Flood Severity Estimation in News Articles using Deep Learning Approaches

Dan Bînă, George-Alexandru Vlad, Cristian Onose, Dumitru-Clementin Cercel
Faculty of Automatic Control and Computers
University Politehnica of Bucharest, Romania
{binadanvalentin96,georgealexandruvlad,onose.cristian,clementin.cercel}@gmail.com

ABSTRACT

The aim of this study is to detect flooding events by analyzing both texts published by African online news outlets as well as the accompanying article images. The data is provided by MediaEval 2019 within the Multimedia Satellite Task. Our contributions are related to the image- and text-based subtasks. In order to solve the required classification subtasks, we build models capable to extract features from images and texts separately, and then combine them to obtain a complex classifier, providing a better evidence of flooding. Specifically, we adopt the MobileNet architecture which is based on convolutional layers for image processing and also employ a robust text processing method based on long short-term memory cells. The results of our final models on the official test sets are promising, 85.26% average F1-score on the first subtask and 66.19% average F1-score on the second subtask.

1 INTRODUCTION

The flood phenomenon is a problem that humanity is still facing, especially in some disadvantaged territories from Africa. An automated method of estimating the severity of the flood phenomenon could significantly improve prevention policies and rescue measures taken in addressing its consequences.

The Multimedia Satellite Task is organized as part of the MediaEval benchmarking initiative[1] and comprises of various subtasks in the flooding event analysis domain. After the success of the two previous editions [2, 3], the 2019 edition of Multimedia Satellite Task [1] includes three new challenging subtasks as follows: (i) Image-based News Topic Disambiguation (INTD), (ii) Multimodal Flood Level Estimation from News Articles (MFLE), and (iii) Binary Classification of city-centered satellite sequences. In our work, we only investigate the first two subtasks. While the INTD subtask refers to the detection of flooding events with the help of on-ground images using visual attributes, the MFLE subtask involves the detection of people standing in water below the knee using features extracted from both images and texts.

In recent years, the field of flooding event analysis has attracted a lot of interest by researchers. For example, Rizk et al. [22] reported an accuracy of 91.10% for a Support Vector Machine classifier [5] taking as input a concatenated vector of handcrafted semantic features, such as color histogram, gradient direction histogram, hue saturation intensity, in a classification task of disaster-related Twitter images. A better performance of 92.62% is achieved in classifying damage-related social media posts by Mouzannar et al. [18] using a multimodal approach based only on convolutional neural networks for both images and texts. Lopez-Fuentes et al. [17] proposed a multimodal solution for the flood detection from social media posts on Flickr. The architecture of this solution consists of combining a convolution neural network, InceptionV3 [24], with a BiLSTM network [10] for the extraction of visual and textual features respectively.

[1] http://www.multimediaeval.org/

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

2 PROPOSED APPROACH

We define a neural architecture for image processing for both subtasks. For the MFLE subtask, we also define a neural model for text processing and a multimodal one which combines all obtained features. The details of our neural network architectures are further explained in this section.

Visual Feature Extraction. We use a deep convolutional neural network, namely MobileNet [14], with weights pre-trained on the ImageNet dataset [7] in order to extract visual features. The MobileNet model is relatively small compared with the state-of-the-art models [12], but this is rather important since the competition data set has a small size for a deep learning approach. The MobileNet network has been proven to be effective in several tasks such as vehicle recognition from short-range aerial images [13], traffic density estimation [4], skin cancer classification [6], and underwater pipeline damage identification [23]. During the learning process, we freeze all the MobileNet layers until the next to last convolutional layer because the pre-trained network can extract high-level features, for instance edges, and add two fully connected layers capable of detecting task-specific features.

Textual Feature Extraction. First, we obtain a vector representation of the words in texts using the GloVe model [20]. The GloVe embeddings are built based on an occurrence matrix which stores the number of times a word is found within the context of a given word. The context is limited to a window size within the input sequence. The word embeddings are generated based on the probability of co-occurrence of two words in the same context.

