

SUPPLEMENTARY MATERIALS FOR COMPLEMENTARITY MATTERS: A CLOSER LOOK AT NEAREST NEIGHBOR GUIDANCE FOR OOD DETECTION

Anonymous authors

Paper under double-blind review

1 APPENDIX

1.1 IMPLEMENTATION DETAILS

In this section, we clarify the details about the implementation of our methodology CoNNGuide for reproductions, and the code for the methods is also included in the zip file.

1.1.1 ACTIVATION CLIPPING.

Sun et al. Sun et al. (2021) suggested that activation clipping to feature vectors can increase the score gap between data in distribution (ID) and Out-Of-Distribution (OOD) data to facilitate OOD detection. Hence, we initially clip the feature vectors (i.e., penultimate layer activations) with a threshold δ_a determined as the value ranked at p_{clip} percentile of the activations of the training set examples \mathbf{H}_{train} . The clipped feature vectors $\hat{\mathbf{h}}(x) = [\hat{a}_1, \hat{a}_2, \dots, \hat{a}_d]$. This process is illustrated in Equation 1.

$$\begin{aligned} \delta_a &= percentile(\mathbf{H}_{train}, p_{clip}) \\ \hat{a}_i &= min(a_i, \delta_a) \end{aligned} \quad (1)$$

The logits and probabilities computed with clipped feature vectors are noted as $\hat{f}_\theta(x) = \mathbf{W}^T \hat{\mathbf{h}}(x) + \mathbf{b}$ and $\hat{p}_\theta(x) = softmax(\hat{f}_\theta(x))$.

1.1.2 ACTIVATION AND WEIGHT PRUNING.

We follow the strategy proposed in Ahn et al. (2023) to apply the activation and weight pruning for removing noises and obtaining a better energy score. As described in the main paper, we replace the way of identifying the important activations (i.e., neurons) with the P-Shapley indices; we recall the equation for the P-Shapley indices in Equation 2:

$$\begin{aligned} s_i^{(l)} &= \left| \hat{a}_i \cdot \frac{\partial(\hat{f}_\theta(x^{(l)}) \cdot \hat{p}_\theta(x^{(l)}))}{\partial \hat{a}_i} \right| \\ &= |\hat{p}_l(x^l) \hat{f}_l(x^l) - \hat{p}_l(x^l; \hat{a}_i \rightarrow 0) \hat{f}_l(x^l; \hat{a}_i \rightarrow 0)| \end{aligned} \quad (2)$$

Where \hat{f}_l and \hat{p}_l represent the logit and probability for the class l for which we compute the activation importance of each neuron. The important activations are then identified through an activation pruning rate p_n that keeps only the activations having a significance above the value at p_n percentile position (i.e., the mask $\mathbf{m}^{(l)}$ mentioned in the section ‘‘Preliminaries’’ of the paper).

Besides this modification, we also change the original method for evaluating the weight contribution matrix \mathbf{C} to the method proposed in Sun & Li (2022), which gives more precise weight contribution identification based on our experiments. The contributions are only evaluated for the identified important activations, and the weights related to other neurons receive a direct contribution of 0. With the computed contributions, we determine a mask \mathbf{M}_l with a weight pruning rate p_w for utilizing only the important weights, which help us to obtain the final pruned weights $\hat{\mathbf{W}}$ for the logit evaluation (i.e., $\hat{\mathbf{W}} = \mathbf{M}_l \odot \mathbf{W}$). The logits computed with the activation clipping and the importance pruning mentioned in this step are used for the energy score computation and denoted as $\hat{f}_{CoNNGuide}(x)$.

