Explore Visual Concept Formation for Image Classification

Shengzhou Xiong [1] Yihua Tan [1] Guoyou Wang [1]

Abstract

Human beings acquire the ability of image classification through visual concept learning, in which the process of concept formation involves intertwined searches of common properties and concept descriptions. However, in most image classification algorithms using deep convolutional neural network (ConvNet), the representation space is constructed under the premise that concept descriptions are fixed as one-hot codes, which limits the mining of properties and the ability of identifying unseen samples. Inspired by this, we propose a learning strategy of visual concept formation (LSOVCF) based on the ConvNet, in which the two intertwined parts of concept formation, i.e. feature extraction and concept description, are learned together. First, LSOVCF takes sample response in the last layer of ConvNet to induct concept description being assumed as Gaussian distribution, which is part of the training process. Second, the exploration and experience loss is designed for optimization, which adopts experience cache pool to speed up convergence. Experiments show that LSOVCF improves the ability of identifying unseen samples on cifar10, STL10, flower17 and ImageNet based on several backbones, from the classic VGG to the SOTA Ghostnet. The code is available at https://github.com/elvintanhust/LSOVCF.

1. Introduction

Image classification is an important task in computer vision that aims to classify the unseen samples into the correct categories. In the era of deep learning, the general idea to this problem is to assign a unique code to each category, and then learn a mapping between training samples and codes in order that unseen samples also fit this mapping. But for humans, the ability of classification is gained through concept learning, which learns the generalized concept description from sample observations such that a given observation can be identified as a learned concept (Seel, 2012). During concept learning, the process of concept formation involves two intertwined searches: one for the best common properties of instances, and the second for the best concept descriptions based on that properties. To generalize the idea of concept formation to deep learning, the common properties are considered as progressively abstracted deep features while concept descriptions correspond to output response in the last layer. These two parts should be explored together as a whole according to human concept formation, which means we shouldn't optimize the feature extraction parameters by presetting output response of training sample. However, in literature almost all the algorithms of image classification based on deep learning violate the principle: each training sample is given a description code, such as one-hot code. Inspired by this fact, we try to explore the visual concept formation by optimizing both parts in training process of image classification based on deep learning.

Concept learning has been an important research area in machine learning, but rarely scholars consider both two part of concept formation. The research goal of the most works is to construct the representation space by finding the best common properties of instances. For example, a model that combines reinforcement learning and clustering assumption is proposed to extract better deep features for concepts separation (Shi et al., 2019). Recently, the other research direction is to build cognitive structures of concepts by lots of labeled samples, and then new concepts can be obtained with few samples based on the existing cognitive structures. The typical example of this type of research is few-shot learning (Snell et al., 2017). In this paper, we focus on the learning of concept description motivated by visual concept formation, which is carried out in image classification based on deep convolutional neural network (ConvNet).

Existing image classification algorithms based on ConvNet only concern about the properties of instances while the learning of concept descriptions is ignored. Even though many effective preprocessing methods, network architectures and loss functions have been proposed (DeVries &

[1] National Key Laboratory of Science & Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074 , China. Correspondence to: Yihua Tan <yhtan@hust.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Taylor, 2017; Li et al., 2019; Cao et al., 2019), one-hot codes are directly assigned to categories as the concept descriptions (He et al., 2016; Hu et al., 2020; Han et al., 2020b). One-hot codes strictly correspond to the points on the coordinate axes in the output space of ConvNet so that relationship between different categories are exactly equal. Another kind of method that takes fixed hard-code as concept descriptions during property exploration is label distribution learning (LDL) (Geng, 2016), in which descriptions correspond to non-axes points which has taken relationship between categories into account by human experience. However, label distribution in LDL is actually obtained by some predefined rules, which is separated from the feature extraction process without learning. Both one-hot code based algorithms and LDL take concept descriptions as known knowledge during learning while ignoring that common properties and concept descriptions are explored as a whole in human's concept formation. In fact, there are some internal relations between concepts while they are separated from each other. For example, bicycle and car are two visual concepts in the image classification, but they have the same property wheels. These relations imply different losses when samples are misclassified into different categories. However, the relation between categories is also fixed when the concept descriptions are fixed. For example, one-hot coding makes the sample misclassified into any category contribute the same loss, which will limit the mining of some important properties and lead to the decline of the ability to identify new samples. Even though LDL considers the different loss of misclassification in such case, label distribution is only designated by experience in some special applications. Concept formation is a process of interwoven exploration of both two parts, in which the concept description should dynamically adjusted with common properties.

