electronics

MDPI

Communication

Can Deep Generative Models Explain Brain Function in People with Developmental Dyslexia?

Hiroto Ogawa [1], Sakiko Ogoshi [1,*], Yasuhiro Ogoshi [2] and Akio Nakai [3]

[1] National Institute of Technology, Fukui College, Fukui 916-8507, Japan
[2] The Department of Human and Artificial Intelligent Systems, Graduate School of Engineering, University of Fukui, Fukui 910-8507, Japan; y-ogoshi@u-fukui.ac.jp
[3] Graduate School of Clinical Education & The Center for the Study of Child Development, Institute for Education, Mukogawa Women's University, Hyogo, Nishinomiya 663-8558, Japan; a_nakai@mukogawa-u.ac.jp
* Correspondence: ogoshi@fukui-nct.ac.jp; Tel.: +81-778-62-8280

Abstract: Many developmental disorders are diagnosed based on symptoms, which may result in lumping together multiple causes. This is thought to be a factor that complicates the research and treatment of developmental disorders. The purpose of this study is to provide hypotheses on the causes of brain functions in developmental dyslexia (DD) by constructing and analyzing a simple computational model of visual information processing using a deep generative model. We then analyze three symptoms observed in DD and investigate their functions and causes.

Keywords: developmental disabilities; developmental dyslexia; artificial intelligence; computational neuroscience

Citation: Ogawa, H.; Ogoshi, S.; Ogoshi, Y.; Nakai, A. Can Deep Generative Models Explain Brain Function in People with Developmental Dyslexia? *Electronics* **2023**, *12*, 2305. https://doi.org/10.3390/electronics12102305

Academic Editor: Eiji Takada

Received: 31 March 2023
Revised: 12 May 2023
Accepted: 15 May 2023
Published: 19 May 2023

1. Introduction

Difficulties in reading and writing in the absence of intellectual or visual abnormalities characterize developmental dyslexia (DD). Specific DD people cannot recognize written characters correctly. For example, some DD people might see blurred characters [1], whereas others might perceive mirror characters [2]. Studies have reported individual differences in the visual perception of DD people. However, many have excellent non-verbal skills, and some are active in fields requiring visuospatial cognition, such as art and architecture. DD people display a well-known trade-off in reinforcement learning: strong exploration and weak exploitation [3]. Conversely, people with autism spectrum disorder (ASD) are described as having strong exploitation and weak exploration [3]. Over 80% of people with hyperlexia, which has an opposite profile to DD, are also on ASD [4]. We believed that such a trade-off could be explained by a computational model [5]. DD affects a large proportion of the population, estimated between 5% and 20% [6,7].

DD is a broad disorder category that includes a variety of symptoms, and several hypotheses have been proposed on their mechanisms [3,8], including the magnocellular deficit hypothesis [1,9], the phonological deficit hypothesis, and the cerebellar dysfunction hypothesis. However, no single hypothesis explains all DD cases [3,8], and explaining the mechanisms of DD might require a multifactorial model [3].

Diagnosing developmental disorders is typically based on interviews with patients to determine if they meet the diagnostic criteria when they describe their symptoms [10]. This method is easy to understand; however, it can lead to ambiguous diagnoses and lumping together different causes of DD, resulting in individual differences in treatment efficacy and inconsistent results of studies on DD characteristics. As a result, it is necessary to consider not only symptom-based but also cause-based diagnoses. However, previous studies have shown that a bottom-up approach to identifying the causes of DD based on clinical findings is challenging. Therefore, we attempted a top-down approach using

a simple computational model to identify the causes of DD by modeling different DDs characterized by unique visual perceptual characteristics.

Artificial intelligence (AI) has made significant strides by deep neural networks (DNNs) [11] that originated from the neurosciences [12], as implied by the name "neural networks". Recent research has focused on the similarity of representations between DNN models and the brain. For example, a recent study reported a one-to-one correspondence between "latent variables" in β-variational autoencoder (β-VAE) [13], a type of DNN-based generative model, and neurons in the IT cortex of macaques [14]. β-VAE is a derivative of a deep generative model called variational autoencoder (VAE) [15,16], which has the objective of learning disentangling representations. Studies have suggested that disentangling is a plausible learning objective for the visual brain [14]. We experimentally adopt β-VAE as a model of human visual information processing based on previous research with primates, the macaque monkey [14].

The VAE, which has an encoder/inference model and a decoder/generative model, has been focused on as an unsupervised representation learning method [17]. The encoder compresses inputs to obtain a low-dimensional representation called latent variables, and the decoder reconstructs the original input from latent variables. When representations are disentangled, single latent variables are sensitive to changes in single generative factors. The trained decoder can be used as a generative model by providing arbitrary latent variables. Figure 1 shows the images generated by manipulating the value of a single latent variable arranged horizontally. As shown in Figure 1, each latent variable encodes skin color, age/gender, and image saturation, among others, when β-VAE is trained on face images [13]. VAE has been widely applied as a generative model and a representation learning method because of these characteristics [18–20].

