BUAA-iCC at ImageCLEF 2015 Scalable Concept Image Annotation Challenge

Yunhong Wang and Jiaxin Chen
Intelligent Recognition and Image Processing Lab, Beihang University, Beijing 100191, P.R.China
yhwang@buaa.edu.cn; chenjiaxinX@gmail.com.
http://irip.buaa.edu.cn/

Ningning Liu and Li Zhang
School of Information Technology and Management, University of International Business and Economics, Beijing 100029, P.R.China
ningning.liu@uibe.edu.cn

Abstract. In this working note, we mainly focus on the image annotation subtask of ImageCLEF 2015 challenge that BUAA-iCC research group participated. For this task, we firstly explore textual similarity information between each test sample and predefined concept. Subsequently, two different kinds of semantic information are extracted from visual images: visual tags using generic object recognition classifiers and visual tags relevant to human being related concepts. For the former information, the visual tags are predicted by using deep convolutional neural network (CNN) and a set of support vector machines trained on ImageNet, and finally transferred to textual information. For the latter visual information, human related concepts are extracted via face and facial attribute detection, and finally transferred to similarity information by using manually designed mapping rules, in order to enhance the performance of annotating human related concepts. Meanwhile, a late fusion strategy is developed to incorporate aforementioned various kinds of similarity information. Results validate that the combination of the textual and visual similarity information and the adopted late fusion strategy could yield significantly better performance.

Keywords: Textual similarity information, visual similarity information, deep convolutional neural network (CNN), face and facial attribute detection, late fusion, ImageCLEF

1 Introduction

For the ImageCLEF 2015 Scalable Concept Image Annotation Challenge [1, 2], we aim to develop a scalable image annotation approach, which could also yield high performance.

As shown in Fig.1, our proposed framework mainly consists of three components: exploration of textual similarity information between each testing sample

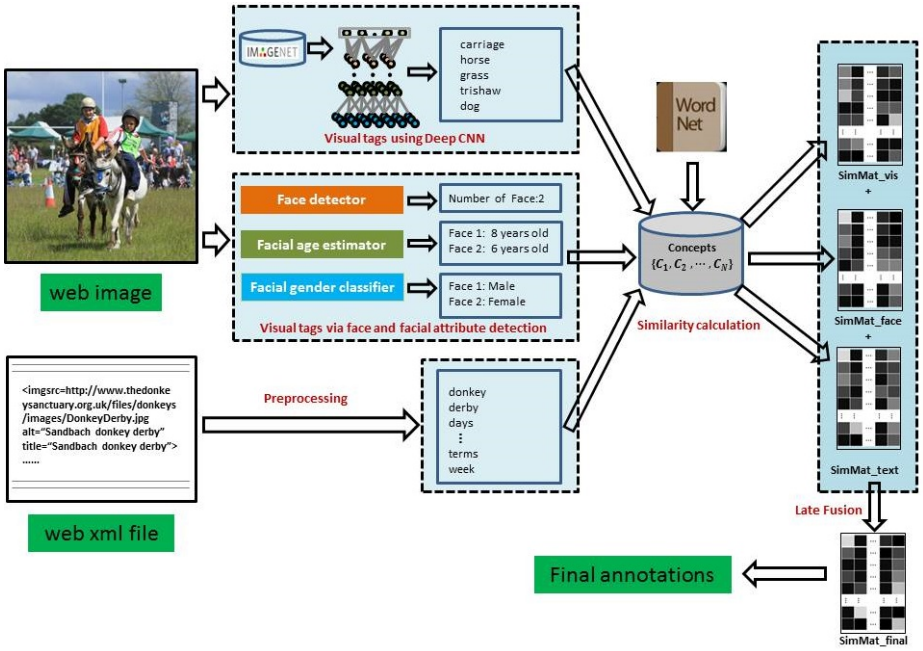

Fig. 1. The overall flowchart of the proposed framework.

and predefined concept, visual similarity information extraction and fusion of various kinds of similarity information.

For textual similarity information, we directly utilize the *path*[4] semantic distance that is based on WordNet[5] ontology to construct textual similarity matrix $SimMat\_text$, of which the element in row i, column j indicates the similarity score between the $i-th$ sample and the $j-th$ predefined concept.

