Author Guidelines for ECCV Submission

Anonymous ECCV submission

Paper ID ***

Abstract. Automatic image organization of personal photos is a problem with many real world applications, and can be ~~segmented~~ into two main tasks: recognizing event types of the photo collections, and selecting interesting images from the collections. The two tasks are both challenging, in that the photos from the same event type are highly ~~various,~~ and there is ~~high ambiguity of photos~~ from different event types. In this paper, we looked into the possibility ~~to simultaneously solve~~ both tasks: album-wise event recognition and single image event-specific importance score prediction. We collected an event album dataset with both event type labels and image importance labels, refined from the existing CUFED dataset. We propose a hybrid system consisting of three parts: Siamese network based event-specific image importance prediction, Convolutional Neural Network(CNN) based event recognition, and ~~Long Term Short Memory~~(LSTM) based sequence level event recognition, and we propose an iterative updating procedure for event type and image importance score prediction. We show with experiments that the proposed method outperforms the classical approach based on static image classification, and more importantly, we verified that image importance score prediction and event type recognition can in turn help the performance of each other.

Keywords: Event Recognition, Image Importance, Convolutional Neural Network, Long Term Short Memory

1 Introduction

~~With the rapid progress in portable photo taking devices~~ ~~and the increasing volume of storage devices and services,~~ it becomes painless to take photos frequently in daily life, resulting in the explosion of personal photo collections. However, the oversized image collections ~~makes~~ it difficult for us to organize the photos, and thus automatic organization algorithms ~~is~~ highly desirable. The organization of personal photo collections can be decomposed into two stages: recognition of the event types of ~~an~~ photo collection, and suggesting the most interesting/important images in the photo collection to represent the album, e.g. making ~~a~~ album cover, or ~~to suggest~~ to the user for ~~further usage,~~ e.g. making a photo book. The two stages assist users ~~to keep~~ the photo collections organized and ~~clean.~~

Both image importance prediction and event recognition have been studied independently.

Studies ~~on~~ event recognition can be separated into several types. The most popular branches of study use videos as input [1–3]. Spatial and temporal features of videos are usually used for this task. The success of ~~Long Term Short Memory~~ (LSTM) ~~network~~ for sequence tasks has also extended to video based event recogntion. In [4], the LSTM network and visual feature extraction network is stacked, and the network can deal with both event recognition and description. Reiter *et al.* [5] also combine LSTM and ~~HMM~~ for video meeting analysis. ~~On~~ the other end of the spectrum, event recognition for single static ~~image~~ ~~is also studied~~ [6–8]. In contrast to video based event recognition, there is no temporal information to ~~explore~~, and there is no need to consider relevant frame importance or contribution to the event, since there is only one ~~"frame".~~ This problem can be viewed as a special ~~class~~ of scene recognition, and both object level features and scene level features are ~~utilized~~ [6]. ~~Recent~~ success of Deep Neural Networks, especially ~~the~~ Convolutional Neural Networks(CNN) provides us with an outstanding visual ~~representation, and is used~~ for static image event recognition task [7, 8].

Album-wise event recognition lies in the middle between single-image-based and video-based event recognition. Images in an album can be thought of very sparse samples from the event video. Photo albums differ from videos in that consecutive images from the photo album are ~~not anymore~~ continuous, and can have very different visual and semantic information. However, there is still sequential information in ~~the time ordered~~ albums, and the images in an album are of varied importance. In [9], an HMM-based model is proposed to ~~utilize~~ the sequential information ~~of albums for recognition task.~~ It ~~is~~ shown that the ~~proposed~~ model outperforms simple aggregation of ~~prediction~~ from all the images in ~~an~~ album. This indicates that the sequential information ~~of~~ an album is helpful for ~~the~~ album-wise event recognition.