In this paper, we explore visual concept formation based on ConvNet, in which common properties (feature extraction) and concept descriptions will be optimized as a whole. Therefore, the first question to be addressed is what the concept description should look like. Since concept description corresponds to output response in the last layer of the ConvNet which is the induction of the attributes of the samples belonging to the same concept, the output response should follow the clustering assumption. It means the responses of the congeneric samples in response space should gather in the same area while the heterogeneous samples gather in the different areas. For simplicity, we assume the response follows Gaussian distribution with a fixed variance. The second question we have to face is how we drive the visual concept formation. During training, the responses of batch samples can be obtained through forward propagation of ConvNet, which is indicated as the temporary concept descriptions (formally as exploration response). With exploration response in each iteration, the network parameters can be optimized by minimizing the difference between the sample responses and the exploration response. By repeating these steps over and over again, the two intertwined searches of visual concept formation will be performed continuously. Moreover, the dynamic change of exploration response will bring convergence difficulty. To overcome this problem, the using of experience is adopted which is inspired by the fact that exploration and experience coexist in human's learning.

According to above analysis, a learning strategy of visual concept formation (LSOVCF) is proposed in this paper. First, the exploration and experience loss (EE-Loss) based on temporary concept descriptions following clustering assumption is designed to drive the visual concept formation, which refers to the weighted summation of exploration loss and experience loss. Specifically, exploration loss of a sample means the KL divergence between its sample response and the corresponding exploration response while experience loss considers experience response from the past experiences. To calculate experience loss, the experience cache pool (ECP) is adopted to save experiences. Second, stochastic gradient descent is adopted here for optimization. Finally,the effectiveness of experience has been demonstrated in ablation experiments. Furthermore, the improvement of identifying unseen samples of the overall LSOVCF has been proved by several backbones and datasets in the experiment section.

Contributions: To handle the image classification task, we try to explore visual concept formation by learning deep features and output response together, considering humans' concept learning process. Our contributions can be summarized as follows:

- We propose LSOVCF for image classification that learn both deep features and output response together, which is inspired by the human concept learning process
- We propose to use ECP to store experience, and then a EE-Loss is designed for rapid and stable formation of visual concept.
- We demonstrate the effectiveness of our LSOVCF through experiments based on several backbones and datasets.
- The experimental results and the limitations of proposed LSOVCF are discussed.

Section Arrangement: The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 details the methodology of proposed LSOVCF. Section 4 shows the experiments and results. Section 5 discusses the experiment results and limitations of proposed LSOVCF. Section 6 makes a conclusion.

2. Related Work

Deep Classification: Image classification are basically realized based on ConvNet since the rise of deep learning in the past decade. The existing researches of ConvNet based image classification can be basically divided into three parts. The first one mainly improves the model's generalization ability by increasing sample diversity (DeVries & Taylor, 2017; Cubuk et al., 2019). The second part is the most researched one, many impressive network architectures have been proposed, including the design of basic modules (Jeon & Kim, 2017; Kobayashi, 2019), connection between layers (He et al., 2016; Huang et al., 2017), reorganization of feature maps (Hu et al., 2020; Han et al., 2020b), attention mechanism (Wang et al., 2017; Kim et al., 2020; Zoran et al., 2020), neural structure search (Zoph et al., 2018; Li et al., 2020; Howard et al., 2019), etc. And the last one focuses on the design of loss function, mainly about the data imbalance (Cao et al., 2019; Lin et al., 2020). However, such algorithms mainly focus on exploring better deep features with fixed concept descriptions. It is common to directly assign one-hot code to each category, which restricts that relationship between categories must be same. Of course, there are also some studies named LDL taking the connections between categories into account. But there is still a fixed code assigned to each sample during learning which is generated by some predefined rules. In addition, the LDL is only appropriate for some special classification tasks, such as emotion classification (Yang et al., 2017), age estimation (He et al., 2017), skin disease severity grading (Wu et al., 2019), etc.

Deep Clustering: Clustering is one of the basic tasks of unsupervised learning and the proposed strategy adopts the similar assumption as clustering. The response of congeneric samples in output space should gather in same area while heterogeneous samples gather in different areas. In age of deep learning, various traditional clustering algorithms are combined with deep features while features are learned alone or optimized together with clustering. For deep features extraction, auto encoder, variational auto encoder and other variants are most common choice, for example, they are combined with subspace clustering (Dang et al., 2020), K-means (Xie et al., 2016), hierarchically clustering (Shin et al., 2020), etc. In addition, ConvNet is also a popular choice for representation learning (Hsu & Lin, 2018; Zeng et al., 2020; Zhan et al., 2020). Although having to abide by clustering assumption, the goal of this paper is to classify unseen samples by exploring visual concept formation with supervised learning while clustering algorithms aim to find a reasonable division of unlabeled samples.

