Regim Lab Team at ImageCLEF Lifelog Moment Retrieval Task 2018

Fatma Ben Abdallah, Ghada Feki, Mohamed Ezzarka, Anis Ben Ammar, and Chokri Ben Amar

Regim-LAB, REsearch Groups in Intelligent Machines, University of Sfax, National Engineering School of Sfax (ENIS), Sfax, Tunisia
{ben.abdallah.fatma, ghada.feki, mohamed.ezzarka, anis.ben.ammar, chokri.benamar}@ieee.org

Abstract. In this paper we describe our approach for the ImageCLEFlifelog Moment Retrieval task. A total of five runs were submitted, which used visual features, textual features or combination. The first run was based only on the concepts gived by the organizers. In the second and third runs, we used respectively fine-tuned Googlenet and Alexnet for images description. The fourth run was based on the fusion of the two previous runs. For the fifth run, we crossed the results of our best run (based on Alexnet model) with the result of XQuery FLWOR expression applied to the XML file containing the semantic location and activities data. Our architecture is implemented using Neural Network Toolbox, Parallel Computing Toolbox and GPU coder which generates CUDA from MATLAB. The results obtained are promising for a first participation to such a task, with F1-measure@10=0.424 which placed us at third behind AILabGTi Team with 0.545 and HCMUS Team with 0.479.

Keywords: Deep-learning · CNN · LSTM · fine-tuning · lifelog· moments retrieval.

1 Introduction

The goal of the ImageCLEFlifelog LMRT task is to retrieve a number of specific moments in a lifelogger's life. Moments are semantic events, or activities that happened throughout one or several days. For example, participants in this subtask should return the relevant moments for the query 'Find the moment(s) when I was shopping for wine in the supermarket' [6] [18].
ImageCLEFlifelog data consists of anonymised lifelogs gathered by one user over 50 days. The dataset is based on the data available for the NTCIR-13-Lifelog 2 [16].

The ImageCLEFlifelog LMRT task can best be compared to a known-item search task with one (or more) relevant items per topic [16]. By analysing more closely the topics, we find that the query is based on finding a location, an activity or both. The major difficulty to retrieve relevant images to a specific query from

the ImageCLEFlifelog dataset is to analyse the multiple multimodal source of information : the images, the semantic content and the biometric information. There is no precise format for the data. The accuracy of the images returned depends extremely on the exploitation of these data.
The main objective of experiments is to find an automatic way to extract and analyse these multimodal data. Also, it is necessary to find a way to translate each query (topic) into a set of concepts to match with image concepts.

2 Approaches used and progress beyond state-of-the-art

Despite the rapid increase in the number of publications in multimedia retrieval [3,8,9,10,11,12,13,15,21,29,31], the problems related to the semantic gap are not yet solved. Use low-level descriptors via sophisticated algorithms can not model in an efficient way the semantics of an image or a video [19]. Indeed, this approach has many limitations especially when it is dealing with large dataset [23] because there is no direct link between the low level and the semantic level [25] . Deep learning have exposed encouraging results and performance in several multimedia research domain [2,7,4,5]. Convolutional neural networks (CNN) are nowadays the most powerful models for classifying images [14,20]. Given this fact, we focus on works which used deep-learning to retrieval egocentric images. Authors in [30] adopted the text retrieval method where each document has a document ID and its content is a set of labels assigned to the corresponding image. They counted the document frequency for each distinct label to get the idf (inverse document frequency). The image labels were stem from Deep Neural Network (DNN). They used GoogleNet [28] and AlexNet [20] trained on the Imagenet[1] dataset for the object recognition. They used GoogleNet, AlexNet, VGG [26] and Resnet [17] trained on Places365[2] for scene recognition. [22] based their approach on the use of convolutional neural networks to transform egocentric lifelog images into concepts. To generate this transformation, they used Resnet152 trained on Imagenet1K and Places365 and a fast region-based convolutional network method [24] with Inception-Resnet [27] pre-trained on MSCOCO[3]. The important concepts were then learned with the conditional random field. [32] based their works on translating manually (by a human-in-the-loop) the query into specific required pieces of information.
Thus, the proposed works depends on the concepts gave by benchmarks' organizers. They make use of manual annotation to overcome this gap. Some of them, use only CNN trained on IMAGENET to extract concepts. According to [1], performance can be enhanced when the CNN is retrained on images that are more related to the retrieval dataset.
Creating a new convolutional neural network is expensive in terms of expertise, equipment and the amount of annotated data needed. The training can take several weeks for the best CNNs, with many GPUs working on hundreds of

