Flood level estimation from news articles and flood detection from satellite image sequences

Yu Feng, Shumin Tang, Hao Cheng, Monika Sester
Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany
{yu.feng,hao.cheng,monika.sester}@ikg.uni-hannover.de,shumin.tang@outlook.com

ABSTRACT

This paper presents the solutions of team *EVUS-ikg* for the *Multimedia Satellite Task* at *MediaEval 2019*. We addressed two of the subtasks, namely *multimodal flood level estimation (MFLE)* and *city-centered satellite sequences (CCSS)*. For *MFLE*, a two-step approach was proposed, which retrieves flood relevant images based on global deep features and then detects severe flood images based on self-defined distance features, which can be extracted from human body keypoints and semantic segments. For *CCSS*, a neural network, which combines CNN and LSTM, was used to detect floods in satellite image sequences. Both methods have achieved a good performance on the test set, which shows a great potential to improve the current flood monitoring applications.

1 INTRODUCTION

Flood, as one of the great natural disasters, endangers people's safety and their property. Satellite images are one of the most often used data sources for flood mapping. However, this is not sufficient to obtain enough evidence for the estimation of local flood severity. Crowdsourcing, as a rapidly developing method for data acquisition, has been proved to be beneficial for such a purpose. From social media, flood relevant posts can be retrieved with image and text classifiers trained from deep neural networks (e.g. [8, 11]). However, the information retrieved by now are mostly evidences, further details, such as flood severity, is still desired for many emergency response applications. In the previous *Multimedia Satellite Tasks (MMSat)* at the *MediaEval* benchmarking initiative, several tasks have been proposed regarding flood detection from satellite and social media data. *MMSat'18* [3] provided binary labels showing road passability for tweets with photos, which can be regarded as an early step for extracting local flood severity information. Our solution [9], which simply used early fusion of several pre-trained CNN features, has achieved an average performance compared with the other teams. In *MMSat'19* [1], the subtask *multimodal flood level estimation (MFLE)* goes one step further, which aims to extract news articles only with severe flood situation based on textual and visual information. As for the satellite data, most of the previous research applied semantic segmentation indicating which pixels are water. In order to confirm if it is flooding, an extra water boundary is always needed for a comparison. The differences are not only caused by flood, but also can be caused by the mapping errors or season change. In *MMSat'19* [1], sequences of satellite images are provided for a binary classification of the appearance of flooding events, which could be a more reliable data source.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

2 APPROACH

In *MFLE*, corresponding image and text pairs were annotated with binary labels, which indicate whether the image contains at least one person standing in water above the knee. For **run 1**, where only visual information is allowed, a two-step approach was proposed. A first classifier was trained for extracting flood relevant images and a second classifier was then used to detect the images containing people standing in water above the knee from these relevant images. We concatenated the features extracted from four CNN models, namely *InceptionV3* [15], *DenseNet201* [10], *InceptionResNetV2* [14] pre-trained on ImageNet and VGG16 pre-trained on Place365 [16]. Then, we trained a classifier on these features with Xgboost [7]. Subsequently, all positive predicted images are processed with OpenPose [5] pre-trained on Microsoft COCO [13] dataset for multi-person body keypoint detection and DeeplabV3+ [6] pre-trained on ADE20K [17] for semantic segmentation. During this step, images without persons, or all persons in the image who are without adjacency to ground or water segments, were directly marked as negative. Afterwards, we detected the water line based on the body keypoints and segments, with the steps shown in Figure 1. Finally, the pixel distances from each keypoint to the water line in vertical direction are divided by the body length to calculate the relative distances. These distances were used as the features to represent the relationship between water and single person. After all, we assigned each image annotations to all of the persons in the image and trained a second binary classifier with Xgboost. As for images with multiple persons, the image would be considered "positive" if at least one of the persons is predicted as "positive" by this model.

