Transfer Learning and Mixed Input Deep Neural Networks for Estimating Flood Severity in News Content

Pierrick Bruneau[1], Thomas Tamisier[1]
[1]LIST, Luxembourg
{pierrick.bruneau,thomas.tamisier}@list.lu

ABSTRACT

This paper describes deep learning approaches which use textual and visual features for flood severity detection in news content. In the context of the MediaEval 2019 Multimedia Satellite task, we test the value of transferring models pre-trained on large related corpora, as well as the improvement brought by dual branch models that combine embeddings output from mixed textual and visual inputs.

1 INTRODUCTION

Identifying news items related to a catastrophic event such as a flood, and assessing the severity of the event using the collected information can provide timely input to support the victims. The *News Image Topic Disambiguation* (NITD) and *Multimodal Flood Level Estimation* (MFLE) subtasks of the *Multimedia Satellite* (MMSat) MediaEval 2019 task foster the application of machine learning research to this context. The MMSat overview paper [2] discloses further information on this matter. To address these subtasks, we do not propose any specialized model (e.g. combined use of pose detection and occlusion detection [3] that could be used for MFLE subtask). Rather, we reuse existing general-purpose text and image classification models. In particular, the value of adapting pre-trained models to these subtasks is estimated in this work.

Part of the literature on multimodal neural networks aims at learning similarities between modalities such as image and text e.g. for automatic image captioning [8]. In the present work, multimodality is understood as the joint usage of several modalities (i.e. text and image) as means to improve prediction capabilities. In other words, embeddings derived from each modality are merged, and fed forward to a sigmoid function typical of classification models [1, 9]. Runs submitted to the MFLE subtask are meant to measure the improvement brought by such multimodality. In the remainder, after shortly introducing the addressed subtasks, implementation rationale and details are presented, and the respective experimental results are disclosed and commented.

2 DATA

The NITD subtask aims at predicting the flood-relatedness of news articles using their featured images as input. The training set contains 5180 images, with $\sim$ 10.1% flood-related images. The test set contains 1296 images. The MFLE subtask aims at classifying news articles w.r.t. flood severity using both the news text and featured images as input. The training set features 4932 news articles, with $\sim$ 3.2% of instances from the positive class (i.e. high severity). The test set features 1234 articles. For details about the subtasks, e.g. regarding the annotation of training and test sets, the reader may refer to the workshop overview paper [2].

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

3 PROPOSED APPROACH

For given training and validation sets, each model in this work was trained for 50 epochs. Model selection was performed by monitoring a validation metric at epoch end, and retaining the best model w.r.t. this metric. As both subtasks are significantly imbalanced, instead of the accuracy, we used the F1 metric. We optimized the binary cross-entropy loss using the Adam solver [7]. Batches of 32 elements were used. For each run, we used stratified 5-fold cross-validation, hence retaining 20% of validation data in each fold. A model ensemble was built using the model selected for each fold. Majority voting or score averaging was selected depending on the F1-Score on the full training data. For handling class imbalance in the context of neural networks, we used instance weighting. In the Multimedia Satellite overview paper [2], runs that use only the provided textual and visual information are distinguished from those that can use any external information. In the remainder, pre-trained models (e.g. pre-trained word embeddings or convolutional models pre-trained using tier image collections) are understood as external information: runs using textual or visual information only were obtained with models trained from random initializations.

3.1 Textual Information

Classification of textual content is usually carried out by considering the text as a sequence of words, and using recurrent neural models such as the LSTM [6] for classifying these sequential inputs. We compared the baseline LSTM to several variants (e.g. Attention-based BiLSTM [17], Multi-Head Attention model [14]), but no option yielded results significantly better than the baseline LSTM with random initialization. Hence the baseline LSTM model was the only one considered for textual processing in the MFLE subtask. This setting is suitable to MFLE run 2, as it does not require any pretraining. For the latter run, we performed a grid search for hyper parameters. In the end, we retained 50 for the fixed text size, 100 for the hidden vector size, and 32 for the word embedding size. Taking inspiration from data augmentation techniques used with images (see Section 3.2), we tried to augment the training set by setting a random offset to the extracted textual sequences (instead of always taking the 50 first words in the text). We did not observe improved results by doing so. We hypothesize that the starting words in a text carry a lot of its overall meaning.

3.2 Visual Information

For image classification in both subtasks, we focused on 3 well known model architectures: InceptionV3 [13], MobileNetV2 [11] and VGG16 [12]. InceptionV3 has served as a building block for

Table 1: F1-Scores (%) for NITD and MFLE subtasks. We refer to Sections 3.1, 3.2 and 3.3 for details about specific runs. ∅ indicates the model has been trained from a random initialization.

NITD					MFLE				
Run 1	Run 2	Run 3	Run 4	Run 5	Run 1	Run 2	Run 3	Runs 4	Run 5
(MNV2)	*(InceptionV3)*		*(VGG16)*		*(MNV2)*	*(LSTM)*	*(MNV2 & LSTM)*	*(IV3 & LSTM)*	*(VGG16 & LSTM)*
∅	*ImageNet*	*fine-tuned*	*Places365*	*fine-tuned*	∅	∅	∅	*ImageNet*	*Places365*
85.1	81.0	79.6	89.0	**89.6**	56.6	56.5	57.6	**67.1**	66.0

many recent contributions (e.g. [4, 9, 10]). Parameter sets pre-trained on the data from the ImageNet challenge [5] are also widely available. VGG16 yields performance close to the state of the art on the Places365 scene recognition task [16]. As NITD and MFLE can be understood as recognizing certain types of scenes, we hypothesize that transferring a VGG16 model pre-trained on Places365 can be valuable. MobileNetV2 has a comparatively small number of parameters (~ 2M, when InceptionV3 defines ~ 20M parameters, and VGG16 ~ 130M), and is hence more suitable for being trained from scratch. For all models, we rescaled images to 224×224 pixels. We applied image augmentation methods commonly used in above-mentioned papers, i.e. each image in the training batches is modified by a combination of random transformations. When no external data can be used (i.e. run 1 of NITD and MFLE), we used randomly initialized MobileNetV2 models. InceptionV3 models pre-trained using the ImageNet dataset were used. In a first stage, only the last dense layer was trained while freezing all other layers (NITD run 2). The last 2 convolutional layers were then fine-tuned (NITD run 3). Also, we jointly trained the two last fully connected layers of a pre-trained VGG16 model (NITD run 4). We tried to fine tune the large fully connected ante penultimate layer and the last convolutional layer after this first stage (NITD run 5).

