On convex decision regions in deep network representations

Lenka Tětková, Thea Brüsch, Teresa Karen Scheidt, Fabian Martin Mager, Rasmus Ørtoft Aagaard, Jonathan Foldager, Tommy Sonne Alstrøm, Lars Kai Hansen

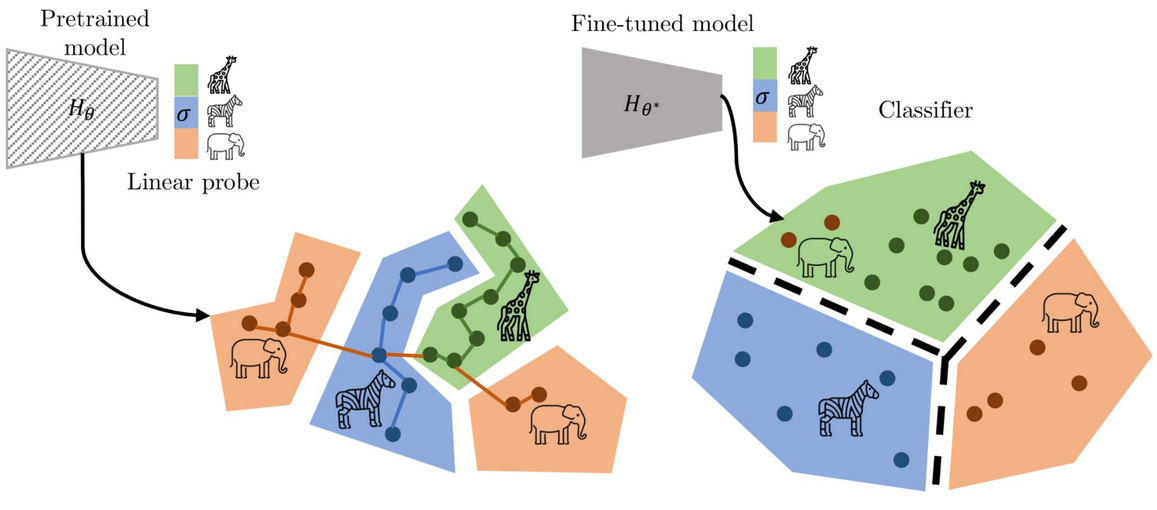
Technical University of Denmark

Motivation

In cognitive sciences, it has been shown that:

- Natural concepts form **convex regions** in human geometrical representations [1, 2].
- Convexity is closely related to **generalization** in cognitive systems [3, 4].
- Convexity supports **few-shot learning** [3].

Are decision regions implemented as convex regions in machine-learned representations as well?



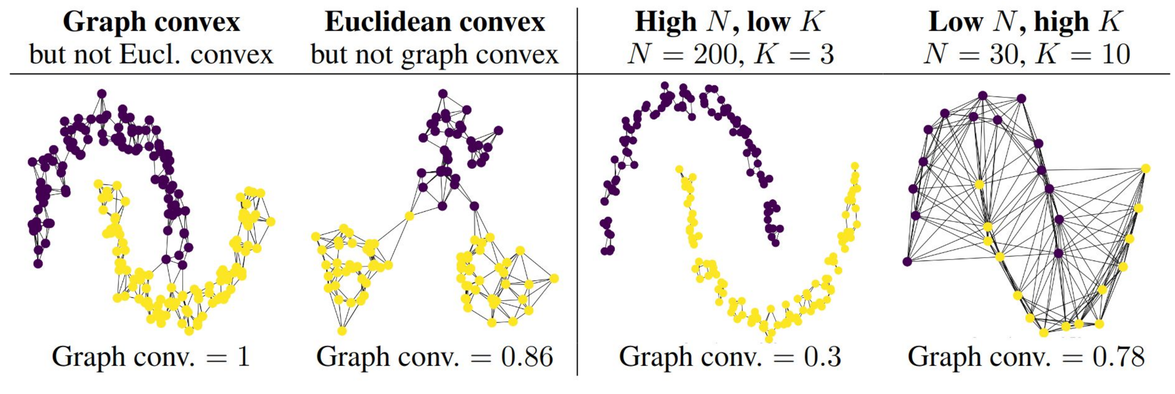
Workflows

Definition 1 (**Euclidean convexity**). A subset $S \subset \mathbb{R}^D$ is convex iff $\forall x, y \in S\ \forall t \in [0, 1]$, $z(t)= tx+ (1 - t)y$ is also in S

