Revealing Similar Semantics Inside CNNs: An Interpretable Concept-based Comparison of Feature Spaces

Georgii Mikriukov[1,2], Gesina Schwalbe[1], Christian Hellert[1], Korinna Bade[2]

AIMLAI: Advances in Interpretable Machine Learning and Artificial Intelligence

Workshop, ECML 2023, Torino, Italy

[1] Continental AG, Germany | [2] Hochschule Anhalt, Germany

Goal & Contribution

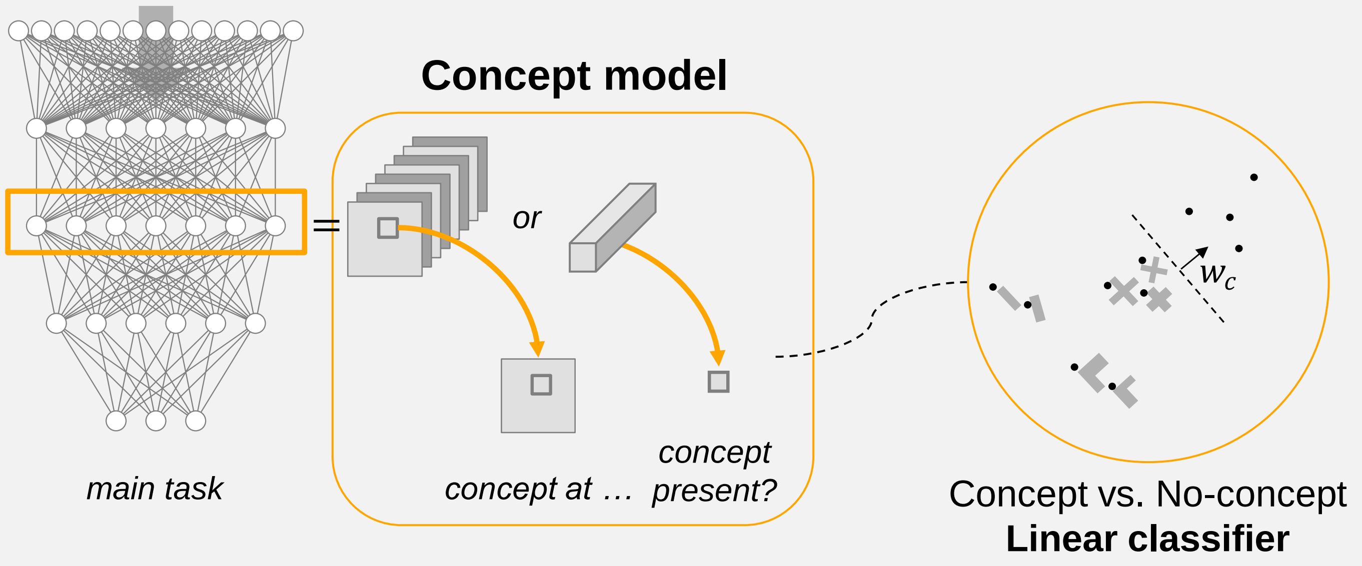
Goal: Semantic Comparability of DNNs for **informed model selection**

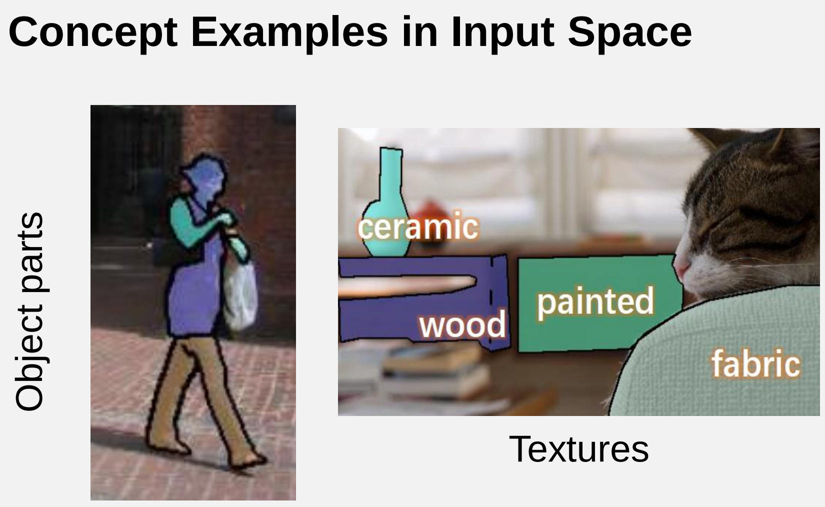
- **Architecture-agnostic** concept-based **comparison of feature space semantics** across models
- Supervised (ranking-based) and unsupervised (saliency-based) **semantic similarity metrics**
- Study and **comparison of learned semantics in** layers of several **object detection** CNNs

