Convolutional Neural Networks for Disaster Images Retrieval

Sheharyar Ahmad[1],Kashif Ahmad[2], Nasir Ahmad[1], Nicola Conci[2]
[1]DCSE, UET Peshawar, Pakistan
[2]DISI-University of Trento, Trento
engr_sheharyar@yahoo.com,kashif.ahmad@unitn.it,n.ahmad@uetpeshawar.edu.pk
nicola.conci@unitn.it

ABSTRACT

This paper presents the method proposed by MRLDCSE team for the disaster image retrieval task in Mediaeval 2017 challenge on Multimedia and Satellite. In the proposed work, for visual information, we rely on Convolutional Neural Networks (CNN) features extracted with two different models pre-trained on ImageNet and places datasets. Moreover, a late fusion technique is employed to jointly utilize visual and the additional information available in the form of meta-data for the retrieval of disaster images from social media. The average precision for our three different runs with visual information only, meta-data and combination of meta-data and visual information are 95.73%, 18.23% and 92.55%, respectively.

1 INTRODUCTION

In recent years, social media emerged as an important source of information and communication; especially in disaster situations where usually news agencies are unable to provide information in time due to unavailability of reporters in the area. For instance, the authors in [9, 14] have proved social network as an effective medium of mass communication in emergency situations. A rather recent trend is to infer events from information shared through social media [2, 15]. The analysis of recent literature reveals that social media platforms, particularly Twitter and Flickr have been heavily exploited for inferring information about different types of events, such as social and sports events. In this regards, an interesting application is to collect and analyze information about natural disasters available on social network. To this aim, a number of interesting solutions have been proposed to effectively utilize social media for information collection and analyzing the impact of a natural disaster [4, 12].

On the other hand, satellite images have also been proved very effective to explore and monitor the surface of the earth and its environment [13]. In this regards, Joyce et al. [10] provides a detailed review of different techniques developed to efficiently utilize remote-sensed data for the monitoring and assessment of damage due to natural hazards and disasters.

A rather recent trend is to combine remote-send data with social media information allowing to provide a better overview of a disaster [3, 5]. For instance, in [4], a system called "JORD" is introduced to automatically collect information from different platforms of social media and link it with remote-sensed data to provide a more detailed story of a disaster. Similarly, a task to automatically link social media with satellite images was introduced as a challenge in ACM MM 2016[1].

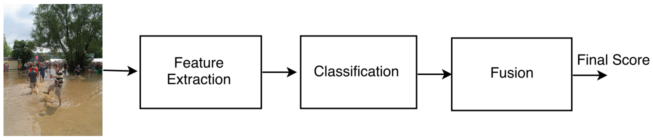

Figure 1: Block diagram of the proposed methodology for DIRSM Task.

This paper provides a detailed description of the method proposed by team MLRDCSE for the first task of Mediaeval2017 Multimedia and Satellite challenge [6]. The basic insight of the task is to jointly utilize satellite imagery and social media as a source of information to provide a detailed story of the disaster. The proposed challenge is composed of two sub-tasks namely (i) Disaster Image Retrieval from Social Media (DIRSM) and (ii) Flood Detection in Satellite Images (FDSI). Detailed description of the tasks are provided in [6].

2 PROPOSED APPROACH

Figure 1 provides a block diagram of the proposed methodology. As can be seen, the proposed approach is composed of three main phases namely feature extraction, classification and fusion. In the next sub-sections, we provide a detailed description of the each phase.

