Hindawi
Mobile Information Systems
Volume 2023, Article ID 9857590, 1 page
https://doi.org/10.1155/2023/9857590

Retraction

Retracted: Multicategory Image Recognition Based on Image Semantic Features and Transformer

Mobile Information Systems

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] H. Chen, C. Xiang, D. Qiu, and X. Huang, "Multicategory Image Recognition Based on Image Semantic Features and Transformer," *Mobile Information Systems*, vol. 2022, Article ID 4508507, 8 pages, 2022.

Hindawi
Mobile Information Systems
Volume 2022, Article ID 4508507, 8 pages
https://doi.org/10.1155/2022/4508507

Research Article

Multicategory Image Recognition Based on Image Semantic Features and Transformer

Hewei Chen, Chen Xiang, Dong Qiu, and Xuxiang Huang

School of Computer Science and Information Engineering, Hubei University, Wuhan, 430062 Hubei, China

Correspondence should be addressed to Hewei Chen; heywechen@stu.hubu.edu.cn

Received 25 February 2022; Accepted 6 April 2022; Published 9 May 2022

Academic Editor: Mian Ahmad Jan

In the field of computer vision research, generative adversarial networks (GAN) are used for general object recognition. In recent years, however, GAN have learned only from image data without using label information. In recent years, however, unsupervised learning, which learns GAN only from image data without using label information, where GAN are learned from image data alone without using label information, has been introduced. In this paper, we describe research on unsupervised learning of GAN since the introduction of transformer, reviewing trends in computer vision/artificial intelligence-related research since the introduction of transformer from a visual neuroscience perspective.

1. Introduction

With the widespread use of the Internet, it has become possible to use crowdsourcing to collect large amounts of image data and build a large-scale database of tagging information [1]. This database was provided as a benchmark for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [2]. In the context of this proliferation of large-scale image data for training purposes, a new type of computer vision system has emerged whose basic architecture is based on GAN. In most conventional approaches, one designs image features that are effective for general object recognition and learns to recognise them. Transformer, on the other hand, uses classical convolutional algorithms. In contrast, transformer is a classical convolutional neural network, which is only multilayered.

The architecture of GAN with hierarchical convolutional processing was originally inspired by the way information is processed in the visual cortex of the brain. GAN trained from scratch using large amounts of image data for general object recognition have been reported to show a layered representation of information homologous to the ventral visual pathways of the brain. That is, the convolutional weights of neurons in the first layer of the GAN exhibit a Gabor filter-like weight distribution consisting of various orientations and spatial frequencies, like neurons in area V1, while neurons at higher levels exhibit object class selectivity, like neurons in the inferior temporal lobe [3].

Although GAN have received much attention for their similarity and homology to visual information processing and human cognitive abilities in the brain, their differences have also been pointed out. In supervised learning of GAN based on labelled information, discrimination criteria are learned based on the training data. As a result, discrimination errors (generalisation problems) can occur with untrained data, even for images in which humans do not make mistakes [4, 5].

Supervised GAN are fragile because they are unable to learn a properly informative representation of natural images of natural images. If the representation is inappropriate, it is possible that images that should be easily distinguishable are represented near the boundaries because the distinctions there are too subtle. Therefore, using a larger image database, it is possible to reflect the statistical properties of natural images without relying on labelling information [6]. If GAN can obtain internal representations that reflect the statistical properties of natural images, GAN can be highly resistant to adversarial attacks. In addition to being robust to adversarial attacks, GAN can be adapted to a variety of vision tasks other than object recognition. If GAN can

obtain internal representations that reflect the statistical properties of natural images without relying on labelling information, they are not only robust to adversarial attacks but also more adaptable to a variety of vision tasks other than object recognition [7].

A topic related to the generalisation problem is the use of teacher-trained GAN with low discrimination accuracy on new datasets or new tasks. In addition, GAN have been reported to fail to maintain high recognition performance without directly learning such changes, even though image changes are easily processed by humans. It has also been reported that GAN do not maintain high discrimination performance without directly learning changes.