[1] www.image-net.org/
[2] http://places2.csail.mit.edu/download.html
[3] http://cocodataset.org

thousands of annotated images. The complexity of creating CNN can be avoided by adapting publicly available pre-trained networks. We exploit the knowledge acquired on a general classification problem to apply it again to the lifelog context. We choose Alexnet and Googlenet and adapt the last three layer.
The proposed approaches which is based on five phases follow the schema as illustrated in Figure 1. It is divided into two parts: one online and the other offline. The offline part contains the (1)query analysis, (2)the fine-tuning CNN, (3)the data extraction and the (4)image inverted index generation. The online part use these several steps to (5)retrieve relevant images according to a specific query.
The first phase consists in analyzing the query using LSTM to match concepts with queries. The second phase is based on fine-tuning CNN to improve the search performance of the neural network. In the third phase, we use XQuery FLWOR expression to extract relevant images related to location, activity or time. The fourth phase consists in image inverted index generation to facilitate and speed up the processing time of the retrieval. The fifth phase which based on doc2sequence aims at retrieving the data that is matching the query. We detailed in the following each of these phases.

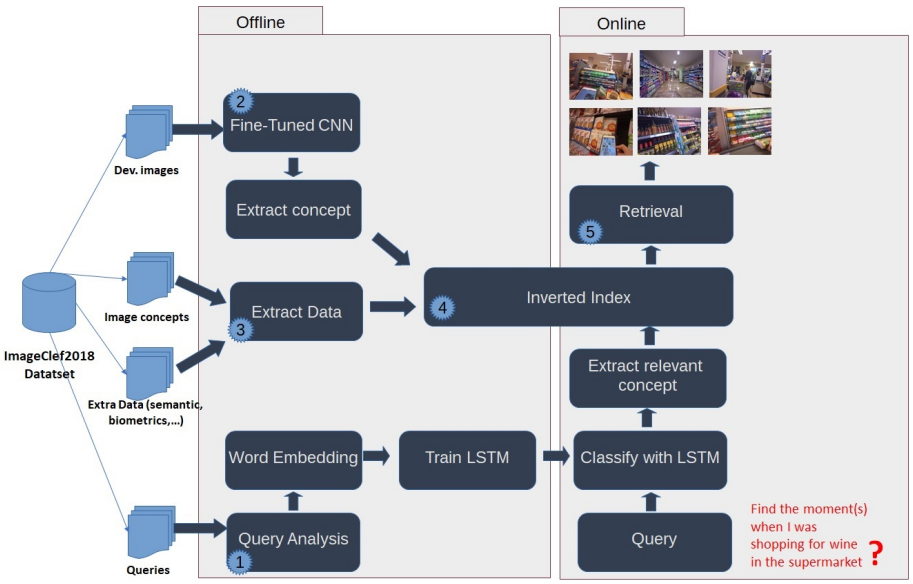

Fig. 1. Proposed architecture

2.1 Query Analysis

As first step, we build a document which contains labeled textual descriptions of queries moments. Then we convert the words to numeric vectors by training a word embedding with dimension 100 and 50 epochs. After that, we create and train an LSTM network based on the sequences of word vectors using the stochastic gradient descent with momentum (SGDM) optimizer with learning rate 0.05.

