DOI: Registering DOI

A Biomimetic Covering Learning Method Based on Principle of Homology Continuity

Xin Ning^{1,2} , Yuebao Wang^{2,*} , Weijuan Tian² , Liang Liu^{2,*} , Weiwei Cai³

¹ Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China

² Cognitive Computing Technology Joint Laboratory, Wave Group, Beijing 10008, China

³ Central South University of Forestry and Technology, Changsha 410004, China

* Corresponding author: Yuebao Wang and Liang Liu (wangyuebao, liuliang@wavewisdom-bj.com)

Manuscript Revised 21 April 2021; Accepted 22 April 2021; Published 22 April 2021

Academic Editor: Zhenghua Huang

Abstract: The Principle of Homology Continuity (PHC) based covering learning method is an effective method to solve the pattern recognition problem. However, PHC and the existence of optimal coverage are not mathematical proven. To address this issue, we firstly give the mathematical description and theoretical proof of PHC. On this basis, the theoretical definition of optimal coverage is introduced. Optimal coverage can determine the internal connections among samples as prior knowledge and use covering neurons to learn prior knowledge. Finally, we propose a kind of covering neuron model, and the effectiveness of which is demonstrated through extensive experiments conducted on the CIFAR-10, LFW, and YTF datasets.

Index Terms: Brain-inspired, covering neuron models, deep learning, homology continuity, optimal coverage.

1. Introduction

Building “a bridge between computer science and biology” is the key to overcoming the limitations of artificial intelligence (AI) [1]. Since the development of back propagation algorithm, brain science, psychology, and neuro science have made pivotal contributions to AI development. Currently, deep learning has become the main method for AI development. Investigating the biomimetic modeling of neural networks has also become crucial [2]. In recent years, numerous excellent achievements have been made in the realization of bionic functions of neural networks. To improve the performance of convolutional neural networks (CNNs), some studies proposed the use of human visual mechanism models [3], [4], [5], [6], [7], [8], [9], [10]. Capsule neural networks inspired by human brains convert the scalar input and output features in the original neuron structure into a vector output that can store different attributes of features to realize bottom-up learning through dynamic

routing mechanisms [11], [12], [35]. Transformers mechanism is also a method inspired by human brain, which injects strong impetus into the development of deep learning [13], [14], [33]. The classical biomimetic conditioned reflex impulse neural network model [34] contains brain regions, neural circuits, cognitive functions, neuron-scale mechanisms, computer mechanisms, etc. However, current conventional deep learning algorithms are not adequately inspired by the actual human brain. Therefore, biomimetic learning models require further exploration.

Brain cognitive mode and brain neural cell structure are two important aspects that inspire biomimetic modeling. First, in terms of functional mechanism, brain neuroscience suggests that anything seen by humans will converge on a continuous attractor in different ways, or the human brain stores information in the form of manifolds [15], [16]. The attractor may be a curve, surface, or hypersurface that can extrapolate different angles and dimensions [17]. This concept is similar to

Principle of Homology Continuity (PHC) hypothesis. PHC hypothesis assumes that there is a nonquantized gradient process between any two samples of the same class and that any intermediate state in the gradient process belongs to the same category [17], [4]. PHC hypothesis has been widely used as a prior knowledge, and some research achievements have been obtained in many neighborhoods [18], [19], [20], [21]. However, PHC hypothesis does not have a mathematical proof of rationality thus far. The PHC-based optimal coverage model has also not been effectively demonstrated.

For these problems, we firstly provided a rational theoretical proof of PHC hypothesis. Secondly, we introduced a biomimetic covering learning method based on PHC. PHC hypothesis forms the theoretical basis of the method for the construction of covering neuron models. This kind of prior knowledge is used to improve the pattern recognition ability. Subsequently, a hyper-sausage covering (HSC) neural method, which is constructed based on the biomimetic covering learning method, is proposed. The effectiveness of the model is demonstrated through extensive experiments conducted on CIFAR-10, LFW, and YTF datasets.

