Simulating Latent Space Representation of Novice and Expertise using Supervised VAE (SVAE)

Seungmin Baek
Department of Computer Science
Vanderbilt University
seung.min.baek@vanderbilt.edu

Jinhyeok Jeong
Department of Psychology
Vanderbilt University
jinhyeok.jeong@vanderbilt.edu

Abstract

One long-standing question in vision science and cognitive psychology is how people learn and recognize categories of objects. Objects can be categorized in various levels of abstraction. As individuals gain expertise within a certain category domain, they become adept at differentiating exemplars at the subordinate levels. Whether such perceptual expertise is a result of changes in representations and/or other cognitive processes like decision making remains inconclusive. In this study, we simulate how visual experience in a certain category modulates the object category representations across basic- and subordinate-levels by using a Supervised Variational Autoencoder (SVAE), which isa variance of VAE with a classification component. To simulate a visual experience of a novice, we initially trained the model to learn the basic categories of objects with the CIFAR-10 dataset. After that, we simulated the visual experience of a bird expert by subsequently training the model to learn subordinate-level categories of birds. To examine the influence of supervision signals on the development of expertise, we tested several different methods to test the development of subordinate-level category representations of a bird with a subset of CUB-200-2011 dataset. We succeeded in training both novice and expert models to reconstruct the images, but the latent representations that models learned did not show clear evidence of categorical representation. Possible reasons why it was difficult to learn the meaningful latent representations while preserving decent reconstruction performance and how the current framework could be improved are discussed.

1. Introduction

Visual categorization, which refers to the process of classifying objects into a particular category based on their visual properties, is a crucial skill for survival as it allows us to generalize our knowledge into the new instances of objects (Barsalou, 1985; Mervis & Rosch, 1981; Rosch & Mervis, 1975; Rosch et al., 1976). How people learn and recognize categories of objects has been one of the central questions in vision science and cognitive psychology (Logothetis & Sheinberg, 1996; Mervis & Rosch, 1981; Palmeri & Gauthier, 2004; Serre, 2016). One interesting property of categorization is that an object can be classified in various levels of abstraction. For example, a dog can be categorized as a “dog” (basic-level category), but also be categorized as an “animal” or

as a specific breed (superordinate-level category), like “labrador retriever” (subordinate-level category).

Humans’ ability to recognize object categories could differ across individuals depending on their visual experience and knowledge about objects. It is usually easy to recognize a basic-level category of an object (e.g., dog), but recognizing its subordinate-level category (e.g., Boston terrier) tends to be difficult and requires some extent of expertise in that category. To fully understand how perceptual expertise arises from visual experience of objects, both empirical work and computational modeling work are necessary to address several possible mechanisms underlying the development of expertise (Palmeri et al., 2004). Although there have been computational modeling work to investigate mechanisms underlying perceptual expertise, whether such a development of expertise is related to the changes in representations and/or to changes in other cognitive processes like decision making remain elusive.

In the current work, we aimed to explore one of the possible mechanisms of the development of perceptual expertise with deep neural network. Although cognitive psychologists have been developed articulated computational models of how categorization and object recognition work (Estes, 1993; Kruschke, 2008; Pothos et al., 2011), it has been difficult to study how representations change based on visual experience with realistic objects because it is difficult to figure out how complex objects with multiple attributes are encoded in human brains. Following the recent success utilizing a deep neural network as a front-end of cognitive models to approximate the object representations that humans may have (Annis et al., 2021; Battleday et al., 2020), we simulate how visual experience changes objects and category representations across different levels of abstraction with a neural network. Specifically, we train a Supervised Variational Autoencoder (SVAE), which is a variant of VAE that can learn the latent representations of natural images to reconstruct images, for the CIFAR-10 dataset to simulate the visual experience of a novice. CIFAR-10 datasets consist of 60,000 images in 10 different basic categories (airplane, automobile, bird, cat, deer, dog ,frog, horse, and truck), with 6,000 images per each category. We subsequently simulate visual experience of a perceptual expert to explore possible changes associated with extensive visual experience in a certain category domain. Due to the architecture of SVAE, which has both supervised and unsupervised components, latent representations that SVAE learned could possibly used for developing cognitive models of how people learn category and objects, as well as be used for testing what aspects of visual experience (e.g., whether supervised learning is necessary for humans to learn categories) are necessary to develop perceptual expertise.

