Image Memorability Using Diverse Visual Features and Soft Attention

Marco Leonardi[1], Luigi Celona[1](✉), Paolo Napoletano[1], Simone Bianco[1],
Raimondo Schettini[1], Franco Manessi[2], and Alessandro Rozza[2]

[1] University of Milano - Bicocca, Milan, Italy
m.leonardi6@campus.unimib.it,
{luigi.celona,paolo.napoletano,simone.bianco,schettini}@unimib.it
[2] lastminute.com, Chiasso, Switzerland
{franco.manessi,alessandro.rozza}@lastminute.com

Abstract. In this paper we present a method for still image memorability estimation. The proposed solution exploits feature maps extracted from two Convolutional Neural Networks pre-trained for object recognition and memorability estimation respectively. The feature maps are then enhanced using a soft attention mechanism in order to let the model focus on highly informative image regions for memorability estimation. Results achieved on a benchmark dataset demonstrate the effectiveness of the proposed method.

Keywords: Memorability · Residual Neural Network · Convolutional Neural Network · Deep learning

1 Introduction

A remarkable feature of human cognition is the ability to remember different images that have been seen only once [14]. Furthermore, different people tend to remember or forget same pictures. This result suggests that people encode and discard very similar types of information. Precisely, images that are usually forgotten seem to lack distinctiveness and a fine-grained representation in human memory [14]. Taking into account the aforementioned considerations, it seems that memorable images have some kind of intrinsic visual features, making them easier to remember. Indeed, past studies have shown that memorability is a measurable stationary property of an image shared across different viewers [9] and that it is possible to determine a compact set of attributes characterizing the memorability of any individual image [8]. These results led researchers wondering how to predict accurately which images will be remembered and which will be not, resulting in the first large scale visual memorability estimation with near-human performance [11].

Nowadays, we are continuously being exposed to photographs when browsing the Internet or leafing through a magazine. Exploiting memorable pictures can have a huge impact in many applications, also thanks to the relationship

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11752, pp. 171–180, 2019.
https://doi.org/10.1007/978-3-030-30645-8_16

between emotions and memorability [3]. Just to give some examples, estimating the memorability can help to automatically select the images that can have a key role in optimizing the conversion rate for media advertisement and online shopping, or in improving the communication of a specific concept. More recently, researchers started to show interest in how to make an image more memorable, by exploiting deep architectures for generating memorable pictures by exploiting style-transfer techniques [24].

In this work, we propose a novel approach to compute memorability that exploits the combination of feature computed by different Convolutional Neural Networks (CNNs) and an attention map extracted from a caption generation model with visual attention. In details, the main contributions of the approach presented in this paper are:

- it achieves comparable or better results with respect to state-of-the-art approaches, respectively in terms of Spearman's rank correlation and Mean Squared Error (MSE);
- it reduces the amount of parameters with respect to the best performing technique in terms of Spearman's rank correlation.

The paper is organized as follows: in Sect. 2 the related works are summarized; in Sect. 3 the proposed method is described; in Sect. 4 the experimental results are presented; finally, conclusions and future works are summarized in Sect. 5.

2 Related Works

In the first works on image memorability, Isola *et al.* [8,9] showed the ability of our mind to remember certain images better than others and also that memorability is a stable property of an image shared across different viewers. They introduced a database for which they collected the probability that each image will be remembered after a single view as well as image attribute annotations (such as spatial layout, content and aesthetic properties) in order to:

- understand which features are highly informative about memorability;
- demonstrate that memorability is not influenced by content frequency or familiarity, namely the presence of particular objects, scene categories, relatives or famous monuments. However, some contents like faces are memorable, while vistas and peaceful settings are not;
- prove that memorability is not correlated with aesthetics, interestingness, and simple image features.

Furthermore, they developed a method to predict the memorability of an image involving the use of Support Vector Regressor machines on the combination of global image features – GIST [19], SIFT [15], HOG [4], SSIM [23], and color histogram. Following the intuition of Isola *et al.* [8] that memorability and visual attention are correlated, Mancas and Le Meur [17] demonstrated that attention-related features can effectively replace some of the low-level features used by Isola *et al.* [9] and thus reducing the dimensionality of the feature set. Afterwards, Bylinskii *et al.* [2] proved that the interplay between intrinsic image properties (the fact that some scene categories are more memorable than others) and extrinsic factors, such as image context and observer behavior, are necessary to build an improved image memorability model. The effectiveness of the proposed solution has been assessed on FIne-GRained Image Memorability (FIGRIM) dataset that is composed by more than 9K images.

