Exploiting ConvNet Diversity for Flooding Identification

Keiller Nogueira[1], Samuel G. Fadel[2], Ícaro C. Dourado[2], Rafael de O. Werneck[2], Javier A. V. Muñoz[2], Otávio A. B. Penatti[3], Rodrigo T. Calumby[4], Lin Tzy Li[2,3], Jefersson A. dos Santos[1], Ricardo da S. Torres[2]

[1]Departamento de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG)
{keiller.nogueira,jefersson}@dcc.ufmg.br
[2]Institute of Computing, University of Campinas (Unicamp)
{samuel.fadel,icaro.dourado,rafael.werneck,lintzyli,rtorres}@ic.unicamp.br
[3]Advanced Technologies Group, SAMSUNG Research Institute
o.penatti@samsung.com
[4]Department of Exact Sciences, University of Feira de Santana (UEFS)
rtcalumby@ecomp.uefs.br

November 13, 2017

Abstract

Flooding is the world's most costly type of natural disaster in terms of both economic losses and human causalities. A first and essential procedure towards flood monitoring is based on identifying the area most vulnerable to flooding, which gives authorities relevant regions to focus. In this work, we propose several methods to perform flooding identification in high-resolution remote sensing images using deep learning. Specifically, some proposed techniques are based upon unique networks, such as dilated and deconvolutional ones, while other was conceived to exploit diversity of distinct networks in order to extract the maximum performance of each classifier. Evaluation of the proposed algorithms were conducted in a high-resolution remote sensing dataset. Results show that the proposed algorithms outperformed several state-of-the-art baselines, providing improvements ranging from 1 to 4% in terms of the Jaccard Index.

Keywords: Flooding identification, Natural disaster, Remote Sensing, Inundation, MediaEval, Satellites.

1 Introduction

Natural disaster monitoring is a fundamental task to create prevention strategies, as well as to help authorities to act in the control of damages, coordinate rescues, and help victims. Among all kinds of natural hazards, flooding is possibly the most extensive and devastating one, destroying buildings, roads, bridges; tearing out trees; devastating agriculture; causing mudslides; and threatening human lives [1]. All these consequences make such events to be considered as the world's most costly type of natural disaster in terms of both economic losses and human causalities as pointed out by the disaster statistical review released every year by the Centre for Research on the Epidemiology of Disasters [2]. Some of these disasters happen annually [3], for example in humid tropics and subtropical climates, where river flooding is a recurrent natural phenomenon due to excessive rain within a short period of time. However, other events may occur atypically [4], which is the case of flooding caused by hurricanes, such as the recent triple hurricanes (e.g., Harvey, Maria, and Irma) that brought massive flooding in several countries [5].

Although extremely important, floods are difficult to monitor, because they are highly dependent on several local conditions, such as precipitation, slope of terrain, drainage network, protective structures, and land cover [6]. A first and essential step towards such monitoring is based on identifying

areas most vulnerable to flooding, helping authorities to focus on such regions while monitoring inundations. Remotely sensed data play a crucial role in identifying such areas, since it allows the capture of whole inundated regions during a flooding event, allowing a better understanding of what and how it is flooded. Although there are lots of works [7,8] performing flooding detection using remote sensing data integrated with elevation maps in order to augment the amount and type of information available for an efficient flood management, as far as we know, there are no works that focus on identifying flooding areas using only remote sensing data. Because of the importance of such task and the lack of works dealing with it, a subtask (called Flood-Detection in Satellite Images) of the 2017 Multimedia Satellite Task [9], which was part of the traditional MediaEval Benchmark, was proposed to leverage the development of methods for identifying flooding events in high-resolution remote sensing images.

In this paper, we present our proposed methods, which won the aforementioned task, to automatic identify flooding areas in high-resolution remote sensing images using deep learning paradigm. Deep learning [10], commonly represented as multi-layered neural networks, can learn simultaneously the features and the classifiers. In other words, during the training process, a deep neural network is able to learn both the features and the classifier in a unified manner, adjusting itself to better represent the characteristics of the data and their labels. Among all deep learning-based networks, a specific type, called Convolutional (Neural) Networks, ConvNets or CNNs [10], is the most popular for learning visual features in computer vision applications, including remote sensing [11,12]. This type of network relies on the natural stationary property of an image, i.e., the statistics of one part of the image are assumed to be the same as those of any other part. Furthermore, deep ConvNets can be considered as an inherently multiscale approach since they usually obtain different levels of abstraction for the data, ranging from local low-level information in the initial layers (e.g., corners and edges), to more semantic descriptors, mid-level information (e.g., object parts) in intermediate layers, and high-level information (e.g., whole objects) in the final layers.