2. Related work

2.1. Combining deep neural network and cognitive models for studying human cognitions

Many cognitive models in psychology typically assume that objects/stimuli are represented in multidimensional psychological space, and then elaborate the subsequent operations that occur on top of those psychological representations (e.g Ashby, 2014; Ashby & Maddox, 2005; Love et al., 2004; Nosofsky, 1986). To avoid the problem of complexity of inferring the nature of object representations, it has been common to use artificial stimuli with relatively simple attributes or to approximate the representations by using multidimensional-scaling method. Such an approach has been useful to advance the mechanistic models of how cognitive operation occurs on those representations, whether and how such an approach could be extended to complex and natural objects wasn't clear until recently. Nowadays, psychologists and cognitive scientists have started to use deep neural networks as a front-end of cognitive models that approximate the representations that are used for further cognitive operations (e.g., Annis et al., 2021; Battleday et al., 2020).

In recent efforts combining the representations of deep neural networks with cognitive models, most works have been using supervised deep convolutional neural network that were pre-trained for larger natural imagesets (e.g., ImageNet) for classification. Although such an approach is reasonable given the similarity between human brains' ventral stream and the architecture of deep convolutional neural network (Cichy et al., 2016; Güçlü & van Gerven, 2015; Kriegeskorte, 2015), there has been a debate which one between generative models and discriminative models is a better model for human cognition/perception. Despite such a recent trend, it has not been tested whether a generative model learned natural objects could be used as a front-end of cognitive models.

2.2. Generative neural networks

There are various contemporary state-of-the-art generative models by leveraging techniques from Generative Adversarial Networks (GAN), and diffusion models. Examples of such models include Deep Convolutional GAN (DC-GAN) (Radford, Metz & Chintala, 2016), Wasserstein-GAN (w-GAN) (Arjovsky, Chintala & Bottou, 2017), Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain & Abbeel, 2020), and Latent diffusion model (LDM) (Rombach, 2022), among others. Despite the prevalence of these models, our selection has gravitated towards the Variational Autoencoder, specifically Supervised VAE (SVAE). The rationale for this choice lies in our objective which is to discern cognitive process differentials between novices and experts through a comprehensive analysis of latent space visualization. The decision to employ SVAE is predicated on the confidence that VAE will provide insightful representation within the latent space.

3. Methodology

3.1. Supervised VAE (SVAE)

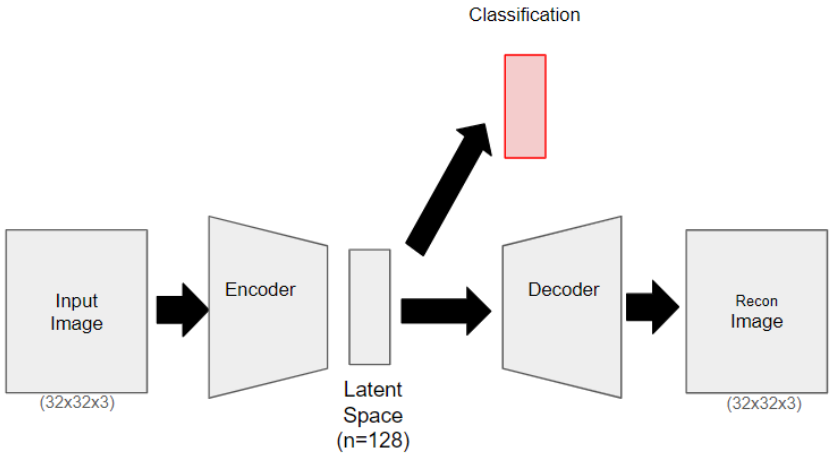

Figure 1. Architecture of SVAE.

Variational Autoencoder (VAE) is a representational learning algorithm that generates input by learning model parameters to maximize the likelihood of training data, denoted as $p_\theta(x)$ where $p_\theta(x) = \int p_\theta(z)\, p_\theta(z) dz$. VAE comprises an encoder network, responsible for mapping input data to a probability distribution in latent space, and a decoder network, generating data through the reconstruction of samples in the latent space. The primary objective of VAE training is the simultaneous ELBO reconstruction of input data and regularization of the latent space with KL Divergence to follow a Gaussian distribution.

Given the intractability of $\int p_\theta(z) dz$, Evidence Lower Bound, known as ELBO, is vital. Leveraging the ELBO technique and the reparameterization trick which enable model to be differentiable, facilitates the maximization of the likelihood of training data, leading to achieve optimal parameters θ^*, ϕ^* through $argmin_{\theta,\phi} \sum_{i=1}^{N} - L(x^{(i)}, \theta, \phi) = \sum_{i=1}^{N} - log\ log\ (p_\theta(g_\theta(z)) + KL(q_\phi(z|x^{(i)})|p(z))$, where $p(z) = N(0, I)$ and L: Mean Squared Error.