Concept analysis

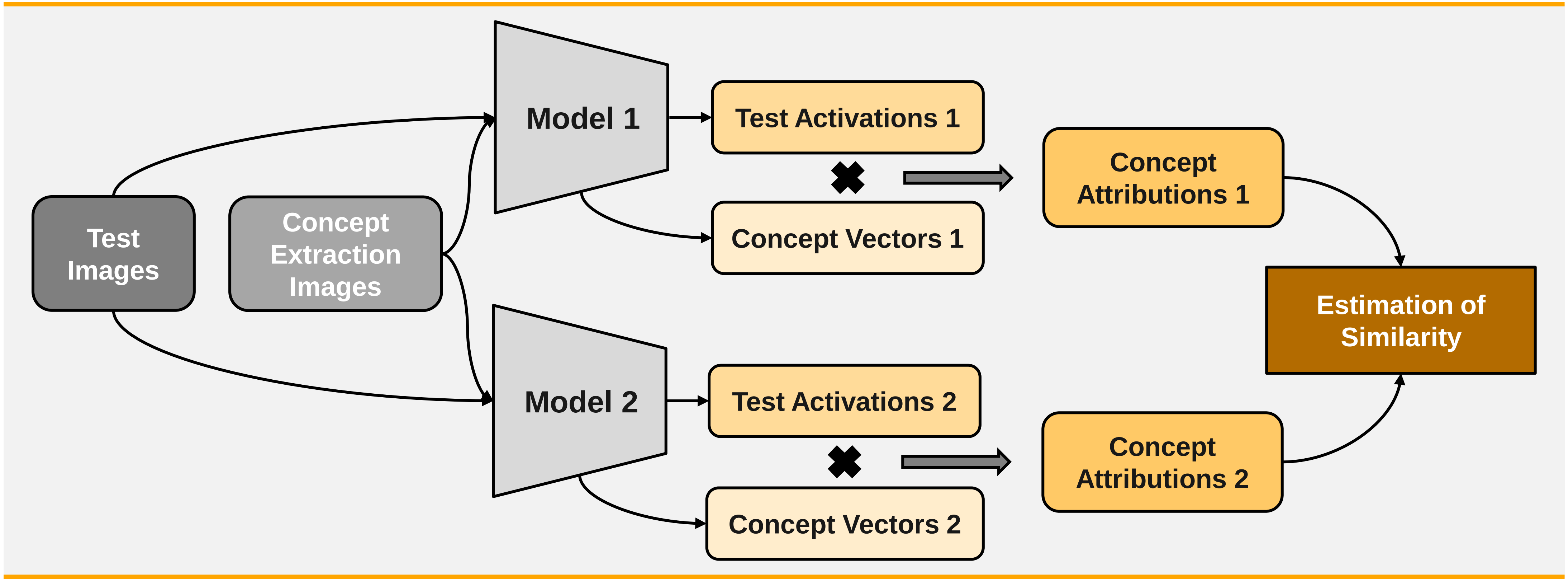
(Semantic) **concept associates vectors** in latent space **to input regions**

- **Concept Activation Vector** (CAV) [1] indicates the **orientation of the concept** within the feature space
- Concepts vectors enable the measurement of **concept attribution in samples**





Concept-based semantics comparison



Indirect feature space comparison via semantic concepts and sample attributions

Concept-based semantics comparison

Unsupervised Concept Similarity

- Saliency-based
- Are there similar concepts in feature spaces of different layers?
- How similar are concepts?

Supervised Feature Space Similarity

- Based on similarity ranking
- How similar is the arrangement of feature spaces in compared layers with respect to given concepts?

Saliency-based Unsupervised Concept Similarity

Aim:

