Latent space navigation

interpretation, probing and steering

Lenka Tětková, Technical University of Denmark

Disclaimer

- I work on explainability, human-machine alignment and representation learning (postdoc at Technical University of Denmark).
- I am **not** a verification person.
- Let’s find how our fields can benefit each other!

Outline

- Motivation: why latent spaces?
- Levels of granularity
  - Micro
  - Meso
  - Macro
- Limitations & Discussion

Motivation & Background on Latent Spaces

Motivation

The problem: Neural networks are black boxes—how can we understand and control their outputs?

What are latent spaces?

- Latent space = internal representation a neural network learns to encode meaning.
- Compressed encodings of input/output (e.g., embeddings, hidden layers).
- Every input (image, sentence, sound) is mapped into a vector in this space.

What are latent spaces?

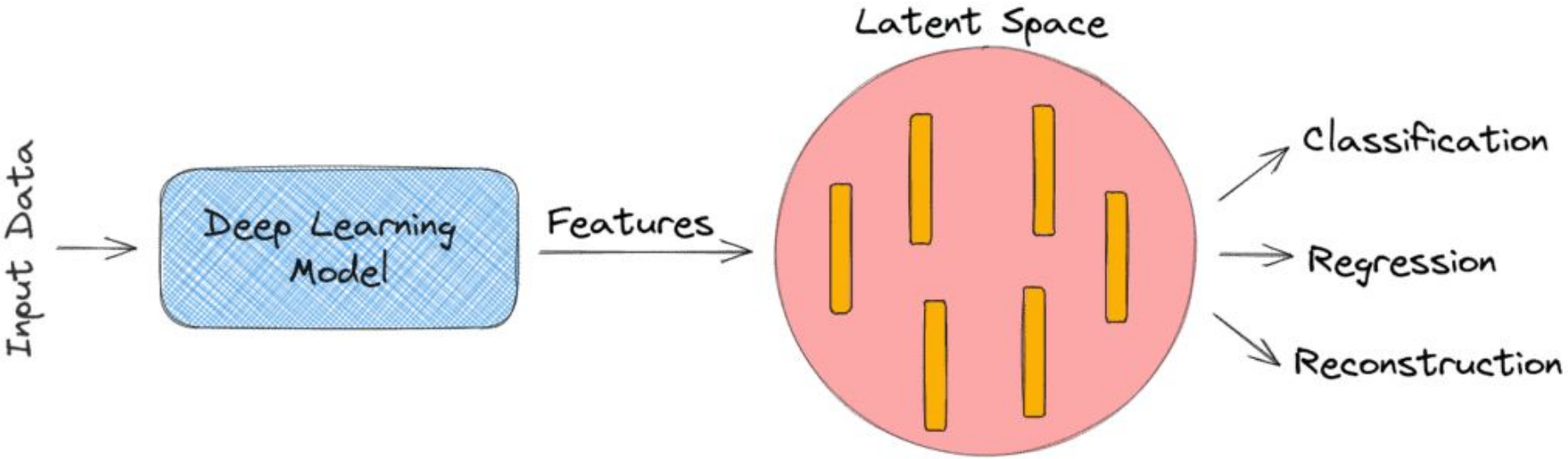


Figure from https://www.baeldung.com/cs/dl-latent-space.

Why latent spaces?

Heatmap-like explanations (e.g. attributions over pixels) need not reflect what the model actually computes – so look directly at the internal representations.

- Understanding → interpretability, science, transparency.
- Improvement → debugging, pruning, transfer learning
- Control → steering, disentanglement, creative uses.
- Responsibility → bias mitigation, safety, alignment, explainability.

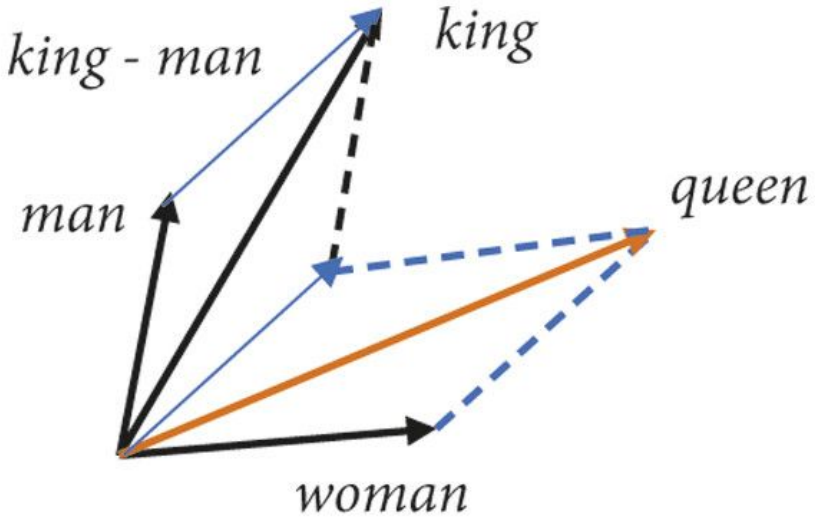
Properties

- Latent = hidden: not directly observable, but inferred.
- Compressed representation: captures essential features, ignoring raw noise/details.
- Structured geometry: distances and directions in this space often correspond to semantic relationships.
- Emergent structure: not explicitly programmed

Key takeaway: **Latent spaces are *where meaning lives*.**

Steering them = influencing outcomes.