Image importance is a complex image property which is related to various factors, such as aesthetics [10], interestingness [11] and image memorability [12]. It ~~is~~ shown [13] that the ~~image importance property~~ is ~~also relevant to~~ the context the image ~~is in~~, making ~~the~~ image importance ~~album dependent,~~ or ~~event specific~~. ~~The event~~-specific image importance is a highly subjective ~~problem~~, and learning it is a very challenging task, due to the very high intra-class variability, and the ~~underlining~~ uncertainty caused by the subjectiveness of the property. Nevertheless, it is still ~~predictive~~.

Return to the task of photo collection organization, we ask the question: ~~Can~~ we simultaneously recognize the event type for an album, and discover important images in the event album? To answer this question, this paper makes ~~contributions as follows:~~ 1. We refine the existing event curation dataset CUFED by collecting more human annotations for ~~event types of albums~~ for more reliable ground-truth, and allow for multilabel for an album. The multilabel and ground-truth ~~confusion~~ between event types provides us with more training information, and allows for a ~~more fair~~ evaluation ~~at testing~~ stage; 2. We propose a joint event ~~recognition - image~~ importance prediction algorithm. We use CNN for image level event recognition, and Siamese Network for event-specific image importance prediction. An iterative update scheme is conducted during test stage, and ~~it is found~~ that event recognition and image importance prediction can ~~improve the performance~~ ~~in turn for each other;~~ 3. We further boost the performance of event recognition with LSTM network that ~~learns sequential information.~~

2 Related Works

Our work is partly inspired by [13], ~~which~~ proposed a novel image property: event-specific image importance. In this work, it is claimed that image importance or interestingness is contextual and is related to the album it is in. For example, a photo of a beautiful architecture is important in an album of urban trip, yet not so important in a wedding event. A Siamese network is used to predict relative score difference between an input image pair, and ~~it~~ is jointly trained on all event types. However, in this work, event-specific image importance score is predicted given the ground-truth event type of the album. In our work, we extend this idea to training a system ~~for~~ simultaneously event recognition and importance image curation, so that additional user input of event type for testing is ~~omitted~~.

Our work is closely related to the study of event recognition for a personal album. The model in [14] classifies a personal album into 8 social events and 10 sports events simply by aggregating the SVM classification result from single images in ~~this~~ album. In [15], Tsai *et al.* ~~exploits~~ object level patterns for event type recognition. Object patterns are learnt from single images, and then an album-wise SVM is trained on the frequency distribution of different object patterns appearing in an album. Similarly, Imran *et al.* [16] use Pagerank technique to mine the most useful features for an event, and an album-wise SVM classifier is used for recognitoin. The above works treat albums as an unordered collection of images. On the other hand, in [9], Bossard *et al.* exploit the sequential ~~property~~ of personal albums and use HMM based sub-event approach for event recognition. They use ~~time information~~ ~~of~~ images, and model an album with successive latent sub-events to boost the recognition performance. They collected a 14 class dataset consisting of 807 albums for the task.

Event recognition for single photos has also been studied. Li *et al.* [6] use a generative graphical model to recognize event types of a database with 8 sports events. Their model integrates cues from scene and object categorization to classify ~~sports~~ events. Salvador *et al.* [8] focus on cultural event recognition. They integrate cues from visual features extracted by ~~CNN~~ and from time stamp of ~~the~~ photo, inspired by the fact that photos of a cultural event are mostly taken in the same period of time. However, in personal photo collections, the relevance of an image within an event album varies a ~~lot~~. These approaches for single image are useful, but not sufficient for album-wise event recognition.