2. Related Work

Although the improvement of GAN objective function improves the generation effect, the improvement of generation quality alone is not enough to meet the demand of generated data in practical applications. The emergence of Conditional GAN (CGAN) [8] addresses how to generate samples with specified labels based on multilabel data. Info GAN (mutual information) [9] is a method to split the structured implicit encoding from the input noise on the generator based on CGAN, which makes the generation process with a certain degree of controllability and the generation results with a certain degree of interpretability. Pix2Pix (map pixels to pixels), i.e., pixel-to-pixel mapping [10], based on CGAN is used to solve numerous problems in the field of image translation. However, the drawback of this model is that the training of Pix2Pix requires mutually paired images, yet such data is extremely scarce [11]. Using CycleGAN requires one-to-one training one by one, which is obviously inefficient. StarGAN [12], as a further extension of CycleGAN, emanates the mapping relationship between one-to-one into a mapping between multiple domains [13].

In summary, generative adversarial networks are widely used and have the potential and value for continued research.

3. Transformer

Transformer is an 8-layer neural network consisting of 5 convolutional and 3 fully combinatorial layers, while the later emergence of VGGNet is a 16-19 layer network. As a result, the state-of-the-art models approach human visual function in terms of object recognition performance. In this paper, we present a new brain model based on brain scores, which is a predictive performance on neural activity data. In this study, we quantitatively assessed the effectiveness of the brain model based on the brain score, which is the predictive performance of neural activity data [14]. It has been proposed to quantitatively assess the validity of brain models based on brain scores, which are predictive performance on neural activity data, rather than simply based on object recognition performance.

We refer to factors that explain large changes in the external world as "meaningful" factors. And in a narrow sense, disentanglement means that "each dimension of the underlying variable z is independent or uncorrelated."

$$I(\mathbf{x};\mathbf{z}) = \sum_{x,z} p(\mathbf{x},\mathbf{z}) \log \frac{p(\mathbf{x}|\mathbf{z})}{p(\mathbf{x})}. \tag{1}$$

Unsupervised learning methods differ in the way they design the loss function or objective function for learning. Unsupervised learning methods can be classified as predictive or control methods. The generate/predictive model as shown in Figure 1. On the contrary, when data x_0 and x_1 are input, their relationship in the output is between z_0 and z_1 as a loss function/objective function.

3.1. Programme Improvements. Another learning method known as instance learning, in which each training image is identified as a separate class, has also attracted attention. This method is implemented as a contrastive learning and achieves robust learning of the internal representations of natural images, although it has the limitation that the index of each training image and its internal representation must be stored in memory. It is implemented as a form of contrastive learning [15]. It is implemented as a contrastive learning and achieves a robust internal representation of natural images. It has also been pointed out that it has a better correspondence with brain information representations than traditional supervised GAN.

3.2. General Object Recognition. From 2019, contrast learning unsupervised learning models were used. Unsupervised learning models can achieve highly accurate general object recognition comparable to GAN. From 2019 onwards, unsupervised learning models using contrastive learning have been reported to achieve high accuracy in general object recognition comparable to supervised GAN [16]. As of January 2021, the best performing model is SimCLR35, and other methods have been implemented based on very similar ideas. In this paper, we will focus on the implementation of SimCLR and describe contrast learning as a representation learning method that can be effective in improving the accuracy of object recognition in general [17, 18].

Contrast learning is the process of determining how an image is in the latent variable space. As a loss function/objective function, in the latent variable space. In contrast, in general object recognition tasks, an object is considered to have a very different appearance in an image, depending on the viewing conditions [19].

In the general object recognition task, an object can be judged to be the same object even if it looks very different in the image due to different observation conditions. In a general object recognition task, it is required to be able to judge an object as the same object and distinguish it from different in a general object recognition task; it is required to be able to judge an object as the same object and distinguish it from different images of objects, even if they look very different in different images due to different observation conditions. It is therefore necessary in this paper that we describe how to create a positive sample from the original image using various image processing methods (different

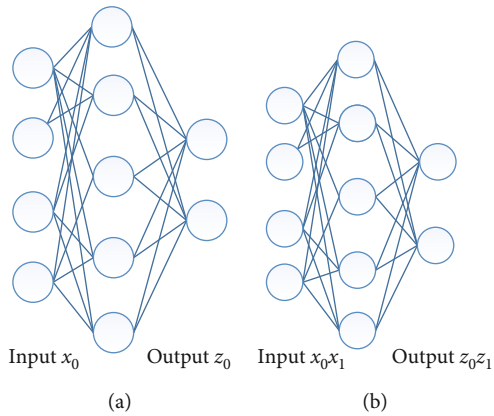

FIGURE 1: Classification of learning methods: (a) generative/predictive methods; (b) contrastive methods.

