Journal of Computers Vol. 32 No. 2, 2021, pp. 47-60
doi:10.3966/199115992021043202005

Learning to Different Attribute Examples with Deep Transfer for Object Categorization

Szu-Yun Pai, Yu-Cheng Liu, Ta-Hsiang Hu*

Department of Electrical Engineering, Da-Yeh University, Changhua 51591, Taiwan, ROC
{d0403002, f0603005}@cloud.dyu.edu.tw; thhu@mail.dyu.edu.tw

Received 21 January 2020; Revised 5 June 2020; Accepted 17 August 2020

Abstract. This paper proposes a small sample dataset, regarded as the specific-task dataset in deep transfer learning, in order to improve the performance of transfer learning. Each single-frame image is divided into three top-down sub-images in the dataset. Object features, such as sharp and texture information, are enhanced to capture the features of each class in the target domain, to reduce the loss of the network function caused by one-way transfer. Therefore, the network can efficiently learn more accurate information, and help to reduce softmax cross-entropy loss and generalization error. In addition, we explore the knowledge transfer among different attributes, such as photos to paintings, and proposes a two-phase training method to improve the loss function and its generalization error. From the experimental results, transfer learning between different attributes is not as effective as the proposed two-phase training used in knowledge transfer. Especially in VGG-11 with batch normalization (BN), our method can effectively improve the accuracy of 11.78 % and reduce softmax cross-entropy loss by 1.283 and generalization error by 1.496, respectively. Therefore, the multi-scale small sample dataset can improve the information loss caused by one-way transfer, thereby improving the overall network performance, and making its prediction closer to human recognition results.

Keywords: knowledge transfer, the small sample dataset, deep convolutional neural networks (DCNNs), object categorization in different attributes

1 Introduction

Deep convolutional neural networks (DCNNs) have provided great progress in image recognition for different tasks [1-4]. However, most deep learning models use supervised learning as the main training mode, whereas most data in the physical world have no relevant defining tags, such as pathological image recognition, information security monitoring, etc., and correlations between sample information and tags are highly dependent on a manual definition, with consequentially significant cost [5]. In fact, transfer learning (TL) based on information similarity within the dataset and the combination with the desired prediction model could improve specific target domain accuracy [6-7]. In particular, Hu et al. [8] and Triantafillou et al. [9] suggested that constructing a small sample dataset can incorporate relevant knowledge from different perspectives, which can be transferred to specific-tasks to provide optimal state-of-the-art (SOTA) results.

DCNNs is mainly a data-driven processing model, which cannot understand the relevance of features extracted, so the prediction results in the real world are often not as good as humans[10-12, 36]. Therefore, it is necessary to combine transfer learning appropriately and focus on specific feature information to prevent over-fitting due to the small number of training samples. Since object classification with dissimilar attributes mainly focuses on the correlation between images and relative semantics (for example, the images of a dog correspond to the vocabulary of a dog) [14-15]. But object categorization for similar objects with dissimilar attributes in related works, which is still not much. Despite [16] mainly discusses object categorization and object detection in painting. Namely, it does not

* Corresponding Author

include related research on similar objects with dissimilar attributes (for example, the photo of a dog corresponds to the painting of a dog).

From our points of view, the small sample dataset is similar to the task-specific dataset in transfer learning. It allows the network to efficiently capture the features of each class in the target domain and then reduce the loss of the network function caused by one-way transfer. In this paper, we proposed a way to construct a task-specific dataset to help recognize objects with two different attributes, for example with transfer learning the dogs are categorized in real-world and painting. The main contributions of this paper are summarized as follows,

- Based on the human visual attention mechanism, we proposed a multi-scale small sample dataset, in which each single-frame image is to divide into three top-down sub-images. Object features, such as sharp and texture information, had been enhanced easily to capture the features of each class in the target domain and reduce the loss of the network function caused by one-way transfer.
- In order that we explore the knowledge transfer between two different attributes for photos to paintings and to photos, which proposes a two-phase training method that can for the object categorization in different attributes have been improved in prediction results and reducing network generalization error, making it similar to human visual recognition.

The rest of the paper is organized as follows: Section 2 discusses the background and related work. Section 3 describes the proposed approach. Section 4 conducts experiments to evaluate the performance of the proposed approach. Finally, Section 5 conclusions and future work.

2 Background and Related Work

2.1 Deep Convolutional Neural Networks

Deep Learning (DL) is a feature learning algorithm [2], unlike most classical machine learning (ML) algorithms, these DCNNs can perform automatic feature extraction without intervention. When untagged data is trained, each node layer in such a network automatically performs the function of extracting features by repeating the input samples and correcting the error and then predicts their probability distributions. The content is represented by these features and the relationship between these features is captured by these networks. This network, which makes the connection between feature information and the content represented by these features, could be applied for unstructured data. From AlexNet, deep convolutional neural networks have become generalized models by adding the Relu activation function and local response normalization (LRN). Furthermore, in its last two layers, fully-connected (*fc*) layers 6 and 7 dropout functions employed in the *fc* layer effectively prevent network overfitting to intensify the network generalization ability [2]. Unlike AlexNet, Inception V3 draws on the concept of Network-In-Network (NIN) and selects 1x1, 3x3 and 5x5-scale Gabor Filters in the inception module to achieve similar frequency and direction of human vision [17]. In addition, both VGG-11 with BN and Inception V3 can be regarded as the applications of the Hebbian Principle. In these neural networks, the higher the layers in these networks, the sparser and the stronger the representation ability of these networks [18-19]. It can also avoid that the problem of overfitting is in excessive neural network calculations.

2.2 Transfer Learning with the Small Sample Datasets

Transfer learning is an important skill to solve the shortage of training data in ML. The main reason is to relax the assumption that training data and test data must be independent, and both map knowledge from the source domain to the target domain, making it similar to the mechanism of human learning new things. In fact, transfer learning can map the relevant knowledge of the source domain to the target domain, thereby saving the time of labeling training samples and improving the classification effect of the target domain [20].