For **run 2**, where only textual information is allowed, we used a TextCNN model [12] with fasttext [4] word embeddings, which is same as our solution for *MMSat'18* [9]. For **run 3**, where the visual and textual information are fused, only the articles predicted "positive" by both models in run 1 and 2 would be considered as "positive" by this fused model. In **run 4**, we introduced extra data for the visual based model. Since *MMSat'17* [2] provided binary labeled images for training a binary classifier showing flood relevancy. We trained on this augmented dataset to replace the first classifier in run 1.

In *CCSS*, the sequences are collected from 12-bands Sentinel-2 satellite images with date and time. As a pre-processing step, we performed a normalization by calculating the Z-score (subtracting the mean and then dividing by the standard deviation) for each band individually and then clipped the normalized image into the range from -1 to 1. We used DenseNet121 as a feature extractor, and then connected the features using a LSTM with 32 cells in the temporal direction (Figure 2). We used a many-to-many LSTM, where we required an output for each input image individually. The weights of this DenseNet were initialized with the weights pre-trained on

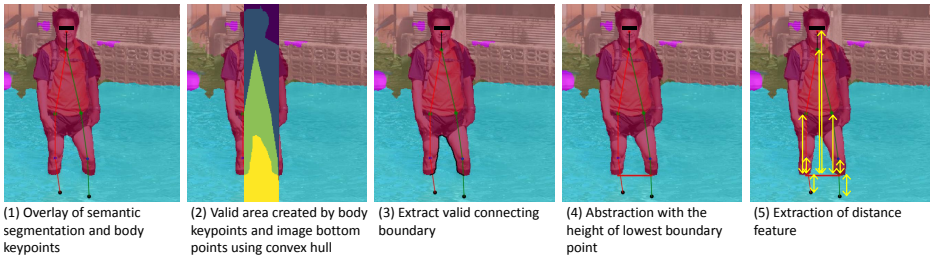

Figure 1: Steps for feature extraction (adapted on image under CC BY-NC-SA 2.0)

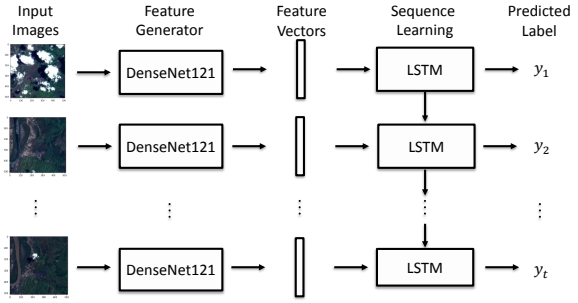

Figure 2: Model for subtask *city-centered satellite sequences*

Table 1: Evaluation on multimodal flood level estimation

Macro-avg. F1-score	Run 1	Run 2	Run 3	Run 4
Development set	73.99%	52.96%	75.56%	73.23%
Test set	68.16%	48.86%	67.27%	68.28%

Table 2: Evaluation on city-centered satellite sequences

Micro-avg. F1-score	Run 1	Run 2	Run 3	Run 4
Development set	97.00%	98.50%	97.75%	68.16%
Test set	92.65%	94.12%	97.06%	60.29%

ImageNet. Since the image sequence lengths are different and do not exceed 24 layers, we padded the sequence to 24 layers with same size tensors with zeros and generate a mask indicating which layers are padded. Then, we excluded these padded images when calculating the categorical cross-entropy loss and accuracy metric. During the training, we used only the RGB channels with a reduced image size 256×256. Since there are 267 sequences available in devset, we used 170 for training, 43 for validation and 54 as an internal testset. Since some of the images in the sequence may be broken or incomplete, the field *FULL-DATA-COVERAGE* in the *timeseries* files can help us to filter these images.