views). (1) Creating a frontal sample from the original image using various image processing (different views), samples created by image processing of different images are called negative samples. Positive samples = similar images are close to each other, negative samples = similar images are close to each other, and dissimilar images are far from each other. The procedure for mapping the internal representation space is shown in Figure 2.

3.3. Data Enhancement. For image processing, we use data enhancement methods (cropping, rotation, scaling, Gaussian noise, colour distortion, etc.), and these methods are also used for supervised learning of GAN. A suitable GAN, such as ResNet50, is prepared as a coding model/encoder, and its output is used as the latent variable Z. We use the information noise contrast loss shown in equation (2). The latent variable representation for each image sample is shown in equation (2). The latent variable representation for each image sample is normalized to a criterion and distributed over a multidimensional hypersphere, where the exponent of the distance between positive samples is the exponent of the distance between positive samples in the numerator and negative samples in the numerator.

$$L_N = E_{x\in X}\left[-\log \frac{\exp\left(\mathrm{sim}\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[i!=k]} \exp\left(\mathrm{sim}\left(z_i, z_k\right)/\tau\right)}\right]. \quad (2)$$

As shown in the schematic diagram in Figure 3, similar images are located near each other, while dissimilar images are located far away from each other. Similar images are mapped to the nearest neighbourhood, and dissimilar images are mapped to the furthest neighbourhood. The loss function is designed in such a way that, by this way, the data is augmented, and it is invariant to the predetermined disturbances of the data augmentation.

The implementation in SimCLR has the advantage of learning object recognition performance efficiently, but it is not as powerful as a purely computational model implementation of the brain, as it requires an implementation in SimCLR that is difficult to implement as a purely computational model of the brain.

It turns out that the loss function is best learned by maximising mutual information. As equation (3) shows, the size N of the image samples also needs to be increased for better representation learning, and how to retain a large number of negative samples is a problem for computational models of the brain. It seems to be a challenge for computational models of the brain to retain a large number of negative samples.

N is the number of training samples or the batch size during training. τ is a temperature constant. Function $1[i! = k]$ is a function that is 0 when $i = k$. Function $1[i! = k]$ is a function that is 0 when $i = k$ and 1 otherwise.

$$I(x; z) \geq \log(N) - L_N. \quad (3)$$

3.4. Short Textbook Categories. In the field of natural language processing, unsupervised learning models based on the transformer architecture, such as BERT and GPT-n, dominate the research. The input data is a sequence of words used as markers (Figure 4). By repeating this process for many layers, we can learn the next sentence based only on cooccurrence relations based on the word order in the sequence. The method is based on unsupervised learning, predicting the next occurrence of a sentence or sentence filler based only on cooccurrence relationships based on the word order in the sequence (see research papers related to natural language processing for details). By scaling up the amount of data used for learning and the size of the network parameters, the method can achieve significant improvements in accuracy and even high performance with only a few samples for tasks where there is no direct training (Few-Shot Learning). It also shows a very high degree of generalisability. The structure of the transformer can also be applied to image processing by transforming images into one-dimensional arrays. In fact, there are many reports on the use of transformers in image processing. For example, if we simply divide an image into patches and assign them directly to the transformer, we can achieve the performance of a traditional supervised learning GAN when trained on a very large database of labelled images [20].

Images are arranged in one dimension at the pixel level, and unsupervised prediction of the next pixel or missing pixel is learned unsupervised. It has been reported that unsupervised learning of the prediction of the next pixel or missing pixel can lead to an internal representation suitable for image recognition.

The unsupervised learning-based image generation frameworks GAN and VAE are two widely used image generation frameworks based on unsupervised learning.