1.1.3 MULTI-GUIDANCE DETAILS.

For CoNNGuide, we use the multi-guidance from the last two layers (i.e., $\{G_{n-1}, G_n\}$), but there are some specific details for our way of applying guidance. Firstly, we remove the implication of the base score $S_{base}(x)$ in the similarity computation and rescale the cosine similarity values into the interval of $[0, 1]$. The equation describing the exact guidance we use is Equation 3.

$$G_i(\mathbf{x}) = \frac{1}{k} \sum_{j=1}^k \frac{1 + \cos(z_i^{(j)}, z_i)}{2} \quad (3)$$

We use the original penultimate layer activations for G_{n-1} to obtain the normalized feature vectors (i.e., without activation clipping). For the logit guidance term (i.e., G_n), we use the logits that are computed differently from the ones used for energy scores, because we would like to keep more information from the original logits for the Nearest-Neighbor (NN) search rather than just increasing the difference between ID and OOD examples. We introduce a new sensitivity index named “entropy index” that measures the global importance of the neurons across different classes. The entropy index defines each neuron’s significance as the Shannon entropy difference that we obtain when setting its value to 0, as demonstrated in Equation 4.

$$s_i^x = |\text{entropy}(p(x)) - \text{entropy}(p(x; a_i \rightarrow 0))| \quad (4)$$

We chose to use global significance over identified neuron importance for each class because we aim to retain as much information as possible across all classes in the logits, rather than focusing solely on the essential information within one class for more precise similarity guidance.

The final logits used for NN guidance are computed with activation clipping and global activation and weight pruning, utilizing entropy indices and weight contributions as described in Sun & Li (2022). The masks used for these pruning operations are denoted as \mathbf{M} and \mathbf{m} . We denote the pruning rates used in this step as p_n^{lo} and p_w^{lo} . In other words, the logits used for multi-guidance (i.e., G_n) are $f_{lo}(x) = (\mathbf{M} \odot \mathbf{W})^T (\mathbf{m} \odot \hat{h}(x)) + \mathbf{b}$.

1.1.4 CONNGUIDE SCORE.

To summarize, the final score for the OOD detection is the energy score Liu et al. (2020) (i.e., $\text{energy}(f(x)) = \log \sum_i^K e^{f_i(x)}$) with multi-guidance from the last two layers, as described in Equation 5.

$$\begin{aligned} S_{base}^{co} &= \text{energy}(f_{\text{CoNNGuide}}(x)) \\ S_{\text{CoNNGuide}} &= S_{base}^{co} \cdot (G_{n-1} \cdot G_n) \end{aligned} \quad (5)$$

1.1.5 NNGUIDE++ SCORE.

As described in the paper, the strong baseline NNGuide++ we built uses the same pipeline as CoNNGuide but without P-Shapley indices (i.e., using the original Shapley indices) and the guidance from the last layer (i.e., G_n) layers. The logit computed with this pipeline is noted as $f_{\text{NNGuide++}}(x)$, and the whole equation for the NNGuide++ score is displayed in Equation 6.

$$\begin{aligned} S_{base}^{NN} &= \text{energy}(f_{\text{NNGuide++}}(x)) \\ S_{\text{NNGuide++}} &= S_{base}^{NN} \cdot G_{n-1} \end{aligned} \quad (6)$$

1.2 HYPERPARAMETER TUNING

CoNNGuide contains 5 pruning rate hyperparameters which should be adjusted for the optimal performance (i.e., p_a, p_n, p_w, p_n^{lo} and p_w^{lo}). We report the best values of these hyperparameters for different datasets and models below:

1.2.1 COMMON VALUE LISTS FOR ALL THE DATASETS

For p_n and p_w that are used for pruning activations and weights to obtain a more separable energy score, we use the same potential value list for both hyperparameters on all the datasets (i.e., CIFAR10, CIFAR100 and Imagenet), which is $[0, 10, 20, 30, 40, 50, 60, 70, 80, 90]$. This potential value list allows for complete tuning for pruning and is suggested in Ahn et al. (2023).