Example: Word2Vec



$$king - man + woman \approx queen$$

Figure 1(a) from Liang, Wentao, Wang, Lu, She, Jialuo, Liu, Yuqing, Detecting Resource Release Bugs with Analogical Reasoning, Scientific Programming, 2022, 3518673, 9 pages, 2022. https://doi.org/10.1155/2022/3518673

Techniques for Discovering & Steering Latent Directions

Levels of granularity

Micro: individual neurons

Meso: vectors, concepts

Macro: society, values, power dynamics

Steering = shifting a latent vector in a particular direction to cause a controlled change in the output.

- freezing/removing some neurons
- vector arithmetics to increase/decrease certain attribute

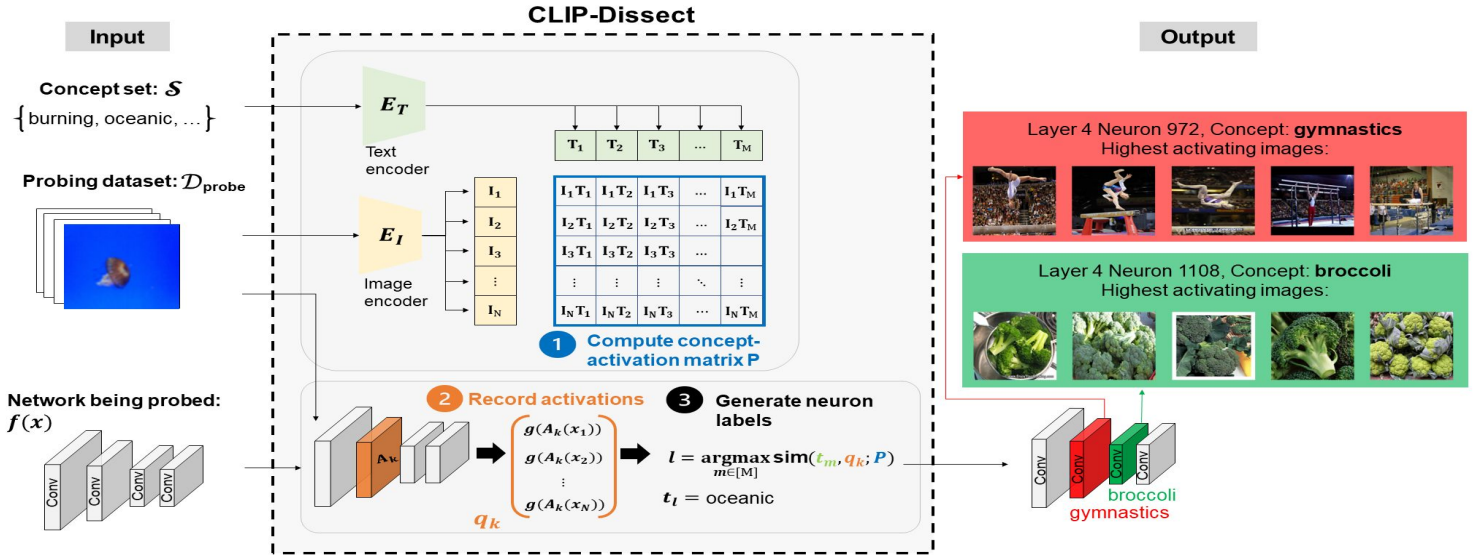
Micro level: neuron labeling (methods)

- Feature visualization [1] (Optimize an input image/text to maximally activate a neuron)
- Activation maximization [2] (patches from real images) + manual labelling

[1] Olah, et al., "Feature Visualization", Distill, 2017.
[2] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Cham: Springer International Publishing.

Micro level: neuron labeling (methods)

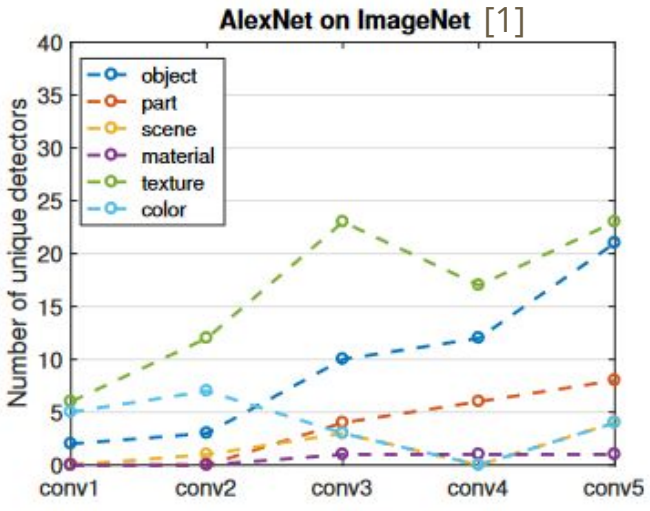
LLMs for labelling, e.g. CLIP-dissect [1]

