Activation Subspaces for Out-of-Distribution Detection

Supplementary Material

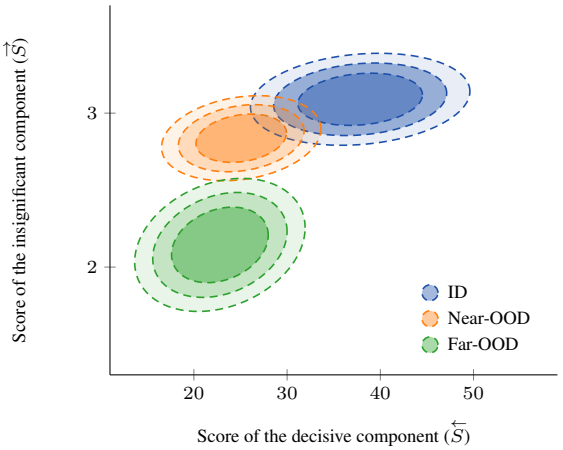

Figure 5. *Score distributions of ID and OOD data.* We fit multivariate Gaussians to the concatenated (along a new second dimension) score values from the decisive ($\overleftarrow{S}$) and insignificant ($\vec{S}$) components of the sampled ID (ImageNet-1k [9, 44]), Near-OOD (NINCO [3]), and Far-OOD (Textures [7]) data. We use ResNet-50 [16]. Score values are negated to visualize in the positive domain.

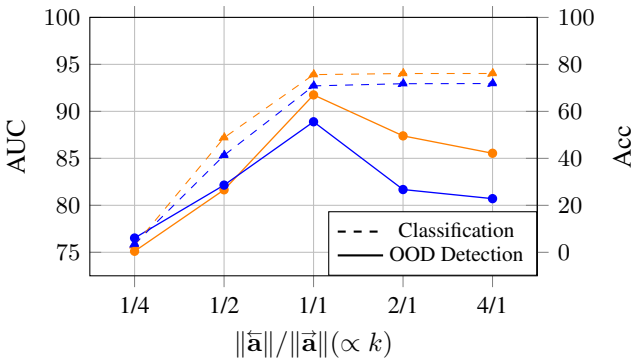

Figure 6. *Classification and OOD detection accuracy (%, ↑) for different ratios of the norms of decisive and insignificant components with ResNet (orange) and MobileNet (blue).* The AUC is calculated by the average of datasets in OpenOOD [63], and ID is ImageNet-1k [9, 44].

A. Additional Experiments

We extend our experiments in the setting of Sec. 4.3. We use the OpenOOD [63] benchmark and report Near-OOD and Far-OOD results separately.

ViT. Tab. 8 reports the accuracy of our method ActSub with ViT-B/16 [12], ResNet-50, and their average. Since activation shaping methods (including SCALE) are known to perform worse on ViTs [11, 51, 57], we additionally combine ActSub with GEN [37] on the decisive component. For ViT, we use the prototype variant of ID data (see below: "Dependence on the data volume") for the insignificant component, which we empirically found to work better. For ViT, ActSub improves GEN and SCALE [57], particularly in Near-OOD, improving the AUC by 2.15 %/18.09 % and reducing the FPR by 6.68 %/28.10 %, respectively. Remarkably, when considering the average of ResNet and ViT, our variant with SCALE reaches SotA accuracy.

Visualization of the score distributions. Fig. 5 visualizes the score distributions of our decisive ($\overleftarrow{S}$) and insignificant ($\vec{S}$) components for ID, Near-OOD, and Far-OOD data. For ID, Near-OOD, and Far-OOD, we use ImageNet-1k [9, 44], NINCO [3], and Textures [7], respectively. We observe that the Far-OOD data is better separated from the ID data with the score of the insignificant component. Inversely, Near-OOD and ID are better separated by the score of the decisive component. With this observation, we emphasize the complementary effect of our two score functions.

