ICCV OCT 19-23, 2025 HONOLULU HAWAII

Activation Subspaces for Out-of-Distribution Detection

Barış Zöngür[1]

Robin Hesse[1]

Stefan Roth[1,2]

[1]TU Darmstadt [2]hessian.AI

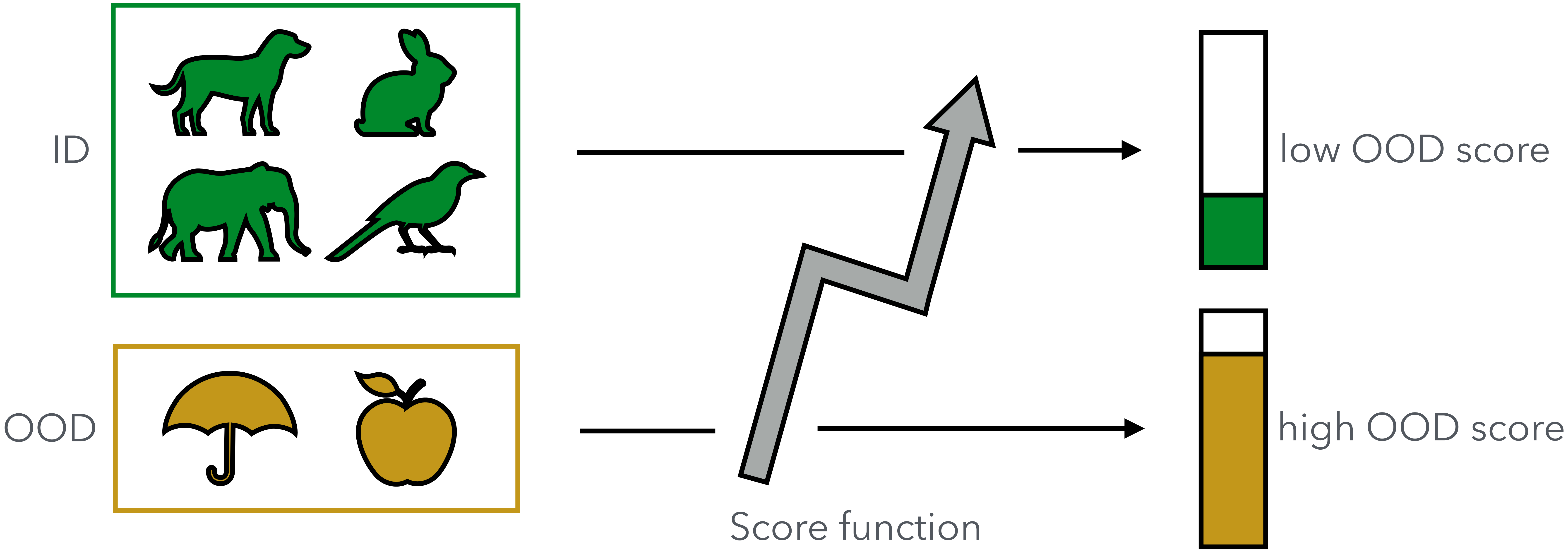
ICCV 2025 Video Presentation

hessian.AI

vi visual inference

TECHNISCHE UNIVERSITÄT DARMSTADT