Convolutional Neural Network(CNN) methods have greatly boosted ~~the~~ performance in image understanding tasks, such as image classification, object detection and scene recognition [17–20]. Many studies ~~move~~ their focus ~~on~~ to higher-level image ~~properties such~~ as event recognition [21], semantic segmentation [22], multilabel image annotation [23], and image captioning [24], ~~etc.~~ Long Short Term Memory(LSTM) networks [4] ~~were~~ proposed for sequence prediction and sequence labeling. ~~It is advantageous~~ over traditional Recurrent Neural Networks (RNN) in that ~~it~~ can ~~keep the long range context information of the data sequence~~. LSTMs have achieved success for ~~different~~ tasks such as handwritten text recognition [25] and speech recognition [26]. Relevant to our work, Long-term Recurrent Convolutional Network (LRCN) model [24] ~~is~~ proposed to stack CNN feature extractor and LSTM networks for sequential learning of videos or images.

3 The Refined CUFED Dataset

In order to train and evaluate the joint curation-recognition model, we use the Curation of Flickr Events Dataset (CUFED), and refine it by collecting more human opinions on the event types ~~of~~ the ~~available~~ dataset. In this section, we describe the dataset, and provide ~~the~~ consistency analysis of the labels collected from Amazon Mechanical Turk (AMT). The dataset will be ~~-~~available to public.

3.1 ~~Problem of the~~ CUFED Dataset

The CUFED Dataset is an image curation dataset ~~collected~~ from the Yahoo Flickr Creative Commons 100M Dataset (YFCC100M). 20,000 albums ~~are~~ segmented by user tags and timestamp, and then their event types ~~are~~ collected from workers on AMT. Each album ~~receives~~ 3 workers' labels. The dataset contains 23 ~~most~~ common event types ~~in our daily life,~~ ranging from nature trips to wedding. For each event type, 50-200 albums are further randomly sampled from the 20,000 albums, ~~consisting~~ the CUFED dataset of 1883 albums. Importance score for each image in the albums ~~is~~ then decided by workers from AMT. The final ground-truth event-specific importance scores ~~of images~~ ~~are~~ obtained from the average of 5 workers' ~~voting~~.

One problem the CUFED Dataset has is that the event type of an album is decided ~~by 3 workers,~~ and ~~it is restricted to be single label.~~ However, for an album with ambiguous event types or with multiple event types, it is ~~not enough~~ to give the album a single event label. For example, the two albums in Fig 1 are both birthday event, but they can also fall into the category of casual family/friends gathering. Those two event types are not mutually exclusive. Moreover, intuitively, we would say the album on the right is a more typical birthday event, with most images focusing on the little boy celebrating his ~~birthday; meanwhile,~~ the album on the left is more of a casual family/friends gathering rather than an obvious birthday event. Therefore, collecting the event types and their proportion in one album is necessary. This ~~will result~~ in a multi-label event recognition ~~dataset.~~

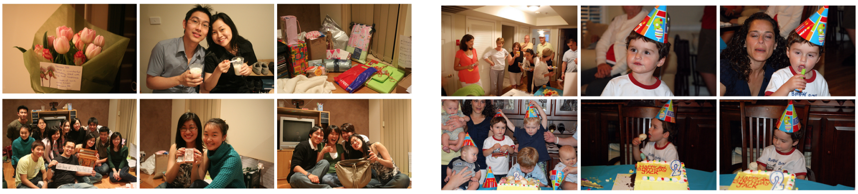

Fig. 1: Example of two birthday albums (both ~~has~~ the photo uploader's tag "birthday").

3.2 Data collection

On top of the three votes the dataset already has, we collected 9 more workers' opinion, and ~~allow~~ for multiple choices for ~~one~~ album. One worker ~~can~~ select up to three event types for an album. There ~~are totally~~ 299 distinct workers ~~participating~~ in the task.

The quality of different AMT workers' ~~submission varies.~~ Therefore, we need to do quality control in order to collect high quality annotations. Before the real task, only workers who passed an album event recognition test (which is very similar to the ~~real task)~~ ~~could~~ proceed to work on the ~~real~~ task. During the tasks, there was another round of quality control. After workers submitted the tasks, the results they turned in were compared with other workers' ~~submission,~~ and submissions highly diverged from others were further manually inspected. If the divergence ~~is~~ unreasonable, the submission ~~is~~ rejected. After all the annotations from workers ~~are~~ collected, we further ~~clean~~ the annotations by eliminating the labels with minor votes: for ~~one~~ album, all the event types with only one vote ~~are~~ discarded.