Concept Learning: Concept learning has been an important research area in machine learning. The formation of a concept involves two intertwined searches in human concept learning, that is, feature expression and concept description. However, almost all literatures concern about better feature expression only rather than both two parts, such as the algorithms that try to learn visual concepts with hierarchy (Jia et al., 2013; Divvala et al., 2014) and algorithm that combines reinforcement learning and clustering assumption for normal image classification task (Shi et al., 2019). Another direction of concept learning aims to acquire new concepts based on the relevant concepts originally existing in the cognitive structure. The main strategy is to gather statistics of labeled samples for cognitive structure construction, and then to take this as prior information to handle new concepts. In addition, the cognitive structure is usually expressed as Bayesian probability (Lake et al., 2015) or neural network (Han et al., 2020a; Yang et al., 2020). In fact, such research of concept learning has several terms with different task descriptions which are well known, including one-shot learning (Xue & Wang, 2020), few-shot learning (Snell et al., 2017), new categories discovering (Han et al., 2020a), etc. In this paper, we try to explore visual concept formation by learning both two intertwined parts, and to acquire new concepts based on formed concepts will be the future work.

3. Methodology

3.1. Preliminaries

Existing ConvNet based image classification algorithms only concern about deep features of instances while the learning of concept descriptions is ignored. One-hot code is usually directly adopted as figure 1. For C categories, the concept description of each category is fixed to a C-bit binary code. As for the ConvNet, the last layer is a linear layer with C neurons which takes softmax as activation function. The optimization goal of such algorithms is that samples of same category will activate one specified neuron. So the output response of each category is limited to a point on the coordinate axis, making the relationship between categories completely equal. Noting that neurons of the last layer is unordered, so it can be considered as a C-dimension space. Further, although there are some advanced loss functions according to data distribution (Cao et al., 2019; Lin et al., 2020), cross entropy loss (CE-Loss)is the most commonly

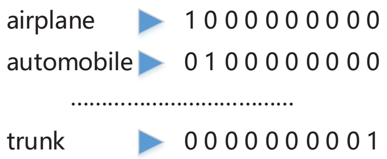
airplane ▶ 1 0 0 0 0 0 0 0 0 0
automobile ▶ 0 1 0 0 0 0 0 0 0 0
....
trunk ▶ 0 0 0 0 0 0 0 0 0 1

Figure 1. The one-hot codes of cifar10 dataset.

used here. And in section 4, CE-Loss and One-hot code are selected for comparison. The CE-Loss is calculated as follows:

$$L_{CE} = -\sum_{c=1}^{C} p_c \log q_c \tag{1}$$

in which p means one-hot code of the sample and q is the output response.

Although keeping the same form with existing ConvNet, we consider the last linear layer as one-dimension space with size *K* in our proposed learning strategy, that is, the neurons are in order. The location of target output response is not fixed until concepts formed and only the clustering assumption is required. We assume the total response intensity of output as 1 which is achieved through the softmax activation function. Since the total response intensity is limited to 1, the response intensity in each position can be viewed as probability that response center locates in here, then expectation of response position can be calculated. In the proposed LSOVCF the expectation of response position in output space is viewed as the response center, which is calculated as (2).

$$E(y) = \sum_{k=1}^{K} y_k k, \quad \sum_{k=1}^{K} y_k = 1 \tag{2}$$

in which y is the output response. Thus, the target response of LSOVCF is determined by such expectation of congeneric samples.

3.2. Learning Strategy of Concept Formation

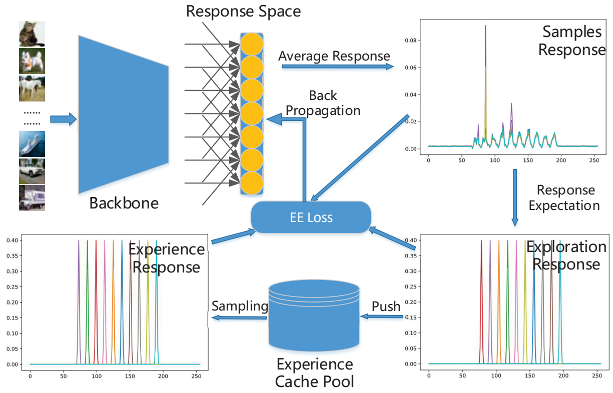

Figure 2. The proposed learning strategy of concept formation. The backbone here can be any ConvNet, and response space is the last linear layer of the selected ConvNet.

The proposed strategy aims to explore the formulation of concepts which mainly focuses on the output response, so the backbone can be any ConvNet. The main idea of LSOVCF is shown in figure 2. Besides ConvNet, another key component ECP is proposed to store recent exploration results. As mentioned earlier, the last layer of the network can be considered as one-dimension space and samples response can be obtained through forward propagation in each iteration during training. Then we can get the response center $RC(C_i)$ of each category C_i as follows:

$$RC(C_i) = \tfrac{1}{N}\textstyle\sum_{j=1}^{N} E(y^j), y^j \in C_i \tag{3}$$

where y^j is the output response of single sample which belongs to C_i and there are N samples for C_i in such batch. After all response centers are obtained, they are shifted by a certain distance when response centers of different categories are too close, according to the clustering assumption. Furthermore, for simplify, the target response of category C_i is assumed to follow a Gaussian distribution that takes $RC(C_i)$ as the mean value. In addition, the variance of the Gaussian distribution is set to a constant. The target responses of categories obtained from each batch images are called exploration response. And such exploration response will be pushed into the ECP as an experience sample. The capacity of ECP depends on the size of training dataset, and we need to fill the pool before network parameters updating. At each iteration of training, exploration response can be obtained based on sample response and experience response can be calculated through randomly sample from ECP. Then the EE-Loss can be calculated and both common properties and concept descriptions will be optimized by stochastic gradient descent during the visual concept formation. The detail of ECP and EE-Loss will be presented later.

