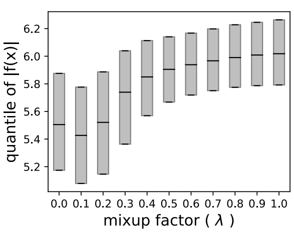
1. More Experiments for Cosine-based Softmax and Feature Norm

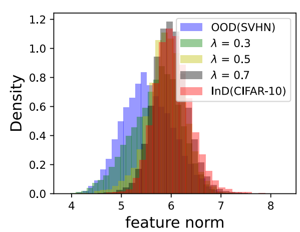
We present more experimental results to support our observation that the feature norms of OOD samples tends to decrease when the cosine-based softmax is used. We experiment with 2 models and 3 datasets. The models are RestNet34, DenseNet3 and the datasets are CIFAR-10, CIFAR-100, and SVHN. We use SVHN as the OOD dataset for CIFAR-10, and CIFAR-10 for SVHN. We also observe the feature norms of mixup samples by changing the mixup factor, λ, as we did in Section 3.1.

Various experiments also support our observation. When models are trained with the standard softmax, the average of feature norms is not always monotonically increase as shown (a), (c) and (e) of Figures 1 and 3. In the figures, the top of each bar represents the 1-quantile, and the bottom represents the 3-quantile of feature norms of mixup images. We can notify that the bar when $\lambda = 0$ overlaps much with the bar when $\lambda = 1$, which mean that the distributions of OOD samples and normal samples overlaps much. We can also verify this from (b), (d), and (f) of Figures 1 and 3. The distribution of feature norms of OOD samples and that of normal samples overlap much each other. It means that the feature norms by the standard softmax are hardly useful for OOD detection.

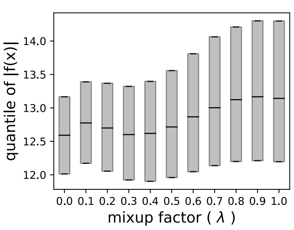
However, the models are trained with the cosine-based softmax, the average of feature norms is monotonically increase as shown in (a), (c) and (e) of Figures 2 and 4. The overlaps between distributions of OOD samples and normal samples are small as shown in (b), (d), and (f) of Figures 2 and 4, which means that the feature norms by the cosine-based softmax are useful for OOD detection.



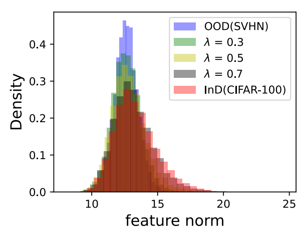
(a) Quantiles (CIFAR-10)



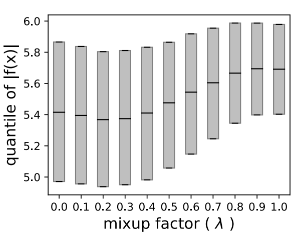
(b) Dist. (CIFAR-10)



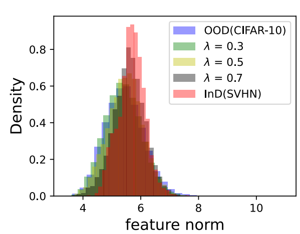
(c) Quantiles (CIFAR-100)



(d) Dist. (CIFAR-100)

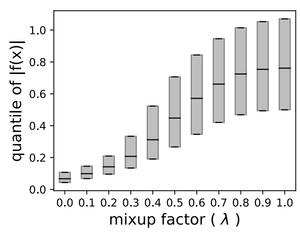


(e) Quantiles (SVHN)

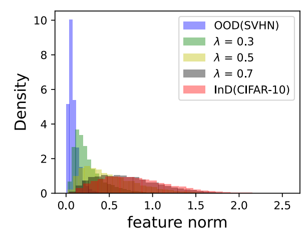


(f) Dist. (SVHN)

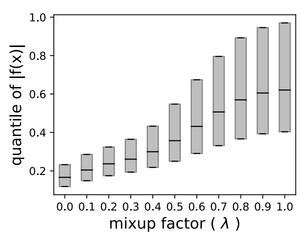
Figure 1: The quantile value of feature norms (a, c, e) and the distribution of feature norms (b, d, f) according to the mixup factor. (a) and (b) are for CIFAR-10 (InD) vs. SVHN (OOD), (c) and (d) are for CIFAR-100 vs. SVHN, and (e) and (f) are for SVHN vs CIFAR-10. The model is ResNet34 trained with the standard softmax.