For the final ground-truth event types and their proportion ~~of one~~ album, we ~~use~~ the Softmax outputs of the workers' voting.

3.3 Dataset Analysis

~~To make sure the dataset we collected is valid,~~ we ~~analyze~~ the annotations in several ways. ~~For all 1883 albums, each~~ album can get ~~12 or more votes~~ (because we allow for multiple choices from one worker). 76% of the albums ~~get all the~~ votes for two or fewer event types. This suggests the high ~~confidence~~ of those albums. 95% of the albums ~~get~~ votes for three or fewer event types. To check the consistency among workers, we randomly split ~~all~~ the 299 workers into two halves, and for each album we ~~check~~ whether the annotations from ~~each~~ half ~~is~~ consistent with the other half. For ~~one~~ album, we ~~examine~~ whether the top event types suggested by these two independent groups ~~are~~ the same. We ~~repeat~~ the random split ~~for~~ 100 times, and on average, for 89.6% of the albums, the event type receiving most votes ~~are~~ the same ~~from~~ both groups. This suggests that despite the ambiguity of some album types, we ~~get~~ consistent opinion from ~~different~~ AMT workers.

4 Approach

In this section, we describe our approach to jointly attain image importance prediction and album event-type recognition. The system is shown in Figure 2.

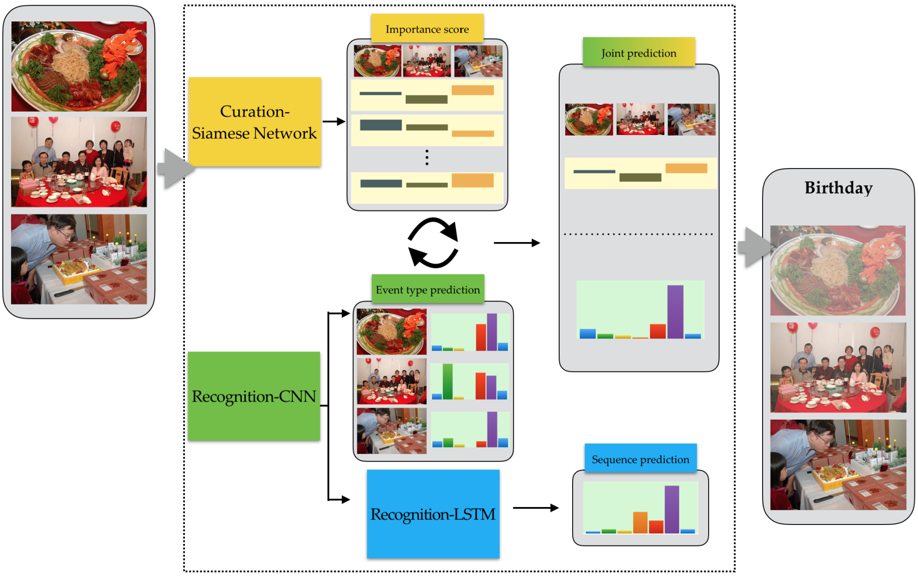

Fig. 2: The joint album recognition-curation system. The system consists of three parts: Siamese network for image importance score prediction, CNN for single image event-type recognition, and LSTM network for album event-type recognition. During test stage, the three components ~~are merged~~ and jointly produce the prediction of album event type and image importance score.

