

Teresa Dorszewski<sup>1</sup>, Lenka Tětková<sup>1</sup>, Robert Jenssen<sup>2, 3, 4</sup>, Lars Kai Hansen<sup>1</sup>, Kristoffer Knutsen Wickstrøm<sup>2</sup>

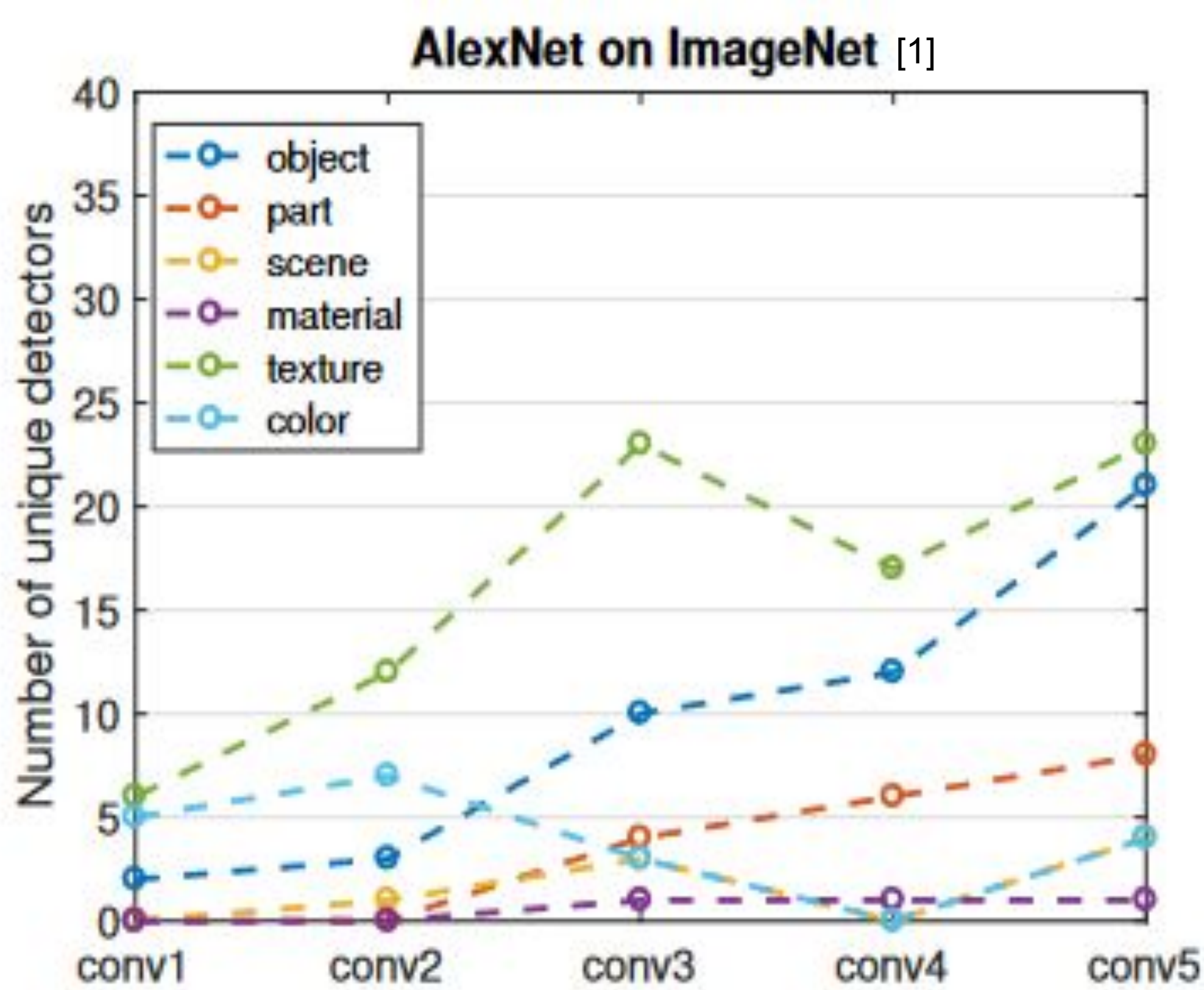
<sup>1</sup> Technical University of Denmark, <sup>2</sup> UiT The Arctic University of Norway,

<sup>3</sup> Pioneer Centre for AI, University of Copenhagen <sup>4</sup> Norwegian Computing Center, Oslo, Norway