2.1 Feature Extraction

DIRSM is composed of three mandatory runs involving (i) Visual information only (ii) Meta-data only and (iii) combination of meta-data and visual information. For visual information, we extract Convolutional Neural Network (CNN) features via AlexNet [11] pre-trained on ImageNet [8] and Places dataset [16] from each image at hand. AlextNet is composed of 8 fully connected layers including five convolutional and three fully connected layers. The ultimate insight of the proposed scheme for visual information is to utilize both object specific and scene-level information for the representation of the disaster related images. A model pre-trained on ImageNet corresponds to object specific information while the one pre-trained on the places dataset is intended to extract scene-level information. This scheme has also been proved very effective in social event detection in single images [1]. We extract a 4096-dimensional feature vector from each model in caffe toolbox [2]. On the other hand, we also consider user tags, title and GPS information from the available meta-data.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

[1] http://www.acmmm.org/2016/wp-content/uploads/
[2] http://caffe.berkeleyvision.org/tutorial/

2.2 Classification and Fusion

In the proposed methodology, next steps correspond to classification and fusion of the classification results obtained in the previous step. For the classification purposes, we rely on Support Vector Machines (SVM) based on its proven performances in object recognition and classification [7]. We train separate Support Vector Machines (SVM) classifiers for both CNN models on the complete development dataset. Subsequently, test images are classified with the trained classifiers providing results in the form of posterior probabilities. On the other hand, for meta-data, we rely on Random Forest classifier in WEKA Machine Learning library [3]. The trained classifier provide the results in terms of posterior probabilities.

In the subsequent phase, we fuse the scores of the individual classifiers in a late fusion method as shown in Equ. 1 where $w1$, $w2$ and $w3$ represent the weights used for each type of information while $p1$, $p2$ and $p3$ represent the posterior probabilities obtained with classifiers trained with features obtained with AlexNet pre-trained on ImageNet, AlexNet pre-trained on Places dataset and meta-data, respectively.

$$S = w1 * p1 + w2 * p2 + w3 * p3 \tag{1}$$

In the current implementation, we use equal weights for each classifier. Results can be further improved if some optimization techniques, such as Genetic Algorithms (GA), are used. It is important to mention that fusion method is used in both run 1 (fusion of two classifiers trained on features extracted with both CNN models) and run 3 (fusion of all types of information including meta-data and visual information).

3 RESULTS AND ANALYSIS

Table 1 provides the experimental results of our method proposed for the Mediaeval2017 Multimedia and Satellite task in terms of average precision at cut-toffs 480. As can be seen, we achieve best results in run 1, where we use visual information extracted with two different CNN models of AlexNet pre-trained on ImageNet and Places dataset. On the other hand, in Run 2, which is mainly based on meta-data, we achieve the worst results among all runs by having an average precision of just 22.83%. A significant difference of around 64% can be noticed in the performances of meta-data and visual information. This huge difference in the performances shows a clear advantage of visual information over the meta-data in this particular application. Moreover, run 2 also shows the limitations of meta-data. Some common problems with meta-data includes missing time stamps and geo-location information. Moreover, the ambiguous meaning of user's tags also affects the performance of the model.

The third run requires to combine meta-data and visual information. In this experiment, our team achieves an average precision of 83.73%, which is significantly lower than the performance with visual information only. This is mainly caused by equally treating the classifiers trained on visual information and meta-data. This fact can also be concluded from the results of run 2, where the classifier trained on meta-data achieves very low precision.

In Table 2, we provide the experimental results of the proposed method in terms of mean average precision at different cutoffs namely 50, 100, 250 and 480. Again, better results are reported for run 1 relying on visual information only. Similarly, worst results are achieved with meta-data. It can also be noticed in Table 2 that mean average precision at different cutoffs for meta-data is lower than the average precision at maximum cutoff 480. However, in the case of run 1 and run 3 different behaviour can be noticed by achieving better performances at lower cutoffs, which shows the strength of visual information in differentiating among flooded and non-flooded images.

Table 1: Evaluation of the proposed approach in terms of average Precision at cutoff 480

Run	Feature	Avg. Precision
1	Visual Information only	86.81
2	Meta-data only	22.83
3	Meta-data and Visual Information	83.73

Table 2: Evaluations of the proposed approach in terms of mean over avg. precision at different cutoffs (50,100,250,480)

Run	Features	Mean precision
1	Visual Information only	95.73
2	Meta-data only	18.23
3	Meta-data and Visual Information	92.55

Moreover, a significant increase can be noticed in the precision using lower cutoffs (mean of 50,100 and 250, 480), which shows that increasing the cutoff allows false positive to be included in the threshold.

