Event Recognition Based on Classification of Generated Image Captions

Andrey V. Savchenko^{1,2(✉)} and Evgeniy V. Miasnikov¹

¹ Samsung-PDMI Joint AI Center, St. Petersburg Department of Steklov Institute of Mathematics, Fontanka Street, St. Petersburg, Russia

² National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia
avsavchenko@hse.ru

Abstract. In this paper, we consider the problem of event recognition on single images. In contrast to conventional fine-tuning of convolutional neural networks (CNN), we proposed to use image captioning, i.e., a generative model that converts images to textual descriptions. The motivation here is the possibility to combine conventional CNNs with a completely different approach in an ensemble with high diversity. As event recognition task has nothing serial or temporal, obtained captions are one-hot encoded and summarized into a sparse feature vector suitable for the learning of an arbitrary classifier. We provide the experimental study of several feature extractors for Photo Event Collection, Web Image Dataset for Event Recognition and Multi-Label Curation of Flickr Events Dataset. It is shown that the image captions trained on the Conceptual Captions dataset can be classified more accurately than the features from an object detector, though they both are obviously not as rich as the CNN-based features. However, an ensemble of CNN and our approach provides state-of-the-art results for several event datasets.

Keywords: Image captioning · Event recognition · Ensemble of classifiers · Convolutional neural network (CNN)

1 Introduction

Nowadays, social networks and mobile devices create a vast stream of multimedia data because people are taking more photos in recent years than ever before [1]. To organize a large gallery of personal photos, they may be assigned to albums according to some events. Social events are happenings that are attended and shared by the people [2,3] and take place in a specific environment [4], e.g., holidays, sports events, weddings, various activities, etc. The album labels are usually assigned either manually or by using locations from EXIF data if the GPS tags in a camera are switched on. However, content-based image analysis has been recently introduced in photo organizing systems. Such analysis can be

used to selectively look for photos for a particular event in order to keep nice memories of some episodes of our lives [4] or to gather our specific interests for personalized recommender systems.

There exist two different event recognition tasks [2]. In the first task, the event categories are recognized for the whole album (a sequence of photos). However, the assignments of images of the same event into albums may be unknown in practice. Hence, in this paper, we focus on the second task, namely, event recognition in single images from social media. As an event here is a complex scene with large variations in visual appearance [4], deep learning techniques [5] are widely used. It is typical to fine-tune existing convolutional neural networks (CNNs) on event datasets [4]. Sometimes CNN-based object detection is applied [6] for discovering particular categories, e.g., interior objects, food, transport, sports equipment, animals, etc. [7, 8].

However, in this paper, a slightly different approach is considered. Despite the conventional usage of a CNN as a discriminative model in a classifier design [9], we propose to borrow generative models to represent an input image in the other domain. In particular, we use existing methods of image captioning [10] that generate textual descriptions of images. Our main contribution is a demonstration that the generated descriptions can be fed to the input of a classifier in an ensemble in order to improve the event recognition accuracy of traditional methods. Though the proposed visual representation is not as rich as features extracted by fine-tuned CNNs, they are better than the outputs of object detectors [8]. As our approach is completely different than traditional CNNs, it can be combined with them into an ensemble that possesses high diversity and, as a consequence, high accuracy.

The rest of the paper is organized as follows. In Sect. 2, the survey of image captioning models is given. In Sect. 3, we introduce the proposed pipeline for event recognition based on generated captions. Experimental results for several event datasets are presented in Sect. 4. Finally, concluding comments and future works are discussed in Sect. 5.