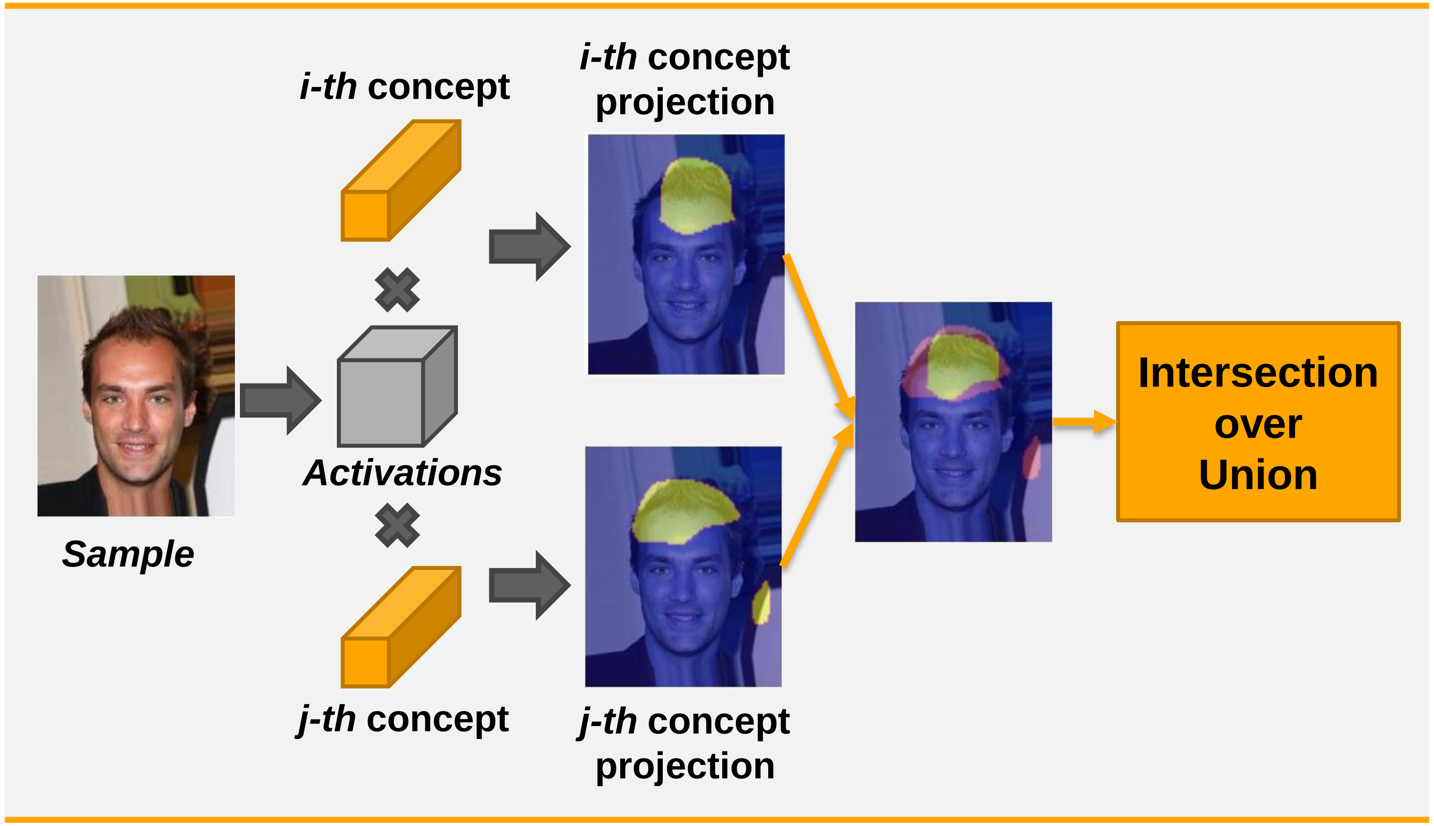
- **Discover & compare important concepts**

Similarity Estimation:

- Concept Attribution → Projection (mask) [2]
- Unsupervised Concept Similarity → IoU:

$$\mathrm{UCS}_{i,j} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{IoU}(M_i^k, M_j^k)$$

M – concept projection mask, N – number of test samples, $\mathrm{IoU}(\text{-},\text{-})$ – Intersection over Union



Similarity of concepts i and j for a single sample

Ranking-based Supervised Feature Space Similarity

Aim:

