POLITECNICO DI TORINO
Repository ISTITUZIONALE

Deep learning models for road passability detection during flood events using social media data

Original
Deep learning models for road passability detection during flood events using social media data / Lopez Fuentes, Laura; Farasin, Alessandro; Zaffaroni, Mirko; Skinnemoen, Harald; Garza, Paolo. - In: APPLIED SCIENCES. - ISSN 2076-3417. - ELETTRONICO. - 10:24(2020). [10.3390/app10248783]

Availability:
This version is available at: 11583/2855075 since: 2020-12-08T17:40:19Z

Publisher:
MDPI

Published
DOI:10.3390/app10248783

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
default_article_editorial [DA NON USARE]

-

(Article begins on next page)

16 April 2024

applied sciences

MDPI

Article

Deep Learning Models for Road Passability Detection during Flood Events Using Social Media Data

Laura Lopez-Fuentes [1,2,*], Alessandro Farasin [3,4], Mirko Zaffaroni [4,5], Harald Skinnemoen [2] and Paolo Garza [3]

1 Department of Mathematics and Computer Science, University of the Balearic Islands, Ctra. de Valldemossa, Km.7.5, 07122 Palma, Spain
2 AnsuR Technologies AS, Martin Linges vei 25, 1364 Fornebu, Norway; harald@ansur.no
3 Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy; alessandro.farasin@polito.it (A.F.); paolo.garza@polito.it (P.G.)
4 Department of Data Science for Industrial and Societal Applications, LINKS Foundation, Via Pier Carlo Boggio, 61, 10138 Turin, Italy; mirko.zaffaroni@unito.it
5 Department of Computer Science, Universita' degli Studi di Torino, Via Pessinetto, 12, 10149 Turin, Italy
* Correspondence: l.lopez@uib.es

Received: 4 November 2020; Accepted: 2 December 2020; Published: 8 December 2020

Abstract: During natural disasters, situational awareness is needed to understand the situation and respond accordingly. A key need is assessing open roads for transporting emergency support to victims. This can be done via analysis of photos from affected areas with known location. This paper studies the problem of detecting blocked/open roads from photos during floods by applying a two-step approach based on classifiers: does the image have evidence of road? If it does, is the road passable or not? We propose a single double-ended neural network (NN) architecture which addresses both tasks simultaneously. Both problems are treated as a single class classification problem with the use of a compactness loss. The study was performed on a set of tweets, posted during flooding events, that contain (i) metadata and (ii) visual information. We studied the usefulness of each data source and the combination of both. Finally, we conducted a study of the performance gain from ensembling different networks. Through the experimental results, we prove that the proposed double-ended NN makes the model almost two times faster and the load on memory lighter while improving the results with respect to training two separate networks to solve each problem independently.

Keywords: flood detection; road passability; image classification; emergencies; social media; deep learning

1. Introduction

In the last decades, the frequency and intensity of natural disasters has risen significantly. According to worldwide data from the Centre for Research on the Epidemiology of Disasters [1], about 12 times more natural disasters events were registered in 2017 compared to in 1950. In Figure 1, the dramatic increase in these events is shown, which comprise mass movements, volcanic activities, wild-fires, landslides, earthquakes, extreme temperatures, droughts, extreme weather, floods and epidemics. It can be noted that flood events were particularly frequent; in 2017 alone, floods represented approximately 39% of global natural disasters. On top of the tragic loss of human lives and infrastructures, natural disasters come at a large cost to governments.