2.2 Fine-Tuned CNN

In this section, we describe how we fine-tuned Googlenet and Alexnet.
We organize 2184 images into 48 classes. To retrain GoogLeNet to classify new images, we replace the last three layers of the network: a fully connected layer, a softmax layer, and a classification output layer. We set the final fully connected layer to have the same size as the number of classes in the new data set. We freeze the weights of the first 110 layers in the network by setting the learning rates in those layers to zero to speed up network training and to prevent those layers from overfitting to the new data set. We also generate data augmentation to prevent the network from overfitting and memorizing the exact details of the training images. We use 70% of the images for training and 30% for validation. To retrain Alexnet, we followed the same steps as Googlenet except for freezing initial layers. Finally, we classify each images from the IMAGECLEFLifelog2018 dataset using the fine-tuned network and generate a csv file which contains image, concepts and scores. The scores are obtained from the classification of the CNN.

2.3 Extract data

We need to extract the location, the activity or the time coming from sensor readings on mobile devices described in XML format. Indeed, some topics like 'Find the moments when I was taking a road vehicle in foreign countries' or 'Find the moments when I was having dinner at home' need other information than those contained in the images themselves. Therefore, we use XQuery FLWOR expression to extract relevant images related to location (home and longitude/latitude), to activity (transport) and to time (night).

2.4 Image Inverted Index Generation

Building an image inverted index is an important step in our approach. In this file, the images are organized in a matrix which represents the occurrence (or not) for each concept. Indeed, each image contains one or more concepts which describes the content. If a concept is contained in the image, a score between 0 and 1 is assigned to the image. Since the Narrative Clip wearable camera captures one image every 30 seconds, we obtain at the end of each day about 1440 images. So, we chose to generate an image inverted index for each day of

the lifelogger. Create one image inverted index for all the images will cause a considerable loss of time in the generation of the file and also a slowness in the retrieval. To build this matrix, we first extracted all the concepts contained in the dataset and we sorted them alphabetically. Then, we generated a matrix with images names as column and concepts as rows. For our first run, we used the data provided by the organizers to build the image inverted index. For the other runs, we generated the matrix using the output file of the fine-tuned CNN step previously described.

2.5 Retrieval

All the steps described above are done offline. Only the retrieval is online. The proposed approach for the retrieval process is based on the trained LSTM networks, the fine-tuning and the XQuery results. Firstly, we classify the query using the trained LSTM network. We then obtain the concepts that we are working on in the inverted index. Secondly, we search the concepts in the inverted index then we extract the relevant images with scores. After that, we realized an aggregation between the results obtained by the fine-tuning and those obtained by Xquery. Finally, we sort the results based on highest scores.

3 Resources

Our approach is implemented using Intel(R) Core(TM) i5-4430 CPU @3.00Ghz with 16Go RAM. We work on Windows 10 Professional using Matlab 2018a.
We use Neural Network Toolbox with GPU coder which generates CUDA from MATLAB code for deep learning.
We train the fine-tuned Alexnet and Googlenet using Intel(R) Core(TM) i5-4200M CPU @2.50Ghz with 6Go RAM.
It lasts respectively 215 and 751 minutes for a dataset containing 2184 images.

4 Results obtained

We submitted 5 runs on the retrieval LMRT subtask summarized in Table 1.
The first run is exploiting only the concepts provided by the ImageCLEFlifelog organizers. We generate an inverted index for each lifelogger's day based on the information gived by the organizers of ImageCLEFlifelog2018. The retrieval returns the images that contains the concepts extracted from the topics based on the inverted index and the trained LSTM.
The second run is using the Googlenet network fine-tuned with batch size 10 and learning rate 0.0001. We train the network for 6 epochs with 912 iterations. The third run is using Alexnet network fine-tuned with same parameters as Googlenet. The fourth run is based on the result of the two previous runs. We merged the two results, sorted the scores from highest to lowest and take the n first results where n=50. For the fifth run, we crossed the results of our best run

(that of Alexnet) with the result of XQuery FLWOR expression applied to the XML file containing the semantic location and activities data.
The figure 2 shows the official ranking metrics F1-measure@10 for the five runs, which gives equal importance to diversity (via CR@10) and relevance (via P@10) [6].