- **Compare latent spaces around given concepts**

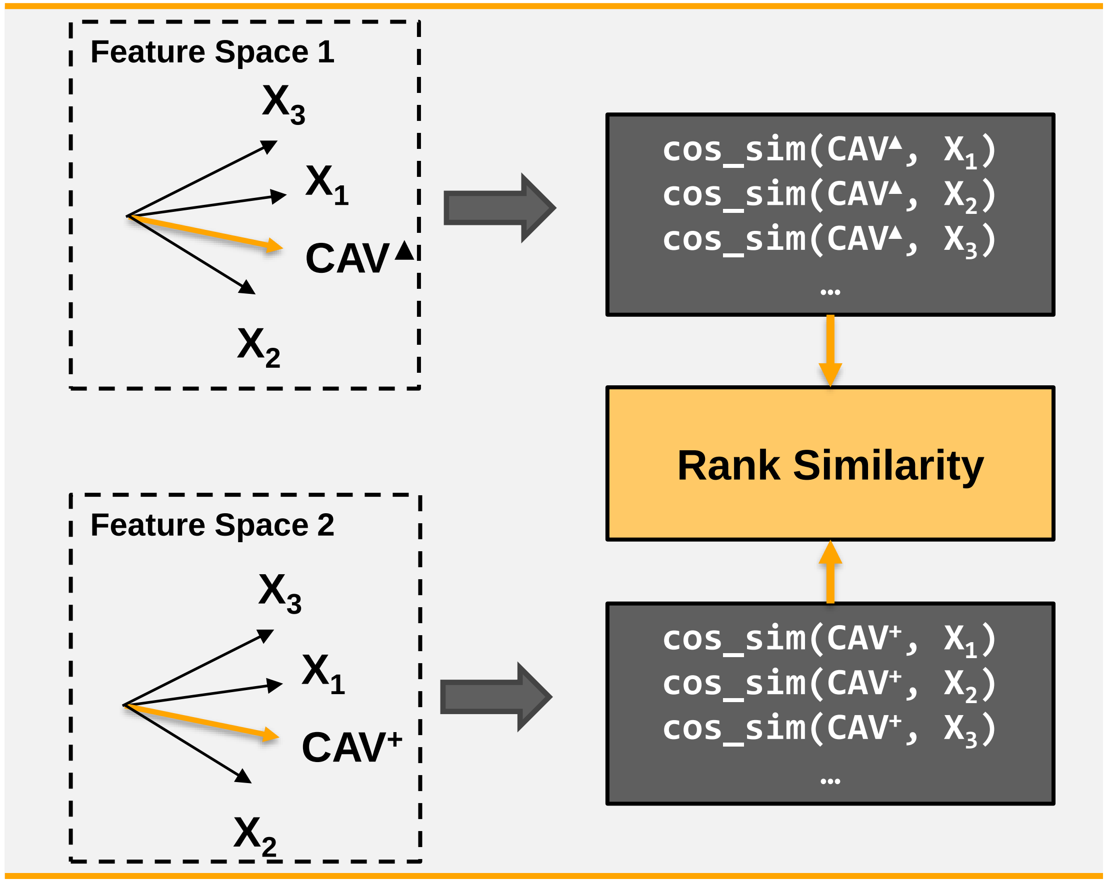
Similarity Estimation:

- Concept Vector [1] → Pivot
- Concept Attribution → Cosine Similarity
- Supervised Feature Space Similarity → Rank Order:

$$\mathrm{SFSS}_{u,v} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{PCC}\left(\{\mathrm{CS}^{i}_{u,k}\}_{k=1}^{N}, \{\mathrm{CS}^{i}_{v,k}\}_{k=1}^{N}\right)$$

$$\mathrm{CS}^{i}_{*,k} = \cos\left(\mathrm{CAV}^{i}_{*}, x_{*,k}\right), * \in \{u, v\}$$

u,w – layers, M – number of concepts, N – number of test samples, $\mathrm{PCC}(\text{-},\text{-})$ – Pearson Correlation Coefficient, $\cos(\text{-},\text{-})$ – cosine similarity, CAV^{i}_{*} – concept vector, $x_{*,k}$ – sample



Similarity of feature spaces around CAVs

Experimental Setup

Concept Analysis Methods:

- **TCAV [1]** – Supervised Similarity
- **ICE [2]** – Unsupervised Similarity, Concept Discovery

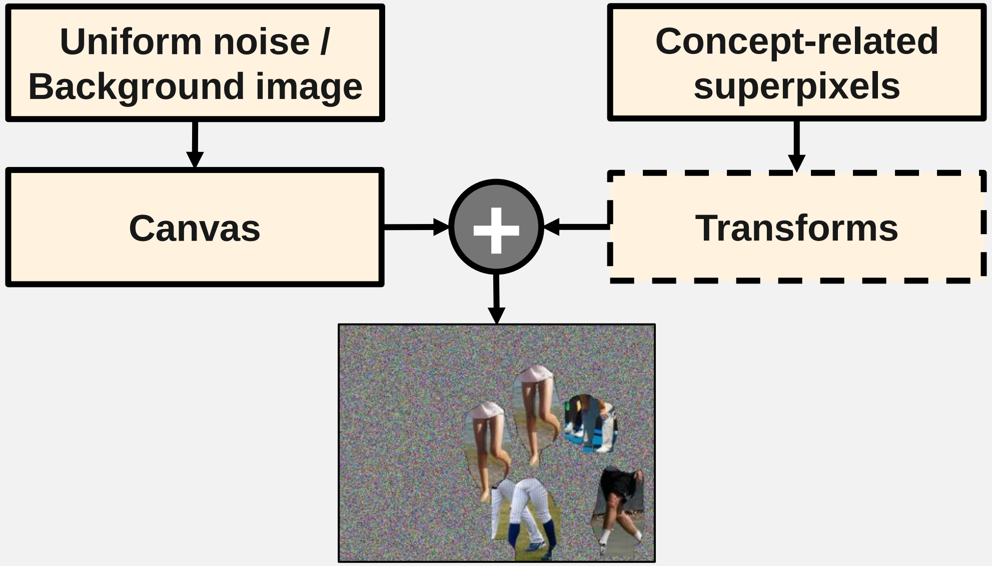
Data:

