COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Karan Desai

Language Supervision for Computer Vision

Wednesday, January 24, 2024
1:30pm – 3:30pm
3725 Beyster
Hybrid – Zoom
Passcode: 969459

ABSTRACT: Representation learning lies at the core of modern Artificial Intelligence. In computer vision, labeled image datasets like ImageNet have been the standard choice for representation learning. Despite being empirically successful, this approach is expensive to scale due to labeling costs. Moreover, the representation quality is limited by the size and diversity of datasets and their associated label ontologies.

My research explores using natural language supervision for computer vision. Using natural language allows us to go beyond fixed label ontologies and scale up to more general sources such as internet data. Toward this goal, my dissertation explores four problems -- (1) Learning representations: I propose one of the first methods for language-supervised visual learning that uses image captioning as the training objective, showing its efficacy compared to ImageNet-trained methods on downstream tasks like object detection and segmentation. (2) Scaling data: I explore social media as a rich source of high-quality image descriptions and curate a dataset of 12 million image-text pairs while ensuring responsible curation practices. (3) Understanding data: It is difficult to comprehend the diversity of visual concepts present in millions of image-text pairs. I posit that images and text naturally organize into a tree-like hierarchy, and propose an approach for learning representations that capture this hierarchy using tools from hyperbolic geometry. (4) Transfer to downstream tasks: Large vision-language models show impressive zero-shot transfer capabilities on image-level tasks like classification and retrieval. However, their transferability to pixel-level tasks like object detection and segmentation has so far relied on expensive labeled mask annotations. I propose an object detector to efficiently transfer pre-trained vision models to segment and classify visual objects without any fine-tuning, unlike existing detectors that train using orders of magnitude more labeled masks to achieve high performance. In summary, my research affirms that using language supervision can drive the next leap of progress in computer vision, and has immense utility in practical applications.

CHAIR: Prof. Justin Johnson