Choice of hyperparameter k. Ideally, the decisive subspace exclusively captures the classification signal while the insignificant subspace captures all information that does not aid classification. We plot the OOD-detection and ID-classification accuracy against the norm ratio of the two components – which is proportional to k – in Fig. 6. For ID, we use ImageNet-1k [9, 44]), and for OOD, we consider the average accuracy of the datasets in OpenOOD [63]. The ID-classification accuracy already saturates with a balanced decomposition where the norms of the two components are equal (1/1), corresponding to the ratio we propose in our experiments. Shifting the balance towards $\vec{\mathbf{a}}$ (smaller k), causes classification-related directions to be captured in the insignificant subspace, which has a detrimental effect on OOD detection (*cf*. Tab. 6). Inversely, shifting the balance towards $\overleftarrow{\mathbf{a}}$ (larger k) results in insignificant directions being captured by the decisive subspace, and consequently, increasing interference. To summarize, deviations in either direction from the 1/1 ratio reduce the OOD-detection accuracy, confirming our choice in Eq. (6).

Sensitivity of hyperparameter λ. The hyperparameter λ weights the information of the decisive component ($\overleftarrow{S}$) that excels in Near-OOD and the information of the insignificant component ($\vec{S}$) that excels in Far-OOD. In other words, λ balances the discriminative information of each subspace that targets a different aspect of the distribution shift. In practice, not only the scale but also the variation and the margin of each score distribution might vary. To account

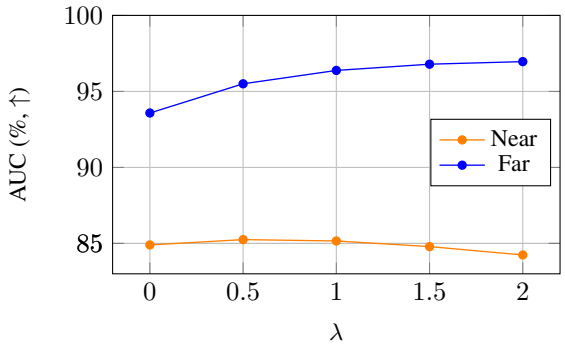

Figure 7. *Sensitivity of hyperparameter* λ. We report the AUC for our method ActSub ($\overleftrightarrow{S}$) for different λ values for Near-OOD and Far-OOD of the OpenOOD [63] benchmark. We use a ResNet-50 [16] model trained on ImageNet-1k [9, 44].

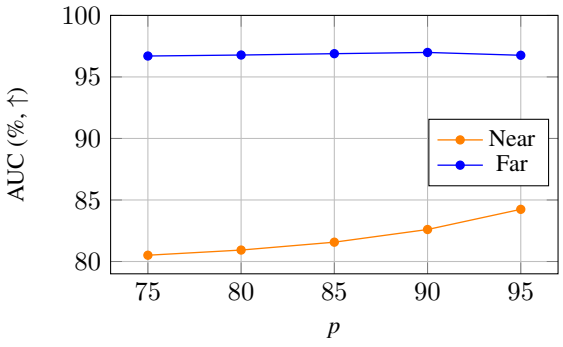

Figure 8. *Sensitivity of pruning percentage p of SCALE.* We report the AUC for our method ActSub ($\overleftrightarrow{S}$) for different percentages of pruning (p) for Near-OOD and Far-OOD from the OpenOOD [63] benchmark. We use a ResNet-50 [16] model trained on ImageNet-1k [9, 44].

for this, we define λ as an exponent (instead of a simple weighting factor) and ensure that the score function of each component contributes sufficiently to the final score ($\overleftrightarrow{S}$).