- **CelebA [3]** – Faces of celebrities
- **MS COCO [4]** – «Person» class
- **Synthetic concepts** (generated from MS COCO)

Models (MS COCO):

- **SSD** – VGG backbone
- **YOLOv5** – Residual backbone (DarkNet)
- **FasterRCNN** – Inverted residual backbone (MobileNetV3)

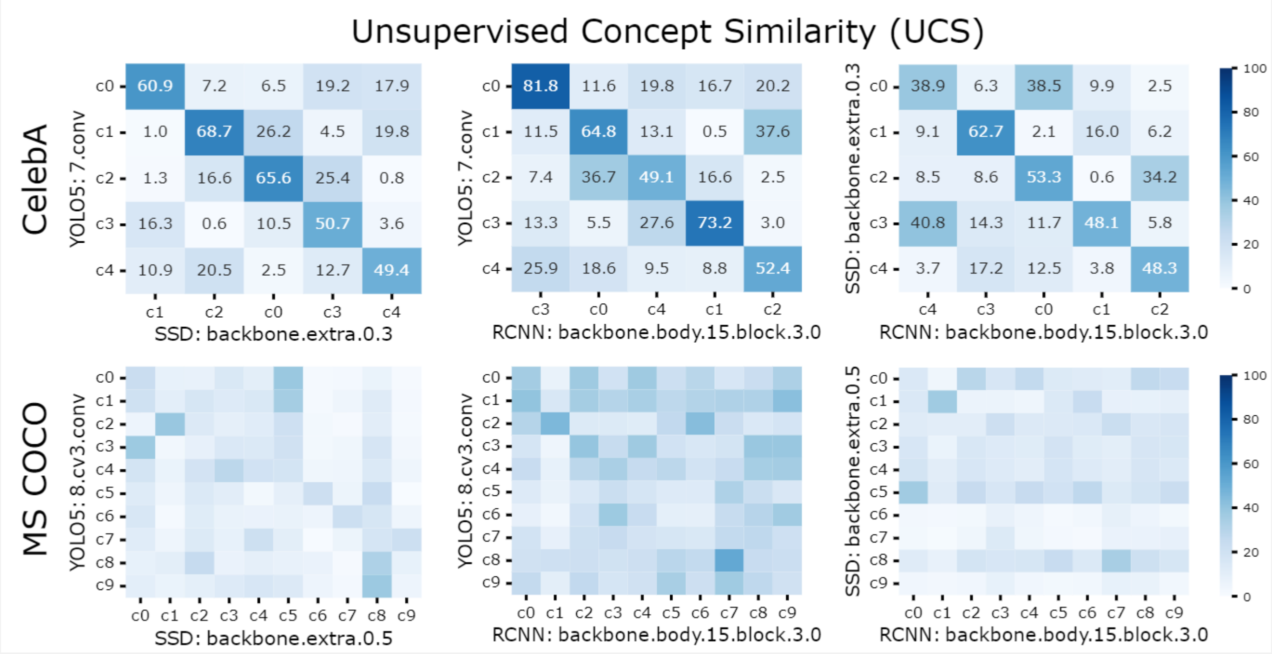
CelebA Examples



«Legs» Synthetic Concept Sample

Results: Unsupervised Saliency-based Similarity

- Test **data diversity impacts the complexity** of further **inspection.**
- Different (architecture-wise) **networks learn similar concepts:**
  - Trained on MS COCO, discovered similar concepts in CelebA