We introduce several approaches to identify flooding areas of remote sensing images exploiting the advantages of ConvNets. Some methods are based uniquely on networks with distinct properties, including: (i) dilated convolutions [13], which, unlike standard ConvNets, process the input without downsampling it, and (ii) deconvolution layers, such as SegNet [14], in which a coarse feature map is upsampled outputting a dense map with the same resolution of the original image. Another approach exploits the diversity of distinct networks in order to extract the maximum performance of each classifier. In summary, the contributions of the paper are: (i) novel ConvNet architectures specialized in identifying flooding areas; and (ii) a new strategy to exploit network diversity for inundation identification. Obtained results of the proposed methods represent the state of the art, in terms of the Jaccard Index, in a remote-sensing-based flooding detection task. These results made us the winner of the Flood-Detection in Satellite Images, a subtask of 2017 Multimedia Satellite Task [9].

2 Proposed Methods

In this section, we present the proposed techniques to perform flooding identification. The first approaches (Section 2.1) are based upon new and specific architectures to handle such important task. The other method (Section 2.2) was conceived to better combine the diversity of distinct, but complementary networks.

2.1 Network Architectures

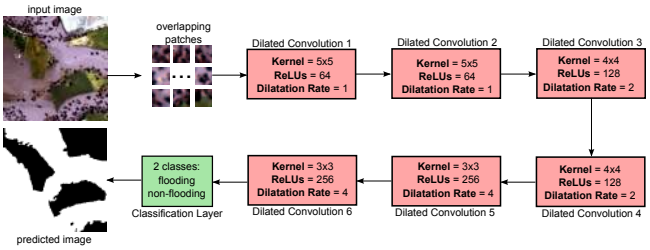
All networks conceived specifically for flooding identification are based on ConvNets and have architectures illustrated in Figure 1. Some use dilated convolutions while others are based on deconvolutional networks.

Specifically, two architectures are based on the concept of dilated convolutions [13]. In these layers, the convolution filter is expanded by dilation rate. Given this rate, the weights are placed far away at given intervals and the kernel size increases by allowing gaps (or "holes") inside their filters. Therefore, networks composed of these layers allow the receptive field to expand but preserving the resolution, i.e., without downsampling the input data. This procedure represents a great advantage in terms

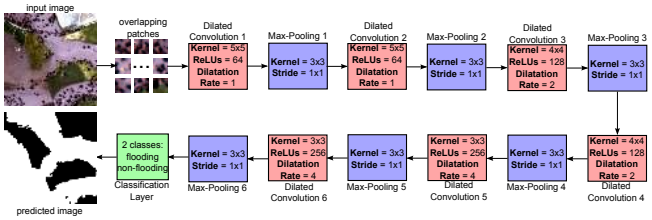
of computational processing, as well as in terms of learning, since internal feature maps do not lose resolution (and maybe useful information).

The first architecture, presented in Figure 1a, is composed of seven layers: six dilated convolutions (that are responsible for learning the patterns of the input images) and a final 1×1 convolution layer, which is responsible for identifying the flooding areas. There are no pooling or normalization operations inside this network. Specifically, the first two convolutions have 5×5 filters with dilation rate 1. The following two convolutions have 4×4 filters and rate 2 while the last two convolutions have smaller filters (3×3) with 4 dilation rate.

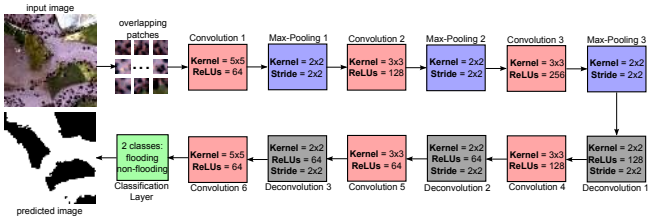
The second proposed network, presented in Figure 1b, is also based on dilated convolution layers. This network shares the same architecture of the first one, except for the fact that there are additional max-pooling layers between each dilated convolutions layer. Each pooling has 3×3 kernel and stride of 1. Although pooling usually reduces the resolution of the input data, in this case, it also preserves the resolution because of zero padding used in the input. The core idea of this network is to conserve the benefits of dilated convolutional layers but adding known advantages of pooling layers, such as invariance to small rotations and translations, as well as the preservation of the most important features.