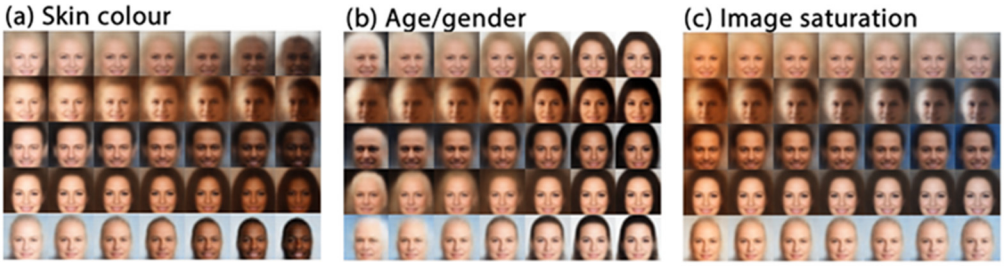

Figure 1. Images generated by β-VAE manipulating single latent variables [13]. (**a**) Skin colour; (**b**) Age/gender; (**c**) Images saturation.

According to the Bayesian brain hypothesis [21], the free energy principles [22], and the world model [20], inference and generative models are essential for humans to recognize the world. Marr indicated that the function of vision is to infer a three-dimensional structure from a two-dimensional retinal image [23]. Therefore, innumerable inference solutions (visual perceptions) are possible, and a generative model, as opposed to an inference model, is needed to determine a single solution. This concept is critical in the free energy principle [21], which is the focus of the unified brain theory. The objective function of the free energy principle and VAE is derived from an identical framework and have similar shapes. What we see is not the state of the world itself but the result of our inferences. Moreover, phenomena such as visual illusions are the result of our inference.

This study aimed to develop a simple computational model of visual information processing using VAE and propose hypotheses on the causes of DD. The study used VAE to analyze the brain at the top two levels of Marr's tri-level hypothesis [23], which is required to understand any complex biological system, the goals of the system (the computational level), and the representations and processes used by the system (the algorithmic/representational level). Marr stated that the computational level is indispensable. At this level, it is necessary to identify the system's inputs and outputs and clarify the purpose of the system's computations. The results of this study, as well as previous stud-

ies [14], suggest that disentangling is a plausible learning objective in DD people's visual information processing. We needed to clarify the representations of inputs and outputs and the information processing algorithm at the algorithmic/representational level. Then, we can analyze the representation of latent variables based on the one-to-one correspondence between latent variables and IT cortex neurons reported in previous studies [14]. On the other hand, our model did not deal with the implementation/physical level. Therefore, this study does not attempt to explain visual information processing in the human brain but proposes functions of visual information processing as one possible form. We proposed four hypotheses on the causes of DD characterized by perceiving blurred characters and also analyzed types of DD characterized by perceiving mirror characters and confused characters by conducting three experiments using VAE as a visual information processing model.

2. Materials and Methods

We applied VAE, β-VAE, and conditional VAE [23], widely utilized as deep generative models, to visual information processing models. We have only described the essential details for understanding the current study because we used standard VAE. For theoretical details of VAE, we refer the reader to the original papers [13,15,16,24]. Our use of VAE and the interpretation of the results are unique to this study. Figure 2 shows the correspondence between VAE architecture and visual information processing proposed in this study. We mapped the inputs to visual information, the latent variables to neurons in the IT cortex and the outputs to visual perception. However, we did not design this study to map the internal processing of DNNs to actual brain processing but to investigate brain functions by interpreting DNN, i.e., changing objective function and visualizing representation of latent variables and inputs and outputs.

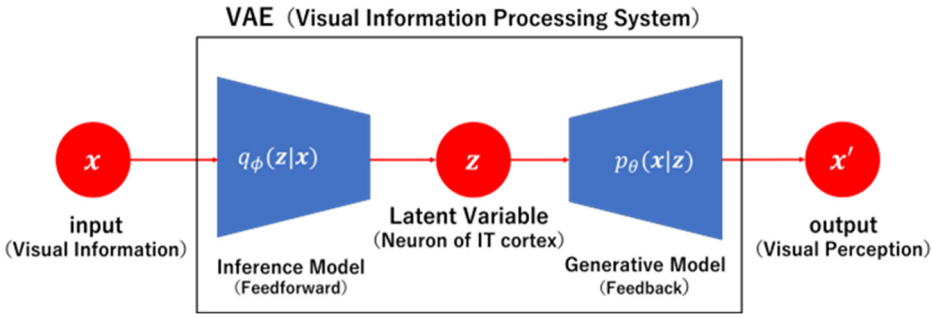

Figure 2. Correspondence between VAE and visual information processing.