As for classification of the unseen sample, the average response P^i of each category C_i in the training set needs to be calculated first, then KL divergence between P^i and the output response of unseen sample will be calculated, the unseen sample will be finally assigned into the category with minimum KL divergence. P^i and KL divergence are calculated as follows:

$$P^i = \frac{1}{N}\sum_{j=1}^{N} y^j, y^j \in C_i \tag{4}$$

$$KL(P^i, y) = \sum_{k=1}^{K} P^i_k \log(P^i_k/(y_k + \epsilon)), \epsilon = 1e-8 \tag{5}$$

in which, y^j is output response of single sample from category C_i and the sample count of C_i is N.

3.3. Experience Cache Pool

In each iteration of the training process, the samples of each category need to be fitted into a target response. But the target response of proposed LSOVCF is calculated based on samples response of each batch images rather than fixed codes. So the target response of each iteration is different before the concepts formed. Due to the uncertainty of fitting target, it is difficult for the model to converge stably and quickly, despite the tendency of concept formation. Therefore, the ECP is proposed to cope with this challenge, which is inspired by the fact that exploration and experience coexist in human's learning. In each iteration, exploration response can be calculated, and we take it as the temporary target response so that we can obtain the prediction accuracy of current batch samples. Then the accuracy will be saved when the exploration response is pushed into the ECP. In addition, the earliest sample in the ECP will be popped if the pool is full when new sample is pushed. During training, a batch of samples will be randomly selected from the ECP in each iteration and the weighted average of them will be calculated as experience response with the accuracy as weight.

Algorithm 1 Learning Strategy of Concept Formation

```
Initialization: fill the ECP.
for $i = 1$ to $epoches$ do
    for $j = 1$ to $iterations$ do
        Get output response $y$;
        Calculate exploration response $P_{epl}$ and take it as target response to calculate accuracy of this batch;
        Pop the earliest sample in ECP;
        Push exploration response and accuracy into ECP;
        Random sampling from ECP and calculate experience response $P_{epr}$;
        Calculate EE-Loss;
        Optimized by stochastic gradient descent;
    end for
end for
```

3.4. Exploration and Experience Loss

We expect that sample's output response satisfies the clustering assumption, so the target response of each iteration can be calculated and loss function is designed based on it. According to the statement above, we can obtain the exploration response and the experience response in every iteration. Considering the design motivation for the ECP, the target response of each sample contains both exploration response and experience response. And such two parts of target response can be obtained respectively based on the sample's label, denoted as P_{epl} and P_{epr}. Then the designed EE-Loss is defined as follows:

$$L_{EE} = \alpha KL(P_{epl}, y) + (1-\alpha) KL(P_{epr}, y) \tag{6}$$

in which α means exploration ratio (ER) and aims to tradeoff between exploration and experience, which is related to the learning ability of selected ConvNet. Furthermore, the EE-Loss is differentiable, so stochastic gradient descent can be used for optimization. The overall process of the proposed LSOVCF is shown in algorithm 1.

4. Experiments

Datasets: Experiments are conducted on four datasets, including cifar10 (Krizhevsky, 2012), STL10 (Coates et al., 2011), flower17 (Nilsback & Zisserman, 2006) and a subset of ImageNet (Russakovsky et al., 2015). Cifar10 dataset consists of 60000 images in 10 classes, with 6000 images per class and 5000 of them are for training. And in our experiments, we randomly select 500 images from each class's training set for validation. STL10 dataset have 1300 labeled images per class, for each class, the training set, test set and validation set are divided with a ratio of 750:500:50. Flower17 dataset consists of 1360 labeled images that belong to 17 kinds of flowers, we take 70 images of each class as training set and 10 image as test set, there is no validation set in flower17 dataset. As for Imagenet, 100 categories are randomly selected for experiments because the time consumption on the complete dataset is too high. And 100 samples of each category are randomly selected as test set.

Approaches: There are seven backbone ConvNets being selected for experiments, including VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), SENet (Hu et al., 2020), MobileNet v3 (Howard et al., 2019), ShuffleNet v2 (Ma et al., 2018), EfficientNet (Tan & Le, 2019) and GhostNet (Han et al., 2020b). VGG is one of the most classical ConvNet for image classification while GhostNet is one of the SOTA which is proposed last year. ResNet and SENet are famous for the widely used residual structure and "Squeeze-and-Excitation" block. As for MobileNet v3, ShuffleNet v2 and EfficientNet, such models based on different ideas are representatives of models that aspire to speed and lightness, note that MobileNet v3 is one of the best examples of neural architecture search techniques. Except ShuffleNet v2 and GhostNet, there are several architectures for each backbone to choose. For VGG, the 11-layer model with batch normalization is adopted here. And 18-layer ResNet is selected while SENet is based on it. In addition, mobileNetV3-Small and efficientnet-b0 is selected here.

