Ensembled Convolutional Neural Network Models for Retrieving Flood Relevant Tweets

Yu Feng, Sergiy Shebotnov, Claus Brenner, Monika Sester
Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany
{yu.feng,claus.brenner,monika.sester}@ikg.uni-hannover.de,shebotnov@gmail.com

ABSTRACT

Social media, which provides instant textual and visual information exchange, plays a more important role in emergency response than ever before. Many researchers nowadays are focusing on disaster monitoring using crowd sourcing. Interpretation and retrieval of such information significantly influences the efficiency of these applications. This paper presents a method proposed by team *EVUS-ikg* for the *MediaEval 2018* challenge on *Multimedia Satellite Task*. We only focused on the subtask *"flood classification for social multimedia"*. A supervised learning method with an ensemble of 10 Convolutional Neural Networks (CNN) was applied to classify the tweets in the benchmark.

1 INTRODUCTION

Crowdsourcing is a rapidly developing method for acquiring information from many users in real time. Many applications nowadays are focusing on monitoring natural disaster events such as earthquakes, fires and flooding. The retrieved information can improve the situation awareness for citizens. At the same time, it helps the rescuers to provide a better emergency response. Flooding is one of the topics which attracts lots of attention. With the development of information retrieval and deep learning techniques, instead of using pre-defined keywords for extracting flood relevant information, deep learning models can achieve much better performance for visual and textual information understanding.

In our previous work [5], a method which considers both predictions from separately trained text and image classifiers was used to extract flood and heavy rainfall relevant information from twitter data. However, an end-to-end classification approach, which can directly fuse the information, seems more attractive. Some of the teams [1, 8, 10] from the *Multimedia Satellite Task* [2] at *MediaEval 2017* have already achieved end-to-end solutions. Well-performing end-to-end classifiers have been trained for flickr data with binary labels (with evidence and non evidence for flooding).

More information regarding the floods, such as severity, is still desired. *Multimedia Satellite Task* at *MediaEval 2018* [3] provided the binary labels for the tweets (with evidence and non evidence for road passability). For the tweets with road passability evidence, the benchmark dataset also provided the labels for road passability. Most tweets in this dataset were labeled as non evidence (3,685 tweets). The number of tweets labeled as passable and not passable are 946 and 1,179, respectively.

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

2 APPROACH

In this section, our approach is introduced. All the models are trained using the Tensorflow and Keras frameworks. We randomly selected 10% from the given dataset (582 tweets) as an independent internal test set. Moreover, 60 tweets from each label were randomly selected and used as a validation set. All the remaining tweets were used to train the models. The network architectures and parameters were tuned and compared internally based on the performance of the models on the internal test set. The validation set was used for early stopping during training. Data augmentation, such as rotation, shift, and zoom, was also performed during the training process.

Run 1 allows only visual information to be used for the classification task. Pre-trained models *DenseNet201* [6], *InceptionV3* [12] and *InceptionResNetV2* [11] were used as basic feature extractors. They were all trained based on the ImageNet dataset and achieved a top-5 accuracy of 0.936, 0.937 and 0.953, respectively. We froze the weights of these pre-trained models and concatenated the nodes at the layers before the output layers. Followed by two dense layers with batch normalization and dropout of 50%, we produced an output of three nodes with the softmax function. The architecture of our model is shown in Figure 1.

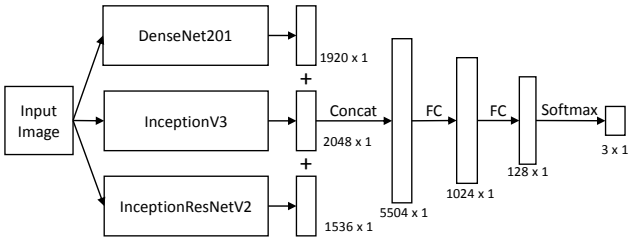

Figure 1: CNN model using visual information only

Run 2 allows only metadata information to be used for the classification task. In this case, only user provided tweet texts are used as inputs. We applied a classic CNN model [7] for natural language processing, combined with the word embeddings of *fasttext* [4]. This embedding contains one million word vectors trained on Wikipedia 2017, UMBC webbase corpus, and statmt.org news dataset [9]. Each word vector contains 300 dimensions. Since most of the texts in the training dataset are no more than 21 words after pre-processing (e.g. removing stop words, url, emojis), we limited the maximum allowed sentence length n to 21. For sentences with less than 21 words, we used zero padding to obtain a fixed size input (21 x 300) for the network. Convolutional filters were then applied on this embedding matrix to extract feature maps for each sentence. After experiments with different filter sizes, the combination of filter

sizes 1, 3, 5 performed best on our internal test set. The architecture of the model is shown in Figure 2.

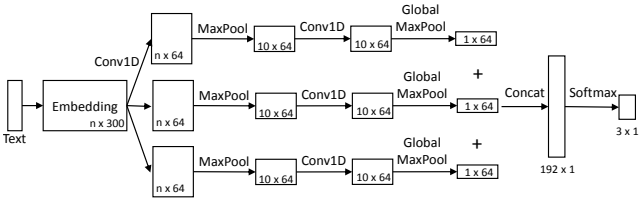

Figure 2: CNN model using metadata information

Run 3 allows metadata-visual fused information to be used for tweets classification. Pre-trained models on ImageNet were used as visual feature extractors in the same way as in Run 1. We cut the visual model at the dense layer with 1024 nodes. For the metadata information, we used the same CNN structure as presented in Run 2. The output of the text classifier was the dense layer with 192 nodes. We simply concatenated the features from both modalities and used the softmax function to derive an output of 3 nodes.

