Neurocomputing 328 (2019) 29–38

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A self-adaptive cascade ConvNets model based on label relation mining

Zhihua Wei, Wen Shen*, Cairong Zhao, Duoqian Miao

Department of Computer Science and Technology, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 7 November 2017
Revised 7 March 2018
Accepted 22 March 2018
Available online 18 August 2018

Keywords:
Image classification
Cascade
Three-way decision
Label relation

ABSTRACT

Uncertainty is a fundamental and unavoidable feature in daily life, which is the same for a single classifier. Thus, combining the predictions of many different classifiers is a very successful way to reduce the uncertainty. In this paper, we present a Correcting Reliability Level (CRL) supervised three-way decision (3WD) cascade model to implement image classification tasks. Our model simulates the human decision process by using 3WD to judge "certainty" or "uncertainty" of the classification result. When judged as "uncertainty", CRL will supervise the 3WD and learn more information to make the final prediction. In addition, we introduce two Class Grouping methods to mining the relation between labels, which help us to train several expert ConvNets for different types of images. Experimental results show that our model can effectively reduce the classification error rate compared with the base classifier.

1. Introduction

How to improve the classification accuracy is the focus of the image classification task. Recent advances in deep learning have shown that convolutional neural network is very good at discovering intricate structures in high-dimensional images [1–3]. In recent years, there have been many well known models that have beaten records in image recognition, such as AlexNet [4], VGGNet [5], GoogLeNet [6], ResNet [7] and so on.

Before the ConvNets prevailing, conventional image classification techniques generally design a feature extractor that transformed the raw data (the pixel values of an image) into a suitable internal representation or feature vector from which the classifier could classify patterns in the input [3,8–11]. SIFT [12], HOG [13] LBP [14] and Haar [15] are four famous traditional feature description methods. However, they have different application scenarios and difficult to be promoted to different tasks. For instance, HOG features combined with SVM classifier [16–18] has been widely used in image recognition, especially in pedestrian detection has been a great success, while Haar-like features perform well at face detection tasks.

Different from conventional image classification techniques, ConvNets do not need to design feature extractors and can be used to identify different objects. ConvNets can automatically discover the representations from raw data. There are multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) [3]. Although the recent model [7] has been proved that the classification error rate on ImageNet dataset [19] is lower than the error rate of humans, we are still groping for more efficient or more interesting models. In addition to designing a convolution neural network with higher learning ability, combining the predictions of many different models is a very successful way to reduce test errors [4,20,21].

To this end, in this work, we propose a method for combining classifiers in a "cascade" which allows the information to be passed between classifiers and the output information of the upper classifier is used as the additional information of the next classifier. More creative, we introduce the theory of three-way decisions (3WD) to determine how information is passed between classifiers. We build a 3WD-based cascade model (3WD-CM) including 3 layers, the first layer is a base ConvNet for all categories, the second layer is a three-way decision layer and the third layer consists of several experts ConvNets. Therefore, we propose a class grouping algorithm to mining the relation between labels and then train several deep ConvNets become experts of different types of images. To better play experts classifiers roles, we further develop a "Correcting Reliability Level" (CRL, which will be defined in Section 3.4) supervised 3WD cascade model (CRL-CM). Furthermore, in order to automatically mine the relation between labels and construct better expert classifiers, we propose a new class grouping algorithm based on the Latent Dirichlet Allocation [22] topic model, which can automatically determines the number of expert classifiers and the categories that make up the expert classifier. Experimental

* Corresponding author.
E-mail address: 1810068@tongji.edu.cn (W. Shen).

https://doi.org/10.1016/j.neucom.2018.03.082
0925-2312/

results on JD clothing dataset and Stanford Dogs Dataset [23,24] demonstrate the effectiveness of the proposed methods.

This paper is an extension of our conference paper of the 2017 Chinese Conference on Computer Vision (CCCV2017)[25]. The new contribution of this paper are summarized as follows: (1)We have proposed one new class grouping algorithm based on the Latent Dirichlet Allocation and replace the original class grouping algorithm which can not automatically determines the number of expert classifiers. The 3WD-CM with LDA-CG (LDA-based class grouping algorithm) adopt the topic model to learn the hidden association between categories so that construct expert classifiers with similar categories. (2)We have conducted more experiments and analysis on image classification task to show the effectiveness of the proposed methods. Specifically, we include the experiments of the CRL-CM on JD clothing dataset and Stanford Dogs dataset.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work. Section 3 details the proposed class grouping algorithm and CRL-CM. Section 4 reports the experimental results and analysis, and Section 5 concludes this paper.

2. Related work

2.1. Convolutional neural networks cascade

ConvNets are now the most commonly used large-scale image classification models. As early as 1990, ConvNets was trained for the task of classifying low-resolution images of handwritten digits [1]. Later, a gradient-based optimization was applied to ConvNets, laying the foundation for learning with gradient descent [2]. Since Krizhevsky et al. [4] trained a large, deep ConvNet (named AlexNet) to win over other contestants in the ILSVRC-2012 competition, ConvNets appear in each session of ILSVRC competition and the record was broken again and again. However, the ability of a single classifier is limited. A single classifier may be good at predicting some certain categories, while bad at predicting others. Thus, many researchers try to improve the accuracy by combining the predictions of many different models [26–32]. For example, Viola and Jones [30] combined classifiers in a cascade which allows background regions of the image to be quickly discarded while spending more computation on promising face-like regions. Simonyan and Zisserman [27] proposed a two-stream ConvNet architecture which incorporates spatial and temporal networks. Each stream was implemented using a deep ConvNet, softmax scores of which were combined by late fusion. Qin et al. Inspired by these works, we use cascaded ConvNets to classify images. We first use a base classifier to do classification roughly. If the classification result is not reliable, an expert classifier will do a second time classification meticulously.

2.2. Three-way decisions

The notion of three-way decisions was originally introduced by the needs to explain the three regions of probabilistic rough sets [33–35]. A theory of three-way decisions is constructed based on the notions of acceptance, rejection and noncommitment [36], whenever it is impossible to make an acceptance or a rejection decision, the third noncommitement decision is made [37]. Three-way decisions play a key role in everyday decision-making and have been widely used in many fields and disciplines. Three-way spam filtering systems [38,39], for example, add a suspected folder to allow users make further examinations of suspicious emails, thereby reducing the chances of misclassification. Three-way decisions are also commonly used in medical decision making [40,41]. In the threshold approach to clinical decision making proposed in [40], by comparing the probability of disease with a pair of a "testing" threshold and a "test-treatment" threshold, doctors make one of three decisions: (a) no treatment no further testing; (b) no treatment but further testing; (c) treatment without further testing. In this work, we borrow the idea of three-way decision and proposed a 3WD-based ConvNets cascade model for image classification tasks. When there is doubt about the base classifier's classification result, the model will make a noncommitement decision and learn more information from expert classifiers to make the final prediction.

3. A self-adaptive cascade ConvNets model

We use GoogLeNet model throughout the paper. GoogLeNet, as defined in [6], was the winner of ILSVRC 2014 with a top 5 error rate of 6.7% and was one of the first ConvNet architectures that really strayed from the general approach of simply stacking convolution and pooling layers on top of each other in a sequential structure. GoogLeNet does not use the full connection (FC) layer any more, since the FC layer occupies almost 90% of the network parameters. Thus, GoogLeNet has 12x fewer parameters than AlexNet. GoogLeNet uses 9 Inception modules in the whole architecture, which is a network in the network structure [42]. By using the Inception models, the network is deeper and wider and the performance can improved 2–3 times.