We investigate the effect of the hyperparameter λ on the accuracy of our method ActSub in Fig. 7. Note that $\lambda = 0$ is equivalent to only using the score of the decisive component. We observe that Near-OOD performance is relatively unaffected by λ. On the other hand, for Far-OOD, introducing the cosine similarity-based score of our insignificant component with increasing λ significantly increases the accuracy. We use $\lambda = 0.5, 1, 2$ for MobileNetV2 [45], ViT, and ResNet-50, respectively. For ViT with GEN [37], we use $\lambda = 0.5$. In the main experiments with ImageNet-1k [9, 44] (*cf*. Tabs. 1 and 2), we use NINCO [3] as the held-out validation set. For CIFAR [29], we use MNIST [30]. For OpenOOD [63] experiments, we use the OpenOOD validation split.

Sensitivity of the decisive component. We design the decisive component ($\overleftarrow{S}$) to capture the information related to the classification task, which we assume to be mostly semantic information. To reflect this, when tuning the hyperparameters of the decisive component, *i.e*., the hyperparameters of SCALE [57] or GEN [37], we incorporate ImageNet-R [19], which includes stylized renditions of ImageNet classes, along with ImageNet-1k as an ID dataset. By doing so, we introduce non-semantic variations to the ID data and guide the decisive component to focus on the semantic variations between ID and OOD distributions. We select the parameter that maximizes the difference between AUC and FPR.

In Fig. 8, we present how the pruning percentage p of the activation shaping method SCALE [57] affects the accuracy of our combined score function ($\overleftrightarrow{S}$). For Far-OOD, the effect of p is insignificant. However, for Near-OOD, the accuracy increases with increasing p. For the original SCALE, *i.e*., when applied on the whole activation, the accuracy has been shown to decrease after 85 % [57]. We believe this is because SCALE needs to keep more channels to capture discriminative information when applied to the whole representation. In contrast, when only applied to the decisive component ($\overleftarrow{S}$), SCALE benefits from higher pruning percentages – with the interference of the insignificant component being removed from the activation, a higher pruning percentage still captures the important channels for the model's prediction.

		Near-OOD (Avg.)		Far-OOD (Avg.)	
Volume	**Sampling**	AUC (↑)	FPR (↓)	AUC (↑)	FPR (↓)
10% (Tab. 4)	Random	84.24	52.60	96.96	14.29
1%	Random	83.33	54.99	96.69	15.62
1%	Averaging	83.48	54.75	96.90	14.50
0.1%	Clustering	83.73	54.41	96.85	14.77

Table 7. *Accuracy (in %) of our unified score function with different data volumes and sampling strategies.* We use ResNet-50 [16] trained on ImageNet-1k [9, 44], and evaluate on OpenOOD [63].

Dependence on the data volume. Tab. 7 shows the OOD detection accuracy of our method using different amounts of ID data used to compute the cosine similarity of the insignificant component. Compared to 10% used in our main method, reducing the data volume to 1% either by random sampling or sample averaging maintains strong accuracy. Furthermore, we evaluate a prototype approach, where each sample is a cluster center obtained by applying k-means to the insignificant component sampled from the training split. The number of prototypes corresponds to 0.1% of the training data volume. This shows that we can *(i)* significantly reduce memory requirements with a sparse set of representations (0.1% of the data volume), *(ii)* reduce privacy vulnerability, as reconstructing individual samples from cluster centers is harder, and *(iii)* retain the significant improvement over previous methods (*cf*. Tab. 4).