In VAE, the basic components are the encoder, which is responsible for the representation transformation from the image data x to the latent variable z, and the decoder, which is responsible for the recovery from the latent variable z to the original image data x'. In VAE, the latent variable z (also known as the bottleneck due to its hourglass structure) is constrained to a normal distribution,

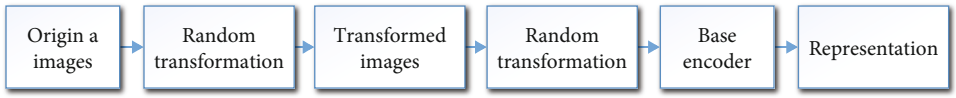

Figure 2: SimCLR framework.

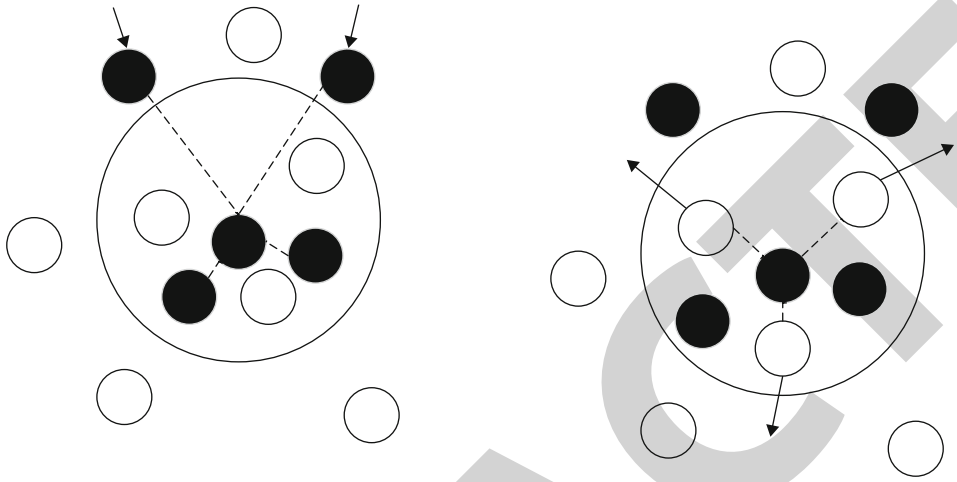

Figure 3: Schematic representation of contrasting losses.

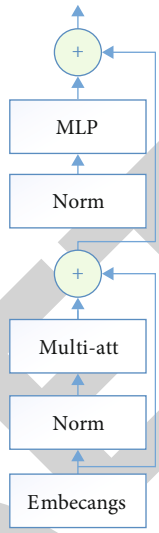

Figure 4: Transformer-based natural language processing framework.

and the encoder and decoder are trained to minimise the restoration error between the training/raw image x and the generated/restored image x'. To obtain a generative model distribution $p_\theta(\mathbf{x})$ that approximates the true data distribution $p_{\text{data}}(\mathbf{x})$ and to approximate the true data distribution $p_{\text{data}}(\mathbf{x})$, the Kullback-Leibler (KL) distance between the two distributions can be calculated. It can be seen from equation (4) that the parameter that maximises the expected value of the log-likelihood function of the model is the parameter that finds the expected value that maximises the model. The first term is a constant term determined by the data sample.

$$\begin{aligned} D_{\text{KL}}(p_{\text{data}}(\mathbf{x})\|p_\theta(\mathbf{x})) &= E_{p_{\text{data}}(\mathbf{x})}\left[\log \frac{p_{\text{data}}(\mathbf{x})}{p_\theta(\mathbf{x})}\right] \\ &= E_{p_{\text{data}}(x)}[\log p_{\text{data}}(\mathbf{x})] \\ &\quad - E_{p_{\text{data}}(x)}[\log p_\theta(\mathbf{x})]. \end{aligned} \tag{4}$$

The objective function of VAE is to maximise the ELBO (Evidence Lower Bound) of the log-likelihood function of the log-likelihood function. The objective function of VAE is to maximise the ELBO (Evidence Lower Bound) of the log-likelihood function. VAE is represented by a reconstruction-related error term, also known as the negative reconstruction error, and a regularisation term. It is represented by a KL distance term, which is a regularisation term and a reconstruction-related error term, also known as the negative reconstruction error (equation (5)).

