From Colors to Classes: Emergence of Concepts in Vision Transformers

We present the first comprehensive layer-wise analysis of concept-learning in Vision Transfomers.

Teresa Dorszewski[3], **Lenka Tětková**[3], **Robert Jenssen**[1,2,4,5,6], **Lars Kai Hansen**[3], **Kristoffer Knutsen Wickstrøm**[1,2]
UiT The Arctic University of Norway[1]
Visual Intelligence[2], Technical University of Denmark[3]
Norwegian Computing Center[4], Pioneer Centre for AI[5], University of Copenhagen[6]

Concept analysis in deep learning

- 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) [1].

Concept analysis of ViTs

- ViTs can attend entire image in first layer.
- Are concepts still learned from simple to complex?
- Prior studies focused on final layer of ViTs.

CLIP-dissect for neuron labeling

- Label each neuron in ViT-layers with CLIP-Dissect [2].
- Concept are the 20k most common words in English (See Tab. 2).
- Broden dataset as image probing.
- Threshold concept discovery for reliable analysis.

model	threshold τ	# of labeled neurons
ViT	0.17 ± 0.01	317 ± 13
DINOv2	0.21 ± 0.04	323 ± 11
CLIP	0.17 ± 0.03	309 ± 13
MAE	0.17 ± 0.02	331 ± 17
ResNet50	0.21 ± 0.05	$32, 92, 209, 423, 893$

Table 1: Mean threshold τ across layers for each analyzed model and the number of neurons with similarity scores above τ.

Experiments

- Investigate numerous vision models.
- Measure concepts across layers (Fig. 2).
- Measure complexity of concepts (Tab. 3).
- Investigate effect of finetuning (Fig.1).

Results

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

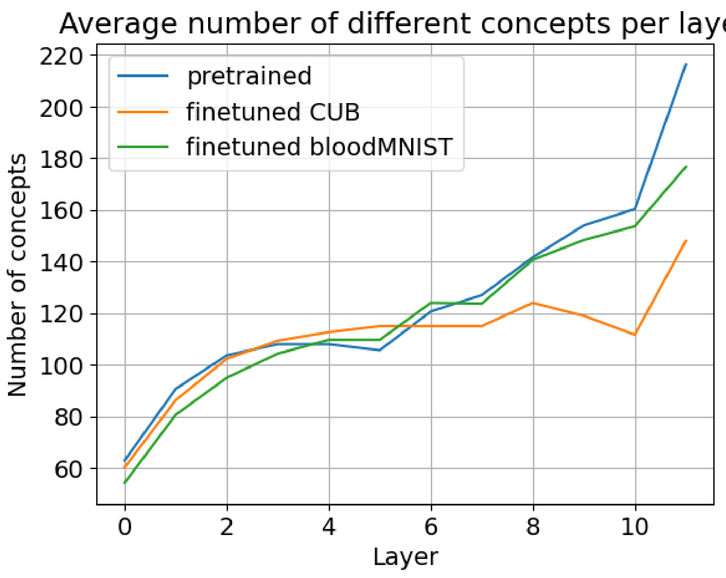

Figure 1: Number of different concepts averaged over three ViT models in their pretrained and finetuned versions.