We need to use written characters or language datasets to investigate the characteristics of DD by examining their distinctive visual perceptions of written characters. We used MNIST, a widely used dataset of handwritten digit images. The encoder and decoder of VAE were DNNs. The MNIST image ($28 \times 28 = 784$ pixels) was inputted into the encoder and compressed into 10-dimensional latent variables. Then, using the 10-dimensional latent variables, the decoder reconstructed 784-pixel images that closely resembled the inputted image. The VAE architecture we used was a simplified version in previous studies [14]. The encoder consisted of 2 convolutional layers ($32 \times 4 \times 4$ stride 2 and $64 \times 4 \times 4$ stride 2), followed by a 3136-dimensional fully connected layer and 10-dimensional latent variables. The decoder's architecture was the reverse of the encoder.

The objective function of β-VAE is expressed by the following equation [13]:

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta D_{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] \quad (1)$$

The first term (the reconstruction error term) brings inputs and outputs closer together, and the second term (the regularization term) disentangles the representations. Here, $\beta = 1$

is equal to VAE's objective function. Usually, $\beta \gg 1$ is used for disentangling. The second term can be decomposed as follows [25,26]:

$$\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[D_{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]\right] = I(\boldsymbol{x};\boldsymbol{z}) + D_{KL}\left[q_\phi(\boldsymbol{z})||p(\boldsymbol{z})\right] \quad (2)$$

The first term is the mutual information between inputs and latent variables and is the term that we want to maximize. On the other hand, the second term is related to disentangling and is the term that we want to minimize. In other words, there is a trade-off between disentangling and encoding quality.

Conditional VAE is a method of specifying data generated by VAE using labels or other data. For example, as shown in Figure 3, we can specify a particular character to be generated using labels. The characters generated by a single label such as 2, 3, and 4 with continuously changing the two-dimensional latent variables are arranged.

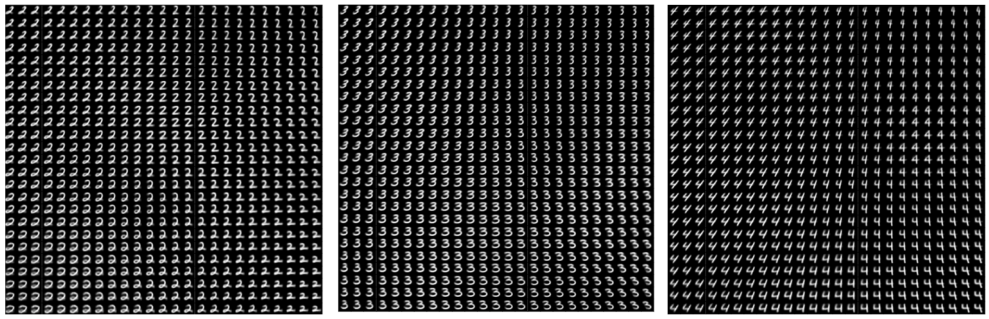

Figure 3. Images generated by conditional VAE specifying a single label.

The brain is optimized for a specific objective function based on predictive coding [27] and the free energy principle [22]. We developed the following hypothesis as a premise: the brain attempts to optimize under a specific objective function, and differences in brain functions are caused by differences in the initial state, the objective function, constraints, and the learning algorithm, among others. This hypothesis implies that the trade-off noted by Taylor et al. is caused by optimization. This study had a fixed initial state and algorithm.

When learning, we can impose constraints on the inputs, the model, and the outputs, to optimize the model to display functions such as DD. However, we did not constrain the inputs in this study because we were developing a model of learning disorders rather than visual impairments. We optimized the model by constraining the model and/or its outputs, such that constraining the outputs when learning indirectly constrained the model. One possible method of accomplishing this was to transform the outputs based on the finding related to DD's visual perception in cognitive psychology. For example, we can train the model so that its output is mirror characters and analyze its function based on the findings that specific DD people perceive mirror characters. Alternatively, we can give the model multiple labels and ask it to analyze their functions based on the findings that other DD people cannot distinguish between characters. Changing the model's objective function or constraining learning based on neuroscientific findings on DD peoples' brain structures could directly constrain the model, for example, by changing the value of β in β-VAE.

We conducted the following experiments:

1. Changing β in β-VAE (similar to a previous study [14]).
2. Developing mirror character perception using β-VAE.
3. Developing character confusion using conditional VAE.

We used the following analytical methods in these experiments.