In this section, we first introduce the notation used in this work. Then, we present the class grouping algorithm based on confusion matrix, the class grouping algorithm based on topic model, the 3WD-based cascade model, the CRL-supervised 3WD cascade model.

3.1. Notation

Let *Img* present the input image. Let $CAT = \{c_1, c_2, \cdots, c_i, \cdots, c_C\}$ be the class set, which includes C classes. Let $P = (p_{ij})_{N\times C}$ be the classification result of test set, where p_{ij} is the probability that i is classified as category c_j. Let $Conf = (n_{ij})_{C\times C}$ be the confusion matrix of ConvNet test result, where n_{ij} is the number of images being classified as class c_j, while their true label is c_i. The bigger the n_{ij}, the easier that images of class c_i are classified as class c_j. Let *Top-1 class* (referred to as c_{top}) be the class considered the most probable by the model, the probability is p_{top}.

It is obvious that we do not need to consider the situation that *Img* belonging to class c_i if p_i is very small. Therefore, we need a *threshold of possible classes (Th-pos)* to determine the possible classes of *Img*. If p_i is no less than *Th-pos*, we think that *Img* may belong to class c_i. We stipulate that *Th-pos* is no less than $\frac{1}{C}$.

3.2. Class grouping based on confusion matrix

For commodity images of web-based platforms, many commodity classes are similar to each other, which is difficult for both humans and machines to distinguish them (see Fig. 1). Therefore, a classifier trained for all the classes is not enough for distinguishing those similar classes, we need some more specified classifiers trained for certain similar classes. It is unreliable to rely on human observation to determine which classes are similar. In this paper, we propose a Class Grouping (CG) algorithm (see Algorithm 1) based on the feedback of the classification results.

A confusion matrix, also known as an error matrix [43], is a specific table layout that allows visualization of the performance of an algorithm. We can learn how does the classifier confuse a class with another class. We train a base ConvNet for all the C classes and compute the similarity between classes based on confusion matrix of this base ConvNet.

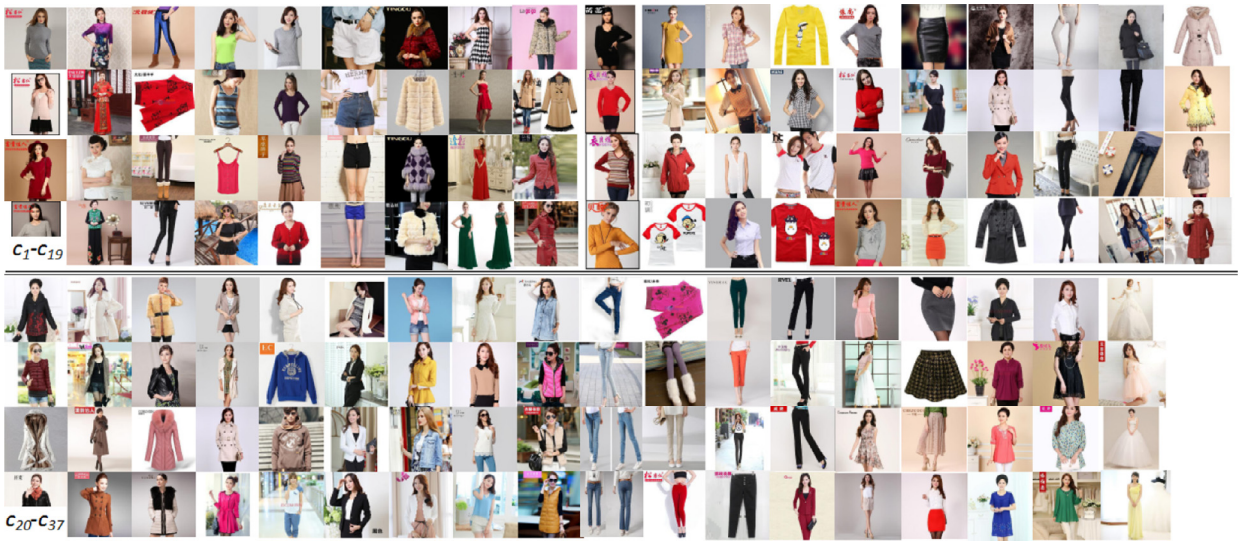

Fig. 1. JD clothing dataset, 4 samples each class. Classes c_1, c_5, c_{10} and c_{14} are of similar features, they are all kinds of sweaters. And classes c_3, c_{18}, c_{29}, c_{30}, c_{31} and c_{32} are kinds of trousers.

Algorithm 1 Class Grouping.

Input: *CAT*; *S*; cluster number *K*;
Output: *K* clusters;
Make each class in *CAT* a cluster;
Compute pair-wise distance of all clusters, $d_{mn} = min(s_{ij}), \forall\ c_i \in clt_m\ \&\ \forall\ c_j \in\ clt_n$;
repeat
  find two clusters that are closest to each other;
  merge the two clusters form a new cluster clt_{new};
  compute the distance form clt_{new} to all other clusters.
until there are only *K* clusters

We define s_{ij} the similarity of class c_i to class c_j, see Eq. (1).

$$s_{ij} = \frac{n_{ij}}{\sum_{t=1}^{C} n_{it}} \tag{1}$$

We define S_{ij} the similarity between class c_i and class c_j, see Eq. (2).

$$S_{ij} = s_{ij} * s_{ji}, i = 1, 2, \cdots, C-1; j = i+1, \cdots, C \tag{2}$$

Then, we get the similarity matrix S of all classes,

$$S = \begin{bmatrix} 0 & S_{12} & S_{13} & S_{14} & \cdots & S_{1C} \\ 0 & 0 & S_{23} & S_{24} & \cdots & S_{2C} \\ 0 & 0 & 0 & S_{34} & \cdots & S_{3C} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & S_{(C-1)C} \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

After class grouping, we can divide class set *CAT* into several subsets ($cat\text{-}k, k = 1, 2, \cdots, K$). We train an expert ConvNet *ExpConvNet-k* for *cat-k*. These ExpConvNets will be used to build the cascade model.

Based on class grouping experimental results, we introduce *similar-classes*. If class c_i and c_j are belong to the same subset *cat-k*, we call that c_i and c_j are similar-classes. Similar-classes are of high probability to be wrongly classified to each other and therefore require special treatment.

3.3. Class grouping based on topic model

The above class grouping algorithm has a drawback, the number of subsets needs to be given. We give the number of subsets based on experience, but this can not reflect the latent relationship between categories. Therefore, we need a method to automatically determine the number of subsets. Inspired by Latent Dirichlet Allocation [22] topic model, which can extract latent topics form documents, we proposed a new class grouping algorithm based on topic model, which can extract latent subsets from classification results.

A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. In this work, we compare category to word, subset to topic and probability to word frequency. For example, in this work, p_{ij} can be interpreted as the frequency of the word j appears in the document i. Thus, the class grouping task can be interpreted as extracting K topics from N documents.

Different from class grouping algorithm based on confusion matrix, the class grouping algorithm based on topic model can automatically determine the number of subsets via calculating perplexity values of models. Perplexity [44–46] is a measurement of how well a probability model predicts a sample and is often used to compare probability models. A low perplexity indicates that the probability model is good at predicting samples. The following is the modeling process.

In order to compare image to document, we need to do pretreatment. We define a function that calculates "term frequency" (*tf*)

$$tf = round(A * x), \tag{3}$$

where A is a constant which represents the total "words" number of a "document" and x is a probability value. For instance, $tf_{ij} = round(A * p_{ij})$ means the number of times word(category) j occurs in document(the classification result of image) i. Sometimes the probability of the real category is really small, which leads to the *tf* being 0. This will lost the relevance of some categories. Thus, we define another constant B to represent the initial *tf* of the real category. The final *tf* of the real category is $tf_{iREAL} + B$.

