SCIENCE CHINA
Information Sciences

• LETTER •

January 2020, Vol. 63 119101:1–119101:3
https://doi.org/10.1007/s11432-018-9567-8

Large margin deep embedding for aesthetic image classification

Guanjun GUO[1], Hanzi WANG[1*], Yan YAN[1], Liming ZHANG[2] & Bo LI[3]

[1]*Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen* 361005*, China;*
[2]*Faculty of Science and Technology, University of Macau, Macau* 999078*, China;*
[3]*Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing* 100191*, China*

Received 25 May 2018/Accepted 3 August 2018/Published online 12 September 2019

Citation Guo G J, Wang H Z, Yan Y, et al. Large margin deep embedding for aesthetic image classification. Sci China Inf Sci, 2020, 63(1): 119101, https://doi.org/10.1007/s11432-018-9567-8

Dear editor,
The problem of aesthetic image classification has attracted much attention during the past few years. The recently proposed methods [1–4] based on the deep convolutional neural network (CNN) [5,6] have achieved large improvements over the methods based on the handcrafted aesthetic features. Although the existing CNN-based methods have shown superior performance, the task of aesthetic image classification is still very challenging. This is mainly due to the fact that aesthetic images are usually captured in complex environments and they have different subjects and styles.

Deep embedding methods have achieved impressive performance on different computer vision applications. They usually use deep neural networks to map input images to a compact space, where the intra-class distances of the features of the input images belonging to the same class are smaller than the inter-class distances of the features of those from different classes. Generally speaking, deep embedding methods can be roughly divided into siamese-based deep embedding methods and triplet-based deep embedding methods. For the task of image classification, the triplet-based deep embedding methods usually obtain better performance than the siamese-based deep embedding methods, because the former methods explicitly encourage the inter-class separability. However, the performance of the traditional triplet-based deep embedding methods greatly drops when they are applied to the task of aesthetic image classification. The main reason is that aesthetic images usually contain large intra-class variations. Therefore, it is difficult to map aesthetic images belonging to difference classes to a compact space by using the traditional triplet-based deep embedding method.

In the deep embedding methods, only minimizing a triplet loss function (belonging to the class of metric loss functions) results in a problem; i.e., it is hard to map a few challenging images to a compact space. We call such images as hard samples. As a result, the distance-based classifiers cannot obtain satisfactory performance owing to the existence of hard samples in aesthetic images. However, hard samples can be correctly classified with a hyperplane-based classifier, which minimizes a hinge loss function (belonging to the class of margin-based loss functions). In this study, we propose a large margin deep embedding (LMDE) method, which minimizes a joint loss function by combining a triplet loss function and a hinge loss function. In the training stage, the minimization of the joint loss function ensures that the intra-class variability of the features from the same class is reduced and the inter-class separability of the features from different classes is increased. Thus, hard samples can be correctly classified. In the test stage, a linear classifier (such as SVM) is used

[*] Corresponding author (email: hanzi.wang@xmu.edu.cn)

for aesthetic image classification.

Model and methodology. We employ the triplet net [7] in the proposed LMDE method. The triplet net uses three CNN networks, which share the same network structure and parameters (i.e., weights and biases) as feature extractors. The three CNN networks in the triplet net respectively take three samples from a triplet (consisting of an anchor input, a positive input and a negative input) as the input. Let x^a, x^p and x^n denote the anchor input, the positive input and the negative input, respectively. Let f denote the feature extractor. Given an input triplet x, its features are written as $f(x) \in R^d$, where d is the feature dimensionality. The triplet loss L_T of the triplet net is defined as [7]

$$L_T = \sum_{i=1}^{N} \max\big(0, \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \gamma\big), \quad (1)$$

where N denotes the number of triplets, i denotes the index of the triplet, and γ is the distance parameter that enforces the margin between a positive pair ($f(x^a)$ and $f(x^p)$) and a negative pair ($f(x^a)$ and $f(x^n)$). Minimizing the triplet loss encourages the triplet net to learn the embedding function $f(\cdot)$ (i.e., the feature extractor), where the distance between the positive pair is less than that between the negative pair plus the distance γ.

Based on the features $f(x_i)$ obtained by the triplet net, we design a two-class linear classifier $g(x_i, W, b)$ (abbreviated as $g(x_i)$) to evaluate the linear separability of the features. The classifier is defined as

$$g(x_i) \equiv g(x_i, W, b) = Wf(x_i) + b, \quad (2)$$

where W and b are the parameters of the linear classifier. The linear classifier can be trained by minimizing the hinge loss L_H, which is defined as

$$L_H = \sum_{i=1}^{3N} \max(0, \beta - g(x_i)y_i), \quad (3)$$

where y_i is either 1 or -1, indicating the class to which the sample x_i belongs; β denotes the geometric margin between support vectors and the linear classifier $g(\cdot)$. The value of L_H is zero if the sample x_i lies on the correct side of the geometrical margin. For the samples lying on the wrong side of the geometric margin, the loss values of L_H are proportional to their distance from the geometric margin. In our implementation, $f(x^a)$ and $f(x^p)$ are labelled as 1, and $f(x^n)$ is labelled as -1.