(a) Visualizing outputs decoded from random latent variables.
(b) Visualizing of inputs that maximized latent variables.
(c) Visualizing of latent variables' distribution (latent space) encoded in test images.

The correspondences between VAE and brain when VAE as a visual information processing model we proposed in this study were as follows. The result of method (a) is expected to indicate visual perception corresponding to neuron activity in the IT cortex; the result of method (b) is the optimal stimulus for IT cortex neurons, and the result of method (c) is the activity of a group of IT cortex neurons. In method (c), compressed 10-dimensional latent variables into two dimensions using t-distributed stochastic neighbor embedding (t-SNE) [28] for visualization.

In method (c), the concepts of between-class variance (S_b) and within-class variance (S_w), which are used in Fisher's discriminant analysis [29] and other methods, are introduced as clustering evaluation indexes for the latent variable. In terms of clustering, S_b should be larger and S_w should be smaller. Therefore, maximizing S_b/S_w is generally considered to be the objective function. In this paper, it is formulated as follows:

$$S_w = \sum_c \sum_{\boldsymbol{z}_c \in \mathcal{D}_c} N_c \left(\boldsymbol{z}_c - \bar{\boldsymbol{z}_c} \right)^2 \tag{3}$$

$$S_b = \sum_c \left(\bar{\boldsymbol{z}_c} - \bar{\boldsymbol{z}} \right)^2 \tag{4}$$

where $\boldsymbol{z}$ is the latent variable, c is the class to which the data belong, $\mathcal{D}$ is the data set, N is the number of data, and $\bar{\boldsymbol{z}}$ is the mean vector. These indexes are interpreted as follows:

- If the between-class variance S_b is large, the latent variable strongly encodes differences in character.
- If the within-class variance S_w is large, the latent variable strongly encodes differences in handwriting.

Since we are considering a model of the brain, S_b does not need to be large, and S_w does not need to be small.

3. Results

3.1. Experiment 1

In this experiment, we used three values of β: 0.05, 1, and 20, which we referred to as low, medium, and high β values, respectively. The latent variables were 10-dimensional. Figure 4 shows the combined results of Experiment 1.

3.1.1. Method (a)

The visual perception was blurred at high β, clear at medium β, and clear at low β while some are not perceived.

The blurred visual perception at high β might match a characteristic of DD. The results at low β suggested that the latent space was sparse, i.e., more representations were available for encoding new visual information, and thus it was easier to learn, which matches a characteristic of hyperlexia.

3.1.2. Method (b)

The optimal stimulus at high β was clear with comparatively little noise. We could identify a number-like pattern.

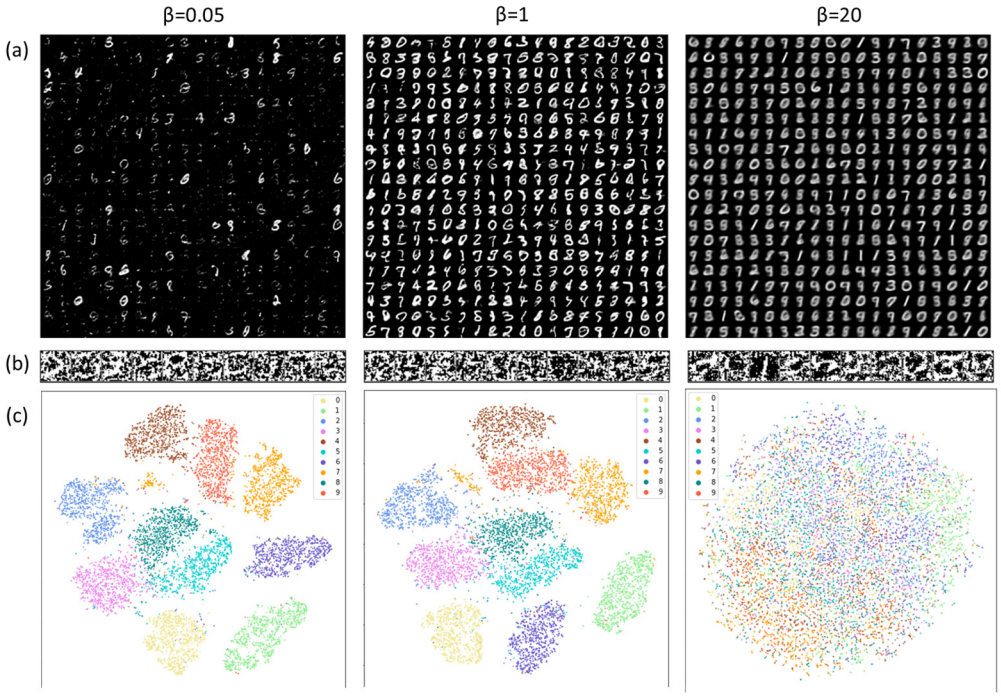

Figure 4. Combined results of Experiment 1. (**a**) The visual perception. The results of decoding randomly sampled latent variables are arranged in 20 × 20, 400 images. (**b**) The optimal stimuli for IT cortex neurons. Ten optimal stimuli are arranged because the latent variables are 10-dimensional. (**c**) The activity of an IT cortex neuron group. The latent variables are compressed to 2 dimensions by t-SNE because they are 10-dimensional. The color coding is based on input image labels.