$$L_\beta = \max_{\varphi,\theta} \frac{1}{N}\sum_{n=1}^{N}\left(E_{q_q(z|x_n)}[\log p_\theta(\mathbf{x}_n|z)] - \beta D_{\text{KL}}\left(q_\varphi(z|\mathbf{x}_n)\|p_z(z)\right)\right). \tag{5}$$

Many studies on the use of VAE to unwind internal representations have theoretical support. One of the most widely used is β VAE49, which improves the untangling of representations by adjusting the penalty of the KL distance term in the VAE objective function (equation (6)), $\beta > 1$. By adjusting the penalty of the KL distance term

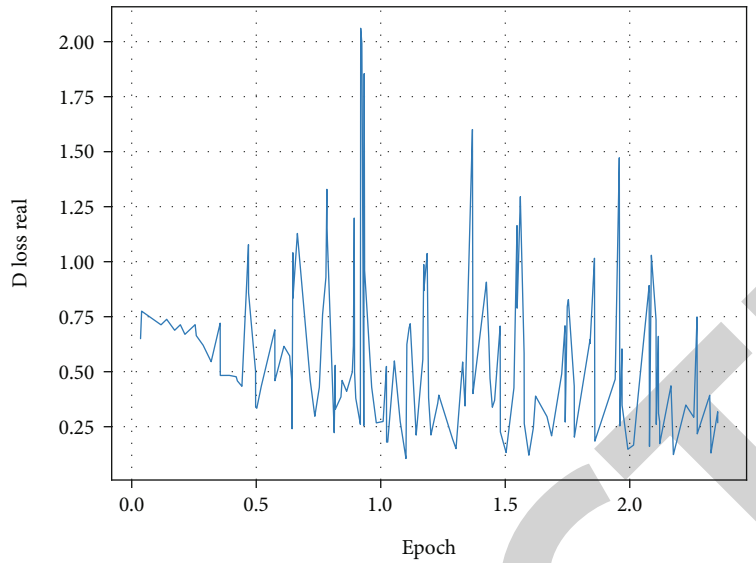

Figure 5: Trend of d loss real on MNIST.

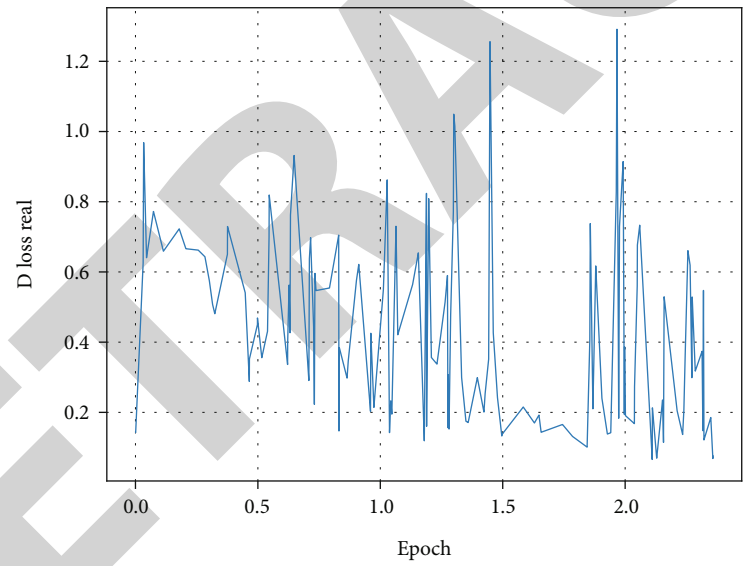

Figure 6: Trend of d loss fake on MNIST.

in the VAE objective function (equation (6)), $\beta > 1$ (in normal VAE, $\beta = 1$). In this study, we compared the internal representation of the β-VAE with the neural representation of the monkey visual cortex. Since the first term is an error term related to image reconstruction, the penalty for β is somewhat larger. However, the first term is the error term associated with image reconstruction. (β-TCVAE51) further decomposes the KL distance term into the equation; (β-TCVAE51) further decomposes the KL distance term into the equation; (β-TCVAE51) further decomposes the KL distance term into the equation and imposes a penalty only on the total correlation term ($\beta > 1$) to avoid correlation between the dimensions of the potential variable z. The quality of the generated images and the separation of representatives was improved [21].