2 Literature Survey

Most existing methods of event recognition on single photos tend to applications of the CNN-based architectures [2]. Four layers of fine-tuned CNN were used to extract features for LDA (Linear Discriminant Analysis) classifier in the ChaLearn LAP 2015 cultural event recognition challenge [11]. The iterative selection method [4] identifies the most relevant subset of classes for transferring representations from CNN learned from the object (ImageNet) and scene (Places2) datasets. The bounding boxes of detected objects are projected onto multi-scale spatial maps in the paper [6]. An ensemble of scene classifiers and object detectors provided the high accuracy [12] for the Photo Event Collection (PEC) [13]. Unfortunately, there is a significant gap in the accuracies of event classification in still photos [4] and albums [14], so that there is a huge demand in all-the-more accurate methods of single image processing.

That is why in this paper, we proposed to concentrate on other suitable visual features extracted with the generative models and, in particular, image captioning techniques. There is a wide range of applications of image captioning: from the automatic generation of descriptions for photos posted in social networks to image retrieval from databases using generated text descriptions [15]. The image captioning methods are usually based on an encoder-decoder neural network, which first encodes an image into a fixed-length vector representation using pre-trained CNN, and then decodes the representation into captions (a natural language description). During the training of a decoder (generator), the input image and its ground-truth textual description are fed as inputs to the neural network, while one hot encoded description presents the desired network output. The description is encoded using text embeddings in the Embedding (look-up) layer [5]. The generated image and text embeddings are merged using concatenation or summation and form the input to the decoder part of the network. It is typical to include the recurrent neural network (RNN) layer followed by a fully connected layer with the Softmax output layer.

One of the first successful models, “Show and Tell” [16], won the first MS COCO Image Captioning Challenge in 2015. It uses RNN with long short-term memory (LSTM) units in a decoder part. Its enhancement “Show, Attend and Tell” [17] incorporates a soft attention mechanism to improve the quality of the caption generation. The “Neural Baby Talk” image captioning model [18] is based on generating the template with slot locations explicitly tied to specific image regions. These slots are then filled in by visual concepts identified in the object detectors. The foreground regions are obtained using the Faster-RCNN network [19], and LSTM with attention mechanism serves as a decoder. The “Multimodal Recurrent Neural Network” (mRNN) [20] is based on the Inception network for image features extraction and deep RNN for sentence generation. One of the best models nowadays is the Auto-Reconstructor Network (ARNet) [21], which uses the Inception-V4 network [22] in an encoder, and the decoder is based on LSTM. There exist two pre-trained models with greedy search (ARNet-g) and beam search (ARNet-b) with size 3 to generate the final caption for each input image.

3 Proposed Approach

Our task can be formulated as a typical image recognition problem [9]. It is required to assign an input photo X from a gallery to one of $C > 1$ event categories (classes). The training set of $N \geq 1$ images $\mathbf{X} = \{X_n | n \in \{1, \dots, N\}\}$ with known event labels $c_n \in \{1, \dots, C\}$ is available for classifier learning. Sometimes the training photos of the same event are associated with an album [13, 14]. In such a case, the training albums are unfolded into a set $\mathbf{X}$ so that the collection-level label of the album is assigned to labels of each photo from this album. This task possesses several characteristics that makes it extremely challenging compared to album-based event recognition. One of these characteristics is the presence of irrelevant images or unimportant photos that can be associated with

any event [2]. These images can be detected by attention-based models when the whole album is available [1] but may have a significant negative impact on the quality of event recognition in single images.

As N is usually rather small, transfer learning may be applied [5]. A deep CNN is firstly pre-trained on a large dataset, e.g., ImageNet or Places [23]. Secondly, this CNN is fine-tuned on $\mathbf{X}$, i.e., the last layer is replaced to the new layer with Softmax activations and C outputs. An input image X is classified by feeding it to the fine-tuned CNN to compute C scores from the output layer, i.e., the estimates of posterior probabilities for all event categories. This procedure can be modified by the extraction of deep image features (embeddings) using the outputs of one of the last layers of the pre-trained CNN [5, 24]. The input image X and each training image $X_n, n \in \{1, \dots, N\}$ are fed to the input of the CNN, and the outputs of the last-but-one layer are used as the D -dimensional feature vectors $\mathbf{x} = [x_1, \dots, x_D]$ and $\mathbf{x}_n = [x_{n;1}, \dots, x_{n;D}]$, respectively. Such deep learning-based feature extractors allow training of a general classifier $\mathcal{C}_{emb}$, e.g., k-nearest neighbor, random forest (RF), support vector machine (SVM) or gradient boosting [9, 25]. The C -dimensional vector of $\mathbf{p}_{emb} = \mathcal{C}_{emb}(\mathbf{x})$ confidence scores is predicted given the input image in both cases of fine-tuning with the last Softmax layer in a role of classifier $\mathcal{C}_{emb}$ and feature extraction with general classifier. The final decision is made in favor of a class with maximal confidence.

