A multimodal approach in estimating road passability through a flooded area using social media and satellite images

Anastasia Moumtzidou[1], Panagiotis Giannakeris[1], Stelios Andreadis[1], Athanasios Mavropoulos[1], Georgios Meditskos[1], Ilias Gialampoukidis[1], Konstantinos Avgerinakis[1], Stefanos Vrochidis[1], Ioannis Kompatsiaris[1]
[1]CERTH-ITI, Greece
{moumtzid,giannakeris,andreadisst,mavrathan,gmeditsk,heliasgj,koafger,stefanos,ikom}@iti.gr

ABSTRACT

This paper presents the algorithms that CERTH-ITI team deployed to tackle flood detection and road passability from social media and satellite data. Computer vision and deep learning techniques are combined in order to analyze social media and satellite images, while word2vec is used to analyze textual data. Multimodal fusion is also deployed in CERTH-ITI framework, both in early and late stage, by combining deep representation features in the former and semantic logic in the latter so as to provide a deeper and more meaningful understanding of the flood events.

1 INTRODUCTION

The high popularity of social media around the world and the large streams of satellite data that are openly available can be considered as useful sources in the case of natural disasters, such as floods, hurricanes and fires. Several H2020 projects, such as beAWARE [2] and EOPEN [3], already apply their technologies on one or both of these kind of sources to extract knowledge and assist civil protection agencies to monitor a flood event and have a holistic view of an area during an emergency event.

Evidence and road passability recognition is performed in "Flood classification for social multimedia" dataset within Multimedia Satellite Task 2018, which contains a list of tweets with images for the three big Hurricane events of 2017. While, flood detection in satellite images from the same events was also performed in the compiled satellite dataset [6]. CERTH contribution involves the implementation of recent computer vision and deep learning techniques, which analyze social media and satellite images to identify evidence of flooded regions and perform road passability classification. CERTH also deploys deep learning algorithms for textual recognition, while low and high level fusion was also performed in the acquired textual and visual data in order to leverage both contexts and get a more meaningful flood classification outcome.

Flood detection from social media images has also been presented in our previous work [5] that took place in last year's MediaEval Satellite task competition and more recently in [7]. For textual classification a more recent work is adopted, based on word2vec [9] representation, which includes the use of novel architectures and models for producing word embeddings (i.e. representation of words from a given vocabulary as vectors in a low-dimensional space), based on deep neural networks (NN), namely the Continuous Bag-of-Words (CBOW) and the Skip-gram models. Semantic networks and lexical knowledge bases, such as WordNet [4] and ConceptNet [1], provide useful, multilingual representations and interconnections among terms, named entities, concepts, and relations. They can be used for word-sense disambiguation and context enrichment, creating graph-based semantic interpretations by linking candidate meanings, such as the outputs of visual analysis, with lexical resources, creating semantic signatures. Regarding the analysis of satellite image, our approach is based on applying DCNN on satellite data, following e.g. [10] on with Sentinel-1 imagery for oil spill identification.

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

2 APPROACH

2.1 Analyzing social media images

Two separate Deep Convolutional Neural Networks (DCNN) were trained and evaluated in order to carry out each one of the 'evidence' and 'passability' analysis levels, while the VGG architecture [13] was adopted in both of them for the sake of extracting deep features of images in a holistic manner. The first model seeks for evidence in the images and classifies them between relevant or non-relevant in the context of road passability. Thereafter any images that pass the first check are fed to the second model that classifies between images showing passable or non-passable roads.

During the learning phase, we initialized our models with the previously learned weights of a VGG architecture acquired from Places365 scene recognition dataset [15]. Furthermore, 5 splits of the MediaEval 2018 development set were made so as to perform cross validation and select the best epoch to stop training our models. The best parameter results were 6 epochs for the 'evidence' model, and 15 epochs for the 'passability' model.

2.2 Textual analysis of social media

The textual analysis initially involves a preprocessing step of the given text by applying tokenization, stop word removal, and word stemming, and then text representation by using word2vec [9] method. Parameter selection was deployed both for selecting vector dimension (i.e. 50, 100, 200, 300, 400, 500, 600) and window size (i.e. 2, 3, 4). Furthermore, we exploited a set of Twitter posts that have been collected inside the scope of the beAWARE project.

Before finalizing the content of the corpus, we tried to filter out tweets that are irrelevant to actual events of floods. First, we removed texts in which the keyword "flooding" is used metaphorically, by defining phrases often met in Twitter, e.g. "flooding my timeline". Next, we removed all texts where words of hateful communication or sexual intent appear, based on an available list of

dirty words onlineThe final step of textual analysis involves serving the text feature vector as input to a classifier (i.e. SVM, Naïve Bayes or Random Forests) which is tuned.

2.3 Early fusion of visual and textual features

A multimodal analysis approach is also explored here by combining the information that is provided in the text and the accompanied images in social media tweets.

A novel scheme was designed in order to fuse Deep CNN visual features and text features and produce a single feature per tweet that will then be used for classification purposes. For the extraction of the feature vector from the text, we followed the same procedure as described in the previous section. As far as the visual analysis, the activations of the last fully connected layer of the VGG network was chosen as the feature extractor. The feature map there is a 4096-dimensionality vector. Our scheme follows closely the bi-modal stacked AutoEncoder of [14], but with the addition of an extra fully connected layer attached to the DCNN framework, used for classifying the fused feature vector.