The remainder of this paper is organized as follows: The section 2 introduces the principle of homologous continuity, including the description of PHC hypothesis, the corresponding theorem and proof. Based on PHC, a new biomimetic covering learning method and a hyper-sausage covering (HSC) neural method were propose in section 3. The experimental results on LFW and YTF datasets were discussed in Section 4. The conclusion was presented in Section 5.

2. Principle of Homology Continuity

PHC hypothesis believes that if two objects in nature are of a similar type, their relationship can be described by a gradual nonquantization process, and any of these intermediate states belong to the same type as these two objects. PHC is the theoretical basis of Biomimetic covering learning.

2.1 PHC Hypothesis

In image recognition, any image is encoded into a feature vectors, and the space in which the feature vectors exist is called a hidden space(feature space). Hidden space is denoted by E^V . On the basis of [17], we restate the mathematical description of the cognate continuity theorem as follows:

PHC hypothesis. 1 In hidden space E^V , the set M_A consists of all samples of category (label) A , $M_A \subset E^V$. $\forall X, Y \in M_A$, where $X \neq Y$, $\forall \varepsilon > 0$, we think exists a set B , expressed as:

$$B = \left\{ X_1, X_2, \dots, X_s \mid \begin{array}{l} X_1 = X, X_s = Y \\ s \in N \\ \rho(X_m, X_{m+1}) < \varepsilon \\ 1 \leq m \leq s+1 \end{array} \right\} \quad (1)$$

When $s \rightarrow \infty$, $\bar{B} \subset M_A$. There $\rho(X_m, X_{m+1}) < \varepsilon$ is the Euclidean distance defined on E^V .

2.2. Propaedeutics of Proving PHC Hypothesis

Proving the PHC hypothesis requires some preliminary knowledge, which will be given in this section. Without loss of generality, we use face recognition as a scene to demonstrate PHC hypothesis. Take person A as an example. The identity attribute of each face image of Person A is A . Therefore, the category label of any face image of person A can be recorded as A .

Photography is influenced by many external conditions, such as time, angle, illumination, etc. When you take a picture of the same person, different conditions produce different images. But the labels of these face images remain the same. Without considering the influence of noise, we take the process of a camera capturing a picture of a human face as our simulation scene and model it as a function. Assuming that the process of capturing a facial image is affected by n conditions, these n conditions are modeled as n dimensions, and the value range of each dimension is expressed as $[a_i, b_i]$, $i = 1, 2, \dots, n$. When person A is photographed, the identity attribute (label) of each picture is A . Each image can be expressed as a function value $f_A(x_1, x_2, \dots, x_n)$, where $(x_1, x_2, \dots, x_n) \in D_A$, Cartesian defines $D_A = (a_1, b_1) \times (a_2, b_2) \times \dots \times (a_n, b_n)$ as the domain function, f_A is called a representational mapping of A . Any point $\mathbf{x}$ in D_A corresponds to a unique image $f_A(\mathbf{x})$ labeled A . $f_A(D_A)$ recorded all the face images of person A .

In order to prove the PHC principle hypothesis, we first need to prove that f_A is continuous on D_A . For this reason, We need to introduce the basic hypothesis(1). **Fig. 1** shows two groups of face pictures of the same person. In group (1), only brightness changes are adjusted and images are taken one by one; group (2) only adjusts the shooting angle and takes images one by one. Both groups of images showed a tendency of gradual change. Take two images, keeping all other shooting conditions unchanged but varying the brightness. Then, the closer the brightness of the shots, the smaller the difference between the matrix norm of the two images. More generally, we assume that when only the dimensional value of the i -th shooting condition changes and the other-dimensional values

Fig. 1. Two groups of picture of the same person. Group (1) only adjusts the brightness change and takes the image one by one. Group (2) only adjusts the shooting Angle and takes images one by one.

are fixed, we regard f_A as a continuous mapping on the i th dimension. $f_A(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ is used to represent the mapping of variable x_i , where $1 \leq i \leq n$ and $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ are fixed, which can be recorded briefly as $f_A(x_i | *)$. We have put forward the basic hypothesis (1):

When $|x'_i - x''_i| \rightarrow 0$, $\|f_A(x'_i | *) - f_A(x''_i | *)\| \rightarrow 0$ where $x'_i, x''_i \in [a_i, b_i]$ and $\|\cdot\|$ denotes the norm of f_A . That is to say, $f_A(x_i | *)$ is continuous on $[a_i, b_i]$ for each i .