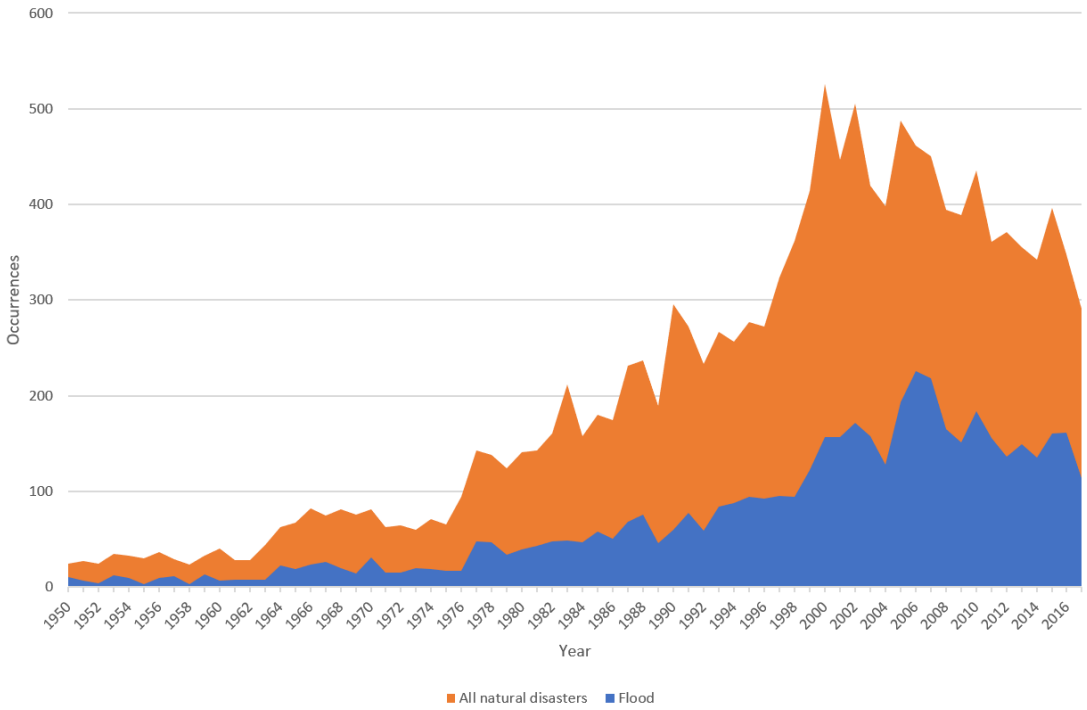


Figure 1. The occurrence of natural disasters and flood events since 1950. The natural disasters set is characterized by the following phenomena: mass movements, volcanic activities, wildfires, landslides, earthquakes, extreme temperatures, droughts, extreme weather, floods and epidemics. The chart shows a sensible increase in this kind of events, which, by 2017 has increased by over a factor of 10 with respect to 1950. Another important aspect is the percentage of incidence of flood events, representing, on average, 30% of the overall natural disasters [2].

The European Commission (EC) estimated that, since 2005, natural disasters have cost the European Union (EU) close to 100 billion euros. However, this cost can be significantly reduced by investing in risk prevention: the EC stated that for every 1 euro spent on prevention, 4 euros or more could be saved in response. In this respect, with the "cohesion policy", the EU allocated 8 billion euros for climate change adaptation, risk prevention and management over the 2014–2020 period [3]. Those investments generated several projects and research opportunities, from which this work is taken. The ability to detect the insurgence of such problems in a timely manner is a powerful tool for the bodies in charge of protecting and guaranteeing citizens' safety.

In this work, we are going to focus on flood events and specifically, on assessing the status of roads after floods, since knowing the best route to access the affected areas is crucial to transport emergency support to victims. This can be done via analysis of photos from affected areas with known location. Such photos can be: (i) solicited via dedicated apps, such as UN-ASIGN [4,5] and I-REACT [6,7], or (ii) harvested from unsolicited sources such as social media, as people frequently share pictures during emergencies. The use of apps and social media to engage the civil population is of increasing interest and can be useful for first responders.

Certainly, social media is not largely known and adopted as an emergency reporting tool, but there is evidence [8] of a large number of posts that provide direct proof of occurring natural disasters, which, if properly processed, could help in the handling of the emergencies. The need for sensibility regarding natural disasters and the variety of data to deal with make the research community itself an active player in those topics and it generates numerous important and effective conferences.

The objective is, given a collection of posts (including images) related to floods, determining whether: (i) there is *Evidence of Roads* and, in a positive case, (ii) there is *Evidence of Road Passability*. In the first case, we are more interested in asserting the presence of a road in the picture: this means the road can be directly visible or there are enough elements justifying its existence, such as the presence

of traffic lights or vertical signs. On the other hand, the second goal aims to determine whether the identified road is in good condition to be transited. In the flood context, the evidence of road passability means that the road is completely clean or it can be partially or totally covered by water, but there must be evidence that vehicles or people can still pass through it.

This paper has been inspired by the *Flood classification challenge* from MediaEval 2018, where we presented an algorithm that predicted if there was evidence of road and if so, if the road was passable. We presented an algorithm which achieved the best results in the challenge [9]. The work presented in this paper has three goals:

- Provide a more detailed explanation of the work presented in [9], which was published in the form of an extended abstract due to the page limitation.
- Contextualize our results with all the results from the challenge participants.
- Introduce two major modifications to the algorithm, namely a new loss and a new architecture which combines the two problems into a single network, which introduces an almost 10% gain in performance for the passability task while maintaining the road evidence task performance. Moreover, we make the problem end-to-end and the solution almost 90 times faster and lighter, obtaining a model that can be feasibly integrated into a real-life solution.

The paper is organized as follows. Section 2 introduces state-of-the-art techniques, focusing on the ones presented in the same competition. Then, Section 3 focuses on the quantitative and qualitative analysis of the available data and how they are used to build the dataset. The approaches developed specifically to deal with textual and picture information are explained in Section 4 and evaluated in Section 5, where the results are compared with those of the other techniques presented in the MediaEval 2018 competition. Finally, conclusions and improvements planned for the future are described in Section 6.

2. Related Work

In the twenty-first century, our social interactaction habits mainly revolve around smartphones and IoT-devices. The Internet in general and social media represent a new way for us to learn and communicate. In a personal Facebook or Twitter profile, it is easy to find personal information about daily activities, but also news about real-time events. During natural disasters, social media represents a huge source of information from which, if properly processed, it is possible to extract valuable data for emergency management organizations. Indeed, the research literature presents several studies to detect, collect and process valuable information [10–12]. Concerning flood events, general approaches aim to detect flood events [13–15], to segment water regions [16], or to estimate water level [17]. Other approaches aim to examine details, such as the presence of people [18,19], the identification of the most affected areas [11], or the identification of flooded roads and their viability. This last topic is addressed in our work, and it is thought to be an extension of the approaches presented in the MediaEval 2018 conference. Therefore, this section continues introducing the methods submitted to the MediaEval challenge. The techniques are developed to deal with the two main kinds of data available from social media—metadata and images. As extensively described in Section 3, metadata are composed by textual information (e.g., text of the post, title) and punctual information (e.g., coordinates, post creation date, post author reference), while the images are PNG or JPG pictures. The metadata information was approached in many ways. A simple approach was proposed by Zhao et al. [20]. They manually created a set of rules which, leveraging on the textual part of the tweets, look for n-grams (subset of n contiguous words in the same sentence) representing strings of lexical items they would expect to occur in tweets related to road passability. Other works, such as the ones proposed by Hanif et al. [21] and Moumtzidou et al. [22] started with a pre-processing of the tweet texts—first removing hyperlinks, punctuations and symbols and performing the word tokenization, then removing the stop-words and performing word stemming. The processed information was enriched by adding other metadata features, such as user tags. Another work by Kirchknopf et al. [23] proposed to check the metadata