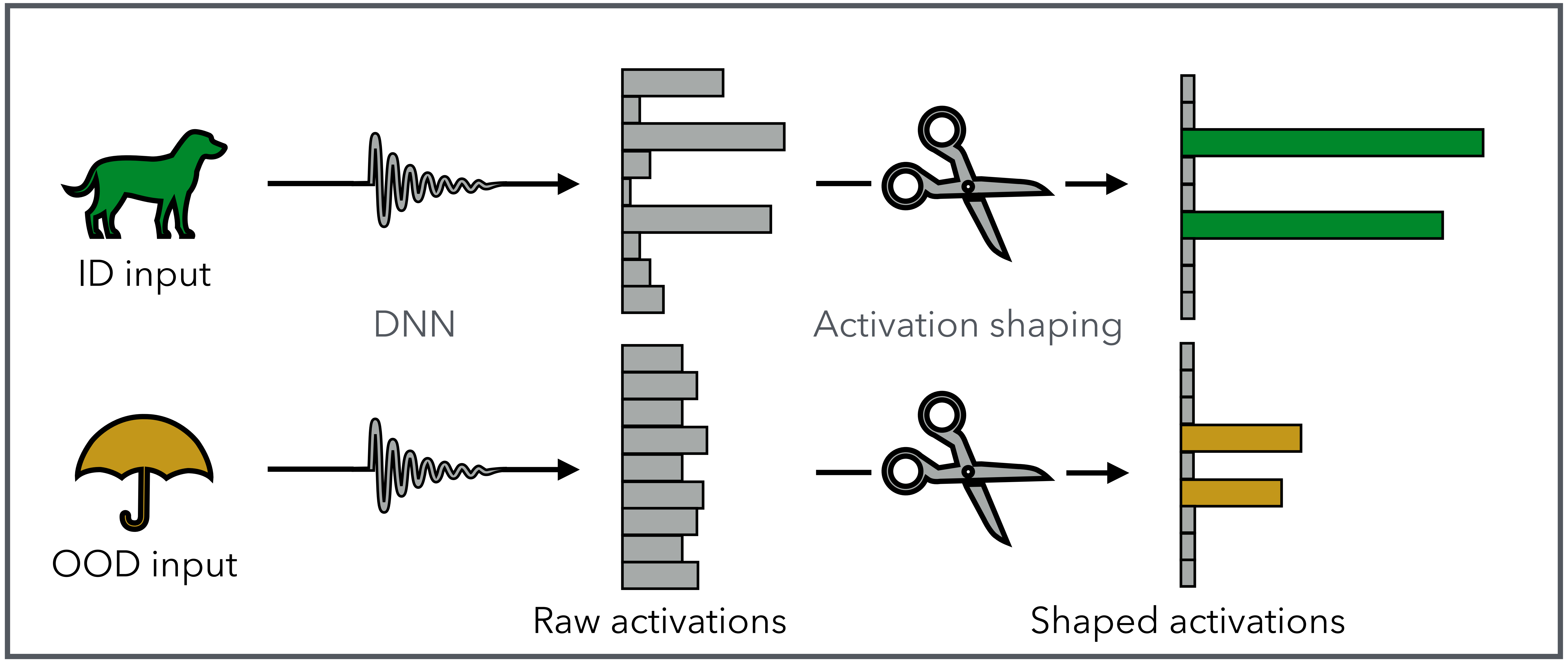
OOD Detection



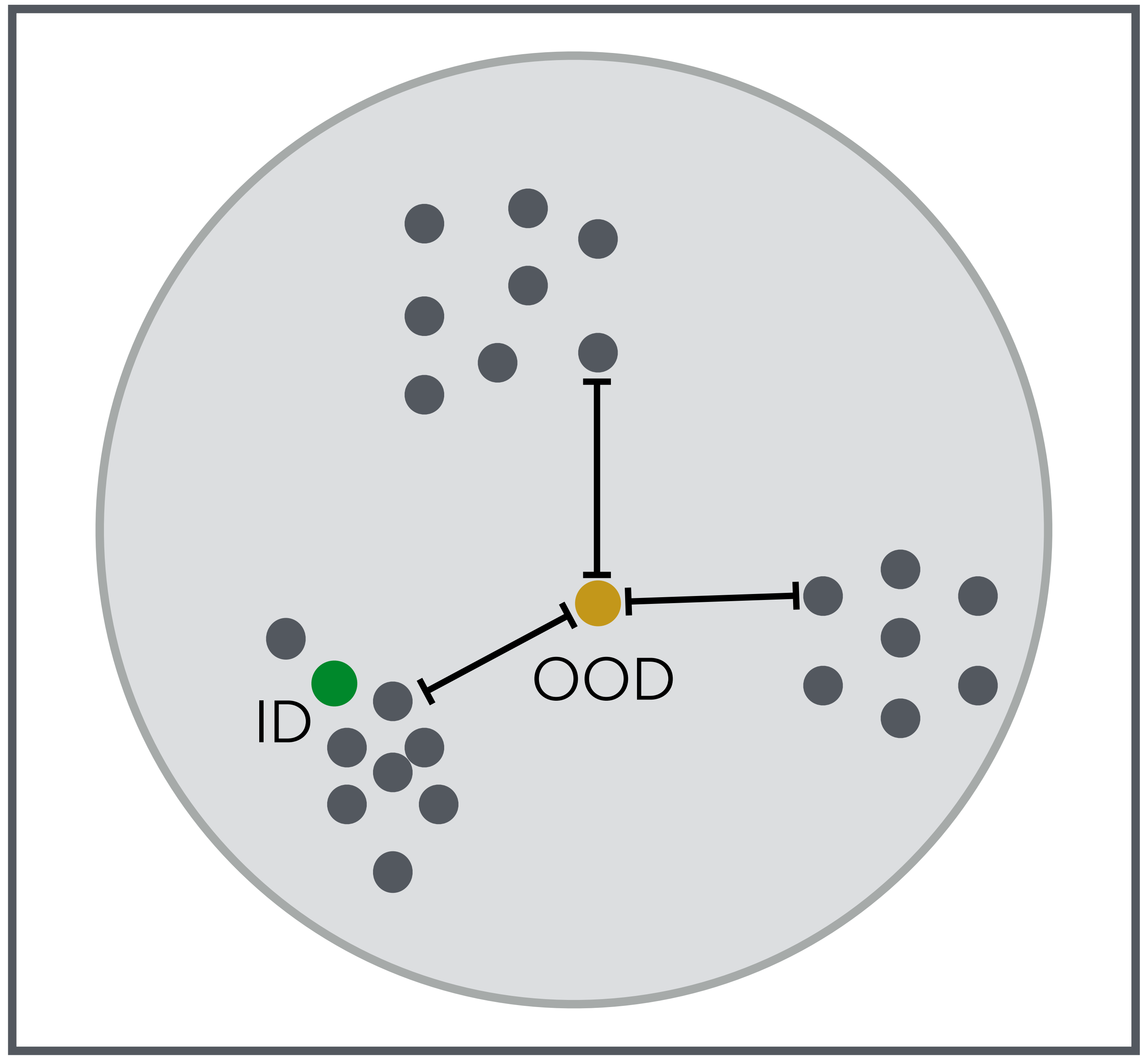
- ***ID (in-distribution)*** inputs belong to the training distribution.
- ***OOD (out-of-distribution)*** inputs deviate from the predefined class taxonomy [1].

[1] Hendrycks et al. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In ICLR, 2017.