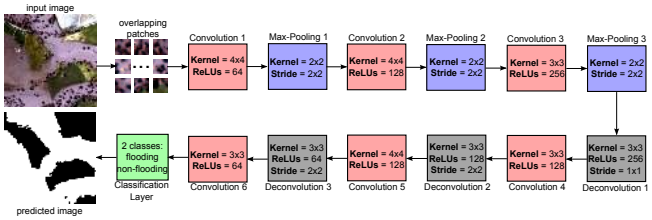
(a) Dilated ConvNet #1



(b) Dilated ConvNet #2



(c) Deconvolution ConvNet #1



(d) Deconvolution ConvNet #2

Figure 1: ConvNet architectures proposed in this work.

The two remaining networks are based on deconvolutional networks [14]. This type of network has two modules: the first receives input images, learns the visual features by using standard convolution and pooling layers, and outputs a coarse feature map while the second receives this map as input,

learns to upsample these features by using several deconvolution layers, and outputs a dense prediction map with the same resolution of the original image. Both modules work together without distinction and can be trained end-to-end by using standard feedforward and backpropagation algorithms.

The two architectures, presented in Figures 1c and 1d, are based on the solution discussed in [14]. The encoder of both is based on three standard convolution layers intercalated with max-pooling operations. The only difference is the kernel of the convolutions, i.e., while the first layer uses a larger kernel in the first layer (in order to learn more information about flooding regions directly from the input image), the other one bets on similar filter sizes, which should extract the same amount of information in all layers resulting in a balanced learning of the flooding patterns. The deconvolutional part has similar architecture with the same number of layers (three deconvolutional ones intercalated with standard convolutions) differing only in the size of deconvolutional and convolutional kernels. The premise was conserved, i.e., while the first deconvolutional network bets on larger kernel, the second tries to use similar filter sizes.

2.2 Combination

As previously explained, we also proposed another strategy to solve the flooding detection task, which aims to exploit the diversity of distinct ConvNets. The main premise of the proposed method is that the previous presented ConvNets learn and produce distinct outcomes, which are dense prediction maps. This difference should make ConvNets complementary to each other. Therefore, a clever combination of such outcomes should improve the final prediction map if compared with the ConvNets individual results.We propose a combination method based upon Support Vector Machines (SVM).