$$L_{\beta-TC} = \max_{\varphi,\theta} E_{q_q(z|n)p(n)}[\log p_\theta(n|z)] - I_{q_\varphi}(z;n) - \beta D_{\mathrm{KL}}\left(q_\varphi(z)\left\|\prod_j q_\varphi(z_j)\right.\right) - \sum_i D_{\mathrm{KL}}\left(q_\varphi(z_j)\|p_z(z_j)\right). \tag{6}$$

4. Experimental Analysis

The dataset in this paper is MNIST, and the experimental results are shown in Figures 5–8. In recent years, it has been

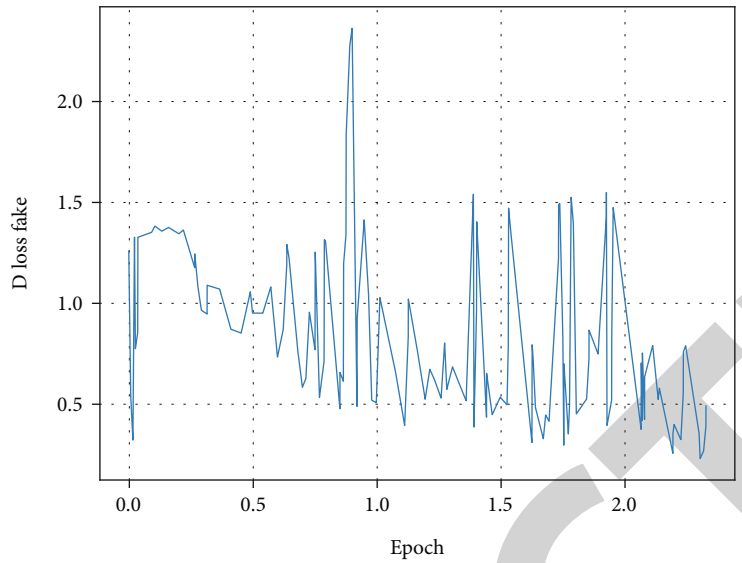

Figure 7: Trend of d loss on MNIST.

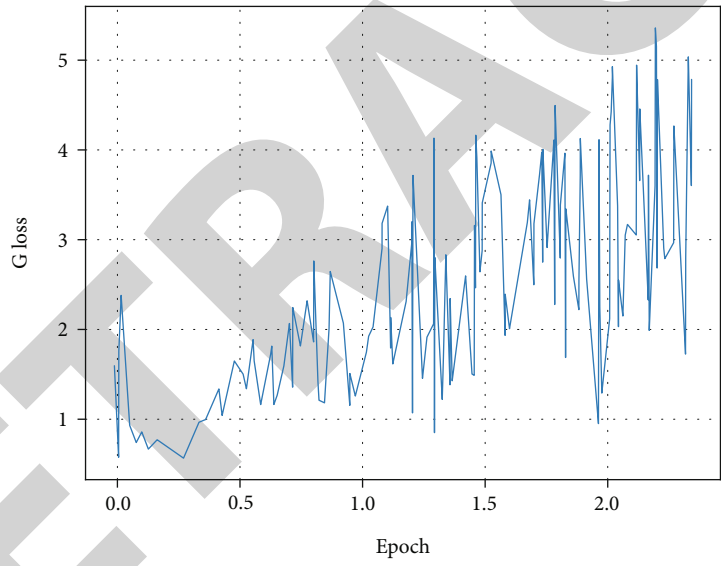

Figure 8: Trend of g loss on MNIST.

reported that it is not necessary to set up direct data augmentation based on unsupervised learning models as a brain learning model. However, as a brain-based learning model, it seems necessary to devise a method to handle negative samples without directly setting up data augmentation. After learning the decomposed latent variable representations, it is necessary to learn the invariant representations of each factor. It would be interesting to propose a mechanism to learn the invariant representation of each factor, after learning the decomposed latent variable representation.