Fig. 2. After feature mapping, the input image is encoded as a vector in hidden space.

$f_A(D_A)$ contains all the face images of person A . We also need to discuss that face images be encoded as vectors. As shown in Fig. 2, the face image is encoded into the vector in the hidden space through the backbone network. The backbone network is called feature mapping and denoted as F . In deep learning, the CNN model is optimized via back-propagation. The differentiable function must be continuous. Therefore, We have put forward the basic hypothesis (2): The feature operator F is continuous on the sample set $f_A(D_A)$.

2.3. The Proof of PHC Hypothesis

In this section, we will prove that PHC hypothesis holds. In Theorem 1, we're going to prove that the representational mapping $f_A(x_1, x_2, \dots, x_n)$ of A is continuous on D_A . $F(f_A(D_A))$ is the set of all feature points labeled A , the proof that $F(f_A(D_A))$ is a connected set in E^V provided in Theorem 2. In Theorem 3, based on Theorem 1-2, we prove that PHC hypothesis holds. At this point, we have completed the proof of PHC.

Theorem 1. The representational mapping $f_A(x_1, x_2, \dots, x_n)$ of A is continuous on D_A .

Proof: $\forall (x'_1, x'_2, \dots, x'_n), (x''_1, x''_2, \dots, x''_n) \in D_A$, let

$$S = \|f_A(x'_1, x'_2, \dots, x'_n) - f_A(x''_1, x''_2, \dots, x''_n)\| \quad (2)$$

Adding some items to Eq. (2),

$$\begin{aligned} S &= \|f_A(x'_1, x'_2, \dots, x'_n) - f_A(x'_1, x'_2, x'_3, \dots, x'_n) + \\ &\quad f_A(x'_1, x'_2, x'_3, \dots, x'_n) - f_A(x''_1, x''_2, x'_3, \dots, x'_n) + \dots + \\ &\quad f_A(x''_1, x''_2, \dots, x''_{n-1}, x'_n) - f_A(x''_1, x''_2, \dots, x''_n)\| \end{aligned} \quad (3)$$

Expand Eq. (3),

$$\begin{aligned} S &\leq \|f_A(x'_1, x'_2, \dots, x'_n) - f_A(x'_1, x'_2, x'_3, \dots, x'_n)\| + \\ &\quad \|f_A(x'_1, x'_2, x'_3, \dots, x'_n) - f_A(x''_1, x''_2, x'_3, \dots, x'_n)\| + \dots + \\ &\quad \|f_A(x''_1, x''_2, \dots, x''_{n-1}, x'_n) - f_A(x''_1, x''_2, \dots, x''_n)\| \end{aligned} \quad (4)$$

Simplifying Eq. (4),

$$\begin{aligned} S &\leq \|f_A(x'_1 | *) - f_A(x''_1 | *)\| \\ &\quad + \|f_A(x'_2 | *) - f_A(x''_2 | *)\| \\ &\quad + \dots + \|f_A(x'_n | *) - f_A(x''_n | *)\| \end{aligned} \quad (5)$$

According to the Hypothesis (1), $\forall x'_i, x''_i \in [a_i, b_i]$, $i = 1, 2, \dots, n$, $\forall \varepsilon > 0$, $\exists \sigma_i > 0$, when $|x'_i - x''_i| < \sigma_i$,

$$\|f_A(x'_i | *) - f_A(x''_i | *)\| < \frac{\varepsilon}{n} \quad (6)$$

Thus, $\forall \varepsilon > 0$, $\exists \sigma > 0$,

$\|f_A(x'_1, x'_2, \dots, x'_n) - f_A(x''_1, x''_2, \dots, x''_n)\| < n \cdot \frac{\varepsilon}{n} = \varepsilon$ when $|(x'_1, x'_2, \dots, x'_n) - (x''_1, x''_2, \dots, x''_n)| < \sigma$, which means that f_A is continuous on D_A .

