Global-Local Feature Fusion for Image Classification of Flood Affected Roads from Social Multimedia

Benjamin Bischke[1, 2], Patrick Helber[1, 2], Andreas Dengel[1, 2]
[1] TU Kaiserslautern, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Germany

ABSTRACT

This paper presents the solution of the DFKI-team for the Multimedia Satellite Task 2018 at MediaEval. We address the challenge of social multimedia classification with respect to road passability during flooding events. Information about road passability is an important aspect within the context of emergency response and is not well studied in the past. In this paper, we primarily investigate into the visual classification based on global, local and global-local fused image features. We show that local features of objects can be efficiently used for road passability classification and achieve similar good results with local features as with global features. When we fused global and local visual features, we did not achieve a significant outperformance against global features alone but see a lot of potential for future research into this direction.

1 INTRODUCTION

The Multimedia Satellite Task 2018 [3] continues to focus on flooding events as in last year's Task 2017 [2], since, among high-impact natural disasters, flooding events represent, according to the United Nations Office for the Coordination of Humanitarian Affairs, the most common type of disaster worldwide. The task looks at road passability, namely whether or not it is possible to travel through a flooded region. This work focuses on social multimedia and is based on the benchmark dataset, that contains 7.387 tweets with accompanying images and labels for evidence of road passability as well as the actual road passability (passable vs. non passable).

2 APPROACH

Our solution for classifying Tweets with respect to road passability follows a two-step approach. We first categorize all images that provide evidence for road passability during a flooding event and then classify the relevant images with respect to road passability. Our approach is only based on the visual modality, since we could not obtain any meaningful results by taking the metadata of Tweets (e.g. text, location) into consideration.

2.1 Evidence classification of flood passability

The approach for the evidence classification of images relies on last year's solution [1] for the Multimedia Satellite Task 2017 [2]. The goal of the challenge was to retrieve all images from a Flicker dataset that provide evidence of a flooding event. We applied a pre-trained CNN to obtain the feature representation of images and used a SVM, with a radial basis function (RBF) kernel, as classifier. One important insight of our approach was the importance of the dataset on which the network was pre-trained on. We achieved a significant improvement when relying on a network that was trained on scene-level information rather than object classes as in the ImageNet dataset. Building upon this approach, we evaluted models pre-trained on different datasets containing scene-level and object-level classes for the visual classification of flood passability evidence. We achieved the best results on our internal valiation set with features extracted from a Wide-Resnet38 pre-trained on Places365 [8], and obtained an improvement of 3% against the features of ResNet152 pre-trained on ImageNet [6]. These findings are in line with the insights from last year's solution [1].

Figure 1: Local image features of objects provide a strong evidence for the classification of road passability (left: passable, right: non passable).

2.2 Flood passability image classification

In this paper we investigate three strategies for the road passability classification of images. We use a SVM (RBF kernel) as classifier and visual features based on the following approaches:

(1) Global features of CNNs pre-trained on Places365 [8], ImageNet [6] and the Visual Sentiment Ontology (VSO)[4]
(2) Local features of objects extracted with Faster R-CNN [7] pre-trained on Pascal VOC [5]
(3) Fusion of global and local features

Global Features.
We follow the same approach as described in section 2.1 and extract global image features with pre-trained CNNs. We analyzed models pre-trained on ImageNet [6], Places365 [8] and VSO [4] datasets and obtained the best results set with scene-level features (VSO and Places365) on the internal validation (see Table 1).

Local Features.
In our second stragey we investigated into local image features. Our hypothesis is that local features corresponding to objects and its surrounding context such as cars, persons, traffic signs shown in Figure 1 provide a high evidence for the discrimination of road passability. We trained the object detection network Faster R-CNN [7] on the dataset Pascal VOC [5] and applied it on the images of the provided Twitter dataset. Whenever Faster-RCNN identified an instance for one of the following classes C={*bus, boat, person, car*},

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

we croped based on the bounding box of the particular object a small patch out of the image. We combined the patches for *bus, car* and *boat* classes into one dataset and resized all patches to the same size of 224x244 pixels. The three classes covered 45% of images in the development with at least one object. Based on the created dataset, we trained a CNN with two road passability classes that follows the same arichtecture as LeNet with a kernel size of 7x7 on all convolutional layers. In the case, that Faster-RCNN detected multiple objects in the image, we followed a late-fusion approach, in which we calculated the mean of the predictions and mapped values above 0.5 to the passable road class.