3.3 Mixed Input

Figure 1 shows the generic architecture for our mixed input models. The proposed multimodal approach is very similar to that proposed in [9]. The embedding of textual and visual models is taken as their penultimate layer output. For MFLE run 3, we reused previously trained LSTM (run 2) and MobileNetV2 (run 1), and trained only the additional fully connected layers. Similarly, for run 4 we used a pre-trained InceptionV3 model, adapted and fine-tuned to the MFLE subtask data as described in 3.2. Run 5 used a pre-trained and adapted VGG16 model instead. The textual and visual models performing the best w.r.t all training data were selected. Ensembles of 5 mixed input models were trained from this common basis. To balance the influence of text and image, we defined a bottleneck fully connected layer, that reduces the generally high-dimensional convolutional embedding (e.g. 2048 for InceptionV3) to size 100, the same size as the LSTM hidden vector.

4 ANALYSIS

The F1-Scores resulting from our runs are displayed in Table 1. For the NITD subtask, VGG16 with fine-tuning gets the best results (F1 = 89.6%). VGG16 models performed better than others with a significant margin. This means NITD is close to a scene recognition task. However, fine-tuning brought at best minor improvement

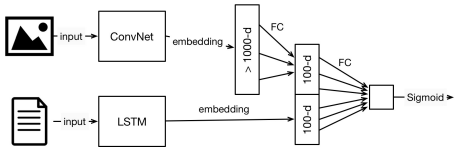

Figure 1: Generic mixed input model architecture.

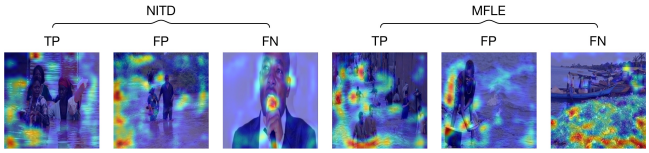

Figure 2: Class Activation Maps for true positives (TP), false positives (FP) and false negatives (FN) in both subtasks.

(+0.6% for VGG16, runs 4 and 5), at worse performance degradation (-1.4% for InceptionsV3, runs 2 and 3). Perhaps surprisingly, MobileNetV2 trained from scratch offers solid performance, in between InceptionV3 and VGG16. For the MFLE subtask, the mixed input model benefits from a small positive combination effect between modalities (+1.0% between runs 1 and 3). Using pre-trained vision models (runs 4 and 5) yields a significant performance boost ($\approx$ 10%). The best results are obtained with the combination of LSTM and InceptionV3 (run 4). As all images in the MFLE data set depict flooded scenes, we can hypothesize that object level features are more relevant than scene level features. Figure 2 displays Class Activation Maps [15] of examples with high positive or negative activations w.r.t. InceptionV3 models trained on both subtasks. We see that the positive class is frequently associated to the detection of water patterns. This possibly explains for the limited performance in MFLE, as these patterns are not very discriminative then. On the other hand, false negatives are often associated to misleading elements in the image (e.g. microphone for NITD, pile of rubbish for MFLE).

5 CONCLUSION

In this paper, we tested several approaches to the detection of flood severity in multimodal news content. We highlighted the relevance of considering closely related tasks for pre-training, rather than general-purpose image datasets such as ImageNet. Mixed-input architectures in the MFLE task yielded an improvement w.r.t. modalities taken separately, but this improvement was limited in comparison to the influence of using relevant pre-trained models.

REFERENCES

[1] N. Audebert, C. Herold, K. Slimani, and C. Vidal. 2019. Multimodal deep networks for text and image-based document classification. *arXiv:1907.06370 [cs]* (2019).

[2] B. Bischke, P. Helber, S. Brugman, E. Basar, Z. Zhao, M. Larson, and K. Pogorelov. 2019. The Multimedia Satellite Task at MediaEval 2019. In *Proc. of the MediaEval 2019 Workshop*.

[3] A. Bulat and G. Tzimiropoulos. 2016. Human pose estimation via Convolutional Part Heatmap Regression. In *European Conference on Computer Vision*. 717–732.

[4] P. Chen, Y. Sharma, H. Zhang, J. Yi, and C. Hsieh. 2018. EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[5] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and F. Li. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[6] A. Graves. 2012. Supervised Sequence Labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Berlin Heidelberg, 5–13.

[7] D. Kingma and J. Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations*.

[8] R. Kiros, R. Salakhutdinov, and R. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *NIPS Deep Learning Workshop*.

[9] L. Lopez-Fuentes, J. van de Weijer, M. Bolaños, and H. Skinnemoen. 2017. Multi-modal Deep Learning Approach for Flood Detection. In *Proc. of the MediaEval 2017 Workshop*.

[10] R. Poplin, A. Varadarajan, K. Blumer, Y. Liu, M. McConnell, G. Corrado, L. Peng, and D. Webster. 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* 2, 3 (2018), 158–164.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[12] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference for Learning Representations*.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*. 5998–6008.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.

[16] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464.

[17] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 207–212.