Next, we define an architecture with an initial GloVe embedding layer, followed by a BiLSTM layer [9, 10], which is capable to recognize complex features. The BiLSTM layer consists in two LSTM layers concatenated together which are processed forward and backward over the input sequence of word embeddings in order to capture past and future information likewise. We focus on the BiLSTM architecture since it has achieved good results in different tasks including dialect identification [19], answer selection [25], essay scoring [16], rhyme detection [11], and spelling correction [27]. After the BiLSTM layer, we use a max polling layer with the role of keeping most important features for each dimension. Finally,

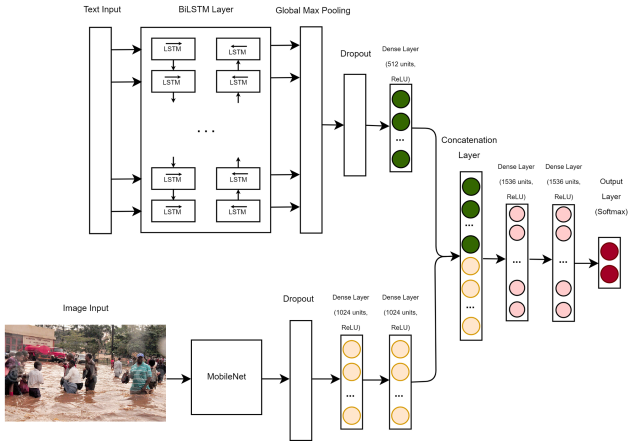

Figure 1: An overview of our multimodal architecture.

we use a fully connected layer to combine previous characteristics using a softmax activation function.

Fusion of Textual and Visual Features. In order to deal with the multimodal fusion, we use the two unimodal neural networks previously specified in the same architecture, without the prediction layers. As shown in Figure 1, the concatenated outputs of the two neural networks are fed into two fully connected layers to get a deeper representation of the text-level and image-level features at the same time. The visual feature vector extracted by the network is double the size of the textual one so that the visual input would have a greater contribution to the final categorical class inference. Our choice is based on the better performance obtained using only visual information in comparison to the result obtained considering solely text information.

3 EXPERIMENTS AND RESULTS

Next, we describe the experiments that we performed using the defined models in order to solve the two subtasks.

Data Augmentation. We have randomly split the provided dataset in train data, 80%, and validation data, 20%. Because the dataset is unbalanced and rather small for a deep learning approach, only 564/5181 entries for the INTD subtask, and 157/4932 entries for the MFLE subtask, we augment the positive items in order to improve our model performance. For each image and text, we added four more items that we generated as follows. In case of the image augmentation, we randomly applied a number of geometric transformations: (i) rotation with 20 degrees, (ii) horizontal reflection and changing, (iii) translation with maximum 10%, and (iv) the range of brightness between 0.9 and 1.2. For text augmentation, we use the Google Translate service in order to translate each text from English to two other intermediary languages, i.e., French, Spanish, Portuguese, German and then back to English, which results in different worded texts but with the same semantics. Because the title of a publication might contain relevant information on the subject of the writing, we choose to concatenate it to the content of the corresponding news article.

Text Pre-processing. Before introducing data in the learning process, the news articles must be pre-processed considering the fact that their content might contain insignificant information. We clean the text removing unused symbols, and special strings such as e-mails, links or abbreviations. Lastly, we lowercase the text, remove stop words, apply lemmatization and tokenize the words.

Experimental Settings. Regarding hyperparameters, we use a grid search in order to get the best configuration. In case of pre-trained models, we keep the recommended values. For the BiLSTM model, we use 300-dimensional word vectors pre-trained on the Common Crawl corpus[2] and the LSTM size of 300. We use Dropout layers with 20% deactivated neurons to prevent overfitting. As optimization method, we use Stochastic Gradient Descent (SGD) with a learning rate of 1e-5 for the only image processing, training for 30 epochs. In both approaches, i.e., text unimodal and multimodal, we use Adaptive Moment Estimation (Adam) [15] with the same learning rate, but training for 10 epochs. The batch size used in all cases is 64.

Results. It is important to note that we submitted all our neural networks trained on the whole provided dataset in order to increase their performance. Table 1 presents the results we obtained on the test dataset. The best score for the MFLE subtask is obtained by the MobileNet model. We consider that the lower performance on the MFLE subtask, compared to the INTD subtask, is attributed to the more complex features that are required to detect people standing in water below the knee. We expected the combined image and text network to perform the best. Unfortunately, the learned semantic features offer little to no improvement over the only image model.