However, [2, 10, 37] suggested that DCNNs usually has excellent inductive characteristics and has a good impact on target recognition, but DCNNs is mainly data-driven processing and cannot identify the relevance of features, thereby increasing the possibility of the wrong judgment. In fact, the quantity in most public database categories in unbalancing (for example, Stanford Dogs and Stanford Cars), or the image quality is poor (for example, CIFAR-10 and CIFAR-100), which in turn can lead to extremely poor predictions [31-33]. Therefore, we must pay attention to factors such as whether the number of

samples is consistent and whether the image quality is good during the training process.

Since DCNNs are susceptible to interference from uncertain factors, such as noise, light, and color, it reduces the generalization ability of the network [32-33, 51]. Therefore, transfer learning based on a small sample dataset contains the advantages of transfer learning and provides enhanced object features in the task-specific dataset, allowing DCNNs to effectively capture the features of each class in the target domain and reduce the network softmax-cross entropy loss.

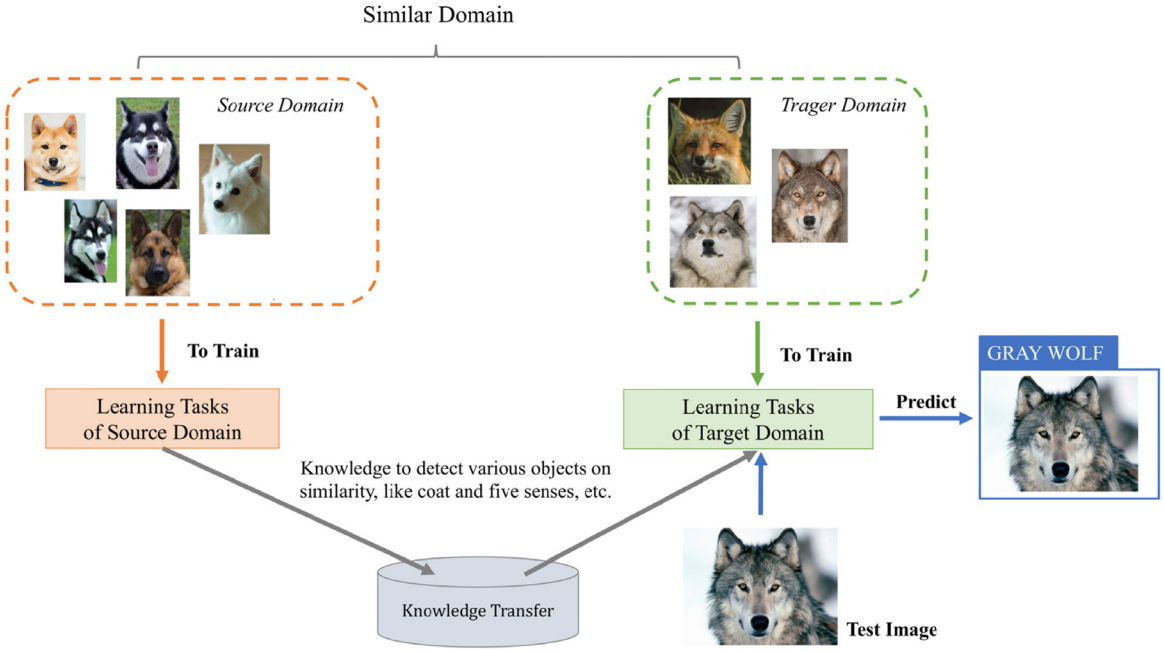

Fig. 1. Illustration of detailed operations in transfer learning

2.3 Object Categorization in Dissimilar Attributes

In recent years, different attributes based on knowledge transfer have been used for object classification re-search. It mainly focuses on the relative vocabulary corresponding to the image, so that the computer can reach the perception ability of the 3-year-old child, thus initiating research in Visual Question Answering (VQA) related fields [14-15].

Crowley and Zitherman are inspired by the Pascal Visual Object Class (PASCAL VOC) challenge [36] and proposed a painted version of the PASCAL VOC dataset, which is significantly improved in the classification and detection [16]. In the field of image classification and detection, however, classification and detection seem to be the same, but they are actually different. In fact, classification is mainly to identify a single object in a single frame image through related algorithms or models, while detection is to identify more than two objects in a single frame image [35]. Despite it can bring a whole new field to the classification of objects in the painting. However, compared with the PASCAL VOC Dataset and ImageNet [24], the richness of this dataset is still insufficient, and the data collection needs to rely on the designated institution. Undoubtedly, it will bring great restrictions to related work in the future. Although the related citation rate of the Oxford Paintings Dataset [16] is not as good as the PASCAL VOC Dataset, it still makes a great contribution to image classification and detection in painting. Therefore, we use this dataset as an experimental control group for object classification in dissimilar attributes.

Since [16] mainly discusses object detection, it lacks the related work of image classification. For this reason, we continue with the results of [33, 38] and use the proposed dataset as a task-specific dataset in transfer learning. The results in Table 6 showed that in classifying objects with two different attributes, the proposed method improves the accuracy of the entire network, and reduces softmax cross-entropy loss function, thereby improving the generalization ability of the network. In other words, our method makes the network very transferable, and the prediction result similar to human vision.

3 The Proposed Approach

In this section, we will detail the processing of the proposed in Fig. 2, the proposed approach contains three parts into (1) the pre-trained with ImageNet for DCNNs; (2) the proposed dataset; and (3) training tricks and prediction results.