language feature and it increased the number of English tweets by translating the ones written in other languages. This simple step avoids the need to handle multiple languages simultaneously, which is still an open problem in the field of Natural Language Processing. To be properly processed by classifiers, words in tweets are then translated into numerical features. This step was made through the use of (i) pre-trained word embeddings, which convert words into numerical vectors, such as fasttext [20,24], Word2Vec [25] or GloVe [26], and/or (ii) statistical features, like Term Frequency—Inverse Document Frequency (TF-IDF) [27]. Numerical features are then used to train models such as Support Vector Machines (SVM) or Convolutional Neural Networks (CNNs) [28] for the final classification.

Regarding the visual information, two approaches were mainly adopted on pictures: (i) using visual descriptors and (ii) extracting features from pre-trained CNNs. In the first case, the aim was to describe the images through a set of discrete information that could lead the classifier to improve the performances on the tasks. Several descriptors were already available from the dataset: Color and Edge Descriptor (CEDD) [29], Color Layout (CL) [30], Fuzzy Color and Texture Historgram (FCTH) [31], Edge Histogram (EH) [32], Joint Composite Descriptor (JCD) [33] and Scalable Color Descriptor (SCD) [34]. In the latter case, hidden layers of CNNs are used as feature descriptors. State-of-the-art CNNs such as AlexNet [35], DenseNet201 [36], InceptionV3[37], InceptionResNetV2 [38], ResNet [39], VGG [40] or YOLOv3 [41] were taken after they had been pre-trained on popular and wide datasets such as ImageNet [42], Places365 [43] or VOC [44] , and then fine-tuned on the dataset of this work. Leveraging on pre-trained networks is a common practice in deep learning research: training a single model from scratch requires prohibitive computational performances, nearly inaccessible to most research centres or universities. Within the context of this paper, using CNNs that were pretrained on datasets containing a variety of places, environments, and buildings, enables them to represent and recognize objects and shapes in their internal layers. Fine-tuning such networks on a smaller dataset for specific tasks, such as the ones used in this work, allows them to reuse the pre-trained knowledge to achieve the goal more effectively. Most of the works proposed during the MediaEval competition used the aforementioned CNNs to extract visual features from the last layers of the networks. Moreover, besides extracting *global* features (pertinent to the whole image), Bischke et al. [45] and Zhao et al. [46] also combined information related to single entities (i.e., cars, boats, persons), named *local* features.

Then, extracted features were used for classification in several manners. One option [20,22,47] was to feed them as input for a neural network having few fully connected layers and using softmax for classification. Other approaches used other state-of-the-art machine learning algorithms, such as Support Vector Machine (SVM) [21,23,45,46,48], Multinomial Naive-Bayes, Random Forest and SRKDA [21]. The approaches which performed best exploited ensemble models, where their final output was determined by majority voting or averaging each model's prediction. Ensemble models were also used for feature extraction, combining features extracted from the same picture by several CNNs [20]. Finally, two strategies were used to merge metadata and visual information: *early* and *late fusion*. The *early fusion* combines the features before being computed by the classifier(s), while *late fusion* averages the prediction of the approaches separately developed for the two domains.

In our work, we introduce a novel lightweight network architecture, able to achieve comparable results of the winner-approach, namely an ensemble model of 45 CNNs per task. The proposed approach leverages on a custom loss function, and accomplishes both the tasks at once, reducing the number of needed parameters.

3. Dataset

The dataset used to train, validate and test the algorithms was distributed by MediaEval 2018 for the Multimedia Satellite Challenge [49,50]. It consists of 7387 tweet ids for the development set and 3683 tweet ids for the test set. By the time the images were downloaded for this competition, a significant number of tweets were no longer available, which resulted in a development set of 5818 tweets and a test set of 3017 tweets. However, since the work done in this paper corresponds to an extension to the work done for the Multimedia Satellite Challenge and we do not have the

ground truth for the test set data, we will divide the training set into training (4074 images), validation (872 images) and test (872 images). The images for the test set will only be used to report the final results to make the setup as close as the original challenge.

The tweets have been collected by retrieving all the tweets with images containing the tags *flooding*, *flood* and *floods* during the hurricanes *Harvey*, *Irma*, and *Maria*. Since the image information is crucial to this work and a lot of images were duplicate in the tweets, a process to remove duplicate images was carried out by the dataset distributors. This process is described in detail in [49].

The provided ground truth for the tweets was manually generated through a crowdsourcing task and consists of a binary class label for the evidence of road presence and only for those images classified as containing a road, a second binary class label for the actual passability of the road. Positive road passability is considered when the road is practicable by conventional means (no boats, off-the-road vehicles, monster trucks, Hummer, Landrover, farm equipment) and it is, therefore, related to the water level and the surrounding context.