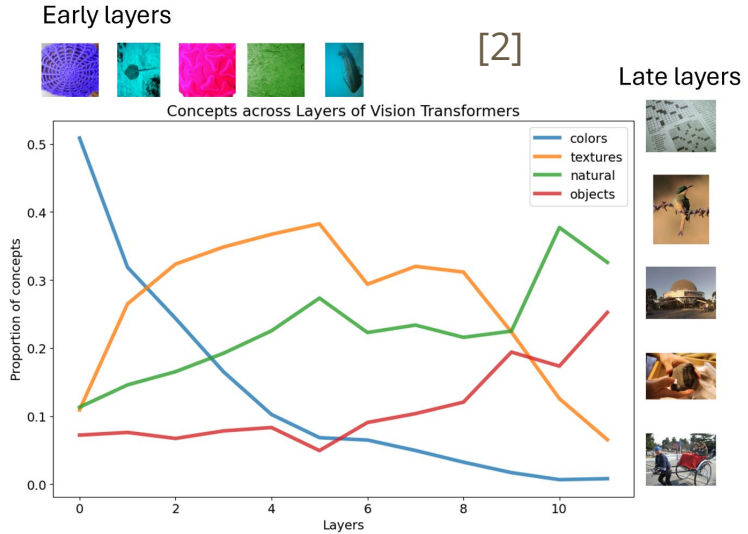


[1] Oikarinen, Tuomas, and Tsui-Wei Weng. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks." ICLR 2023.

Micro level: neuron labeling (use cases)

- Understanding: where are specific concepts encoded [1, 2]

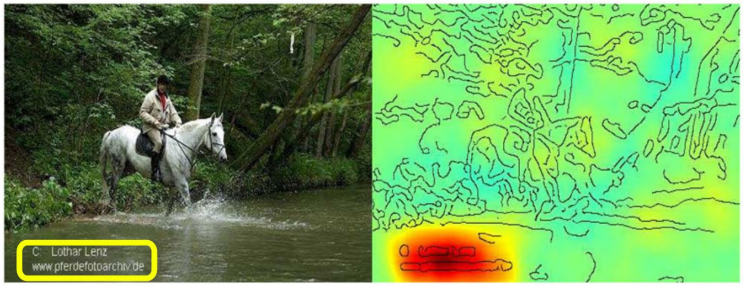




[1] Bau, David, et al. "Network dissection: Quantifying interpretability of deep visual representations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[2] Dorszewski, T., Tětková, L. et al. From Colors to Classes: Emergence of Concepts in Vision Transformers. The 3rd World Conference on eXplainable Artificial Intelligence (2025).

Micro level: neuron labeling (use cases)

- Pruning [1] - unhelpful or redundant neurons
- Bias mitigation [1]
- Detecting spurious correlations [2]
- Safety auditing [3]



[1] Dreyer, M., Berend, J., Labarta, T., Vielhaben, J., Wiegand, T., Lapuschkin, S., & Samek, W. (2025). Mechanistic understanding and validation of large AI models with SemanticLens. Nature Machine Intelligence, 1-14.
[2] Singla, S., & Feizi, S. (2022). Salient ImageNet: How to discover spurious features in Deep Learning?. In International Conference on Learning Representations.
[3] Bricken Trenton, Templeton Adly, Batson Joshua, Chen Brian, Jermyn Adam, Conerly Tom, Turner Nick, Anil Cem, Denison Carson, Askell Amanda, et al. 2023.Towards Monosemanticity: Decomposing Language Models With Dictionary Learning.Transformer Circuits Thread (2023).
Figure 2a from Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), 1096.

Micro level: sparse autoencoders (SAEs)

Idea: individual neurons often encode multiple concepts (superposition).

SAEs [1] can disentangle these into sparse, interpretable features.

