Exploiting Local Semantic Concepts for Flooding-related Social Image Classification

Zhengyu Zhao, Martha Larson, Nelleke Oostdijk
Radboud University, Netherlands
z.zhao@cs.ru.nl,m.larson@cs.ru.nl,n.oostdijk@let.ru.nl

ABSTRACT

In this paper, we present an approach to identification of the images that depict passable and non-passable roads, from a collection of flood-related tweet images. Our key insight is that the local information from domain-specific concepts (‘boat’, ‘person’ and ‘car’) can be exploited to help determine whether an image depicts a location that is passable. We use concept detection as the basis for features that encode local information. We use conventional features, i.e., presence of concepts and visual features extracted from the concept region, but also a novel light-weight feature, i.e., the aspect ratio of the bounding box. Experimental results show that integrating local semantic information yields slightly better performance than only using image-level CNN representation. Text features are not competitive.

1 INTRODUCTION

Despite achieving impressive performance in various visual recognition tasks, convolutional neural network (CNN) representations do not fully capture local-level discriminative information when only trained at a single scale, i.e., input size of 224x224 for most conventional CNNs. In order to complement global CNN features, recent work on fine-grained object classification [5, 6, 8, 12] and scene recognition [2, 9–11] has also tried to exploit discriminative information from local semantic regions. Building on these insights, here, we demonstrate that the task of differentiating two road conditions (passable vs. non-passable) [1] will also benefit from local semantic information. Our starting point is the observation that images with similar global appearance have differentiable local patterns, as shown in Figure 1. Intuitively, we consider that three specific concepts (‘boat’, ‘person’ and ‘car’) will show different properties in the context of road passability. Moreover, based on our exploratory experiments, we observed that the images containing the three concept classes account for a large proportion (46%) of the passability-relevant images. As shown in Figure 2, the images with these three concepts span over the entire passability-relevant dev-set without any specific bias related to time order, which is reflected by the numerical order of the tweet ID. These two observations indicate that using local information from these concepts is not accidental but can be generally applicable.

2 APPROACH

We start with a light-weight approach by only using text information. By manual inspection of the patterns in the dev-set, we created a set of rules that apply to a vocabulary that has been annotated with basic part-of-speech and semantic-word class information. On the basis of these rules, we create a set of ngrams, which represents strings of lexical items that we would expect to occur in tweets related to road passability. Whenever any created ngram is encountered in the text, the associated class label is assigned (either passable or non-passable). As we target mostly texts indicating that roads are not passable, there are only few ngrams that yield the label passable. In the case of no matching, the image will be regarded not relevant to road passability.

For the visual-based approach, the basic pipeline is hierarchical classification with two SVM classifiers. The first classifier is applied to differentiate the images that are relevant to road passability from the others. Here we only use image-level features extracted from a ResNet50-based CNN model, which is pre-trained on the large-scale

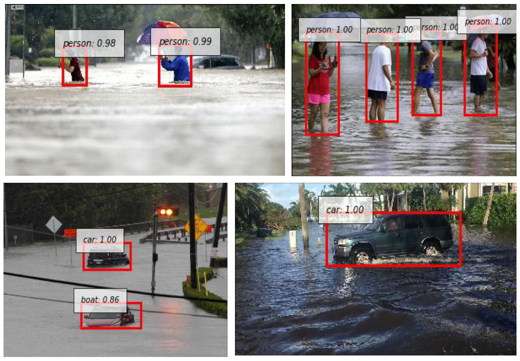

Figure 1: Image examples showing the contrast in visual properties of ‘person’ and ‘car’ in non-passable (left column) vs. passable (right column) classes.

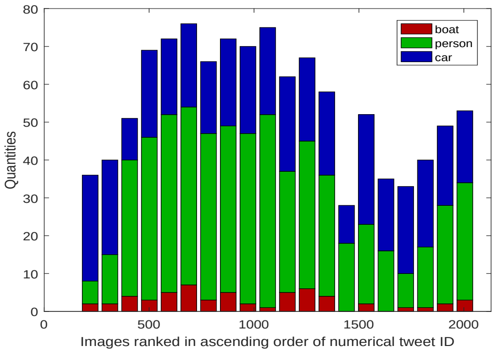

Figure 2: Distribution of the three concept classes over the whole subset of passability-relevant images in the dev-set.

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

scene-centric database Places2 [13]. Exploratory experiments on the dev-set showed that this option performed better than using the object-centric ImageNet [3] as pre-training data. Then, the second classifier will further predict the images that have been classified as relevant into passable or non passable classes. Here, we use both the Places2 and ImageNet as the pre-training data, resulting in better performance than using only one of them. This result suggests that discriminative information from scene-level and object-level will complement each other for differentiating passable vs. non-passable images.

Alternatively, we add a pre-filtering step before the second classifier that allows test images containing the three concepts to be treated differently. We adopt the state-of-the-art YOLOv3 [7], which is pre-trained on the union of VOC2007 and VOC2012 trainval set [4], for automatic concept detection. In order to capture differences accurately, we exclude the image candidates that have incomplete bounding boxes in the image area, or a confidence score below 0.9. When multiple instances are detected in one image, we use the average values of their features as the final feature.

Since 'boat' is not a conventional means to pass a road, the presence of any boat in the image indicates the road is very likely to be non-passable. So we use the +/- presence of 'boat' in an image as a feature. The experiments on the dev-set show that boats can be detected in 46 of 1179 non-passable images, and only in 5 of 951 passable images.