The annotators made the decisions based only on the content of the analyzed images. The dataset is significantly imbalanced towards the non-evidence of road, having only ∼36% of the tweet images containing roads. In ∼45% of the tweets labeled as containing roads, there is evidence of positive road passability. In Table 1, the absolute number of images pertaining to each class is displayed.

Table 1. Information about the number of images of each set and the number of images with evidence of road and among those, the number of images with passable roads.

Dataset	# Tot. Imgs	# Evid. of Roads		# Passable Roads	
		YES	NO	YES	NO
development set	5818	2130	3688	951	1179
test set	3017	-	-	-	-

3.1. Metadata

Each tweet has a set of metadata associated with it, including the user who tweeted it and the text shared by the user. In Table 2, we briefly describe the most relevant fields (metadata) contained in each tweet. Since many of them are empty or semi-empty, we only report the fields (16 out of 29) without missing values in the MediaEval 2018 tweets.

Table 2. Brief description of the metadata fields that have non-empty values in at least 90% of the given tweets.

Field	Description	Type
Created at	UTC time when this tweet was created	object
Entities	Dictionary of the entities which have been parsed out of the text, such as the hashtags	object
Extended entities	Dictionary of entities extracted from the media, such as the image size	object
Favorite count	Indicates how many times the tweet has been liked	int64
Favorited	Indicates whether the tweet has been liked	bool
Id	Unique identifier of the tweet	int64
Id str	String version of the unique identifier	object
Is quote status	Indicates whether this is a quoted tweet	bool
Lang	Indicates the language of the text (machine generated)	object
Possibly sensitive	When the tweet contains a link it indicates if the content of the URL is identified as containing sensitive content	object

Table 2. *Cont.*

Field	Description	Type
Retweet count	Indicates how many times has the tweet been retweeted	int64
Retweeted	Indicates whether the tweet has been retweeted	bool
Source	Utility used to post the tweet	object
text	Text written by the user	object
Truncated	Whether the value of the text parameter was truncated	bool
User	Dictionary of information about the user who posted the tweet	object

3.2. Images

Since the tweets have been retrieved using flood-related tags, most of the images contained in the dataset are actually related to floods. Among the images that have been classified as not containing roads, some of them contain charts or weather maps, others contain information about floods which is not related to roads, whereas some images contain no flood information. The images containing evidence of passable roads, in many cases, show cars crossing the road or have enough surrounding contextual information that allows it to be inferred that the water level is not very high, whereas the images containing evidence of roads with negative passability contain cars stuck in roads and people crossing the street with boats in many cases. Some examples of the images contained in the dataset are given in Figure 2. Sometimes the differences between positive and negative road passability are very subtle and subjective (e.g., see Figure 2i,j), while we believe others are wrongly classified (e.g., see Figure 2k,l).

(**a**) ER: no

(**b**) ER: no

(**c**) ER: no

(**d**) ER: no

(**e**) ER: yes; EPR: yes

(**f**) ER: yes; EPR: yes

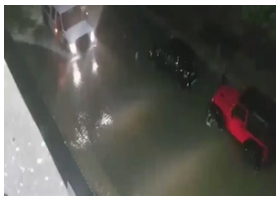
(**g**) ER: yes; EPR: no

(**h**) ER: yes; EPR: no

(**i**) ER: yes; EPR: yes

(**j**) ER: yes; EPR: no

(**k**) ER: no



(**l**) ER: no

Figure 2. Examples of images from the dataset. The **first row** (**a**–**d**) contains images classified as not containing Evidence of Roads (ER), while the **second row** (**e**–**h**) contains images classified as containing evidence of roads and their corresponding Evidence of Road Passability (ERP). The **third row** (**i**–**l**) corresponds images that were difficult to classify or wrongly classified ones.

4. Proposed Solutions

In this section, we describe a solution using only metadata information, a solution using only the tweeted image and a solution that combines both sources of information.

4.1. Algorithm Based on Metadata Only

As explained in Section 3, each tweet contains 29 different fields, but only 16 of them had non-empty values in at least 90% of the tweets. Therefore, the other 13 features were discarded since they do not contain enough information to give any statistical significant information. Moreover, we discarded the following features: (i) "*Created at*", which contained the date in which the tweet was posted. Since the tweets were collected during specific hurricane events (namely, Harvey, Irma and Maria), we considered this field to have a very limited time coverage with the risk of being biased and, therefore, not useful. Specifically, the development set contains tweets from 38 different days. (ii) "*Extended entities*", which contains structural information about the tweet, such as the icon and image sizes, their urls and ids and, therefore, it does not provide any relevant information; (iii) "*Id*" and "*Id str*" fields are automatically generated to guarantee uniqueness to the tweet thus, do not contain any meaningful information; (iv) "*Truncated*" contains a constant value, which is equal for each tweet in the development set; (v) "*Source*" and "*User*": contained features pertinent to Twitter and the user profile, such as "id", "profile image url", "friends count", which is information not relevant to our purposes. Additionally, we verified that the development set rarely contained multiple posts from the same user: this lack of information prevented the extraction of data for determining a possible positive (or negative) influence toward our goals.

As for the "*Lang*" feature, since most of the tweets were in English and all the other languages were very minoritary, we transformed it into a binary value "*originally_en*" to state whether the language of the tweet was English. To ensure that all features would contribute equally to the loss function used to train our proposed approaches, we normalized the features "*Favorite count*" and "*Retweet count*" between 0 and 1, which we named "*favorited_norm*" and "*retweeted_norm*", respectively. Finally, we also discarded the features corresponding to "*Favorited*" and "*Retweeted*" since they are subsumed by the former ones.