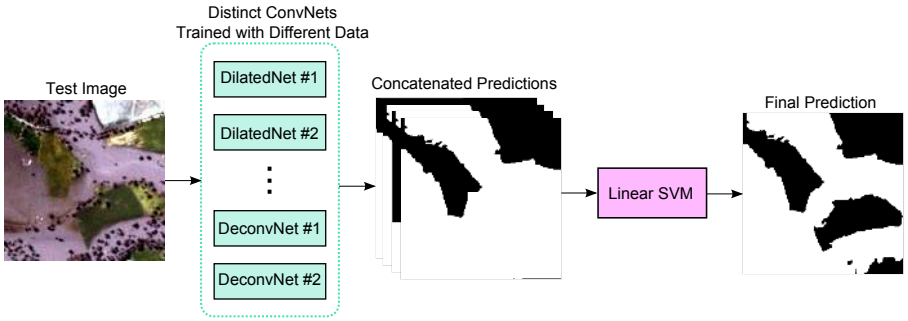


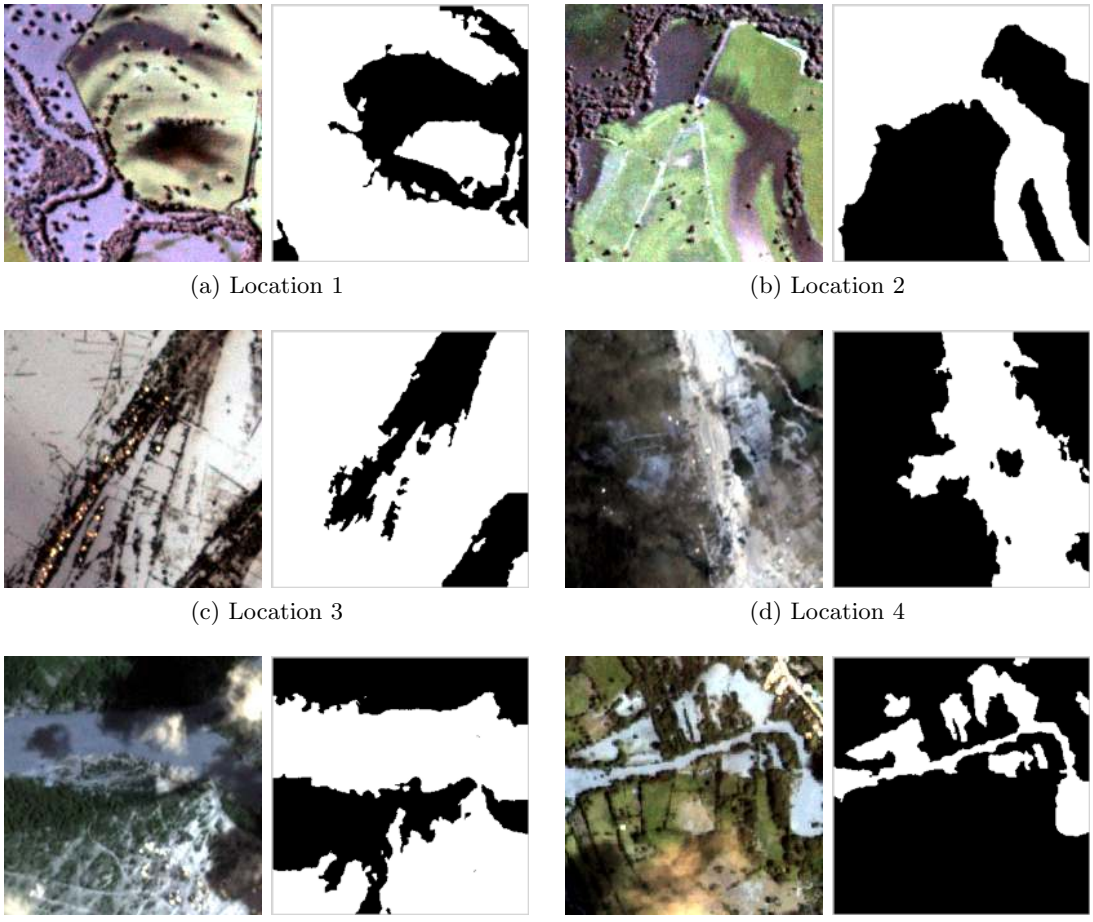
Figure 2: Pipeline for the prediction phase of the proposed combination approach.

The proposed method is divided into three main steps: (i) *extraction:* In this phase, an image is processed by all proposed network, which produce distinct outcomes (i.e., different probability or prediction maps). All these maps (that have the same resolution of the original input image) are then concatenated creating a feature vector that, in fact, represents the input image. (ii) *learning:* In this step, the SVM receives the aforementioned feature vector as well as the ground-truth flooding map for all training data. Then, it independently process each pixel of these images, learning which and when each classifier is better; and (iii) *prediction:* This final step receives feature vectors of testing images and, using the trained SVM, outputs the improved prediction map for each test image. This final step is illustrated in Figure 2.

3 Experiments

In this section, we present the experimental setup. Also, we present and discuss the obtained results of the proposed methods comparing them with the best performing teams of the Flood-Detection in Satellite Images subtask of the 2017 Multimedia Satellite Task.