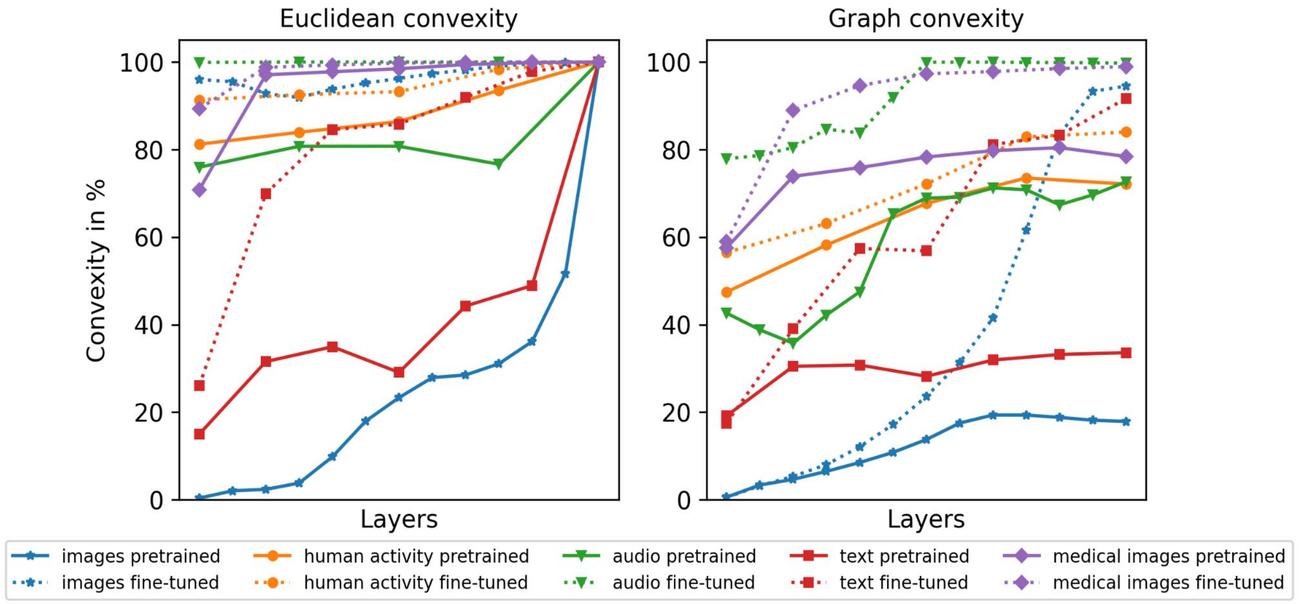
Definition 3 (**Graph convexity**, see e.g., [5]). Let (V, E) be a graph and $A \subseteq V$. We say that A is convex if for all pairs $x, y \in A$, there exists a shortest path $P = (x=v_0 , v_1 , v_2 , \ldots , v_{n-1} , y=v_n)$ and $\forall i \in \{0, \ldots , n\} : v_i \in A$.

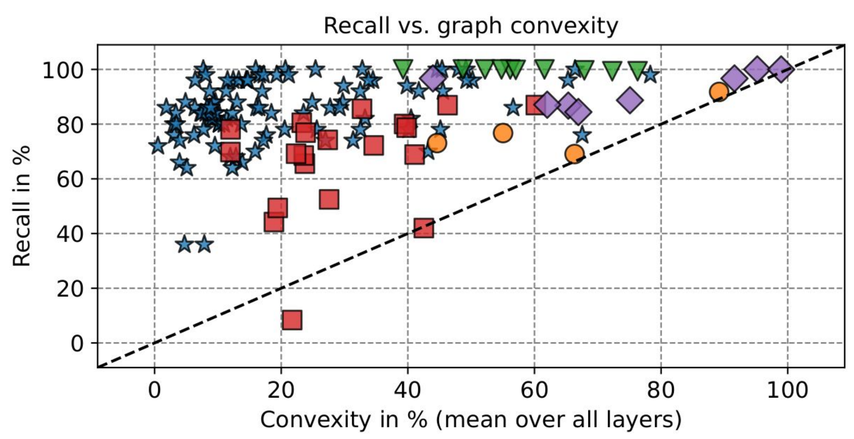
- Euclidean is the "classical" convexity, graph convexity is relevant for data on curved manifolds, resembles connectivity.
- Pretrained and fine-tuned models.
- Five modalities: images, text, audio, human activity recognition, medical images.