Khosla *et al.* [11] released LaMem, the first large scale dataset for image memorability containing 60K images. Alongside the dataset, they proposed MemNet, a CNN for memorability score estimation. The model is based on the fine-tuning of Hybrid-CNN [28], a CNN trained using 3.5 million images from 1,183 categories, obtained by merging the scene categories from Places database [28] and the object categories from ImageNet [22]. They achieved near human consistency rank correlation (0.68) for memorability. Fajtl *et al.* [5] proposed AMNet, a model consisting of a ResNet50 [7] pre-trained on ImageNet, a soft attention mechanism, and a Long Short-Term Memory [6] for memorability score regression. The AMNet model achieved a performance of 0.677 in terms of Spearman's rank correlation on LaMem dataset. Recently, Squalli-Houssaini *et al.* [26] approached the task of image memorability estimation as a classification problem instead of a regression one. They developed a model combining features extracted from both a VGG16 [25] pre-trained on ImageNet and an image captioning system [13] and outperformed both state-of-the-art and human consistency correlation (0.72) on LaMem dataset.

3 Proposed Method

Image memorability is influenced by some intrinsic image properties, namely *what* kind of objects and scenes are present and *what* are their characteristics, but also by extrinsic factors such as the image locations *where* humans focus their attention. Our approach tries to model memorability according to the aforementioned aspects by using a CNN for encoding intrinsic characteristics of objects, and a soft attention mechanism for estimating attention maps that highlight salient regions. Furthermore, we include in the proposed model a CNN pre-trained on image memorability for mapping *how* features encode memorability.

3.1 Architecture

The proposed model, depicted in Fig. 1, estimates a memorability score given as input an RGB image of size 256 × 256 pixels. It consists of two CNNs trained on two different tasks, and a soft attention mechanism based on a system originally designed for caption generation [27]. The aforementioned blocks (i.e. soft attention and memorability) are followed by two convolution layers preceding the last regressor module, which estimates the memorability score.

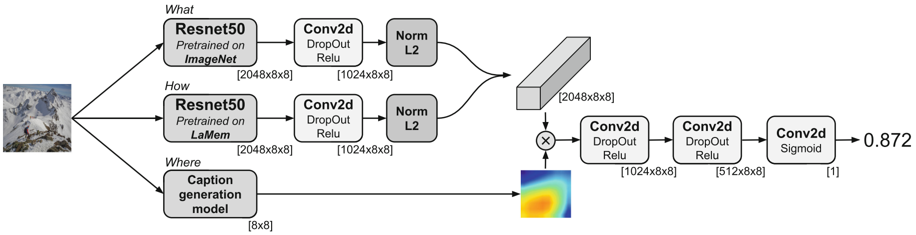


Fig. 1. Overview of the proposed model for image memorability estimation. The attention map produced by the caption generation model is combined channel-wise with the feature volume.

Feature Extraction. The two considered CNNs are two ResNet50 architectures pre-trained respectively for: image memorability estimation on LaMem dataset [11] and object recognition on Imagenet [22] dataset. We consider these two CNNs to provide the model prior information over the memorability of the image as well as knowledge of the image context. Both the architectures are truncated before their last average pooling layer in order to obtain two feature maps of size $2048 \times 8 \times 8$. These feature maps are first passed through a convolution layer which halves their channel dimension, then they are L2-normalized by dividing the feature map by its L2-norm, and finally stacked together obtaining a new feature map having a dimension equal to $2048 \times 8 \times 8$.

Soft Attention Mechanism. To focus the model attention on salient regions that are highly informative for memorability estimation, we include a state-of-the-art captioning generation approach [27] for extracting attention maps. This model is trained on the MS COCO dataset [16] and produces at most 50 attention maps with spatial size 8×8 pixels, each one focusing on a particular detail of the image. We exploit these maps by averaging them in order to get a single and global attention map.

Memorability Estimation. The feature map extracted from the two CNNs is weighted with the attention map generated from the captioning model replicated channel-wise. Finally, the resulted weighted feature map is given as input to a three-layer CNN to predict the memorability score.

3.2 Training Procedure

In order to improve the generalization of the model and minimize the risk of overfitting, we use data augmentation techniques during the training phase. Specifically, random scaling in the range $[0.8, 1.2]$ is first applied to the image, which then is randomly flipped along the vertical axis. Subsequently, random crop (0.8 to 1.0) of the image is applied before sub-sampling it to a size of 256×256 pixels. Finally, the image is normalized by subtracting and dividing each image by the mean and standard deviation estimated on the ImageNet training set [22] in order to limit the variability of the input range.