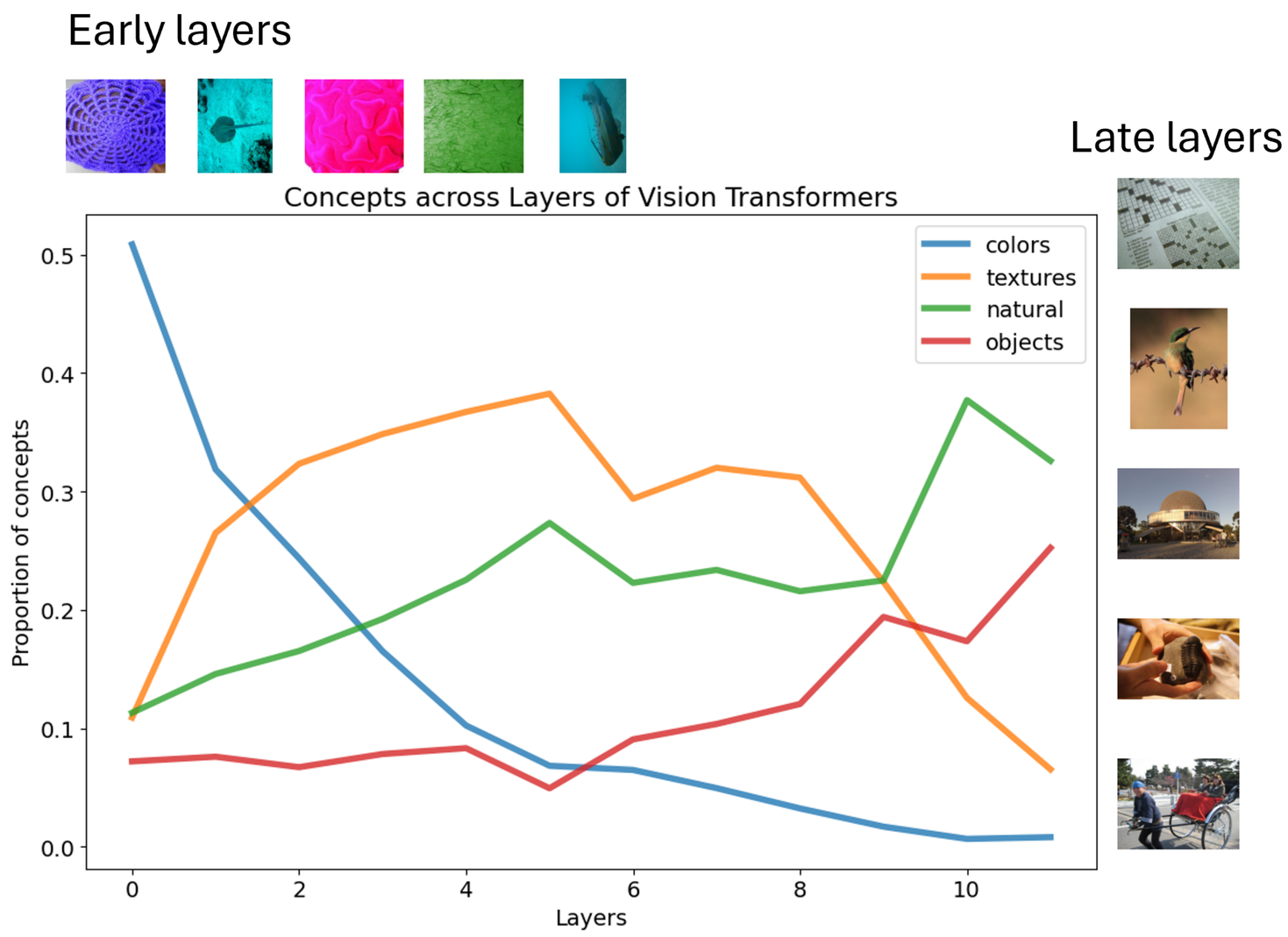

Figure 2: We analyze how different concepts develop across the layers of vision transformers. Early layers tend to process simpler concepts and images while late layers focus on more complex and diverse concepts.

