taking the maximum value. Then samples with MSP exceeding a certain threshold λ are classified ID and the rest are OOD. The threshold λ is usually chosen so as to have a true positive rate of 95% over the input dataset.

3.2. Guiding Principle for Image Classification: A Feature Perspective

In image classification, the images are classified into the appropriate categories depending upon the presence of categorical features. Considering $\vec{f_r} = (f_r^1, f_r^2, ...)$ to be relevant in-distribution features and $\vec{f_i} = (f_i^1, f_i^2, ...)$ to be irrelevant in-distribution features for an image of k^{th} category, then appropriate linear combination (assuming non-negative setup) of in-distribution features with corresponding importance vectors $\vec{w_{r_k}} = (w_{r_k}^1, w_{r_k}^2, ...)$ and $\vec{w_{i_k}} = (w_{i_k}^1, w_{i_k}^2, ...)$ gives the decision score D_k:

$$D_k = \vec{f_r} \cdot \vec{w_{r_k}} + \vec{f_i} \cdot \vec{w_{i_k}} \quad (2)$$

This means the relevant semantic features, which ID samples contain, must be learned and activated while suppressing irrelevant features for k^{th} category classification making decision score D_k higher than that of other category m, $D_m, m \neq k$. Roughly speaking, $\vec{w_{r_k}} \approx \vec{1}$ and $\vec{w_{i_k}} \approx \vec{0}$. However, OOD samples, by definition, inherently don't contain such a combination of relevant semantic features that form any category m. Hence, a fundamental difference in feature representation exists between ID and OOD, which can potentially be exploited for OOD detection. As such, in this work, we aim to capture feature representation for ID and OOD samples differently, such that they are separable. Towards this, we argue for imposing a hypersphere in our embedding space such that the network learns to produce high-level relevant semantic ID features lying on the hypersphere due to normalization performed during training. However, this happens only for ID samples as the network was trained with them, while for OOD samples, high-level relevant semantic ID features are not activated because of their absence, causing OOD feature representation to lie significantly beneath hypersphere's surface.

3.3. Significance of Feature Norm

As observed by recent works [10, 11, 35], generally ID samples have a more significant feature norm in comparison to OOD data. In CNN models, high-level spatial features are generated by convolution operations. The penultimate feature is derived from globally pooling post-ReLU spatial features. ReLU activation signifies presence of specific in-distribution features, while their absence corresponds to smaller norms, often seen in out-of-distribution samples. Therefore, a neural network having better ID/OOD separability should demonstrate a higher relative norm for in-distribution versus out-of-distribution samples. Quantitatively, we can formalize such discriminability as the ratio of mean ID norm to mean OOD norm, resulting in a novel OOD detection metric which we term as ***separability ratio*** ($\mathcal{S}$). Given n_{ood} and n_{id} refer to the number of OOD and ID samples, the separability ratio $\mathcal{S}$ can be formulated:

$$S = \frac{\frac{1}{n_{\text{id}}} \sum_{\text{x} \in X_{\text{id}}} \|\phi(\text{x})\|_2}{\frac{1}{n_{\text{ood}}} \sum_{\text{x'} \in X_{\text{ood}}} \|\phi(\text{x'})\|_2} \tag{3}$$

However, overconfident features induced by the one-hot cross-entropy objective don't allow the model to exhibit the ideal behavior of near-zero activation in feature representation for OOD samples thereby compromising OOD detection performance.

3.4. T2FNorm: Train-Time Feature Normalization for OOD detection

Our work proposes a method **T2FNorm** to improve the robustness of the network itself for OOD detection which can be used in conjunction with any downstream classification-based scoring function. Our method introduces two simple operations: Normalization and Scaling of the features during training of a classification network. We discuss their role and significance for OOD detection in the following sections.

