Multi-modal Deep Learning Approach for Flood Detection

Laura Lopez-Fuentes[1,2,3], Joost van de Weijer[2], Marc Bolaños [4], Harald Skinnemoen[3]
[1]University of the Balearic Islands, Palma, Spain
[2]Autonomous University of Barcelona, Barcelona, Spain
[3]AnsuR Technologies, Oslo, Norway, [4]Universitat de Barcelona, Barcelona, Spain
l.lopez@uib.es,joost@cvc.uab.es,marc.bolanos@ub.edu,harald@ansur.no

ABSTRACT

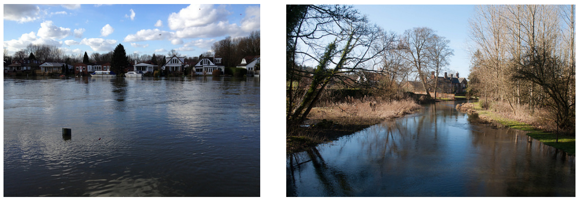
In this paper we propose a multi-modal deep learning approach to detect floods in social media posts. Social media posts normally contain some metadata and/or visual information, therefore in order to detect the floods we use this information. The model is based on a Convolutional Neural Network which extracts the visual features and a bidirectional Long Short-Term Memory network to extract the semantic features from the textual metadata. We validate the method on images extracted from Flickr which contain both visual information and metadata and compare the results when using both, visual information only or metadata only. This work has been done in the context of the MediaEval Multimedia Satellite Task.

1 INTRODUCTION

The growth in smart phone ownership and the almost omnipresent access to Internet has empowered the rapid growth of social networks such as Twitter or Instagram, where sharing comments and pictures has become part of our daily lives. Using the vast amount of data from social media to extract valuable information is a hot topic nowadays [8]. In this work we will focus on extracting information to facilitate the task of emergency responders during floods. Images coming from citizens during a flood could be essential for emergency responders to have situational awareness. However, given the tremendous amount of information posted in social networks, it is necessary to automatize the search of relevant information corresponding to floods. Therefore, in this work we propose an algorithm for the retrieval of flood-related posts. As stated in [4] algorithms for flood detection have received little attention in the field of computer vision. There exist two major trends in this direction: algorithms based on satellite images [5–7] and algorithms based on on-ground images [3]. In this work we will focus on on-ground images taken by humans in the flooded regions and posted on social networks and therefore containing metadata. To the best of our knowledge, there is no published previous work on multi-modal flood detection. However, combining image and text features has recently received great attention to solve tasks such as image captioning, multimedia retrieval or visual question answering (VQA). The work presented in this paper has been inspired by the VQA model presented in [1].

2 DATA

The dataset used in this work was introduced for the MediaEval 2017 Multimedia Satellite Task [**?**], and contains 6600 images extracted from the YFCC100M-Dataset [10] which have been classified as having evidence of flood or not. All the images are associated with metadata information, from which we will take into account for the task the name of the photo and the description and user tags, if available. In Figure 1 we give an example of an image classified as having evidence of flood and one which has been classified as not having evidence of flood and the metadata that we keep for the analysis associated to both.



(a) Classified as containing evidence of flood. Metadata: Image title: "Floods in Walton on Thames", image description: "Most of those houses looked like they had been flooded.", tags: None

(b) Classified as not containing evidence of flood. Metadata: Image title: "The closest we have got to the flooding disaster", image description: None, tags: "freefolk"

Figure 1: Example of an image with evidence of flood and an image with no evidence of flood with the associated metadata that has been considered relevant for the task. Note that although the second image has no evidence of flood it also contains water and the word "flood" in the metadata, which makes the classification harder.

Moreover, since the initial dataset was unbalanced and had approximately 60% of images with no evidence of flood we have downloaded the three top images from the Google Similar Images search engine using as input images from the dataset classified as having evidence of flood. Then we have manually removed incorrect images ending up with 989 extra images with evidence of flood. We do not have the corresponding metadata of these images, so they will only influence the visual part of the algorithm.

3 APPROACH

In this section we will discuss the deep learning algorithm design for the task of flood evidence retrieval in social media posts. The problem will be approached under a probabilistic framework. As explained in Section 2, the posts contain an image and/or metadata. To extract rich visual information we apply the convolutional InceptionV3 network, using the pre-trained weights on ImageNet [2] and fine-tune the last inception model of the network. For the metadata we use a word embedding to represent the textual information in a continuous space and feed it to a bidirectional LSTM. The word embedding is initialized using Glove [9] vectors, which we fine-tune with our metadata. Finally, we concatenate the image

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

Table 1: Average Precision (AP) on the test set on the first 480 images retrieved as flood and the Mean AP at 50, 100, 250 and 480 cutoffs

	AP @ 480	Mean AP @ 50,100,250 and 480
Metadata only	67.54	70.16
Image only	61.58	66.38
Metadata and Image	81.60	83.96
Metadata and Image with extra images	68.40	75.96

and text features followed by a fully connected layer and a softmax classifier to give a final probability of the sample containing relevant information about a flood. In Figure 2 we show a sketch of the multimodal system, which can also be applied using only one of the modalities.