Table 1. Submitted Runs

Run	RunID	Name	Parsing	Type of information
Run1	#Run4	Baseline-concepts of the organizers	Automatic	Textual
Run2	#Run2	Fine-Tuning with Googlenet	Automatic	Visual
Run3	#Run5	Fine-Tuning with Alexnet	Automatic	Visual
Run4	#Run1	Fine-Tuning with Alexnet/Googlenet	Automatic	Visual
Run5	#Run3	Fine-Tuning with Alexnet + XQuery	Automatic	Visual + Textual

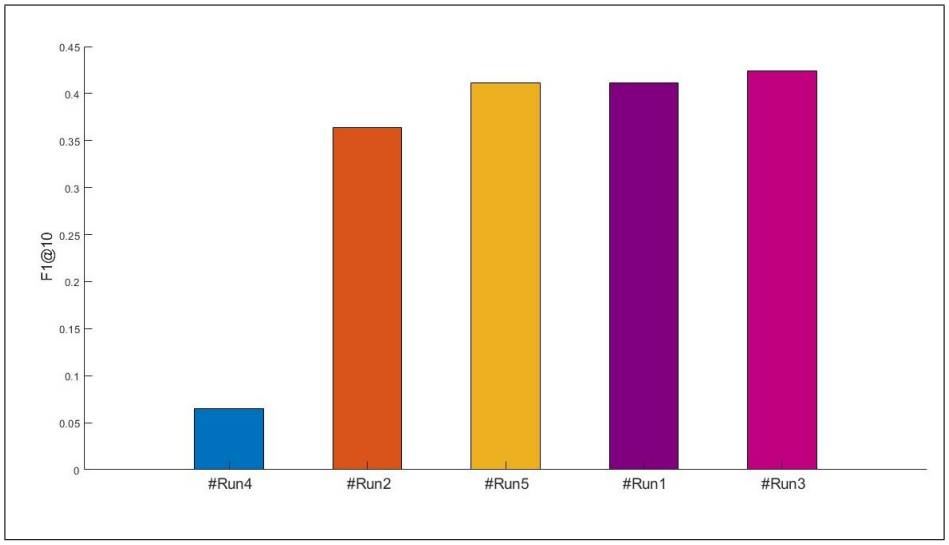

Fig. 2. F1-measure@10

5 Analysis of the results

The tables 2, 3 and 4 show the results of the runs submitted to the retrieval LMRT subtask. The results confirm that relying only on the textual concepts provided by the organizers does not give good results. We then focus on working directly on images so we choose to apply fine-tuning. We first fine-tune

with AlexNet, then with Googlenet which improved significantly retrieval performance. After that, we compare the results from the two previous runs and we take the images with highest scores. The fifth run which gave the best result in term of F1@10 is based on fine tuning with Alexnet and extracted information from the XML files which contains semantic locations and semantic activities. We extracted this information using XQuery.

Table 2. Precision at X (P@X)

Run	RunID	Name	@5	@10	@20	@30	@40	@50
Run1	#Run4	Baseline-concepts of the organizers	0.1	0.06	0.03	0.02	0.015	0.012
Run2	#Run2	Fine-Tuning with Googlenet	0.38	0.34	0.295	0.263	0.227	0.198
Run3	#Run5	Fine-Tuning with Alexnet	0.48	0.39	0.315	0.28	0.24	0.212
Run4	#Run1	Fine-Tuning with Alexnet/Googlenet	0.48	0.39	0.305	0.273	0.237	0.21
Run5	#Run3	Fine-Tuning with Alexnet + XQuery	0.46	0.43	0.33	0.28	0.24	0.214

Table 3. Cluster Recall at X (CR@X)