For image $i, \theta_i = \{pt_1, pt_2, ...pt_k, ..., pt_K\}$ represents the distribution of topics(subsets) in document(image) i. Where

$$pt_k = \frac{nt_k}{n_i}, \tag{4}$$

represents the probability of topic k occurring in document i for any given word. Where nt_k is the number of words (in document i) occur in topic k and n_i is the total number of words in document i.

For subset k, $\varphi_k = \{pL_1, pL_2, ...pL_c, ..., pL_C\}$ represents the distribution of words(categories) in topic(subset) k. Where

$$pL_c = \frac{nL_c}{n_k}, \tag{5}$$

represents the probability of word c occurring in topic k. Where nL_c is the number of word c occurring in all the documents and n_k is the total number of words (of topic k) in all the documents.

The relationship of document(image), topic(subset) and word(category) can be represented as

$$p(c|i) = p(c|k)p(k|i) = \frac{nL_c}{n_k}\frac{nt_k}{n_i}. \tag{6}$$

Topic is the hidden layer between document and word, similarly, subset is the hidden layer between image and category.

Then we calculate perplexity value to evaluate the model.

$$perp = \exp\frac{\sum_{i=1}^{N}\log p(w_i)}{\sum_{i=1}^{N} n_i}, \tag{7}$$

where

$$p(w_i) = \sum_i \prod_c \sum_k p(c|k)p(k|i)p(i). \tag{8}$$

3.4. Three-way decision based cascade model

A theory of 3WD is constructed based on the notions of acceptance, rejection and noncommitment. It is an extension of the commonly used binary-decision model with an added third option [36]. 3WD is a decision model of human-cognitive. 3WD holds that people can make quick decisions about what they can accept or reject with certainty in real decision-making, while for those uncertain things, people tend to postpone the judgment, namely delay decision-making. There are many reasons for delayed decision-making, such as the available information is not sufficient, the assessment of risk is not comprehensive, the cognition of the event is not thorough enough and so on. When people have a certain level of information, risk and cognition, they will make the final judgment. 3WD is an intermediate step of the final realization of binary-decision. For instance, we move from true/false into true/unsure/false. The third option "unsure" can also be referred to as a deferment decision that requires further information or investigation. Ultimately, the answer we need remains true/false.

For the classification result, we no longer directly accept it. Instead, we make one of two decisions: (a) accept it if it is reliable; (b) opt for a noncommitment if it is not reliable. Since this is not a binary-decision problem with two options, but a multiclass classification problem, there is no "reject" option. Thus, decision (b) is the "third option". We judge the classification result is not reliable if meeting two conditions: (i) existing class c_a and $p_a \geq Th\text{-}pos$; (ii) c_a and c_{top} belonging to the same subset $cat\text{-}k$. The condition (i) guarantees that class c_a is possible for Img, condition (ii) guarantees that c_a and c_{top} are similar-classes. We define these conditions under the hypothesis that if meeting these two conditions, it means that the classifier is confused. The classifier considers that Img can be predicted as c_a and c_{top}, which means that we need to put the image into the expert classifier $ExpConvNet\text{-}k$ for further judgment. Eq. (9) is the 3WD process.

$$3wd = \begin{cases} \text{delay}, & \text{satisfiying condition (i) and (ii)} \\ \text{accept}, & \text{otherwise} \end{cases} \tag{9}$$

In practical applications, misclassification is very common since the ability of a single classifier is limited. Therefore, combining several different models is a way to reduce the error rate [4]. Cascade [47] is a special case of ensemble learning. The basic idea of cascade is the connection of multiple classifiers. The information is passed between layers and the output information of the upper classifier is used as the additional information of the next classifier.

Under the guidance of 3WD theory, we establish a self-adaptive cascade ConvNets model, including 3 layers (see Algorithm 2). The

Algorithm 2 3WD-CM.

Input: image Img,size $n * n$
Output: prediction P,size C
Input Img into the first layer (a base ConvNet), get P^1;
Send the classification result P^1 to the 3WD layer;
if $\exists\ c_a,\ \&\ p_a \geq Th\text{-}pos$ **then**
  if $c_a \in cat\text{-}k$ & $c_{top} \in cat\text{-}k$ **then**
    put Img into $ExpConvNet\text{-}k$, get P^2;
    calculate the final P with Eqn. 10.
  else
    $P = P^1$
  end if
else
  $P = P^1$
end if

first layer is a base classifier (a base ConvNet), the second layer is a 3WD layer and the third layer is expert layer, including several experts classifiers (ExpConvNets). We put Img into the first layer (a base classifier) and send the classification result P^1 into the 3WD layer. Next, the 3WD layer will make one of two decisions: (a) accept P^1 as the final P; (b) opt for a noncommitment and put Img into $ExpConvNet\text{-}k$ (the classification result is denoted by P^2). Finally, we calculate probability P based on P^1 and P^2, see Eq. (10).

$$p_i = \begin{cases} p^2_{c_i}, & \text{if } c_i \in cat\text{-}k \\ p^1_{c_i}, & \text{otherwise} \end{cases} \tag{10}$$

Where $p^1_{c_i}$ represents the probability of the image being predicted as class c_i by the base classifier and $p^2_{c_i}$ represents the probability of the image being predicted as class c_i by the expert classifier.

3.5. CRL-supervised 3WD cascade model

3WD decides which images may need expert judgments. But the experimental experience tells us that blindly following the decision of 3WD is an aiduous but fruitless matter. It not only wastes time but also has no obvious help to reduce the error rate. Suppose that the base classifier considers c_i as c_{top} and the final c_{top} considered by 3WD-CM is c_j, experimental experience tells us that there are two situations: (1) in most cases, c_i is the correct class while c_j is wrong; (2) on the contrary, in most cases, c_i is wrong while c_j is the correct class. Situation (1) tells us that the 3WD makes the classification result from right to wrong; while situation (2) tells us that 3WD makes the classification result from wrong to right. Obviously, we welcome the latter situation. Thus, we need to supervise the 3WD process. In other words, we need to further determine that which image really needs expert judgement instead of blindly following the 3WD.

We define *Correcting Reliability Level (CRL)* to measure the necessity of the expert judgments. A high CRL means that c_{top} considered by the base classifier has high probability of being wrong result (in other words, the expert judgment has a high probability of correcting the wrong result).

We define TF_i (Truth to False) to describe the situation (1) above, and FT_i (False to Truth) to describe the situation (2) above.

Table 1
Class grouping results and ExpConvNets error rates.

Subset	Labels	ExpConvNet	Top-1 error rate(%)
cat-1	c_2, c_8, c_{15}, c_{27}, c_{33}, c_{37}	ExpConvNet-1	9.16
cat-2	c_1, c_5, c_{10}, c_{14}, c_{25}, c_{35}	ExpConvNet-2	41.17
cat-3	c_3, c_6, c_{17}, c_{18}, c_{21}, c_{29}, c_{30}, c_{31}, c_{32}	ExpConvNet-3	23.50
cat-4	c_4, c_{12}, c_{13}, c_{34}	ExpConvNet-4	16.67
cat-5	c_7, c_9, c_{11}, c_{16}, c_{19}, c_{20}, c_{22}, c_{23}, c_{24}, c_{26}, c_{28}, c_{36}	ExpConvNet-5	33.48

CRL computing see Eq. (11).

$$CRL_i = \begin{cases} \frac{N_{FT_i} - N_{TF_i}}{N_{FT_i} + N_{TF_i}} & \text{while } N_{FT_i} > N_{TF_i} \ \& \ N_{FT_i} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The denominator indicates the total number of expert judgement when c_i is considered as c_{top} by the base ConvNet, and the numerator represents the number of net effective expert judgement.