Related Work



Activation Shaping

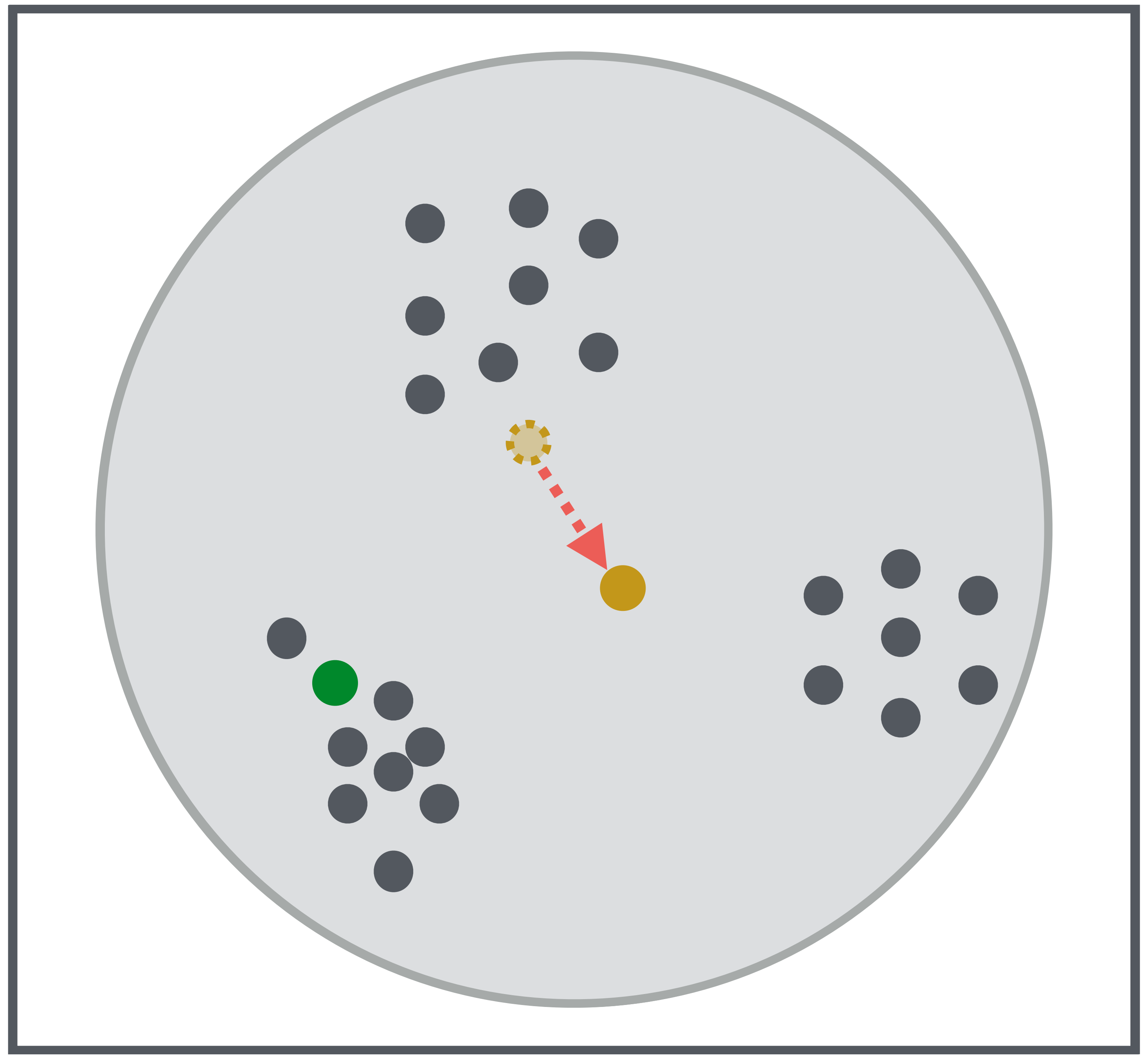


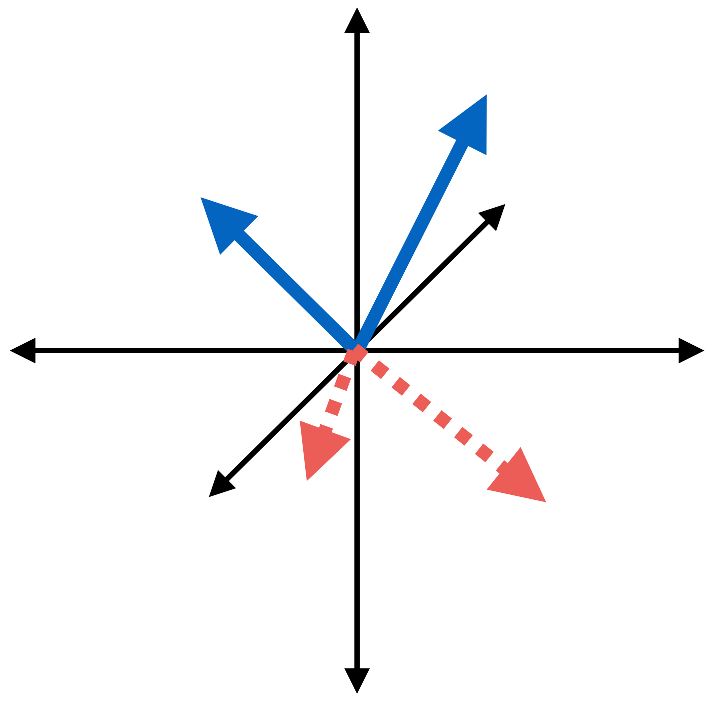
Distance-based

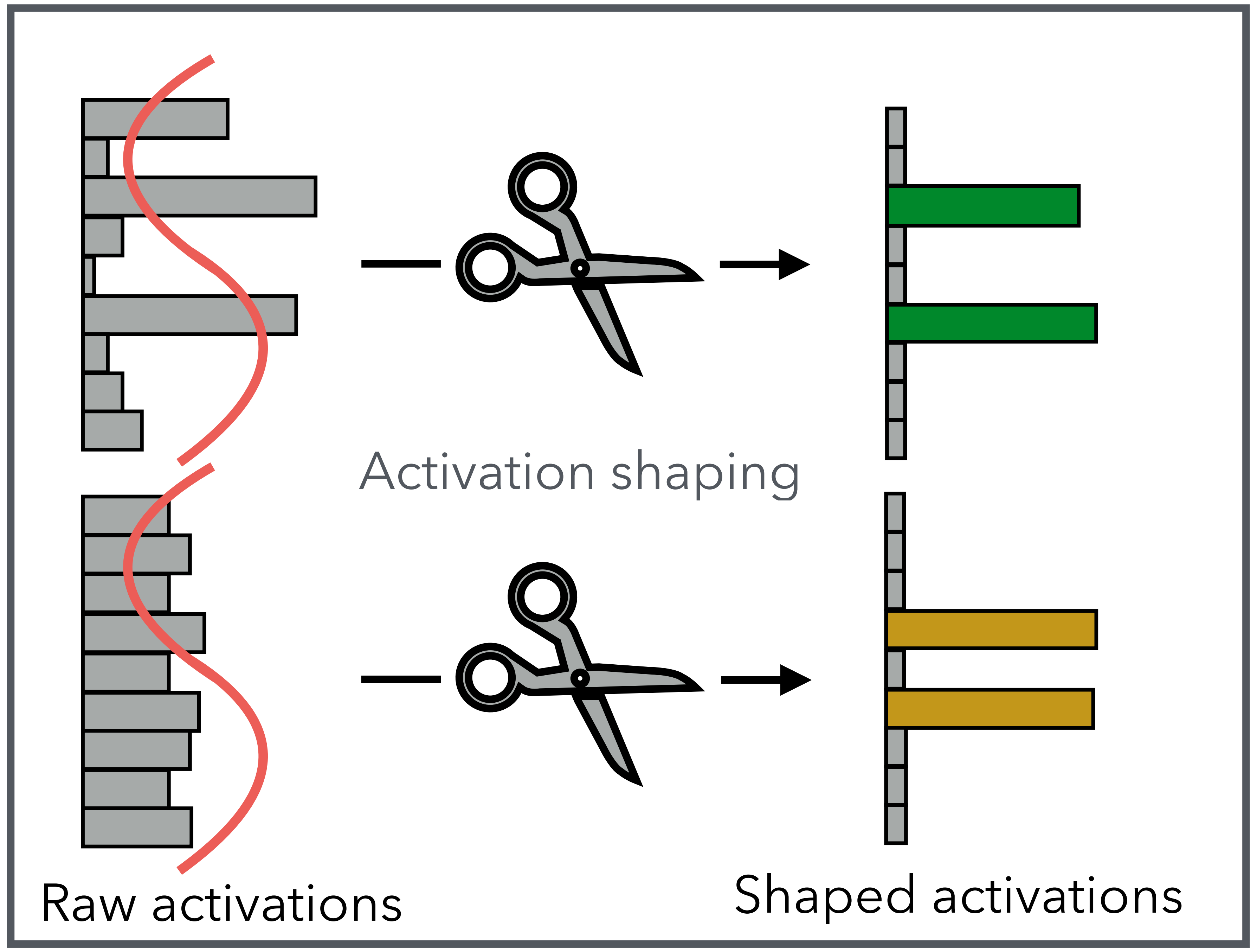
- ***Activation shaping*** methods prune or scale the activation channels [1].
- ***Distance-based methods*** utilize the position of the activations [2, 3].

[1] Djurisic et al. Extremely simple activation shaping for out-of-distribution detection. In ICLR, 2023.
[2] Park et al. Nearest neighbor guidance for out-of-distribution detection. In ICCV, 2023.
[3] Sun et al. Out-of-distribution detection with deep nearest neighbors. In ICML, 2022.

Motivation







- OOD inputs can produce activations that ***highly align*** with the class vectors.
- ***Insignificant*** directions can be useful for OOD detection.
- To identify the insignificant directions, we examine the ***right singular vectors*** of the weight matrix.

- The insignificant directions can interfere with the ***decisive information*** on activations.