Theorem 2. $F(f_A(D_A))$ is a connected set in E^V and is not a zero measure set.

Proof: From Theorem 1, it can be seen that $f_A(x_1, x_2, \dots, x_n)$ is continuous in D_A , F is continuous, and D_A is a connected set; thus, $F(f_A(D_A))$ is connected in E^V . The measure of D_A is not zero measure set, and the composite mapping $F \cdot f_A$ is continuous in D_A ; hence, the measure of $F(f_A(D_A))$ is not zero measure set.

Theorem 3. The finite sample point set M with label A in E^V satisfies Eq. (1).

Proof: Based on Theorem 2, $F(f_A(D_A))$ is connected in E^V , and it is not a zero measure set. M can be regarded as a set of discrete points embedded in $F(f_A(D_A))$. The measure of feature space denotes $\rho(x, y)$.

According to the properties of connected sets, $\forall x, y \in M$,

$$I_{xy} = \{a_1, a_2, \dots, a_m | a_1 = x, a_m = y, m \rightarrow \infty, \rho(a_i, a_{i+1}) \rightarrow 0\} \quad (7)$$

there must be $I_{xy} \subset F(f_A(D_A))$. which render

$\bar{I}_m \subset F(f_A(D_A))$ true. That is, $\forall x, y \in M$, there are paths connecting x, y in $F(f_A(D_A))$.

Theorem 3 is the equivalent form of Eq. (1), which shows that the characteristic points of the same kind satisfy PHC.

3. Covering Learning Model

3.1. Optimal Coverage

According to PHC, there is a continuous gradient relationship between two homogeneous samples, and the sample points in the gradient process still belong to the same category. The goal of the biomimetic covering

Fig. 3. The triangles represent the sample to be recognized, while the circles and the crosses represent the other two samples. The broken line is the traditional BP network pattern recognition division method. The larger circles represent the division method of radial basis RBF network. The elongated ellipse represents the division of biomimetic covering learning.

learning is to realize continuous coverage of similar samples distributed in the feature space, and the learner can learn a connected geometry to cover the same type of discrete samples.

Fig. 3 shows the mechanism of biomimetic covering learning, which recognizes samples from the perspective of cognition, and it is different from the BP network and radial basis function network (RBF) generated to divide hyperplanes. Definition 1 provides the definition of the optimal coverage of the same class of sample points and presents the proof of its existence in Theorem 4. There are various coverings in the hidden space for the point set M made up of analogous samples. We will define optimal coverage and prove that it exists in M .

Definition 1: Let M be the set of feature points with label A in E^v . If there is a constant S that makes $M \subset \bigcup_i^s S_i$ true, it is said that $\bigcup_i^s S_i$ is the S -coverage of M (S_i is a hypersphere with a radius of r_i centered at point O_i). Among

all the connected S -coverage of M , the coverage, which has the smallest volume and every point in the coverage that is labeled as A , is called the optimal coverage of M .

Theorem 4. In E^v , the optimal coverage exists in the finite sample point set M labeled A .