Run	RunID	Name	@5	@10	@20	@30	@40	@50
Run1	#Run4	Baseline-concepts of the organizers	0.05	0.083	0.083	0.083	0.083	0.083
Run2	#Run2	Fine-Tuning with Googlenet	0.438	0.487	0.512	0.537	0.557	0.557
Run3	#Run5	Fine-Tuning with Alexnet	0.5	0.567	0.64	0.653	0.657	0.661
Run4	#Run1	Fine-Tuning with Alexnet/Googlenet	0.5	0.567	0.616	0.648	0.648	0.672
Run5	#Run3	Fine-Tuning with Alexnet + XQuery	0.45	0.578	0.624	0.648	0.657	0.681

Table 4. F1-measure at X (F1@X)

Run	RunID	Name	@5	@10	@20	@30	@40	@50
Run1	#Run4	Baseline-concepts of the organizers	0.067	0.065	0.042	0.031	0.025	0.02
Run2	#Run2	Fine-Tuning with Googlenet	0.359	0.364	0.324	0.299	0.268	0.24
Run3	#Run5	Fine-Tuning with Alexnet	0.413	0.411	0.365	0.326	0.288	0.261
Run4	#Run1	Fine-Tuning with Alexnet/Googlenet	0.413	0.411	0.352	0.321	0.285	0.259
Run5	#Run3	Fine-Tuning with Alexnet + XQuery	0.36	0.424	0.368	0.324	0.286	0.263

6 Conclusion and Perspectives

This paper focuses on the problem of retrieving specific moment in lifelogger's life during ImageCLEFlifelog2018 LMRT task. We proposed a deep-learning based approach established on five phases using fine-tuning and LSTM. The first phase

consists in analyzing the query using LSTM to match concepts with queries. The second phase is based on fine-tuning CNN to improve the search performance of the neural network. In the third phase, we use XQuery FLWOR expression to extract relevant images related to location, activity or time. The fourth phase consists in image inverted index generation to facilitate and speed up the processing time of the retrieval. The fifth phase which based on doc2sequence aims at retrieving the data that is matching the query. Promising results has been officially reported, demonstrating the effectiveness of the proposed retrieval approach. As future work, we will focus on fine-tuning with other CNNs like Inception or Resnet. Moreover, we will consider face detection based on training cascade object detector.

7 Acknowledgments

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number LR11ES48.

References

1. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European conference on computer vision. pp. 584–599. Springer (2014)
2. Boughrara, H., Chtourou, M., Amar, C.B., Chen, L.: Facial expression recognition based on a mlp neural network using constructive training algorithm. Multimedia Tools and Applications **75**(2), 709–731 (2016)
3. Bouhlel, N., Feki, G., Ammar, A.B., Amar, C.B.: A hypergraph-based reranking model for retrieving diverse social images. In: Computer Analysis of Images and Patterns - 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part I. pp. 279–291 (2017)
4. Brahimi, S., Aoun, N.B., Amar, C.B.: Very deep recurrent convolutional neural network for object recognition. In: Ninth International Conference on Machine Vision (ICMV 2016). vol. 10341, p. 1034107. International Society for Optics and Photonics (2017)
5. Chaabouni, S., Benois-Pineau, J., Tison, F., Amar, C.B., Zemmari, A.: Prediction of visual attention with deep cnn on artificially degraded videos for studies of attention of patients with dementia. Multimedia Tools and Applications **76**(21), 22527–22546 (2017)
6. Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
7. ElAdel, A., Ejbali, R., Zaied, M., Amar, C.B.: Deep learning with shallow architecture for image classification. In: High Performance Computing & Simulation (HPCS), 2015 International Conference on. pp. 408–412. IEEE (2015)
8. Fakhfakh, R., Ammar, A.B., Amar, C.B.: Fuzzy user profile modeling for information retrieval. In: KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014. pp. 431–436 (2014)