The subtle differences in local information can also be encoded by a single value derived from concept bounding boxes. Specifically, we look at the height-width aspect ratio of the bounding box, since we observed that the person or car will be more likely to be stuck in water on the non-passable road, resulting in a lower aspect ratio. In this paper, we set two empirical thresholds for 'person'. We classify images with aspect ratios lower than the first threshold (T1=1.37) as non-passable, and images with aspect ratios higher than the second threshold (T2=2.98) as passable. Since the aspect ratio of the front/back view of a car could be plausibly with a respectively high value, we only apply one threshold (T3=0.30). We classify images with aspect ratios lower than this threshold as non-passable. Figure 3 shows the precision-recall curves of passable vs. non-passable classification as T1 and T3 change. For better visualization, we balanced the number of images from the two classes by upsampling the minority class.

Furthermore, we conjecture that the local information could also be learned by a CNN based on the visual content enclosed by the concept bounding box. We apply this for 'car', for which the subtle differences of appearance are not well reflected by the aspect ratios as described above.

3 EXPERIMENTS

3.1 Run submissions

Run 1 is our text-based approach. Run2, run3 and run4 only use visual information and also use SVM classifiers for a two-stage classification. We use the same method for the first stage of each of these three runs. For run 2, in the second stage, only image-level features are leveraged. For run 3, in the second stage, we add a pre-filtering step, which use the +/- presence of 'boat' and aspect ratio-based method for both 'person' and 'car' as local features. Run 4 follows the same process as in run 3, but instead of aspect ratios, we use deep visual features extracted from the bounding box region of 'car' to train a SVM classifier for pre-filtering. Note that no local features for 'boat' and 'car' are used for this run.

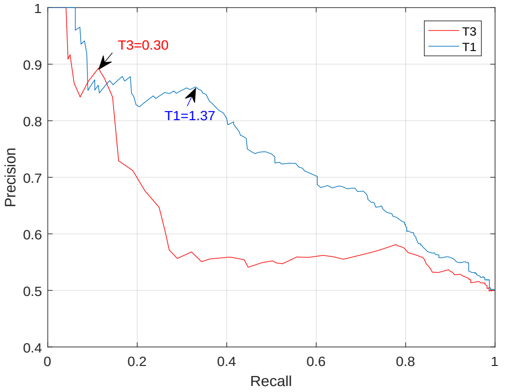

Figure 3: Precision-Recall curves with varying values for T1 (person) and T3 (car), where the arrows point out the positions, where the specific values were set.

Table 1: Average testset scores for evidence vs. non-evidence (Ave-F1_1) and passable vs. non-passable (Ave-F1_2)

	Run 1	Run 2	Run 3	Run 4
Ave-F1_1	0.3260	**0.8758**	**0.8758**	**0.8758**
Ave-F1_2	0.1286	0.6313	**0.6389**	0.6388

3.2 Experimental analysis

Table 1 shows the evaluation results of our 4 runs. Since the annotation of 'road passability' is based on visual inspection of the images associated with the tweets, it is not surprising the tweet text did not make strong contribution. In particular, we noticed that people often discuss in the tweet whether it is legally allowed to pass a road rather than whether the road is physically passable. Also, the text does not necessarily pertain to the type of the image or what is depicted in the image. For the visual information, we can observe that slightly better performance could be achieved by exploiting additional local information in the two methods that we applied.

4 CONCLUSION

In this paper, a new approach was proposed to capture local-level information from specific semantic concepts for better identification of Twitter images that depict passable and non-passable roads. Specifically, we explored two different types of features based on the light-weight summary of the output of the concept detector, i.e., aspect ratio of the bounding box, or visual features derived from the bounding box. From the analysis of the text-based approach, we concluded that the text information might be useful if we would in the future, be looking at other aspects of evidence about road passability.

REFERENCES

[1] Benjamin Bischke, Patrick Helber, Zhengyu Zhao, Jens de Bruijn, and Damian Borth. 2018. The Multimedia Satellite Task at MediaEval 2018. In *Proc. of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29-31 October 2018*.

[2] Xiaojuan Cheng, Jiwen Lu, Jianjiang Feng, Bo Yuan, and Jie Zhou. 2018. Scene recognition with objectness. *Pattern Recognition* 74 (2018), 474–487.

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

[4] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The PASCAL visual object classes challenge: A retrospective. *International journal of computer vision (IJCV)* 111, 1 (2015), 98–136.

[5] Xiangteng He, Yuxin Peng, and Junjie Zhao. 2017. Fine-grained discriminative localization via saliency-guided Faster R-CNN. In *ACM International Conference on Multimedia (ACM MM)*. 627–635.

[6] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. 2016. Part-stacked CNN for fine-grained visual categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1173–1182.

[7] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[8] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. 2018. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition* 76 (2018), 704–714.

[9] Ruobing Wu, Baoyuan Wang, Wenping Wang, and Yizhou Yu. 2015. Harvesting discriminative meta objects with deep CNN features for scene classification. In *International Conference of Computer Vision (ICCV)*. 1287–1295.

[10] Guo-Sen Xie, Xu-Yao Zhang, Shuicheng Yan, and Cheng-Lin Liu. 2017. Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 27 (2017), 1263–1274.

[11] Zhengyu Zhao and Martha Larson. 2018. From Volcano to Toyshop: Adaptive Discriminative Region Discovery for Scene Recognition. In *ACM International Conference on Multimedia (ACM MM)*.

[12] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. *International Conference of Computer Vision (ICCV)* (2017), 5219–5227.

[13] 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 (TPAMI)* 40 (2018), 1452–1464.