Toddler-inspired learning induces hierarchical object representations

Arthur Aubret, Céline Teulière, Jochen Triesch

Toddler-inspired learning induces hierarchical object representations

1st Arthur Aubret
CNRS, Pascal institute
University Clermont-Ferrand
Clermont-Ferrand, France
arthur.aubret@uca.fr

2nd Céline Teulière
CNRS, Pascal institute
University Clermont-Ferrand
Clermont-Ferrand, France
celine.teuliere@uca.fr

3rd Jochen Triesch
Frankfurt Institute for Advanced Studies
Frankfurt am Main, Germany
triesch@fias.uni-frankfurt.de

***Abstract*—Humans learn to both visually recognize individual objects and categorize them at different levels of abstraction. Such multi-semantic representation is crucial to efficiently reason about the world. However, it is currently unclear how such representations could be learned with the very sparse labeling available to human learners. To answer this question we let an artificial agent play with objects while occasionally "hearing" their category label. Our agent assigns similar representations to a) similarly labelled and b) close-in-time visual inputs. We show that our agent learns a 2-level hierarchical representation that first aggregates different views of objects and then brings together different objects to form categories. Interestingly, we do not observe a trade-off between each semantic content. Our work suggests that the temporal structure of visual experience during object play together with occasional labeling suffice for learning a hierarchically structured object/category representation.**

***Index Terms*—hierarchical representation, representation learning, embodiment**

I. Introduction

Children quickly start to learn to both recognize a specific object independently of its viewpoint/distance (object recognition) and build categories that support generalization to novel exemplars (category recognition). On the one hand, several works suggest that object recognition is learnt by semantically associating views that are close in time in an unsupervised way [Wood and Wood, 2018]. On the other hand, the functional use that defines objects' human-assigned categories may come from the intertwining between their similarity based on their global shape, associated word (label) or the later acquisition of conceptual knowledge [Landau et al., 1998]. In practice, infants better categorize with labels even though they only need few labels to generalize objects to categories [LaTourrette and Waxman, 2019].

Machine learning methods have been used to model how toddlers learn each of these semantic representations [Zhai et al., 2019], [Schneider et al., 2021]. Yet, it remains unanswered how these two semantic representations can coexist. Here, we investigate the relation of two semantic contents inside one learnt representation. We take inspiration from how toddlers interact with their world and let an agent interact with 3D toy objects, with an unseen social partner who sometimes gives the category label of the object being manipulated ("banana"). We learn the representations with a contrastive learning (CL) algorithm fed with image/label (cross-modal consistency) and image/next-image (temporal consistency) pairs. In CL, inputs that are often paired develop similar representations. We assume the brain uses similar principles to learn an internal representation of the world. We experimentally show that our bio-inspired agent builds a 2-level hierarchical representation: the first level aggregates different views from a single object; the second levels groups objects into categories.

II. Method

a) Environment: In order to marginalize the impact of colors and backgrounds on category recognition [Aubret et al., 2022], we use the simplest environment introduced in [Aubret et al., 2022] (cf. Figure 1c). We place an agent in an empty environment where it can interact with more than 2,000 untextured 3D toys distributed among 105 common categories [Stojanov et al., 2021]. The agent acts on two timescales: 1) At each timestep, the agent rotates the object in front of it (between 0 and 360 degrees on the yaw axis) and potentially receives the category label of this object according to a probability p_{lab}; 2) every 10 timesteps, it replaces its current object by a new randomly sampled one.

b) Contrastive learning: In order to learn from the different inputs, we combine two different loss functions. The first one guarantees the cross-modal label-vision consistency of the representation; the second one ensures the temporal consistency of the representation. For both, we train a neural network f_θ to minimize the SimCLR loss [Chen et al., 2020], which is given, for each sample x_i in the minibatch, by:

$$\mathcal{L}(x,z) = \frac{1}{N}\sum_{x_i \in x} -\log \frac{e^{-||f_\theta(x_i)-z_i||_2/\tau}}{\sum_{k=1}^{N} e^{-||f_\theta(x_i)-z_k||_2/\tau}}, \quad (1)$$

where x refers to the visual inputs, N is the size of the mini-batch and τ is the temperature hyper-parameter. The difference between the two loss functions comes from the computation of z. For temporal consistency, $z_i = f_\theta(prev(x_i))$ is the representation of the previous image. For cross-modal consistency, $z_i = g_\omega(l_i)$ is the representation of the one-hot label l_i. Therefore, the whole loss is $\mathcal{L}(x, f_\theta(prev(x))) + \mathcal{L}(x, g_\omega(l))$.