For all the hidden layers *tanh* activations are used and for the output reconstruction layers linear activations are used so that the network is able to reconstruct accurately the input features that are of arbitrary range. Stochastic gradient descent (SGD) is used for the optimization with a learning rate of 0.01 and momentum equal to 0.9. A separate model was trained for 5000 epochs.

2.4 Flood detection and Semantic enrichment

The semantic event fusion is based on the use of concepts annotating the images. Specifically, each image is annotated with concepts from a predefined concept pool of 345 concepts (TRECVID SIN concepts) and each concept is accompagnied with a score that indicates the probability that it appears in the image. To obtain such scores, we used a DCNN that was trained according to the 22-layer GoogLeNet architecture on the ImageNet 2011 dataset for 5055 categories. Then, we fine-tuned the network on the 345 concepts by using the extension strategy proposed in [12].

In an effort to semantically enrich the context of the predefined concept pool, we mapped each concept to WordNet and ConceptNet resources. For each term t_w in WordNet, we create a vector with the synsets that belong to the hierarchy of hypernyms of t_w (up to the third level). For each ConceptNet term t_c, the pertinent vectors contain all the terms in the knowledge graph that are considered relevant to t_c with a plausibility score above 80%. These vectors are then used to semantically enrich the annotations derived by visual analysis, by adding semantically relevant concepts.

2.5 Road passability from satellite images

In order to classify satellite images to the class "road passability" we built models by using a pretrained ResNet-50 [8] DCNN. ResNet-50 uses residual functions to help add considerable stability to deep networks, and its input are $224x224$ images. Then we fine-tuned it by removing the last pooling layer and attached a new pooling layer with a softmax activation function with size 2. The NN was trained on 1000 images and validated on the remaining 437 images. It should be noted that several experiments were run in order to find the best performing model. The parameters that were tuned were the following; the learning rate, the batch size and the optimizer function. The epoch value was set to 15 and the loss function considered was the sparse categorical crossentropy. To evaluate the performance of the different networks we considered accuracy as the evaluation metric and the results showed that the parameters of best performing network were the following: SGD as optimizer function, 0.001 as learning rate and 10 as batch size.

Table 1: Evaluation Results

Run submissions	Averaged F1-Score (%)
Flood classification for social multimedia	
Visual	66.65
Textual	30.17
Early fusion	66.43
Semantic Enrichment (ConceptNet)	55.12
Semantic Enrichment (WordNet)	54.48
Road passability estimation from satellite images	
Visual	56.45

3 RESULTS AND ANALYSIS

At this point, it should be noted that our system was tested using embeddings not only based on the word2vec model (predictive), but also on the Glove [11] model (count). The latter, while it performed quite well on our problem, it did not manage to outclass the word2vec results. The most probable reason is that in order to perform optimally, Glove needs to be trained over more data than what is available in our dataset. The size of the text is also an issue for the text classification part.As far as visual analysis, it is safe to assume that some of the most difficult samples were pictures that didn't contain any civilians or vehicles inside flooded roads, as features of large water bodies are not informative enough to provide information about water depth. In a higher level of conclusions, we can see that the visual component overpassed all the others, including early and late fusion of the low level data. That infers that the visual indications can provide more meaningful and less ambiguous information than the text that accompanies Twitter.

4 DISCUSSION AND OUTLOOK

Our participation in Social Media Satellite Task, gave CERTH the opportunity to test and enhance its algorithms in computer vision, textual analysis and semantic fusion in realistic datasets. The results of these challenge highlight to us that DCNN can provide very meaningful results for flood detection in both tasks and especially when visual context is taken under consideration. We plan to deploy more sophisticated fusion techniques that will be able to leverage the low level (text, visual) information in a more efficient way.

ACKNOWLEDGMENTS

This work was supported by EC-funded projects H2020-700475-beAWARE and H2020-776019-EOPEN.

REFERENCES

[1] ConceptNet - An open, multilingual knowledge graph. =http://conceptnet.io.

[2] H2020, beAWARE project. https://beaware-project.eu/.

[3] H2020, eOPEN project. https://eopen-project.eu/.

[4] Princeton WordNet 3.1. http://wordnet-rdf.princeton.edu.

[5] Konstantinos Avgerinakis, Anastasia Moumtzidou, Stelios Andreadis, Emmanouil Michail, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2017. Visual and textual analysis of social media and satellite images for flood detection@ multimedia satellite task MediaEval 2017. In *Proceedings of the Working Notes Proceeding MediaEval Workshop, Dublin, Ireland*. 13–15.

[6] 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.

[7] Panagiotis Giannakeris, Konstantinos Avgerinakis, Anastasios Karakostas, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2018. People and vehicles in danger-A fire and flood detection system in social media. In *IEEE Image, Video, and Multidimensional Signal Processing (IVMSP) Workshop*.

[8] 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.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[10] Georgios Orfanidis, Konstantinos Ioannidis, Konstantinos Avgerinakis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2018. A Deep Neural Network for Oil Spill Semantic Segmentation in SAR Images. In *ICIP*. IEEE, 3773–3777.

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[12] Nikiforos Pittaras, Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. 2017. Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In *International Conference on Multimedia Modeling*. Springer, 102–114.

[13] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[14] Pengfei Zhang, Xiaoping Ma, Wenyu Zhang, Shaowei Lin, Huilin Chen, Arthur Lee Yirun, and Gaoxi Xiao. 2015. Multimodal fusion for sensor data using stacked autoencoders. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*. IEEE, 1–2.

[15] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).