4.1 Event curation network

For event curation purpose, we followed the approach in [13], using Piecewise Ranking (PR) loss to jointly train a Siamese network to predict the importance score difference between an image pair given the ground-truth event type of the input image pair. The architecture can predict the relative event-specific importance score for a set of images, and is found to perform better than traditional CNN that directly predict absolute important score of an image. One difference between our implementation and [13] is that the ground-truth event type for the input image pair ~~now~~ is not one-hot representation ~~anymore~~; instead, the ground-truth event type label is a soft distribution. ~~E~~vent type label is used to gate the output and gradient of the Siamese network: for an input image pair of certain event type $c \in C$, only the part of network which corresponds to event type c is back-propagated. Here, to train the Siamese network on the multi-event type label, we change the 0/1 gating to a soft gating, thus for input image pair with multiple event type labels, the error signals from all the possible labels are back-propagated, but with a weight: the ground-truth probability of that event type. Another scheme is to still use 0/1 gating, but for input image pair with multiple event type labels, there is equal probability for the network to view the input image pair as from one of the possible labels. This gives more relaxation on the ground-truth event type label.

4.2 Event recognition network

One of the properties of an "event album" that ~~make~~ it different from ~~just a~~ collection of images is that it is a sequence, and this provides us with ~~more information than just the images themselves, but also the order and relationship between the images.~~ ~~LSTM is~~ successfully applied ~~for~~ sequential tasks, and its ability ~~of~~ long-range context memorization is suitable for our task of album-wise event recognition. Therefore, we use the LSTM network to capture the sequential information, in ~~additional~~ to a classical CNN that captures visual features of single image.

We start with a CNN pre-trained on ImageNet [27] [17], and fine-tune it on the CUFED Dataset to recognize single image's event type. We then extract the high level CNN features for each image from the adapted network, and use them as the input features to train the LSTM network for album-wise event recognition. The LSTM network consists of single LSTM layer, a mean pooling layer, and ~~the Softmax~~ prediction layer.

The target for both fine-tuned CNN and LSTM network is the soft distribution of the ground-truth event ~~label~~. ~~Again, another scheme is to use the one-hot target, but treat each training example as from one of the possible event types with equal probabilities.~~

During testing stage, the album-wise prediction from ~~LSTM~~ and the image-wise prediction from ~~CNN~~ are combined to produce the final prediction for the album type.

4.3 Iterative curation-recognition procedure

For an "event album", more interesting images or important images give us more information about the event type of this album. For example, although a candle blowing image may only appear in an album once, it is very helpful for deciding the event type of ~~this~~ album. However, as shown in [13], the importance of an image is related to the context it is in, and is event-type dependent. Therefore, we propose that the image importance score can help with event recognition of an album, while event recognition of an album will in turn improve~~s~~ the image importance score prediction.

We denote an N-image album as $\mathbf{S} = \left\{I^1, ..., I^N\right\}$. From the event curation network in Section 4.1, we obtain the importance score of an image $I^n \in S$ given its event type: $W^n = \left[w_1^n, ..., w_C^n\right]^T$ where C is the number of event types. From the event recognition CNN, we can also get the prediction of ~~event~~ type for single image $P^n = [p_1^n, ..., p_C^n]^T$. We then conduct the iterative curation-recognition procedure:

$$\begin{cases} Q(k+1) = \left[P^1, P^2, ..., P^N\right] \cdot \max(m_1, V(k))^\alpha \\ V(k+1) = \left\{[W^1, W^2, ..., W^N]^T \circ \mathrm{I}\left\{p_c^n \geq m_2 \cdot \max_{c'}(p_{c'}^n)\right\}_{n,c}\right\} \cdot Q(k+1) \end{cases} \tag{1}$$

Or equivalently, it can be written in vector form:

$$\begin{cases} q_c(k+1) = \max(m_1, V^T(k))^\alpha \cdot P^n \\ v^n(k+1) = \left\{W_n^T \circ \mathrm{I}\left\{p_c^n \geq m_2 \cdot \max_{c'}(p_{c'}^n)\right\}_{(c,1)}\right\} \cdot Q(k+1) \end{cases} \tag{2}$$

where the N-dimensional column vector $V(k) = [v^1, ..., v^N]^T$ is the k-th step prediction for all images' importance score in album S, and the C-dimensional column vector $Q(k) = [q_1, q_2, ..., q_C]^T$ is the k-th step prediction for the album's event type.