$$\boxed{\mathbf{l} = \mathbf{W}\mathbf{a}}$$

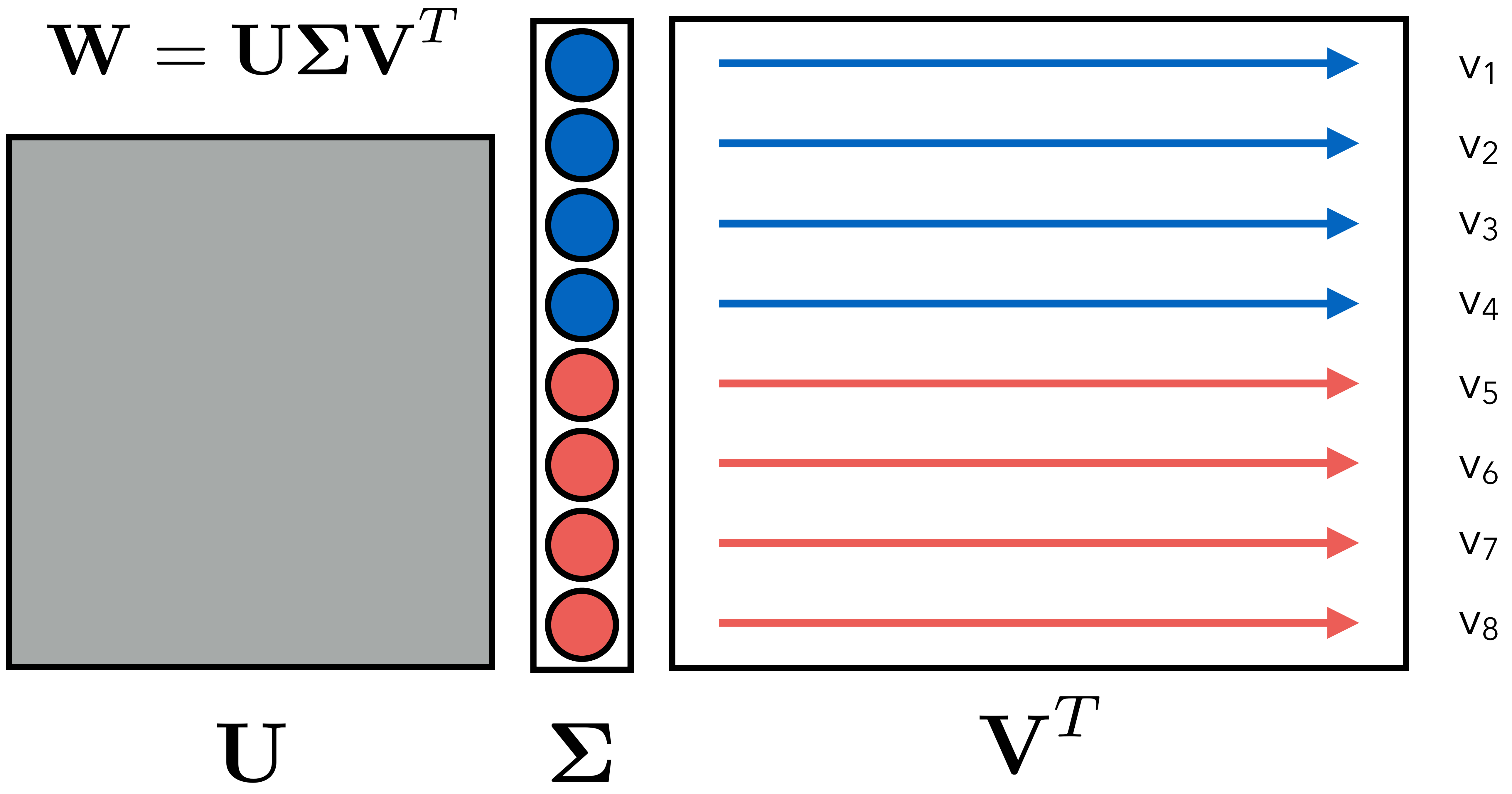
linear classification head (penultimate later)

logits

activation

weight matrix

Activation Subspaces



Decisive Subspace

$$\overleftarrow{\mathbf{V}}^T = \begin{cases} \mathbf{v}_i^T & \text{if } i \leq k \\ \mathbf{0}^T & \text{otherwise} \end{cases},$$

$$\overleftarrow{\mathbf{a}} = \overleftarrow{\mathbf{V}}\overleftarrow{\mathbf{V}}^T\mathbf{a}$$

Insignificant Subspace

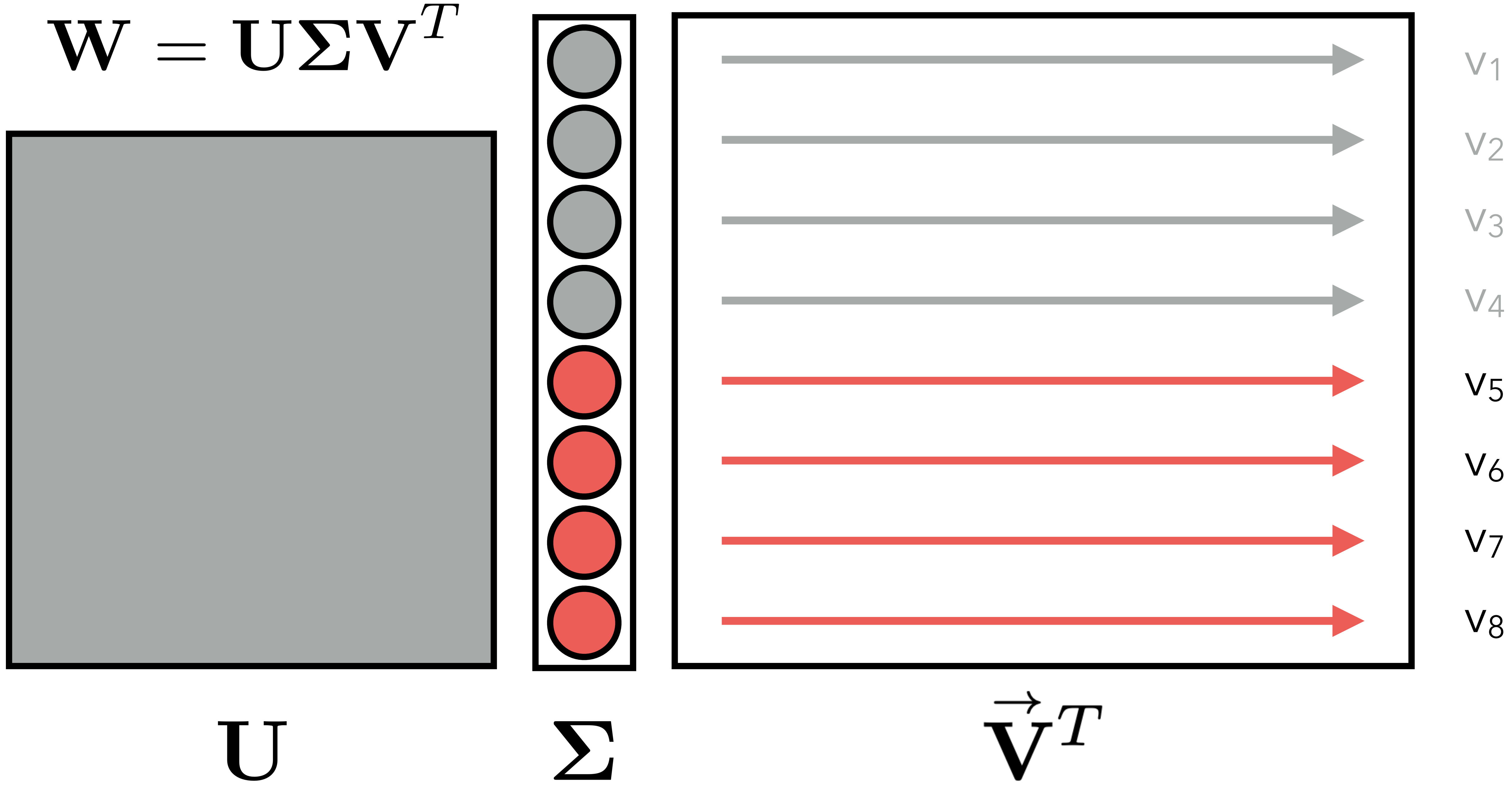
$$\overrightarrow{\mathbf{V}}^T = \begin{cases} \mathbf{0}^T & \text{if } i \leq k \\ \mathbf{v}_i^T & \text{otherwise} \end{cases},$$