As shown in Figure 1, Supervised VAE constitutes a simple modification of Vanilla-VAE, incorporating principles from supervised learning (Nguyen & Martínez, 2020). In the supervised learning, the model is trained on a labeled dataset, where each input is associated with a corresponding target output which has an objective to map from inputs to outputs for making predictions on unseen data. SVAE combines elements of both VAEs and supervised learning methodologies. Specifically, encoder in SVAE adapted to map input data not only to a distribution in latent space but also to a space for classification. The training process encompasses reconstruction loss, KL Divergence, and in addition to conventional VAE, a classification loss, thereby enriching VAE's capacity to reconstruct and perform classification.

3.2. Metric for the evaluation of latent representations

Within the analytical framework for evaluation of latent representation, we have employed the Calinski-Harabasz (CH) index (Calinski & Harabasz, 1974) as a metric for

measuring distribution within the latent space. The CH index represents similarity by evaluating a ratio between inter cluster dispersion and intra cluster dispersion across all clusters. Leveraging the CH index within analysis of latent representation enables transparent quantification of the optimal distribution of embedded data in latent space.

4. Simulating basic category representations of novice

To simulate the basic category representations of a novice, we trained the SVAE to learn the latent representations of CIFAR-10 dataset. The CIFAR-10 dataset consists of 32x32 pixels of images that belong to 10 basic-level categories (airplane, automobile, bird, cat, deer, dog ,frog, horse, and truck). As images are blurry and exemplars within each category are diverse enough, training the models to learn the latent representations of CIFAR-10 would be challenging enough. After training the novice models to learn the CIFAR-10 dataset, we analyzed their latent space to test whether exemplars within each basic level category are clustered well and also analyzed the performance of image reconstruction. In the case of humans, it is often assumed that variability of within-basic level category exemplars is much lower than that of between-basic level category exemplars. In addition, boundaries of categories for natural objects tend to be fuzzy, so that some exemplars might be difficult to decide its category membership (Rosch et al., 1976). Therefore, it would be ideal if we could observe that exemplars are clustered based on their categories but with some overlapping after training the novice models.

```
SupervisedVAE_basic(
  (vae): VAE(
    (encoder): Sequential(
      (0): Conv2d(3, 32, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1))
      (1): LeakyReLU(negative_slope=0.01)
      (2): Conv2d(32, 64, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1))
      (3): LeakyReLU(negative_slope=0.01)
      (4): Conv2d(64, 128, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1))
      (5): LeakyReLU(negative_slope=0.01)
      (6): Flatten(start_dim=1, end_dim=-1)
      (7): Linear(in_features=2048, out_features=256, bias=True)
      (8): LeakyReLU(negative_slope=0.01)
    )
    (fc_mu): Linear(in_features=256, out_features=128, bias=True)
    (fc_logvar): Linear(in_features=256, out_features=128, bias=True)
    (decoder): Sequential(
      (0): Linear(in_features=128, out_features=256, bias=True)
      (1): LeakyReLU(negative_slope=0.01)
      (2): Linear(in_features=256, out_features=2048, bias=True)
      (3): LeakyReLU(negative_slope=0.01)
      (4): Unflatten(dim=1, unflattened_size=(128, 4, 4))
      (5): ConvTranspose2d(128, 64, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1))
      (6): LeakyReLU(negative_slope=0.01)
      (7): ConvTranspose2d(64, 32, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1))
      (8): LeakyReLU(negative_slope=0.01)
      (9): ConvTranspose2d(32, 3, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1))
    )
  )
  (classifier): Linear(in_features=128, out_features=10, bias=True)
)
```

Figure 2. SVAE architecture for building a novice model.

After testing several different versions of SVAE models with different number of latent dimensions and with different weights assigned to each component of loss function (reconstruction loss, KL loss, and classification loss), we've decided to use the architecture whose latent dimension

is 128 based on their relatively better reconstruction performance than the others. As shown in Figure 3 and Table 1, our initial results showed a typical pattern of posterior collapsing, which means the failure of image reconstruction (Figure 3a). Ironically, visualized latent representations with t-SNE suggested that object exemplars were well clustered based on the basic-level categories, consistent with a high CH index of 942 (higher values indicate well separated clusters). To improve the reconstruction performance, we increased the weights assigned to reconstruction loss (MSE loss) during training, and we found that this procedure indeed improved the reconstruction performance (Figure 3b). However, this improvement of reconstruction seemed to be at the expense of category clustering in the latent representations: as different from the first version, objects were not clustered well based on their categories. We've tried to find a balanced point where both category-based clustering and reasonable reconstruction performance can be achieved simultaneously, but it was difficult in the current setting to find a situation where category representations are well clustered while reconstruction performance is preserved. Although we slightly improved the classification index (CH index) from 43 to 85 by giving a high weight to classification loss, the visualization result suggests that clustering was not enough to be considered as a reasonable simulation of human novice representations.