$\mathrm{I}\left\{p_c^n \geq m_2 \cdot \max_{c'}(p_{c'}^n)\right\}_{n,c}$ denotes the binary mask that eliminates the event type predictions with low confidence, and makes sure only event types with high probability contribute to the image importance prediction. m_1 is the threshold to only allow for important images to contribute to event type prediction of the album. α is the weight factor that reflect our emphasis on the image importance score.

By iteratively conducting the procedures in Equation 1, we obtain the album-wise event prediction Q and image importance score prediction V.

Note that Equation 1 is not guaranteed to converge. In case of oscillation between states, we set a maximum number of iterations, and when the iteration number hits the threshold, predictions for Q and V are obtained by averaging over previous steps.

4.4 Joint prediction

Iterative curation-recognition procedure takes image importance into account for event type prediction, while the LSTM network ~~learns sequential information of the album.~~ These two processes stress two distinct properties of event albums, and are complementary to each other. Therefore, we average the predicted probability density from LSTM network Q_{LSTM} and from iterative curation-recognition procedure Q_{iter}, and get the final event type prediction for an album. The entire system is illustrated in Figure 2.

5 Experimental Results

6 Conclusion

References

1. Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A.F., Quéenot, G., Ordelman, R.: Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2015, NIST, USA (2015)
2. Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: CVPR. (2012)
3. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative cnn video representation for event detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1798–1807
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997)
5. Reiter, S., Schuller, B.W., Rigoll, G.: A combined LSTM-RNN - HMM - approach for meeting event segmentation and recognition. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14-19, 2006. (2006) 393–396
6. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (2007)
7. Park, S., Kwak, N.: Cultural event recognition by subregion classification with convolutional neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2015)
8. Salvador, A., Zeppelzauer, M., Manchon-Vizuete, D., Calafell-Orós, A., Giró-i Nieto, X.: Cultural event recognition with visual convnets and temporal models. In: CVPR ChaLearn Looking at People Workshop 2015. (06/2015 2015)
9. Bossard, L., Guillaumin, M., Van, L.: Event recognition in photo collections with a stopwatch hmm. In: Computer Vision (ICCV), 2013 IEEE International Conference on. (2013)
10. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: Rating pictorial aesthetics using deep learning. In: Proceedings of the ACM International Conference on Multimedia. (2014)
11. Dhar, S., Ordonez, V., Berg, T.L.: High level describable attributes for predicting aesthetics and interestingness. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
12. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
13. Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., Cottrell, G, W.: Event-specific image importance. ~~In:~~ In Proc. Computer Vision and Pattern Recognition Conference (CVPR). (2016)
14. R.MattiviandJ.R.R.UijlingsandF.deNataleandN.Sebe: Exploitation of time constraints for (sub-) event recognition. In: ACM Workshop on Modeling and Representing Events (J-MRE 11). (2011)
15. Tsai, S., Cao, L., Tang, F., Huang, T.S.: Compositional object pattern: a new model for album event recognition. In: Proceedings of the 19th International Conference on Multimedia 2011. (2011)
16. Imran, N., Liu, J., Luo, J., Shah, M.: Event recognition from photo collections via pagerank. In: Proceedings of the 17th ACM International Conference on Multimedia. MM '09, New York, NY, USA, ACM (2009) 621–624
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. Curran Associates, Inc. (2012)

18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
19. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
20. 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 27 (NIPS). (2014)
21. Xiong, Y., Zhu, K., Lin, D., Tang, X.: Recognize complex events from static images by fusing deep channels. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CVPR (to appear) (November 2015)
23. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. CoRR **abs/1312.4894** (2013)
24. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. (2015)
25. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2009)
26. Sak, H., Senior, A.W., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014
27. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)