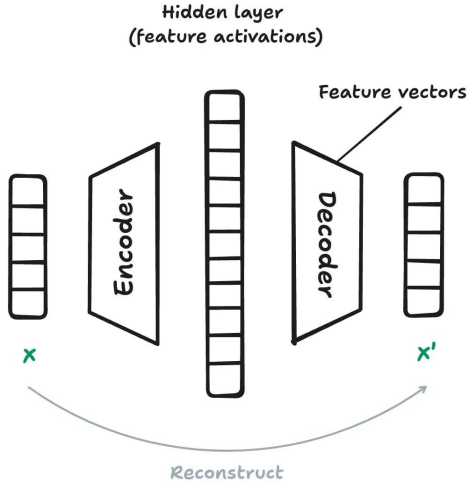
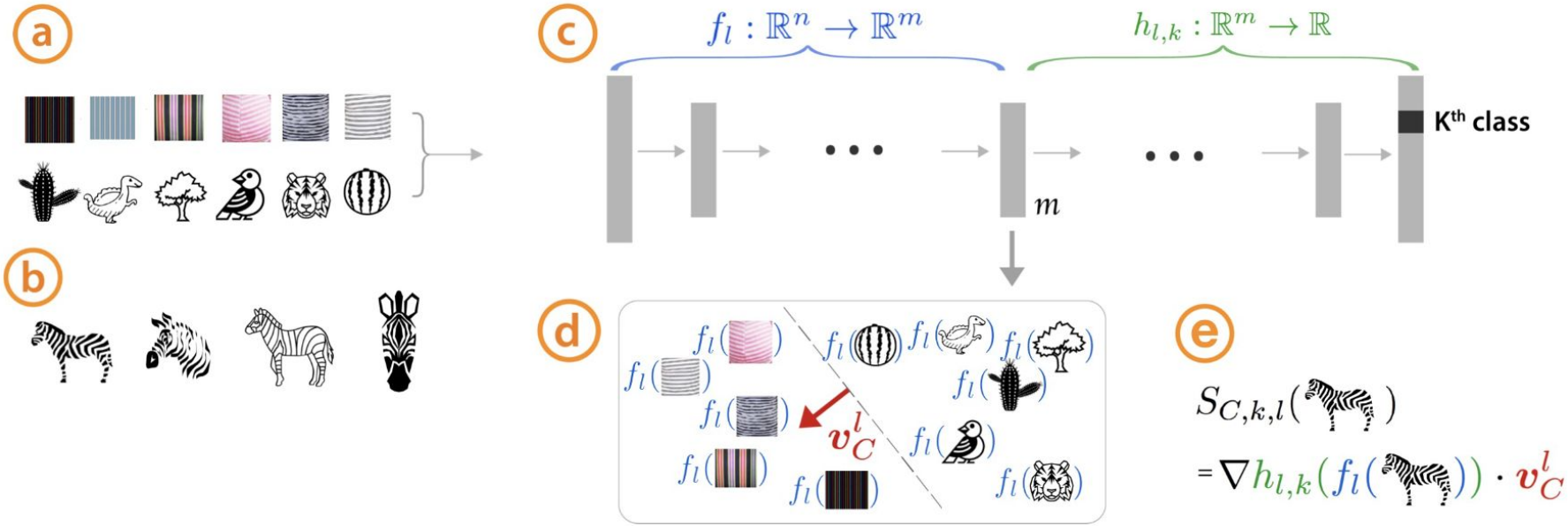


Figure from https://www.lesswrong.com/posts/8YnHuN55XJTDwGPMr/a-gentle-introduction-to-sparse-autoencoders

[1] Bricken Trenton, Templeton Adly, Batson Joshua, Chen Brian, Jermyn Adam, Conerly Tom, Turner Nick, Anil Cem, Denison Carson, Askell Amanda, et al. 2023.Towards Monosemanticity: Decomposing Language Models With Dictionary Learning.Transformer Circuits Thread (2023).

Meso level: linear probes [1]/concept activation vectors (CAVs) [2]

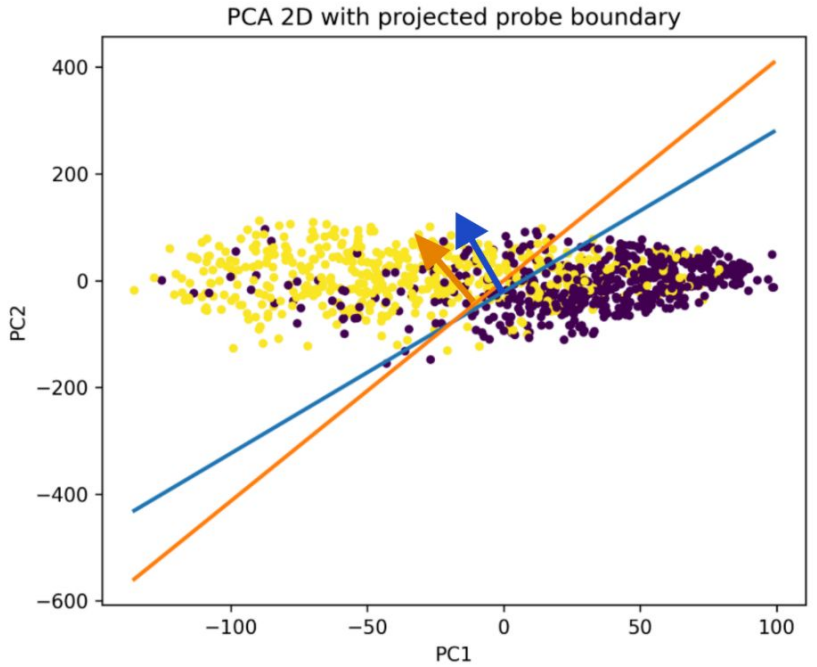


[1] Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644
Figure 1. from [2] Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018..