$$\overrightarrow{\mathbf{a}} = \overrightarrow{\mathbf{V}}\overrightarrow{\mathbf{V}}^T\mathbf{a}$$

- ***Decisive Subspace:*** Directions corresponding to high singular values.
    - Contributes ***maximally*** to the final classifier output.
- **Insignificant Subspace:** Directions corresponding to low singular values.
    - Contributes ***minimally*** to the final classifier output.

Insignificant Component

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$



Insignificant Subspace

$$\vec{\mathbf{V}}^T = \begin{cases} \mathbf{0}^T & \text{if } i \le k \\ \mathbf{v}_i^T & \text{otherwise} \end{cases},$$

$$\vec{\mathbf{a}} = \vec{\mathbf{V}}\vec{\mathbf{V}}^T\mathbf{a}$$

$$\vec{S} = -\log\left(1 - \frac{1}{N}\sum_{i=1}^{N} \text{cos\_sim}(\vec{\mathbf{a}}^{(i)}, \vec{\mathbf{a}})\right)$$

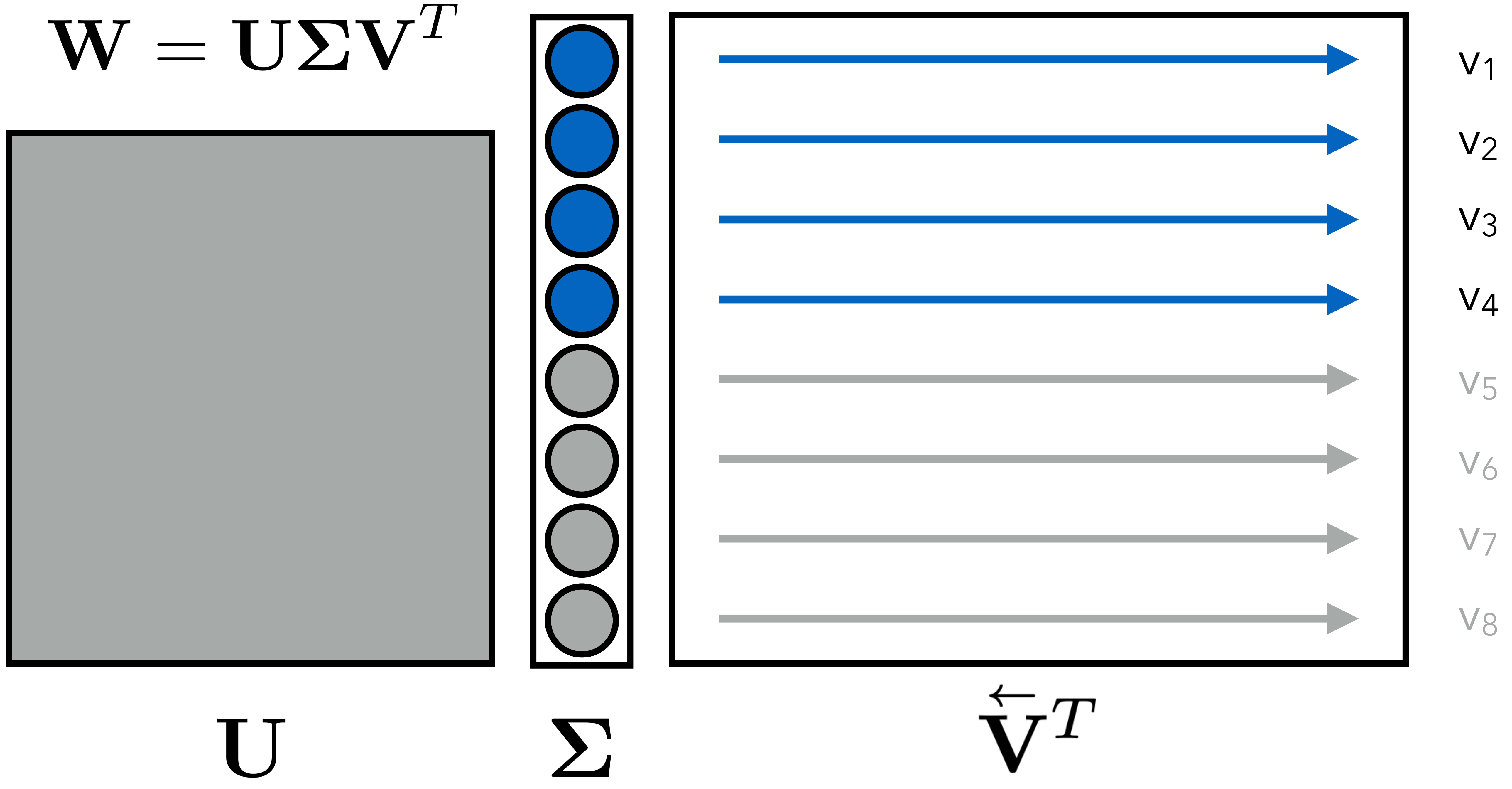
Key Observation

- The ***insignificant component*** yields powerful features, akin to a random neural network [1, 2], that is discriminative for OOD detection!
- $\vec{S}$: The ***average cosine similarity*** to nearest neighbors from the training data.