In general, our attempts to train the SVAE model to achieve both reconstruction performance and category-based clustering of representations were not successful. We will later discuss what were the possible reasons for failure and how the model could be improved to achieve the desired goal.

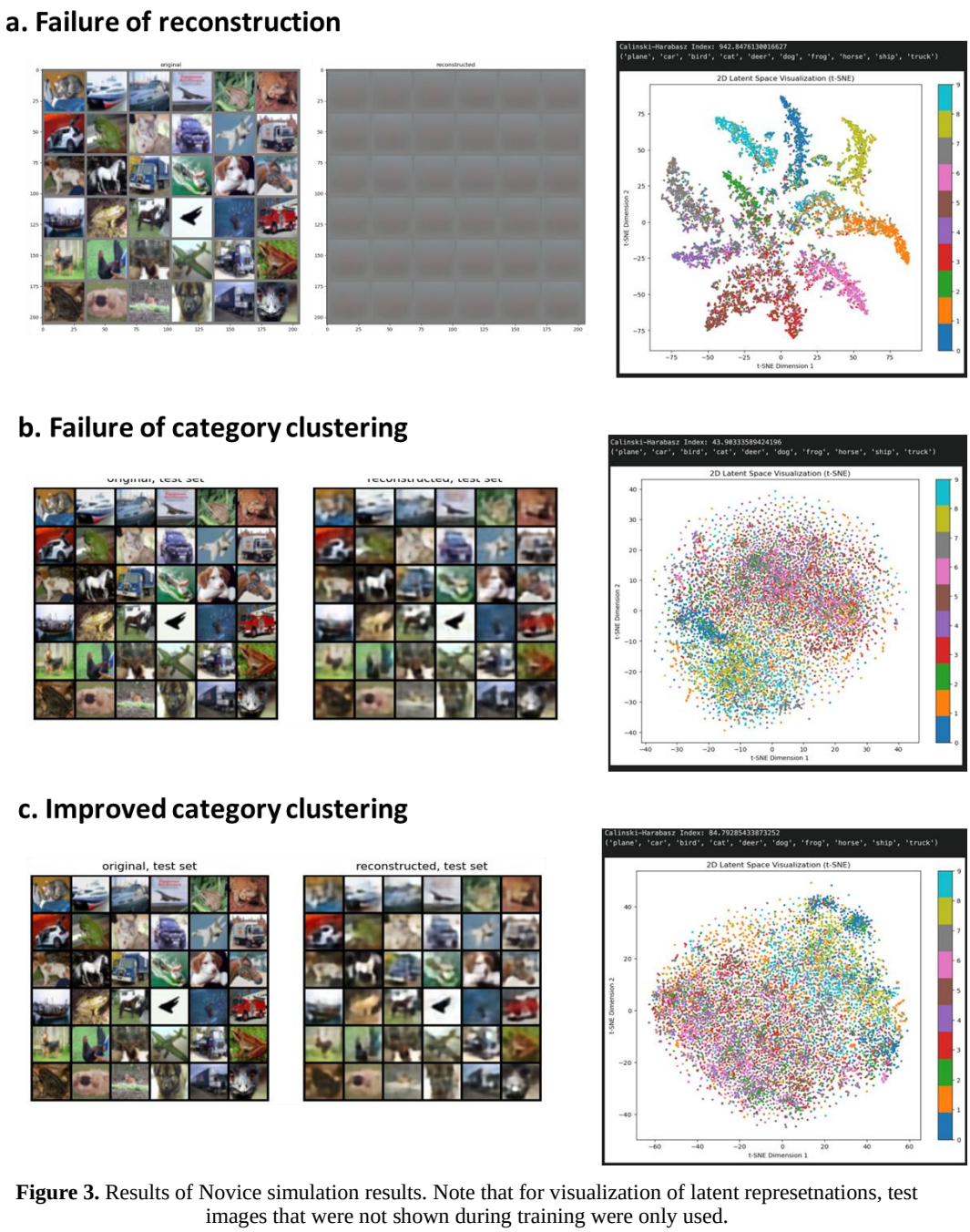

Figure 3. Results of Novice simulation results. Note that for visualization of latent represetnations, test images that were not shown during training were only used.

Table 1. Clustering performance of Novice simulation results.