The training procedure consists of two phases. We first train one ResNet50 from scratch on LaMem [11] dataset for image memorability. Then we fine-tune the whole model on the same dataset freezing the weights of the two ResNet50 and the weights of the caption generation model with visual attention [27]. Both of the training processes are trained to minimize the mean squared error between the ground-truth and the predicted image memorability scores. For the first stage, we train the model for 150 epochs due to a larger number of parameters to learn, with a batch size of 10 images. For the second phase, the model is trained for only 50 epochs with a bigger batch size of 16 images.

During both the training processes, we use the technique of early stopping analyzing the Spearman's rank correlation see Sect. 4.2 for the definition) on the validation set. For both stages we use the ADAM optimizer [12] with starting learning rates respectively of 5×10^{-7} and 5×10^{-5} for the first and the second stage. Both the learning rates are decreased every epoch as follows:

$$LR(epoch) = \left[1 - \left(\frac{epoch}{total\ epochs}\right)^{0.9}\right] * LR_0, \tag{1}$$

where *epoch* is the 0-based index of the actual epoch, LR_0 is the initial learning rate, and *total epochs* is the total number of epochs for the training process.

4 Experiments

In the following sections, the dataset and metrics adopted for evaluating the proposed method are described. Experimental results are then reported. We develop the proposed approach using the PyTorch framework [20], and we run experiments on a NVIDIA GTX 1070 GPU.

4.1 Dataset

We evaluate our model on the LaMem dataset [11], a massive collection of 58,741 images annotated with a memorability score. The images were sampled from different existing datasets and cover various indoor and outdoor scenes. Figure 2 shows some samples from the dataset. The provided memorability score were collected on Amazon Mechanical Turk using an improved version of the memorability game introduced in [9]. The data are divided into five random training, validation and test set splits. Each of these splits has respectively 45k images as training set, 3741 as validation and 10k as test set. For each split, training and validation sets are labeled from the same group of people while the test is labeled from a different group.

4.2 Evaluation Metrics

Following the previous work [11], we evaluate the performance of our method using the Spearman's rank correlation coefficient, ρ, [21] and the Mean Squared

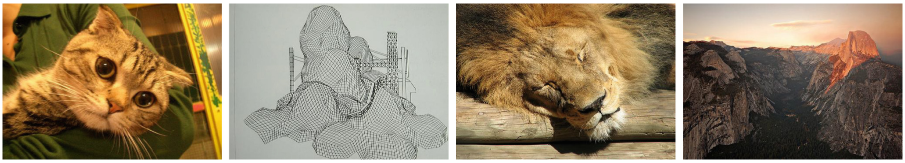


Fig. 2. Sample images from the LaMem dataset [11].

Error (MSE). The Spearman's rank correlation coefficient is a value, ranging from −1 to +1, which measures the monotonic relationships between the predicted and ground-truth ranking. A value of ρ equal to zero indicates no correlation between the two variable while values close to ±1 indicate relatively a strong positive (+1) or strong negative (−1) correlation. Spearman's rank correlation coefficient is defined as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}, \tag{2}$$

where d_i is the difference between the two ranks of each variable and N is the number of samples.

The MSE measures the goodness of fit between reference and observations in terms of absolute numerical errors as shown in the following equation:

$$MSE(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y_i} - y_i)^2, \tag{3}$$

where $\hat{y_i}$ are the ground-truth values for the memorability, y_i the predicted memorability score and N is the number of samples.

4.3 Results

In this subsection we evaluate the performance of the proposed model by averaging both the Spearman's rank correlation and MSE over the five splits of LaMem dataset. The proposed model reaches an average rank correlation of 0.687 and a MSE of 0.0079 over the five splits of LaMem.

Table 1. Results of the ablation study on the LaMem dataset reported in terms of Spearman's rank correlation (ρ) and Mean Squared Error (MSE).

Method	$\rho \uparrow$	MSE ↓
ResNet50-LaMem	0.680	0.0083
ResNet50-LaMem + ResNet50-ImageNet	0.686	0.0080
Whole model	0.687	0.0079

In Table 1, we report the results of an ablation study investigating how each module of the proposed model affects the overall performance. In particular, a single ResNet50 [7] trained on the task of image memorability achieves a Spearman's rank correlation of 0.680 and a MSE of 0.0083. The model involving the combination of the feature maps extracted from the two ResNet50 without the use of the attention map increases the correlation by 0.006 and lowers the MSE by 0.0003. Finally, we can see that the whole model, i.e. the addition of the soft attention mechanism, increases performance by 0.001 for the Spearman's rank correlation and decreases the MSE by 0.0001.