We use a random function

$$R(p) = binomial(1, p) \tag{12}$$

to move CRL value to a Boolean value, which tells the model whether put *Img* into the expert classifier. The return value of the function R is a Boolean value, "True" means that *Img* needs expert judgment and "False" means that *Img* dose not need expert judgment. Wherein, p is the probability of return "True". $R(CRL_i)$, for instance, has a probability of CRL_i to return "True". Therefore, the larger the CRL value is, the more likely that *Img* will be passed into the expert classifier.

On the basis of 3WD-CM, we add a CRL table after the 3WD layer. CRL table is used to determine whether *Img* is worthy of expert judgment (see Algorithm 3).

Algorithm 3 CRL-CM.

Input: image *Img*,size $n * n$
Output: prediction P,size C
Input *Img* into the first layer (a base ConvNet), get P^1;
Send the classification result P^1 to the 3WD layer;
if $\exists\ c_a,\ \&\ p_a \geq Th\text{-}pos$ **then**
  if $c_a \in cat\text{-}k$ & $c_{top} \in cat\text{-}k$ **then**
    Check CRL table and get $crl = CRL_{top}$
    if R(crl) is TRUE **then**
      put *Img* into *ExpConvNet-k*, get P^2;
      calculate the final P with Eqn. 10.
    else
      $P = P^1$
    end if
  else
    $P = P^1$
  end if
else
  $P = P^1$
end if

4. Experiments

In this section, we evaluate the CRL-supervised 3WD cascade model with two class grouping methods on two datasets: JD clothing dataset and Stanford Dogs dataset. The followings describe the detailed settings and experimental results.

4.1. Commodity image classification

4.1.1. JD clothing dataset

The experimental data of this paper is JD clothing dataset, examples see Fig. 1. JD is one of the most famous B2C shopping site in China and the first large-scale integrated business platform to be listed in the United States. JD has a strong market share; therefore, it has accumulated a large number of commodity image data, which provides researchers with a lot of resources. Our experimental dataset has about 400,000 clothing images, including 37 classes. The dataset is divided into training set, validation set and test set at a ratio of 8: 1: 1.

In this work, we report two error rates: top-1 and top-5, where the top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered most probable by the model [4].

4.1.2. Experiments with confusion matrix based class grouping method

In this work, we do experiments on a deep learning framework named Caffe. Yosinski et al. [48] pointed out that fine-tuning is better than randomly initialize parameters. Our experimental results also confirm this. We start from a pre-trained model (GoogLeNet on ImageNet LSVRC-2014) and fine-tune it. The top-1 error rate is 44.59% and the top-5 error rate is 9.40%, which are better than randomly initializing (the top-1 error rate is 57.11% and the top-5 error rate is 21.16%). We call the fine-tuned model Base Model (BM) below and the latter experiments will fine-tune models on the basis of it.

Test results confirm that images of many classes are easily to be misclassified to each other, like class c_1 and c_{10} (see Fig. 2, readers can view the electronic draft and enlarge the figure to see the details). Thus, we group those similar classes into the same subset with confusion matrix based class grouping method introduced in Section 3.2.

We set $K = 5$ and divide *CAT* into 5 subsets. After class grouping, we train an expert ConvNet for each subset. We fine-tune the BM and adapt most of the architecture (only change the output number of the last layer), and resume training from the B weights. Table 1 shows the class grouping result and the ExpConvNets error rates. After obtaining 5 expert classifiers, we can establish 3WD-CM with 5 expert classifiers, referred as 3WD-CM(5) below.

In order to get the CRL table, we first need to test images with 3WD-CM(5). We set Th-pos 0.1 in this experiment. The top-1 error rate of 3WD-CM is 44.401%, reducing by 0.189% compared with BM; top-5 error rate is 9.475%, increasing by 0.075% compared with BM. The results are worse than we expected, the top-1 error rate reduces a little bit, and the top-5 error rate does not drop but increase.

Table 2 shows the counting results of TF and FT. We can see that there are totally 1195 cases of situation (1) and 1,273 cases of situation (2) (two situations see Section 3.5). Therefore, in fact only 78 samples are modified correctly by 3WD-CM, the accuracy increases by only 0.189%. When c_1 is considered as c_{top} by the base classifier, there are totally 613 images considered needing expert judgement by 3WD, wherein, 135 images are modified correctly and 478 images are modified incorrectly. Thus, there are 343 images being modified incorrectly in total. This shows that when c_1 is considered as c_{top}, we should better ignore the decision of 3WD that *Img* needing expert judgment. We should better accept c_1 as the prediction result. On the contrary, if the base classifier

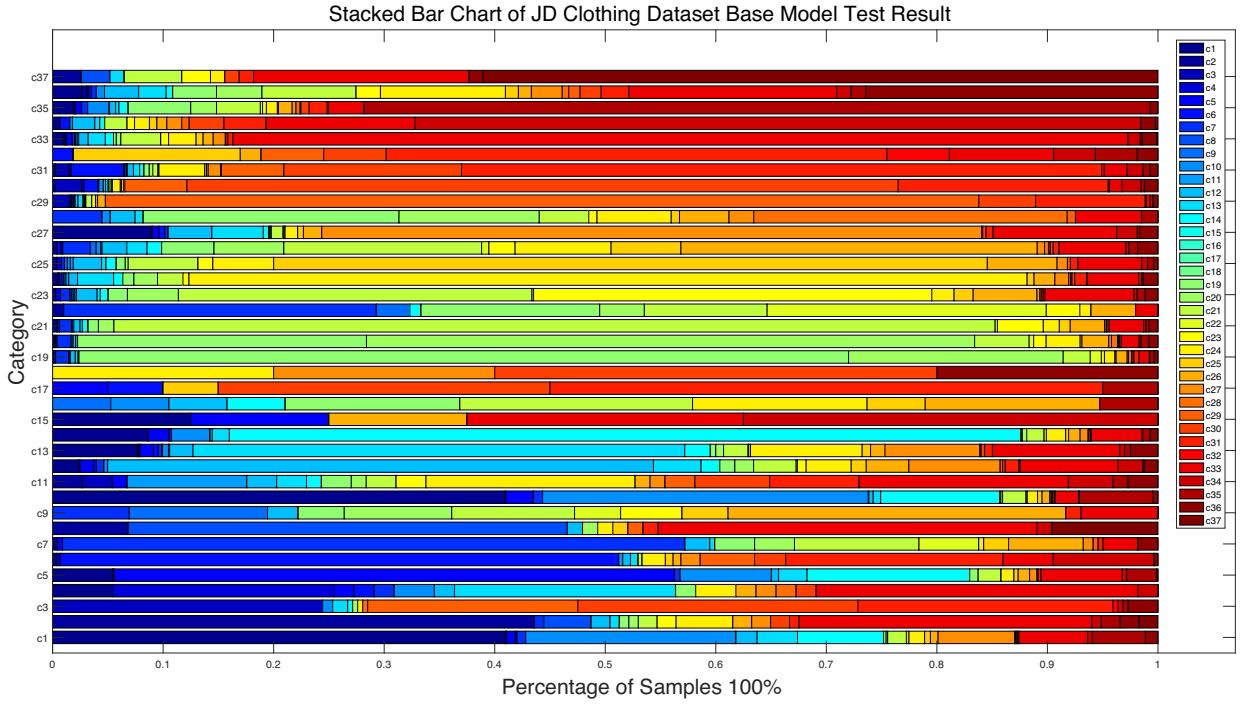

Fig. 2. Stacked bar chart of test result of BM. Take class c_1 as example, there is about 20% of test images misclassified as class c_{10}. Similarly, there is about 40% of test images of class c_{10} misclassified as class c_1.