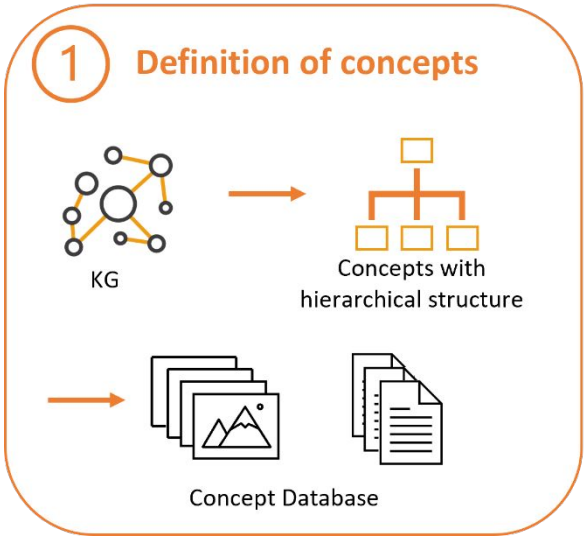
Meso level: linear probes/concept activation vectors (CAVs)

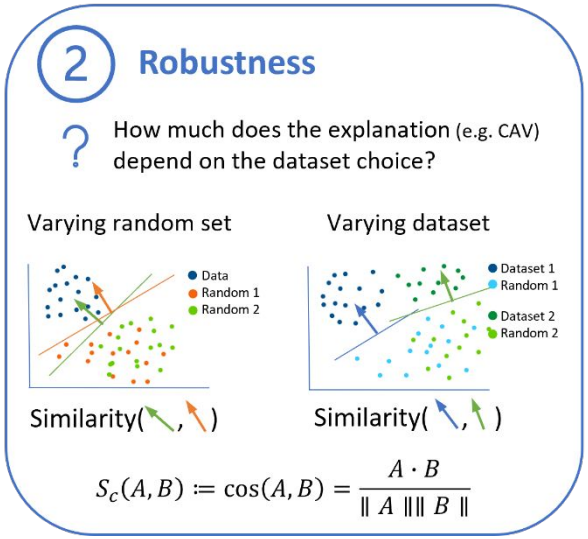
Methods for obtaining the concept activation/steering vector:

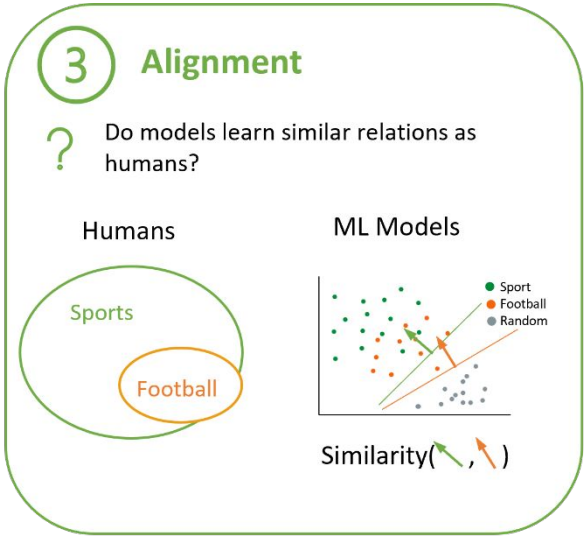
- Linear classifier
- PCA
- Mean difference
- …



Meso level: where do concept datasets come from?







Tětková, L., Scheidt, T. K., Fogh, M. M., Jørgensen, E. M. G., Nielsen, F. Å., & Hansen, L. K. (2024, July). Knowledge graphs for empirical concept retrieval. In World Conference on Explainable Artificial Intelligence (pp. 160-183). Cham: Springer Nature Switzerland.

Meso level: steering generative models

Attribute vector arithmetic.

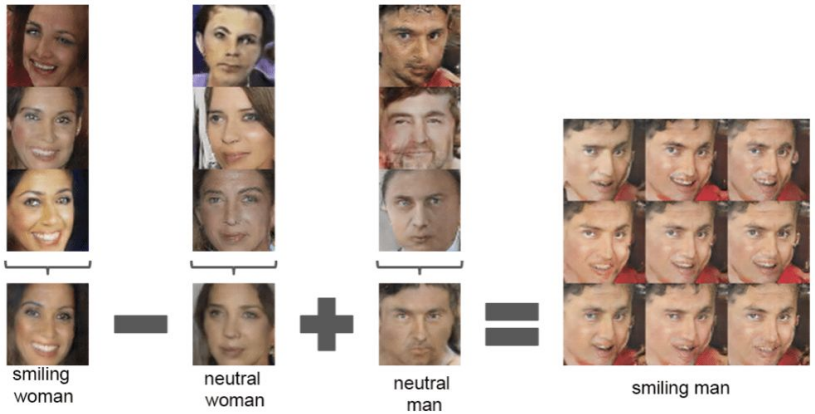


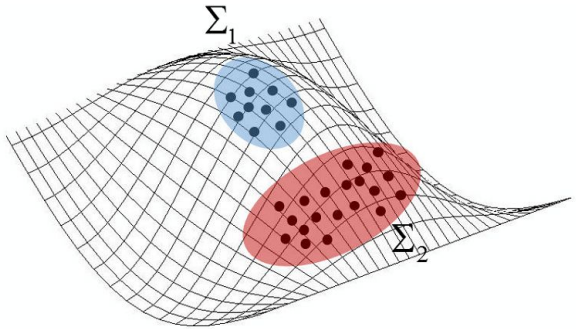
Figure 7 from Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Unsupervised concept discovery – PCA or ICA to find axes of variation.

