公開

研 究 終 了 報 告 書

「Learning categories grounded in sensation without supervision」

研究期間： 2021 年 10 月～2023 年 12 月
研 究 者： Mingbo Cai

1. 研究のねらい

Current deep learning heavily relies on labeled data for image classification and object segmentation, limiting its application when such labeling is costly. As it learns a different way of representing objects from how the brain represents them, its trustworthiness is also compromised. This project pursues a novel framework for learning object-centric representations and basic categories from sensory data with essentially no supervision. It takes inspiration from infant learning and enforces the networks to learn to perceive objects with similar constraints faced by infants, namely, with no direct supervision for object segmentation, spatial localization or depth perception. The central principle of the proposal is that the ability to perceive objects can be a consequence of learning to predict the future sensory input. I envisioned that by predicting the trajectories of moving objects in videos, representations of basic categories that are distinct in the ways they interact with the environment would emerge as well. If successful, the project will provide a new direction for learning the building blocks of symbolic representation essentially only from sensory data and information of self-motion, which would have application in robotics.

2. 研究成果

(1)概要

The project has gained important insights for learning representation of objects similar to that of the brain, towards developing more trustable AI.

I have successfully developed a set of unsupervised learning deep neural networks that simultaneously learn to segment objects from visual scenes, perceive the depth of the scene, and localize each object in 3D space. The network only learns from scenes including moving objects observed by a moving camera, with access to the information of the self-motion of the camera. No parts of the network receive any supervision or pre-training on other labeled datasets.

I further attempted to relax the requirement of the built-in knowledge of rigid-body motion. Although the performance of 3D localization and depth perception degraded after the relaxation, surprisingly, object segmentation remains almost intact.

I also found that unsupervised learning to form representation of basic categories of geometric shapes purely by predicting object trajectories appears to be challenging. The reasons for this awaits further examination.

(2)詳細

(1) **Establishment of novel unsupervised object-centric representation learning model**

公開

inspired by the brain

I have developed a custom database for evaluating the framework I proposed. The dataset contains streams of videos taken by a camera making translational and panning motion while objects in the scenes move horizontally with constant speeds or rotate with constant angular velocities. The creation of this database predated a recent new dataset named MOVi. This contribution will allow researchers in this field to focus more on tackling environments with complex texture on object surfaces.

The model I proposed is mainly composed of three networks. An object extraction network sequentially extracts 3D locations and identity codes for all objects and generates segmentation maps for them from each frame of the video. A depth perception network that infers depth of the scene. The combination of the two networks allows predicting most of the pixels for the next frame by warping pixels visible in the current frame (essentially predicting optical flow). Because not all pixels in a new frame could have been seen in its preceding frame, additionally, an imagination network is included that implicitly predicts the newly appearing part of the next frame and its depth, to augment the warping-based prediction. By minimizing the prediction error for the new input image, all the three networks are jointly trained without direct supervision.

Comparing against several recently proposed unsupervised object-centric representation networks, my model achieves the best segmentation performance. I found that several existing works that do not explicitly model 3D structures of the objects often utilize inductive biases (such as clustering colors) that do not work for scenes with complex texture as the ones I tested on.

Furthermore, I have confirmed that the model can infer depth with high correlation with the ground truth depth, and infer object 3D locations with high accuracy as well, for the objects that the model manages to segment.

(2) **Evaluating the minimal set of assumptions needed to learn 3D object representation grounded in sensation**

Researchers in developmental psychology found important insights on the development of object perception in infants: 3D perception matures before and may be used by object segmentation, while the general ability of object segmentation is acquired without knowledge of object categories; infants appear to honor the principles of *cohesion* (two surface points on the same object should be linked by a connected path of surface points and move continuously in time), *boundedness* (two objects cannot occupy the same place at the same time), *rigidity* (objects move rigidly) and *no action at a distance* (independent motions of separated objects make infant interpret them as two objects instead of one) when perceiving objects, reflecting basic constraints of physical objects.