Presented at The 3rd World Conference on eXplainable Artificial Intelligence (July 2025, Istanbul)

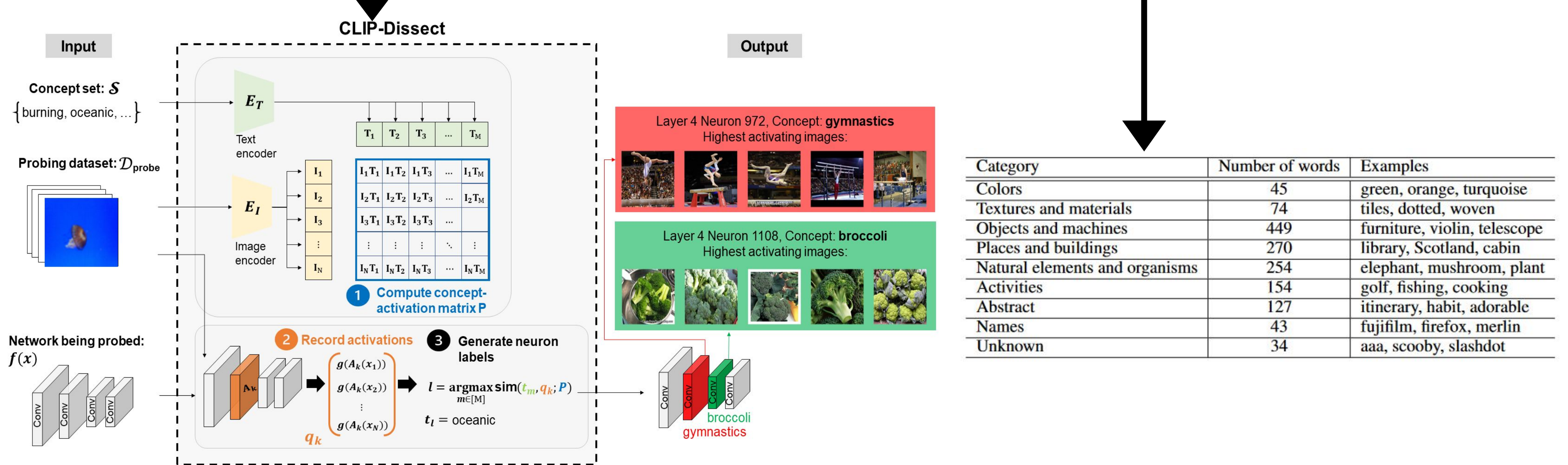
## Motivation

- Understand what neural networks have learned through human-interpretable concepts.
- Heavily studied in CNNs, showing that early layer learns simple concepts and later layers learn complex concepts.
- Less studied in Vision Transformers (ViTs).