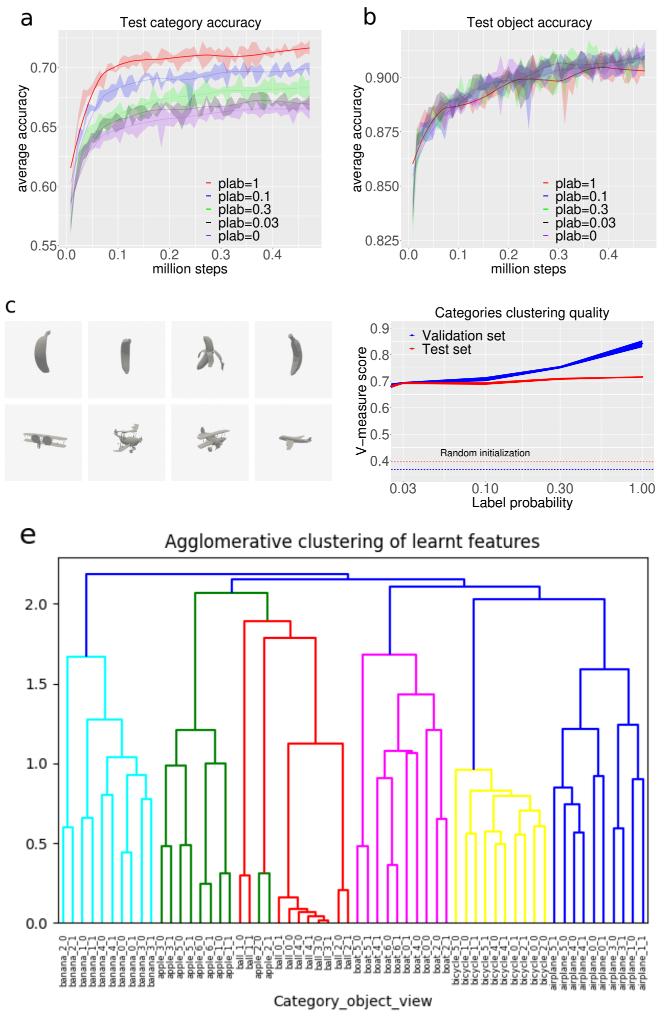

Fig. 1. Curves show the mean +/- standard deviation over 3 random seeds. a) and b) respectively show the test category and test object as a function of label probability. (c) Examples of different bananas and airplanes used in (e). Neither object recognition (bananas) nor category recognition (airplanes) are always visually obvious. (d) V-measure score as a function of label probability, computed using true category labels and the last clusters of a 105-clusters agglomerative clustering applied on the features output by f. (e) Dendogram showing the hierarchy of clusters generated by a smaller agglomerative clustering. Same color indicates the same category.

c) Training and evaluation: we use the same neural network f as [Aubret et al., 2022]. Our label encoder g_ω is a fully connected neural network with 2 hidden layers of 256 units. During training, the agent stores its interactions and labels in a replay buffer and learns on a randomly sampled minibatch between each interaction with the environment. To evaluate the representation, we freeze the weights of our trainable networks and train a linear classifier on top of the representation using the true labels. The linear classifier is trained and evaluated on the same objects as in training (validation set) or different objects (test set).

III. Experiments

Our experiments aim to assess whether we manage to learn a representation endowed with a structure that reflects several semantic contents, *i.e.*, objects' identity and their category.

We observe in Figure 1(a) that the more labels are given, the higher the test category accuracy. More importantly, in 1(b) we see that using labels does not hurt the quality of individual object recognition, suggesting that both the object identity and its category can be reliably retrieved from the representation. The V-measure score displayed in Figure 1(d) (closer to one is better) validates that this improvement of accuracy is associated with a representation whose structure better reflects the labeling.

To give an intuitive illustration of these quantitative results, we randomly selected 60 images (2 views $\times$ 5 objects $\times$ 6 categories) in the validation set and applied an agglomerative clustering on them. The features were learnt **without** labels. In Figure 1(e), we observe that the algorithm quickly clusters different views of the same object (distance close to 0). Even without labels, different objects are clustered slightly later according to their category (same color, distance close to ~ 1.3). We conclude that the learnt representation has captured the hierarchical object/category structure.

IV. Conclusion

We investigated the representations learnt by an agent that interacts with objects in a toddler-inspired way. The agent exploits temporal consistency and sparse labeling through a SimCLR loss function. We showed that this results in a hierarchical representation retaining object identity and class information. Our results offer an explanation how infants may effectively learn hierarchical object/category representations despite receiving only sparse label information.

References

[Aubret et al., 2022] Aubret, A., Ernst, M., Teulière, C., and Triesch, J. (2022). Time to augment contrastive learning. *http://arxiv.org/abs/2207.13492*.

[Chen et al., 2020] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[Landau et al., 1998] Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1):19–24.

[LaTourrette and Waxman, 2019] LaTourrette, A. and Waxman, S. R. (2019). A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental science*, 22(1):e12736.

[Schneider et al., 2021] Schneider, F., Xu, X., Ernst, M. R., Yu, Z., and Triesch, J. (2021). Contrastive learning through time. In *SVRHM 2021 Workshop @ NeurIPS*.

[Stojanov et al., 2021] Stojanov, S., Thai, A., and Rehg, J. M. (2021). Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1798–1808.

[Wood and Wood, 2018] Wood, J. N. and Wood, S. M. (2018). The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, 42(4):1391–1406.

[Zhai et al., 2019] Zhai, X., Oliver, A., Kolesnikov, A., and Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485.