For visual similarity information, two different kinds of semantical information from visual images are firstly extracted: visual tags using generic objection recognition classifier and human-being related visual tags. As to the former one, a deep convolutional neural network is trained to extract discriminative features, and a set of support vector machine are trained on on ImageNet [6], based on which each visual image from ImageCLEF 2015 testing data are tagged by using object categories with top 5 probabilities. Through this way, a given visual image is then transferred into textual semantic information, based on which one kind of visual similarity matrices $SimMat\_vis$ could be calculated by following the same method for constructing $SimMat\_text$. As to the latter one, we use existing face and facial attribute detectors [17] to obtain the following information from visual images: the number of faces, facial age and facial gender, in order to enhance the image annotation of human related concepts such as '*female-child*', '*man*' and '*woman*'. Subsequently, these face and facial attribute detection re-

sults are transferred into another kind of visual similarity matrix $SimMat\_face$ via manually designed mapping rules.

Each kind of similarity information could individually yield an image annotation result. However, in order to further enhance the performance of the proposed annotation method, a late fusion strategy is adopt to learn the optimal weight coefficient using develop set. Results demonstrate that the incorporation of the proposed textual and visual similarity information could boost the performance, and the late fusion strategy could further enhance the annotation accuracy compared with trivial fusion schemes.

The remainder of this working notes is organized as follows. Section 2 describes the details of the textual/visual similarity information extraction, and late fusion strategy of fusion multiple modal similarity information. In Section 3, we summarize the implementation details of the submitted runs, and demonstrate the experimental results together with the corresponding experimental analysis. Finally, in section 4, we draw the conclusion of this working note.

2 The proposed Framework

In this section, we will elaborate the details of each components. We firstly describe the textual similarity information exploration in Section 2.1, followed by the visual similarity information exploration via deep convolutional neural networks in Section 2.2. In section 2.3, we present the visual similarity information exploration by using face and facial attribute detection, and finally detail the late fusion strategy of fusing multiple sources of similarity information.

2.1 Textual Similarity Information Exploration

As provided by ImageCLEF 2015, each sample is composed by a visual image and corresponding web textual xml file. Textual similarity is therefore an important information for image annotation. In our submission, the textual similarity is explored by the strategy elaborated in Algorithm 1.

Specifically, given the textual description of the textual xml files of all 500000 test samples $\{W_i\}_{i=1}^{500000}$ and the predefined 251 concepts $\{D_j\}_{j=1}^{251}$, our goal is to calculate the similarity score s_{ij} between the $i-th$ sample W_i to the $j-th$ concept D_j, which finally consists of the similarity matrix $SimMat\_text$ with $SimMat\_text(i,j) = s_{ij}$. s_{ij} is calculated as the following:

$$s_{ij} = \sum_{k=1}^{N_i} \sum_{m=1}^{N_j} dist(w_{i,k}, d_{j,m}),$$

where $W_i = \{w_{i,k}\}_{k=1}^{N_i}$ and $D_j = \{d_{j,m}\}_{m=1}^{N_j}$.

In our implementation, we follow the method depicted in [3] and utilize $path$[4] distance that is based on the WordNet ontology[5] to measure the semantic similarity $dist(w_{i,k}, d_{j,m})$ between two synsets $w_{i,k}$ and $d_{j,m}$:

$$dist(w_{i,k}, d_{j,m}) = d_{path}(w_{i,k}, d_{j,m}) = \frac{1}{1 + spl(w_{i,k}, d_{j,m})},$$

where $spl(w_{i,k}, d_{j,m})$ returns the distance of the shortest path linking the two synsets (if one exists).

Finally, we normalized the similarity matrix $SimMat\_text$ to [0,1] as the following:

$$SimMat\_text(i,j) = (SimMat\_text(i,j) - v_{min})/(v_{max} - v_{min}),$$

where $v_{max} = max_{i,j} SimMat\_text(i,j)$ and $v_{min} = min_{i,j} SimMat\_text(i,j)$.