We perform feature normalization to alleviate the issue of over-confident predictions at the feature level. Specifically, the model f_θ, consisting of the encoder ϕ and classification layer FC, is trained with the cross-entropy objective $\mathcal{L}_{\text{CE}}$. In the training loop, the encoder encodes the images into one-dimensional feature representations h^*. The feature representations are then L_2-normalized and scaled with temperature $1/\tau$ to produce hyperspherical embeddings h. The embeddings are then finally classified with the FC layer. The optimized loss is given in Equation 4.

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}_{\mathcal{XY}}} \left[\mathcal{L}_{\text{CE}} \left(FC(\hat{\phi}(\boldsymbol{x})/\tau), y\right)\right] \tag{4}$$

Importantly, the normalization is performed (Algorithm 1) only during training and inference time, however, we skip the normalization part for OOD detection (Algorithm 2). Using MSP scoring, the scoring function SC avoiding normalization (using ϕ instead of $\hat{\phi}$) is given in Equation 5.

$$SC = MSP(FC(\phi(\boldsymbol{x})/\tau)) \tag{5}$$

Algorithm 1 T2FNorm: Training

```
Input: Dataset D, Feature Extractor φ,
classification layer FC
function train(D)
  for (x_i, y_i) ← D do
    h* ← φ(x_i)
    h ← h*/τ||h*||_2
    L ← L_CE (FC(h), y_i)
    L.backward()
  end for
end function
```

Algorithm 2 T2FNorm: Inference

```
function classify(x)
  h* ← φ(x_i)
  if SC(x; h*/τ) < γ then
    return OOD
  else
    h ← h*/τ||h*||_2
    logits ← FC(h)
    return argmax_i logits_i
  end if
end function
```

As shown in the Algorithm 2, for the OOD detection, a given sample is classified as OOD if the scoring function

SC yields a value higher than γ. Otherwise, the sample is determined as originating from in-distribution and is then classified with the FC layer. Figure 3 shows the schematic diagram for our method. The proposed approach is simple and easy to implement, and as we will show later, it produces improved performance for OOD detection while maintaining predictive abilities.

High Dimensional normalization Given that feature-level normalization implies normalization within a higher-dimension space than logit-level normalization, we postulate that high-dimension normalization of training ID samples would enable the network to significantly reduce OOD norm relative to ID norm while preserving ID-specific features. As a result, we anticipate a substantial decrease in overconfidence, which is intrinsically linked to logit and feature norms, primarily since overconfidence is addressed at the penultimate feature level, indirectly tackling the norm at the logit level. Confirming this, recent work, ReAct [33], has observed that the penultimate layer is most effective for OOD detection due to the distinct activation patterns between ID and OOD data.

Intuition of avoiding normalization at OOD scoring Should feature normalization be adopted during OOD scoring, it erroneously activates features for OOD samples (L_2 norm of feature = 1), causing OOD samples to mimic the behavior of ID samples within feature space. However, the removal of normalization helps to preserve the difference in response of the network towards OOD and ID samples.

4. Experiments

In this section, we discuss the experiments performed in various settings to verify the effectiveness of our method.

Datasets: We use CIFAR-10 [14] and CIFAR-100 [15] as in-distribution datasets. Texture [17], TinyImageNet (TIN) [16], MNIST [4], SVHN [24], Places365 [48], iSUN [44], LSUN-r [46], LSUN-c [46] are used as out-of-distribution datasets. Following [45], we use CIFAR-10 as OOD if CIFAR-100 is used as in-distribution and vice versa.

Metrics and OOD scoring: We report the experimental results in three metrics: FPR@95, AUROC and AUPR. FPR@95 gives the false positive rate when the true positive rate is 95%. AUROC denotes the area under the receiver operator characteristics curve and AUPR denotes the area under the precision-call curve. We use multiple OOD scoring methods, including parameter-free scoring functions such as maximum softmax probability [7], parameter-free energy score [22] and GradNorm [10] as well as hyperparameter-based scoring functions such as ODIN [20], ReAct [32], Activation Reshaping [5], and DICE [31]. The hyperparameters are optimized with respect to the validation set. A part of CIFAR10 is used as a validation set when CIFAR100 is used as an ID dataset and vice-versa.

