Abstract:

Large Vision & Language models pretrained on web-scale data provide representations invaluable for numerous V&L problems. However, it is unclear how they can be used for reasoning about user-specific visual concepts in unstructured language. We introduce a new learning setup called Personalized Vision & Language (PerVL) with two new benchmark datasets for retrieving and segmenting user-specific "personalized" concepts in the wild. We propose an architecture for solving PerVL that operates by extending the input vocabulary of a pretrained model with new word embeddings for the new personalized concepts. We demonstrate that our approach learns personalized visual concepts from a few examples and can effectively apply them in image retrieval and semantic segmentation using rich textual queries.

(Published as an oral presentation at ECCV2022. This work was done during an internship at NVIDIA Research Tel Aviv.) Bio: Niv is a Ph.D. student at The Hebrew University of Jerusalem, advised by Dr. Yedid Hoshen. He received his BSc. in mathematics with physics, and M.Sc. in physics, both from the Technion. He's interested in computer vision and representation learning with a focus on anomaly detection and scientific data.