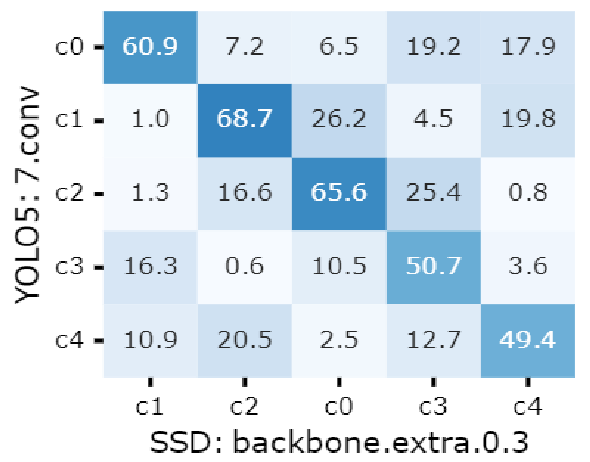


Similar concepts in CelebA

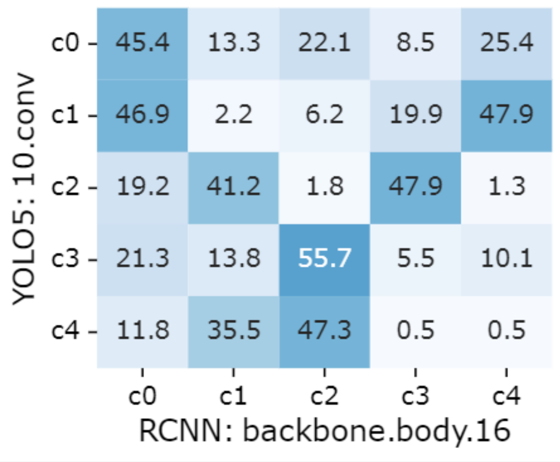
Similar concepts in MS COCO

Results: Unsupervised Saliency-based Similarity

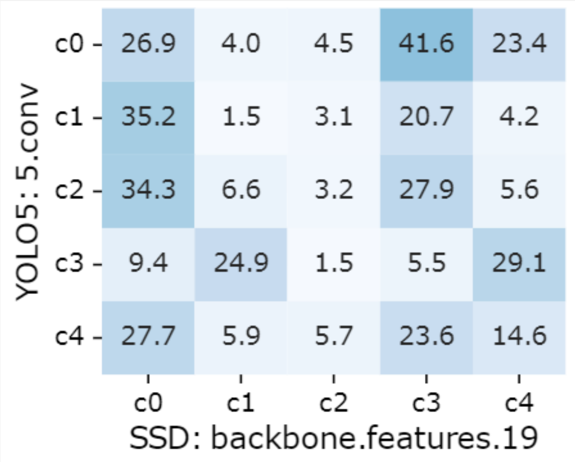
- Discovery of semantically identical layers:
  - Some **layers may have one-to-one correspondence** in discovered concepts.