From Figures 7 and 8, it can be seen that the generator and discriminator are smooth at the beginning of training, but as the number of training increases, the model becomes more stable, and the two network structures fight against each other, showing a large oscillation in the figure. The data is efficient, generalisable, and robust. The reasons for this are

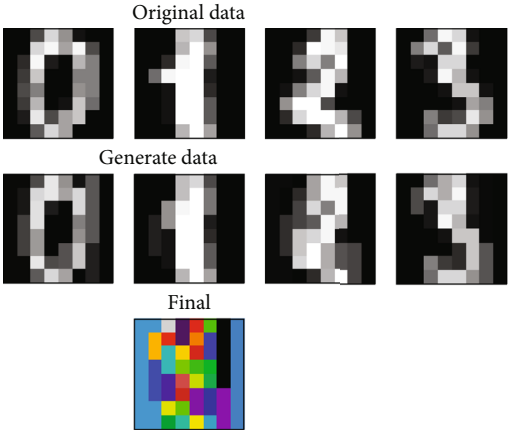

Figure 9: MNIST generated sample.

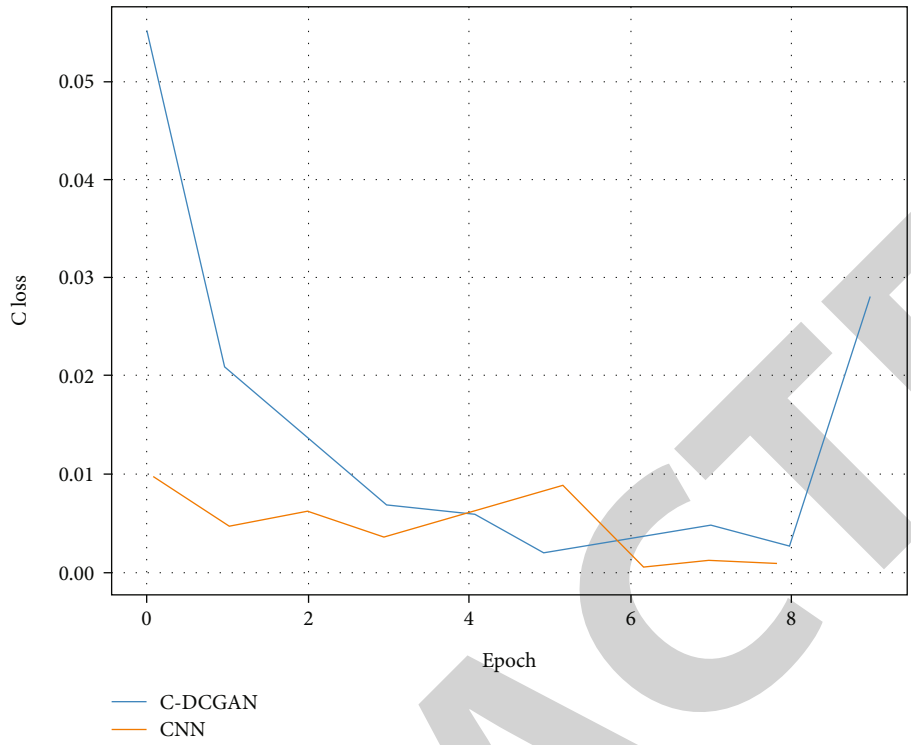

Figure 10: Trend of *c* loss on MNIST.

high data efficiency, extended generalisability, and improved robustness. This is due to its high data efficiency, extended generalisability, and improved robustness. Labelling required for supervised learning the collection of information requires manual labelling. This is an expensive process, and it does not scale with the size of the database.

Figure 9 shows the change of the generated samples for the first 64 samples of the dataset as the number of epochs increases.

In order to verify the advantages of GAN algorithm for image classification, a separate CNN model is trained as a comparison, which is structurally identical to the GAN discriminator in this paper, and the data is normalized before training. Figure 10 shows the comparison of the loss function of the two models with the number of iterations. The cues used by GAN for object recognition also differ from those used by humans. For example, when a method known as image style transfer is used to transform only the texture of an image into another image texture while retaining the shape information contained in the image, most GAN models tend to recognise objects based on the transformed texture.

5. Conclusions

In this paper, we propose disentanglement in representational learning that is often discussed qualitatively without a clear definition because it is difficult to define truly "meaningful" elements or factors. This is because it is difficult to define elements and factors that are truly "meaningful". In a broader sense, disentanglement is the presence of separate representations of "meaningful" factors in the space of potential variables. We conclude that, in a narrow sense, it refers to the separate representation of "meaningful" factors for each dimension of the latent variable (i.e., for each axis in the latent variable space).