Experimental Details: We perform experiments with three training methods: a) Baseline (cross-entropy), LogitNorm [41], and T2FNorm (ours) by following the training procedure of open-source framework OpenOOD [45]. Experiments were performed across ResNet-18, WideResnet(WRN-40-2), and DenseNet architectures with an initial learning rate of 0.1 with weight decay of 0.0005 for 100 epochs based on the cross-entropy loss function. We set the temperature parameter $\tau = 0.04$ for LogitNorm as recommended in the original setting [41] and $\tau = 0.1$ for T2FNorm. Please refer to Figure 11 for the sensitivity study of τ. Five independent trials are conducted for each of 18 training settings (across 2 ID datasets, 3 network architectures, and 3 training methods). We trained all models on NVIDIA A100 GPUs.

5. Results

Superior OOD Detection Performance Quantitative results are presented in Table 1. It shows that our method is consistently superior in FPR@95, AUROC as well as AUPR metrics. Our method reduces FPR@95 metric by 34% compared to Baseline and 7% compared to LogitNorm using DICE Scoring for ResNet-18. Interestingly, for both ID datasets, we can also observe the incompatibility of LogitNorm with DICE scoring in DenseNet architecture where it underperforms even when compared to the baseline. On the other hand, our method is more robust regardless of architecture or OOD scoring method.

Architecture Agnostic without Compromising Accuracy Our experiments across three architectures as reported in Table 1 show the compatibility of our method with various architectures evidencing the agnostic nature of our method to architectural designs. An essential attribute of OOD methods employing regularization during training is the preservation of classification accuracy in ID datasets, independent of their OOD detection performance. The evidence supporting these assertions can be found in Table 2.

Significant Reduction in Overconfidence In Figure 6, we show the comparison between Baseline, LogitNorm, and T2FNorm in terms of distribution of maximum softmax probability. It can be observed that overconfidence has been addressed by T2FNorm to a greater extent in comparison with the baseline. Though the issue of overconfidence is also reduced in LogitNorm, the separability ratio is significantly higher in T2Norm, as we show in Figure 7.

Norm and Separability Ratio The statistics of norm and separability ratio for ResNet-18 model trained with CIFAR-10 datasets are given in Table 4. The average ID norm of $0.9 \sim 1$ for the penultimate feature implies empirically that, ID samples approximately lie on the hypersphere even at the pre-normalization stage. Again, the average norm for OOD samples is found to be 0.15 implying OOD samples lie significantly beneath the hypersphere as ID-specific features are not activated appreciably. This depicts a clear difference

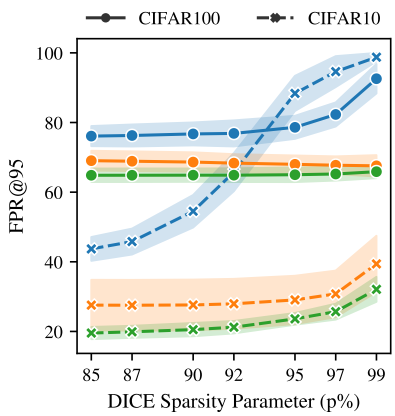


Figure 4. Sensitivity study of sparsity parameter p [31] shows superior performance and robustness of T2FNorm.

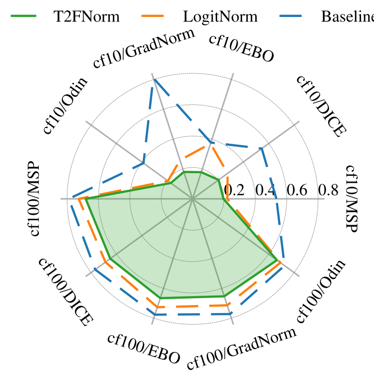


Figure 5. FPR@95 across different scoring functions in CIFAR100 (cf100) and CIFAR10 (cf10).