We tried to classify the 3275 patches beloning to objects of the person class, but our classifier was not siginicantly better than random guessing. By visual inspection we noticed that there are a lot of variations in the image patches of *persons* that made it also very difficult for us to classify single patches with respect to road passability. The dataset contains, for example, images with persons being fully visible (evidence of passability) and at the same time persons walking in the water at hip height (evidence of no passability). Since we were not able to achieve sufficient results for patches of *persons*, we suppressed this class in our current approach and leave it open for future research.

Global-Local Feature Fusion.
In our third strategy, we combined global and local features. We extracted the global features as described in section 2.1 and appened to this vector the prediction for local features from section 2.2. In case that no local feature could be extracted from the image, we appended a special label to the vector. The resulting feature vector was classified with a SVM (RBF kernel) as described in 2.1.

3 EXPERIMENTS AND RESULTS

We first evaluated our three approaches on the internal validation set. Table 1 shows the results for the classification of road passability using the F1-score as metric. In the table, we can see that (1) for global features the best results are achieved with scene-level features from the VSO, followed by Places365 and then ImageNet. (2) The classification using local features performed with 77.24% similar good as the global features (in the range between 73.29% – 79.16%) and better compared to features extracted from ImageNet pre-trained models. However, it is also worth mentioning that this comparison is not completly fair, since the dataset using local features was smaller as not very image contained the local features. (3) For the global local feature fusion, we see an small improvement when using features of *Places365 Wide-ResNet38* and an decrease of the performance using *VSO X-ResNet50 Adjective* features.

The final results on the private test set are shown in Table 2. **Run 1** are the results for the global feature *VSO X-ResNet50 Adjective*, **run 2** for the same feature but fused with local predictions and **run 3** for features from *Places365 Wide-ResNet38* fused with local predictions. The offcial metric is the average F1-scores of (C1) images with the evidence and passable roads as well as (C2) images with evidence and non passable roads. In the table, we can see that the fusion of local and global feature information slightly decreased the results for the *VSO X-ResNet50 Adjective* feature whereas on the *Places365 Wide-ResNet38* feature the opposite can be observed, simialar as on the internal validation set. Results show that the global local fusion did neither significantly improve nor worsen the results. We believe that this can be achieved with a more sophisticated fusion strategy and better local features. The resizing of extracted patches to 244x244 pixels for different objects could have a negative influence on the classification, since the aspect ratio of objects can get distorted. A deeper network for classifying local patches could additionally improve the results.

Table 1: Road passability classification based on global, local and fused features on the internal validation set

Approach	F1 score
ImageNet ResNet101 (global)	73.29%
ImageNet ResNet152 (global)	75.98%
VSO X-ResNet50 Adjective (global)	79.16%
VSO X-ResNet50 Noun (global)	78.97%
Places365 Wide-ResNet38 (global)	77.31%
local features based on Faster R-CNN	77.24%
Places365 Wide-ResNet38 (global) + local features	78.60%
VSO X-ResNet50 Adjective (global) + local features	78.45%

Table 2: Results on the internal test set for the F1-score of evidence vs. no evidence for passability (row 1) and average F1-score of evidence passable and evidence non passable (row 2)

	Run 1	Run 2	Run 3
evidence vs. no evidence	87.70%	87.70%	**87.70%**
passable vs. non passable	65.21%	64.96%	**66.48%**

4 CONCLUSION

In this paper, we presented our approach for the Multimedia Satellite Task 2018 at MediaEval. In line with previous research [1], we also observed the advantages of scene-level features compared to object related features when classifying images with respect to road passability. We could confirm our hypothesis in this work and showed that local features of a few object instances corresponding to classes in Pascal VOC can be used for the visual classification for road passability. We achieved similar good results with local features as with global features and see a lot of potential to improve our current approach.

When we fused global and local visual features, we did not achieve a significant outperformance against global features alone. We strongly believe that better and more general local features, which can be extracted from more than only half of the images, play an important role in this context. We will continue this work with additional classes that are not covered by Pascal VOC dataset. One direction would be to extract semantic segmentation classes from a model pre-trained on the Cityscape dataset. This dataset contains additional classes such as traffic signs, poles, and that could be important for road passability classification as well.

ACKNOWLEDGMENTS

The authors would like to thank NVIDIA for support within the NVAIL program. Additionally, this work was supported BMBF project DeFuseNN (01IW17002).

REFERENCES

[1] Benjamin Bischke, Prakriti Bhardwaj, Aman Gautam, Patrick Helber, Damian Borth, and Andreas Dengel. 2017. Detection of flooding events in social multimedia and satellite imagery using deep neural networks. In *Proceedings of the Working Notes Proceeding MediaEval Workshop, Dublin, Ireland*. 13–15.

[2] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth. The Multimedia Satellite Task at MediaEval 2017: Emergency 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] 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.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[6] 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.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[8] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2018), 1452–1464.