3.1.3. Method (c)

The model with medium and low β could distinguish between different characters, whereas the model with high β could not. Table 1 shows the clustering indexes. It can be observed that S_b and S_b/S_w become smaller as β increases, i.e., as the disentangling increases. Table 1 also suggests that the representation is sparse at low β.

Table 1. Clustering index in Experiment 1.

β	S_b	S_w	S_b/S_w
0.05	98,544	64,022	1.539
1.0	35,889	53,145	0.675
20.0	10,606	78,613	0.134

3.1.4. Summary

It is suggested that by changing the value of β, characteristics between DD and hyperlexia can be found. We could identify characteristics common to DD and hyperlexia by changing the value of β. The trade-off caused by the nature of objective function in β-VAE [25,26] led to the results of method (a). The higher β resulted in, the worse encoding because of the trade-off in disentangling and mutual information between inputs and latent variables. In generative models, this trade-off is a factor interrupting the high-quality generation. Various improvements have been proposed to overcome this interruption. Nevertheless, this trade-off might be an essential brain characteristic.

A medium-β model would be the highest quality model from the generative models' perspective because of the relatively good balance between encoding and decoding quality. However, both low and high β have merits for the brain, suggesting that we can observe DD and hyperlexia characteristics because of optimization. Therefore, we might be able

to explain the characteristics of DD and hyperlexia as a continuum based on different objective optimization functions.

The results of method (b) indicated that the optimal stimulation of neurons was independent in high β. According to McCulloch-Pitts' classical model that formed the basis of neural networks [30], we assumed that disentangling is related to inhibiting the synchronous firing of neurons and proposed four hypotheses on the type of DD causing blurred visual perceptions. We suggest that future research clinically validate these hypotheses and examine the possibility that opposite hypotheses might be valid for hyperlexia.

1. A small number of common inputs.
2. A low synaptic weight.
3. A high firing threshold.
4. Many inhibitory synaptic inputs and/or few excitatory synaptic inputs.

3.2. Experiment 2

We used an experimental procedure similar to Experiment 1 and trained the model to horizontally flip the visual perception at a probability of 30%, which resulted in mirror characters. To be precise, we calculated the reconstruction error for the mirror inputs. Figure 5 shows the results of the experiment.