In this paper, we use another approach to event recognition based on generative models and image captioning. The proposed pipeline is presented in Fig. 1. At first, the conventional extraction of embeddings $\mathbf{x}$ is implemented using pre-trained CNN. Next, these visual features and a vocabulary V are fed to a special RNN-based neural network (generator) that produces the caption, which describes the input image. Caption is represented as a sequence of $L > 0$ tokens

Fig. 1. Proposed event recognition pipeline based on image captioning

$\mathbf{t} = \{t_0, t_1, \dots, t_{L+1}\}$ from the vocabulary ($t_l \in V, l \in \{0, \dots, L\}$). It is generated sequentially, word-by-word starting from $t_0 = \langle START \rangle$ token until a special $t_{L+1} = \langle END \rangle$ word is produced [21].

The generated caption $\mathbf{t}$ is fed into an event classifier. In order to learn its parameters, every n -th image from the training set is fed to the same image captioning network to produce the caption $\mathbf{t}_n = \{t_{n;0}, t_{n;1}, \dots, t_{n;L_n+1}\}$. Since the number of tokens L_n is not the same for all images, it is necessary to either train a sequential RNN-based classifier or transform all captions into feature vectors with the same dimensionality. As the number of training instances N is not very large, we experimentally noticed that the latter approach is as accurate as the former, though the training time is significantly lower. This fact can be explained by the absence of anything temporal or serial in the initial task of event recognition in single images. Hence, we decided to use one-hot encoding and convert the sequences $\mathbf{t}$ and $\{\mathbf{t}_n\}$ into vectors of 0s and 1s as described in [26]. In particular, we select a subset of vocabulary $\tilde{V} \subset V$ by choosing the top most frequently occurring words in the training data $\{\mathbf{t}_n\}$ with the optional exclusion of stop words. Next, the input image is represented as the $|\tilde{V}|$ -dimensional sparse vector $\tilde{\mathbf{t}} \in \{0, 1\}^{|\tilde{V}|}$, where $|\tilde{V}|$ is the size of reduced vocabulary $\tilde{V}$ and the v -th component of vector $\tilde{\mathbf{t}}$ is equal to 1 only if at least one of L words in the caption $\mathbf{t}$ is equal to the v -th word from vocabulary $\tilde{V}$. This would mean, for instance, turning the sequence $\{1, 5, 10, 2\}$ into a $\tilde{V}$ -dimensional sparse vector that would be all 0s except for indices 1, 2, 5 and 10, which would be 1s [26]. The same procedure is used to describe each n -th training image with $\tilde{V}$ -dimensional sparse vector $\tilde{\mathbf{t}}_n$. After that an arbitrary classifier $\mathcal{C}_{txt}$ of such textual representations suitable for sparse data can be used to predict C confidence scores $\mathbf{p}_{txt} = \mathcal{C}_{txt}(\tilde{\mathbf{t}})$. It is known [26] that such an approach is even more accurate than conventional RNN-based classifiers (including one layer of LSTMs) for the IMDB dataset.