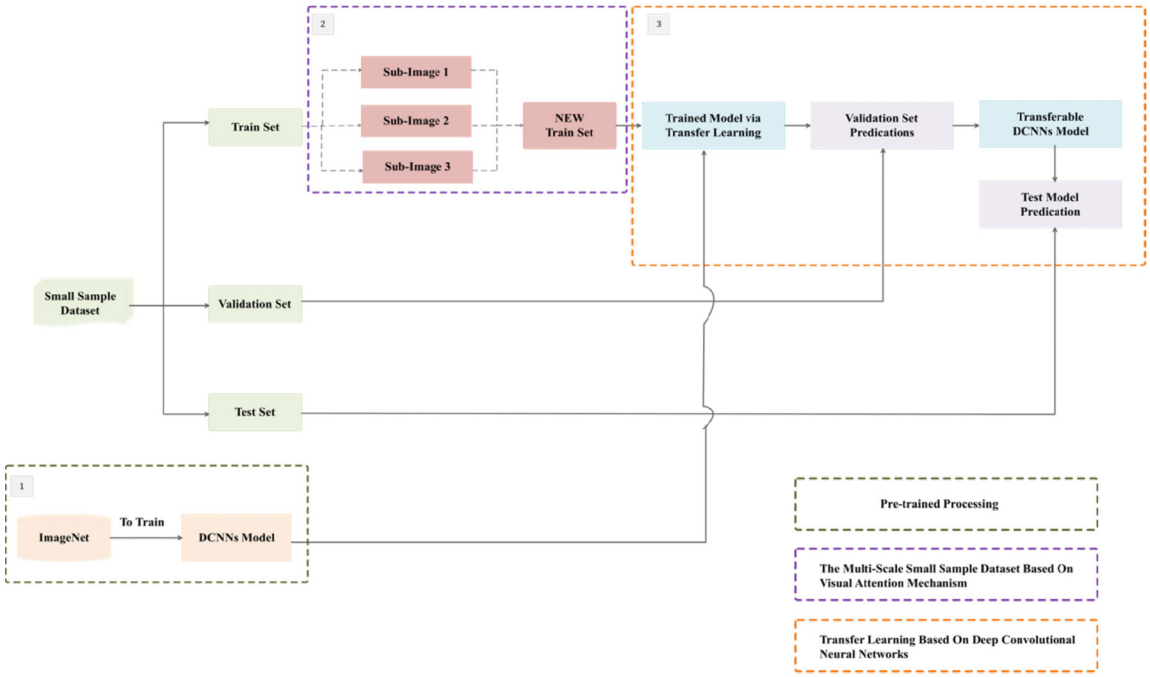

Fig. 2. Illustration of the proposed approach

3.1 Human Visual Attention Mechanism-Based Multi-Scale Small Sample Dataset

Pre-trained Deep Convolutional Neural Networks. Zador [23] suggested that human beings embed knowledge obtained in long-term evolutionary processes within their genomes, whereas DCNNs mostly aim to solve specific-tasks without integration capability. Therefore, distinct pre-training solutions can be constructed for corresponding problems, enabling learning specific skills quickly, which is similar to metric learning [25] and inductive bias concepts [26]. For example, ImageNet richness can simulate human genomes inherited from parents [24]. Hence we used ImageNet as the pre-training dataset to meet transfer learning requirements.

Multi-Scale Sub-Images with Division. The foveal region in the human retina has the strongest sensitivity. To maximize visual information processing, particular visual regions are selected and focused, commonly referred to attention mechanisms [27]. Although the attention mechanism is often used for visual information processing in the field of computer vision, there is no strict mathematical definition for the mechanisms. In image processing, however, local image feature extraction and filter mask movement direction can be both regarded as the attention mechanism. In contrast, a multi-scale regional sub-image processing can extract more meaningful feature information due to the smaller processing range, compared with single-frame image processing, improving accuracy and reducing computational overheads [28-30]. Since most photos are different from specific images (e.g., facial recognition databases and medical images), the background of their images is more complex, and there is no clear specification for the size of the target object. Therefore, we divide target objects into three zones from top to bottom that contain unique feature information as much as possible, increasing sample diversity and calculation speed [13, 29]. The detailed operations of the proposed method are shown in Fig. 3.

Fig. 3. Illustration of detailed of the proposed method

3.2 The Small Sample Dataset for Transfer Learning

Multi-Category Dataset with Balanced Properties. In fact, DCNNs are susceptible to noise or an unbalanced number of training samples for multi-category training, increasing the probability of false predictions [31-32, 51]. Therefore, following [6], we selected 4200 images for 3 classes, such as birds, cows and dogs, from The Oxford Pet-IIIT Dataset [34], The Pascal VOC 2011 Dataset [35] and The Caltech-UCSD Birds-200-2011 Dataset [36]. We set 50% of the dataset as training, 20% as validation and 30% as testing datasets. These datasets were independent and selected randomly to minimize false predictions. Since the input image scale of the network is considered to be 224 x 224 by default [2], images in the training and validation datasets were sub-images obtained by the proposed method and resize each image to 224 x 224.

Knowledge Transfer. Although DCNNs is superior to common feature extraction algorithms, such as HOG, PCA-SIFT, etc., due to the increase in the number of network layers, the rich information closest to the input convolutional layer cannot be passed to the top network, thereby assisting the softmax layer to combine related features [2]. Thus, deep learning has some limitations for practical applications: (1) DCNNs internal design does not consider spatial hierarchy between simple and complex objects, nor can it explain correlations between feature information [37-38, 50]; and (2) humans adapt top-down and bottom-up processing [27-28, 38] to recognize objects (i.e., the brain builds a perception of reality based on prior experience) and the assumption of storing information. However, Pitkow and Angelaki [39] suggested that perception is the causality obtained among several complex sensory data, and estimated the probability for particular nonlinear and dynamic tasks. Therefore, a small sample dataset extracts specific features of information and narrows differences between features for different classes in the target domain, consequently improving network prediction results. The following Fig. 4 illustrates the use of knowledge transfer in the small sample dataset.

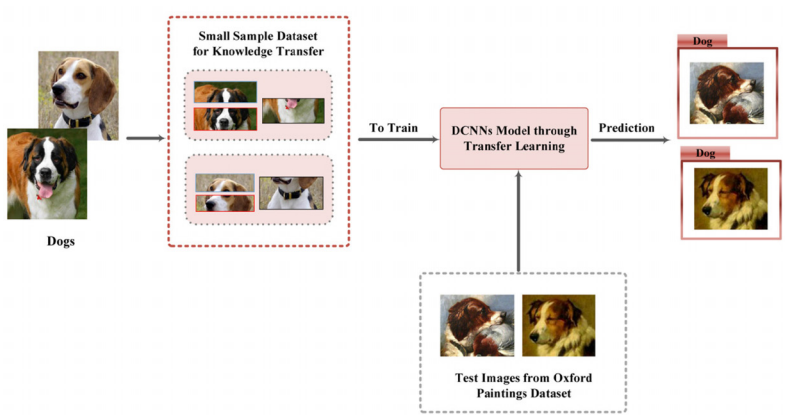

Fig. 4. An example is the classification of a dog in the painting. The target image classification has learned specific features through relevant prior knowledge