It is difficult to further answer with infant studies whether some of these principles are hard-wired in the brain by evolution or learned through experience due to the limitation of experimental approach. Therefore, I further sought to shed insight on the minimal set of

assumptions that need to be incorporated in a learning system to achieve 3D object perception from 2D visual input, utilizing the networks I have developed. I focused on the rigidity assumption.

I found that, by replacing hard-coded knowledge of rigid-body motion with a learnable neural network, the model does not degrade in its performance of object segmentation. However, the performance on depth perception is reduced. This suggests that additional teaching signals or learning principles may be needed to achieve what infant brains achieve.

(3) **Examination of learning to form basic categories by predicting future trajectories of objects.**

In my original hypothesis, basic categories of geometric shapes may be learned without supervision because each type of objects follows different dynamics when interacting with the environment. For example, if a ball is thrown on the floor, the bouncing pattern is typically more predictable and it turns to move along the same horizontal direction, while a cube or pyramid will follow much less predictable bouncing trajectories and may move back and forth depending on the parts that touch the ground. In order to make efficient prediction of object trajectories, the brain might group the representations of objects that follow similar dynamics in its representational space such that it can generalize the learned ability of predicting trajectories to similar objects, naturally clustering objects into categories of shapes while ignoring irrelevant features such as colors and textures. In fact, humans also have a bias of using shapes to name objects, while deep neural networks have a bias to rely on texture.

I tested this hypothesis on a simple dataset. However, the networks I designed did not form a representation that groups objects with the same shape into clusters, as I have initially hypothesized. The reason for this deserves further investigation.

3. 今後の展開

(1) **Learning to disentangle properties of objects from lighting effects.**

The brain has an impressive ability to disentangle the intrinsic appearance property of objects (e.g., the vase is white) from the contextual conditions that produce variations in the appearance of objects (lighting angle, reflection, shadows etc.). An example illustrating this ability is the illusion shown on the right in which humans perceive A as darker than B even though the pixels in A have the same color as the pixels in B. This is achieved without supervision and also largely ignored in the current deep learning (even though unsupervised learning of object categories can become invariant to such variations, it is not guaranteed that it understands that all parts of a white vase are fundamentally of the same white color regardless of the shades). How to achieve this in deep networks in an unsupervised way is an

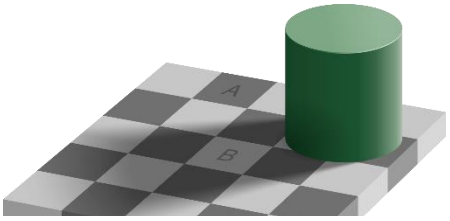

Figure. A visual illusion illustrates the brain's ability of disentangling the intrinsic colors of objects in the face of variation in lighting.

interesting topic for future research.

(2) **Investigating the cause of difficulty of unsupervised learning of object categories by predicting object trajectories.**

One of the original goals of spontaneously learning categories grounded in sensory data was not successful in the experiment. I will investigate the reason for this and propose new methods of unsupervised learning of categories with inspiration from development psychology.

4. 自己評価

I consider the proposed framework of learning 3D object-centric representations as novel and promising for developing symbolic representations from deep networks. Such evaluation is also received during peer review processes. I am very excited that I have the opportunity to explore this research direction, thanks to the funding of ACT-X.

The initial goal of achieving unsupervised object classification based on prediction appears to be overly ambitious. However, I accumulated many first-hand experiences in experimenting with variations of network architecture and training neural networks with partial gradients, which will significantly facilitate future research.

The early stage of the work on object-centric learning was accepted as an oral presentation at the Shared Visual Representations in Human & Machine Intelligence (SVRHM) workshop of NeurIPS in 2021.

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：1件

1. Tushar Arora, Li Erran Li, Ming Bo Cai, Learning to perceive objects by prediction, SVRHM2021 workshop at the conference of Neural Information Processing Systems, 2021

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものは件数にのみ含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. Invited talk: Learning internal models of the world: brain and machine. Workshop on Mechanism of Brain and Mind, Japan (online), 2022