Algorithm 1 (Textual Similarity Computation)

Input: textual xml file $\{W_i\}_{i=1}^{500000}$ of the complete samples; textual description $\{D_j\}_{j=1}^{251}$ of predefined concepts

Output: A similarity matrix $SimMat\_text \in R^{500000 \times 251}$, of which the element in the $i-th$ row and $j-th$ column is the similarity score between the textual file of each sample and the predefined concept

Steps:

1. Preprocess $\{W_i\}_{i=1}^{500000}$ and $\{D_j\}_{j=1}^{251}$ by using a stop-words filter.

2. For each textual description of concept $i = 1$ to 50000

For each textual xml file of sample $j = 1$ to 251

$SimMat\_text(i,j) = \sum_{k=1}^{N_i} \sum_{m=1}^{N_j} dist(w_{i,k}, d_{j,m})$,

end

end

where $W_i = \{w_{i,k}\}_{k=1}^{N_i}$ and $D_j = \{d_{j,m}\}_{m=1}^{N_j}$, and $dist(w_{i,k}, d_{j,m})$ is the semantic similarity distance defined by WordNet.

3. Calculate the maximal value v_{max} and minimal value v_{min} of $SimMat\_text$, and normalize $SimMat\_text$: $SimMat\_text(i,j) = (SimMat\_text(i,j) - v_{min})/(v_{max} - v_{min})$.

2.2 Visual Similarity Exploration Based on Objection Recognition Via Deep Convolutional Neural Network

In the past few years, significant progress in generic visual object recognition has been achieved, by virtue of the availability of large scale datasets such as ImageNet [6] and advances in recognition algorithms such as deep learning [7, 8]. Current visual recognition systems based on deep learning are capable of recognizing thousands of object categories with promising accuracy. For instance, by using the deep convolutional neural network (CNN), He et al. [9] has reduced the visual recognition error rate on ImageNet 2012 dataset to 4.94%, which has amazingly surpassed human-level performance (with error rate 5.1% for comparison).

It is therefore reasonable to adopt deep CNN trained on ImageNet to help automatically annotate a visual image with a list of terms representing concepts depicted in the image. In our proposed framework, we follow the similar way as depicted in [11, 12] for visual objection recognition using deep CNN. Specifically, we use the 1,571,576 ImageNet images in the 1,372 synsets as our training set, and the 500,000 images in the image annotation task of imageCLEF 2015 as our test set. For images with different sizes, we uniformly wrapped all training and testing images into 256x256. For each image in both sets, we extracted activation of a pre-trained CNN model as its feature. The model is a reference

implementation of the structure proposed in Krizhevsky et al. [7] with minor modifications, and is made publicly available through the Caffe project [13].

Once the feature is extracted for both training and test sets, 1,372 binary classifiers are trained and applied using LIBSVM [14], which give probability estimates for the test images. For each image, the 1,372 classifiers are then ranked in order of their probability estimates. In order to reliably capture the semantic information contained in the test image, we only choose the categories with top 5 probabilities, through which visual images are finally transferred into textual tags $\{T_i\}_{i=1}^{500000}$. We therefore could construct visual similarity information $SimMat\_vis \in R^{500000\times 251}$ via the same way as the textual xml file, i.e., by using Algorithm 1 in Section 2.1.

2.3 Visual Similarity Exploration Via Face and Facial Attribute Detection

Human being is one of the most frequently occurred objects in visual images, of which face is one of the most representive and critical biometrics. Recent years have seen the substantial progress in face detection, face recognition together with facial attribute recognition. For example, on the largest unconstrained face dataset Label Face in the Wild (LFW) [15], the most state-of-the-art approach [16] has archived 99.63% accuracy using deep learning. Many open access cloud planform for face recognition have also merged, such as Face++ [17], which could provide free API services for face and facial attributes (such as age, gender, race and etc.) detection.

In our framework, we assume that the face and facial attribute detection could explore useful semantic information from visual images. Considering its promising performance [18, 19] and open access, we adopt Face++ as a tool for face and facial attribute detection, and finally utilize the following three results to enhance performance of visual image automatic annotation:

1). FA_{num_face}: number of face detected.

2). FA_{age}: age of each detected face.

3). FA_{gender}: gender of each detected face.

It should be noted that $FA_{num\_face} \in N, FA_{age} \in N, FA_{gender} \in \{$'*female*', '*male*'$\}$, where FA_{num_face} and FA_{age} are numeric variables, and could not be directly used for image annotation. Here, we manually design a mapping from FA_{num_face}, FA_{age}, FA_{gender} to the similarity score between each visual image and the predefined 251 concepts.