In Table 2 we compare the proposed method with respect to the state-of-the-art on LaMem [11] dataset. We report the performance provided in terms of correlation and MSE as the average results over the five dataset splits. From the results reported in Table 2 we can observe that in terms of Spearman's rank correlation, our model performs slightly worse with respect to the best state-of-the-art model [26]. Given that Squalli *et al.* [26] do not provide the MSE, we have implemented their solution and obtained an error of 0.00923. Based on this result, our approach reduces the MSE by 0.0013 using a number of parameters equal to less than half of those used by [26]. We conduct an analysis of the efficiency of proposed solution respect to previous methods. To this end, in Fig. 3a we compare the Spearman's rank correlation and the number of parameters, while in Fig. 3b we plot the MSE and the number of parameters. Among the methods that outperform the human consistency correlation (0.68), our model achieves lower performance by using a reduced amount of parameters. Instead in terms of MSE, the proposed method is the solution exploiting more efficiently its parameters by obtaining the smallest MSE with the fewest parameters. In Fig. 4 we show samples from LaMem dataset with memorability scores estimated by the proposed solution as well as ground-truth memorability scores. Furthermore, we provide the corresponding attention maps for each image to highlight how these maps in most cases focus on the relevant subjects in the scenes.

Table 2. Comparison with state-of-the-art methods in terms of Spearman's rank correlation and MSE on the LaMem dataset. For each model the number of its parameters (in millions) is also reported.

Method	$\rho \uparrow$	MSE $\downarrow$	# parameters
AMNet [5]	0.677	0.0082	**39M**
MemNet [11]	0.640	N/A	62M
Squalli *et al.* [26]	**0.720**	0.0092*	280M
Ours	0.687	**0.0079**	130M

*Estimated by the authors.

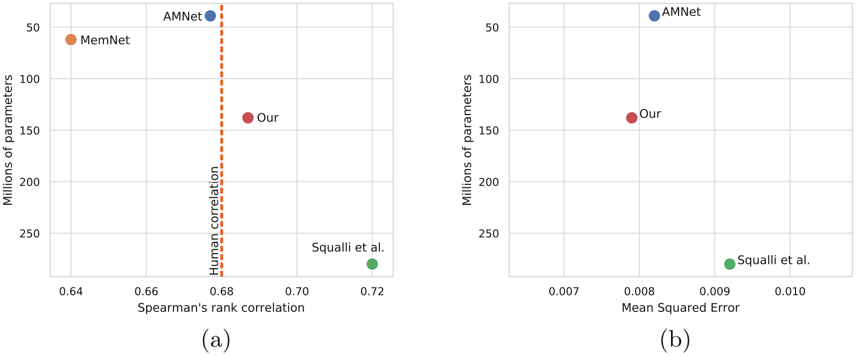
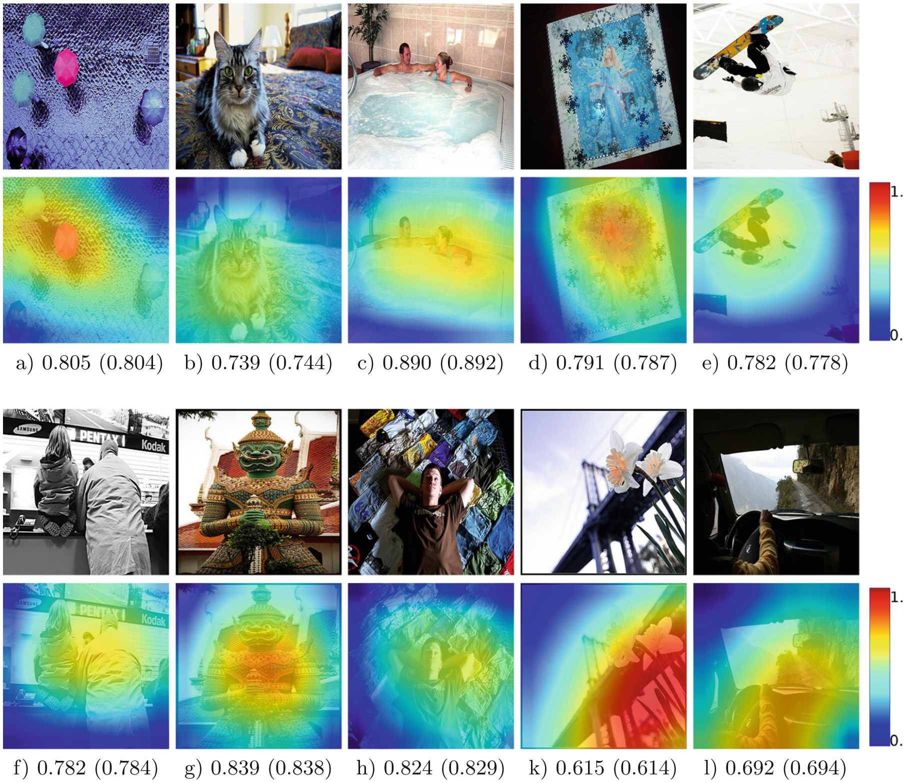


