CNN and GAN Based Satellite and Social Media Data Fusion for Disaster Detection

Kashif Ahmad[1], Pogorelov Konstantin[2], Michael Riegler[2],Nicola Conci[1],Pal Holversen[2]
[1]DISI-University of Trento, Italy
[2]Simula Research Labs Oslo, Norway
kashif.ahmad@unitn.it,konstantin@simula.no,michael@simula.no,nicola.conci@unitn.it,paalh@ifi.uio.no

ABSTRACT

This paper presents the method proposed by team UTAOS for the Mediaeval 2017 challenge on Multi-media and Satellite. In the first task, we mainly rely on different Convolutional Neural Network (CNN) models combined with two different late fusion methods. We also utilize the additional information available in the form of meta-data. The average and mean over precision at different cut-offs for our best runs are 84.94% and 95.11%, respectively. For challenge two, we utilize a Generative Adversarial Network (GAN). The mean Intersection-over-Union (IoU) for our best run is 0.8315.

1 INTRODUCTION

Linking social media information to remote sensed data holds large possibilities for society and research [1–3]. The Multimedia and Satellite task in Mediaeval 2017 [4] aims to integrate information from both sources, sensed data and social media, to provide a better overview of a disaster. This paper provides a detailed description of the methods developed by the UTOS team for the Mediaeval 2017 Multimedia Satellite Task. The challenge consists of two sub tasks, (i) Disaster Image Retrieval from Social Media (DIRSM) and (ii) Flood Detection in Satellite Images (FDSI).

2 PROPOSED APPROACH

2.1 Methodology for DIRSM Task

To tackle challenge (i), we rely on Convolutional Neural Network (CNN) features. In detail, we first extract CNN features for seven different models from state-of-the-art architectures pre-trained on the ImageNet [5] and places datasets [14]. These models include AlexNet [8] (pre-trained on both ImageNet and places datasets), GoogleNet [12] (pre-trained on ImageNet), VGGNet 19 [10] (pre-trained on both ImagNet and places datasets) and different configurations of ResNet [7] with 50, 101 and 152 layers. For feature extraction from Alexnet and VGGNet19 we use the Caffe toolbox[1] while in the case of GoogleNet and Resnet we exploited Vlfeat Matcovnet[2].

All in all, we extract eight feature vectors through four different network architectures from the same image. AlexNet and VGGNet16 provide a feature vector of size 4096 while GoogleNet and Resnet provide feature vectors of 1024 and 2048, respectively. Subsequently, the extracted features are fed into ensembles of Support Vector Machines (SVMs), which provide classification scores in terms of posterior classification probabilities. We also consider user's tags, date taken along with GPS information from the available meta-data. For the meta-data we rely on Random Tree classifier provided by the WEKA toolbox [6]. Finally, the classification scores obtained through Random Trees and SVM trained on meta-data and visual features are fused using late fusion. For the late fusion we propose two different methods, namely, (i) Induced Ordered fusion scheme inspired by Induced Ordered Weighting Averaging Operators (IOWA) by Yager et al. [13] and (ii) Particle Swarm Optimization (PSO). Figure 1 provides a block diagram of the proposed methodology for the Disaster Images Retrieval from Social Media (DIRSM) task.

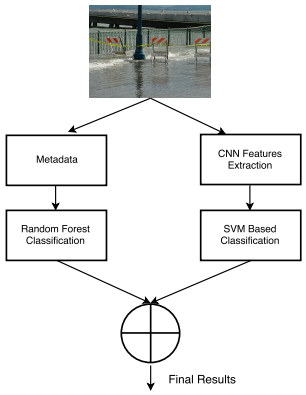

Figure 1: Block diagram of the proposed methodology for DIRSM task.

2.2 Methodology for FDSI Task

For the challenge (ii), we started from the visual analysis of the provided development set. We observed that it is not possible to use any already existing open-source framework due to the nature of the provided satellite data. Furthermore, we observed that the used four-channel 16-bit TIFF file format is too specific and cannot be correctly processed and even viewed by existing libraries.

To perform the visual analysis we developed a conversion code which provide a conversion from geo-TIFF to a pair of images: RGB and infrared (IR). For the RGB images we used the per-three-channels normalization which fits all the R, G and B pixel values of the input geo-image into standard 0-255 RGB region. Normalization coefficients are the same for all three channels to achieve real color balance even in cases of low variations in one of the components. The normalization of the IR component is performed separately.

$$rgb_{min} = \min(\min_{i \in R} r_i, \min_{i \in G} g_i, \min_{i \in B} b_i)$$
$$rgb_{max} = \max(\max_{i \in R} r_i, \max_{i \in G} g_i, \max_{i \in B} b_i)$$

[1]http://caffe.berkeleyvision.org/
[2]http://www.vlfeat.org/matconvnet/

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

$$ir_{min} = \min_{k \in IR} ir_k, \quad ir_{max} = \max_{k \in IR} ir_k$$

$$\forall i \in \{R|G|B\} \quad \{r|g|b\}_i^* = \frac{(\{r|g|b\}_i - rgb_{min}) * 255}{rgb_{max} - rgb_{min}}$$

$$\forall k \in IR \quad ir_i^* = \frac{(ir_k - ir_{min}) * 255}{ir_{max} - ir_{min}}$$

Moreover, we performed a human-expert-driven visual analysis of the images and found them all to be non-contrast, blurry and color-range-limited. From our previous experience [9] we decided to use a generative adversarial network (GAN). GANs[3] are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks contesting with each other in a zero-sum game framework.