9. Fakhfakh, R., Feki, G., Ammar, A.B., Amar, C.B.: Personalizing information retrieval: a new model for user preferences elicitation. In: Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on. pp. 002091–002096. IEEE (2016)
10. Feki, G., Ammar, A.B., Amar, C.B.: Towards diverse visual suggestions on flickr. In: Ninth International Conference on Machine Vision (ICMV 2016). vol. 10341, p. 103411Z. International Society for Optics and Photonics (2017)
11. Feki, G., Fakhfakh, R., Ammar, A.B., Amar, C.B.: Knowledge structures: Which one to use for the query disambiguation? In: Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on. pp. 499–504. IEEE (2015)
12. Feki, G., Fakhfakh, R., Bouhlel, N., Ammar, A.B., Amar, C.B.: REGIM @ 2016 retrieving diverse social images task. In: Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016. (2016)
13. Feki, G., Ksibi, A., Ammar, A.B., Amar, C.B.: Improving image search effectiveness by integrating contextual information. In: Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on. pp. 149–154. IEEE (2013)
14. Garcia-Garcia, A., Orts, S., Oprea, S., Villena-Martinez, V., Rodríguez, J.G.: A review on deep learning techniques applied to semantic segmentation. CoRR **abs/1704.06857** (2017)
15. Guedri, B., Zaied, M., Amar, C.B.: Indexing and images retrieval by content. In: High Performance Computing and Simulation (HPCS), 2011 International Conference on. pp. 369–375. IEEE (2011)
16. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Gupta, R., Albatal, R., Nguyen, D.T.D.: Overview of ntcir-13 lifelog-2 task. Proceedings of the Thirteenth NTCIR conference (NTCIR-13), Tokyo, Japan (December 5-8 2017)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
19. Khodaskar, A., Ladhake, S.: Content based image retrieval using quantitative semantic features. In: International Conference on Human Interface and the Management of Information. pp. 439–448. Springer (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. pp. 1097–1105. NIPS'12, Curran Associates Inc., USA (2012)
21. Ksibi, A., Feki, G., Ammar, A.B., Amar, C.B.: Effective diversification for ambiguous queries in social image retrieval. In: International Conference on Computer Analysis of Images and Patterns. pp. 571–578. Springer (2013)
22. Lin, J., del Molino, A.G., Xu, Q., Fang, F., Subbaraju, V., Lim, J.H., Li, L., Chandrasekhar, V.: Vci2r at the ntcir-13 lifelog-2 lifelog semantic access task (2017)
23. Mojsilovic, A., Rogowitz, B.: Capturing image semantics with low-level descriptors. In: Proceedings 2001 International Conference on Im-

age Processing (Cat. No.01CH37205). vol. 1, pp. 18–21 vol.1 (2001). https://doi.org/10.1109/ICIP.2001.958942
24. 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. pp. 91–99 (2015)
25. Sethi, I.K., Coman, I.L., Stan, D.: Mining association rules between low-level image features and high-level concepts. In: Data Mining and Knowledge Discovery: Theory, Tools, and Technology III. vol. 4384, pp. 279–291. International Society for Optics and Photonics (2001)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. vol. 4, p. 12 (2017)
28. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR) (2015)
29. Wali, A., Aoun, N.B., Karray, H., Amar, C.B., Alimi, A.M.: A new system for event detection from video surveillance sequences. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 110–120. Springer (2010)
30. Yamamoto, S., Nishimura, T., Akagi, Y., Takimoto, Y., Inoue, T., Toda, H.: Pbg at the ntcir-13 lifelog-2 lat, lsat, and lest tasks. Proceedings of NTCIR-13, Tokyo, Japan (2017)
31. Zarka, M., Ammar, A.B., Alimi, A.M.: Fuzzy reasoning framework to improve semantic video interpretation. Multimedia Tools Appl. **75**(10), 5719–5750 (2016)
32. Zhou, L., Duane, A., Nguyen, D., Tien, D., Gurrin, C.: Dcu at the ntcir-13 lifelog-2 task. NTCIR (2017)