3.1 Dataset



(a) Location 1 (b) Location 2

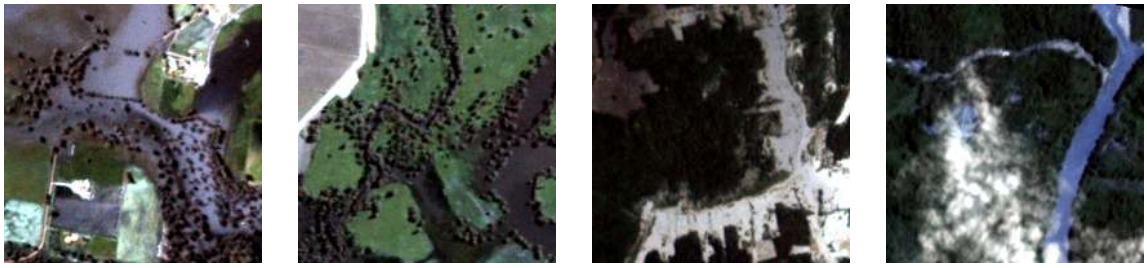
(c) Location 3 (d) Location 4

(e) Location 5 (f) Location 6

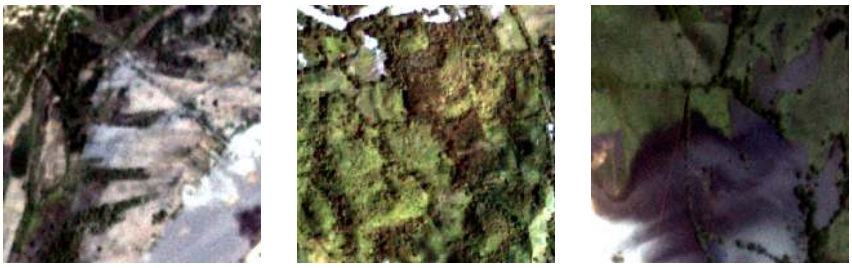
Figure 3: Examples of patches for all locations of the training set with its respective ground-truth, in which white regions refer to flooded area and black ones correspond to background.

The dataset consists of satellite image patches collected from eight different flooding events June 1st, 2016 to May 1st, 2017. Each image patch is composed of four bands (red, green, blue, and near infrared bands) and has resolution of 320×320 pixels, with a ground-sample distance (GSD) of 3.7 meters and an orthorectified pixel size of 3 meters [9].

The training set is composed of 462 image patches unevenly extracted from **six** locations. Among these images, 92 (20%) were employed as internal validation set to evaluate the proposed algorithms while the remaining 370 images were used to train the proposed methods. Some examples of image patches for each of these locations are presented in Figure 3. Two test sets were released in this dataset: the **Same Locations** test set contains 216 unseen patches unevenly extracted from the same region presented in the training set, while the **New Locations** test set contains 58 unseen patches extracted from a region not present in the training set. Figure 4 presents some examples of both test sets. It is important to highlight that, until the submission of this current paper, ground-truth of the test sets were not released by the organization of the competition.



(a) Location 1 (b) Location 2 (c) Location 3 (d) Location 4



(e) Location 5 (f) Location 6 (g) New Location

Figure 4: Examples of patches for both test sets.

3.2 Experimental Evaluation

In order to assess the performance of generated segmentation masks for flooded areas in the satellite image patches, the intersection-over-union metric (also known as Jaccard Index), was used. The metric measures the accuracy for the pixelwise classification, and as defined as $IoU = \frac{TP}{(TP+FP+FN)}$, where TP, FP, and FN are the numbers of true positive, false positive, and false negative pixels, respectively, determined over the whole test set.

3.3 Experimental Protocol

First, we trained all ConvNets (presented in Section 2.1) using overlapping patches of size 25×25 extracted from all training images (independent of the location). In the prediction phase, we also extracted overlapping patches with the same resolution from the testing images and averaged the probabilities outputted by the network. Among all networks, the best one (in our internal validation set) is reported as **ConvNet** 25×25.

Another proposed method relied on training the aforementioned ConvNets using larger overlapping patches, with 50×50 pixels, also extracted from all training images. The motivation behind this strategy is based on the entire context that could be extract from the input patches and improve the learning process. The prediction phase is similar to the previous strategy. Considering this configuration, the best network (in our internal validation set) is referred to in the next section as **ConvNet** 50×50.

The **Location ConvNets** strategy is based on the idea of creating specialized ConvNets for each flooding event. Since the dataset has six distinct flooding event locations, we propose to train a specific Dilated ConvNet #1 (using patches of 25×25) for each location. The prediction is similar to the other proposed protocols, except for the fact that, in this case, each ConvNet was used in its respective location. For the **New Locations** test set, we combined the outcomes extracted from each ConvNet (trained specifically for each location) using a linear SVM, as proposed in Section 2.2.

