Learning from an Infant's Visual Experience

Deepayan Sanyal
Department of Computer Science
Vanderbilt University,
Nashville, TN, USA
deepayan.sanyal@vanderbilt.edu

Abstract

Infants see a selective view of the world: they see some objects with high frequency and from a wide range of viewpoints (e.g., their toys during playing) while a much larger set of objects are seen much more rarely and from limited viewpoints (e.g., objects they see outdoors). Extensive, repeated visual experiences with a small number of objects during infancy plays a big role in the development of human visual skills. Internet-style datasets that are commonly used in computer vision research do not contain the regularities that result from such repeated, structured experiences with a few objects. This has led to a dearth of models that learn by exploiting these regularities. In my PhD dissertation, I use deep learning models to investigate how regularities in an infant's visual experience can be leveraged for visual representation learning.

Introduction

Regular, repeated, multi-view experience with individual objects characterizes infants' visual experiences during object play (Herzberg et al. 2022). In my dissertation, I characterize infant's visual experiences as **instance-limited, viewpoint-rich**, while internet-style datasets are **instance-rich, viewpoint-limited**. Fig 1 shows the distinctions in characteristics of the two. The research questions addressed in my dissertation are described below (overview in Fig 1):

- **RQ 1: Learning From Exposure to Individual Objects** When interacting with individual objects, infants receive rich visual information about that object; for instance, they see each object from multiple viewpoints and under different conditions of occlusion. I hypothesize that this, together with object permanence, provides access to signals for visual learning. *To what extent can regularities in continuous visual experiences during individual object interactions be exploited to learn visual representations?*
- **RQ 2: Joint Learning Under Distribution Shift** A right-skewed distribution with a long tail characterizes the visual experience of embodied agents; some objects are seen with very high frequency, while a long tail of objects are seen much more rarely. In machine learning terms, this pattern of exposure is an example of learning from multiple distributions, i.e. when the training data are sampled from two or more distributions. *How do different learning signals affect outcomes such as classification accuracy when learning jointly from these two distributions?*
- **RQ 3: Learning From Instance-Viewpoint Combinations** Learning about object categories requires finding the similarities across different objects in the category and dissimilarities between objects from different categories. Analogies provide a way of systematically capturing such similarities and dissimilarities between objects. *To what extent can representations learned via analogical inference over instance-viewpoint combinations support image classification?*

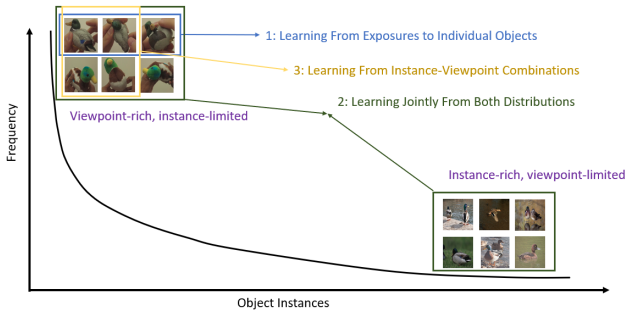

Figure 1: Long-tailed distribution showing frequency of seeing different instances of a category. The top-left part of the distribution signifies infants' visual experiences during object play. The bottom-right part represents the characteristics of internet-style datasets. The RQs, as well as the source of regularities each RQ uses are shown.

The Toybox Dataset

To mimic patterns of visual experience infants get when playing with objects, the Toybox dataset (Wang et al. 2018) was curated in our lab. The dataset contains egocentric videos of several toy objects from a small number of categories being manipulated in different ways. Due to the presence of a diverse and large set of viewpoints from a relatively small number of objects from a developmentally relevant object set (toys), I use the Toybox dataset in my experiments.

Current Work I have led the effort to add videos depicting additional object rotations to the dataset. I am currently working on collecting detailed annotations for bounding box and viewpoints for objects in the Toybox dataset.

Anticipated Progress The annotations will be collected in the next few months and will be released in the form of Toybox-V2 in addition to the extra transformation videos.