To determine a correlation between the normalized fields: "*favorited_norm*", "*is_quote_status*", "*originally_en*", "*possibly_sensitive*" and "*retweeted_norm*" and the task at hand, we built a point-biserial correlation matrix between each feature and the "ER" and "ERP" ground truth using the Pearson correlation coefficient. As seen on the point-biserial correlation matrix from Figure 3, none of the features have a very strong correlation with the ground truth; however, we decided to keep the fields "*favorited_norm*", "*originally_en*" and "*retweeted_norm*" since they are the highest correlated features.

We expected the text written by the user ("*Text*") and the hashtags of the tweet ("*Entities*") to be the most informative features, which we concatenated, obtaining a single sentence. To help the training, we translated all the texts into English, tokenized the words, filtered stopwords (i.e., emojis, urls, special characters, articles, conjunctions) and lemmatized the sentence. Finally the sentences were transformed into a matrix using a word embedding initialized with GloVe [26] weights, transforming each word into a vector of 200 dimensions. To be processed by a neural network, the matrices generated from *Text* and *Entities* have been standardized to have the same number of word vectors—sentences shorter than 30 words (the maximum length of a processed sentence in the dataset) have been filled with zero padding. As other state-of-the-art works [51], the 30×200 matrices have been fed in a Bidirectional Long Short-Term Memory (BiLSTM) network. Then, the output was concatenated with the *extra fields* and fed into two parallel fully-connected (FC) layers with a softmax classifier, one per task. In each FC layer, we used the cross entropy $H(y, \overline{y})$ as loss function, where y is the class annotation and $\overline{y}$ is the model prediction. Denoting by $H_{ER}(y, \overline{y})$, the loss function for the ER task and $H_{ERP}(y, \overline{y})$ the loss function for the ERP task, the overall loss $H_{TOT}(y, \overline{y})$ is set to be the sum of the preceding two. Finally, the outputs from the two FC networks have been thresholded (with the threshold set to 0.5, which is the typical threshold taken in these contexts since the output ranges between 0 and 1). The first FC layer output is the prediction for the *ER* task, while the second FC layer output, which represents the prediction for the

ERP task, is combined with the first output through a logical AND operation. This operation avoids the network to predict inconsistent situations, such as having Evidence of Roads Passability while there is No Evidence of Roads. A representation of the architecture is shown in Figure 4.

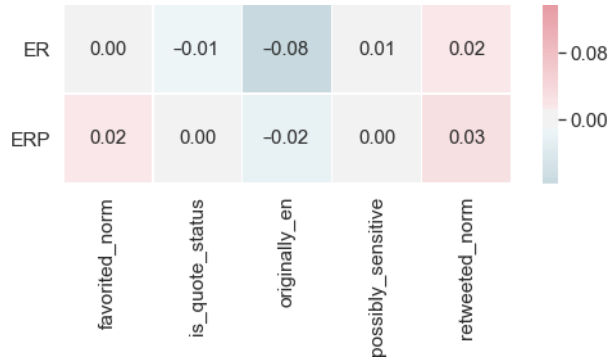


Figure 3. Correlation matrix between the ground truth features, "*ER*" (Evidence of Roads) and "*ERP*" (Evidence of Roads Passability) as two separate binary values, with the numerical features that were not discarded.

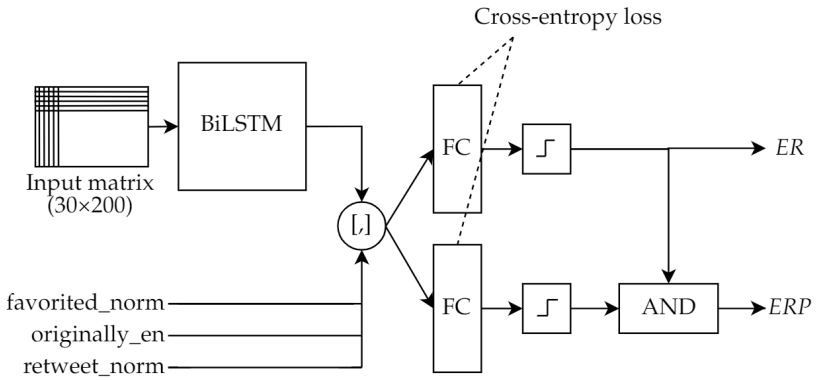


Figure 4. Architecture of the neural network to process metadata. The input matrix, composed by stacked word embeddings representing the tweeted text and hashtags is processed by a Bidirectional Long Short-Term Memory network (BiLSTM). Its output is concatenated with other metadata information representing whether: (i) the tweet has been favorited, (ii) the tweet was originally written in English, (iii) it was retweeted. Then, two Fully Connected (FC) layers are dedicated to deal with each task: the FC on the top will determine the Evidence of Roads (ER), while the other one will determine the Evidence of Roads Passability (ERP). The classification is obtained by thresholding their output. Finally, to guarantee consistent classifications, the output of the ERP classifier is combined with the output of the ER classifier by a logical AND operation.

4.2. Algorithms Based on Image Only

In this subsection, we explain first the "ensemble image base architecture", which is the solution that we presented in the *Flood classification challenge*. Then, we present the new proposed architecture which will be referred to as "double-ended network" and finally, we will introduce the extra loss that we have applied to the learning process, which will be referred as "compactness loss".