Table 1. Mean OOD metrics in the form of Baseline/LogitNorm/T2FNorm with hyperparameter-free (MSP, EBO) as well as hyperparameter-based OOD scoring (DICE). **Bold** numbers are superior.

		CIFAR-10			CIFAR-100		
	Network	FPR@95 ↓	AUROC ↑	AUPR↑	FPR@95↓	AUROC↑	AUPR↑
MSP	ResNet-18	53.4 / 22.1/ **19.7**	90.7 / 96.0 / **96.5**	90.8 / 95.7 / **96.4**	78.9 / 72.6 / **68.2**	79.0 / 80.1 / **83.2**	79.8 / 79.4 / **82.4**
	WRN-40-2	53.4 / 22.6 / **22.4**	90.1 / 95.9 / **95.9**	90.2 / 95.8 / **95.9**	81.8 / 63.5 / **63.2**	74.7/ 83.8 / **83.9**	76.6 / 83.7 / **84.2**
	DenseNet	48.8 / 24.0 / **21.0**	91.7 / 95.4 / **96.1**	91.6 / 95.3 / **96.2**	77.4 / 66.8 / **64.1**	77.6 / 82.1 / **84.1**	79.8 / 82.6 / **84.6**
	Mean	51.9 / 22.9 / **21.0**	90.9 / 95.8 /**96.2**	90.9 / 95.6 / **96.2**	79.4 / 67.6 /**65.1**	77.1 / 82.0 /**83.7**	78.7 / 81.9 / **83.7**
DICE	ResNet-18	54.5 / 27.6 / **20.5**	86.0 / 94.4 / **96.3**	87.4 / 94.1 / **96.1**	76.7 / 68.6 / **64.9**	81.0 / 73.5/ **83.1**	81.2 / 74.6 / **81.9**
	WRN-40-2	36.5 / 32.5 / **26.0**	89.0 / 92.6 / **95.1**	90.9 / 92.8 / **95.1**	76.4/ 59.1 / **55.7**	74.8 / 81.6 / **84.6**	76.0 / 81.6 / **84.6**
	DenseNet	30.8 / 38.0 / **23.0**	92.3 / 90.3 / **95.4**	93.3 / 90.7 / **95.4**	63.4/ 68.2 / **61.2**	82.8 / 75.7 / **82.9**	**83.5** / 75.8 / 82.6
	Mean	40.6 / 32.7 / **23.2**	89.1 / 92.5 / **95.6**	90.5 / 92.6 / **95.5**	72.2 / 65.3 / **60.6**	79.6 / 76.9 / **83.5**	80.2 / 77.3 / **83.0**
EBO	ResNet-18	37.7 / 37.0 / **17.9**	91.5 / 88.9 / **96.7**	92.7 / 89.4 / **96.6**	77.6 / 72.6 / **66.6**	81.0 / 75.1 / **83.3**	81.2 / 75.3 / **82.2**
	WRN-40-2	35.3 / 54.9 / **22.5**	91.1 / 85.0 / **95.8**	92.1 / 84.1 / **95.7**	78.0 / 62.6 / **60.0**	77.0 / 81.7 / **84.2**	78.3 / 81.9 / **84.4**
	DenseNet	30.3 / 73.9 / **20.0**	93.3 / 86.3 / **96.1**	93.8 / 83.2 / **96.1**	69.2 / 70.3 / **62.2**	82.4 / 75.7 / **83.4**	83.6 / 77.0 / **84.0**
	Mean	34.5 / 55.3 / **20.1**	92.0 / 86.7 / **96.2**	92.9 / 85.6 / **96.2**	75.0 / 68.5 / **63.0**	80.1 / 77.5 / **83.6**	81.0 / 78.1 / **83.5**

Table 2. Accuracy in % with (Baseline / LogitNorm / T2FNorm)