In general, we do not expect that classification of short textual descriptions is more accurate than the conventional image recognition methods. Nevertheless, we believe that the presence of image captions in an ensemble of classifiers can significantly improve its diversity [27]. Moreover, as the captions are generated based on the extracted feature vector $\mathbf{x}$, only one inference in the CNN is required if we combine the conventional general classifier of embeddings from pre-trained CNN and the image captions. In this paper, the outputs of individual classifiers are combined in simple voting with soft aggregation. In particular, we compute aggregated confidences as the weighted sum of outputs of individual classifier:

$$\mathbf{p}_{ensemble} = [p_1, \dots, p_C] = w \cdot \mathbf{p}_{emb} + (1 - w) \mathbf{p}_{txt}. \quad (1)$$

The decision is taken in favor of the class with maximal confidence:

$$c^* = \underset{c \in \{1, \dots, C\}}{\operatorname{argmax}} p_c. \quad (2)$$

The weight $w \in [0, 1]$ in (1) can be chosen using a special validation subset in order to obtain the highest accuracy of criterion (2).

Let us provide qualitative examples for the usage of our pipeline (Fig. 1). The results of (correct) event recognition using our ensemble are presented in Fig. 2. Here the first line of the title contains the generated image caption. In addition, the title displays the result of event recognition using captions $\mathbf{t}$ (second line), embeddings $\mathbf{x}_{emb}$ (third line), and the whole ensemble (last line). As one can notice, the single classification of captions is not always correct. However, our ensemble is able to obtain a reliable solution even when individual classifiers make wrong decisions.

4 Experimental Results

In the experimental study, we examined the following event datasets:

1. PEC [13] with 61,000 images from 807 collections of $C = 14$ social event classes (birthday, wedding, graduation, etc.).
2. WIDER (Web Image Dataset for Event Recognition) [6] with 50,574 images and $C = 61$ events (parade, dancing, meeting, press conference, etc.).
3. ML-CUFED (Multi-Label Curation of Flickr Events Dataset) [14] contains $C = 23$ common event types. Each album is associated with several events, i.e., it is a multi-label classification task.

We used standard train/test split for all datasets proposed by their creators. In PEC and ML-CUFED, the collection-level label is directly assigned to each image contained in this collection. Moreover, we completely ignore any metadata, e.g., temporal information, except the image itself similarly to the paper [4]. As a result, the training and validation sets are not ideally balanced. The majority classes in each dataset contains 5-times higher number of training images when compared to the minority classes. However, the class distribution in the training and validation sets remains more or less identical, so that the number of validation images for majority classes is also 5-times higher than the number of testing examples for minority classes.

As we mainly focus on the possibility of implementing offline event recognition on mobile devices [12], in order to compare the proposed approach with conventional classifiers, we used MobileNet v2 with $\alpha = 1$ [28] and Inception v4 [22] CNNs. At first, we pre-trained them on the Places2 dataset [23] for feature extraction. The linear SVM classifier from the scikit-learn library was used because it has higher accuracy than other classifiers from this library (RF, k-NN, and RBF SVM) on the considered datasets. Moreover, we fine-tuned these CNNs using the given training set as follows. At first, the weights in the base part of the CNN were frozen, and the new head (fully connected layer with C outputs and Softmax activation) was learned using the ADAM optimizer (learning rate 0.001) for 10 epochs with an early stop in the Keras 2.2 framework with the TensorFlow 1.15 backend. Next, the weights in the whole CNN were learned during 5 epochs using the ADAM. Finally, the CNN was trained using SGD during 3 epochs with 10-times lower learning rate.

In addition, we used features from object detection models that are typical for event recognition [6, 12]. As many photos from the same event sometimes contain identical objects (e.g., ball in the football), they can be detected by contemporary CNN-based methods, i.e., SSDLite [28] or Faster R-CNN [19]. These methods detect the positions of several objects in the input image and predict the scores of each class from the predefined set of $K > 1$ types. We extract the sparse K -dimensional vector of scores for each type of object. If there are several objects of the same type, the maximal score is stored in this feature vector [8]. This feature vector is either classified by the linear SVM or used to train a feed-forward neural network with two hidden layers containing 32 units. Both classifiers were learned using the training set from each event

dataset. In this study, we examined SSD with the MobileNet backbone and Faster R-CNN with the InceptionResNet backbone. The models pre-trained on the Open Images Dataset v4 ($K = 601$ objects) were taken from the TensorFlow Object Detection Model Zoo.