4.2.1. Ensemble Image Base Architecture

For this solution, we considered both tasks as two, separated, two-class classification problems. Since performance is prioritized over computation and operational time, we created an ensemble of networks, using 9 state-of-the-art networks: InceptionV3, Xception, VGG16, VGG19, InceptionResNetV2, MobileNet, DenseNet121, DenseNet201, NaSNetLarge. Since the dataset was too small to train the networks from scratch, we pre-trained all the networks on ImageNet [35], freezed

the parameters from the first half of the network and fine-tuned the parameters from the second half. All nine models were separately trained for both classification problems—one network for the classification of road passability and one for the classification in positive or negative road passability. The networks trained on the detection of positive or negative road passability were trained using only the images with evidence of road passability according to the ground truth. Moreover, in order to prevent overfitting, the dataset was randomly divided into training (75%) and validation (25%) sets. The validation set was used to prevent the networks to overfit to the training set—after training the networks in the training set, we stored the models that performed better in the validation set. Finally, in order to prevent overfitting while exploiting the whole dataset, we performed cross-validation using 5 different train-validation folds. Each fold was generated using a random split of the development set into 75% train and 25% validation, this means that there is some overlapping between splits. Each network was trained in each fold and for each task separately, resulting in a total of 90 networks, 45 convolutional neural networks for each task (or 5 networks per network architecture and task). The network architecture is shown in Figure 5a.

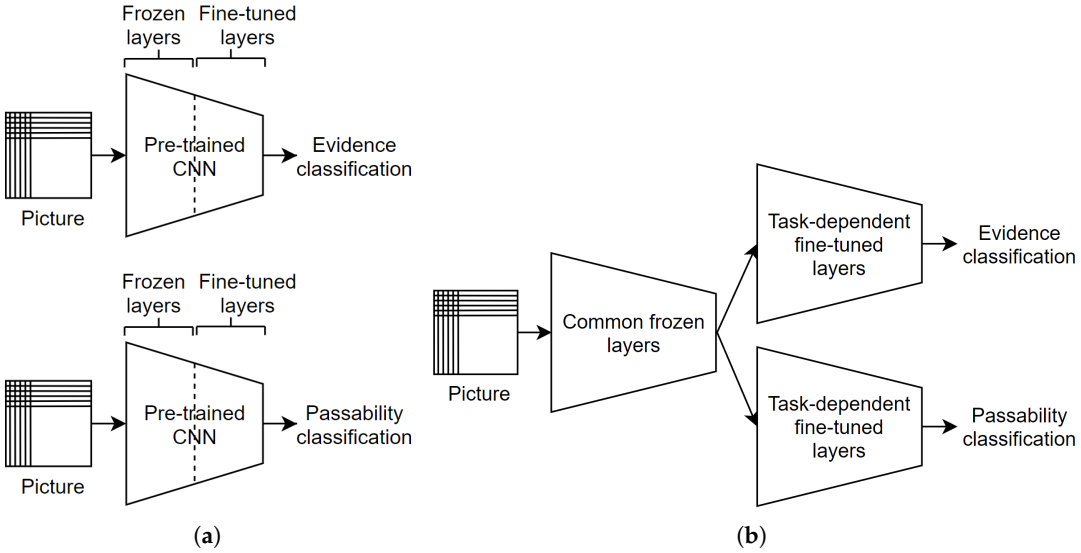


Figure 5. (**a**) Schematic of the "ensemble base image architecture" which is composed by two separate pre-trained networks in which their last layers were fine-tuned on their respective task. (**b**) An equivalent architecture which shares the first layers and then diverges into two separate branches for each task.

The output of each network is a number between 0 and 1 which represents the probability of the picture containing evidence of road passability and whether the road has a positive or a negative passability, respectively. In order to ensemble the results of all the networks, we decided to allocate the same weight to each sub-model and we applied two different aggregation methods. On the one hand, we applied a classical average aggregation prediction. On the other hand, we applied a combination of an average aggregation with a majority voting aggregation. These two aggregation methods are defined, respectively, by

$$\text{pred}_1(p_1, \ldots, p_n) = (\overline{p} > 0.5),$$

and $\text{pred}_2(p_1, \ldots, p_n) =$

$$\begin{cases} 1 & \text{if } \left(\overline{p} > 0.45 \text{ and } voting(p_1, \ldots, p_n) \geq \frac{n}{2}\right) \text{ or} \\ & \left(\overline{p} > 0.5 \text{ and } voting(p_1, \ldots, p_n) > \frac{n}{2} - 2\right), \\ 0 & \text{otherwise,} \end{cases}$$

where n is the number of networks, p_i is the probability given by the i^{th}-network of the picture belonging to Class 1, which corresponds to having positive Evidence of Road (ER) for the first task and having positive Evidence of Passable Road (EPR) for the second task. $\overline{p}$ is the average of p_i for all $1 \leq i \leq n$ and $voting$ is given by $voting(p_1, \ldots, p_n) = \{i \; p_i > 0.5, 1 \leq i \leq n\}$, where . is the set cardinality.