Table 2
Classification results of 3WD-CM(5).

Label	N_{FT}	N_{TF}	NET	Label	N_{FT}	N_{TF}	NET
c_1	135	478	−343	c_{20}	58	54	4
c_2	4	7	−3	c_{21}	0	22	−22
c_3	11	3	8	c_{22}	3	8	−5
c_4	1	1	0	c_{23}	22	10	12
c_5	33	18	15	c_{24}	37	32	5
c_6	16	13	3	c_{25}	3	8	−5
c_7	5	2	3	c_{26}	11	19	−8
c_8	5	2	3	c_{27}	6	20	−14
c_9	3	2	1	c_{28}	9	1	8
c_{10}	602	150	452	c_{29}	21	26	−5
c_{11}	1	2	−1	c_{30}	30	55	−25
c_{12}	9	9	0	c_{31}	64	52	12
c_{13}	3	13	−10	c_{32}	1	1	0
c_{14}	73	29	44	c_{33}	15	42	−27
c_{15}	0	0	0	c_{34}	0	2	−2
c_{16}	0	0	0	c_{35}	17	43	−26
c_{17}	0	0	0	c_{36}	12	10	2
c_{18}	0	0	0	c_{37}	3	3	0
c_{19}	60	58	2	Total	1273	1195	78

Note: $NET = N_{FT} - N_{TF}$.

Table 3
CRL table of CRL-CM(5).

Label	CRL	Label	CRL	Label	CRL	CRL	Label
c_1	0	c_{11}	0	c_{21}	0	c_{31}	0.103
c_2	0	c_{12}	0	c_{22}	0	c_{32}	0
c_3	0.571	c_{13}	0	c_{23}	0.375	c_{33}	0
c_4	0	c_{14}	0.431	c_{24}	0.072	c_{34}	0
c_5	0.294	c_{15}	0	c_{25}	0	c_{35}	0
c_6	0.103	c_{16}	0	c_{26}	0	c_{36}	0.091
c_7	0.429	c_{17}	0	c_{27}	0	c_{37}	0
c_8	0.429	c_{18}	0	c_{28}	0.8		
c_9	0.2	c_{19}	0.017	c_{29}	0		
c_{10}	0.601	c_{20}	0.036	c_{30}	0		

Table 4
Average error rates of CRL-CM(5) underdifferent *Th-pos*.

Th-pos	top-1 error rate(%)	top-5 error rate(%)
0.1	43.50 (1.09)	9.386 (0.017)
0.2	43.62 (0.97)	9.398 (0.005)
0.3	43.68 (0.91)	9.415 (-0.013)
0.4	43.77 (0.82)	9.413 (-0.011)

Note: The numbers in parentheses are reduced error rates (%) compared with BM(GoogLeNet).

considers c_{10} as c_{top}, we would better follow the 3WD and do an expert judgment for *Img*. Because, for c_{10}, the number of images which are modified correctly is greater than the number of images which are modified incorrectly. Therefore, we use CRL to supervise 3WD process.

Then, we use Eq. (11) to calculate CRL of each class, see Table 3, and establish CRL-CM(5).

Now, we can establish CRL-CM(5). We test images with CRL-CM(5) and Table 4 is the classification performance of CRL-CM(5) under different *Th-pos* values. We set *Th-pos* with 0.1, 0.2, 0.3 and 0.4. We test 30 times for each *Th-pos* value and take the average error rate. Compared with BM, the top-1 error rate reduces by about 1.09% when $Th\text{-}pos = 0.1$. Top-5 error rate does not reduce obviously. With the increasement of *Th-pos*, the error rate reduces less, because the greater the *Th-pos* is, the more harsh that 3WD judge a result being "uncertainty", thus, those misclassified samples lose the chance of being modified correctly. The experimental results show that the CRL-CM(5) can effectively reduce the classification error rate compared with a single base ConvNet.

4.1.3. Experiments with topic model based class grouping method

In this section, we use topic model based class grouping method to learn expert subsets and then construct a new CRL-CM.

First, we calculates the *tf* of JD Clothing test set. As introduced in Section 3.3, we treat the prediction result (output of Softmax layer, a *K*-d vector of probability) of test image *i* as a document *i*. p_{ij} is the probability that *i* is classified as category c_j. We

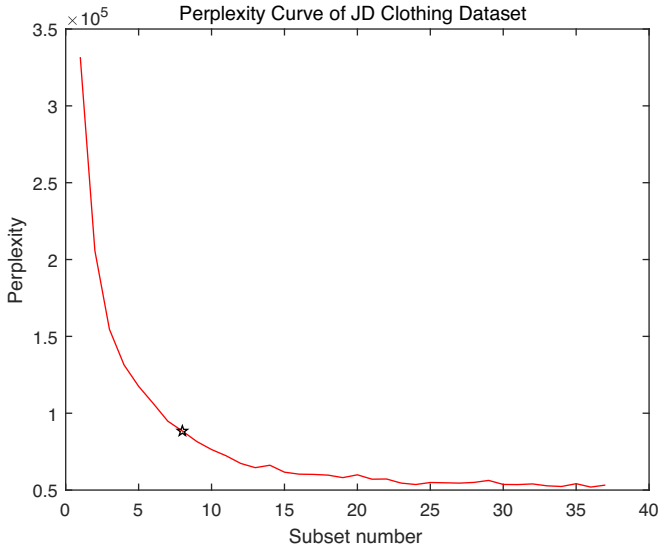

Fig. 3. Perplexity curve of experiment on JD Clothing test set.

Table 5
Clustering result of topic model based class grouping method.

Subset	Labels	ExpConvNet	top-1 error rate(%)
cat-1	c_3, c_6 ,c_{29}, c_{30}, c_{31}, c_{34}	ExpConvNet-1	32.58
cat-2	c_{12}, c_{33}	ExpConvNet-2	6.04
cat-3	c_1, c_{10}	ExpConvNet-3	41.27
cat-4	c_5, c_{25}, c_{35}	ExpConvNet-4	8.03
cat-5	c_{13}, c_{27}, c_{36}	ExpConvNet-5	23.91
cat-6	c_{10}, c_{14}	ExpConvNet-6	14.11
cat-7	c_{21}, c_{24}, c_{26}	ExpConvNet-7	13.86
cat-8	c_7, c_{19} ,c_{20}, c_{21}, c_{23}	ExpConvNet-8	27.92

calculate the *tf* with Eq. (3). Thus, the *tf* of word c_j in document i is $round(A^* p_{ij})$, we set $A = 10$ in this work. If c_j is the real label, the *tf* is $round(A * p_{ij}) + B$. We set $B = 5$ in this work.

After calculating *tf* of each word in all the documents, we run LDA topic model to mining the latent relationship of all the words. Fig. 3 shows the perplexity curve of experiment on JD Clothing test set. When subset number bigger than 10, the perplexity value no longer reduce significantly. Thus, at the beginning, we select the topic model with 10 topics. However, we found that when the number of topic is greater than 8, there is a phenomenon in which the labels overlap. Therefore, we select the topic model with 8 topic numbers.