Our preliminarily experimental study with the pre-trained image captioning models discussed in Sect. 2 demonstrated that the best quality for MS COCO captioning dataset is achieved by the ARNet model [21]. Thus, in this experiment, we used ARNet’s encoder-decoder model. However, it can be replaced with any other image captioning technique without modification of our event recognition algorithm.

Unfortunately, event datasets do not contain captions (textual descriptions), which are required to train or fine-tune the image captioning model. Due to this reason, the image captioning model was trained on the Conceptual Captions dataset. Today this dataset is the largest dataset used for image captioning. It contains more than 3.3M image-URL and caption pairs in the training set, and about 15 thousand pairs in the validation set. While there exist other smaller datasets, such as MS COCO and Flickr, in our preliminary experiments, the image captioning model, which were trained on the Conceptual Captions Dataset, provided better worst-case performance in the cross-dataset evaluation.

The feature extraction in the encoder is implemented not only with the same CNNs (Inception and MobileNet v2). We extracted $|\tilde{V}| = 5000$ most frequent words except special tokens $\langle START \rangle$ and $\langle END \rangle$. They are classified by either linear SVM or a feed-forward neural network with the same architecture as for the object detection case. Again, these classifiers are trained from scratch, given each event training set. The weight w in our ensemble (Eq. 1) was estimated using the same set.

The results of the lightweight mobile (MobileNet and SSD object detector) and deep models (Inception and Faster R-CNN) for PEC, WIDER and ML-CUFED are presented in Tables 1, 2, 3, respectively. Here we added the best-known results for the same experimental setups.

Certainly, the proposed recognition of image captions is not as accurate as conventional CNN-based features. However, the classification of textual descriptions is much better than the random guess with accuracy $100\%/14 \approx 7.14\%$, $100\%/61 \approx 1.64\%$ and $100\%/23 \approx 4.35\%$ for PEC, WIDER and ML-CUFED, respectively. It is important to emphasize that our approach has a lower error rate than the classification of the features based on object detection in most cases. This gain is especially noticeable for lightweight SSD models, which are 1.5–13% less accurate than the proposed classification of image captions due to the limitations of SSD-based models to detect small objects (food, pets, fashion accessories, etc.). The Faster R-CNN-based detection features can be classified more accurately, but the inference in Faster R-CNN with the InceptionResNet backbone is several times slower than the decoding in the image captioning model (6–10 s vs. 0.5–2 s on MacBook Pro 2015).

Table 1. Event recognition accuracy (%), PEC

Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	59.72	61.82
	Objects	42.18	47.83
	Texts	43.77	47.24
	Proposed ensemble (1), (2)	60.56	62.87
Fine-tuned CNN	Embeddings	62.33	63.56
	Objects	40.17	47.42
	Texts	43.52	46.89
	Proposed ensemble (1), (2)	63.38	65.12
Aggregated SVM [13]			41.4
Bag of Sub-events [13]			51.4
SHMM [13]			55.7
Initialization-based transfer learning [4]			60.6
Transfer learning of data and knowledge [4]			62.2

Table 2. Event recognition accuracy (%), WIDER

Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	48.31	50.48
	Objects	19.91	28.66
	Texts	26.38	31.89
	Proposed ensemble (1), (2)	48.91	51.59
Fine-tuned CNN	Embeddings	49.11	50.97
	Objects	12.91	21.27
	Texts	25.93	30.91
	Proposed ensemble (1), (2)	49.80	51.84
Baseline CNN [6]			39.7
Deep channel fusion [6]			42.4
Initialization-based transfer learning [4]			50.8
Transfer learning of data and knowledge [4]			53.0