Figure 1 from Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). Ganspace: Discovering interpretable gan controls. Advances in neural information processing systems, 33, 9841-9850.

Meso level: geometry of latent spaces

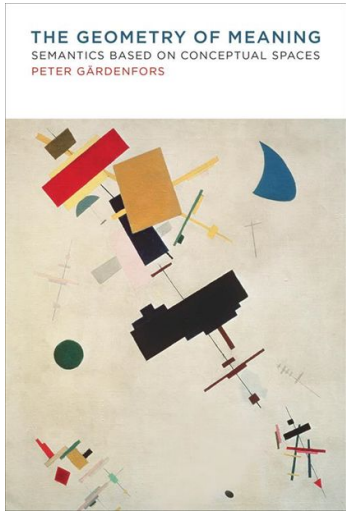
Representations concentrate on curved, low-dimensional structures (manifolds) [1, 2].

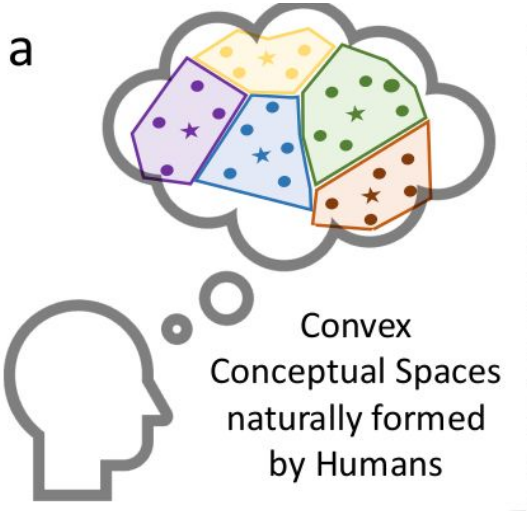


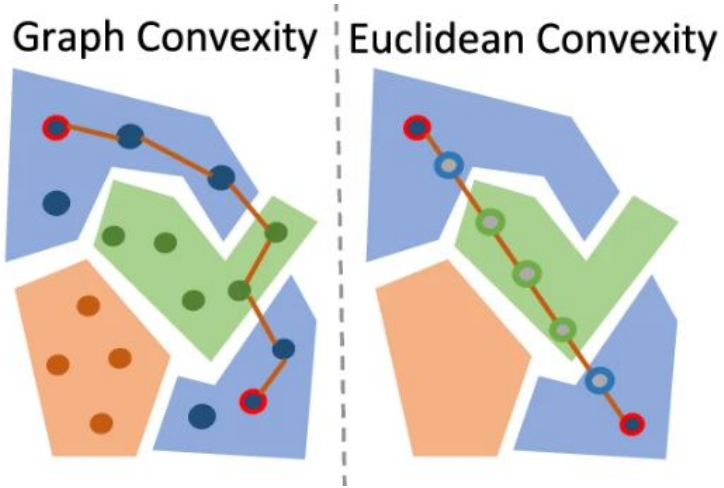
What shape does a concept have?

[1] Cosentino, R., Shekkizhar, S., Soltanolkotabi, M., Avestimehr, S., & Ortega, A. (2022). The geometry of self-supervised learning models and its impact on transfer learning. arXiv preprint arXiv:2209.08622.
[2] Arvanitidis, G., Hansen, L. K., & Hauberg, S. (2018, January). Latent space oddity: On the curvature of deep generative models. In 6th International Conference on Learning Representations, ICLR 2018.
Figure 4b from Lahav, A., Talmon, R., & Kluger, Y. (2019). Mahalanobis distance informed by clustering. Information and Inference: A Journal of the IMA, 8(2), 377-406.