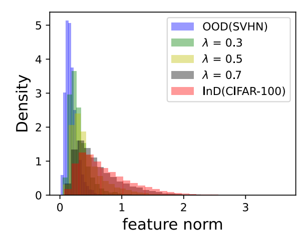
(a) Quantiles (CIFAR-10)



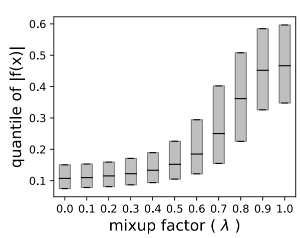
(b) Dist. (CIFAR-10)



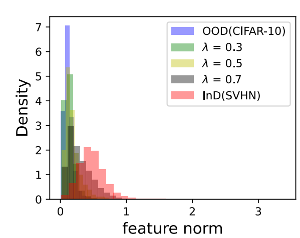
(c) Quantiles (CIFAR-100)



(d) Dist. (CIFAR-100)

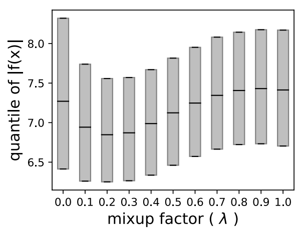


(e) Quantiles (SVHN)

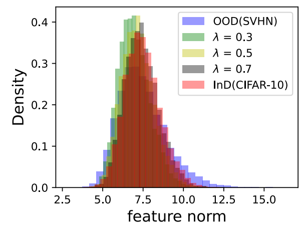


(f) Dist. (SVHN)

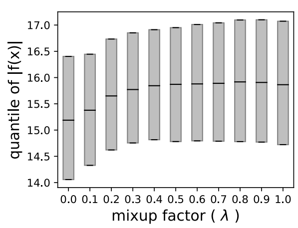
Figure 2: The quantile value of feature norms (a, c, e) and the distribution of feature norms (b, d, f). The model is ResNet34 trained with the cosine-based softmax.



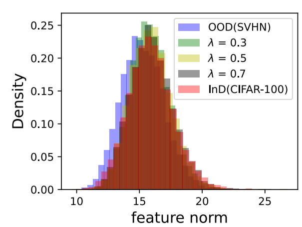
(a) Quantiles (CIFAR-10)



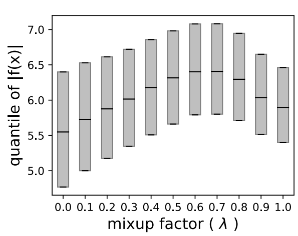
(b) Dist. (CIFAR-10)



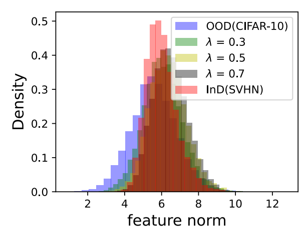
(c) Quantiles (CIFAR-100)



(d) Dist. (CIFAR-100)

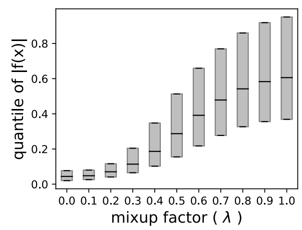


(e) Quantiles (SVHN)

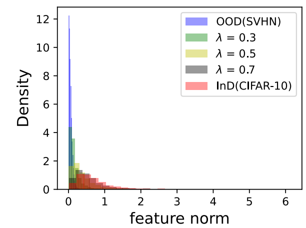


(f) Dist. (SVHN)

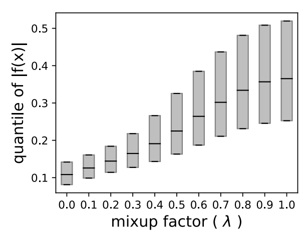
Figure 3: The quantile value of feature norms (a, c, e) and the distribution of feature norms (b, d, f). The model is DenseNet3 trained with the standard softmax.



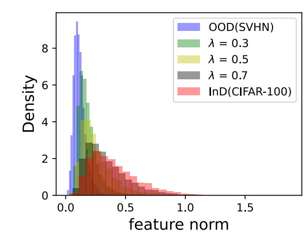
(a) Quantiles (CIFAR-10)



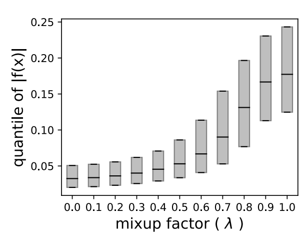
(b) Dist. (CIFAR-10)



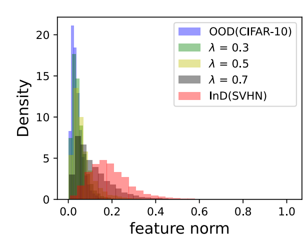
(c) Quantiles (CIFAR-100)



(d) Dist. (CIFAR-100)



(e) Quantiles (SVHN)



(f) Dist. (SVHN)

Figure 4: The quantile value of feature norms (a, c, e) and the distribution of feature norms (b, d, f). The model is DenseNet3 trained with the cosine-based softmax.