Run 4 is a general run, where we only used the visual information. Same architecture as Run 1 was used. Since the flood related tweets have only a very small proportion among the daily user-sent tweets, we wanted to investigate whether introducing a large amount of negative labeled images can improve the robustness of the classifiers for the current task. Therefore, the 3,349 negative labeled flickr photos from the *Multimedia Satellite Task* at *MediaEval 2017* were used as an extra data source to improve the robustness of the model against the tweets with no evidence.

Run 5 is also a general run, where we only used the visual information. Same architecture as Run 1 was used. We observed that many of the positive labeled images from the *Multimedia Satellite Task* at *MediaEval 2017* were describing severe flood situations. Therefore, in this run we assigned the label (2) with evidence "not passable" to these 1,916 positive labeled photos. Together with the 3,349 negative labeled photos, we trained an image classifier. In this way, we introduced imprecise labeled training examples. Our intention is to investigate how much the performance of classifiers is affected through introducing more data but with imprecise labels.

Since the random initialization of the weights in neural networks often leads to unstable performance of the models, during **Run 1, 2 and 3**, each model was trained 10 times on the same training set. In this case, we regard these 10 models as weak classifiers. We ensemble the predictions and take the majority voting of all the 10 predictions as the final prediction. In this case, we hope the ensemble learning improves the robustness of the classifiers. For **Run 4 and 5**, the models were trained only once due to the much longer training time.

3 RESULTS AND DISCUSSION

As for the subtask "flood classification for social multimedia", we firstly tested the performance of the classifiers based on our internal test set. The results are shown in Table 1. The evaluation on the private test set, which was provided by the organizer, is shown in Table 2. The metrics used for evaluation are the averaged F1-scores for the classes (1) with evidence passable and (2) with evidence not passable. Additionally, we also listed the F1-scores separately for both classes, and the accuracy on our internal test set.

Table 1: Evaluation on internal test set

	Run 1	Run 2	Run 3	Run 4	Run 5
Avg. F1-score	66.94%	41.05%	61.23%	54.67%	49.89%
F1-score (1)	61.08%	43.37%	57.14%	53.02%	33.10%
F1-score (2)	72.80%	38.73%	65.32%	56.31%	66.67%
Accuracy	81.93%	55.33%	75.90%	75.73%	75.90%

Table 2: Evaluation on private test set

	Run 1	Run 2	Run 3	Run 4	Run 5
Avg. F1-score	64.35%	32.81%	59.49%	52.16%	51.59%

From the results above, the classifiers have generally similar performance on both internal and private test set. The visual based classifier (Run 1) can achieve the best performance, compared to all the other runs. The models trained only on metadata information do not achieve a good performance on both test sets. The fusion of both models did not make any significant improvements according to the evaluation. From the results of Run 4 and 5, we noticed that introducing more negative examples or more examples with imprecise labels, lead to a significantly worse performance. Therefore, we conclude that the balance of the training examples plays an important role for the classifier performance.

Since the textual descriptions from users rarely address the severity of the flood situation, it is very hard to achieve a reasonable classification performance only based on textual information. Textual information may contain informative words or phrases regarding flood evidence. However, distinguishing whether the road is passable or not from a single tweet text, which is less than 280 characters, is a challenging task. In our case, we concluded that introducing textual information did not help for the current task.

4 CONCLUSIONS AND OUTLOOK

In this paper, an ensemble of CNN models was trained for retrieving flood relevant tweets. Our best model is the one trained only on visual information. Using only metadata, the classifiers were not able to produce high quality predictions. Using photographs is reasonable, since nowadays people are likely to share photographs to address their current situation, rather than detailed textual descriptions. The analysis of video sequences would be a promising extension for extracting more information regarding flooding events, such as rainfall intensity, flow speed or even water depth.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from the BMBF funded research project "EVUS - Real-Time Prediction of Pluvial Floods and Induced Water Contamination in Urban Areas" (BMBF, 03G0846A). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of a GeForce Titan X GPU used for this research.

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.*

[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] 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.

[6] 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).*

[7] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[8] Laura Lopez-Fuentes, Joost van de Weijer, Marc Bolanos, and Harald Skinnemoen. 2017. Multi-modal deep learning approach for flood detection. In *Proc. of the MediaEval 2017 Workshop (Sept. 13–15, 2017). Dublin, Ireland.*

[9] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

[10] Keiller Nogueira, Samuel G Fadel, Ícaro C Dourado, Javier AV Muñoz, Otávio AB Penatti, Rodrigo Tripodi Calumby, Lin Li, and Jefersson Alex dos Santos. 2017. Data-Driven Flood Detection using Neural Networks.. In *Proceedings of the Working Notes Proceeding MediaEval Workshop, Dublin, Ireland.*

[11] 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.

[12] 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.