Settings: According to the count of categories, the size of last linear layer is set to 256 on the cifar10, STL10 and flower17, and it is set to 1800 on the subset of ImageNet. The response centers of different categories should be shifted to a certain distance $D = 13$ if they are too close, because we assume that the response follows a Gaussian distribution with a variance of 1. $D = 13$ is reasonable considering the effective length of response region, and the

Table 1. Train ResNet on cifar10 with/out ECP.

ECP	1	2	3	4	5	MEAN	STD
×	93.45	92.76	89.07	93.56	93.79	92.53	1.76
√	93.65	93.88	93.63	93.78	93.71	93.73	0.09

target response intensity lists as:

$$\begin{gathered}[0.000134, 0.00443, 0.0540, \\ 0.242, 0.399, 0.242, \\ 0.0540, 0.00443, 0.000134]\end{gathered} \tag{7}$$

The capacity and sampling batch size of ECP are set to (1500, 500) for cifar10, (600, 200) for STL10, (300, 100) for flower17 and (1300, 500) for ImageNet. And the training batch size of cifar10, STL10, flower17 and ImageNet are set to 100, 100, 170, 500. The proposed LSOVCF will be compared with the selected backbones which adopt the one-hot code and cross-entropy. All models will be optimized by stochastic gradient descent and the initial learning rate is set to 0.1, except that learning rate of VGG is set to 0.01 for convergence. In addition, all models will be trained 200 epochs with the same data preprocessing, and the learning rate will decay by a factor of 10 in 100th and 150th epoch. Finally, all experiments are based on the PyTorch and performed on RTX 2080 Ti GPU.

4.1. Ablation Experiments

With/out ECP: Take ResNet and cifar10 as example, we have demonstrated the effect of the ECP. In figure 3, the model can converge quickly and stably with the ECP according to sub-figure (c) and (d). Although the model without ECP achieves good results at the end of the training in sub-figure (b), there is a lucky element because it keeps stable for only a small segment. Furthermore, with and without the ECP, the ResNet has been trained five times on the cifar10 dataset and the accuracy has been calculated on the test set, the results are shown in table 1. According to the result, higher and more stable accuracy can be obtained via ECP.

Exploration Ratio: Different neural architectures leads to different learning abilities just like that different person behaves differently in learning, so the most appropriate ER values are different for selected backbones. To find the most appropriate values, take ER range from 0.1 to 0.9 with step of 0.1, the models based on each backbone have been trained 5 times with each ER value, and flower17 is selected for this experiment to save time. The result of ResNet is shown in table 2, and the most appropriate ER is 0.3. Although it gets highest accuracy when ER is set to 0.7, but the standard deviation of 0.7 is nearly three times as that of 0.3. Furthermore, the most appropriate ER values of all backbones are shown in table 3, and them will be used in the following experiments.

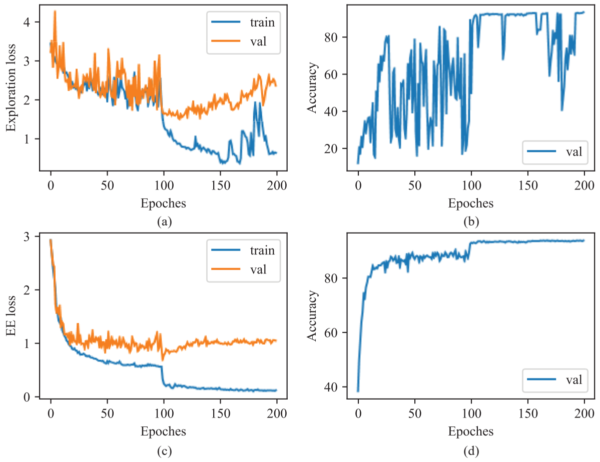

Figure 3. Training process of model. The ER is set to 1.0 in (a) and (b) while 0.3 in (c) and (d), $ER = 1.0$ means that ECP doesn't work. (a) and (c) show the loss curve of training set and validation set, (b) and (d) show the accuracy curve of validation set.

Table 2. Find most appropriate ER of ResNet on flower17.