Finally, we extract eight new subsets from the original 37 categories, see Table 5. Compared with Table 1, topic model based class grouping method can put aside some less relevant categories, only focus on mining closely related categories. Most of the new subsets contain only 2 to 3 categories, which means the work of each subset is clearer. Then we establish 3WD-CM with 8 expert classifiers, referred as 3WD-CM(8) below.

Table 6 shows the counting results of TF and FT. We can see that there are totally 1,573 cases of situation (1) and 1,306 cases of situation (2). There are totally 267 samples are modified correctly by 3WD-CM(8). Then, we use Eq. (11) to calculate CRL of each class, CRL table see Table 7.

Table 8 shows the comparison of the two class grouping method. Experiments show that topic model based class grouping method do better on class grouping task. CRL-CM(8) performs better than CRL-CM(5), it reduces top-1 error rate from 43.50% to 43.21%.

Table 6
Classification results of 3WD-CM(8).

Label	N_{FT}	N_{TF}	NET	Label	N_{FT}	N_{TF}	NET
c_1	274	469	−195	c_{20}	82	53	29
c_2	0	0	0	c_{21}	57	32	25
c_3	23	4	19	c_{22}	0	2	−2
c_4	0	0	0	c_{23}	22	13	9
c_5	2	17	−15	c_{24}	20	9	11
c_6	13	15	−2	c_{25}	0	7	−7
c_7	10	1	9	c_{26}	14	12	2
c_8	0	0	0	c_{27}	17	5	12
c_9	0	0	0	c_{28}	0	2	−2
c_{10}	652	157	495	c_{29}	25	39	−14
c_{11}	0	1	−1	c_{30}	50	56	−6
c_{12}	10	8	2	c_{31}	55	59	−4
c_{13}	39	23	16	c_{32}	0	0	0
c_{14}	96	39	57	c_{33}	9	25	−16
c_{15}	0	0	0	c_{34}	8	2	6
c_{16}	0	0	0	c_{35}	5	201	−196
c_{17}	0	0	0	c_{36}	4	4	0
c_{18}	0	0	0	c_{37}	0	0	0
c_{19}	86	51	35	Total	1573	1306	267

Note: $NET = N_{FT} - N_{TF}$.

Table 7
CRL Table of CRL-CM(8).

Label	CRL	Label	CRL	Label	CRL	Label	CRL
c_1	0	c_{11}	0	c_{21}	0.281	c_{31}	0
c_2	0	c_{12}	0.111	c_{22}	0	c_{32}	0
c_3	0.704	c_{13}	0.258	c_{23}	0.257	c_{33}	0
c_4	0	c_{14}	0.422	c_{24}	0.379	c_{34}	0.6
c_5	0	c_{15}	0	c_{25}	0	c_{35}	0
c_6	0	c_{16}	0	c_{26}	0.077	c_{36}	0
c_7	0.818	c_{17}	0	c_{27}	0.545	c_{37}	0
c_8	0	c_{18}	0	c_{28}	0		
c_9	0	c_{19}	0.255	c_{29}	0		
c_{10}	0.612	c_{20}	0.215	c_{30}	0		

Table 8
Experiments results of JD clothing test set.

Model	top-1 error rate(%)	top-5 error rate(%)
BM	44.59	9.40
CRL-CM(5)	43.50	9.39
CRL-CM(8)	43.21	9.42

Note: The *Th-pos* here is 0.1.

4.2. Animal image classification

In order to verify the proposed method can work on different dataset, we do experiments on another dataset—Stanford Dogs dataset.

4.2.1. Stanford Dogs dataset

The Stanford Dogs dataset contains images of 120 breeds of dogs from around the world. This dataset has been built using images and annotation from ImageNet for the task of fine-grained image categorization. The smallest category has 148 images, the largest category has 252 images. Each category leaves 40 images as test data.

4.2.2. Topic model based class grouping

We first trained a base model BM-DOG by fine-tuning GoogLeNet on ImageNet LSVRC-2014. Then we test its classification performance. The top-1 error rate is 35.54% and the top-5 error rate is 21.46%. Fig. 4 is the stacked bar chart of test result of BM-DOG (readers can view the electronic draft and enlarge the Fig. 4 to see the details). From Fig. 4 we can see that there is obvious confusion between some categories. For example, 7 out of 20 test images with true label c_1 are misclassified into label c_{16},

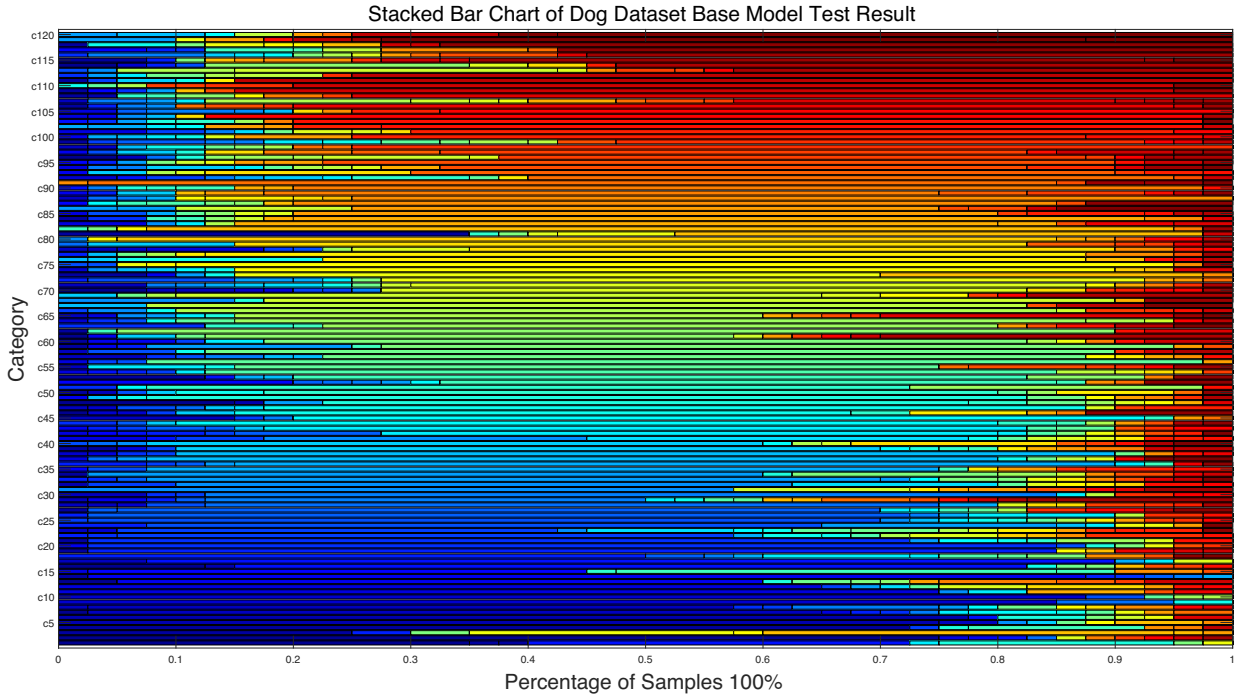

Fig. 4. Stacked bar chart of test result of BM-DOG.