The value list for p_a is also the same for three datasets; the values we considered are [80, 85, 90, 95, 100] because we want to avoid the extreme values and not to eliminate too much information from the activation levels according to the statements in Ahn et al. (2023). We adjust only a small amount for the tested activation clipping rate on CIFAR10, and the best point is found at 96%.

1.2.2 SPECIFIC VALUE LISTS FOR CIFAR DATASETS.

For the remaining two pruning rates p_n^{lo} and p_w^{lo} , the CIFAR datasets consider the full potential pruning values for p_n (i.e., [0, 10, 20, 30, 40, 50, 60, 70, 80, 90]) and small pruning rates for p_w (i.e., [0, 1, 5, 10]). The p_w considers these pruning rates because a wide-range pruning of the weights will cause too much logit information to be lost and is not beneficial for the KNN search.

1.2.3 SPECIFIC VALUE LISTS FOR IMAGENET DATASET.

The Imagenet dataset uses different potential value lists for p_n^{lo} and p_w^{lo} because the logits from models trained on this dataset are more sensitive to value changes, since the vision problem is complex and large-scale. We aim to remove only a small amount of unnecessary noise from weights and activations. Hence, the pruning rates we considered for these two hyperparameters of the logit guidance are [0, 1, 5, 10] on Imagenet.

1.2.4 BEST VALUES FOR THE EVALUATED DATASETS AND MODELS.

We display the best hyperparameter values we found for different models and datasets, along with their corresponding performance, in Table 1.

Table 1: Best pruning rates for different datasets and models.

Method	Dataset	Model	p_a	p_n	p_w	p_n^{lo}	p_w^{lo}	FPR95 ↓	AUROC ↑
NNGuide++	CIFAR10	DenseNet-101	96	90	50	N/A	N/A	13.48	97.4
NNGuide++	CIFAR100	DenseNet-101	85	70	60	N/A	N/A	28.58	90.87
NNGuide++	Imagenet	ResNet50	85	20	10	N/A	N/A	22.53	94.89
NNGuide++	Imagenet	RegNet	85	0	10	N/A	N/A	14.9	96.67
CoNNGuide	CIFAR10	DenseNet-101	96	80	70	90	1	12.77	97.45
CoNNGuide	CIFAR100	DenseNet-101	85	70	50	60	0	26.01	92
CoNNGuide	Imagenet	ResNet50	85	30	40	0	0	21.68	95.12
CoNNGuide	Imagenet	RegNet	85	0	90	10	1	14.15	96.93

1.3 DETAILED OOD PERFORMANCE FOR CONNGUIDE AND NNGUIDE++

This section provides the detailed OOD detection performance for the OOD detection problems we considered in the main paper. The results related to the CIFAR datasets are displayed in Table 2 and 3, and the ones for the Imagenet datasets with ResNet and RegNet are presented in Table 4 and 5.

Table 2: OOD detection performance on CIFAR10 with DenseNet-101. The OOD sets include SVHN, Textures, Places, iSUN, LSUN-Crop, and LSUN-Resize.