The **Fusion-SVM** strategy expands above idea. Differently from previous scheme, in this procedure, an SVM is used to create prediction maps for **both test sets** (and not only for the **New Locations** test set). Based on the premise that distinct ConvNets (trained using different input data) produce distinct (and possible complementary) outcomes, we propose to combine the predictions extracted from all above ConvNets using a linear SVM, as presented in Section 2.2. In this way, the SVM should be able to learn when and how these networks complement each other in order to improve

the final performance. Specifically, the SVM receives as input concatenated probabilities extracted from all previously trained ConvNets, which include: (i) all four ConvNets presented in Section 2.1 trained with overlapping 25×25 patches, (ii) all four networks introduced in Section 2.1 trained with overlapping 50×50 patches, and (iii) six Dilated ConvNet #1 trained specifically for each location of the training set.

Another strategy relied on exploiting the diversity of distinct ConvNets by combining all outcomes of previous methods using a majority voting scheme, which is referred to in following section as **Fusion-MV**.

It is important to emphasize that all proposed methods were created using TensorFlow framework and will have code released upon the acceptance of this paper. When training, all the aforementioned protocol used the same hyper-parameters, i.e., learning rate, weight decay, momentum and number of epochs as 0.01, 0.0005, 0.9 and 20, respectively.

3.4 Baselines

The baselines evaluated in this work were, in fact, the best performing approaches proposed for the Flood-Detection in Satellite Images subtask of the 2017 Multimedia Satellite Task. An overview of the such methods (which includes state-of-the-art methods, such as Generative Adversarial, Deconvolutional and Fully Convolutional Networks) are presented in Table 1.

3.5 Results and Discussion

All results for the test sets are presented in Table 1. These are the official results released by the Mediaeval since no ground-truth for the test set was released yet. Also, it is worth mentioning that, for the two strategies of learning ConvNets using all available training data (but with distinct patch size), the Dilated ConvNet #1 yielded the best results among all experimented networks. Therefore, these were the ones submitted to the competition as well as reported in this work as ConvNet 25×25 and ConvNet 50×50.

For both test sets, the best solution was obtained by combining the probabilities of all trained ConvNets using a Linear SVM. This technique yielded state-of-the-art results in both test set, outperforming all baselines by, at least, 4% in the **Same Locations** test set and 1% in the **New Locations** test set (in terms of Jaccard Index). Some samples of this obtained results are presented in Figure 5.

For the **Same Locations** test set, training a network for each location (Location ConvNets) or training a ConvNet with all available data (ConvNet 25×25) achieved the same result. This may indicate that the proposed architecture can, in fact, extract and interpret all feasible information from the whole data, which is a great advantage given that it reduces the number of networks to train and, consequently, the processing time. This conclusion does not hold for the **New Locations** test set. In this set, training a specific network for each location (Location ConvNets) achieved higher performance (aside the Fusion-SVM strategy) when compared to unique networks trained with the whole training set (such as ConvNet 25×25 and ConvNet 50×50). This indicates that specific Location Network can learn details that may not be useful for classification of a known image, but that is important for unseen data, which is the case.

Another relevant outcome is that increasing the size of the input patch (ConvNet 50×50) decreases the final result, a conclusion that holds for both datasets. We believe that this is because of the amount of training patches generated in each case. More specifically, a large amount of data may be used for training with smaller patch sizes while large patches means less data to train. This corroborates with the fact that deep learning really needs a large amount of labeled data to train [10].

Finally, for both sets, the worst result was obtained using the majority voting scheme. This may be justified by the fact that Majority Voting is not so robust to aggregate information from multiple networks, when they disagree in the classification. This fact can be overcome by using a machine learning technique to capture about the opinions of the ConvNets.

Table 1: IoU (%) results of the proposed method and baselines for both test sets. Higher values of IoU indicates better performance.

Methods		Overview	Test Set	
			Same Locations	**New Locations**
Baselines	WISC [15]	NDVI plus SVM-RBF	80	83
		K-Means to cluster and classify	81	77
	CERTH-ITI [16]	Mahalanobis dist. with stratified cov.	75	56
	BMC [17]	ResNet-152 and random forest	37	40
	UTAOS [18]	Gen. Adv. Net. with 0.78 threshold	82	73
		Gen. Adv. Net. with 0.94 threshold	80	70
		Gen. Adv. Net. with 0.50 threshold	83	74
		Gen. Adv. Net. with 0.35 threshold	83	74
		Gen. Adv. Net. with 0.12 threshold	81	73
	DFKI [19]	VGG13-FCN with RGB data	73	69
		VGG13-FCN with RGB and NIR data	84	70
		VGG13 adapted to be a DeconvNet	84	74
Proposed	Dilated 25×25	Dilated ConvNet #1 (25×25 patches)	87	82
	Dilated 50×50	Dilated ConvNet #1 (50×50 patches)	86	80
	Location ConvNets	Dilated ConvNet #1 trained per location	87	**84**
	Fusion-SVM	SVM over concatenated predictions	**88**	**84**
	Fusion-MV	MV over concatenated predictions	78	49