Table 1. Dataset structure used in our proposed method

Dataset	#Classes	#Train samples	#Valid samples	#Test samples	Total
The Caltech-UCSD Birds-200-2011	Bird	784	196	420	1400
The Pascal VOC 2011	Cow	784	196	420	1400
The Oxford Pet-IIIT	Dog	784	196	420	1400

3.3 Through Deep Convolutional Neural Networks with Transfer Learning

In deep learning, transfer learning mainly solves similar problems (for example, the photos of dog correspond to the painting of dog), thereby reducing the training time of neural networks and reducing generalization errors [21, 40-41]. It has two main implementation tricks: "weight initialization" and "feature extraction". In other words, these usages are not to retrain the network's weight parameters for new problems, but later fine-tuning all weights of the learned network with a small learning rate [6]. However, the richness of the dataset is determined not only by the number of samples in the dataset but also by whether the dataset used for pre-trained can effectively capture the features similar to the dataset in the target domain [21]. This means that it is to propose to construct a dataset composed of the shape and texture information of the object in order to reduce the feature differences in the target domain (see Table 4). Therefore, we used ImageNet to pre-trained the DCNNs to obtain prior knowledge. In the practical training process of DCNNs, the techniques of "loss function", "weight parameter frozen" and "fully connected layers replacement" are applied to achieve transfer learning (or called deep transfer learning). The detailed operations of two-phase training are shown in Fig. 5.

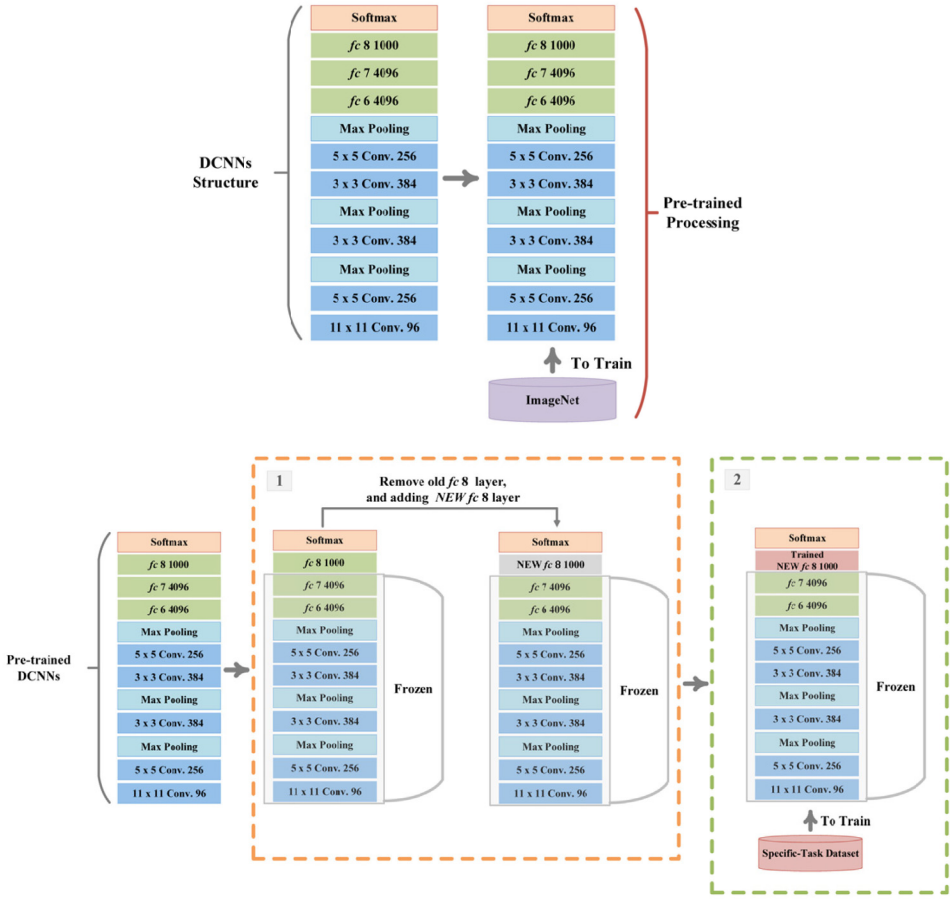

Fig. 5. The proposed method of two-phase training for deep transfer learning. Phase 1: DCNNs is trained with ImageNet (top); Phase 2: First, freezing the weight parameters in the convolutional layers and the *fc* layers. Afterward, replace the old *fc* layer close to the softmax layer with a new *fc* layer. Finally, a specific-task dataset by retrained (bottom)

Loss Function. In deep transfer learning, the training set and the test set are sampled from the distributions *p* and *q*, respectively, and the two types of sample sets are different but related [21]. Thus, Tzeng et al. [6] suggested that can use the feature transformation capabilities of deep neural networks to transform the feature space until the transformed feature distribution matches. In other words, this process can be that the source domain is transformed until it matches the target domain. The matching measure is a maximum mean discrepancy (MMD) distance, it is defined as

$$MMD\left(\mathcal{D}_S,\mathcal{D}_T\right)=\left\|\frac{1}{\left|\mathcal{D}_S\right|}\sum_{x_s\in\mathcal{D}_S}\phi\left(x_s\right)-\frac{1}{\left|\mathcal{D}_T\right|}\sum_{x_t\in\mathcal{D}_T}\phi\left(x_t\right)\right\|_{\mathcal{H}} \tag{1}$$

where $\mathcal{D}_S$ and $\mathcal{D}_T$ denotes the source domain and target domain, which operates on source domain points, $x_s\in\mathcal{D}_S$, and target domain $x_t\in\mathcal{D}_T$. Then $\phi(\cdot)$ is the original variable mapped to Reproducing Kernel Hilbert Space (RKHS) that is reproducible $\left\langle K\left(x\right),K\left(y\right)\right\rangle_{\mathcal{H}}=K\left(x,y\right)$ Hilbert Space.