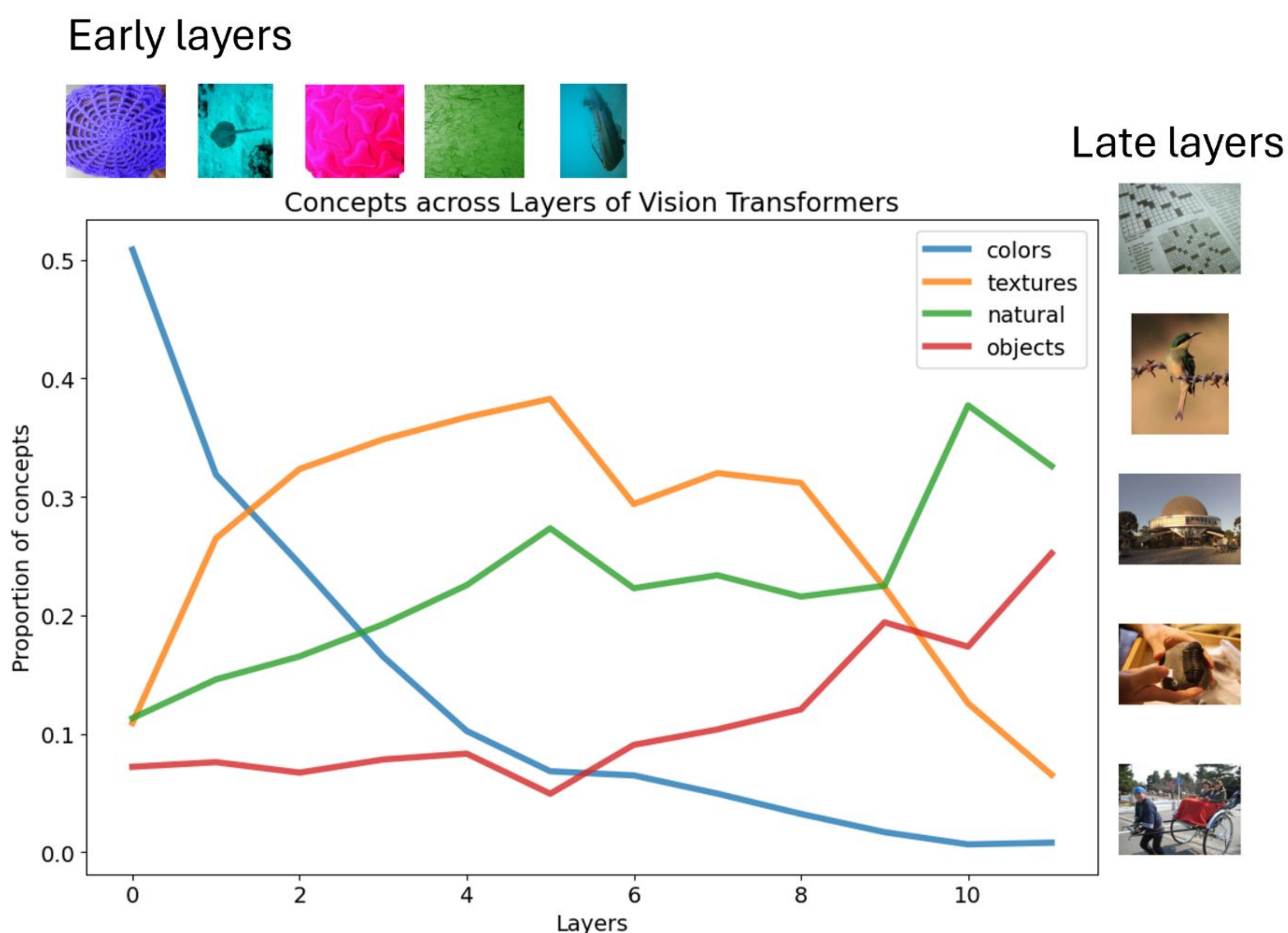
## Methods

- Analyse 4 common ViTs (Supervised ViT, CLIP, Dino v2, Masked Autoencoder (MAE))
  - 12 Layers, 768 output neurons
- CLIP-dissect [2] for neuron labeling
  - Concepts: 20k most common words
  - Probing dataset: ImageNet + Broden
- Fine-tuning: BloodMNIST and CUB
- Labels divided into 9 categories
- Measure image complexity with ICNet [3]