Proof: Let $M = \{e_1, e_2, \dots, e_k\}$, it is obviously clear that $M \subset F(f_A(D_A))$,

- (1) When $k = 1$, the conclusion is apparent.
- (2) When $k \geq 2$, by pairing any two points e_i, e_j in M , we

can see that there are $\frac{k!}{2!(k-2)!}$ combinations. The set of point pairs is denoted as

$$\hat{M} = \{\hat{e}_i \mid i = 1, 2, \dots, \frac{k!}{2!(k-2)!}\}.$$

According to Theorem 3, $\forall \hat{e}_i \in \hat{M}$, there is a connected path between the point pair $\hat{e}_i$, denoted as $\bar{I}_i$, and $\bar{I}_i \subset F(f_A(D_A))$. Therefore, there are S_{ij} points $O_{i1}, O_{i2}, \dots, O_{iS_{ij}}$ on $\bar{I}_i$, S_{ij} radius $r_{i1}, r_{i2}, \dots, r_{iS_{ij}}$, which make $\bar{I}_i \subset \bigcup_{ij}^{S_{ij}} O_{ij} \subset F(f_A(D_A))$ true, and $\bigcup_{ij}^{S_{ij}} O_{ij}$ is connected in E^v . Where O_{ij} is the kick-off with a radius of r_{ij} centered at point O_{ij} .

Therefore, a union U can be obtained by

$$U = \bigcup_i^{\frac{k!}{2!(k-2)!}} \bigcup_{ij}^{S_{ij}} O_{ij} (j = 1, 2, \dots, S_{ij}) \quad (8)$$

And, $M \subset U \subset F(f_A(D_A))$, therefore, U is the optimal coverage of M .

3.2. HSC Neuron Model

The HSC neuron model is introduced based on optimal coverage, as illustrated in **Fig. 4**. First, the low-level features along with the manifold distribution of the class are extracted by stacked multiple parallel convolution groups. Second, the Gaussian mixture model (GMM) implemented in [22] is used to cluster the data. The manifold distribution of samples from every class are decomposed into several sub-manifold distributions. Consequently, every cluster corresponds to a simple

Fig. 4. The complete workflow of the hyper-sausage covering (HSC) neuron model.

manifold distribution, laying a foundation for the hypersausage optimal coverage. Finally, each cluster is covered by the HSC, thus achieving the cognition. After acquiring the sub-manifold distribution of each cluster through GMM, under the assumption that the sub-manifold distributions are independent of each other, the steps in constructing the HSC of each cluster are as follows:

- 1) calculate sample-wise distances and initialize the endpoints q_1 and q_2 of the HSC by the two most distant samples;
 - 2) initialize the radius γ of the hypersphere by the variance of the distances d from samples to the line segment $\overline{q_1 q_2}$;
 - 3) initialize HSC by the topology product between $\overline{q_1 q_2}$ and the hypersphere;
 - 4) update q_1 , q_2 and γ through back propagation algorithm.
- The output of HSC constructed in this way implies the probability of the sample belonging to the cluster, whose mathematic expression is:

$$\hat{y} = \exp\left(-\frac{d^2}{D \cdot r_{sau}^2}\right), \quad d = \|\mathbf{x} - (\lambda \mathbf{q}_1 + (1 - \lambda) \mathbf{q}_2)\| \quad (9)$$

where γ is the radius of the hypersphere used to construct HSC, D is the dimension of the data which is added to balance sparsity of the high-dimension, d is the distance from samples to the line segment $\overline{q_1 q_2}$, $\|\cdot\|$ represents a distance metric which is the cosine distance in this paper, $\mathbf{x}$ is the sample point, q_1 and q_2 are the endpoints of the HSC, λ represents the location of the nearest point on the line segment $\overline{q_1 q_2}$ from the sample point $\mathbf{x}$, the expression of which is

$$\lambda = \begin{cases} 0, & k < 0 \\ k, & 0 \leq k \leq 1 \\ 1, & k > 1 \end{cases}, \quad k = \frac{(\mathbf{q}_2 - \mathbf{x}) \cdot (\mathbf{q}_2 - \mathbf{q}_1)}{\|\mathbf{q}_2 - \mathbf{q}_1\|^2} \quad (10)$$

The construction of the super-sausage model guided by the covering learning theory proposed in this study is based on a one-dimensional simplex, aiming to construct an abstract geometric form to complete the sample manifold coverage. The hyper-sausage constructed in this study has a geometry similar to that of the super-sausage constructed based on the topological product of one-dimensional manifolds and superspheres. It was constructed to complete the coverage of

sample manifolds, thus achieving the purpose of classification and recognition.