Meso level: convexity – a hypothesis from cognitive science





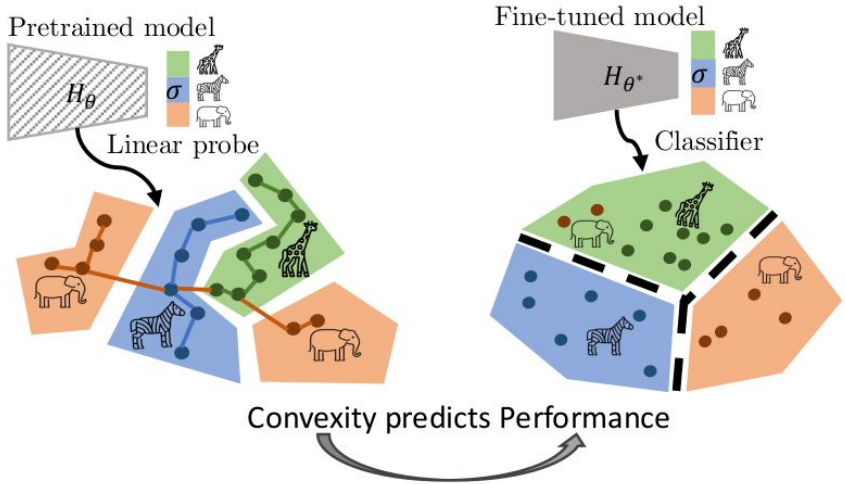


P. Gärdenfors, The geometry of meaning: Semantics based on conceptual spaces. MIT press, 2014.
P. Gärdenfors, Conceptual Spaces: The Geometry of Thought. The MIT Press, 03 2000.

Tětková, L., Brüsch, T., Dorszewski, T., Mager, F. M., Aagaard, R. Ø., Foldager, J., ... & Hansen, L. K. (2025). On convex decision regions in deep network representations. Nature Communications, 16(1), 5419.

Meso level: convexity – findings and applications

- Concept regions are approximately convex – pervasively, across images, text, audio, human activity, and medical data.
- Convexity of pretrained representations predicts fine-tuning performance



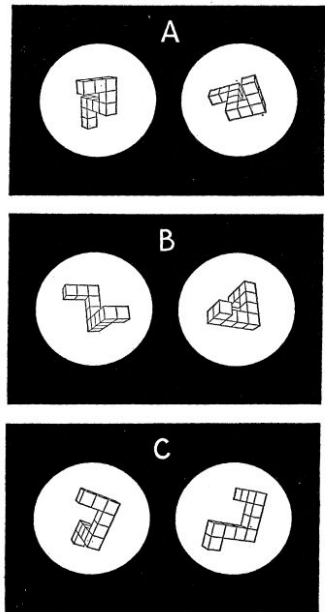
→ Criteria for evaluation of self-supervised learning [1, 2]

→ Layer pruning [3]

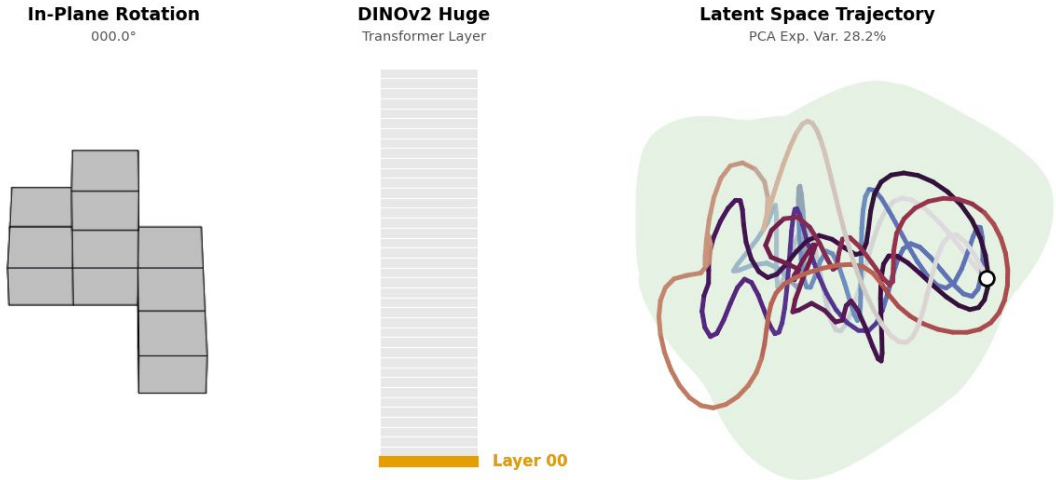
[1] Tětková, L., Brüsch, T., Dorszewski, T., Mager, F.M., Aagaard, R.Ø., Foldager, J., Alstrøm, T.S. and Hansen, L.K., 2025. On convex decision regions in deep network representations. Nature Communications, 16(1), p.5419.
[2] Cosentino, R., Shekkizhar, S., Soltanolkotabi, M., Avestimehr, S., & Ortega, A. (2022). The geometry of self-supervised learning models and its impact on transfer learning. arXiv preprint arXiv:2209.08622.
[3] Dorszewski, T., Tětková, L., and Hansen, L. K. (2024). Convexity-based pruning of speech representation models, 2024 IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2024.

Meso level: mental rotations

Human response time grows with rotation angle – evidence of analog mental rotation [1].



We ask: do large vision models show the same behavior in their latent trajectories? [2]



[1] Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171(3972), 701-703.
[2] Mason, S. R., Gjølbye, A., Højbjerg, P. C., Tětková, L., & Hansen, L. K. (2026, May). Large vision models can solve mental rotation problems. In 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1571-1575). IEEE.