Architectures	CIFAR-10	CIFAR-100
DenseNet	94.94 / 94.05 / 94.62	76.51 / 76.50 / 76.06
WRN-40-2	94.72 / 94.38 / 94.44	75.30 / 74.79 / 75.51
ResNet-18	94.79 / 94.13 / 94.94	77.02 / 75.85 / 76.42

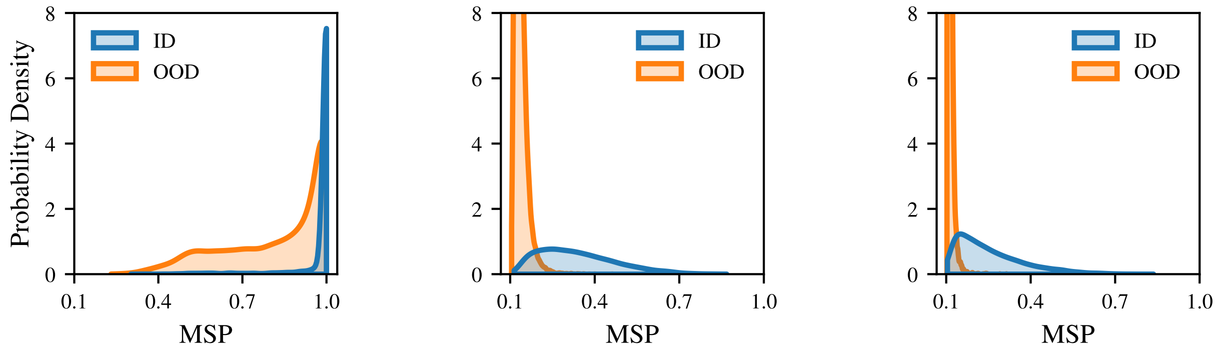
Table 3. Additional results with CIFAR10 (ID) using ResNet-18 network.

Methods	FPR@95↓	AUROC↑	AUPR↑
ReAct	41.58 ± 3.69	90.26 ± 1.65	83.33 ± 1.98
ReAct + T2FNorm	$\mathbf{18.54} \pm \mathbf{0.92}$	$\mathbf{96.59} \pm \mathbf{0.14}$	$\mathbf{96.03} \pm \mathbf{0.13}$
ASH	55.54 ± 10.42	79.91 ± 2.36	66.95 ± 1.52
ASH + T2FNorm	$\mathbf{18.08} \pm \mathbf{0.93}$	$\mathbf{96.64} \pm \mathbf{0.14}$	$\mathbf{96.08} \pm \mathbf{0.16}$

Table 4. Norm of features for ID and OOD samples. (ID / OOD↓ / $\mathcal{S}$ ↑)

Method	Feature
Baseline	6.16 ± 0.27 / 5.43 ± 0.20 / 1.13 ± 0.05
LogitNorm	1.90 ± 0.11 / 0.69 ± 0.12 / 2.83 ± 0.58
T2FNorm	0.90 ± 0.01 / 0.15 ± 0.00 / $\mathbf{6.01} \pm \mathbf{0.18}$

in the response of the network towards OOD and ID. Similar observations can be found on logits as the feature representation has a direct implication on it. More importantly, from the comparison of various methods, we observe that the separability factor $\mathcal{S}$ induced by our method is highly significant. For instance, we achieve ($\mathcal{S} = 6.01$) at the end of training in the penultimate feature. The progression of S over the



(a) Baseline

(b) LogitNorm

(c) T2FNorm

Figure 6. Distribution of Maximum Softmax Probability (MSP) shows that overconfidence is controlled in both T2FNorm and LogitNorm. Overlapping region is also reduced in comparison to the baseline.

epochs in both the feature and logit space can be observed from Figure 7.

