Understanding Video Transformers via Universal Concept Discovery

Matthew Kowal[1,3,4] Achal Dave[3] Rares Ambrus[3]
Adrien Gaidon[3] Konstantinos G. Derpanis[1,2,4] Pavel Tokmakov[3]
[1]York University, [2]Samsung AI Centre Toronto, [3]Toyota Research Institute, [4]Vector Institute
Project page: yorkucvil.github.io/VTCD

Understanding the hidden representations within neural networks is essential for addressing regulatory concerns [3, 10], preventing harms in deployment [2, 9], and can aid innovative model designs [4]. This problem has been studied extensively for images, both for convolutional neural networks (CNNs) [1, 7, 8, 11] and, more recently, vision transformers (ViTs) [13, 15]. However, while video transformers do share their overall architecture with image-level ViTs, the insights obtained in existing works do very little to explain their inner mechanisms. Consider, for example, the recent approach for tracking occluded objects [14] shown in Figure 1 (top). To accurately reason about the trajectory of the invisible object inside the pot, texture or semantic cues alone would not suffice. What, then, are the *spatiotemporal* mechanisms used by this approach? Are any of these mechanisms *universal* across video models trained for different tasks?

In this work we present the Video Transformer Concept Discovery algorithm (VTCD) - the first concept discovery method for interpreting the representations of video transformers. We focus on concept-based interpretability [6–8, 16] due to its capacity to explain the decision-making process of a complex model's distributed representations in high-level, intuitive terms. Our goal is to decompose a representation at any layer into human-interpretable 'concepts' without any labelled data (*i.e*. concept discovery) and then rank them in terms of their importance to the model output.

Concretely, we first group *model features* at a given layer into spatiotemporal tubelets, which serve as a basis for our analysis. Next, we cluster these tubelets across videos to discover high-level concepts. The resulting concepts for an occluded object tracking method [14] are shown in Figure 1 and span a broad range of cues, including spatiotemporal ones that detect events, like collisions, or track containers. To better understand the decision-making mechanisms of video transformers, we then quantify the importance of concepts for the model's predictions and propose a novel, noise-robust approach to estimate concept importance.

Next, we use VTCD to study whether there are any universal mechanisms in video transformer models, that emerge irrespective of their training objective. We extend recent work [5] to discover *important* concepts that are shared between several models. We analyze a diverse set of models (*e.g*. supervised, self-supervised, or video-language) and make several discoveries: (i) many concepts are shared between models trained for different tasks; (ii) early layers form spatiotemporal bases that underlines the information processing; (iii) later layers form object-centric representations, even when trained without supervision.

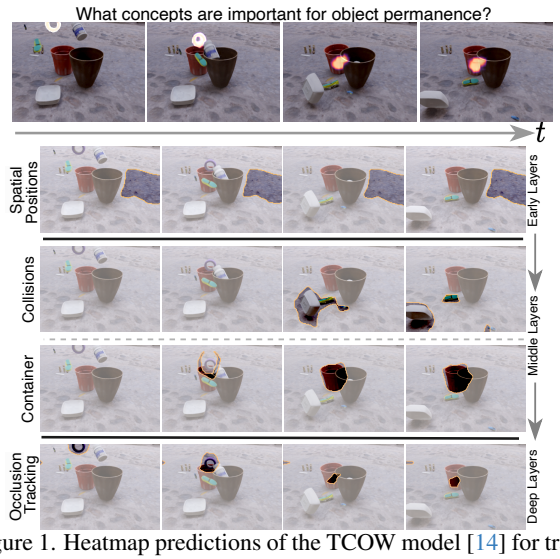

Figure 1. Heatmap predictions of the TCOW model [14] for tracking through occlusions (top), together with concepts discovered by our VTCD (bottom). We can see that the model encodes positional information in early layers, identifies containers and collision events in mid-layers and tracks through occlusions in late layers. Only one video is shown, but the discovered concepts are shared between many dataset samples (see video for full results).

We also show how VTCD can be applied for downstream tasks. Firstly, pruning the heads of an action classification model according to their estimated importance yields a 4.3% *increase* in accuracy while reducing computation by 33%. Secondly, object-centric concepts discovered by VTCD can be used for video-object segmentation (VOS) and achieve strong performance on the DAVIS'16 benchmark [12] even for self-supervised representations.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 1

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM FAccT*, 2018. 1

[3] European Commision. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *European Commision*, 2021. 1

[4] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 1

[5] Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *ICCV*, 2023. 1

[6] Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *NeurIPS*, 2023. 1

[7] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation factorization for explainability. In *CVPR*, 2023. 1

[8] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, 2019. 1

[9] Sven Ove Hansson, Matts-Åke Belin, and Björn Lundgren. Self-driving vehicles-An ethical overview. *Philosophy & Technology*, pages 1–26, 2021. 1

[10] The White House. President biden issues executive order on safe, secure, and trustworthy artificial intelligence. *The White House*, 2023. 1

[11] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018. 1

[12] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1

[13] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021. 1

[14] Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *CVPR*, 2023. 1

[15] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *CVPR*, 2023. 1

[16] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1