Thresholding over the average of the predictions $\overline{p}$ or taking the majority vote are two largely adopted approaches to deal with ensemble model predictions. However, their combination through a logical "and" tends to benefit the prediction of negative outcomes (no ER, nor ERP) with the result of lowering the number of matches with the ground truth. Therefore, we added two variables x and y to weaken the constraints, defining the following function $\text{pred}_2(p_1, \ldots, p_n, x, y) =$

$$\begin{cases} 1 & \text{if } \left(\overline{p} > 0.5 - x \text{ and } voting(p_1, \ldots, p_n) \geq \frac{n}{2}\right) \text{ or} \\ & \left(\overline{p} > 0.5 \text{ and } voting(p_1, \ldots, p_n) > \frac{n}{2} - y\right), \\ 0 & \text{otherwise.} \end{cases}$$

To determine the best values for the two variables, we applied the aggregation methodology with the grid search [52] approach in the validation set: the variable x was set to range from 0 to 0.5 with a step of 0.05, while y was set to range from 0 to $\frac{n}{2}$ with a step of 1. As a result, the assignments that maximized the number of matches between the model's predictions and the dataset annotations were $x = 0.05$ and $y = 2$.

Despite being a simple and effective model, in fact the winning solution of the challenge, this solution requires a long training process as well as high computation cost and time during testing. Moreover, the solution requires a lot of storage space, since parameters trained on 90 different networks are saved.

4.2.2. Double-Ended Network

The ensemble base image architecture relies on two networks that were trained and tested separately to solve each task individually. This architecture is represented in Figure 5a. However, since we are using a pre-trained network and freezing half of the model, both tasks share the first parameters of the model. Thus, we reorganized the solution as a single model where the first part of the model has the shared parameters and then diverges into two branches, each one with the specific parameters learned for each task, as represented in Figure 5b. This solution is equivalent to the two separate networks in terms of performance but it is lighter, end-to-end and computationally less expensive since we do not run the image through the same layers twice. Starting from this idea, and knowing that in the literature it has been stated that networks trained to perform two related tasks simultaneously can achieve better performance on both tasks than if they were trained separately [53], we decided to propose the model represented in Figure 6. This architecture is similar to the one from Figure 5b but the division of the two branches is at the end of the last convolutional layer. After the last convolutional layer, the network is divided into two branches, one for each task with two consecutive fully connected layers per task. In this case, the first half of the shared parameters are frozen during training while the other half are fine-tuned jointly.

To validate that as hypothesized, both tasks are indeed related and what has been learned for one task could benefit the other task, we decided to check the activation maps triggered by the networks trained on both tasks separately. To do so, we used the gradient-weighted class activation mapping (Grad-CAM). This is a technique proposed in [54] which highlights the regions which triggered the Convolutional Neural Network (CNN) to make its classification by analyzing the activations of a convolutional layer of the network. We extracted the heatmap of the gradient activations from the last convolutional layer of the single network trained on both tasks separately. In Figure 7, we present the corresponding activation heatmaps for four images from the validation set. As seen in Figure 7, the activation scale ranges from blue to red in the images, where red corresponds to the greatest activation (MAX) and blue to the minimum one (MIN). The upper row of images corresponds to the

activations of the network trained on evidence of road classification, while the second row corresponds to the activations of the same images on the network trained on road passability classification. As we can observe in the figure, while the activations from both networks are different, the reddish area is located in a different part of the image, because one focuses more on water in general and the surroundings to determine if it corresponds to a road and the other tends to give more importance to the objects in the water to determine if the road is passable or not and even though both networks were trained separately, their activations are highly correlated, since both tasks are also highly correlated. This supports our hypothesis that by training both tasks simultaneously and having both tasks share more parameters, one task could benefit from what the other has learned.

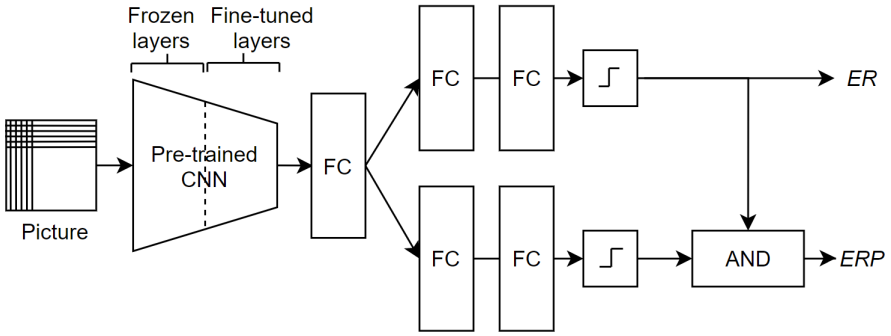


Figure 6. Representation of the double-ended architecture.