The loss function is defined as

$$\mathcal{L}=\mathcal{L}_C\left(\mathcal{D}_S,y_S\right)+\lambda MMD^2\left(\mathcal{D}_S,\mathcal{D}_T\right), \tag{2}$$

where $\mathcal{L}_C\left(\mathcal{D}_S,y_S\right)$ denotes the loss function (i.e., the softmax-cross entropy loss) of classification in the source domain, $\mathcal{D}_S$, and the ground truth labels in source domain, y_S, and $MMD^2\left(\mathcal{D}_S,\mathcal{D}_T\right)$ is adaptive loss function in the network (i.e., target domain). The hyper-parameter λ determines how strongly we would like to confuse the domains.

For this work focuses on improves the information loss caused by one-way transfer, knowledge transfer is performed on two very similar datasets. Therefore, we will of the loss function of the pre-trained network (i.e., the source domain) and the loss function of the retrained network (i.e., the target domain), there is added to each other of the final loss function. According to formula (2), we can define it as

$$\mathcal{L}=\mathcal{L}_S\left(\mathcal{D}_S,y_S\right)+\mathcal{L}_T\left(\mathcal{D}_T,y_T\right). \tag{3}$$

where λ is 1.

Weight Parameters Frozen. For DCNNs can effectively build complete information with several stacked datasets incorporating tiny edge information, we use AlexNet, VGG-11 with BN and Inception V3 by train through ImageNet as phase 1 of transfer learning. Subsequently, prior knowledge required for specific-tasks can be obtained through weighting parameters in the networks to achieve TL [40]. In addition, we froze several convolution layers and the *fc* layers except for replaced to the *fc* layer closest to the softmax layer. Afterward, a proposed dataset is used for training specific-tasks. In contrast to Inception V3, AlexNet and VGG-11 with BN have three *fc* layers. In order to avoid affecting network architecture and effectiveness, the remaining two *fc* layers are set to the same operation as the convolution layers except the one closest to the softmax layer.

Fully Connected Layers Replacement. Since the *fc* layer mainly combines edge information and color extracted in the convolution layers with local information, which regarding the class distinguishing the property from the max-pooling layer into a complex feature. This means the output from the uppermost *fc* layer is transmitted to the softmax layer for classification [41]. Thus, the old *fc* layer in phase 2 of transfer learning must be replaced by a new one and the model retrained again to obtain a more accurate classification.

3.4 Learning Strategy

Batch Normalization, Learning Rates, and Epochs. Generally, the step-wise learning rate decay can achieve similar effects by increasing the batch size [18]. Considering batch normalization size, learning rate, training dataset size is proportionally correlated [42], we know that: (1) if the learning rate is increased, then the batch size is preferably increased accordingly so that the convergence is more stable; and (2) use a large learning rate as much as possible, which is beneficial to improve the generalization

ability. In ML training, we know that the choice of hyper-parameters will vary depending on the dataset and the optimizer. Therefore, we will summarize its related formulas from the training experience

$$Interation = \frac{Number\ of\ Samples\ in\ the\ Dataset}{Batch\ Normalization\ Size}, \tag{4}$$

and

$$\text{Epochs} = \mathcal{N}\ \text{Interations}. \tag{5}$$

The Scheduler of Cyclical Learning Rates. Since the learning rate (LR) mainly determines the step size of each iteration, it also tends to the minimum value of the loss function. Namely, it affects the extent to which newly acquired information overwrites old information. This means that the speed at the machine learning model "learn" can be represented metaphorically [44, 18]. Moreover, the learning rate scheduler changes the learning rate during the learning process, which is mainly by controlling the two parameters of "decay" and "momentum", making the learning curve of the model form a convergence state [44]. For there are no specific specifications for the learning rate scheduler used in different optimizers. Therefore, in addition to the step-wise learning rate decay, we have explored the training results of cyclical learning rates (CLR) [45] in experiments, whose formula is defined as follows,

$$\eta = \eta_{MIN} + \left(\eta_{MAX} - \eta_{MIN}\right)\left(MAX\left(0,\ 1-x\right)\right). \tag{6}$$

where η_{MAX} and η_{MIN} denotes upper and lower learning rate, and 1-x must be positive. Then x and $Cycle$ are

$$x = \left|\frac{Ireratoins}{Step\ Size} - 2\left(Cycle\right) + 1\right|, \tag{6-a}$$

$$Cycle = floor\left[\frac{1 + Iterations}{2\left(Step\ Size\right)}\right], \tag{6-b}$$

3.5 Evaluation Metrics

In this section, we will discuss three evaluation metrics index, such as accuracy, softmax cross-entropy loss function and generalization error, in order to judge the experimental results.

Accuracy and Softmax-Cross Entropy Loss Function. Accuracy (Acc.) is an evaluation metrics of the DCNNs and is a kind of the correct prediction [24], which is defined as follows,

$$Accuracy = \frac{Number\ of\ Correct\ \text{Pr}ediction}{Total\ Number\ of\ \text{Pr}ediction} \times 100\%. \tag{7}$$

However, Severyn and Moschitti [43] suggested that the output of the last fc layer of DCNNs and a combination of the softmax logic function (i.e., softmax layer) and the cross entropy determine the prediction, loss, and learning effects. If this neural networks has K categories of predictive outputs, the process steps involved in the prediction are (1) the last layer of the neural network outputs the score for each category; (2) these scores were employed in the softmax logic function to predict the probability of each category, p_c and $c = 1,...,K$; and (3) through the category output probability p_c and its corresponding one-hot label y_c, calculate the cross entropy to get its loss and predict the accuracy. The softmax logic function and cross entropy (or called softmax-cross entropy loss) are defined as follows,

$$p_c = \frac{e^{\mathbf{z}_c}}{\sum_{i=1}^{K} e^{\mathbf{z}_i}} \quad \text{for} \quad c = 1,\cdots,K \ \text{and} \ \mathbf{z} = \left(\mathbf{z}_1,\cdots,\mathbf{z}_K\right) \in \mathbb{R}^K, \tag{8}$$