[1] Ulyanov et al. Deep image prior. In CVPR, 2018.
[2] Can et al. A random CNN sees objects: One inductive bias of CNN and its applications. In AAAI, 2022.

Decisive Component

$$\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$



Decisive Subspace

$$\overleftarrow{\mathbf{V}}^T = \begin{cases} \mathbf{v}_i^T & \text{if } i \leq k \\ \mathbf{0}^T & \text{otherwise} \end{cases},$$

$$\overleftarrow{\mathbf{a}} = \overleftarrow{\mathbf{V}}\overleftarrow{\mathbf{V}}^T\mathbf{a}$$

$$\overleftarrow{S} = -E\left(\mathbf{U}\boldsymbol{\Sigma}\overleftarrow{\mathbf{V}}^T\Phi(\overleftarrow{\mathbf{a}})\right)$$

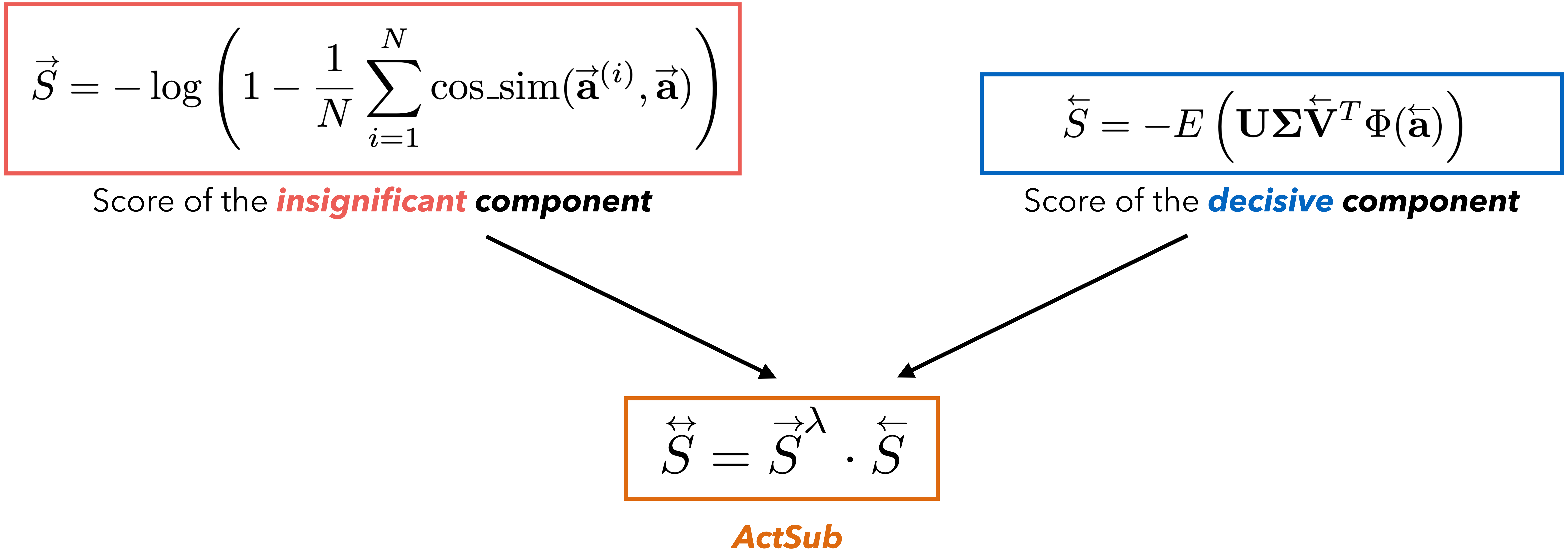
activation shaping

Key Observation

- Activation shaping methods profit from considering the ***decisive component***, as the insignificant component can cause interference.
- $\overleftarrow{S}$: The ***energy function*** [1] on logits from the shaped activation.

[1] Liu et al. Energy-based out-of-distribution detection. In NeurIPS, 2020.

ActSub



$$\vec{S} = -\log\left(1 - \frac{1}{N}\sum_{i=1}^{N} \text{cos\_sim}(\vec{\mathbf{a}}^{(i)}, \vec{\mathbf{a}})\right)$$

Score of the ***insignificant*** ***component***

$$\overleftarrow{S} = -E\left(\mathbf{U}\mathbf{\Sigma}\overleftarrow{\mathbf{V}}^{T}\Phi(\overleftarrow{\mathbf{a}})\right)$$