Method	SVHN		Textures		Places		LSUN-Crop		LSUN-Resize		iSUN		Avg	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	47.27	93.48	64.27	88.13	62	88.95	33.53	95.54	42	94.55	42.5	94.49	48.6	92.52
MaxLogit	39.52	94.13	55.98	86.5	40.15	91.97	4.19	99.09	9.42	98.12	10.08	98.05	26.56	94.65
KL	40.51	94.01	56.29	87.04	39.51	92.04	3.8	99.15	8.72	98.19	9.45	98.11	26.38	94.76
Energy	40.51	93.99	56.29	86.41	39.51	92.01	3.8	99.15	8.72	98.19	9.45	98.11	26.38	94.65
ReAct	31.99	95.29	44.1	91.96	38.6	92.61	6.27	98.8	7.87	98.35	8.62	98.2	22.9	95.87
Mahalanobis	9.3	98	22.64	94.52	63.78	84.76	0.32	99.85	4.07	99.05	5.77	98.84	17.65	95.84
DICE	29.91	94.58	45.76	86.98	44.89	90.33	0.38	99.9	4.2	99.07	5.12	98.99	21.71	94.97
KNN	4.05	99.25	19.59	96.39	46.03	90.1	6.92	98.75	9.91	98.13	10.24	98.21	16.12	96.8
DICE + ReAct	13.85	97.39	25.55	94.65	48.38	90.23	0.47	99.89	3.62	99.22	4.81	99.08	16.11	96.74
NNGuide	26.48	95.68	39.2	92.74	36.57	93	12.75	97.69	9.28	98.23	9.9	98.06	22.36	95.9
LiNe	10.9	97.92	22.38	95.65	45.06	91.18	0.75	99.82	4.03	99.15	5.03	99.05	14.7	97.13
NNGuide++	8.95	98.33	20.05	96.19	43.12	91.69	0.76	99.81	3.54	99.25	4.47	99.15	13.48	97.4
CoNNGuide (Ours)	10.94	97.89	20.9	95.8	38.46	92.38	0.58	99.84	2.67	99.41	3.09	99.33	12.77	97.45

Table 3: OOD detection performance on CIFAR100 with DenseNet-101. The OOD sets include SVHN, Textures, Places, iSUN, LSUN-Crop, and LSUN-Resize.

Method	SVHN		Textures		Places		LSUN-Crop		LSUN-Resize		iSUN		Avg	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	81.82	75.4	84.77	71.41	82.11	74.94	60.49	85.6	87.3	67.14	88.03	68.49	80.75	73.83
MaxLogit	85.42	81.95	83.48	71.19	77.67	78.06	16.92	97.06	76.07	77.42	79.13	76.55	69.78	80.37
KL	87.53	81.9	84.33	71.95	78.26	77.98	14.78	97.43	75.36	77.99	78.69	77.08	69.82	80.72
Energy	87.55	81.85	84.36	71.02	78.26	77.96	14.78	97.43	75.37	77.77	78.7	76.8	69.84	80.47
ReAct	82.23	82.82	78.55	78.74	79.67	76.44	18.54	96.54	64.84	85.53	68.73	84.54	65.42	84.1
Mahalanobis	38.7	92.82	23	94.45	87.71	70.9	4.25	99.03	22.01	96.11	20.74	96.42	32.73	91.62
DICE	59.36	88.57	61.45	77.12	80.41	77.04	0.91	99.74	54.98	88.25	52.45	88.52	51.59	86.54
KNN	17.44	96.41	24.18	93.73	93.03	59.95	31.46	92.85	47.33	90.42	39.64	91.91	42.18	87.54
DICE + ReAct	43.9	90.65	40.12	86.68	91.66	62.65	4.63	99.1	48.1	91.2	38.73	92.8	44.52	87.18
NNGuide	81.15	81.56	68.1	78.75	92.56	61.86	62.03	81.38	32.04	94.18	37.06	91.94	62.16	81.61
LiNe	25.07	93.56	34.04	89.48	89.17	62.33	5.32	98.9	17.92	95.56	18.75	95.45	31.71	89.21
NNGuide++	18	95.61	28.37	92.18	86.63	66.02	4.03	99.12	17.52	96.13	16.91	96.16	28.58	90.87
CoNNGuide (Ours)	15.96	96.18	26.49	92.86	85.29	69.06	4.23	99.17	11.33	97.45	12.77	97.3	26.01	92

Table 4: OOD detection performance on ImageNet-1k with ResNet50. The “curated OODs” are the datasets of Textures, iNaturalist, Places, and SUN.