$$H\left(y, p\right) = -\sum_{c=1}^{K} y_c \log\left(p_c\right). \tag{9}$$

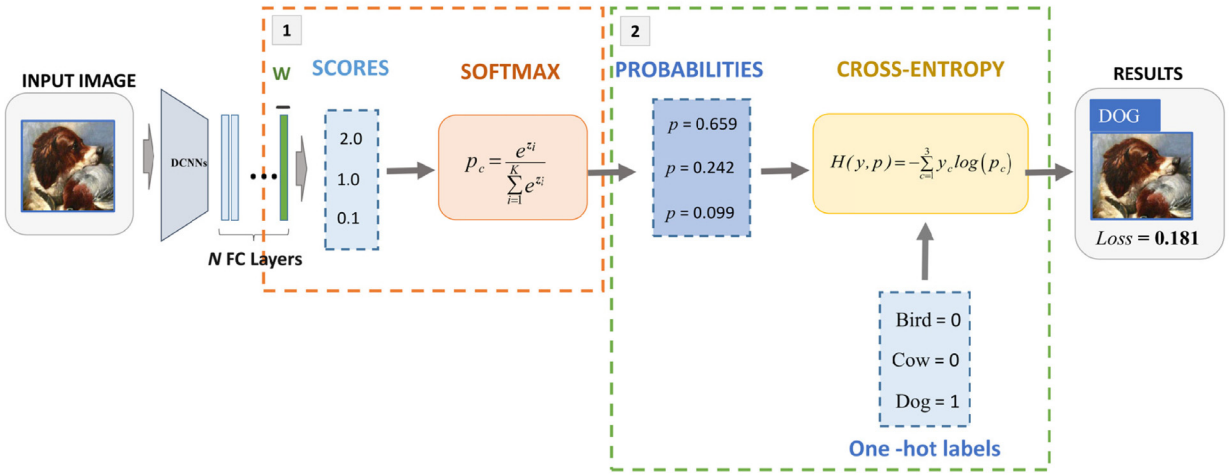

Fig. 6. Illustration of detailed operations in the loss function

Generalization Error. In machine learning, in order to explore the ability of the trained network to adapt to a new dataset, a relevant data set (i.e., test set) is used to test the prediction results of the network. The difference between the two is called generalization error. [47]. The formula definitions following as,

$$E(f;\mathcal{D}) = \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i\right)^2. \tag{10}$$

where $\mathcal{D} = \{(x_1, y_1),(x_2, y_2),...,(x_m, y_m)\}$ denotes test dataset, which x_i is the predicted result and y_i is the true label. Then f is to evaluate the learning performance of the network, the prediction result $f(x_i)$ of the network must be compared with the real label y_i.

4 Experiments

In this section, we have discussed the differences between multi-class image datasets, such as birds, cows and dogs, in dissimilar attributes (e.g., photos and paintings) in transfer learning-based deep convolutional neural networks (e.g., AlexNet, VGG-11 with BN and Inception V3) and used SGD optimizer with momentum. However, to avoid training complexity, we abandoned asynchronous training and using synchronous training to maintain system uniformity.

4.1 Implement Details and Dataset

Hyper-parameter Adjustments. In this experiment, we set up different learning strategies for one-way transfer between datasets with two different attributes. In order to effectively explore the operation of knowledge transfer in dissimilar attributes, we had performed "ImageNet → MS COCO" (i.e., real-world transfers to real-world) and "ImageNet → Oxford Paintings" (i.e., real-world transfers to painting) for experiments. Afterward, batch size and iterations should also be altered corresponding to different datasets and training techniques. All hyper-parameters were listed in Table 2 and Table 3.

Table 2. Hyper-parameters in TL experiment of ImageNet → MS COCO

Network	Epochs		SGD Optimizer with Momentum [42]		The Step-Wise Learning Rate Decay [44]	
	Batch Size	Iterations	Learning Rate	momentum	gamma	step size
AlexNet	48	2625	0.01	0.9	0.1	10
VGG-11 with BN	48	2340	0.1	0.9	0.1	5
Inception V3	16	8750	0.01	0.9	0.1	7

Table 3. Hyper-parameters in TL experiment of ImageNet → Oxford Painting

Network	Epochs		SGD Optimizer with Momentum [42]		Cyclical Learning Rate [45]				
	Batch Size	Iterations	Learning Rate	Momentum	(η_{MAX}, η_{MIN})	$(\text{momentum}_{MAX}, \text{momentum}_{MAX})$	Gamma	Step Size	Type
AlexNet	48	2625	0.01	0.9	(0.1, 0.001)	(0.8, 1)	0.1	2000	triangular
VGG-11 with BN	48	2340	0.1	0.9	(0.1, 0.001)	(0.8, 1)	0.1	2000	triangular
Inception V3	16	8750	0.01	0.9	(0.1, 0.001)	(0.8, 1)	0.1	2000	triangular

Datasets. Despite related art public datasets, such as BAM [48] and Quick [49], these paintings were mostly composed of modern styles or simple lines that were different from the realistic style of the Oxford Paintings dataset. Since we mainly improved the overall performance of the network by enhancing the features of the object. Therefore, whether the similarity of the data structure in the dataset for "Real-World → Painting" was similar to "Real-World → Real-World" was the key to selecting the control group. This experiment employs datasets MS COCO [46] and Oxford Paintings [16] as the test dataset of the comparison group. In order that the number of samples in each dataset is balanced, which were selected 4200 images of three different types of images (birds, cows and dogs) from [46, 16], respectively. After this resized each image to 224-by-224.

Table 4. Test dataset structure of the comparison group

Dataset	Attributes	#Classes	#Train samples	#Valid samples	#Test samples	#Total
MS COCO	Real-World	3	784	196	420	4200
Oxford Paintings	Painting	3	784	196	420	4200

4.2 Experimental Results

In this section, we divided the experiment into two parts into (1) the network architectures; and (2) multi-category categorization in dissimilar attributes. That evaluated whether our method was effective.

The Different Network Architectures. For most related work [7, 14, 24] focuses on assessing the accuracy of the overall network performance, this caused the network to fall into overfitting predictions [26]. In order to avoid this situation, we added "softmax-cross entropy loss function (loss)" and "generalized error (error)" to the evaluation indicators to reduce the occurrence of overfitting. All results were listed in Table 5 to Table 6.