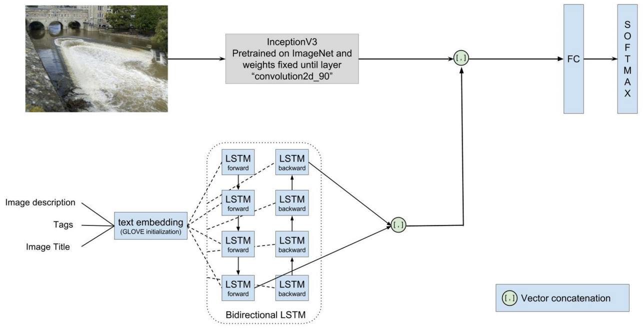


Figure 2: Visual representation of the proposed algorithm.

4 EXPERIMENTS

We have divided the development set in training (3960 + 989 extra flood images) and validation (1320). As for the optimizer we have chosen RmsProp [11] wich uses the magnitude of recent gradients to normalize the gradients, and set an initial learning rate of 0.001.

Since the dataset does not have a very large number of training data it is common to run into overfitting problems. In order to avoid this problem we have used the validation set to determine when to stop the training. Thus, it is stopped when the performance on the validation set stops increasing or starts decreasing over the last two epochs. Then we have used that number to retrain the system using the training and the validation set. We have followed this procedure for all the experiments.

We have trained the system in four different configurations: 1) having images and metadata as input, 2) having only images as input and 3) having only the metadata as input, and 4) having images and metadata in addition to the extra images obtained from Google Similar Images. The results on the test set of these four experiments are given in Table 1. The system has been evaluated as a retrieval task. All the posts from the test set have been given a probability of containing evidence of flood and have been put in order from higher probability to lower. In the first column of Table 1 we show the Average Precision (AP) of posts which have been classified as containing evidence of flood in the first 480 retrieved posts. In the second column we show the mean over average precision when evaluated on the first 50, 100, 250 and 480 posts.

5 RESULTS AND ANALYSIS

As can be seen in Table 1 the metadata that we have selected for the task is certainly relevant to retrieve information about the evidence of flood-related images in social network posts, reaching over 70% mean precision over 4 retrieval cutoffs. Since the classification of the posts as containing evidence of flood or not has been manually done using only the images, the image information should be enough for the retrieval problem. However the performance of the algorithm using only the image goes down to 66% in mean over average precision at different cutoffs. This shows that although images should be more discriminative for this task, due to the difficulty of processing images in comparison to text, the metadata analysis gives better performance. There is also a clear improvement when combining both types of information, reaching almost a 84% accuracy in mean over the average precision in several cutoffs which shows that the metadata and the image complement each other quite well. Surprisingly, when training the system with extra images, the Mean AP drops to 76%, since the images have been manually inspected to make sure that there were no noisy images added to the dataset, this makes us suspect that that result degrades when adding images without metadata, as this performs the weakest among all experiments, however it should be further studied before drawing additional conclusions.

6 DISCUSSION AND OUTLOOK

In this paper we have proposed a multi-modal deep learning approach to retrieve posts from social networks containing valuable information about floods. The system can work using only visual information, only text or combining both types of information.

It has been shown that combining both types of information improves greatly the performance of the system. For future work it would be interesting to check if other type of metadata could also provide useful information for the task, as for example the location or time where the image was taken since there are regions and seasons which are more prone to flooding. It would also be interesting to study why adding more images to the training set has worsened the performance of the system and how well does the system generalize to images outside of the dataset.

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Grants TIN2016-75404-P AEI/FEDER, UE, TIN2014-52072-P, TIN2016-79717-R, TIN2013-42795-P and the European Commission H2020 I-REACT project no. 700256. Laura Lopez-Fuentes benefits from the NAERINGSPHD fellowship of the Norwegian Research Council under the collaboration agreement Ref.3114 with the UIB. Marc Bolaños benefits from the FPU fellowship.

REFERENCES

[1] Marc Bolaños, Álvaro Peris, Francisco Casacuberta, and Petia Radeva. 2017. VIBIKNet: Visual bidirectional kernelized network for visual question answering. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 372–380.

[2] 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.

[3] CL Lai, JC Yang, and YH Chen. 2007. A real time video processing based surveillance system for early fire and flood detection. In *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE*. IEEE, 1–6.

[4] Laura Lopez-Fuentes, Joost van de Weijer, Manuel González-Hidalgo, Harald Skinnemoen, and Andrew D. Bagdanov. 2017. Review on Computer Vision Techniques in Emergency Situations. *arXiv preprint arXiv:1708.07455* (2017).

[5] Sandro Martinis. 2010. *Automatic near real-time flood detection in high resolution X-band synthetic aperture radar satellite data using context-based classification on irregular graphs*. Ph.D. Dissertation. lmu.

[6] David C Mason, Ian J Davenport, Jeffrey C Neal, Guy J-P Schumann, and Paul D Bates. 2012. Near real-time flood detection in urban and rural areas using high-resolution synthetic aperture radar images. *Geoscience and Remote Sensing, IEEE Transactions on* 50, 8 (2012), 3041–3052.

[7] David C Mason, Rainer Speck, Bernard Devereux, Guy JP Schumann, Jeffrey C Neal, and Paul D Bates. 2010. Flood detection in urban areas using TerraSAR-X. *Geoscience and Remote Sensing, IEEE Transactions on* 48, 2 (2010), 882–894.

[8] Michael Mathioudakis and Nick Koudas. 2010. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 1155–1158.

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation.. In *EMNLP*, Vol. 14. 1532–1543.

[10] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[11] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-RMSProp, COURSERA: Neural networks for machine learning. *University of Toronto, Tech. Rep* (2012).