4 Conclusion

In this paper, we propose new approaches based on Convolutional Neural Networks to perform detection of flooding areas in remote sensing images. Specifically, we proposed four distinct architectures based on dilated convolutions [13] and deconvolution layers [14]. Furthermore, different strategies to combine such networks were proposed and evaluated.

Experimental results have showed that the methods are effective and robust. We have achieved state-of-the-art performance, in terms of Jaccard Index, in a specific dataset proposed for the Flood-Detection in Satellite Images subtask of the 2017 Multimedia Satellite Task. The proposed methods outperformed all baselines, winning that subtask challenge. Such results show that our proposed approaches are effective and robust to identify flooding areas (independent if it is for a recurrent or atypical event). This identification process performed by our proposed algorithms may help authorities to keep focus on most vulnerable regions while monitoring forecast inundations, which may aid in coordinate rescues, and help victims.

As future work, we intend to use different post-processing methods, such as Conditional Random Fields, in order to exploit the contextual information.

Acknowledgments

The authors thank FAPESP (grants #2013/50169-1, #2013/50155-0, #2014/50715-9, #2014/12236-1, and #2016/18429-1), FAPEMIG, CNPq, and CAPES.

References

[1] NOAA/NWS (National Weather Service), Floods: The awesome power (2005).

[2] CRED, Centre for research on the epidemiology of disasters (cred) (2017).
URL `http://www.emdat.be/`

[3] T. De Groeve, Flood monitoring and mapping using passive microwave remote sensing in namibia, Geomatics, Natural Hazards and Risk 1 (1) (2010) 19–35.

[4] B. P. Harman, S. Heyenga, B. M. Taylor, C. S. Fletcher, Global lessons for adapting coastal communities to protect against storm surge inundation, Journal of Coastal Research 31 (4) (2013) 790–801.

[5] C. Eric Levenson, 3 storms, 3 responses: Comparing harvey, irma and maria (2017).
URL http://edition.cnn.com/2017/09/26/us/response-harvey-irma-maria/index.html

[6] V. Klemas, Remote sensing of floods and flood-prone areas: an overview, Journal of Coastal Research 31 (4) (2014) 1005–1013.

[7] L. Pulvirenti, N. Pierdicca, G. Boni, M. Fiorini, R. Rudari, Flood damage assessment through multitemporal cosmo-skymed data and hydrodynamic models: The albania 2010 case study, JSTARS 7 (7) (2014) 2848–2855.

[8] A. D'Addabbo, A. Refice, G. Pasquariello, F. P. Lovergine, D. Capolongo, S. Manfreda, A bayesian network for flood detection combining sar imagery and ancillary data, TGRS 54 (6) (2016) 3612–3625.

[9] B. Bischke, P. Helber, C. Schulze, S. Venkat, A. Dengel, D. Borth, The multimedia satellite task at mediaeval 2017: Emergence response for flooding events, in: Proc. of the MediaEval 2017 Workshop, Dublin, Ireland.

[10] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[11] K. Nogueira, W. O. Miranda, J. A. Dos Santos, Improving spatial feature representation from aerial scenes by using convolutional networks, in: Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2015, pp. 289–296.

[12] K. Nogueira, O. A. Penatti, J. A. dos Santos, Towards better exploiting convolutional neural networks for remote sensing scene classification, Pattern Recognition 61 (2017) 539–556.

[13] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: ICLR, 2016.

[14] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, arXiv preprint arXiv:1511.00561.