4. Experiments

A simple two-layer CNN and HSC with FaceNet [23] network and ArcFace network [24] were adopted to demonstrate the effectiveness of the proposed neuron model. In this study, the LFW dataset [25] (<http://vis-www.cs.umass.edu/lfw/>) dataset [26] (<https://www.cs.tau.ac.il/~wolf/ytfaces/>) were adopted. LFW dataset is the de-facto academic test set for face verification. The YTF dataset collects 3,425 YouTube videos of 1,595 subjects (a subset of the celebrities in the LFW). These videos were divided into 5,000 video pairs and 10 splits and then used to evaluate the video-level face verification. Therefore, the dataset is suitable to be used for comparing the classification results of shallow neural networks constructed by M-P and HSC. Finally, evaluation on various parameters such as the classification accuracy, number of parameters, and running time is obtained by applying the M-P and HSC neuron models on the CIFAR-10 and LFW datasets. Some images of the three datasets are shown in Fig. 5. All of the experiments were conducted on an NVIDIA GeForce GTX-1080Ti GPU.

4.1. Recognition on the LFW and YTF datasets

In this section, the recognition results of HSC with FaceNet network and ArcFace network on LFW and YTF datasets are discussed to further demonstrate the superiority of our neuron models for face recognition. We evaluated our model using the standard protocol for unrestricted, labeled outside data. Nine training splits were used to select the L2-distance threshold. We replaced the fully connected layer with our HSC. As shown in Table 1, a high face recognition accuracy is obtained with the loss optimization [28], [29], [24] and various augmentations and decorrelations [30]. The FaceNet [23] with HSC achieves a better performance of 99.85% on the LFW dataset and a 97.94% accuracy on the YTF dataset. We also evaluated the performance of ArcFace with HSC and obtained an accuracy of 99.92% on the LFW dataset and 98.32% on the YTF dataset, which is the best result compared to other algorithms in Table 1. The effectiveness and superiority of

HSC were verified by the above two experiments. The results also show that HSC has strong flexibility and can be combined with different neural networks.

4.2. Evaluation of the Model Complexity

In this section, evaluation on various parameters such as the classification accuracy, number of parameters, and running time is obtained by applying the M-P and HSC neuron models on CIFAR-10 and LFW datasets and the results thus obtained are listed in **Table 2**. It can be seen from the table that the best classification performance without data augmentation on the CIFAR-10 dataset is obtained by the networks based on the HSC neuron model, where an increase by a large margin of 10.16% with a few scarifications of time and number of parameters as compared to the M-P neuron model is observed. The accuracy, time, and the number of parameters used when applying the FaceNet network are also listed in **Table 2**. In particular, for the CIFAR-10 and LFW datasets, the accuracies respectively reach 94.53% and 99.85% respectively, and the number of parameters reaches 0.287 and 26.452 million respectively, with a running time of only 0.217s and 13.1 ms per step respectively. All of these numbers indicate that the proposed HSC neuron model sacrifices only a small amount of time and number of parameters to obtain a higher accuracy improvement.

The reason for the better performance of the HSC neuron model is that it uses a self-adaptive optimization process for approximating the neuron function. This is beneficial for generating a discriminative parameter representation, and thus avoids the rapid increase in the model complexity with increasingly complex training data. Accordingly, the networks based on the HSC neuron model possess better generalization ability as compared to other networks in function approximation. In addition, the superposition results of the multiple neuron models increase the number of degrees of freedom of the HSC neuron models, and the non-linear modeling ability of the network is also greatly improved. In conclusion, the proposed methods possess strong adapt ability, and can be used in the neural networks with arbitrary depth, and meanwhile show better performance and utility.