Fig. 3. Spearman's rank correlation vs. model parameters (the dashed line depicts the human consistency rank correlation [10]) (a). MSE vs. model parameters (b).



a) 0.805 (0.804) b) 0.739 (0.744) c) 0.890 (0.892) d) 0.791 (0.787) e) 0.782 (0.778)

f) 0.782 (0.784) g) 0.839 (0.838) h) 0.824 (0.829) k) 0.615 (0.614) l) 0.692 (0.694)

Fig. 4. Sample images from LaMem dataset with estimated and ground-truth (in brackets) memorability scores. Below each image its depicted the related visual attention map produced by the caption generation model.

5 Conclusion

This work presents a deep learning-based model for image memorability estimation. The proposed approach involves the use of two CNNs trained respectively on image recognition and image memorability. We use the features extracted from these two CNNs in order to exploit the knowledge of the context as well as the information about the memorability of the image. Moreover, we use a soft attention mechanism to focus the model attention on highly informative regions for memorability estimation. Results obtained on the LaMem benchmark dataset are comparable with respect to state-of-the-art approaches demonstrating the effectiveness of the proposed method. Moreover, our solution achieves the smallest MSE with the fewest parameters among the methods that outperform the human consistency correlation (0.68).

As a possible future work, we would like to experiment the approach proposed in [18] to learn combinations of base activation functions (such as the identity function, ReLU, and TanH), thus to improve the overall performance. Furthermore, we would investigate alternative attention mechanisms based on other saliency methods such as [1].

References

1. Bianco, S., Buzzelli, M., Schettini, R.: Multiscale fully convolutional network for image saliency. J. Electron. Imaging **27**, 27 (2018)
2. Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., Oliva, A.: Intrinsic and extrinsic effects on image memorability. Vis. Res. **116**, 165–178 (2015)
3. Cahill, L., McGaugh, J.L.: A novel demonstration of enhanced memory associated with emotional arousal. Conscious. Cogn. **4**(4), 410–421 (1995)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893. IEEE (2005)
5. Fajtl, J., Argyriou, V., Monekosso, D., Remagnino, P.: AMNet: memorability estimation with attention. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6363–6372. IEEE (2018)
6. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016)
8. Isola, P., Parikh, D., Torralba, A., Oliva, A.: Understanding the intrinsic memorability of images. In: Advances in Neural Information Processing Systems, pp. 2429–2437 (2011)
9. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable? In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 145–152. IEEE (2011)
10. Khosla, A., Das Sarma, A., Hamid, R.: What makes an image popular? In: International Conference on World Wide Web, pp. 867–876. ACM (2014)

11. Khosla, A., Raju, A.S., Torralba, A., Oliva, A.: Understanding and predicting image memorability at a large scale. In: International Conference on Computer Vision (ICCV), pp. 2390–2398. IEEE (2015)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2015)
13. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
14. Konkle, T., Brady, T.F., Alvarez, G.A., Oliva, A.: Scene memory is more detailed than you think: the role of categories in visual long-term memory. Psychol. Sci. **21**(11), 1551–1556 (2010)
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 2169–2178. IEEE (2006)
16. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
17. Mancas, M., Le Meur, O.: Memorability of natural scenes: the role of attention. In: International Conference on Image Processing (ICIP), pp. 196–200. IEEE (2013)
18. Manessi, F., Rozza, A.: Learning combinations of activation functions. In: International Conference on Pattern Recognition (ICPR), pp. 61–66 (2018)
19. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
20. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
21. Pirie, W.: Spearman rank correlation coefficient, vol. 8, August 2006
22. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
23. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, vol. 2, p. 3 (2007)
24. Siarohin, A., Zen, G., Majtanovic, C., Alameda-Pineda, X., Ricci, E., Sebe, N.: How to make an image more memorable?: A deep style transfer approach. In: International Conference on Multimedia Retrieval (ICMR), pp. 322–329. ACM (2017)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Squalli-Houssaini, H., Duong, N.Q., Gwenaëlle, M., Demarty, C.H.: Deep learning for predicting image memorability. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2371–2375. IEEE (2018)
27. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML), pp. 2048–2057 (2015)
28. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)