ER	1	2	3	4	5	MEAN	STD
0.1	75.29	77.06	75.29	75.88	73.53	75.41	1.14
0.2	73.53	78.82	72.94	79.41	73.53	75.65	2.85
0.3	75.88	76.47	75.88	77.06	77.06	76.47	0.53
0.4	77.06	74.12	75.88	78.24	74.71	76.00	1.51
0.5	74.71	77.06	74.12	77.06	74.12	75.41	1.36
0.6	76.47	75.29	72.94	74.71	73.53	74.59	1.26
0.7	77.65	74.12	78.24	77.65	75.88	76.71	1.52
0.8	74.71	78.24	70.00	78.24	65.29	73.30	5.02
0.9	70.59	67.65	68.24	73.53	69.41	69.89	2.08

4.2. Experiments with Different Backbones

The models based on all selected backbones will be trained on cifar10, STL10 and flower17, and the proposed LSOVCF will be compared with the strategy that based on one-hot code and CE-Loss. As for the subset of ImageNet, experiments are carried out based on ResNet, ShuffleNet v2, EfficientNet and GhostNet. Each experiment has been performed 5 times and the mean and standard deviation of accuracy on the test set have been calculated, results are shown in table 4, table 5, table 6 and table 7.

On the cifar10 dataset, the accuracy of all backbones in the proposed strategy is improved compared with the previous algorithms, especially ResNet and ShuffleNet v2. In addition, the LSOVCF is more stable in most cases. As for STL10 dataset, we can get similar conclusion, except that the accuracy of VGG network is lower. Flower17 is a dataset of fine-grained recognition which is more difficult to learn concepts, and the lack of labeled samples further enhances

Table 3. Selected ER values of backbones.

backbone	VGG	ResNet	SENet
ER	0.1	0.3	0.7
backbone	MobileNet v3	ShuffleNet v2	
ER	0.2	0.2	
backbone	EfficientNet	GhostNet	
ER	0.3	0.2	

Table 4. Experiments with different backbones on cifar10.

backbone	CE-Loss	EE-Loss
VGG	92.85 ± 0.19	92.97 ± 0.10
ResNet	93.23 ± 0.22	93.73 ± 0.09
SENet	93.51 ± 0.15	93.66 ± 0.18
MobileNet v3	90.00 ± 0.23	90.07 ± 0.13
ShuffleNet v2	90.31 ± 0.35	90.78 ± 0.20
EfficientNet	93.35 ± 0.08	93.62 ± 0.12
GhostNet	91.63 ± 0.25	91.97 ± 0.17

Table 5. Experiments with different backbones on STL10.

backbone	CE-Loss	EE-Loss
VGG	84.39 ± 0.48	84.12 ± 0.31
ResNet	78.74 ± 0.80	80.92 ± 0.29
SENet	79.06 ± 0.35	80.51 ± 0.23
MobileNet v3	75.88 ± 0.49	76.63 ± 0.34
ShuffleNet v2	75.12 ± 0.51	75.73 ± 0.54
EfficientNet	81.58 ± 0.77	84.10 ± 0.31
GhostNet	82.12 ± 0.25	82.43 ± 0.25

Table 6. Experiments with different backbones on flower17.

backbone	CE-Loss	EE-Loss
VGG	75.41 ± 0.94	74.70 ± 1.29
ResNet	70.71 ± 1.87	76.47 ± 0.53
SENet	72.24 ± 1.14	76.35 ± 1.72
MobileNet v3	73.65 ± 2.12	74.94 ± 2.09
ShuffleNet v2	68.24 ± 1.66	72.47 ± 2.12
EfficientNet	74.94 ± 3.04	76.71 ± 1.42
GhostNet	75.18 ± 1.36	74.82 ± 1.55

the difficulty. Nevertheless, the proposed LSOVCF performs better on all backbones except VGG and GhostNet. ImageNet is the most commonly used large-scale image recognition dataset. The complex background of the images will interfere with our LSOVCF which aims to learn single concept of each sample. In spite of this, the proposed LSOVCF performs better on ResNet, EfficientNet and GhostNet, in particular, it improves accuracy by 1.89 percentage on ResNet. In conclusion, the experimental results demonstrate the effectiveness of LSOVCF and indicate that concept formation is more conducive to the identification of unseen samples than the artificial hard codes.

5. Discussion

The above experiments have proved the classification ability of LSOVCF, and there have been some improvements on almost all backbones and datasets. Furthermore, exhilaratingly, an interesting phenomenon emerged in the results, which is shown in figure 4. Conceptually similar categories also have close response locations in the concept space while we don't impose any constraints in such aspect, which is spontaneous in the process of concept formation. In cifar10 dataset, airplane, automobile, ship and truck belong to a higher hierarchy concept of "transportation" while other categories belong to "animal". The results of dozens of experiments show that there is no overlap between "transportation" and "animals" in output space, although the spatial order of lower hierarchy concepts is not fixed. In addition, cat and dog are almost always adjacent in concept space, so does the automobile and truck, airplane and ship. On the other hand, the confusing samples are more likely to be misclassified into similar concepts by humans, so does the LSOVCF. The category that most misclassified into in cifar10 is shown in table 8. For each category, the ”most err” in table means the predicted category that appears most frequently in the misclassified samples while ”ratio” means its percentage in all the misclassified samples. Considering cat and dog which have the minimum accuracy and adjacent output response, about half of cat's misclassified samples are predicted as dog, so does dog. In summary, the LSOVCF is intended to simulate the process of human concept formation and the results also show similar phenomena with human cognition, which means that our strategy successfully achieves the expected results.