Table 1. Recognition accuracy(%) on LFW and YTF datasets compared to state-of-the-art methods

Method	Recognition-accuracy/%	
	LFW [25]	YTF [26]
DeepFace [27]	97.35	91.40
FaceNet [23]	99.63	95.10
CenterFace [28]	99.28	94.90
SphereFace [29]	99.42	95.00
ArcFace [24]	99.83	98.02
CosFace [30]	99.73	97.60
URFace [31]	99.78	97.92
Ours with FaceNet	99.85	97.94
Ours with ArcFace	99.92	98.32

Table 2. Classification accuracy(%), running time(step/ms) and number of parameters (in millions) on CIFAR-10 and LFW datasets

	Accuracy	Time	Parameters
Method	CIFAR-10/ResNet20 [32]		
	M-P	84.37	0.697
HSC	94.53	0.827	0.287
Method	LFW/FaceNet [23]		
	M-P	99.63	0.336
HSC	99.85	0.399	26.452

5. Conclusion

In this paper, we discuss the PHC and provided reasonable theoretical proof. Based on PHC, we introduce the theoretical idea of optimal coverage. Then, we construct a covering neuron model named HSC and conduct experiments on CIFAR-10, LFW, and YTF datasets. Through comparative experiments with M-P neuron on CIFAR-10 dataset, the generalization ability of HSC is verified. It is indicated that the proposed neuron models exhibit superior performance in face recognition.

It is found that the distribution of the same type of sample points in a high dimensional space is a complex manifold structure. The core of biomimetic covering learning method converts to the study of manifold coverage in high dimensional spaces. To obtain reliable approximations of the complex manifold structure, the method is required to select the geometric coverage unit, determine the coverage order, and conduct continuous coverage. These points will be addressed in a subsequent study.

Acknowledgement

This work was supported by the National Nature Science Foundation of China, Grant NO. 61901436, the Key Research Program of the Chinese Academy of Sciences, Grant NO. XDPB22 and Shenzhen Wave Kingdom Co., Ltd.

References

Fig.5. Example images from the (a) CIFAR-10 dataset, (b) LFW dataset, and (c) YTF dataset.