Does higher convexity in a pretrained model create a higher potential for generalizability (i.e., better performance)?



Results





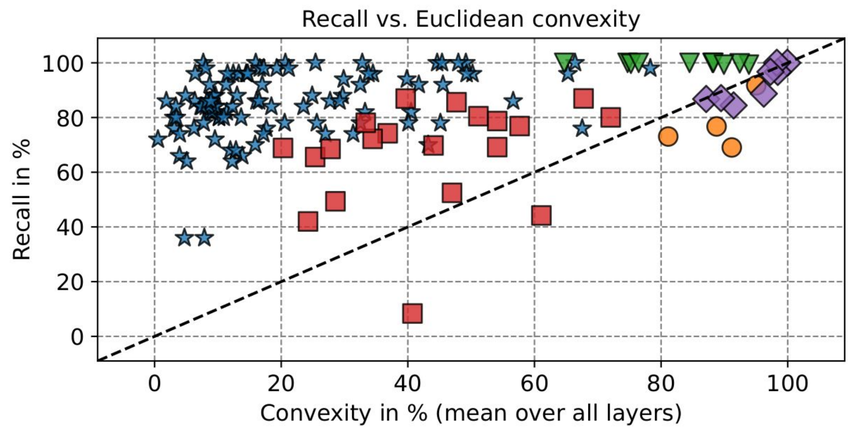


Figure 4: Euclidean and graph convexity scores for all modalities for decision regions of pretrained and fine-tuned networks. Decision regions are determined by model-predicted labels. Prediction in pretrained models is found using a softmax probe, training only the softmax linear layer. In fine-tuning, we train the whole network and softmax output. We find pervasive convexity in all networks and convexity further increases following fine-tuning. The number of layers differs across models but the most-left layer is the first layer that we observe and the right-most layer is the last layer in each model. Error bars are omitted in this plot for clarity (uncertainty estimates are given in figures of individual modalities in Appendices C.1-C.5). Note that the results are not directly comparable across modalities (see Section 2.1). In particular, we find less convexity in the image data containing a high number of classes (C=1000).

Figure 5. Graph convexity (top) and Euclidean convexity (bottom) of a subset of classes in the pretrained models vs. recall rate of these individual classes in the fine-tuned models for all data domains. The number of points for each domain is equal to the number of classes in this domain (except for images, where we take only a subset of classes for clarity – all image classes are shown in and Figure 9 and Figure 10 in Appendix C). The Pearson correlation coefficient is 0.22 ± 0.06 for graph convexity and 0.24 ± 0.06 for Euclidean convexity (the confidence intervals are computed using Fisher transformation).

Conclusions

- Convexity emerges across networks.
- Fine-tuning increases convexity.
- Higher convexity in pre-trained model ⟶ Better performance of the fine-tuned model.

lenhy@dtu.dk
Lenka Tětková
@LenkaTetkova

References

[1] Peter Gärdenfors. Induction, conceptual spaces and ai. Philosophy of Science, 57(1):78–95, 1990.
[2] Peter Gärdenfors. The geometry of meaning: Semantics based on conceptual spaces. MIT press, 2014.
[3] Peter Gärdenfors. Concept learning: a geometrical model. In Proceedings of the Aristotelian Society (Hardback), volume 101, pp. 163–183. Wiley Online Library, 2001.
[4] Peter Gärdenfors, Jürgen Jost, and Massimo Warglien. From actions to effects: Three constraints on event mappings. Frontiers in psychology, 9:1391, 2018.
[5] Tilen Marc and Lovro Šubelj. Convexity in complex networks. Network Science, 6(2):176–203, 2018.