Table 5. Results of the overall network performance for real-world

Architectures	ImageNet → MS COCO			ImageNet → The Proposed Dataset		
	Acc.	Loss	Error	Acc.	Loss	Error
AlexNet	93.33 %	0.5054	0.5045	**98.1 %**	**0.1073**	**0.111**
VGG-11 with BN	95.56 %	0.6759	0.712	**99.21 %**	**0.123**	**0.113**
Inception V3	92.21 %	0.2465	0.2871	**93.33 %**	**0.0261**	**0.041**

Table 6. Results of the overall network performance for painting

Architectures	ImageNet → Oxford Paintings			ImageNet → The Proposed Dataset → Oxford Paintings		
	Acc.	Loss	Error	Acc.	Loss	Error
AlexNet	82.48 %	0.9591	1.142	**87.7 %**	**0.4754**	**0.476**
VGG-11 with BN	87.46 %	1.4056	1.609	**89.87 %**	**0.2827**	**0.303**
Inception V3	74.66 %	0.8561	0.603	**82.96 %**	**0.4086**	**0.481**

Multi-category Categorization in Dissimilar Attributes. In order to intuitively understand the classification effect between categories, we refer to the evaluation metric in [6, 14] to divide the experiment into two parts: (1) the accuracy of three different categories of images (bird, cow and dog);

and (2) the accuracy of the overall network performance (i.e., an average accuracy of the three categories). That evaluated the transferability between the two attributes and determines whether the category information after the transfer was complete. All results were listed in Table 7 to Table 8.

Table 7. Results of multi-category categorization for real-world

Network	ImageNet → MS COCO				ImageNet → The Proposed Dataset			
	Bird	Cow	Dog	Avg. Acc.	Bird	Cow	Dog	Avg. Acc.
AlexNet	98	92.6	89.4	93.33	**99**	**97**	**98**	**98.1**
VGG-11 with BN	97.7	95.3	93.7	95.56	**99.83**	**98.9**	**98.9**	**99.21**
Inception V3	96.83	91	88.8	92.21	**99.8**	**90.32**	**89.9**	**93.33**

Table 8. Results of multi-category categorization for painting

Network	ImageNet → Oxford Paintings				ImageNet → The Proposed Dataset → Oxford Paintings			
	Bird	Cow	Dog	Avg. Acc.	Bird	Cow	Dog	Avg. Acc.
AlexNet	87.6	75.2	84.6	82.48	**90**	**88**	**85**	**87.7**
VGG-11 with BN	93.6	79.5	89.3	87.46	**93.3**	**84.7**	**91.6**	**89.87**
Inception V3	82.6	64.6	76.8	74.66	**86.5**	**76.4**	**86**	**82.96**

4.3 Discussion

Although fine-tuning techniques can overcome the differences between datasets, there is no advantage in passing between two different attributes. From the results of VGG-11 with BN in Table 5 to Table 6 showed that the proposed method not only makes full use of the advantages of fine-tuning, and reduces softmax cross-entropy loss by 0.698 and generalization error by 0.685, respectively. Afterward, it improves the effect of multi-category categorization under dissimilar attributes, thereby achieving excellent transferability (see Table 7 to Table 8). Indeed, this also validates that DCNNs is based on texture information as the main judgment for the benchmark [33]. Nevertheless, it is still limited by the network architecture. The results in Tables 7 and 8 showed that our method improves significant results for the accuracy of AlexNet and VGG-11. However, the effect in Inception V3 is slightly worse. Regardless of AlexNet or VGG-11 with BN, there is mainly a stack type architecture. Thus, it can completely pass the features to the *fc* layers to assist the softmax layer in classification. In other words, the proposed dataset improves the performance of the stack type DCNNs. In addition, Inception V3 of the NIN type architecture, the multi-scale dataset may be regarded as "noise" by the network, thereby reducing accuracy [51]. This means that small data training must develop different training methods for different types of networks. This will maximize the accuracy and performance of the network.

5 Conclusions and Future Work

In this paper we proposed a human visual attention mechanism-based multi-scale small sample dataset, enhancing object features information to improve DCNNs inductive ability, and reducing the computational cost for the softmax cross-entropy loss. In the experiments that we used an SGD with momentum optimizer for DCNNs (AlexNet, VGG-11 with BN, and InceptionV3) by train. Afterward, compare MS COCO with the proposed dataset, our method improves accuracy by 3.18 %, reduces softmax cross-entropy loss by 0.369, and reduces generalization error by 0.413, respectively. Although the multi-scale small sample dataset improves the overall network performance and making its prediction closer to human recognition results. Nevertheless, which cannot clearly explain the relevance of features, making it impossible to play an excellent effect on the different architectural networks. In the future, we will formulate relevant small data learning strategies for DCNNs with different architectures, and improve the interpretability of features, thereby reducing the uncertainty of the network.

Acknowledgements

The authors are grateful to the anonymous reviewers for their constructive comments, which helped improve this paper considerably.

References

[1] J.J. DiCarlo, D. Zoccolan, N.C. Rust, How does the brain solve visual object recognition?, Neuron 73(2012), 415-434.

[2] Y. LeCun, Y. Bengio, G. E. Hinton, Deep Learning, Nature 521(2015) 436-444.

[3] P. Rajpurkar, J. Irvin, A. Park, E. Jones, M. Bereket, K.W. Yeom, K. Shpanskaya, S. Halabi, E. Zucker, G. Fanton, D.F. Amanatullah, C.F. Beaulieu, G.M. Riley, R.J. Stwart, F.G. Blankenberg, D.B. Larson, R.H. Jones, C.P. Langlotz, A.Y. Ng, M.P. Lungren, Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet, PLOS Medicine 15(2018) 1-19.

[4] H. Li, G. Chen, G. Li, Y. Yu, Motion Guided Attention for Video Salient Object Detection, in Proc. 2019 IEEE International Conference on Computer Vision, 2019.

[5] B. Wu, X. Sum, L. Hu, Y. Wang, Learning with Unsure Data for Medical Image Diagnosis, in Proc. 2019 IEEE International Conference on Computer Vision, 2019.

[6] E. Tzeng, J. Hoffman, T. Darrell and K. Saenko, Simultaneous Deep Transfer Across Domains and Tasks, in Proc. 2015 IEEE International Conference on Computer Vision, 2015.