Limitations and Future Work: Although the effectiveness of LSOVCF has been demonstrated through lots experiments and surprising phenomenon emerged in results, there are still many limitations for it to overcome. The limitations are summarized as follows:

- The size of output space is set to 256, which is too small to hold many concepts. And one dimension is not enough too, considering the 1000 categories of ImageNet, at least 10000 neurons is needed in the last linear layer of ConvNet.
- We use KL divergence to measure the difference of response in output space, but the response should be sparse because it is confined to a small area. For such sparse responses, KL divergence is not appropriate very well. Considering three responses without overlap, KL

Table 7. The average accuracy of 100 categories from ImageNet.

BACKBONE	CE-LOSS	EE-LOSS
RESNET	79.62 ± 0.48	81.51 ± 0.52
SHUFFLENET V2	76.66 ± 0.34	75.53 ± 0.29
EFFICIENTNET	83.34 ± 0.29	83.76 ± 0.24
GHOSTNET	79.57 ± 0.27	79.83 ± 0.29

*There are only partial results of supplementary experiment because of the lack of time. The input image size is set to 128*128 and batch_size is set to 512 considering the graphic memory of GPU. The results of ResNet, SeNet, EfficientNet and GhostNet show that the proposed LSOVCF is also work on the large-scale dataset.

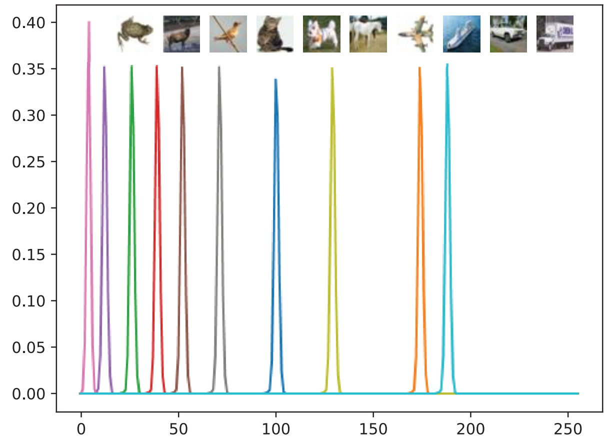

Figure 4. The average output response of cifar10's training set, and backbone is ResNet. The image examples are the same order as the category response centers.

divergence cannot distinguish the relationship between them.

- The output response is constrained to unimodal Gaussian distribution, but some different categories may have common properties in the real world, so multimodal distribution with main peak may be a more appropriate response form.

- The use of ECP here is just a rudimentary form, and it can be improved by referencing to the researches in reinforcement learning, such as ECP with priority.

- We directly calculate the KL divergence based on the statistical results of training set and predict unseen sample as category with minimum KL divergence. There are actually many methods that can be used to predict unseen samples which are not analyzed in this paper, for example, using data in ECP instead of statistical results of the training set.

Table 8. Misclassification statistics. The backbone is ResNet and dataset is cifar10.

CATEGORY	ACCURACY	MOST ERR	RATIO
AIRPLANE	95.7	SHIP	28
AUTOMOBILE	97.3	TRUCK	67
BIRD	91.3	DEER	23
CAT	86.2	DOG	48
DEER	94.8	CAT	25
DOG	87.9	CAT	61
FROG	95.3	BIRD	38
HORSE	95.1	DEER	31
SHIP	96.6	AIRPLANE	41
TRUCK	95.3	AUTOMOBILE	53

In view of these limitations, future work will be carried out from the following aspects:

- To grasp more concepts, the concept space should be extended to higher dimensions, such as 2D, 3D, etc.

- Find an effective metric to measure the difference between sparse distributions, take 2D, 3D, etc. into account.

- Look for a more reasonable response form, such as multimodal distribution.

- Explore more about how to use experience, such as experience form and sampling method.

- Find a more reasonable method to classify unseen samples using learned concepts.

Although the proposed strategy has many limitations, it is still an important direction for concept learning. And we believe machines will actually grasp concepts someday with the breakthrough of these limitations and further research.

6. Conclusion

We have proposed LSOVCF for image classification task. Considering the process of humans' visual concept formation involves intertwined searches of common properties and concept descriptions, the proposed strategy aims to learn both two parts together while previous algorithms directly adopts fixed codes as concept descriptions. In the LSOVCF, ECP is proposed for stable convergence of the model, and then the EE-Loss is designed based on the ECP and clustering assumption. The experiment results have demonstrated the effectiveness of proposed LSOVCF. And the interesting clustering phenomenon of similar concepts shows the prospect of this work that machines have the possibility to grasp concepts rather than just memorize samples. Although there are still many limitations, the proposed strategy is an important direction of concept learning. In addition, our future research will focus on breaking these limitations.