Method	ViT-B/16		ResNet-50		Avg. of ResNet & ViT	
	Near	Far	Near	Far	Near	Far
RMDS [43]	**80.09** / 65.36	92.60 / **28.76**	76.99 / 65.04	86.38 / 40.91	78.54 / 65.20	89.49 / 34.83
ViM [52]	77.03 / 73.73	**92.84** / 29.18	72.08 / 71.35	92.68 / 24.67	74.55 / 72.54	92.76 / 26.93
KNN [55]	74.11 / 70.47	90.81 / 31.93	71.10 / 70.87	90.18 / 34.13	72.60 / 70.67	90.50 / 33.03
SHE [37]	76.11 / 70.88	92.42 / 27.12	73.78 / 73.01	90.92 / 41.45	74.95 / 71.94	91.67 / 34.28
GEN [62]	76.30 / 70.78	91.35 / 32.23	76.85 / 65.32	89.79 / 35.61	76.57 / 68.05	90.57 / 33.92
WeiPer [14]	75.00 / 73.02	90.32 / 38.16	80.05 / 61.39	95.54 / 22.08	77.52 / 67.20	92.93 / 30.12
SCALE [57]	59.03 / 93.94	75.22 / 86.93	81.36 / 59.76	96.53 / 16.53	70.19 / 76.85	85.87 / 51.73
$\overleftrightarrow{S}$ **ActSub** w/ GEN	78.45 / **64.10**	91.62 / 29.19	77.81 / 62.97	95.90 / 18.26	78.13 / 63.54	93.76 / 23.73
$\overleftrightarrow{S}$ **ActSub** w/ SCALE	77.12 / 65.84	90.64 / 31.00	**84.24 / 52.60**	**96.96 / 14.29**	**80.68 / 59.22**	**93.80 / 22.65**

Table 8. *Accuracy of our unified score function reported with ViT and ResNet.* The format is AUC $(\%, \uparrow)$ / FPR $(\%, \downarrow)$. Models trained on ImageNet-1k [9, 44], and evaluated on OpenOOD [63].

B. Expanded Tables

We provide an expanded version of the tables (Tabs. 3 and 4) for our CIFAR [29] and OpenOOD [63] experiments in Sec. 4.2. Tab. 9 shows the accuracy of our method ActSub compared to baselines for each separate OOD dataset for when CIFAR10 and CIFAR100 are ID. Similarly, in Tab. 10, we present the accuracy of ActSub for each individual dataset in Near-OOD and Far-OOD settings of the OpenOOD benchmark.

	Method	SVHN		iSUN		Textures		Places365		Average	
		AUC (↑)	FPR (↓)	AUC (↑)	FPR (↓)	AUC (↑)	FPR (↓)	AUC (↑)	FPR (↓)	AUC (↑)	FPR (↓)
CIFAR10	Energy [36]	93.99	40.61	98.07	10.07	86.43	56.12	91.64	39.40	91.18	54.18
	ReAct [51]	93.87	41.64	97.72	12.72	92.47	43.58	91.03	43.41	93.77	35.31
	DICE [50]	95.90	25.99	99.14	4.36	88.18	41.90	89.13	48.59	93.09	30.21
	LINe [1]	97.75	11.38	99.01	4.90	95.12	23.44	91.17	43.96	95.75	20.88
	ASH-S [11]	98.65	6.51	98.90	5.17	95.09	24.34	88.34	48.45	95.25	21.12
	DDCS [61]	97.95	9.90	99.11	4.45	95.96	20.16	91.19	42.90	96.05	19.35
	SCALE [57]	98.72	5.80	**99.21**	**3.43**	94.97	23.42	91.74	38.69	96.16	17.84
	$\overleftrightarrow{S}$ **ActSub** (ours)	**99.09**	**4.39**	99.17	3.45	**96.76**	**17.38**	**92.47**	**36.27**	**96.87**	**15.37**
CIFAR100	Energy [36]	81.85	87.46	78.95	74.54	71.03	84.15	77.72	79.20	77.39	81.34
	ReAct [51]	81.41	83.81	86.55	65.27	78.95	77.78	74.04	82.65	80.24	77.38
	DICE [50]	88.84	54.65	90.08	48.72	76.42	65.04	**77.26**	**79.58**	83.15	62.00
	LINe [1]	91.90	31.10	94.76	24.12	87.84	39.29	64.18	88.41	84.63	45.74
	ASH-S [11]	95.76	25.02	91.30	46.67	92.35	34.02	71.62	85.86	87.76	47.89
	DDCS [61]	92.58	31.34	**96.17**	**18.46**	90.29	35.30	67.91	87.11	86.73	43.05
	SCALE [57]	96.29	22.05	92.47	42.14	92.34	34.20	72.66	85.04	88.44	45.86
	$\overleftrightarrow{S}$ **ActSub** (ours)	**97.45**	**13.72**	91.43	43.83	**95.07**	**23.44**	73.46	84.06	**89.35**	**41.26**