RQ1: Learning From Individual Objects

Related Work In computer vision, self-supervised learning (SSL) models leverage innovative learning signals to learn visual representations in the absence of image labels; alternative learning signals enable researchers to encode expert knowledge into the model (Chen et al. 2020). However, there has been limited research into studying the effectiveness of SSL signals which exploit naturalistic object interactions.

Progress In work recently published in CogSci 2023 proceedings (Sanyal et al. 2023), I studied the extent to which learning signals that equate different viewpoints of an object supports visual learning. Using the Toybox dataset and modifying the SimCLR framework (Chen et al. 2020), I conducted experiments to investigate the effectiveness of assigning similar representations to different images for a single object for learning high-quality representations. The representations learned using the modified embodied learning signal significantly outperforms baseline methods for downstream image classification on unseen objects from the Toybox dataset. Further, these benefits extend to a wide range of other downstream classification tasks.

Future Work In subsequent work, I found that in learning to equate different views of an object, the model loses information about the geometry of individual objects. Developing methods to retain this information in the model without losing category information would be valuable. I intend to use a multi-task approach, leveraging an additional task that relies on encoding of geometry information in the model.

RQ2: Joint Learning Under Distribution Shift

Related Work In computer vision research, robust generalization across distributions is thought to be mediated by alignment of the representations from the different distributions in feature space. A slew of work in multi-domain learning utilize learning signals which explicitly align features from the different domains in an attempt to learn domain-invariant representations (Pan et al. 2018).

Progress In ongoing work, I proposed metrics for quantitative evaluation of alignment in the feature space. Further, I hypothesize that joint learning from both distributions can use at least two kinds of learning signals: perceptual signals, which indicate the visual similarity of two objects, and categorical signals, which assign a category label to each individual image. Using the metrics, my experiments show a key distinction in the role of categorical and perceptual signals: classification performance on unseen images is driven by the categorical signals, while alignment of features from the two distributions is driven by the perceptual signal.

Anticipated Progress Further experiments are required to understand how these two learning signals interact to produce learning outcomes such as alignment and generalization. Careful variation of these learning signals enables disentangling their effects in strengthening and weakening different learning outcomes.

Future Work I intend to study how convolutional units in deep networks are shared between images from the two distributions. Techniques such as network pruning are useful for identifying convolutional units important for the different distributions. Futher, I intend to study how patterns of unit sharing between distributions are influenced by variations of perceptual and categorical signals.

RQ3: Learning Through Analogical Reasoning on Instance-Viewpoint Combinations

Related Work A recent body of research has trained models to solve visual abstract reasoning tasks directly on pixel inputs. While most existing methods only consider individual images while computing image features, a recent model (Yang et al. 2023) incorporates combined processing of all images within a task item, thus allowing interplay between conceptual and perceptual processing.

Anticipated Progress I intend to define a set of analogical reasoning tasks on the Toybox dataset that seek to test for understanding of both image category as well as image category-viewpoint combinations. Further different test sets will be designed that measure generalization ability at different levels of difficulty.

Future Work Rigorous evaluation of existing methods will be done to evaluate their effectiveness on the new task. Further, I intend to test how suitable representations from the trained network are for image classification by freezing the trained network and using the extracted feature representations to train a linear classifier on image classification.

References

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Herzberg, O.; Fletcher, K. K.; Schatz, J. L.; Adolph, K. E.; and Tamis-LeMonda, C. S. 2022. Infant exuberant object play at home: Immense amounts of time-distributed, variable practice. *Child development*, 93(1): 150–164.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the ECCV*, 464–479.

Sanyal, D.; Michelson, J.; Yang, Y.; Ainooson, J.; and Kunda, M. 2023. A Computational Account Of Self-Supervised Visual Learning From Egocentric Object Play. *CogSci 45*.

Wang, X.; Ma, T.; Ainooson, J.; Cha, S.; Wang, X.; Molla, A.; and Kunda, M. 2018. The Toybox Dataset of Egocentric Visual Object Transformations. arXiv:1806.06034.

Yang, Y.; Sanyal, D.; Ainooson, J.; Michelson, J.; Farhana, E.; and Kunda, M. 2023. A Cognitively-Inspired Neural Architecture for Visual Abstract Reasoning Using Contrastive Perceptual and Conceptual Processing. *arXiv preprint arXiv:2309.10532*.