Acknowledgements

Thanks for the comments from reviewers. This research was funded by Shenzhen Basic Research Program, grant number JCYJ20180305180708546; National Natural Science Foundation of China, grant number 41971284.

References

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 215–223, 2011.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Dang, Z., Deng, C., Yang, X., and Huang, H. Multi-scale fusion subspace clustering using similarity constraint. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6657–6666, 2020.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout, 2017.

Divvala, S. K., Farhadi, A., and Guestrin, C. Learning everything about anything: Webly-supervised visual concept learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3270–3277, 2014.

Geng, X. Label distribution learning. *IEEE Transactions on Knowledge & Data Engineering*, 28(07):1734–1748, 2016.

Han, K., Rebuffi, S. A., Ehrhardt, S., Vedaldi, A., and Zisserman, A. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020a.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. Ghostnet: More features from cheap operations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1577–1586, 2020b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, Z., Li, X., Zhang, Z., Wu, F., Geng, X., Zhang, Y., Yang, M., and Zhuang, Y. Focal loss for dense object detection. *IEEE Transactions on Image Processing*, 26 (8):3846–3858, 2017.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., and Le, Q. Searching for mobilenetv3. In *IEEE International Conference on Computer Vision*, pp. 1314–1324, 2019.

Hsu, C. and Lin, C. Cnn-based joint clustering and representation learning with feature drift compensation for large-scale image data. *IEEE Transactions on Multimedia*, 20(2):421–429, 2018.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(08):2011–2023, 2020.

Huang, G., Liu, Z., Maaten, L. V. D., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, 2017.

Jeon, Y. and Kim, J. Active convolution: Learning the shape of convolution for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1846–1854, 2017.

Jia, Y., Abbott, J., Austerweil, J., Griffiths, T., and Darrell, T. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *International Conference on Neural Information Processing Systems*, pp. 1842—-1850, 2013.

Kim, I., Baek, W., and Kim, S. Spatially attentive output layer for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9530–9539, 2020.

Kobayashi, T. Global feature guided local pooling. In *IEEE International Conference on Computer Vision*, pp. 3364–3373, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. *University of Toronto*, 2012.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332—1338, 2015.

Li, G., Qian, G., Delgadillo, I. C., Muller, M., Thabet, A., and Ghanem, B. Sgas: Sequential greedy architecture search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1617–1627, 2020.

Li, X., Wang, W., Hu, X., and Yang, J. Selective kernel networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.

Lin, T., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(02):318–327, 2020.

Ma, N., Zhang, X., Zheng, H. T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *European Conference on Computer Vision*, pp. 122–138, 2018.

Nilsback, M. and Zisserman, A. A visual vocabulary for flower classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1447–1454, 2006.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Li, F. F. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Seel, N. M. (ed.). *Encyclopedia of the Sciences of Learning*. Springer, Boston, MA, 1st edition, 2012.

Shi, J., Xu, J., Yao, Y., and Xu, B. Concept learning through deep reinforcement learning with memory-augmented neural networks. *Neural Networks*, 110:47–54, 2019.

Shin, S. J., Song, K., and Moon, I. C. Hierarchically clustered representation learning. In *AAAI Conference on Artificial Intelligence*, pp. 5776–5783, 2020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *International Conference on Neural Information Processing Systems*, pp. 4080—4090, 2017.

Tan, M. and Le, Q. Deep residual learning for image recognition. In *International Conference on Machine Learning*, 2019.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6458, 2017.

Wu, X., Wen, N., Liang, J., Lai, Y., She, D., Cheng, M., and Yang, J. Joint acne image grading and counting via label distribution learning. In *IEEE International Conference on Computer Vision*, pp. 10641–10650, 2019.

Xie, J., Girshick, R., and Farhadi, A. Unsupervised deep embedding for clustering analysis. In Balcan, M. F. and Weinberger, K. Q. (eds.), *International Conference on Machine Learning*, pp. 478–487, 2016.

Xue, W. and Wang, W. One-shot image classification by learning to restore prototypes. In *AAAI Conference on Artificial Intelligence*, pp. 6558–6565, 2020.

Yang, J., She, D., and Sun, M. Joint image emotion classification and distribution learning via deep convolutional neural network. In *International Joint Conference on Artificial Intelligence*, pp. 3266–3272, 2017.

Yang, M., Deng, C., Yan, J., Liu, X., and Tao, D. Learning unseen concepts via hierarchical decomposition and composition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10245–10253, 2020.

Zeng, K., Ning, M., Wang, Y., and Guo, Y. Hierarchical clustering with hard-batch triplet loss for person reidentification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13654–13662, 2020.

Zhan, X., Xie, J., Liu, Z., Ong, Y., and Loy, C. Online deep clustering for unsupervised representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6687–6696, 2020.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.

Zoran, D., Chrzanowski, M., Huang, P., Gowal, S., Mott, A., and Kohli, P. Towards robust image classification using sequential attention models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9480–9489, 2020.