Specifically, we firstly select a concept subset C_{14} containing 14 concepts related to human being from the complete 251 concept set C_{All}: $C_{14} =$ { '*arm*', '*eye*', '*face*', '*female-child*', '*foot*', '*hair*', '*head*', '*leg*', '*male*', '*man*', '*mouth*', '*neck*', '*nose*', '*woman*' }. For the $i-th$ input visual image, we obtain FA_{num_face}, $FA_{age}$and FA_{gender} using Face++. Subsequently, we calculate the 251 dimensional vector s_i describing the similarities between the visual image and C_{All} as the following: if $C_{All}(j) \notin C_{14}$, $s(i,j) = 0$; if $C_{All}(j) \in C_{14}$, $s(i,j)$ is evaluated according to FA_{num_face}, FA_{age}, FA_{gender} and the mapping rules described in

Table 1. For instance, assuming that the face and facial attribute detection results of the $i-th$ visual image are $FA^{(i)}_{num\_face}$, $FA^{(i)}_{age}$, $FA^{(i)}_{gender}$ and the $j-th$ concept $C_{all}(j)$ is '*men*', $s(i,j)=1$, if $FA^{(i)}_{num\_face}>0$ and $FA^{(i)}_{gender}=' male'$, and $s(i,j)=0$ otherwise. It is worth noting that different concepts in C_{14} are assigned different values according to their prior tightness to facial attributes. Considering that { '*eye*', '*face*', '*female-child*', '*male*', '*man*', '*mouth*', '*nose*', '*woman*' } could be determined by the face and facial attribute with high confidence, they are assigned the highest similarity score 1 once the condition is satisfied as described in Tabel 1. { '*arm*', '*hair*', '*head*', '*leg*', '*neck*' } are assigned similarity score 0.8, since they are less closely related to facial attributes. { '*foot*' } is evaluated 0.6, which is further less relevant to facial attributes.

For each visual images of the 500000 testing samples, we could calculate a 251 dimensional similarity score vector, which finally consists the similarity score matrix $SimMat\_face \in R^{500000\times 251}$.

Table 1. The mapping rules of transferring face and facial detection results to similarity scores between input visual image and selected 14 human related concepts.

Concept	Description of mapping rule
'arm'	$\begin{cases} 0.8, & if\ FA_{num\_face}>0; \\ 0, & if\ FA_{num\_face}=0 \end{cases}$
'eye'	$\begin{cases} 1, & if\ FA_{num\_face}>0; \\ 0, & if\ FA_{num\_face}=0. \end{cases}$
'face'	$\begin{cases} 1, & if\ FA_{num\_face}>0; \\ 0, & if\ FA_{num\_face}=0. \end{cases}$
'female-child'	$\begin{cases} 1, & if\ FA_{num\_face}>0\ and\ FA_{age}\geq 18\ and\ FA_{gender}=' female'; \\ 0, & otherwise. \end{cases}$
'foot'	$\begin{cases} 0.6, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'hair'	$\begin{cases} 0.8, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'head'	$\begin{cases} 0.8, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'leg'	$\begin{cases} 0.8, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'male'	$\begin{cases} 1, & if\ FA_{num\_face}>0\ and\ FA_{gender}==' male'; \\ 0, & otherwise. \end{cases}$
'man'	$\begin{cases} 1, & if\ FA_{num\_face}>0\ and\ FA_{age}>18\ and\ FA_{gender}=' male'; \\ 0, & otherwise. \end{cases}$
'mouth'	$\begin{cases} 1, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'neck'	$\begin{cases} 0.8, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'nose'	$\begin{cases} 1, & if\ FA_{num\_face}>0; \\ 0, & otherwise. \end{cases}$
'woman'	$\begin{cases} 1, & if\ FA_{num\_face}>0\ and\ FA_{age}>18\ and\ FA_{gender}=' female'; \\ 0, & otherwise. \end{cases}$

2.4 Late Fusion of Various Similarity Information

Until now, we have obtained three different kinds of similarity matrices: $SimMat\_text$, $SimMat\_vis$, $SimMat\_face$, each of which could yield an individual image annotation result. For instance, we could sort each row of $SimMat\_text$, and select the concepts with top k similarity scores as the final annotations. However, $SimMat\_text$, $SimMat\_vis$, $SimMat\_face$ are three different sources of similarity information between each sample and concept. It could be expected that better performance could be archived by fusing these three similarity matrices.