To obtain the embedding function $f(\cdot)$, where $f(x^a)$ and $f(x^p)$ can be separated from $f(x^n)$ by using the linear classifier $g(\cdot)$ with a large geometrical margin β, we minimize the following joint loss function L_J:

$$L_J = \lambda L_T + (1 - \lambda)L_H, \quad (4)$$

where λ is a weighting factor balancing the contribution of each loss. To train the LMDE method with the joint loss function, we use a linear layer (i.e., a fully connected layer) to implement the linear classifier. The linear layer takes the features obtained by the triplet net as its inputs. The outputs of the linear layer are measured by the hinge loss function, which is used to create a large margin between the features from different classes. The implementation details of the fully connected layer are the same as the traditional fully connected layer. The features obtained by the triplet net are directly fed into the fully connected layer. Then the whole net (including the triplet net and the linear layer) is trained in an end-to-end fashion.

The parameters (i.e, the weights and biases) of LMDE can be learned by employing the backpropagation algorithm [8]. The backpropagation algorithm calculates the partial derivatives of a loss function with respect to the weights and the biases in each layer of LMDE by using the chain rule, and it updates the weights and the biases by using a gradient descent method. As the hinge loss function is convex and not differentiable, we use the sub-gradient method [9] to calculate the partial derivatives of the hinge loss function with respect to the weights (W) and the biases (b) of the linear layer.

It is quite fast to train a CNN model with the hinge loss function or the standard softmax loss function. Inspired by the above observation, we propose a fast training strategy for training the triplet net of LMDE by initializing the weights and biases of the triplet net with those of a pre-trained CNN model, by which the training process of the proposed LMDE method is speeded up. An isolated CNN model, which has the same structure as the CNN nets in the triplet net, is firstly trained on large-scale visual aesthetic training datasets with the standard softmax loss function. Then, the triplet net is initialized with the weights and biases of the trained CNN model. Such an initialization strategy offers the advantage that better interclass separability can be obtained for the image features. Then, the backpropagation algorithm is used to fine-tune the LMDE model by minimizing the joint loss function, which further maximizes intra-class compactness and inter-class separability. The input triplets are randomly sampled from

Table 1 The classification accuracy obtained by CAP, CB and the proposed LMDE methods for aesthetic image classification on the CHUKPQ dataset[a)]

Method	Animal	Plant	Static	Architecture	Landscape	Human	Night	Overall
CAP [2] (%)	78.61	76.38	71.74	73.86	77.53	76.94	64.21	77.92
CB [1] (%)	89.37	91.82	90.69	**92.75**	94.68	**97.40**	84.63	92.09
LMDE (%)	**95.54**	**96.70**	**94.93**	92.43	**96.49**	95.35	**92.18**	**94.80**

a) The bold fonts represent the highest classification accuracy.

images in every S epoches to avoid the overfitting problem.

The evaluation of the proposed LMDE method consists of two steps: extracting features and classification. We use the triplet net to extract features from aesthetic images. The extracted features are then classified by using a linear classifier. In our implementation, the linear SVM is simply used as the linear classifier to classify the features, and is trained by using the default settings.

Experiments. We report the experimental results on the CHKUPQ dataset, which consists of seven classes of images: animal, plant, static, architecture, landscape, human and night. Table 1 reports the classification accuracy obtained by the CAP, CB and LMDE methods. As can be seen, the overall average classification accuracy obtained by the proposed LMDE method is 94.8%, which is the highest among the three competing methods and it is higher than those of CAP and CB by 16.9% and 2.7%, respectively. The CAP method obtains the worst average classification accuracy because it only uses several regional features (including color saturation, texture feature, and the depth of field feature), which are less effective for the task of aesthetic image classification. The CB method uses more effective features (including the regional features and the global features) than the CAP method. Thus, the CB method obtains better average classification accuracy than the CAP method. As the proposed LMDE method takes the advantage of CNN to directly learn the features, the proposed LMDE method obtains higher average classification accuracy than the CB method. More specifically, the proposed LMDE method respectively obtains 1%–5% improvements on the classes of animal, plant, static and landscape compared with the CB method, and it also achieves a comparative result to CB on the class of architecture. For more detailed experimental results, please refer to the supplemental materials.

Conclusion. We present an LMDE method with a novel network structure and an effective joint loss function, which takes advantage of both the triplet loss function and the hinge loss function. The minimization of the joint loss function ensures that the intra-class variability of the features belonging to the same class is reduced and the inter-class separability of the features from different classes is increased. As shown in the experiments, the proposed LMDE method significantly outperforms several other state-of-the-art aesthetic classification methods in terms of classification accuracy.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U1605252, 61872307, 61472334, 61571379), National Key R&D Program of China (Grant No. 2017YFB1302400), and UM Multi-Year Research (Grant No. MYRG2017-00218-FST).

Supporting information Experiments. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Tang X O, Luo W, Wang X G. Content-based photo quality assessment. IEEE Trans Multimedia, 2013, 15: 1930–1943

2 Datta R, Joshi D, Li J, et al. Studying aesthetics in photographic images using a computational approach. In: Proceedings of European Conference on Computer Vision, 2006. 288–301

3 Guo G J, Wang H Z, Shen C H, et al. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. IEEE Trans Multimedia, 2018, 20: 2073–2085

4 Pang Y W, Wang S, Yuan Y. Learning regularized LDA by clustering. IEEE Trans Neural Netw Learn Syst, 2014, 25: 2191–2201

5 Krizhevsky A, Sutskever I, Hinton G H. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1097–1105

6 Jiang X H, Pang Y W, Sun M L, et al. Cascaded subpatch networks for effective CNNs. IEEE Trans Neural Netw Learn Syst, 2018, 29: 2684–2694

7 Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 815–823

8 Rojas R. Neural Networks: A Systematic Introduction. Berlin: Springer, 1996

9 Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res, 2011, 12: 2121–2159