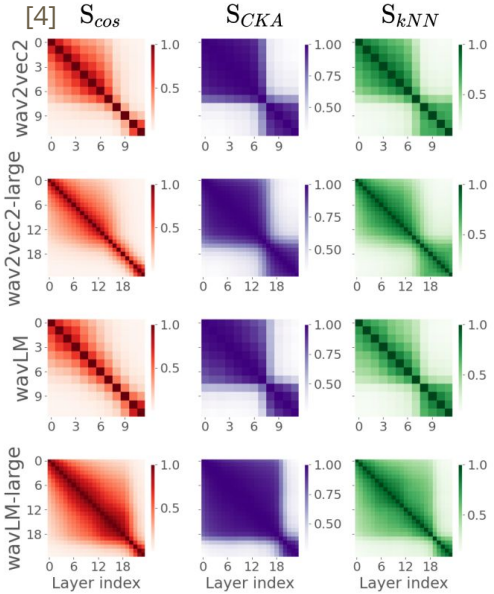
Meso level: similarity of representations, alignment

Measure similarity of representations: compare latent structures across layers, models, or even human brain activity.

- Pairwise similarity of data points → RSA [1]
- Linear projections that maximize correlation → CCA [2]
- Nearest neighborhoods → (mutual kNN) [3]

→ Prune the most similar layers [4]

→ Human-machine alignment [5]



[1] Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 249.
[2] Canonical Correlation Analysis. In Härdle, W., & Simar, L. (2007). Applied multivariate statistical analysis. Berlin, Heidelberg: Springer Berlin Heidelberg.
[3] Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. arXiv preprint arXiv:2405.07987.
[4] Dorszewski, T., Jacobsen, A. K., Tětková, L., & Hansen, L. K. (2025, April). How Redundant Is the Transformer Stack in Speech Representation Models?. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
[5] Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023) Human alignment of neural network representations. In The Eleventh International Conference on Learning Representations.

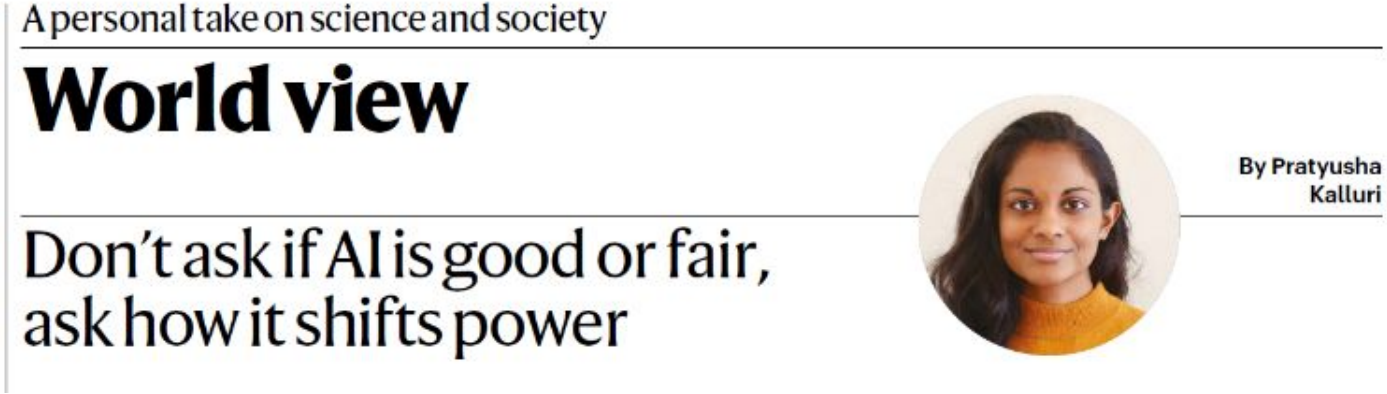
Macro level

Focus: latent spaces as levers for global behavior control & alignment with human goals.

Whose values are encoded?

What values should guide steering?

Who gets to decide?



A personal take on science and society

World view

Don't ask if AI is good or fair, ask how it shifts power

By Pratyusha Kalluri

Limitations, Risks, and Future Directions

Limitations, Risks, and Future Directions

- Entanglement of latent factors → imperfect control
- Linearity assumption
- Dataset dependence (biases)
- Lack of standardized benchmarks
- Risks: malicious use (deepfakes, manipulation)
- Computationally heavy, limited scalability
- Evaluating explanations is hard – ground truth missing or ill-defined

Acknowledgements

Anders Gjølbye Madsen, Sebastian Ray Mason, Mikkel Godsk Jørgensen, Sarah Masud, Lars Kai Hansen

DTU

UNIVERSITY OF COPENHAGEN

lenhy@dtu.dk

Lenka Tětková

lenkatetkova.github.io

PIONEER CENTRE FOR AI

novo nordisk fonden

Danish Data Science Academy

Try it yourself

Colab notebooks, no setup needed

Round table: what I'd like to learn from you

- What would it mean to "verify" a claim about a concept?
- Is approximate structure (e.g., approximate convexity) usable by formal tools?
- How would you evaluate an explanation?
- ...


lenhy@dtu.dk


Lenka Tětková


lenkatetkova.github.io

Try it yourself