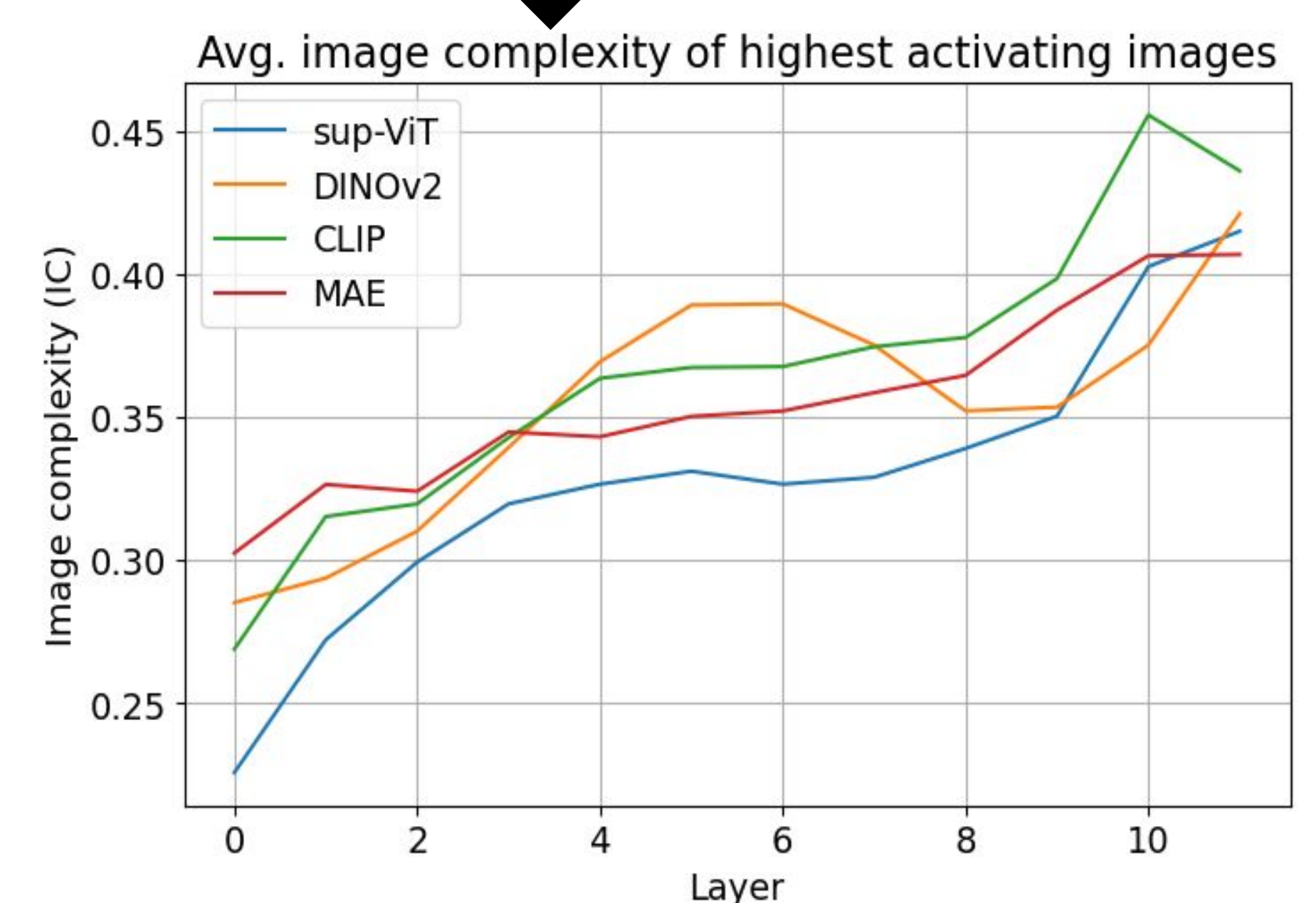
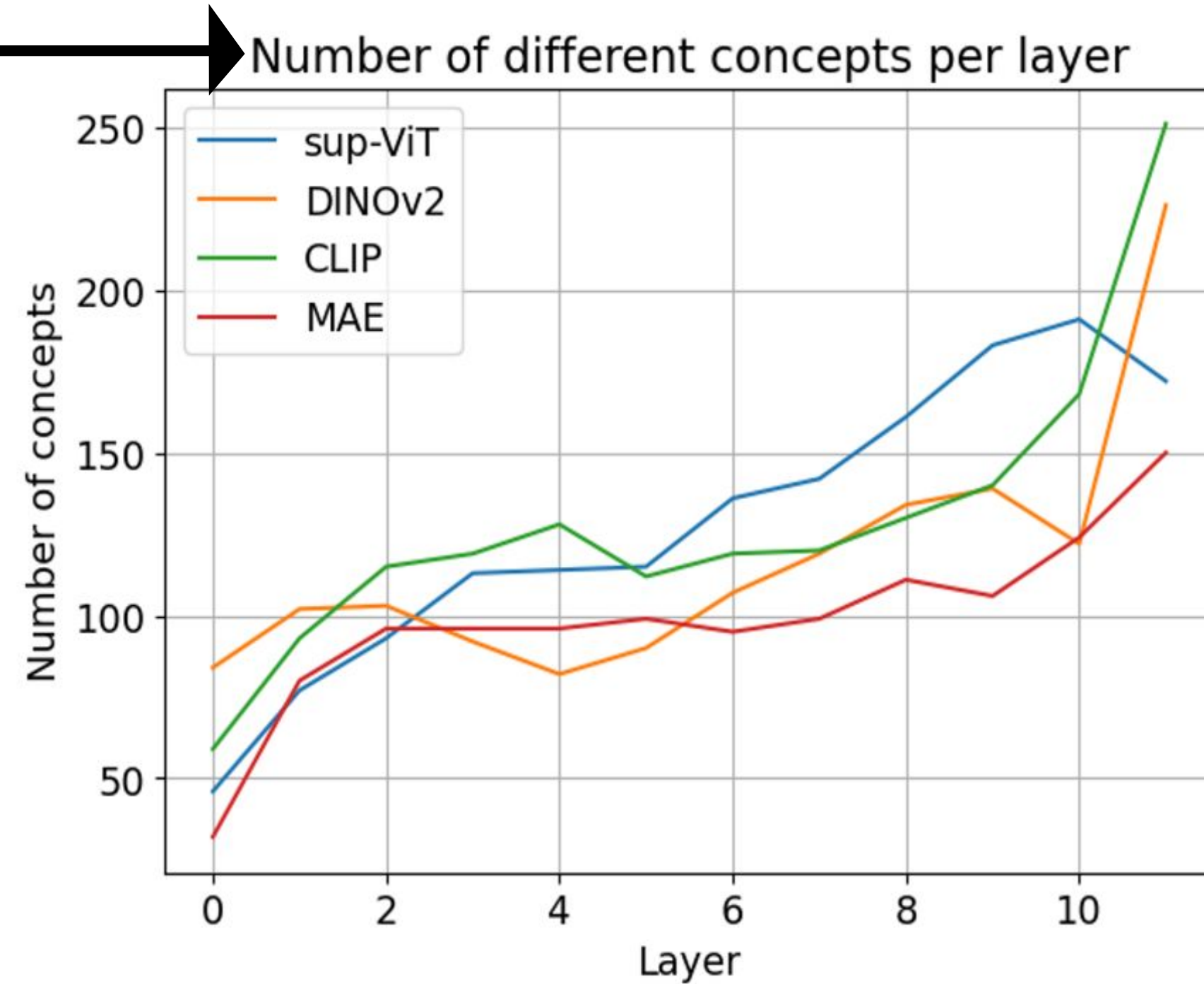