Novice model type	Calinski-Harabasz (CH) index
a	942
b	43
c	85

5. Simulating subordinate category representations of bird expertise

Although training the SVAE models to simulate the visual representations of human novices was not successful, we decided to keep forward to test whether further training for subordinate-level category could change the results. Given the novice model trained in the previous experiment (novice model type c in this case), we did further training for learning the subordinate-level categories of birds. We've decided to use the CUB-200-2011 dataset, which is a challenging dataset consisting of 200 species of birds. Since utilizing all the images and categories of CUB-200-2011 dataset could be time consuming, we've decided to take only 10 categories of them as a starting point (black footed albatross, laysan albatross, sooty albatross, groove billed ani, crested aucklet, least aucklet, parakeet auklet, rhinoceros auklet, brewer blackbird, red winged blackbird). Visual expertise can be defined as an ability to discriminate exemplars at the subordinate-levels (e.g., discriminating black footed albatross and sooty albatross although both are evidently birds). We reasoned that if the models could learn the latent representations that are separable based on those subordinate bird species, then those models could be considered as approximations of human experts for birds.

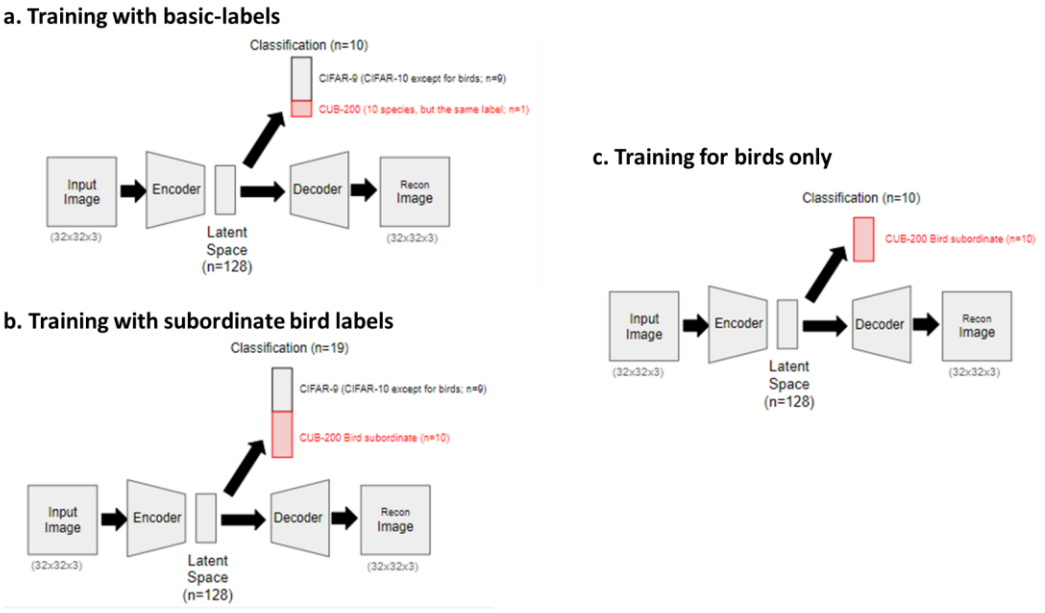

Figure 4. Different training methods for labeling category information.

To simulate bird expertise models, we needed to decide how to provide supervision signals regarding subordinate-level categories. There could be more ways to implement it, we considered three possibilities in the current project. First, we gave the same label ("bird") to all bird images regardless of their subordinate level categories as well as maintaining the labels for the CIFAR-10 imageset except for a bird label (Figure 4a). In this condition is for testing whether having extensive

amounts of visual experience is enough for building subordinate-level categories even if there is only basic-level category information. if the previous training for basic-level category with CIFAR-10 could be transferred, the basic-level category knowledge about birds might be helpful to facilitate this process. Second, we considered a case where subordinate-level category labels (i.e., bird species) are directly given during training as well as the basic-level categories for the other objects (Figure 4b). In this situation, each subordinate-level category (bird species) is treated as a separate category as in the case of other basic-level categories. This condition is relevant to a phenomenon called "entry-level shifting", which refers to the situation where humans start to feel subordinate-level categories in their expertise domain as basic-level category (Palmeri et al., 2004; Tanaka & Taylor, 1991). Lastly, we consider a situation where the novice model is further trained on subordinate bird categories only (Figure 4c). Although this condition may be vulnerable to catastrophic forgetting of basic-level categories that the model learned before, this model may achieve representations that could separate subordinate-level categories well in the expense of basic-level category knowledge.

