Detection of Flooding Events in Social Multimedia and Satellite Imagery using Deep Neural Networks

Benjamin Bischke[1,2] Prakriti Bhardwaj[1,2] Aman Gautam[1,2]
Patrick Helber[1,2] Damian Borth[2] Andreas Dengel[1,2]
[1]University of Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Germany
{Benjamin.Bischke, Patrick.Helber, Damian.Borth, Andreas.Dengel}@dfki.de
{p_bhardwaj14, a_gautam14}@cs.uni-kl.de

ABSTRACT

This paper presents the solution of the DFKI-team for the Multimedia Satellite Task at MediaEval 2017. In our approach, we strongly relied on deep neural networks. The results show that the fusion of visual and textual features extracted by deep networks can be effectively used to retrieve social multimedia reports which provide a directed evidence of flooding. Additionally, we extend existing network architectures for semantic segmentation to incorporate RGB and Infrared (IR) channels into the model. Our results show that IR information is of vital importance for the detection of flooded areas in satellite imagery.

1 INTRODUCTION

Satellite imagery is becoming more and more accessible in the recent years. Programs such as *Copernicus* from ESA and *LandSat* from NASA facilitate this development by providing a public and free access to the data. Large-scale datasets such as the EuroSAT-Dataset [9] or the ImageCLEFremote-Dataset [2] have emerged from these programs and build the foundation for the deeper analysis of remotely sensed data. One major problem when analyzing satellite imagery is the sparsity of data for particular locations over time. Publicly available satellites are mostly non stationary and require several days to revisit the same locations. To overcome this problem, recent work leverages the advances of social multimedia analysis and combines the two data sources [14]. Bischke et. al. [3] demonstrated a system for the contextual enrichment of remote-sensed events in satellite imagery by leveraging contemporary content from social media. Similarly, the work by Ahmad et. al. [1] crawled and linked social media data about technological and environmental disasters to satellite imagery.

Building upon these developments and putting a stronger focus on flooding events, Bischke et. al. [4] released the Multimedia Satellite Task at MediaEval 2017. The goal of this benchmarking task is to augment events that are present in satellite images with social media reports in order to provide a more comprehensive view of the event. The task is divided into two subtasks: (1) The *Disaster Image Retrieval from Social Media Task* has the goal to retrieve social media reports that provide direct evidence of a flooding event. (2) *Flood-Detection in Satellite Images* aims to identify regions in satellite images which are affected by flooding.

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

1.1 Disaster Image Retrieval from Social Media

In this section, we present our solution for first subtask by considering visual, textual modalities as well as their fusion. For all modalities, we train a Support Vector Machine (SVM) with a radial basis function (RBF) kernel on the two classes *flooding* and *no flooding*. We obtain the ranked list of relevant social media reports by computing the distance to the decision boundary of the SVM. The features which we used for the classifier training are discussed in detail in the following section.

1.1.1 Visual Features. Motivated by the recent advances of Convolutional Neural Networks (CNNs) to learn a high-level representation of image content, we apply a CNN to obtain the semantic feature representation of images. In particular, we use a pre-trained network *DeepSentiBank* [6] with the X-ResNet [10] architecture. X-ResNet is an extension of ResNet [8] with cross-residual connections to predict multiple related tasks. We extract the internal representation of X-ResNet's *anptask_pool5* layer, resulting in 1000-dimensional feature vector for each image. Compared to CNNs pre-trained on ImageNet [7], this approach has two advantages: (1) DeepSentiBank was trained to predict adjective noun pairs (ANPs). Unlike ImageNet pre-trained models, this allows to not only rely on information about objects-classes but additionally extract details about the image-scence with adjectives (e.g. *wet road, damaged building, stormy clouds*). (2) The domain change of DeepSentiBank is smaller compared to ImageNet pre-trained models. DeepSentiBank was trained on the Visual Sentiment Ontology (VSO) dataset [5], which contains Flickr images similar to the dataset provided by the task organizers. Such images often include more scenic information whereas images from ImageNet mainly contain objects.

1.1.2 Metadata Features. For the retrieval based on only metadata of social media reports, we relied on the tags given by users. We observed that only relying on the presence of single words such as *'flooding' or 'flood'* is not sufficient and introduces a lot of irrelevant social media reports. We therefore combine individual tags to obtain a document representation for each report.

In the first preprocessing step, we remove numbers and convert all tags to lowercase. We then train a Word2Vec model [12] (with 200 dimensions) on the user tags. For each social media report, we average the word vectors and obtain a document representation. In

order to incorporate the importance of each word into the document representation, we additionally weight each word embedding with the term frequency-inverse document frequency (TF-IDF) of the corresponding word. The intuition behind this approach is fairly straightforward, i.e. document vectors containing semantically similar concepts ('flood', 'river', 'damage') should point to a similar direction in the embedding space as compared to documents with word-vectors of different concepts ('flood', 'book', 'desk', 'drink').

1.1.3 Visual-Textual Fused Features. We extract the visual and textual feature representations using the two approaches as described above. The two modalities are fused by concatenating the feature vectors, resulting in a 1200-dimensional vector.

1.2 Flood Detection in Satellite Imagery

In this section, we explain our approach for the segmentation of flooded areas in satellite images using deep neural networks.