One-to-one concepts



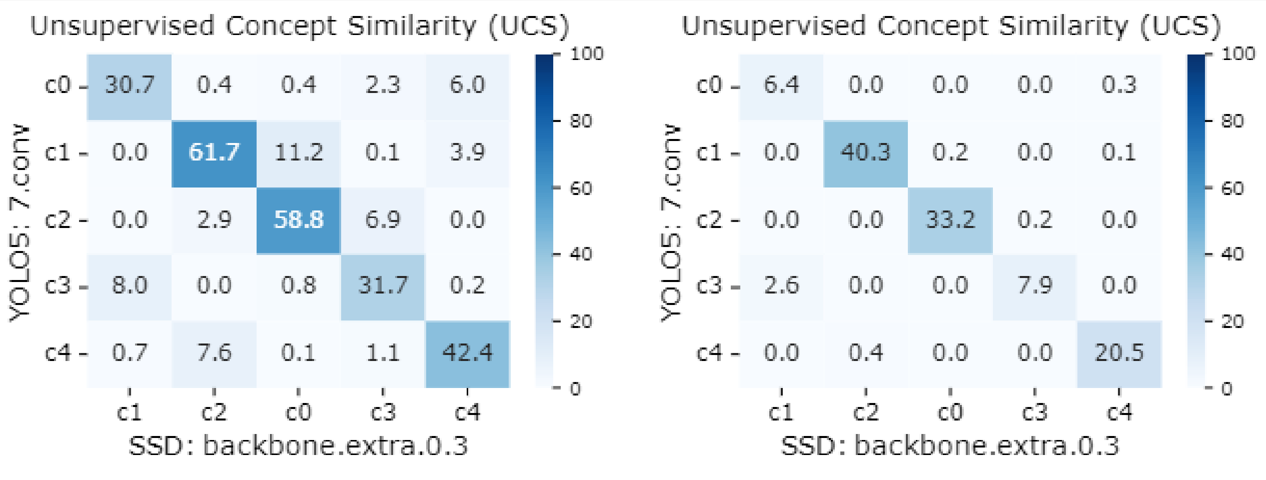
Several similar concepts



No similar concepts

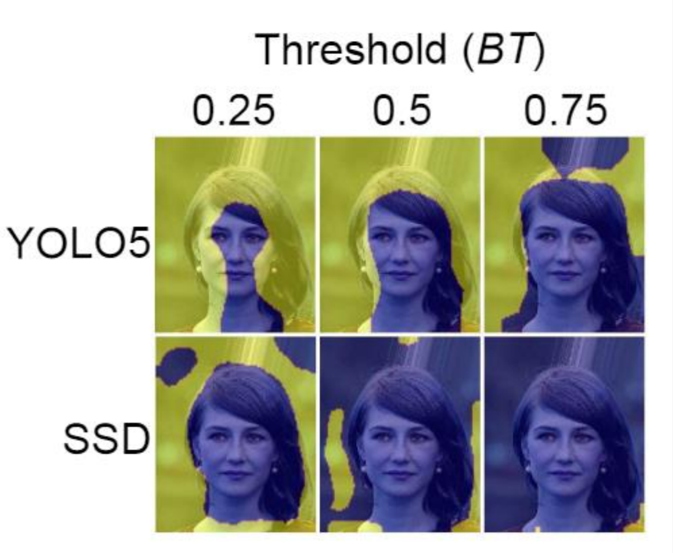
Results: Unsupervised Saliency-based Similarity

- Estimation of **relative concept robustness**:
  - By **changing the binarization threshold (BT)** of concept projection masks



(a) $BT = 0.5$ (b) $BT = 0.75$

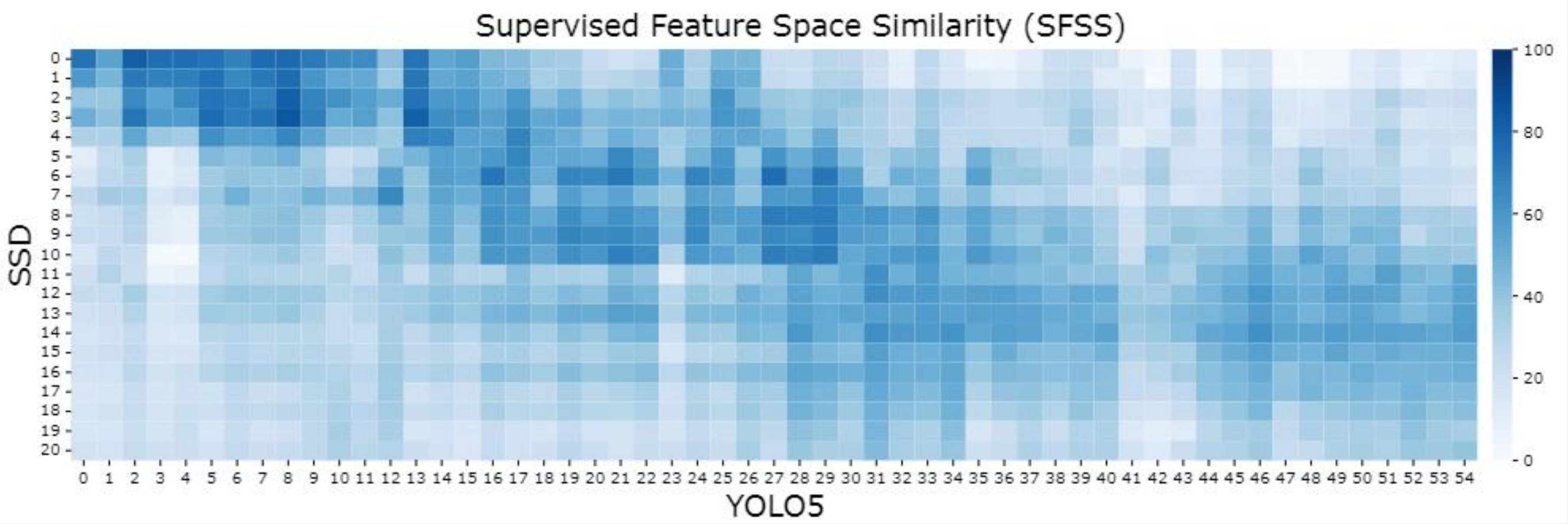
Concept similarity (robustness) for different BT