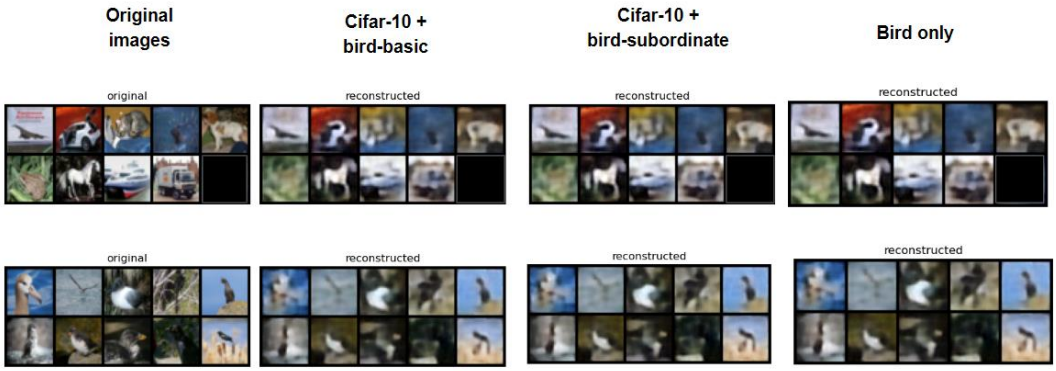

Figure 5. Reconstruction of images from expertise models. First row shows example images of CIFAR-10 (basic-level categories) and the second row contains bird images taken from CUB-200-2011.

At first, we tested whether these expertise models could reconstruct the images and whether there is a difference across categories. As shown in Figure 5, all variants of expertise models tested in this experiment were able to reconstruct images fairly well. For the model trained for bird images only (the rightmost column in Figure 5), it was able to reconstruct CIFAR-10 images even though they were not trained on that during expertise simulation. It suggests that at least catastrophic forgetting did not occur in this condition.

In the previous experiment of simulating basic-level category representations of novice, we failed to achieve both reasonable reconstruction performance and well clustered category representations. To test whether this was also the case for expertise models, we used the same method (2D visualization of multi-dimensional latent representations using t-SNE) to examine the qualitative patterns of representations. As shown in Figure 6, it was difficult to observe a clear

pattern of clusters based on categories, for both basic-level and subordinate-level categories. Consistent with visualization results, clustering index (CH index) indicates that latent representations of expertise models were not clustered well based on their learned category labels (see Table 2). In general, clustering index was lower than the baseline model (novice model without expertise training), indicating that it was not successful to simulate expertise category representations.