1.2.1 Pre-Processing. Before feeding the satellite data to the networks, we perform a location based normalization step. The goal of this step is to remove a location bias due to local changes in images caused by different vegetation, lightning conditions and atmospheric distortions. For each location we compute the mean pixel values of each RGB and IR channel and subtract this value from the corresponding channels of images belonging to the same location. The pixel values in original satellite images are encoded in the 16-bit number format which turned out to be problematic for many frameworks. To overcome this, we additionally scale the *min* and *max* pixel-values channel-wise within the range of 0 and 255.

1.2.2 Network Architectures. We propose three different network architectures for the segmentation problem. All networks use the size of the original image patch (320 x 320 pixels) as input-size and predict classification labels on a pixel-level.

In our first approach, we use a fully convolutional network (FCN) [11] which has a similar architecture as VGG13 [13]. We remove the fully connected layers and attach an up-sampling layer with bilinear interpolation to scale the down-sampled feature maps to the original image-size. An additional convolutional layer is used to predict the class labels for each pixel and classification probabilities are obtained by squashing the network output through a softmax layer. Since the first input layer of VGG13 expects an input tensor with dimension three, we only pass the RGB information of the satellite data into the network. In the second network, we expand our previous architecture by changing the input of the first layer to four channels, allowing the network to incorporate IR information into the prediction. We extend the previous two approaches by investigating into more complex decoders. Therefore, we use the second network as base-model and replace the up-sampling layer with the reversed version of a VGG13 encoder as decoder.

1.2.3 Network Training. In order to train the above described networks from scratch we extend the dataset using data augmentation. Every image patch is flipped (left to right and up down) and rotated at 90 degree intervals, yielding 8 augmentations per image patch. All networks are trained end-to-end with stochastic gradient descend using the negative log likelihood loss, a learning rate of 0.01 and weight decay of 0.0005.

Table 1: Average Precision at 480 and the mean of Average Precisions at different cutoffs for the first subtask (DIRSM).

	Run 1	Run 2	Run 3	Run 4
AP@480	86.64	63.41	**90.45**	74.08
MAP@[50,100,150,240,480]	95.71	77.64	**97.40**	64.50

Table 2: Intersection over Union (IoU) for the second subtask (FDSI). The results are listed for unseen patches covering (i) same locations as in the dev-set and (ii) new locations.

	Run 1	Run 2	Run 3
Same locations	73.56	84.27	**84.36**
New locations	69.32	70.87	**74.13**

2 EXPERIMENTS AND RESULTS

The results for the first subtask are shown in Table 1. *Run 1* is only based on visual information, *Run 2* only on metadata and *Run 3* on the fusion of both modalities as described in Section 1.1. It can be seen that relying on visual information achieves a higher Average Precision (AP) compared to metadata only. At the same time, the fusion of both modalities further helps to improve the retrieval accuracy by 1.7%. *Run 4* uses only visual features from an ImageNet pre-trained ResNet152 model [8]. Compared to *Run 1*, DeepSentiBank (X-ResNet) features perform significantly better.

Table 2 contains the results of the second subtask for unseen satellite images covering same and new locations as in the development set. Each of the three runs corresponds to the three networks as described in Section 1.2.2. Comparing the IoU of the last two networks to the first one (*Run 1*), shows that the IoU increases by more than 10%. This illustrates the importance of the IR-channel for the detection of flooded areas in satellite data. The comparison of the last two networks against each other (*Run 2* vs. *Run 3*) shows that there is a minor improvement of the AP. (0.1% for same and 4% for new locations). The AP's of all runs on new locations demonstrate that the networks generalize to new places.

3 CONCLUSION

In this paper, we presented our approach for the Multimedia Satellite Task 2017 at MediaEval. One major insight is the importance of a multi-modal fusion of text and visual content for the retrieval of social multimedia. In our approach, we analyzed different CNN-features and showed that DeepSentiBank X-ResNet can be used to obtain a powerful image representation. In the second subtask of the challenge, we applied segmentation networks on satellite imagery to extract flooded regions. Our results show that incorporating IR-information is very important. For future work, we would like to extend the satellite imagery to active radar data (Synthetic Aperture Radar) which can "look" through the clouds. We plan to use the results of this work in the future for the monitoring and prediction of flooding events.

ACKNOWLEDGMENTS

The authors would like to thank NVIDIA for support within the NVAIL program.

REFERENCES

[1] 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.

[2] Helbert Arenas, Md Bayzidul Islam, and Josiane Mothe. 2017. Overview of the ImageCLEF 2017 Population Estimation (Remote) Task. (2017).

[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] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 223–232.

[6] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classi□cation with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* (2014).

[7] 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.

[8] 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.

[9] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2017. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classi□cation. *arXiv preprint arXiv:1709.00029* (2017).

[10] Brendan Jou and Shih-Fu Chang. 2016. Deep Cross Residual Learning for Multitask Visual Recognition. In *ACM Multimedia*. Amsterdam, The Netherlands.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Je□rey Dean. 2013. E□cient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

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

[14] Alan Woodley, Shlomo Geva, Richi Nayak, and Timothy Campbell. 2016. Introducing the Sky and the Social Eye. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, Vol. 1739. CEUR Workshop Proceedings.