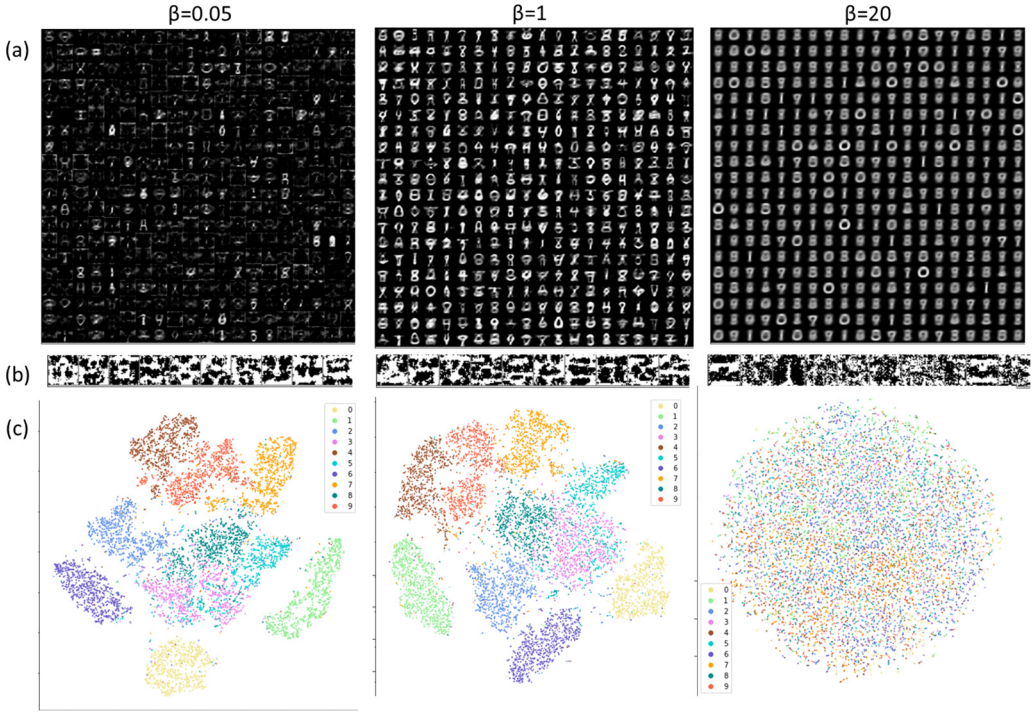

Figure 5. Combined results of Experiment 1. (**a**) The visual perception. The results of decoding randomly sampled latent variables are arranged in 20 × 20, 400 images. (**b**) The optimal stimuli for IT cortex neurons. Ten optimal stimuli are arranged because the latent variables are 10-dimensional. (**c**) The activity of an IT cortex neuron group. The latent variables are compressed to 2 dimensions by t-SNE because they are 10-dimensional. The color coding is based on input image labels.

3.2.1. Method (a)

Figure 5 shows that specific outputs for medium and high β were flipped or overlapped. The effect of flipping at high β was comparatively less.

3.2.2. Method (b)

The optimal stimuli were clearer than in Experiment 1.

3.2.3. Method (c)

No significant differences were observed between Experiments 2 and 1. Table 2 shows the clustering index. The same tendency as Table 1 can be observed. Compared to Table 1, it can be observed that S_w is larger and S_b/S_w is smaller.

Table 2. Clustering index in Experiment 2.

β	S_b	S_w	S_b/S_w
0.05	151,736	108,474	1.398
1.0	33,426	57,530	0.581
20.0	5879	82,856	0.070

3.2.4. Summary

Method (b) results suggested that the perception of mirror characters might be effective for disentangling. Conversely, the perception of mirror characters could be a characteristic arising from the purpose of disentangling. The world consists of many symmetrical natural objects, and DD people might intensively learn symmetry as prior knowledge (generative model) about the natural world. The inputted images are sometimes flipped or rotated in deep learning for data augmentation to improve the model's generalization performance.

However, we forcibly flipped visual perceptions in this experiment and failed to understand the causes of perceiving mirror characters. We suggest that future research develop effective analysis methods for networks.

3.3. Experiment 3

We used a conditional VAE with 10-dimensional latent variables in Experiment 3. The labels were specified by inputting one-hot representation labels into the encoder. In a regular one-hot expression, the number of elements in a vector is equal to the number of labels, and only the label to be specified is set to 1, with all other labels set to 0. In this experiment, however, provided three extra elements and multiple labels. The size of the one-hot vectors was 13. We only used method (a) in Experiment 3.

3.3.1. Providing Multiple Labels to a Trained Model

We first attributed multiple labels to be generated by setting several vector elements to 1. Figure 6 shows an example of a vector with labels 3 and 8 and another with labels 5, 6, and 9. The figure shows that visual perception mixes up the characters' features.

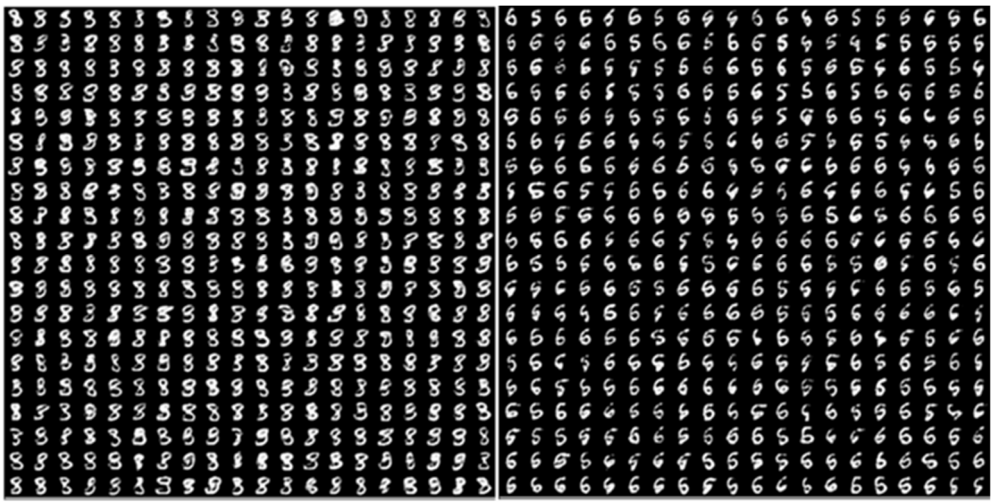

Figure 6. Images generated by conditional VAE specifying multiple labels.