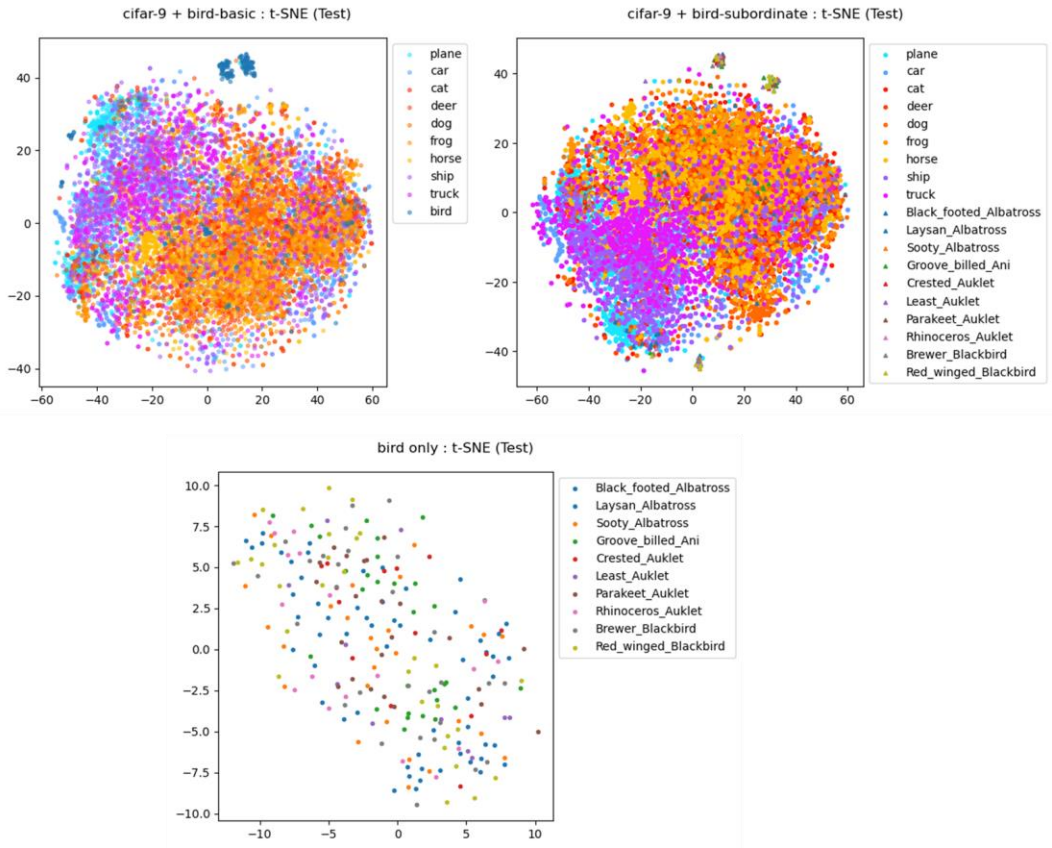

Figure 6. Latent representations of expertise models. Note that only test images that were not shown during training were used to construct latent representations.

Table 2. Clustering performance of Expertise simulation results.

model type	Calinski-Harabasz (CH) index
baseline (novice)	85
cifar-10 + basic-level	76.9615
cifar-10 + subordinate-level	39.2033
birds only	2.1795

Although visualization of latent representations and clustering index provide an insight of

whether latent representations contain useful information for categorization, they may not show the whole picture. To test if they have any information which is relevant for categorization, we examined the performance of classification based on the activations of the classification layer of expertise models. As shown in Figure 7, all variants of expertise models were able to predict categories of given objects more than chance levels. It suggests that although latent representations were not clearly separated, they still have some meaningful information that is useful for categorization. In addition, patterns of train and test accuracies suggest that this result is not the artifact caused by over-fitting. If this was the case, test accuracy should be much lower than train accuracy, but there was not a huge difference between them.

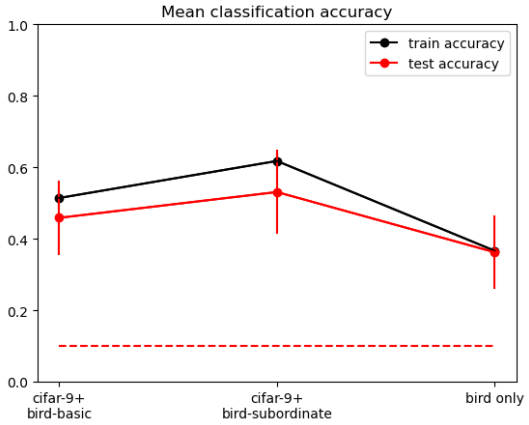

Figure 7. Top-1 classification accuracy.

6. General Discussion

In the current project, we aimed to simulate how visual experience modulates the category representations across different levels of abstraction. By using the SVAE, we tried to answer what kind of supervision signals are necessary to develop subordinate-level knowledge of categories. Given that it is impossible to directly observe how humans represent objects across different categories, it has been challenging to study how humans develop category representations and how perceptual expertise is developed from visual experience. Although the model that we exploited in the current project is too simple to be considered as an ideal model for the human visual system, it was useful to conduct simulations for a proof-of-concept to examine how latent representations change based on visual experiences. Although we failed to successfully train the models to have both reasonable reconstruction performance and well clustered category representations, the current approach may be potentially improved to achieve this goal and used for future study.

One issue of training the VAE is to make a proper balance between the KL loss and the reconstruction loss to achieve a good generalization performance as well as decent reconstruction performance. As SVAE model is a variance of VAE, it also has the same problem. Furthermore, as

SVAE has one more loss function (i.e., classification loss), one needs to find a balance between these three different loss functions to train the model successfully. There are many potential explanations why we failed to train the models successfully although we manipulated various factors including the number of dimensions and weight parameters for losses to achieve it. First, we assumed that SVAE needs to learn the latent space consisting of meaningful information in terms of categorization to achieve the reconstruction performance, but this may not be true. If reconstructing pixel intensities of images was possible without considering categorical information of objects in an image, then making a balance between these three different losses would be much more difficult. Considering that we used low-resolution images of 32 x 32 pixels, information that is crucial for categorization might not be present or difficult to detect.

Another possible reason why we failed to train the models successfully is that the encoder and decoder that we used were too shallow. We only had three convolutional blocks for encoder and decoder, and we did not test if more deeper architecture could achieve better results. Considering that most of the neural networks used for cognitive modeling study are much deeper and extensively trained for larger imagesets, we might need deeper architecture and more images to achieve our goals. It is known that the ventral stream in the human brain, which is responsible for object recognition and categorization, has at least more than 4 layers (e.g., V1 → V2 → V4 → IT cortex; Kubilius et al., 2018). Future study may consider using a deeper network to examine whether it is possible to achieve well clustered representations.