2. Performance Evaluation with ResNet34

OOD	AUROC					TNR@95				
	Use a training dataset					Use a training dataset				
	Use InD validation samples					Use InD validation samples				
	Perform the input processing					Perform the input processing				
	Use OOD samples				(Ours)	Use OOD samples				(Ours)
	ODIN	MAHA	G-ODIN	GRAM	COD	ODIN	MAHA	G-ODIN	GRAM	COD
InD : CIFAR-10										
SVHN	96.7	99.1	97.8	**99.5**	99.4	70.3	87.8	89.5	**97.6**	97.1
TINc	93.1	98.6	96.0	99.2	**99.7**	68.7	92.0	81.1	96.7	**99.0**
TINr	94.0	99.5	96.1	**99.7**	**99.7**	67.9	97.1	81.4	**98.7**	98.4
LSUNc	91.2	96.7	97.2	97.8	**99.9**	92.0	81.3	87.3	89.8	**99.8**
LSUNr	94.1	99.7	98.0	**99.9**	99.8	82.1	98.8	90.9	**99.6**	99.3
iSUN	94.0	99.5	97.6	**99.8**	99.7	73.2	97.8	88.8	**99.3**	99.0
MEAN	93.9	98.9	97.1	99.3	**99.7**	70.7	92.5	86.5	97.0	**98.8**
InD : CIFAR-100										
SVHN	93.9	**98.4**	93.2	96.0	96.0	62.7	**91.9**	55.1	80.8	77.3
TINc	85.4	96.3	95.3	97.7	**98.9**	44.3	80.9	72.6	88.5	**96.2**
TINr	87.6	98.2	95.9	**98.9**	98.7	36.1	90.9	76.5	**94.8**	93.2
LSUNc	82.7	92.0	93.8	92.1	**99.4**	44.1	64.8	95.7	64.8	**97.9**
LSUNr	85.6	98.2	96.1	**99.2**	98.8	23.2	90.9	76.8	**96.6**	94.6
iSUN	85.5	97.9	95.7	**98.8**	98.5	45.2	89.9	75.3	**94.8**	92.5
MEAN	86.8	96.8	95.0	97.1	**98.4**	42.6	84.9	75.3	86.7	**92.0**
InD : SVHN										
CIFAR-10	92.1	99.3	-	97.3	**99.4**	79.8	**98.4**	-	85.8	98.3
CIFAR-100	-	-	-	-	**99.4**	-	-	-	-	**98.1**
TINc	-	-	-	-	**100.0**	-	-	-	-	**100.0**
TINr	92.0	**99.9**	-	99.7	**99.9**	82.0	**99.9**		99.3	99.8
LSUNc	-	-	-	-	**100.0**	-	-	-	-	**100.0**
LSUNr	89.4	**99.9**	-	99.8	**99.9**	77.3	**99.9**	-	99.6	99.8
iSUN	91.4	99.8	-	99.8	**99.9**	79.1	99.7	-	99.4	**99.8**
MEAN	-	-	-	-	**99.8**	-	-	-	-	**99.4**

Table 1: OOD detection performance (%) with a ResNet34 classification model. Our performance is averaged over 5 runs. In the case of SVHN, all baselines did not present their performances. Dash symbols stand for not-available. The mean rows are the averages of the performances over all OOD datasets. For each case, the best results are in bold.

3. OOD Detection with Fine-tuned Models (ResNet34)

InD	OOD	AUROC	TNR95
CIFAR-10 (Acc: 95.2 → 94.8)	SVHN	98.80	94.61
	TINc	99.71	99.71
	TINr	99.56	98.23
	LSUNc	99.89	99.97
	LSUNr	99.70	99.00
	iSUN	99.57	98.26
CIFAR-100 (Acc: 77.7 → 75.2)	SVHN	95.28	71.98
	TINc	99.36	99.89
	TINr	98.58	92.71
	LSUNc	99.60	99.98
	LSUNr	98.97	95.49
	iSUN	98.53	92.24
SVHN (Acc: 96.5 → 96.4)	CIFAR-10	98.96	97.04
	CIFAR-100	99.03	97.26
	TINc	99.99	100.00
	TINr	99.80	99.82
	LSUNc	99.99	100.00
	LSUNr	99.79	99.93
	iSUN	99.79	99.92

Table 2: Classification and OOD detection performance of fine-tuned ResNet34 models. All figures are in %, and are the average results of 5 experiments each.