Finally, the most appropriate way to use image captioning in event classification is its fusion with conventional CNNs. In such case, we improved the previous state-of-the-art for PEC from 62.2% [4] even for the lightweight models (63.38%) if the fine-tuned CNNs are used in an ensemble. Our Inception-based model is even better (accuracy 65.12%). We have not still reached the state-of-the-art accuracy 53% [4] for the WIDER dataset, though our best accuracy (51.84%) is up to 9% higher when compared to the best results (42.4%) from original paper [6]. Our experimental setup for the ML-CUFED dataset is studied for the

Table 3. Event recognition accuracy (%), ML-CUFED

Classifier	Features	Lightweight models	Deep models
SVM	Embeddings	53.54	57.27
	Objects	34.21	40.94
	Texts	37.24	41.52
	Proposed ensemble (1), (2)	55.26	58.86
Fine-tuned CNN	Embeddings	56.01	57.12
	Objects	32.05	40.12
	Texts	36.74	41.35
	Proposed ensemble (1), (2)	57.94	60.01

first time here because this dataset is developed mostly for album-based event recognition. We should highlight that our preliminary experiments in the latter task with this dataset and simple averaging of MobileNet features extracted from all images from an album slightly improved the state-of-the-art accuracy for this dataset, though it is necessary to study more complex feature aggregation techniques [1].

In practice, it is preferable to use pre-trained CNN as a feature extractor in order to prevent additional inference in fine-tuned CNN when it differs from the encoder in the image captioning model. Unfortunately, the accuracies of SVM for pre-trained CNN features are 1.5–3% lower when compared to the fine-tuned models for PEC and ML-CUFED. In this case, an additional inference may be acceptable. However, the difference in error rates between pre-trained and fine-tuned models for the WIDER dataset is not significant, so that the pre-trained CNNs are definitely worth being used here.

5 Conclusion

In this paper, we have proposed to apply generative models in the classical discriminative task [9]; namely, image captioning in event recognition in still images. We have presented the novel pipeline of visual preference prediction using image captioning with the classification of generated captions and retrieval of images based on their textual descriptions (Fig. 1). It has been experimentally demonstrated that our approach is more accurate than the widely-used image representations obtained by object detectors [6, 8]. Moreover, our approach is much faster than Faster R-CNNs, which do not implement one-shot detection. What is especially useful for ensemble models [27] generated caption provides additional diversity to conventional CNN-based recognition.

The motivation behind the study of image captioning techniques in this paper is connected not only with generating compact informative descriptions of images, but also with the wide possibilities to ensure the privacy of user data if further processing at remote servers is necessary. Moreover, as the vocabulary

of generated captions is restricted, such techniques are considered as effective anonymization methods. Since the textual descriptions can be easily perceived and understood by the user (as opposed to a vector of numeric features), his or her attitude to the use of such methods will be more trustworthy.

Unfortunately, short conceptual textual descriptions are obviously not enough to classify event categories with high accuracy even for a human due to errors and lack of specificity (see an example of generated captions in Fig. 2). Another disadvantage of the proposed approach is the need to repeat inference if fine-tuned CNN is applied in an ensemble. Hence, the decision-making time will be significantly increased, though the overall accuracy also becomes higher in most cases (Tables 1 and 3).

In the future, it is necessary to make the classification of generated captions more accurate. At first, though our preliminary experiments of LSTMs did not decrease the error rate of our simple approach with linear SVM and one-hot encoded words, we strongly believe that a thorough study of the RNN-based classifiers of generated textual descriptors is required. Second, the comparison of image captioning models trained on the Conceptual Captions dataset is needed to choose the best model for caption generation. Here the impact on event recognition accuracy arising from erroneous captions being generated should be examined. Third, additional research is needed to check if we can fine-tune a CNN on an event dataset and use it as an encoder for the caption generation without loss of quality. In this case, a more compact and fast solution can be achieved. Finally, the proposed pipeline should be extended for the album-based event recognition [2, 13] with, e.g., attention models [12].