One important issue that was not addressed in the current project is that there could be additional processing on top of those representations in the case of humans. As briefly introduced earlier in this paper, there have been extensive work on cognitive models specifying what kind of operations occur on top of psychological representations of objects. Based on these cognitive processes, the same representations could result in different behavioral outcomes. For example, two people having the exact same representations could have entirely different performance in classification because of the differences in cognitive processes, such as decision making processes and similarity computation (Annis & Palmeri, 2019; Shen & Palmeri, 2016). In addition, it has been shown that categorization performance could be improved and become more similar to humans if we consider additional cognitive processes on top of deep learning representations (Battleday et al., 2020; Jha et al., 2023). Therefore, future studies may consider extending simulations by combining representations obtained from neural networks with cognitive models instantiating how people make a choice given the representations.

In sum, we implemented SVAE models for simulating how visual experience modulates category representations across different levels of abstraction. Although we were not successful to achieve an ideal goal of having reasonable reconstruction as well as interpretable latent

representations, there is a possibility that this approach could be improved to achieve the desired goal. Considering increasing popularity of generative models in cognitive science and its ability to learn the latent representations as well as the ability to generate new samples (Goetschalcks et al., 2021), it is still promising and valuable to keep exploring the possibility of using generative models like VAE to apply it to studying mechanisms of human cognition and perception.

References

[1] Annis, J., Gauthier, I., & Palmeri, T. J. (2021). Combining convolutional neural networks and cognitive models to predict novel object recognition in humans. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(5), 785.

[2] Annis, J., & Palmeri, T. J. (2019). Modeling memory dynamics in visual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(9), 1599.

[3] Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol., 56*, 149-178.

[4] Ashby, F. G. (Ed.). (2014). *Multidimensional models of perception and cognition*. Psychology Press.

[5] Arjovsky, M., Chintala, S., & Bottou, L. (2017, December 6). Wasserstein Gan. arXiv.org. https://arxiv.org/abs/1701.07875

[6] Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. Nature communications, 11(1), 5418.

[7] Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

[8] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports, 6*(1), 27755.

[9] Estes, W. K. (1993). Models of categorization and category learning. In *Psychology of Learning and Motivation* (Vol. 29, pp. 15-56). Academic Press.

[10] Goetschalckx, L., Andonian, A., & Wagemans, J. (2021). Generative adversarial networks unlock new methods for cognitive science. *Trends in Cognitive Sciences, 25*(9), 788-801.

[11] Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35*(27), 10005-10014.

[12] Ho, J., Jain, A., & Abbeel, P. (2020, December 16). Denoising Diffusion Probabilistic models. arXiv.org. https://arxiv.org/abs/2006.11239

[13] Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive science, 47*(1), e13226.

[14] Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science, 1*, 417-446.

[15] Kruschke, J. K. (2008). Models of categorization. *The Cambridge handbook of computational psychology*, 267-301.

[16] Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

[17] Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual review of neuroscience*, *19*(1), 577-621.

[18] Nguyen, A. P., & Martínez, M. R. (2020). Learning invariances for interpretability using supervised VAE. arXiv preprint arXiv:2007.07591.

[19] Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. Psychological review, 111(2), 309.

[20] Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291-303.

[21] Palmeri, T. J., Wong, A. C., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in cognitive sciences*, *8*(8), 378-386.

[22] Pothos, E. M., & Wills, A. J. (Eds.). (2011). *Formal approaches in categorization.* Cambridge University Press.

[23] Pulastya, V., Nuti, G., Atri, Y. K., & Chakraborty, T. (2021, November). Assessing the quality of the datasets by identifying mislabeled samples. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 18-22).

[24] Radford, A., Metz, L., & Chintala, S. (2016, January 7). Unsupervised representation learning with deep convolutional generative Adversarial Networks. arXiv.org. https://arxiv.org/abs/1511.06434

[25] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022, April 13). High-resolution image synthesis with Latent Diffusion Models. arXiv.org. https://arxiv.org/abs/2112.10752

[26] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382-439.

[27] Serre, T. (2016). Models of visual categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(3), 197-213.

[28] Shen, J., & Palmeri, T. J. (2016). Modelling individual difference in visual categorization. *Visual cognition*, *24*(3), 260-283.

[29] Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder?. *Cognitive psychology*, *23*(3), 457-482.

[30] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2022). CUB-200-2011 (1.0) [Data set]. CaltechDATA. https://doi.org/10.22002/D1.20098