Compatibility with existing OOD scoring methods T2FNorm is compatible with various existing OOD scoring functions. Figure 5 shows that existing scoring functions when applied to the model trained with T2FNorm can boost the OOD detection performance. For instance, our model improves the baseline's OOD performance using ODIN from FPR@95 of 38.67 to 17.15 in ResNet18 architecture for CIFAR-10 experiments. Hyperparameter-free energy-based scoring function can also get a boost of 19.86 in comparison to the baseline model. Similarly, DICE [31] exhibits higher compatibility (Figure 4) with our method. Additional comparative results with ReAct and Activation Shaping are presented in Table 3. It further demonstrates the performance-enhancing capability of T2FNorm regularization.

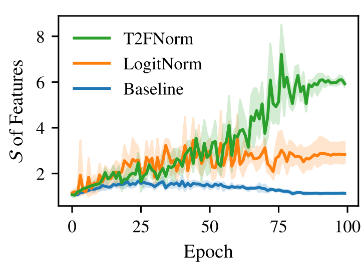


Figure 7. Progression of S at feature space with epochs

Qualitative results The qualitative results are presented in the form of feature activation of OOD samples. The ideal property of OOD detector is to have absolute zero feature activation for OOD samples. We observe the feature activation is minimum in T2FNorm in comparison to others from Figure 8. It further demonstrates the superiority of T2FNorm qualitatively.

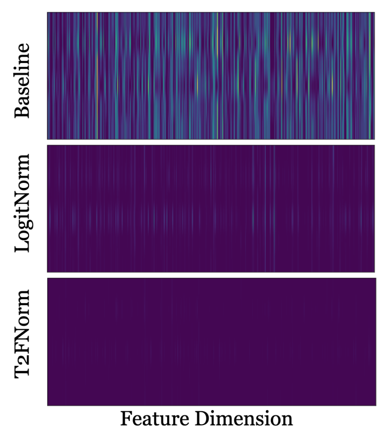


Figure 8. The figure shows qualitative feature activation of randomly sampled OOD images in (a) Baseline, (b) LogitNorm, and (c) T2FNorm. It can be observed that OOD features get activated significantly less in T2FNorm.

6. Discussion

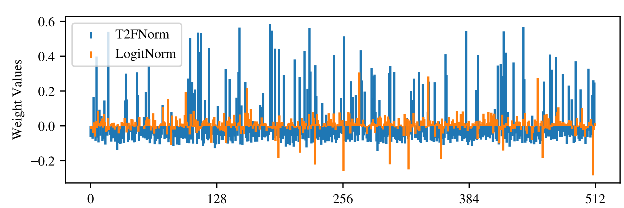


Figure 9. FC layer's weight comparison of Airplane class. Refer to the supplementary for FC layer's weight visualization of all classes.

Table 5. Mean of the FC Layer weights for a single class and for all classes shows that T2FNorm has more distinctly assigned weights

	Mean weights of Airplane Class			Mean weights of All Class		
Method	All Weights	Negative Weights	Positive Weights	All Weights	Negative Weights	Positive Weights
Baseline	0.000	-0.062	0.102	0.000	-0.056	0.107
LogitNorm	0.007	-0.042	0.027	-0.003	-0.027	0.032
T2FNorm	0.005	**-0.075**	**0.261**	0.000	**-0.072**	**0.283**

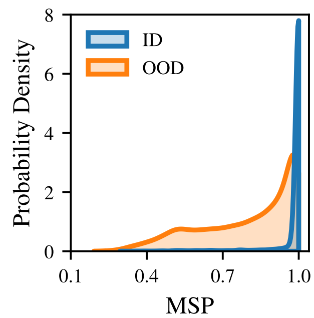


Figure 10. Normalization at OOD scoring

Ablation Study of Normalization As demonstrated in Figure 10, the separability of the nature of input distribution is compromised by normalization during OOD scoring. It results in the trained network incorrectly assuming OOD samples as ID samples. Quantitatively, for trained ResNet-18 architecture with CIFAR-10 as ID, this degrades the mean FPR@95 performance from 19.7% (T2FNorm) to 48.66%.