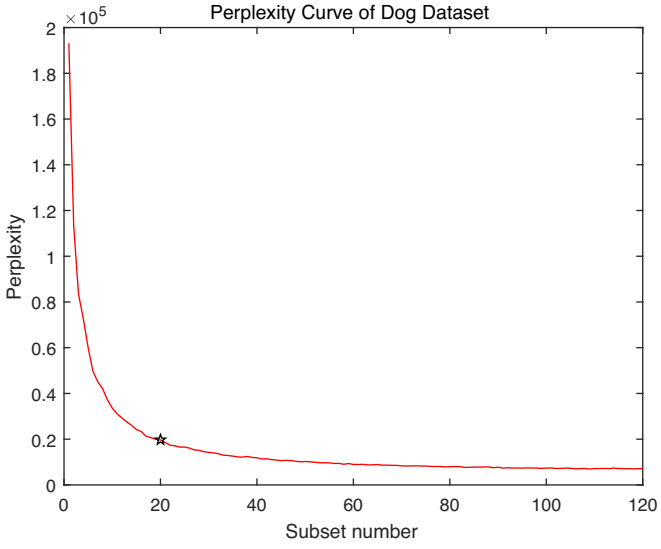

Fig. 5. Perplexity curve.

Table 9
Topic model based class grouping results of Dog dataset.

Subset	Classes	ExpConvNet	Top-1 error rate(%)
cat-1	c_{35}, c_{42}, c_{67}, c_{94}, c_{112},	ExpConvNet-1	17.5
cat-2	c_{28}, c_{65}, c_{83}, c_{107},	ExpConvNet-2	9.0
cat-3	c_{3}, c_{55}, c_{73}, c_{81},	ExpConvNet-3	7.19
cat-4	c_{24}, c_{33}, c_{72}, c_{119}, c_{120},	ExpConvNet-4	5.0
cat-5	c_{40}, c_{79}, c_{89}, c_{93}, c_{97},	ExpConvNet-5	5.71
cat-6	c_{25}, c_{69}, c_{85}, c_{110}, c_{113},	ExpConvNet-6	3.33
cat-7	c_{29}, c_{54}, c_{87}, c_{117},	ExpConvNet-7	10.5
cat-8	c_{21}, c_{44}, c_{51}, c_{60}, c_{63}, c_{116},	ExpConvNet-8	1.5
cat-9	c_{7}, c_{22}, c_{32}, c_{41}, c_{71},	ExpConvNet-9	2.92
cat-10	c_{34}, c_{59}, c_{78}, c_{84},	ExpConvNet-10	6.07
cat-11	c_{1}, c_{16}, c_{48}, c_{61}, c_{96}, c_{114},	ExpConvNet-11	4.38
cat-12	c_{18}, c_{30}, c_{99}, c_{101}, c_{104}, c_{105},	ExpConvNet-12	6.67
cat-13	c_{39}, c_{43}, c_{50}, c_{90}, c_{103},	ExpConvNet-13	7.19
cat-14	c_{4}, c_{12}, c_{31}, c_{53}, c_{68}, c_{70},	ExpConvNet-14	3.5
cat-15	c_{2}, c_{8}, c_{13}, c_{26}, c_{57}, c_{74},	ExpConvNet-15	3.5
cat-16	c_{11}, c_{23}, c_{66}, c_{82}, c_{98}, c_{115},	ExpConvNet-16	4.29
cat-17	c_{15}, c_{36}, c_{52},	ExpConvNet-17	2.5
cat-18	c_{19}, c_{20}, c_{58}, c_{106},	ExpConvNet-18	2.92
cat-19	c_{37}, c_{46}, c_{64}, c_{76}, c_{118},	ExpConvNet-19	5.42
cat-20	c_{86}, c_{95}, c_{100}, c_{102}, c_{109},	ExpConvNet-20	6.0

while 2 out of 20 test images with true label c_{16} are misclassified into label c_1. This means that label c_1 and label c_{16} are easily confused by the classifier. Thus, in theory, the grouping algorithm should put label c_1 and label c_{16} into the same group.

Then we run topic model based class grouping algorithm on the test result $P = (p_{ij})_{N\times C}$. We still set constant $A = 10$ and $B = 5$. Fig. 5 is the perplexity curve. The perplexity value reduce slowly after subset number $= 20$. Thus, we select those 20 subsets as experts, see Table 9. The class grouping result basically matches our expectation. We can see that label c_1 and c_{16} are in the same subset cat-11, which is in line with the expected result. We then establish 3WD-CM(DOG) based on these 20 expert classifiers.

4.2.3. CRL-supervised 3WD cascade model

We test images with 3WD-CM(DOG) and set Th-pos 0.1 in this experiment. And then, we calculate the CRL table based on the classification result of 3WD-CM(DOG), see Table 10. The model works best for label c_{21}, c_{30}, c_{83}, c_{107} and c_{119}, the CRL values are all 1.0. Also, the model works well for label c_{12}, c_{29}, c_{31}, c_{89}, c_{94}, c_{100}, c_{102} and c_{113}, the CRL values are no less than 0.5.

Table 11 compares the classification results of BM-DOG and CRL-CM(DOG), we verify that our CRL-CM can effectively improve the classification performance of different kind of datasets. It is worth noting that the top-5 error rate of CRL-CM(DOG) is half of the BM(DOG). This suggests that the proposed CRL-CM can effectively improve the top-5 prediction performance when subdividing similar samples.

Table 10
CRL table of Dog dataset.

Label	CRL	Label	CRL	Label	CRL	Label	CRL	Label	CRL
c_1	0	c_{25}	0	c_{49}	0	c_{73}	0	c_{97}	0
c_2	0	c_{26}	0	c_{50}	0	c_{74}	0	c_{98}	0
c_3	0.333	c_{27}	0	c_{51}	0	c_{75}	0	c_{99}	0.333
c_4	0	c_{28}	0	c_{52}	0	c_{76}	0	c_{100}	0.5
c_5	0	c_{29}	0.5	c_{53}	0	c_{77}	0	c_{101}	0.333
c_6	0	c_{30}	1.0	c_{54}	0	c_{78}	0	c_{102}	0.5
c_7	0	c_{31}	0.6	c_{55}	0	c_{79}	0	c_{103}	0
c_8	0	c_{32}	0	c_{56}	0	c_{80}	0	c_{104}	0
c_9	0	c_{33}	0	c_{57}	0	c_{81}	0	c_{105}	0
c_{10}	0	c_{34}	0	c_{58}	0	c_{82}	0	c_{106}	0
c_{11}	0	c_{35}	0	c_{59}	0	c_{83}	1.0	c_{107}	1.0
c_{12}	0.5	c_{36}	0	c_{60}	0	c_{84}	0	c_{108}	0
c_{13}	0	c_{37}	0	c_{61}	0.143	c_{85}	0.333	c_{109}	0
c_{14}	0	c_{38}	0	c_{62}	0	c_{86}	0	c_{110}	0
c_{15}	0	c_{39}	0	c_{63}	0.333	c_{87}	0	c_{111}	0
c_{16}	0	c_{40}	0	c_{64}	0.333	c_{88}	0	c_{112}	0
c_{17}	0	c_{41}	0.25	c_{65}	0	c_{89}	0.5	c_{113}	0.5
c_{18}	0	c_{42}	0	c_{66}	0	c_{90}	0	c_{114}	0
c_{19}	0	c_{43}	0	c_{67}	0.143	c_{91}	0	c_{115}	0
c_{20}	0	c_{44}	0	c_{68}	0	c_{92}	0	c_{116}	0
c_{21}	1.0	c_{45}	0	c_{69}	0.2	c_{93}	0	c_{117}	0
c_{22}	0	c_{46}	0	c_{70}	0	c_{94}	0.5	c_{118}	0
c_{23}	0	c_{47}	0	c_{71}	0.429	c_{95}	0	c_{119}	1.0
c_{24}	0	c_{48}	0.333	c_{72}	0	c_{96}	0	c_{120}	0

Table 11
Classification results comparison.