4 CONCLUSIONS AND FUTURE WORK

This paper reports the description of the method proposed by team MRLDCSE along with a detailed description and analysis of the experimental results. For visual information, we rely on the combination of object and scene-level information extracted through two different Convolutional Neural Networks (CNN) models pre-trained on ImageNet and Places datasets. On the other hand, we use user's tags, title, description and geo-location information from the available meta-data. Over all, better results are obtained with visual information only. In contrast, meta-data produce worst results among all the runs we submitted. We also noticed that the inclusion of meta-data degrades the performance of the model when combined with visual information in this particular application.

In the current implementation, we are relying on a single deep architecture, in future, we aim to incorporate multiple deep architectures to better utilize visual information for the retrieval of flooded images. Moreover, very low performance has been noticed with meta-data, in future we aim to employ more sophisticated methods to better utilize the additional information. An other interesting direction can be using some optimization techniques for learning weights of each classifier to fuse them, properly.

[3] http://www.cs.waikato.ac.nz/ml/weka/

REFERENCES

[1] Kashif Ahmad, Nicola Conci, Giulia Boato, and Francesco GB De Natale. 2016. USED: a large-scale social event detection dataset. In *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 50.

[2] Kashif Ahmad, Francesco De Natale, Giulia Boato, and Andrea Rosani. 2016. A hierarchical approach to event discovery from single images using MIL framework. In *Signal and Information Processing (GlobalSIP), 2016 IEEE Global Conference on*. IEEE, 1223–1227.

[3] Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Nicola Conci, Pål Halvorsen, and Francesco De Natale. 2017. JORD: A System for Collecting Information and Monitoring Natural Disasters by Linking Social Media with Satellite Imagery. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 12.

[4] Kashif Ahmad, Michael Riegler, Ans Riaz, Nicola Conci, Duc-Tien Dang-Nguyen, and Pål Halvorsen. 2017. The JORD System: Linking Sky and Social Multimedia Data to Natural Disasters. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 461–465.

[5] Benjamin Bischke, Damian Borth, Christian Schulze, and Andreas Dengel. 2016. Contextual enrichment of remote-sensed events with social media streams. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1077–1081.

[6] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth. The Multimedia Satellite Task at MediaEval 2017: Emergence Response for Flooding Events. In *Proc. of the MediaEval 2017 Workshop* (Sept. 13-15, 2017). Dublin, Ireland.

[7] Hyeran Byun and Seong-Whan Lee. 2002. Applications of support vector machines for pattern recognition: A survey. *Pattern recognition with support vector machines* (2002), 571–591.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[9] Amanda Lee Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *IJEM* 6, 3-4 (2009), 248–260.

[10] Karen E Joyce, Stella E Belliss, Sergey V Samsonov, Stephen J McNeill, and Phil J Glassey. 2009. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography* (2009).

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[12] Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: segment-based event detection from tweets. In *Proc. of ACM IKM*. ACM, 155–164.

[13] Ashbindu Singh. 1989. Review article digital change detection techniques using remotely-sensed data. *International journal of remote sensing* 10, 6 (1989), 989–1003.

[14] Brian Stelter and Noam Cohen. 2008. Citizen journalists provided glimpses of Mumbai attacks. *The New York Times* 30 (2008).

[15] Christos Tzelepis, Zhigang Ma, Vasileios Mezaris, Bogdan Ionescu, Ioannis Kompatsiaris, Giulia Boato, Nicu Sebe, and Shuicheng Yan. 2016. Event-based media processing and analysis: A survey of the literature. *Image and Vision Computing* 53 (2016), 3–19.

[16] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.