- [1] J. Somers, "Is ai riding a one-trick pony?" *Technology review*, vol. 120, no. 6, pp. 29–36, 2017.
- [2] Yuille, A. L., & Liu, C. (2021). Deep nets: What have they ever done for vision?. *International Journal of Computer Vision*, 129(3), 781-802.
- [3] Almahairi, A., Ballas, N., Coijmans, T., Zheng, Y., Larochelle, H., & Courville, A. (2016, June). Dynamic capacity networks. In *International Conference on Machine Learning* (pp. 2549-2558).
- [4] Ning, X., Li, W., Tang, B., & He, H. (2018). BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition. *IEEE Transactions on Image Processing*, 27(5), 2575-2586.
- [5] Ning, X., Nan, F., Xu, S., Yu, L., & Zhang, L. (2020). Multi-view frontal face image generation: A survey. *Concurrency and Computation: Practice and Experience*, e6147.
- [6] Ning, X., Li, W., & Liu, W. (2017). A fast single image haze removal method based on human retina property. *IEICE TRANSACTIONS on Information and Systems*, 100(1), 211-214.
- [7] Ning, X., Li, W., & Xu, J. (2018). The principle of homology continuity and geometrical covering learning for pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(12), 1850042.
- [8] Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 2011-2023.
- [9] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).
- [10] Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 510-519).
- [11] Sabour, S., Frosst, N., & Hinton, G. E. (2017, January). Dynamic Routing Between Capsules. In *NIPS*.
- [12] Cheng, X., He, J., He, J., & Xu, H. (2019). Cv-CapsNet: Complex-valued capsule network. *IEEE Access*, 7, 85492-85499.
- [13] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*.
- [14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [15] Seung, H. S. (1998). Learning continuous attractors in recurrent networks. In *Advances in neural information processing systems* (pp. 654-660).
- [16] Seung, H. S., & Lee, D. D. (2000). The manifold ways of perception. *science*, 290(5500), 2268-2269.
- [17] Shou-Jue, W. (2002). Bionic (topological) pattern recognition-a new model of pattern recognition theory and its applications. *Acta Electronica Sinica*, 10.
- [18] Shoujue, W., & Jiangliang, L. (2005). Geometrical learning, descriptive geometry, and biomimetic pattern recognition. *Neurocomputing*, 67, 9-28.
- [19] Zhao, X. R., He, Q., & Shi, Z. Z. (2006, May). Hypersurface classifiers ensemble for high dimensional data sets. In *International Symposium on Neural Networks* (pp. 1299-1304). Springer, Berlin, Heidelberg.
- [20] Kamgar-Parsi, B., Lawson, W., & Kamgar-Parsi, B. (2011). Toward development of a face recognition system for watchlist surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10), 1925-1937.
- [21] Muraina, I. D., Alobaedy, M. M., & Ibrahim, H. H. (2017). A framework for preserving data integrity during mobile device forensic in open source software environment. In *Proceedings of the Free and Open Source Software Conference (FOSSC)*, Muscat, Oman (pp. 22-26).
- [22] Hinton, G. E., Sabour, S., & Frosst, N. (2018, February). Matrix capsules with EM routing. In *International conference on learning representations*.
- [23] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).
- [24] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4690-4699).
- [25] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008, October). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.
- [26] Wolf, L., Hassner, T., & Maoz, I. (2011, June). Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011* (pp. 529-534). IEEE.
- [27] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).
- [28] Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016, October). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499-515).
- [29] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). SpheroFace: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 212-220).
- [30] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., ... & Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5265-5274).
- [31] Shi, Y., Yu, X., Sohn, K., Chandraker, M., & Jain, A. K. (2020). Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6817-6826).
- [32] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- [33] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213-229).
- [34] Zhao, Y., Zeng, Y., & Qiao, G. (2021). Brain-inspired classical conditioning model. *Iscience*, 24(1), 101980.
- [35] Kosiorek, A. R., Sabour, S., Teh, Y. W., Hinton, G. E., & STAFFORD-TOLLEY, M. J. (2019). Stacked capsule autoencoders. *Advances in Neural Information Processing Systems*, 32.

Biographies

Xin Ning received the B.S. degree in software engineering from Xinjiang University, Wulumuqi, China, in 2012, and the Ph.D. degree in electronic circuit and system from the Institute of Semiconductors, Chinese Academy of Sciences, in 2017. He is currently an Associate Professor with the Laboratory of Artificial Neural Networks and High Speed Circuits, Institute of Semiconductors, Chinese Academy of Sciences. He has authored or co-authored more than 40 papers in journals and refereed conferences. His current research interests include pattern recognition, computer vision, and image processing.

Yuebao Wang received a B.S Degree in mathematics and Applied Mathematics from North University of China in 2013, and received M.S degree in operations research and cybernetics Currently in 2018. He worked as an algorithm research fellow in the Cognitive Computing Technology Joint Laboratory of Wave Group. His current research interests include machine learning, function approximation theory and TDA.

Weijuan Tian received her M.S. degrees from School of Electronic Engineering in Xidian University in 2017. Currently, she worked as an algorithm research fellow in the Cognitive Computing Technology Joint Laboratory of Wave Group. Her research interests include machine learning, DNNs, neuron modeling, video and image processing.

Liang Liu Received dual B.S. degrees from Wuhan University and Wuhan University of Technology, and the Ph.D. degree in North China Electric Power University. He worked as an algorithm research fellow in the Cognitive Computing Technology Joint Laboratory of Wave Group. His current research interests include DNNs, computer vision, and image processing.

Weiwei Cai is currently practicing at the Laboratory of Artificial Neural Networks and High Speed Circuits, Institute of Semiconductors, Chinese Academy of Sciences, Prior to that, he worked with IT industry for more than ten years in the roles of an System Architect and the Program Manager. His research interests include machine learning, deep learning, and computer vision.