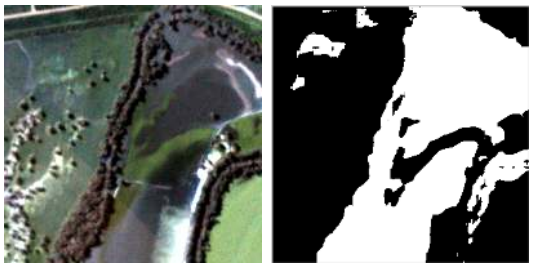
[15] N. Tkachenko, A. Zubiaga, R. Procter, Wisc at mediaeval 2017: Multimedia satellite task, in: Working Notes Proc. MediaEval Workshop, 2017, p. 2.
URL http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_12.pdf

[16] K. Avgerinakis, A. Moumtzidou, S. Andreadis, E. Michail, I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, Visual and textual analysis of social media and satellite images for flood detection @ multimedia satellite task mediaeval 2017, in: Working Notes Proc. MediaEval Workshop, 2017, p. 2.
URL http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_31.pdf

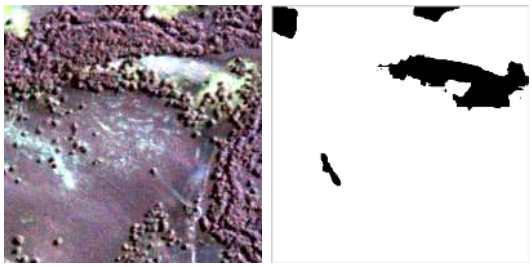
[17] X. Fu, Y. Bin, L. Peng, J. Z. Y. Yang, H. T. Shen, Bmc@mediaeval 2017 multimedia satellite task via regression random forest, in: Working Notes Proc. MediaEval Workshop, 2017, p. 2.
URL http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_46.pdf

[18] K. Ahmad, P. Konstantin, M. Riegler, N. Conci, P. Holversen, Cnn and gan based satellite and social media data fusion for disaster detection, in: Working Notes Proc. MediaEval Workshop, 2017, p. 2.
URL http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_15.pdf

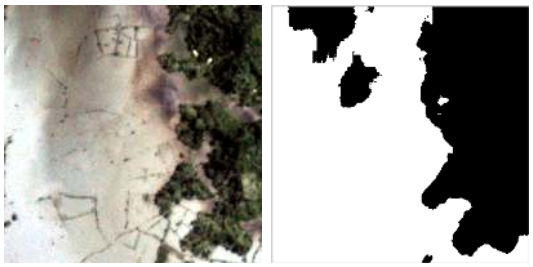
[19] B. Bischke, P. Bhardwaj, A. Gautam, P. Helber, D. Borth, A. Dengel, Detection of flooding events in social multimedia and satellite imagery using deep neural networks, in: Working Notes Proc. MediaEval Workshop, 2017, p. 2.
URL http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_51.pdf



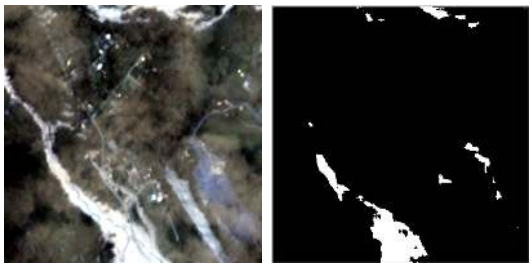
(a) Location 1



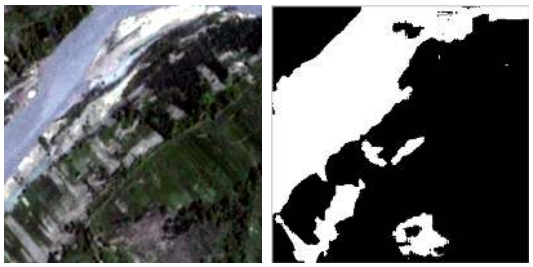
(b) Location 2



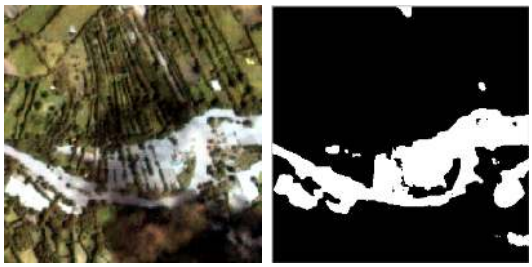
(c) Location 3



(d) Location 4



(e) Location 5



(f) Location 6

(g) New Location

Figure 5: Examples of some test images and the obtained results achieved by using SVM with the aggregated probabilities. White areas refer to flooded regions and black areas correspond to background.