Table 1: Results of our models on the official test data for both Subtasks

Subtask	Run	Averaged F1-Score
INTD	Image	85.26%
MFLE	Image	66.19%
MFLE	Text	52.43%
MFLE	Image and Text	62.38%

4 CONCLUSIONS

This paper presents our solution for two subtasks of the Multimedia Satellite Task: detecting if an image describes a flooding event and if a news article, image and text, depicts a person standing in water below the knee. To solve these subtasks, we propose solutions based on neural networks, namely the BiLSTM model with pre-trained word embeddings GloVe for extracting textual features and the MobileNet network for extracting visual features, respectively. Although the obtained results are competitive, we plan to improve them. For this purpose, we will focus on extending the BiLSTM architecture with an attention mechanism [26] and also contextualized word representations such as Bert [8] or Elmo [21], rather than the Glove embeddings used in our work.

ACKNOWLEDGMENTS

The work was supported by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

[2]https://nlp.stanford.edu/projects/glove/

REFERENCES

[1] Benjamin Bischke, Patrick Helber, Simon Brugman, Erkan Basar, Zhengyu Zhao, Martha Larson, and Konstantin Pogorelov. 2019. The Multimedia Satellite Task at MediaEval 2019: Estimation of Flood Severity. In *Proc. of the MediaEval 2019 Workshop, Sophia-Antipolis, France.*

[2] Benjamin Bischke, Patrick Helber, Christian Schulze, Venkat Srinivasan, 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, Dublin, Ireland.*

[3] Benjamin Bischke, Patrick Helber, Zhengyu Zhao, Jens De Bruijn, and Damian Borth. 2018. The multimedia satellite task at MediaEval 2018: Emergency response for flooding events. In *Proc. of the MediaEval 2018 Workshop, Sophia-Antipolis, France.*

[4] Debojit Biswas, Hongbo Su, Chengyi Wang, Aleksandar Stevanovic, and Weimin Wang. 2019. An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD. *Physics and Chemistry of the Earth, Parts A/B/C* 110 (2019), 176–184.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[6] Saket S Chaturvedi, Kajol Gupta, Prakash Prasad, and others. 2019. Skin Lesion Analyser: An Efficient Seven-Way Multi-Class Skin Cancer Classification Using MobileNet. *arXiv preprint arXiv:1907.03220* (2019).

[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 *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[9] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 6645–6649.

[10] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602 – 610. IJCNN 2005.

[11] Thomas Haider and Jonas Kuhn. 2018. Supervised Rhyme Detection with Siamese Recurrent Networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, Santa Fe, New Mexico, 81–86.

[12] 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.

[13] Yuhang He, Ziyu Pan, Lingxi Li, Yunxiao Shan, Dongpu Cao, and Long Chen. 2019. Real-Time Vehicle Detection from Short-range Aerial Image with Compressed MobileNet. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8339–8345.

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*. 1–13.

[16] Guoxi Liang, Byung-Won On, Dongwon Jeong, Hyun-Chul Kim, and Gyu Sang Choi. 2018. Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture. *Symmetry* 10, 12 (2018).

[17] 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, Dublin, Ireland.*

[18] Hussein Mouzannar, Yara Rizk, and Mariette Awad. 2018. Damage Identification in Social Media Posts using Multimodal Deep Learning.. In *ISCRAM.*

[19] Cristian Onose, Dumitru-Clementin Cercel, and Stefan Trausan-Matu. 2019. SC-UPB at the VarDial 2019 Evaluation Campaign: Moldavian vs. Romanian Cross-Dialect Topic Identification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*. 172–177.

[20] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[21] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[22] Yara Rizk, Hadi Samer Jomaa, Mariette Awad, and Carlos Castillo. 2019. A computationally efficient multi-modal classification approach of disaster-related Twitter images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. ACM, 2050–2059.

[23] Jiajun Shi, Wenjie Yin, Yipai Du, and John Folkesson. 2019. Automated Underwater Pipeline Damage Detection using Neural Nets. In *ICRA 2019 Workshop on Underwater Robotics Perception.*

[24] 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.

[25] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. LSTM-based Deep Learning Models for Non-factoid Answer Selection. In *Proc. of ICLR.*

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[27] Yingbo Zhou, Utkarsh Porwal, and Roberto Konow. 2017. Spelling Correction as a Foreign Language. (2017). arXiv:cs.CL/1705.07371