In our submission, we adopt similar late fusion scheme proposed by [3]. Generally, given K different similarity matrix $\{SimMat\_i\}_{i=1}^{K}$, the overall similarity matrix $SimMat\_final$ is a weighted sum of $\{SimMat\_i\}_{i=1}^{K}$ as follows:

$$SimMat\_final = \sum_{i=1}^{K} w_i * SimMat\_i,$$

where $\sum_{i=1}^{K} w_i = 1$ and $w_i \geq 0$.

In our implementation, the optimal weights $\{w_i\}_{i=1}^{K}$ are determined by using the Selective Weighted Late Fusion (SWLF) algorithm [3] on the develop set with 1980 annotated samples.

After obtained the final similarity matrix $SimMat\_final$, we could assign the label to each sample. In this submission, we mainly adopt two different schemes:

Annotation scheme 1: we select concepts with top N similarity scores as the final annotations;

Annotation scheme 2: we select concepts, of which the similarity score is greater than the given threshold T, as the final annotations.

The optimal N and T could be chosen by maximizing the Mean Average Precision (MAP) on the develop set.

3 BUAA-iCC Runs and Experimental Results

3.1 Description of Submissions

We submitted total ten runs, which are differ from similarity information fused, fusion strategies, and annotation schemes. The brief description of each submission are summarized as follows:

IRIP-iCC_01: SimMat_final=SimMat_vis; adopt Annotation scheme 1, where $N = 6$ is chosen by maximizing MAP on develop set.

IRIP-iCC_02: SimMat_final=SimMat_text; adopt Annotation scheme 1, where $N = 6$ is chosen by maximizing MAP on develop set.

IRIP-iCC_03: SimMat_final=SimMat_text+SimMat_vis+SimMat_face; adopt Annotation scheme 1, where $N = 6$ is chosen by maximizing MAP on develop set.

IRIP-iCC_04: SimMat_final=SimMat_text+SimMat_vis+SimMat_face; adopt Annotation scheme 1, where $N = 7$ is chosen manually.

IRIP-iCC_05: SimMat_final=SimMat_text+SimMat_vis+SimMat_face; adopt Annotation scheme 1, where $N = 251$ is chosen manually.

IRIP-iCC_06: SimMat_final=w*(SimMat_text+SimMat_vis)+(1−w)*SimMat_face, where $w = 0.6$ is chosen by SWLF; adopt Annotation scheme 2, where $T = 0.5$ is chosen manually.

IRIP-iCC_07: SimMat_final=w*(SimMat_text+SimMat_vis)+(1−w)*SimMat_face, where $w = 0.6$ is chosen by SWLF; adopt Annotation scheme 1, where $N = 6$ is chosen by maximizing MAP on develop set.

IRIP-iCC_08: SimMat_final=w*(SimMat_text+SimMat_vis)+(1−w)*SimMat_face, where $w = 0.6$ is chosen by SWLF; adopt Annotation scheme 1, where $N = 7$ is chosen manually.

IRIP-iCC_09: SimMat_final=w*(SimMat_text+SimMat_vis)+(1−w)*SimMat_face, where $w = 0.6$ is chosen by SWLF; adopt Annotation scheme 2, where $T = 0.6$ is chosen by maximizing MAP on develop set.

IRIP-iCC_10: SimMat_final=w*(SimMat_text+SimMat_vis)+(1−w)*SimMat_face, where $w = 0.6$ is chosen by SWLF; adopt Annotation scheme 2, where $T = 0.4$ is chosen manually.

3.2 Results of Submitted runs

As described in [1, 2], the annotation accuracy together with the localization precision are evaluated by Mean Average Precision (MAP) with α overlap. For instance, $mAP\_0\_overlap$ stands for MAP without considering localization overlap, and $mAP\_0.5\_overlap$ stands for MAP with o.5 localization overlap. In our submission, we mainly focus on evaluate the performance of the proposed framework on annotation accuracy. So we mainly analyze the experimental results of $mAP\_0\_overlap$. For $mAP\_0.5\_overlap$, we simply use the objectness detector proposed in [20] for concept localization.