Method	Textures		iNaturalist		Places		SUN		OpenImage-O		Avg curated OODs		Avg all	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	66.28	80.43	52.86	88.39	72.08	80.53	69.13	81.64	66.99	83.89	65.09	82.75	65.47	82.98
MaxLogit	54.93	86.39	50.79	91.15	66.07	84.03	60.41	86.44	63.95	87.38	58.05	87	59.23	87.08
KL	52.45	86.72	53.79	90.62	66.03	83.96	58.84	86.57	64.72	87.08	57.78	86.97	59.17	86.99
Energy	52.45	86.72	53.79	90.62	66.03	83.96	58.84	86.57	64.72	87.08	57.78	86.97	59.17	86.99
DICE	32.52	90.43	26.26	94.56	47.75	87.69	36.25	90.95	54.52	85.67	35.69	90.91	39.46	89.86
Mahalanobis	15.02	95.52	35.04	94.79	70.31	83.92	64.99	86.55	37.52	93.89	46.34	90.19	44.58	90.93
KNN	10.78	97.49	58.32	86.16	77.06	75.66	67.96	81.43	63.62	82.94	53.53	85.18	55.55	84.74
NNGuide	24.93	91.52	12.02	97.47	38.88	90.12	31.62	91.66	31.6	93.66	26.86	92.69	27.81	92.89
LiNe (reproduced)	17.73	95.5	12.05	97.63	31.42	92.11	21.26	94.91	43.99	87.42	20.62	95.04	25.29	93.51
NNGuide++	11.81	97.27	10.27	97.98	29.73	93.41	19.93	95.73	40.9	90.06	17.93	96.1	22.53	94.89
CoNNGuide (Ours)	13.05	97.2	10.85	97.93	25.38	94.23	16.16	96.36	42.97	89.89	16.36	96.43	21.68	95.12

Table 5: OOD detection performance on ImageNet-1k with RegNet. The “curated OODs” are the datasets of Textures, iNaturalist, Places, and SUN.

Method	Textures		iNaturalist		Places		SUN		OpenImage-O		Avg curated OODs		Avg all	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	49.13	86.55	23.64	94.66	56.86	85.2	52.54	86.65	33.9	92.1	45.54	88.27	43.22	89.03
MaxLogit	32.27	91.29	7.73	98.04	40.8	88.21	31.67	91.63	15.92	95.94	28.12	92.29	25.68	93.02
KL	30.66	91.61	6.6	98.3	40.51	88.04	29.44	91.93	15.3	96.12	26.8	92.47	24.5	93.2
Energy	30.67	91.59	6.67	98.29	40.34	88.11	29.38	91.97	15.36	96.1	26.77	92.49	24.48	93.21
DICE	30.67	91.59	6.67	98.29	40.34	88.11	29.38	91.97	15.36	96.1	26.77	92.49	24.48	93.21
Mahalanobis	27.91	93.9	2.22	99.36	61.84	85.77	49.3	89.85	19.5	96.48	35.32	92.22	32.15	93.07
KNN	18.94	94.62	3.48	99.09	46.24	87.5	33.74	91.09	16.4	96.45	25.6	93.08	23.76	93.75
NNGuide	17	95.82	1.83	99.57	31.47	91.87	21.58	94.43	10.79	97.73	17.97	95.42	16.53	95.89
LiNe (reproduced)	30.05	91.68	6.77	98.25	40.13	88.14	29.35	91.93	14.93	96.14	26.58	92.5	24.25	93.23
NNGuide++	7.11	98.58	1.32	99.65	32.65	92.34	23.08	94.89	10.36	97.91	16.04	96.36	14.9	96.67
CoNNGuide (Ours)	7.09	98.55	0.89	99.78	30.27	93.25	21.15	95.35	11.34	97.71	14.85	96.74	14.15	96.93

REFERENCES

- Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. Line: Out-of-distribution detection by leveraging important neurons. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19852–19862, 2023. doi: 10.1109/CVPR52729.2023.01901.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21464–21475, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf.
- Yiyao Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 691–708, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_40. URL https://doi.org/10.1007/978-3-031-20053-3_40.
- Yiyao Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=IBVBtz_sRSm.