Data Availability

The datasets used in this paper are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

References

[1] B. S. Haug and S. M. Mork, "Taking 21st century skills from vision to classroom: What teachers highlight as supportive professional development in the light of new demands from educational reforms," *Teaching and Teacher Education*, vol. 100, article 103286, 2021.

[2] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: a review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[3] R. Ali, M. H. Siddiqi, and S. Lee, "Rough set-based approaches for discretization: a compact review," *Artificial Intelligence Review*, vol. 44, no. 2, pp. 235–263, 2015.

[4] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[5] G. Cai, Y. Fang, J. Wen, S. Mumtaz, Y. Song, and V. Frascolla, "Multi-carrier $ M $-ary DCSK system with code index modulation: An efficient solution for chaotic communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 6, pp. 1375–1386, 2019.

[6] Y. Yang, J. Zhou, J. Ai et al., "Video captioning by adversarial LSTM," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5600–5611, 2018.

[7] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. , 20186964–6974, 2018.

[8] K. Chandra, A. S. Marcano, S. Mumtaz, R. V. Prasad, and H. L. Christiansen, "Unveiling capacity gains in ultradense networks: Using mm-wave NOMA," *IEEE Vehicular Technology Magazine*, vol. 13, no. 2, pp. 75–83, 2018.

[9] D. Wu, C. Zhang, L. Ji, R. Ran, H. Wu, and Y. Xu, "Forest fire recognition based on feature extraction from multi-view images," *Traitement du Signal*, vol. 38, no. 3, pp. 775–783, 2021.

[10] S. Palanisamy, B. Thangaraju, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery, "A Novel Approach of Design and Analysis of a Hexagonal Fractal Antenna Array (HFAA) for Next-Generation Wireless Communication," *Energies*, vol. 14, no. 19, p. 6204, 2021.

[11] S. N. Alsubari, S. N. Deshmukh, A. A. Alqarni et al., "Data analytics for the identification of fake reviews using supervised learning," *CMC-Comput Mater Continua*, vol. 70, no. 2, pp. 3189–3204, 2022.

[12] F. B. Saghezchi, A. Radwan, J. Rodriguez, and T. Dagiuklas, "Coalition formation game toward green mobile terminals in heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 85–91, 2013.

[13] N. Subramani, P. Mohan, Y. Alotaibi, S. Alghamdi, and O. I. Khalaf, "An Efficient Metaheuristic-Based Clustering with Routing Protocol for Underwater Wireless Sensor Networks," *Sensors*, vol. 22, no. 2, p. 415, 2022.

[14] H. S. Gill, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery, "Multi-Model CNN-RNN-LSTM Based Fruit Recognition and Classification," *Intelligent Automation and Soft Computing*, vol. 33, no. 1, pp. 637–650, 2022.

[15] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," in *European Conference on Computer Vision*, pp. 262–277, Springer, Cham, 2020.

[16] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Improved ArtGAN for conditional synthesis of natural image and artwork," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 394–409, 2019.

[17] N. Zakharov, H. Su, J. Zhu, and J. Glaescher, "Towards controllable image descriptions with semi-supervised VAE," *Journal of Visual Communication and Image Representation*, vol. 63, article 102574, 2019.

[18] P. An, Z. Wang, and C. Zhang, "Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection," *Information Processing & Management*, vol. 59, no. 2, p. 102844, 2022.

[19] E. Boran, A. Erdem, N. Ikizler-Cinbis, E. Erdem, P. Madhyastha, and L. Specia, "Leveraging auxiliary image descriptions for dense video captioning," *Pattern Recognition Letters*, vol. 146, pp. 70–76, 2021.

[20] T. Hu, C. Long, and C. Xiao, "A novel visual representation on text using diverse conditional gan for visual recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 3499–3512, 2021.

[21] C. Bai, A. Zheng, Y. Huang, X. Pan, and N. Chen, "Boosting convolutional image captioning with semantic content and visual relationship," *Displays*, vol. 70, article 102069, 2021.