The experimental results are shown in Table 2. From submissions IRIP-iCC_01, IRIP-iCC_02 and IRIP-iCC_03, we can see that visual similarity by using CNN performs better than textual similarity information extracted from xml file, and fusion of visual similarity information and textual similarity information could significantly boost the performance of each single similarity information. It also could be seen that IRIP-iCC_06 and IRIP-iCC_09 yield top 1 and top 2 mAPs, respectively, indicating that the weight w via late fusion strategy using SWLF could enhance the performance. The mAP of IRIP-iCC_09 is about 2% higher than IRIP-iCC_06, which could verify that the performance could be further boosted by choosing optimal threshold T via develop set.

4 Conclusion

In this paper, we described the participation of BUAA-iCC at ImageCLEF 2015 Scalable Concept Image Annotation Challenge. We proposed a novel image annotation framework by fusing textual similarity information, and visual similarity information explored by deep convolutional neural network (CNN), face and facial attribute detection.

Table 2. The results of our submitted runs.

Submitted runs	mAP_0_overlap(%)	mAP_0.5_overlap(%)
IRIP-iCC_01	22.98	10.72
IRIP-iCC_02	16.62	7.67
IRIP-iCC_03	51.40	**14.58**
IRIP-iCC_04	51.56	13.54
IRIP-iCC_05	43.03	4.64
IRIP-iCC_06	58.95	11.15
IRIP-iCC_07	50.65	14.43
IRIP-iCC_08	50.91	13.44
IRIP-iCC_09	**60.92**	11.96
IRIP-iCC_10	51.07	9.00

Experimental results reveals that the visual similarity information extracted by deep CNN, face and facial attribute detection could enhance the performance of the textual similarity information extracted from xml files. The similarity information fusion strategy using selective weighted late fusion could significantly boosts the performance. The annotation scheme by selecting concepts with similarity score larger than an automatically determined threshold, which maximizes the MAP of samples from develop set, yields better performance than other annotation schemes.

Acknowledgement

This work was supported in part by the HongKong, Macao and Taiwan Science & Technology Cooperation Program of China under the grant L2015TGA9004, the Fundamental Research Funds for the Central University in UIBE under grant 14QD21.

References

1. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., Herrera, A., Bromri, S., Amin, M., Mohammed, M., Acar, B., Uskudarli, S., Marvasti, N., Aldana, and J., García., M.: General Overview of ImageCLEF at CLEF2015 Labs. Lecture Notes in Computer Science, Springer International Publishing,(2015).
2. Gilbert, A., Piras, L., Wang, J., and et al.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CEUR Workshop Proceedings (2015).
3. Liu, N., Dellandrea, E., Chen, L., Zhu, C., Zhang, Y., Bichot, C. E., Bras, S., and Tellez, B.: Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme.Computer Vision and Image Understanding, 117(5), (2013) 493-512.
4. Budanitsky, A., and Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources (Vol. 2) (2001).

5. Miller, G. A.: WordNet: a lexical database for English.Communications of the ACM, 38(11), (1995) 39-41.
6. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE (2009) 248-255.
7. Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (2012) 1097-1105.
8. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. arXiv preprint arXiv:1409.4842 (2014).
9. He, K., Zhang, X., Ren, S., and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. arXiv preprint arXiv:1502.01852 (2015).
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge (ILSVRC). arXiv:1409.0575 (2014).
11. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T.L: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013).
12. Wang, J. K., Yan, F., Aker, A., and Gaizauskas, R.: A Poodle or a Dog? Evaluating Automatic Image Annotation Using Human Descriptions at Different Levels of Granularity. V&L Net (2014).
13. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding. http://caffe. berkeleyvision. org (2013).
14. Chang, C. C., and Lin, C. J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3) (2011) 27.
15. Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E: Labeled faces in the wild: A database for studying face recognition in unconstrained environments (Vol. 1, No. 2, p. 3). Technical Report 07-49, University of Massachusetts, Amherst (2007).
16. Schroff, F., Kalenichenko, D., and Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. (2015),
17. Megvii Inc.: Face++ Research Toolkit. http: //www.faceplusplus.com, December 2013.
18. Zhou, E., Cao, Z., and Yin, Q.: Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?. arXiv preprint arXiv:1501.04690 (2015).
19. Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C.: Learning deep face representation. arXiv preprint arXiv:1403.2802 (2014).
20. Alexe, B., Deselaers, T., and Ferrari, V.: Measuring the objectness of image windows. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(11), 2189-2202 (2012).