Labels correspond to numbers and language information. According to the model we used, Experiment 3 corresponds to visually perceiving indistinguishable characters. DD people experience difficulties in reading and writing new characters because some characters, including "b", "d", "ξ", and "ζ", among others, are indistinguishable to them. DD people's visual perceptions might be similar to that shown in Figure 6. We suggest examining the clinical validity of this possibility in future research.

The model in Experiment 3 was trained using labels in the MNIST dataset. However, the labels were unknown, and it is necessary to learn a classifier for the labels from visual information. We believe that semi-supervised learning using conditional VAE [24] is similar to human learning of written characters. Semi-supervised learning uses only a small number of labels given in the dataset. The classifier is trained simultaneously with conditional VAE; if no label is provided, the model uses the label output by the classifier. Therefore, both visual and linguistic information might be involved in character confusion.

3.3.2. Providing Multiple Labels during Model Training

Next, we considered providing multiple labels during the model's training. Specifically, random values were given to the three extra labels during training. Figure 7 shows the results of giving the three extra labels to the model. The three labels produced various results, including some that looked like characters and others that did not, which is one possible form of character confusion. Therefore, multiple labels might be given during training, while the incomplete classifier may cause character confusion.

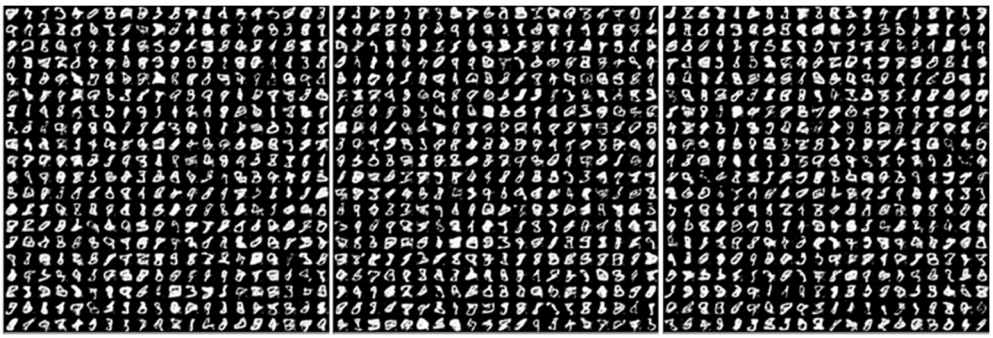

Figure 7. Images generated by conditional VAE specifying extra labels.

4. Discussion

This study focused on disentangling as a plausible learning objective in the visual brain [14,26]. We proposed a correspondence between VAE and visual information processing and conducted three experiments to analyze visual information processing in DD people. Experiment 1 suggested that differences in the value of β in the objective function show characteristics of both hyperlexia and DD. Then, we proposed four hypotheses on the causes of DD based on synchronous firing. In Experiment 2, we suggested that perceiving mirror characters could facilitate disentangling. However, this notion did not explain the cause of perceiving mirror characters. Experiment 3 suggested that the cause of character confusion is related not only to visual information but also to numbers and language information. Therefore, we concluded that this study had achieved its purpose.

This was an exploratory study. The advantage of computational models is they allow top-down designs and analyses of information processing. This study designed the objective functions and the architecture and analyzed the model from different perspectives. Future work is needed to verify these hypotheses and develop improved models and analytical methods. This study's information processing flow analysis was insufficient. We believe that time development and attention, one-shot and semi-supervised learning, memory, multimodal learning, and embodiment must be included in the model to ensure its biological plausibility.

The hypothesis of a "few common inputs" proposed in Experiment 1 is consistent with previous studies [31]. It is known that the minicolumn circuits of DD people have weak local and strong global connectivity than controls or ASD people, which might be explained by top-down approaches using a computational model. It has also been indicated that DD is associated with hyperlexia, attention deficit hyperactivity disorder (ADHD), and ASD [3]. Therefore, we suggest comparing with known anatomical findings of these disorders.

Author Contributions: Conceptualization, H.O., S.O., Y.O. and A.N.; methodology, H.O. and S.O., formal analysis, H.O.; investigation, H.O., S.O., Y.O. and A.N.; data curation, H.O.; analyzed the data, H.O.; writing—original draft, H.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by KAKENHI (Grant-in-Aid for Scientific Research (C):22K12283). The APC was funded by the National Institute of Technology.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or the decision to publish the results.