Score of the ***decisive*** ***component***

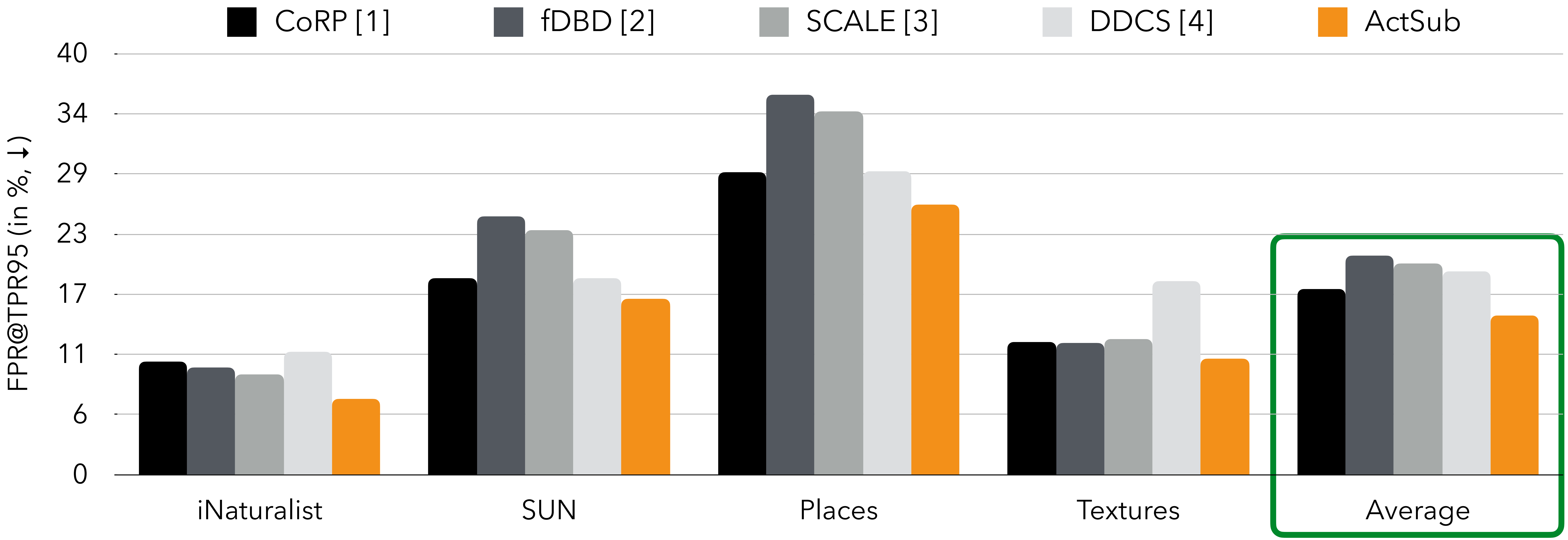
$$\overleftrightarrow{S} = \vec{S}^{\lambda} \cdot \overleftarrow{S}$$

ActSub

- We utilize the discriminative information from both components to define our final score function ***ActSub*****.**

Experiments

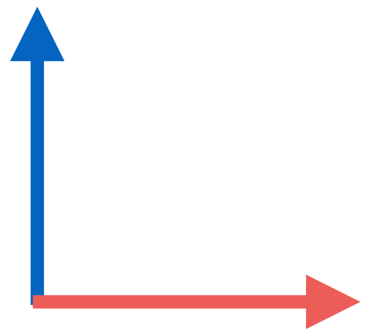
Model: ResNet-50 ID: ImageNet-1K OOD: iNaturalist, SUN, Places, Textures



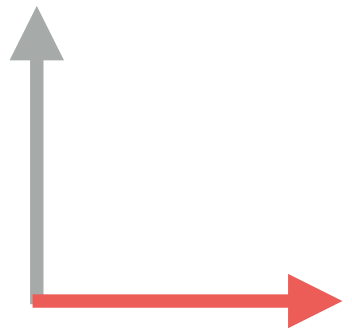
- With the complementary effect of both subspaces, ***ActSub*** achieves SoTA results.
- Note: For distance-based methods CoRP [1] and fDBD [2], we consider the strongest reported variant with an activation shaping method.

[1] Fang et al. Kernel PCA for out-of-distribution detection. In NeurIPS, 2024.
[2] Liu et al. Fast decision boundary based out-of-distribution detector. In ICML, 2024.
[3] Yan et al. Discriminability-driven channel selection for out-of-distribution detection. In CVPR, 2024.
[4] Xu et al. Scaling for training time and post-hoc out-of-distribution detection enhancement. In ICLR, 2024.

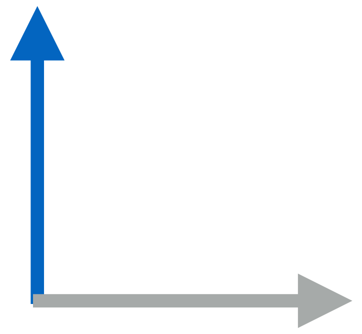
Conclusion



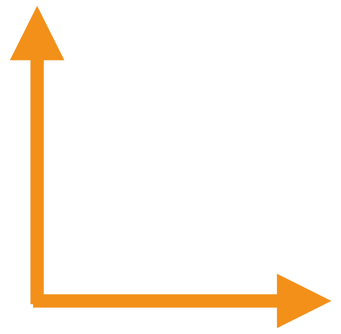
- We define ***two orthogonal subspaces*** of the activation space based on their contribution to the classifier output.



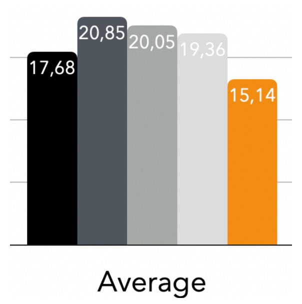
- The ***insignificant component*** yields powerful features untainted by the classification task, discriminative for OOD Detection.



- Selectively applying activation shaping to ***decisive component*** mitigates the channel-wise interference on activations.



- We define ***ActSub*** by combining discriminative information from both subspaces.



- With the complementary effect of the subspaces, ***ActSub*** achieves SoTA results.

https://github.com/visinf/actsub/

https://arxiv.org/abs/2508.21695

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). SR further acknowledges support by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223).

erc
European Research Council
Established by the European Commission

DFG Deutsche Forschungsgemeinschaft

Activation Subspaces for Out-of-Distribution Detection

Barış Zöngür[1] Robin Hesse[1] Stefan Roth[1,2]

[1]TU Darmstadt [2]hessian.AI

hessian.AI

visual inference

TECHNISCHE UNIVERSITÄT DARMSTADT