As the basis for our method we selected a neural network architecture used for retinal vessel segmentation in fundoscopic images with generative adversarial networks (V-GAN) [4]. The V-GAN architecture is designed [11] for processing of retinal images that have comparable visual properties and provides the required output with one-class image segmentation masks.

V-GAN is implemented in Python on top of Keras with Tensorflow GPU-enabled back-end. We have modified the network architecture by changing the top-layers configuration in order to support four-channel floating-point geo-image-compatible input. The final generator network output layer used for creation of probabilistic output segmentation image was extended by the simple threshold activation layer to generate the binary segmentation map.

First, we have performed experiments with the development set only and found that the modified V-GAN is able to perform the segmentation of the provided satellite images, but the estimated performance metrics were below the expected level. Additional visual analysis of the converted RGB and IR images showed that sometimes IR component of the sourced geo-images was irrelevant to the flooding areas that probably caused our GAN to bias during training process and prevent it from the correct flooding areas properties extraction. Thus, we have decided to exclude IR component from the model input and process only the RGB components of the converted normalized geo-images. This resulted in the significant performance improvement and correct segmentation most of the developments set flooding areas except for the some images taken in not-common lighting and cloudy conditions.

3 RESULTS AND ANALYSIS

3.1 Runs Description in DIRSM Task

For DIRSM, we submitted five different runs. Table 1 provides the official results of our methods in terms of average precision at cutoff 480 and mean over precision at different cutoff (50, 100, 250, 480). Run 1 and run 4 are mainly based on visual information extracted with seven different CNN models and jointly utilized in PSO and IOWA based fusions, respectively. As it can be seen in Table 1, the PSO based fusion method outperforms IOWA with a significant gain of 3.79% and 5.34%. On the other hand, run 2 is based on meta-data achieving the worst results among the all runs. Similarly, run 3 and run 5 represents two different variations of our method used for combining meta-data and visual information. Run 3 is based on IOWA while run 5 represents our PSO based fusion of meta-data and visual information. Again, PSO based fusion performs better. One of the main limitations of IOWA based fusion is its mechanism of assigning more weight to a more confident model. In this particular case, we noticed that our classifier trained on meta-data provides more confident decisions with high probabilities causing significant reduction in the performance. This can also be concluded from the results on run 2 where the meta-data obtain worst results. The degradation in the performance due to the inclusion of meta-data shows that the additional information available are not much useful.

[3]http://en.wikipedia.org/wiki/Generative_adversarial_networks
[4]https://bitbucket.org/woalsdnd/v-gan

Table 1: Evaluations of the proposed approach in terms of precision at 480 and mean over average precision at different cutoffs (50, 100, 250 and 480).

Run	Features	Precision at 480	Mean precision
1	Visual only	84.94%	95.11%
2	Meta-data only	25.88%	31.45%
3	Meta-data and Visual	54.74%	68.12%
4	Visual only	81.15%	89.77%
5	Meta-data and Visual	73.83%	82.68%

Table 2: Evaluations of our approach for Flood Detection in Satellite Images (FDSI) task

Run (Thresh.)	Mean IoU per Location							
	01	02	03	04	05	06	Overall	07 (new)
1 (0.78)	0.79	0.81	0.88	0.78	0.75	0.80	0.82	0.73
2 (0.94)	0.77	0.78	0.86	0.74	0.72	0.78	0.80	0.70
3 (0.5)	0.79	0.82	0.88	0.79	0.76	0.81	0.83	0.74
4 (0.35)	0.79	0.82	0.87	0.79	0.77	0.80	0.83	0.74
5 (0.12)	0.78	0.80	0.86	0.78	0.77	0.78	0.81	0.73

3.2 Runs Description in FDSI Task

Table 2 represents the experimental results of our method for FDSI task. In total, we submitted 5 different runs for 7 different target locations that are represented by image patches of satellite images of different regions affected by flooding. We have used the different binarization threshold level for the different runs with the same model in order to find the optimum balance in the number of false-positive and false-negative pixels in the segmented images. The selection of used threshold values was performed based on the visual analysis of the segmentation results in order to maximize the variability of detected flooding area. The best results are reported for location 03 (which have the best ground visibility without clouds and proper lighting with strong light reflections from the water surface in the flooded areas) in all runs. Overall better results are obtained at runs 3 and 4 with mean IoU of 0.83. For the new location (07) better results are obtained at runs 3 and 4.

4 CONCLUSION AND FUTURE WORK

This paper provides a detailed description of the methods proposed by UTAOS for the Mediaeval 2017 challenge on Multimedia and Satellite. During the experimental evaluation of sub-task 1 (DIRSM), we noticed that visual information seems more useful compared to meta-data for the retrieval of disaster images. For sub-task 2 (FDSI), we rely on a Generative Adversarial Network where better results are obtained in 3 and 4. Based on the experiments conducted in this work we believe that a proper fusion of social media information and satellite data can provide a better story of a natural disaster.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] 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.

[9] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. 2017. Efficient disease detection in gastrointestinal videos–global features versus neural networks. *Multimedia Tools and Applications* (2017), 1–33.

[10] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[11] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. 2017. Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. *arXiv preprint arXiv:1706.09318* (2017).

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[13] Ronald R Yager and Dimitar P Filev. 1999. Induced ordered weighted averaging operators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29, 2 (1999), 141–150.

[14] 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.