We trained this model with different annotation settings. Each sequence is annotated with binary labels (hereinafter called seq-label), where the field *FLOODING* in the *timeseries* file of each sequence provides the layer-level labels (hereinafter called layer-label). Our primary observation of the training data shows that, the layer-labels indicate a strong correspondence to the seq-labels, where only the sequences annotated with all negative are annotated with negative in the seq-labels. Thus, in **run 1**, we simply used these layer-labels to train this model. Subsequently, we tested different pseudo labels generated from the seq-labels. A repetition of seq-labels in **run 2** (i.e. seq-label is 1, layer-labels are all 1; seq-label is 0, layer-labels are all 0). Since we observed a strong pattern in the positive labeled sequences, that the first half of the seq-labels are negative while the latter half negative. Thus, we followed this pattern to generate pseudo layer-labels in **run 3**, where seq-label is 1, layer-labels are [0, 0, 0, 1, 1, 1] for a sequence of 6 images. For a comparison, instead of using a many-to-many LSTM, **run 4** applied a many-to-one LSTM, where the model is optimized only based on seq-labels.

3 RESULTS AND DISCUSSION

For ***MLFE*** (Table 1), our image based approach can achieve an averaged F1-score of 68.16% on test set. The text based method performs significantly worse than the image based model. Combining both textual and visual information did not improve the F1-score in our case. In run 4, f1-score was improved slightly by introducing additional images for training of the flood relevance classifier. We further exam the failure examples. They can be categorized into three types, namely failed pose detection, failed semantic segmentation and failed water level estimation, where most of them are caused by drawing a wrong water line. This leads to many false positive detections. For ***CCSS*** (Table 2), comparing the results from the first three runs using many-to-many LSTM and many-to-one LSTM in run 4, the improvement is obvious. Regarding the different annotation settings, run 3 achieved the best performance, where the pseudo labels followed annotation patterns in layer-labels. Run 2 has also achieved a better performance than run 1, which indicates the exact layer-labels may not be necessary to predict if a sequence has flood or not.

4 CONCLUSIONS AND OUTLOOKS

In this paper, separate solutions have been proposed for subtask *MFLE* and *CCSS*. Both models can solve the tasks properly according to their performance on test set. For *MFLE*, the water line estimation can be further improved in order to reduce false positive detections. For *CCSS*, the robustness of model can be tested on further events.

ACKNOWLEDGMENTS

This work is supported by project *TransMiT* (BMBF, 033W105A). The computational resource is provided by *ICAML* (BMBF, 01IS17076).

REFERENCES

[1] Benjamin Bischke, Patrick Helber, Simon Brugman, Erkan Basar, Zhengyu Zhao, Martha Larson, and Konstantin Pogorelov. The Multimedia Satellite Task at MediaEval 2019: Estimation of Flood Severity. In *Proc. of the MediaEval 2019 Workshop* (Oct. 27-29, 2019). Sophia Antipolis, France.

[2] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth. 2017. 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.*

[3] Benjamin Bischke, Patrick Helber, Zhengyu Zhao, Jens de Bruijn, and Damian Borth. The Multimedia Satellite Task at MediaEval 2018: Emergency Response for Flooding Events. In *Proc. of the MediaEval 2018 Workshop* (Oct. 29-31, 2018). Sophia-Antipolis, France.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[5] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1302–1310.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV).* 801–818.

[7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* ACM, 785–794.

[8] Yu Feng and Monika Sester. 2018. Extraction of pluvial flood relevant volunteered geographic information (VGI) by deep learning from user generated texts and photos. *ISPRS International Journal of Geo-Information* 7, 2 (2018), 39.

[9] Yu Feng, Sergiy Shebotnov, Claus Brenner, and Monika Sester. 2018. Ensembled convolutional neural network models for retrieving flood relevant tweets. In *Proc. of the MediaEval 2018 Workshop* (Oct. 29-31, 2018). Sophia-Antipolis, France.

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[11] Xiao Huang, Cuizhen Wang, Zhenlong Li, and Huan Ning. 2018. A visual–textual fused approach to automated tagging of flood-related tweets during a flood event. *International Journal of Digital Earth* (2018), 1–17.

[12] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar.* 1746–1751.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision.* Springer, 740–755.

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In *AAAI*, Vol. 4. 12.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2818–2826.

[16] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

[17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*