Table 9. *Expanded version of CIFAR results for DenseNet-101, showing each dataset individually* (all in %). We report results with CIFAR10 [29] and CIFAR100 [29] as ID and SVHN [40], iSUN [59], Places365 [65], and Textures [7] as OOD.

Method	NINCO(Near)	SSB-Hard(Near)	Near-OOD(Avg.)	iNaturalist(Far)	Textures(Far)	OpenImage-O(Far)	Far-OOD(Avg.)
MSP [17]	79.95 / 56.88	72.09 / 74.49	76.02 / 65.68	88.41 / 43.34	82.43 / 60.87	84.86 / 50.13	85.23 / 51.45
Energy [36]	79.70 / 60.58	72.08 / 76.54	75.89 / 68.56	90.63 / 31.30	88.70 / 45.77	89.06 / 38.09	89.47 / 38.39
ReAct [51]	81.73 / 55.82	73.03 / 77.55	77.38 / 66.69	96.34 / 16.72	92.79 / 29.64	91.87 / 32.58	93.67 / 26.31
RankFeat [48]	55.89 / 89.63	46.08 / 94.03	50.99 / 91.83	40.06 / 94.40	70.90 / 76.84	50.83 / 90.26	53.93 / 87.17
ViM [52]	78.63 / 62.29	65.54 / 80.41	72.08 / 71.35	89.56 / 30.68	97.97 / 10.51	90.50 / 32.82	92.68 / 24.67
SHE [62]	76.49 / 69.72	71.08 / 76.30	73.78 / 73.01	92.65 / 34.06	93.60 / 35.27	86.52 / 55.02	90.92 / 41.45
GEN [37]	81.70 / 54.90	72.01 / 75.73	76.85 / 65.32	92.44 / 26.10	87.59 / 46.22	89.26 / 34.50	89.76 / 35.61
ASH-S [11]	84.54 / 53.26	74.72 / 70.80	79.63 / 62.03	97.72 / 11.02	97.87 / 10.90	93.82 / 28.60	96.47 / 16.86
WeiPer+KLD [14]	85.37 / 48.67	74.73 / 74.12	80.05 / 61.39	97.49 / 13.59	96.18 / 22.17	92.94 / 30.49	95.54 / 22.08
SCALE [57]	85.37 / 51.80	77.35 / 67.72	81.36 / 59.76	98.02 / 9.51	97.63 / 11.90	93.95 / 28.18	96.53 / 16.53
$\overleftrightarrow{S}$ **ActSub** w/ ReAct	83.87 / 51.01	71.53 / 79.71	77.70 / 65.36	97.92 / 8.85	98.13 / 8.64	93.62 / 29.23	96.56 / 15.57
$\overleftrightarrow{S}$ **ActSub** w/ ASH-S	**87.42 / 43.36**	80.70 / **61.58**	84.06 / **52.47**	98.06 / 8.48	98.24 / 9.16	**94.41 / 25.03**	96.91 / **14.22**
$\overleftrightarrow{S}$ **ActSub** w/ SCALE	87.35 / 43.49	**81.14** / 61.71	**84.24** / 52.61	**98.51 / 6.79**	**98.26 / 8.44**	94.11 / 27.54	**96.96** / 14.29

Table 10. *Expanded version of OpenOOD results for ResNet-50 trained on ImageNet-1k, showing each dataset individually* (all in %). Reported results are in the format AUC (↑) / FPR (↓).