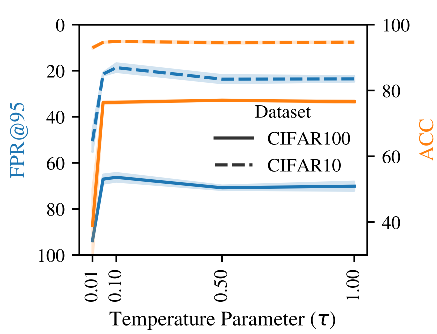


Figure 11. Sensitivity study of temperature τ

Sensivity Study of Temperature τ Figure 11 shows that the classification accuracy and OOD Detection performance (FPR@95) are not much sensitive over a reasonable range of τ. We found the optimal value of τ to be 0.1. And while the performance is good for $\tau \in [0.05, 1]$, both accuracy and FPR@95 score degrades substantially for $\tau > 1$ and $\tau \leq 0.01$.

Implication on FC Layer Weights Figure 9 compares FC layer weights for the "Airplane" class in T2FNorm and LogitNorm. LogitNorm weights show smoother distribution, while T2FNorm weights are more sharper. Quantitatively, T2FNorm has about 10 times higher average variance than LogitNorm. This suggests T2FNorm enforces distinct assignment of important features for category classification, activating such features for ID sample predictions. OOD samples lacking these features fail to activate them, yielding lower softmax probabilities. Table 5 reinforces this, showing T2FNorm with greater mean magnitudes for both negative and positive weights, emphasizing distinct feature assignments.

7. Conclusion

In summary, our work introduces **T2FNorm**, a selective mechanism of adoption and avoidance of normalization, which seeks to mitigate the challenge of overconfidence via enhancing ID/OOD separability. We empirically show that T2FNorm achieves a higher separability ratio than prior works. This study delves into the utility of feature normalization to accomplish this objective. Notably, we apply feature normalization exclusively during the training and inference phases, deliberately omitting its application during the OOD scoring process. This strategy improves OOD performance across a broad range of downstream OOD scoring metrics without impacting the model's overall accuracy. We provide empirical evidence demonstrating the versatility of our method, establishing its effectiveness across multiple architectures and datasets. We also empirically show our method is less sensitive to the hyperparameters.

8. Acknowledgement

This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z]; Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; The Royal Academy of Engineering Chair in Emerging Technologies scheme; and the EndoMapper project by Horizon 2020 FET (GA 863146).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[2] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3

[4] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012. 5

[5] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 5

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 3

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2, 5

[8] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2

[9] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022. 2

[10] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 1, 4, 5

[11] Conor Igoe, Youngseog Chung, Ian Char, and Jeff Schneider. How useful are gradients for ood detection really? *arXiv preprint arXiv:2205.10439*, 2022. 4

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on Machine Learning*, pages 448–456. pmlr, 2015. 2

[13] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022. 2

[14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. (0), 2009. 5

[15] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009. 5

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 5

[17] Gustaf Kylberg. Kylberg texture dataset v. 1.0. Centre for Image Analysis, Swedish University of Agricultural Sciences and . . . , 2011. 5

[18] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 2

[19] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2

[20] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 5

[21] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 3

[22] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 2, 5

[23] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR, 2022. 2

[24] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[25] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, 2015. 2

[26] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 3

[27] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 2

[28] Sudarshan Regmi, Bibek Panthi, Yifei Ming, Prashnna Kumar Gyawali, Danail Stoyanov, and Binod Bhattarai. Reweightood: Loss reweighting for distance-based ood detection. 2023. 2

[29] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020. 2

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 2

[31] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022. 2, 5, 6, 7

[32] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 2, 5, 1

[33] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1, 5

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[35] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in Neural Information Processing Systems*, 33:11839–11852, 2020. 1, 4

[36] Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian conference on computer vision*, 2020. 3

[37] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 3

[38] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 3

[39] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 2

[40] Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*, pages 23615–23630. PMLR, 2022. 2

[41] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022. 1, 2, 5

[42] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[43] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[44] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 5

[45] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022. 5

[46] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[47] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019. 2

[48] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5