Concept masks for different BT

Results: Supervised Ranking-based Similarity

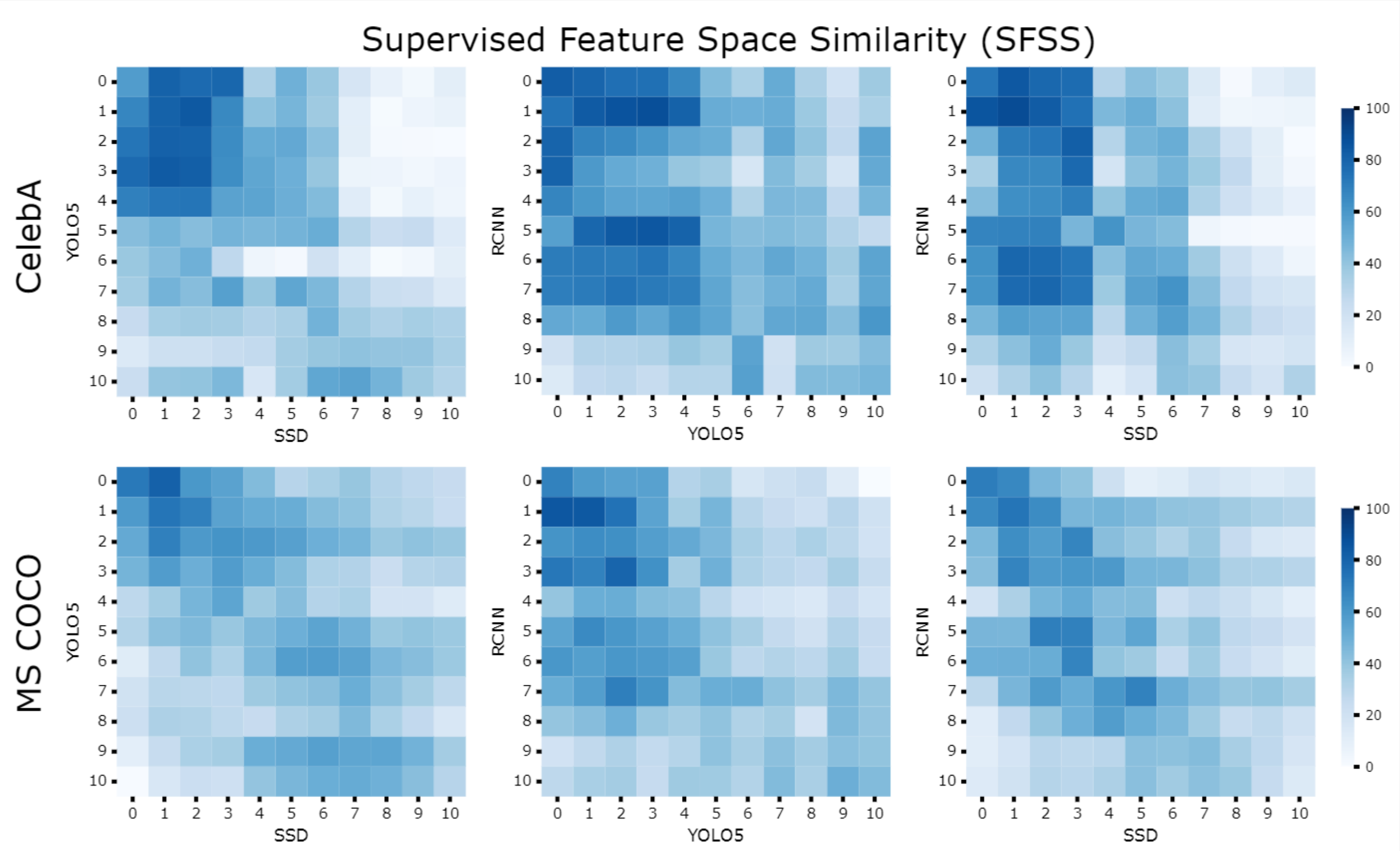
- Semantic **similarity is primarily influenced by the layer's relative depth**
- Networks can be compared by **comparing set of evenly depth-distributed layers**



Layer-wise similarity of every convolutional layer of SSD and YOLO5

Results: Supervised Ranking-based Similarity

- Simpler concepts (from CelebA) result into higher similarity
- Simpler concepts recognized in a wider range of layers
- Network backbones exhibit different semantical behavior
  - RCNN (MobileNetV3) propagates tested semantics more efficiently



Feature Space Similarity: SSD vs. RCNN vs. YOLO5

Conclusion

Summary:

- Proposed **architecture-agnostic methods** and **metrics** for estimating the similarity of feature spaces of CNN backbones.
- Explored how **semantic information is processed in** various **model backbones**.
- Identified similar concepts semantically similar layers
- Discovered that semantic information depends on the relative layer depth.

Future work:

- Apply our approach to further **large NN architectures**, e.g., transformers, and **other visual tasks** than object detection
- Try **alternative methods of concept extraction**

References

[1] **Kim, Been, et al.** "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.

[2] **Zhang, Ruihan, et al.** "Invertible concept-based explanations for cnn models with non-negative concept activation vectors." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 13. 2021.

[3] **Liu, Ziwei, et al.** "Deep learning face attributes in the wild." Proceedings of the IEEE international conference on computer vision. 2015.

[4] **Lin, Tsung-Yi, et al.** "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

Continental

KI WISSEN
Automotive AI Powered by Knowledge

HOCHSCHULE ANHALT University of Applied Sciences