[7] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, IEEE Transactions on Medical Imaging 35(2016) 1285-1298.

[8] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, J. Sun, Meta-SR: A Magnification-Arbitrary Network for Super-Resolution, in Proc. 2019 IEEE International Conference on Computer Vision, 2019.

[9] H. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Zu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, H. Larochelle, Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples, 2020 International Conference on Learning Representations, 2020.

[10] N. Baker, H. Lu, G. Erlikhman, P.J. Kellman, Deep convolutional networks do not classify based on global object shape, PLOS Computational Biology, 2018, 1-14.

[11] Q.-S. Zhang, S.-C. Zhu, Visual Interpretability for Deep Learning: a Survey, Frontiers of Information Technology & Electronic Engineering 19(2018) 27-39.

[12] J. Yang, P.B. Li, Brain Networks of Explicit and Implicit Learning, PLOS ONE, 2012, 1-9.

[13] C.M. Privitera, L.W. Stark, Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence 22(2000) 970-982.

[14] X. Yu, Y. Aloimons, Attribute-Based Transfer Learning for Object Categorization with Zeor/One Training Example, in Proc. the 11th European Conference on Computer Vision, 2010.

[15] L.-J. Li, H. Su, Y. Lim, F.-F. Li, Objects as Attributes for Scene Classification, in Proc. 2010 European Conference on Computer Vision, 2010.

[16] E.J. Crowley, A. Zisserman, The State of the Art: Object Retrieval in Paintings using Discriminative, in Proc. 2014 British Machine Vision Conference, 2014.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[18] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in Proc. 2015 International Conference on Machine Learning, 2015.

[19] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015 International Conference on Learning Representations, 2015.

[20] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22(2010) 1345-1359.

[21] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in Proc. 2014 International Conference on Neural Information Processing Systems, 2014.

[22] F.-F. Li, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Transactions on Pattern Analysis and Machine Intelligence 28(2006) 549-611.

[23] A.M. Zador, A critique of pure learning and what artificial neural networks can learn from animal brains, nature communication 10(2019) 1-7.

[24] O. Russakovsky, D. Jia, S. Hao, J. Krause, S. Satheesh, M. Sean, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, F. -F. Li, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision, 115(2015) 211-252.

[25] J.X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, M. Botvinick, Learning to reinforcement learn, arXiv preprint arXiv: 1611.0576, 2017. <https://arxiv.org/pdf/1611.05763.pdf>.

[26] J. Pearl, Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution, in Proc. 2018 Eleventh ACM International Conference on Web Search and Data Mining, 2018.

[27] J.K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis and F. Nuflo, Modeling visual attention via selective tuning, Artificial Intelligence 78(1995) 507-545.

[28] S. Brandt, L. W. Stark, Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene, Journal of Cognitive Neuroscience 9(1997) 27-38.

[29] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[30] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum Learning, in Proc. 2009 International Conference on Machine Learning, 2009.

[31] J. Feng, Q.-Z Cai, Z. -H. Zhou, Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder, in Proc. 2019 International Conference on Neural Information Processing Systems, 2019.

[32] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, in Proc. 2018 International Conference on Machine Learning, 2018.

[33] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F.A. Wichmann, W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, in Proc. 2019 International Conference on Learning Representations, 2019. <https://arxiv.org/pdf/1811.12231.pdf>.

[34] O.M. Parkhi, A. Vedaldi, A. Zisserman, C. V. Jawahar, Cats and dogs, in Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[35] M. Everingham, L.V. Gool, C.K. I. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes (VOC) Challenge, International Journal of Computer Vision 88(2010) 303-338.

[36] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, The Caltech-UCSD Birds-200-2011 Dataset, California Institute of Technology, CNS-TR-2011-001, 2011. <http://www.vision.caltech.edu/visipedia/CUB-200.html>.

[37] S. Sabour, N. Frosst, G. E. Hinton, Dynamic Routing Between Capsules, in Proc. 2017 International Conference on Neural Information Processing Systems, 2017.

[38] Z.Z. Si, S.-C. Zhu, Learning AND-OR Templates for Object Recognition and Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 35(2013) 2189-2206.

[39] X. Pitkow, D. E. Angelaki, Inference in the brain: Statistics Flowing in Redundant Population Codes, Neuron 94(2017) 943-953.

[40] E. Tzeng, E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep Domain Confusion: Maximizing for Domain Invariance, Computing Research Repository, Vol. abs/1412.3474 (2014). <https://arxiv.org/pdf/1412.3474.pdf>.

[41] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, in Proc. 2014 International Conference on Machine Learning, 2014.

[42] I. Sustskever, J. Martens, G. Dahl, G.E. Hinton, On the importance of initialization and momentum in deep learning, in Proc. International Conference on Machine Learning, 2013.

[43] A. Severyn, A. Moschitti, Learning to rank short text pairs with convolutional deep neural networks, in Proc. 2015 International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015.

[44] Y. Bengio, Practical Recommendations for Gradient-Based Training of Deep Architectures, Neural Networks: Tricks of the Trade, 2nd ed., Springer Berlin Heidelberg, 2012, 437-478.

[45] L.N. Smith, Cyclical Learning Rates for Training Neural Networks, in Proc. 2017 IEEE Winter Conference on Applications of Computer Vision, 2017.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramnan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in Proc. 2014 European Conference on Computer Vision, 2014.

[47] I. Goodfellow, Y. Bengio, A. Courville, Optimization for Training Deep Models, Deep Learning, MIT Press, 2016, 271-325.

[48] M.J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, S. Belongie, BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography, in Proc. 2017 IEEE International Conference on Computer Vision, 2017.

[49] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, N. Fox-Gien, The Quick Draw!, A.I. experiment, Google, 2016. <quickdraw.withgoogle.com>.

[50] Y. Bengio, The Consciousness Prior, arXiv preprint arXiv: 1709.08568v2, 2020. <https://arxiv.org/pdf/1709.08568v2.pdf>.

[51] W. Woods, J. Chen, C. Teuscher, Adversarial explanations for understanding image classification decisions and improved neural network robustness, nature machine intelligence, 1(2019), 508-516.