Vocabulary and complexity of concepts

Category	Number of words	Examples
Colors	45	green, orange
Textures and materials	74	tiles, woven
Objects and machines	449	furniture, violin
Places and buildings	270	library, cabin
Natural elements and organisms	254	elephant, plant
Activities	154	golf, fishing
Abstract	127	itinerary, habit
Names	43	fujifilm, firefox
Unknown	34	aaa, scooby

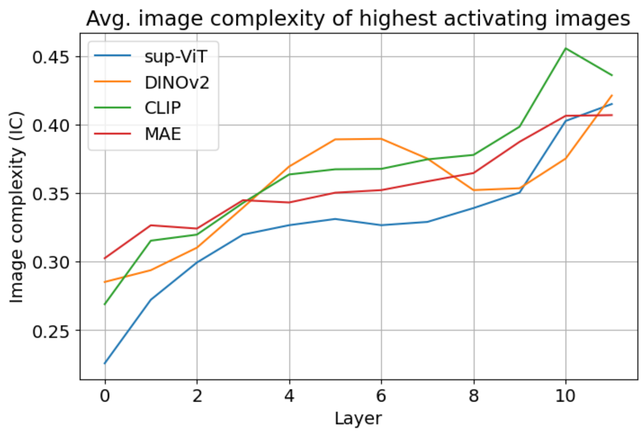

Table 2: (left) Frequency of words in our semi-manual categorization and (right) complexity of the five highest activating images for each neuron averaged across layers measured by ICNet [3]. All models show an increase in image complexity across layers.

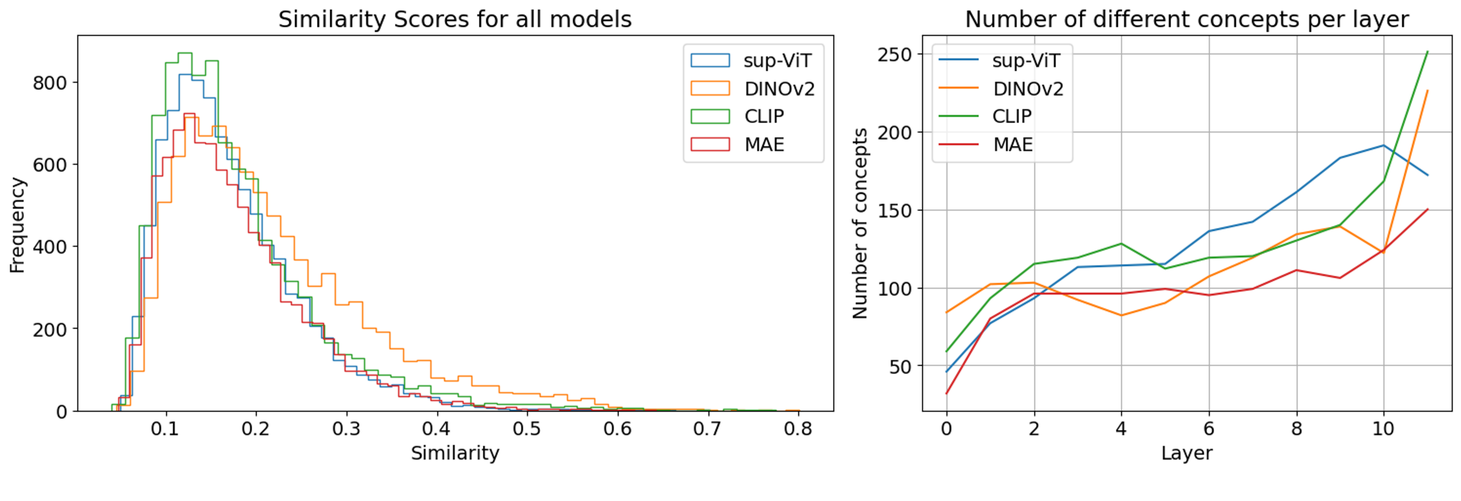

Figure 3: Similarity scores for concepts and number of different concepts encoded in each layer.

Summary

- First layer-wise analysis of concepts in ViTs.
- ViT's learn concepts hierarchically.
- Concepts are "forgotten" after finetuning.

Paper

References

1 Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, ICLR, 2021.
2 Oikarinen et al., *CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks*, ICLR, 2023.
3 Feng et al., *IC9600: A Benchmark Dataset for Automatic Image Complexity Assessment*, TPMAI, 2022.