Acknowledgements. This research is based on the work supported by Samsung Research, Samsung Electronics. The work of A.V. Savchenko was conducted within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE).

References

1. Guo, C., Tian, X., Mei, T.: Multigranular event recognition of personal photo albums. *IEEE Trans. Multimedia* **20**(7), 1837–1847 (2017)
2. Ahmad, K., Conci, N.: How deep features have improved event recognition in multimedia: a survey. *ACM Trans. Multimedia Comput. Commun. Appl.* **15**(2), 39 (2019)
3. Papadopoulos, S., Troncy, R., Mezaris, V., Huet, B., Kompatsiaris, I.: Social event detection at MediaEval 2011: challenges, dataset and evaluation. In: *MediaEval* (2011)
4. Wang, L., Wang, Z., Qiao, Y., Van Gool, L.: Transferring deep object and scene representations for event recognition in still images. *Int. J. Comput. Vis.* **126**(2–4), 390–409 (2018)
5. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge (2016)
6. Xiong, Y., Zhu, K., Lin, D., Tang, X.: Recognize complex events from static images by fusing deep channels. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1600–1609 (2015)

7. Grechikhin, I., Savchenko, A.V.: User modeling on mobile device based on facial clustering and object detection in photos and videos. In: Morales, A., Fierrez, J., Sánchez, J.S., Ribeiro, B. (eds.) *IbPRIA 2019*. LNCS, vol. 11868, pp. 429–440. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31321-0_37
8. Savchenko, A.V., Rassadin, A.G.: Scene recognition in user preference prediction based on classification of deep embeddings and object detection. In: Lu, H., Tang, H., Wang, Z. (eds.) *ISNN 2019*. LNCS, vol. 11555, pp. 422–430. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22808-8_41
9. Prince, S.J.: *Computer Vision: Models, Learning and Inference*. Cambridge University Press, Cambridge (2012)
10. Hossain, M., Soheli, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* **51**(6), 1–36 (2019)
11. Escalera, S., et al.: ChaLearn looking at people 2015: apparent age and cultural event recognition datasets and results. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1–9 (2015)
12. Savchenko, A.V., Demochkin, K.V., Grechikhin, I.S.: User preference prediction in visual data on mobile devices. *arXiv preprint arXiv:1907.04519* (2019)
13. Bossard, L., Guillaumin, M., Van Gool, L.: Event recognition in photo collections with a stopwatch HMM. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1193–1200 (2013)
14. Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., Cottrell, G.W.: Recognizing and curating photo albums via event-specific image importance. In: *Proceedings of British Conference on Machine Vision (BMVC)* (2017)
15. Vijayaraju, N.: Image retrieval using image captioning. Master’s Projects, p. 687 (2019). <https://doi.org/10.31979/etd.vm9n-39ed>
16. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 652–663 (2017)
17. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2048–2057 (2015)
18. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 91–99 (2015)
20. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Deep captioning with multi-modal recurrent neural networks (m-RNN). In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2015)
21. Chen, X., Ma, L., Jiang, W., Yao, J., Liu, W.: Regularizing RNNs for caption generation by reconstructing the past with the present. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018)
22. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *Proceedings of the International Conference on Learning Representations (ICLR) Workshop* (2016)
23. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(6), 1452–1464 (2018)
24. Savchenko, A.V.: Sequential three-way decisions in multi-category image recognition with deep features based on distance factor. *Inf. Sci.* **489**, 18–36 (2019)

25. Savchenko, A.V.: Probabilistic neural network with complex exponential activation functions in image recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(2), 651–660 (2020)
26. Chollet, F.: *Deep Learning with Python*. Manning Publications Company, Shelter Island (2017)
27. Zhou, Z.H.: *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, London (2012)
28. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. IEEE (2018)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.