## Results

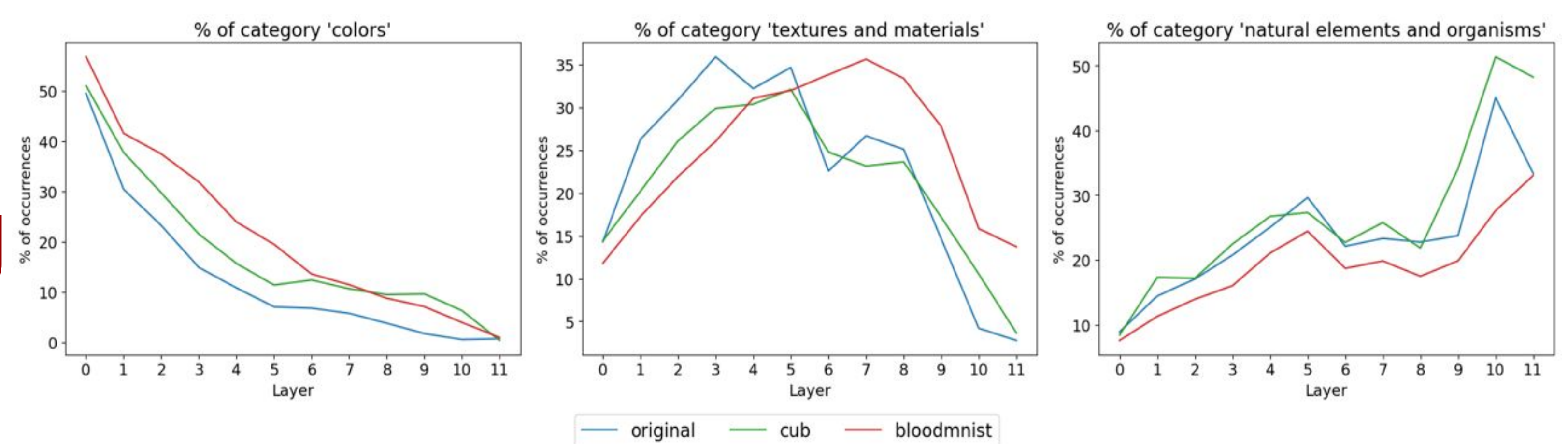
- Neurons in early layers mainly activate on colors and patterns.
- Neurons in late layers activate on complex images with objects or animals.



- Number of concepts increases throughout the network.
- Complexity increases throughout the network.



- Fine-tuning increases the relevant concepts and decreases irrelevant ones (and total number of distinct concepts).



## Conclusions

- ViTs learn concepts similar to CNNs.
- Fine-tuning alters learned concepts
- Complexity of concepts increases.

## References

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

[3] Feng, Tinglei, et al. "Ic9600: A benchmark dataset for automatic image complexity assessment." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.7 (2022)

Paper

Code



lenhy@dtu.dk



Lenka Tětková



@lenkatetkova.bsky.social