References

1. Stein, J. The magnocellular theory of developmental dyslexia. *Dyslexia Chichester Engl.* **2001**, *7*, 12–36. [CrossRef] [PubMed]
2. Fernandes, T.; Leite, I. Mirrors are hard to break: A critical review and behavioral evidence on mirror-image processing in developmental dyslexia. *J. Exp. Child Psychol.* **2017**, *159*, 66–82. [CrossRef] [PubMed]
3. Taylor, H.; Vestergaard, M.D. Developmental Dyslexia: Disorder or Specialization in Exploration? *Front. Psychol.* **2022**, *13*, 889245. [CrossRef]
4. Ostrolenk, A.; Forgeot d'Arc, B.; Jelenic, P.; Samson, F.; Mottron, L. Hyperlexia: Systematic review, neurocognitive modelling, and outcome. *Neurosci. Biobehav. Rev.* **2017**, *79*, 134–149. [CrossRef]
5. Ogawa, H.; Ogoshi, S.; Ogoshi, Y.; Nakai, A. Can Deep Generative Models explain brain function in people with Developmental Dyslexia? *Abstr. Kosen Res. Int. Symp.* **2023**, *2023*, 182.
6. Badian, N.A. Reading disability in an epidemiological context incidence and environmental correlates. *J. Learn. Disabil.* **1984**, *17*, 129–136. [CrossRef]
7. Wagner, R.K.; Zirps, F.A.; Edwards, A.A.; Wood, S.G.; Joyner, R.E.; Becker, B.J.; Liu, G.; Beal, B. The Prevalence of Dyslexia: A New Approach to its Estimation. *J. Learn. Disabil.* **2020**, *53*, 354–365. [CrossRef]
8. Ramus, F.; Rosen, S.; Dakin, S.C.; Day, B.L.; Castellote, J.M.; White, S.; Frith, U. Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain* **2003**, *126*, 841–865. [CrossRef] [PubMed]
9. Stein, J. The current status of the magnocellular theory of developmental dyslexia. *Neuropsychologia* **2019**, *130*, 66–77. [CrossRef] [PubMed]
10. Association, A.P. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed.; American Psychiatric Association: Washington, DC, USA, 2013. [CrossRef]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 7553. [CrossRef]
12. Hassabis, D.; Kumaran, D.; Summerfield, C.; Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **2017**, *95*, 245–258. [CrossRef] [PubMed]
13. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022; Available online: https://openreview.net/forum?id=Sy2fzU9gl (accessed on 28 March 2023).
14. Higgins, I.; Chang, L.; Langston, V.; Hassabis, D.; Summerfield, C.; Tsao, D.; Botvinick, M. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* **2021**, *12*, 6456. [CrossRef] [PubMed]
15. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114. [CrossRef]
16. Kingma, D.P.; Welling, M. An Introduction to Variational Autoencoders. *arXiv* **2019**, arXiv:1906.02691. [CrossRef]
17. Tschannen, M.; Bachem, O.; Lucic, M. Recent Advances in Autoencoder-Based Representation Learning. *arXiv* **2018**, arXiv:1812.05069. [CrossRef]
18. Higgins, I.; Sonnerat, N.; Matthey, L.; Pal, A.; Burgess, C.P.; Bosnjak, M.; Shanahan, M.; Botvinick, M.; Hassabis, D.; Lerchner, A. SCAN: Learning Hierarchical Compositional Visual Concepts. *arXiv* **2018**, arXiv:1707.03389. [CrossRef]
19. Eslami, S.A.; Jimenez Rezende, D.; Besse, F.; Viola, F.; Morcos, A.S.; Garnelo, M.; Ruderman, A.; Rusu, A.A.; Danihelka, I.; Gregor, K.; et al. Neural scene representation and rendering. *Science* **2018**, *360*, 1204–1210. [CrossRef]
20. Ha, D.; Schmidhuber, J. World Models. *arXiv* **2018**, arXiv:1803.10122. [CrossRef]
21. Doya, K. *Bayesian Brain: Probabilistic Approaches to Neural Coding*; MIT Press: Cambridge, MA, USA, 2007.

22. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 2. [CrossRef]
23. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*; MIT Press: Cambridge, MA, USA, 2010.
24. Kingma, D.P.; Mohamed, S.; Rezende, D.J.; Welling, M. Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Available online: https://proceedings.neurips.cc/paper_files/paper/2014/hash/d523773c6b194f37b938d340d5d02232-Abstract.html (accessed on 28 March 2023).
25. Chen, R.T.Q.; Li, X.; Grosse, R.; Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv* **2019**, arXiv:1802.04942. [CrossRef]
26. Burgess, C.P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; Lerchner, A. Understanding disentangling in β-VAE. *arXiv* **2018**, arXiv:1804.03599. [CrossRef]
27. Rao, R.P.N.; Ballard, D.H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **1999**, *2*, 1. [CrossRef]
28. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
29. Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
30. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
31. Williams, E.L.; Casanova, M.F. Autism and dyslexia: A spectrum of cognitive styles as defined by minicolumnar morphometry. *Med. Hypotheses* **2010**, *74*, 59–62. [CrossRef] [PubMed]