Model	Top-1 error rate(%)	Top-5 error rate(%)
BM(DOG)	35.54	21.46
CRL-CM(DOG)	34.65	10.12

Note: The *Th-pos* here is 0.1.

5. Conclusion

In this paper we have proposed a CRL-supervised 3WD cascade model (CRL-CM). By mining label relation from the confusion matrix, we learn a set of expert classifiers to correct the base classifier's prediction result. To better mine the relation between labels, we proposed another class grouping method based on topic model. Experimental results show that the proposed method can achieve better performance than the base classifier(GoogLeNet). The contributions of this paper are: (i) Simulating the human decision process by using 3WD to construct a cascade model with several Exp-ConvNets which become experts on inputs preprocessed in different ways; (ii) introducing topic model to learn relation between labels. Experimental results are presented to show the effectiveness of the proposed methods.

Acknowledgments

The work is partially supported by the National Natural Science Foundation of China (No. 61573259, 61673301, 61573255 and 61673299) and the Special Project of the Ministry of Public Security (No. 20170004).

References

[1] Y. Lecun, B. Boser, J.S. Denker, R.E. Howard, W. Habbard, L.D. Jackel, D. Henderson, Handwritten digit recognition with a back-propagation network, in: Proceedings of the Advances in Neural Information Processing Systems, 1990, pp. 396–404.
[2] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[3] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[4] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the International Conference on Neural Information Processing Systems, 25, 2012, pp. 1097–1105.
[5] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. 2015 (2015). arXiv: 1409.1556
[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 2015, pp. 1–9.
[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2016) 2016, 770–778.
[8] M. Xu, J. Zhu, P. Lv, B. Zhou, M.F. Tappen, R. Ji, Learning-based shadow recognition and removal from monochromatic natural images., IEEE Trans Image Process PP (99) (2017) 5811–5824.
[9] Q. Miao, P. Xu, T. Liu, Y. Yang, J. Zhang, W. Li, Linear feature separation from topographic maps using energy density and the shear transform., IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 22 (4) (2013) 1548–1558.
[10] A. Liu, W. Li, W. Nie, Y. Su, 3d models retrieval algorithm based on multimodal data, Neurocomputing 259 (2017) 176–182.
[11] A. Liu, Y. Lu, W. Nie, Y. Su, Z. Yang, Hep-2 cells classification via clustered multi-task learning, Neurocomputing 195 (C) (2016) 195–201.
[12] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[13] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[14] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, 1, 1994, pp. 582–585.
[15] C.P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: Proceedings of the International Conference on Computer Vision, 2002, pp. 555–562.
[16] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (5) (1999) 988–999.
[17] C. Campbell, Kernel methods: a survey of current techniques, Neurocomputing 48 (1) (2002) 63–84.
[18] D.S.A. V., Advanced support vector machines and kernel methods, Neurocomputing 55 (1) (2003) 5–20.
[19] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F.F. Li, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009, pp. 248–255.
[20] R.M. Bell, Y. Koren, Lessons from the netflix prize challenge, ACM Sigkdd Explor. Newslett. 9 (2) (2007) 75–79.
[21] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32.
[22] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[23] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization, in: Proceedings of the First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO.
[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: CVPR09, 2009.
[25] W. Shen, Z. Wei, C. Zhao, D. Miao, A self-adaptive cascade convnets model based on three-way decision theory, in: Proceedings of the Chinese Conference on Computer Vision, 2017, pp. 433–444.

[26] J. Schmidhuber, U. Meier, D. Ciresan, Multi-column deep neural networks for image classification, in: Proceedings of the Chinese Conference on Computer Vision Computer Vision and Pattern Recognition, 2012, pp. 3642–3649.
[27] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 1 (4) (2014) 568–576.
[28] D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascade face detection and alignment, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 109–122.
[29] H. Qin, J. Yan, X. Li, X. Hu, Joint training of cascaded cnn for face detection, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 3456–3465.
[30] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[31] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.
[32] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the Computer Vision and Pattern Recognition, 2013, pp. 3476–3483.
[33] Y. Yao, Three-way decision: An interpretation of rules in rough set theory, in: Proceedings of the International Conference on Rough Sets and Knowledge Technology, 2009, pp. 642–649.
[34] Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (6) (2011) 1080–1096.
[35] Y. Yao, Rough Sets and Three-Way Decisions, Springer International Publishing, 2015.
[36] Y. Yao, An Outline of a Theory of Three-Way Decisions, Springer Berlin Heidelberg, 2012.
[37] Y. Yao, Three-way decisions with probabilistic rough sets., Inf. Sci. 180 (3) (2011) 341–353.
[38] B. Zhou, Y. Yao, J. Luo, A three-way decision approach to email spam filtering, in: Proceedings of the Advances in Artificial Intelligence, Canadian Conference on Artificial Intelligence, Canadian, Ai 2010, Ottawa, Canada, May 31 - June 2, 2010, pp. 28–39.
[39] B. Zhou, Y. Yao, J. Luo, Cost-sensitive three-way email spam filtering, J. Intell. Inf. Syst. 42 (1) (2014) 19–45.
[40] S.G. Pauker, J.P. Kassirer, The threshold approach to clinical decision making, N. Engl. J. Med. 302 (20) (1980) 1109.
[41] J.D. Lurie, H.C. Sox, Principles of medical decision making., Spine 24 (5) (1999) 493.
[42] M. Lin, Q. Chen, S. Yan, Network in network, Comput. Sci. 2014 (2014). arXiv: 1312.4400
[43] S.V. Stehman, Selectingand interpreting measures of thematic classification accuracy., Remote Sens. Environ. 62 (1) (1997) 77–89.
[44] F. Jelinek, R.L. Mercer, L.R. Bahl, J.K. Baker, Perplexitya measure of the difficulty of speech recognition tasks, J. Acoust. Soc. Am. 62 (544) (1977) S63.
[45] C.H. Chang, Simulated annealing clustering of chinese words for contextual text recognition, Pattern Recognit. Lett. 17 (1) (1996) 57–66.
[46] J. Gao, H.F. Wang, M. Li, K.F. Lee, A unified approach to statistical language modeling for chinese, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP '00., 3, 2002, pp. 1703–1706.
[47] J. Gama, P. Brazdil, Cascade generalization, Mach. Learn. 41 (3) (2000) 315–343.
[48] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, Eprint Arxiv 27(2014)3320–3328.

Zhihua Wei is currently an associate professor Tongji University. She received the double Ph.D degrees from Tongji University, and Lyon2 University in 2010, M.S. degree and B.S. degree from Tongji University in 2005 and 2000. Her research interests include machine Learning, image processing and data mining.

Wen Shen is currently a master student at Tongji University. Her research interests include machine learning, image recognition and deep learning etc. E-mail: 1631596@tongji.edu.cn.

Cairong Zhao is currently an associate professor at Tongji University. He received the PhD degree from Nanjing University of Science and Technology, M.S. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, and B.S. degree from Jilin University, in 2011, 2006 and 2003, respectively. His research interests include computer vision, face recognition, intelligent computing and vision attention etc. E-mail: zhaocairong@tongji.edu.cn.

Duoqian Miao received the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing China in 1997. He is currently a Professor with the School of Electronics and Information Engineering and vice dean of the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China. His current research interests include soft computing, rough sets, granular computing, pattern recognition, machin learning, and intelligent systems. He has published over 200 scientific articles in interational journals, books, and conferences. He is Fellow of International Rough Set Society (IRSS), distinguished member of China Computer Federation(CCF), executive council member of Shnaghai Computer Society (SCA).