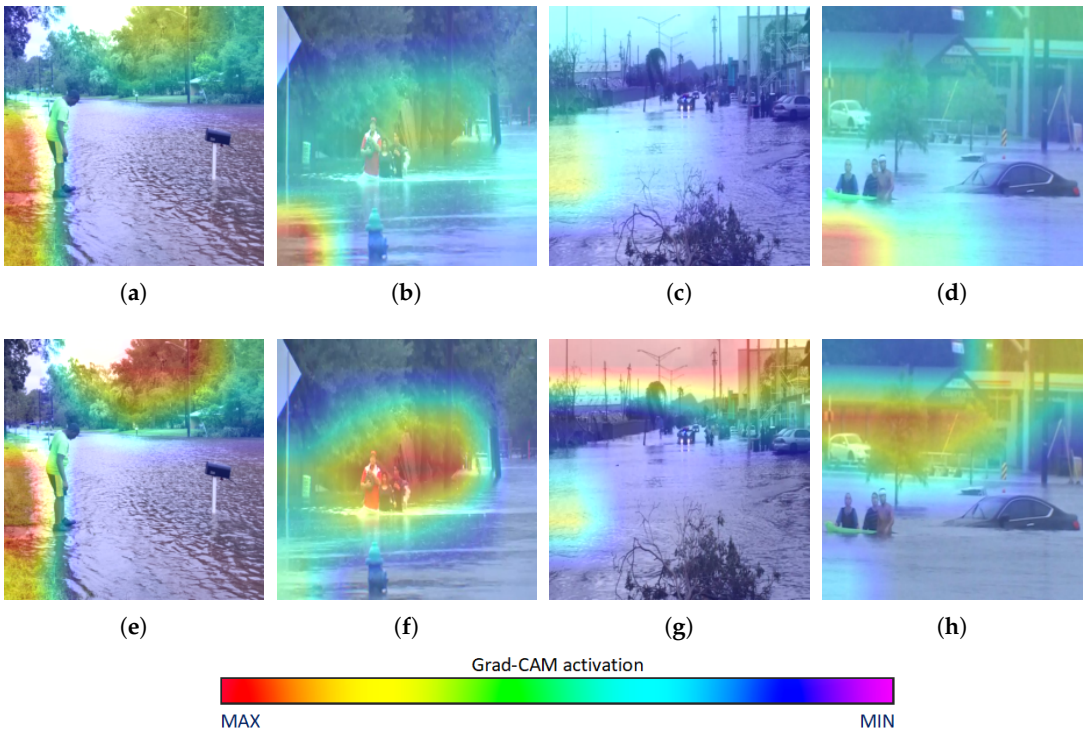


Figure 7. Examples of images created with Grad-CAM. The first row (**a**–**d**) corresponds to the activations triggered by the network trained on the evidence task (ER), while the second (**e**–**h**) has the activations, for the same image, of the network trained on the passability task (ERP). At the bottom of the figure, there is a scale that is used to highlight the activations of the model, where red represents a higher activation and purple means minimal activation. As seen in the first row, the activations for the evidence of the road task are maximal in the water regions. In the second row, which corresponds to evidence of road passability, there is a greater activation for the elements located both in and outside the flooded area. This is because it is necessary to rely on these elements to identify the height of the water.

4.2.3. Compactness Loss

As previously explained, the problem is divided into two tasks: (i) detecting if the image has evidence of roads and (ii) if the image has been classified as containing a road, determining whether the road is passable or not. The first task could be considered as a *binary classifier* ("evidence of road passability" and "no evidence of road passability"), but the concept of "not having any evidence of road passability" could also be subsumed by "anything which is not contained in the first class". Thus, the problem could also be considered as a *one class classification* or as an *out of distribution problem*, where "evidence of road passability" would be the class to classify (or the in distribution class). The advantage of considering the problem as a one-class classification problem rather than a binary classification problem is that one-class classification algorithms take into account that the out of distribution class is not only defined by the images used for the training, but that it could be anything that has not previously been seen during the training phase. Similarly, the second task could be considered as a *binary classification problem* ("passable" and "impassable" roads) or as a *one class classification problem* "passable road".

For this solution, we considered both tasks as a one-class classification problem. Taking inspiration from [55], we wanted the features extracted from the first fully connected layer to be as descriptive as possible for the class at hand, meaning that the feature representation of that class' images will be distinctive from the feature representation for images not belonging to the class and, at the same time, we would like a low intra-class distance, meaning that features from the same class should be as close as possible in the feature space. This optimization can be described as $\hat{g} = \max_g \mathcal{D}(g(t)) + \lambda \mathcal{C}(g(t))$, where: g is the deep feature representation for the training data t, λ is a positive constant and $\mathcal{D}$ is the *Descriptive loss function* (within this approach, we used the cross-entropy) and $\mathcal{C}$ is the *compactness loss function*, which evaluates the batch inter-class deep feature distance to derive objects from the same class. This compactness loss can be applied to either of the two previous models, the ensemble base image architecture or the double-ended architecture. In the first case, we would add the compactness loss to the first fully connected layer. For the double-ended architecture, we would replace the last fully connected layer by a fully connected layer followed by two fully connected layers in parallel—one for each task—and add a compactness loss for each task, as shown on Figure 8. The outputs of both final fully connected layers are two real values in the range (0,1). They represent the percentage of which the evidence is believed to be of roads and of passable roads, respectively. The two outputs are then rounded according to a threshold of 0.5. Therefore, the first output is the classifier for the ER class. On the other hand, to avoid inconsistent classifications (i.e., ER = false and ERP = true), the second output is multiplied by the first one, determining the classifier for the ERP class. Note that the decision of which fully connected layer accomplishes the ER and which the ERP tasks is taken during the network training phase.

For the compactness loss, we implemented the same loss as the one proposed in [55], which is given by

$$l_c = \frac{1}{nk} \sum_{i=1}^{n} \mathbf{z_i}^T \mathbf{z_i} \tag{1}$$

where $\mathbf{z_i} = \mathbf{x_i} - \mathbf{m_i}$, being $\mathbf{x_i} \in \mathbb{R}^k$ the samples of the batch of size n for all $1 \leq i \leq n$ and $\mathbf{m_i} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{x_j}$, the mean of the remaining samples. As it is proved in [55], this compactness loss is in fact a scaled version of the sample variance given by

$$l_C = \frac{1}{nk} \sum_{i=1}^{n} \frac{n^2 \sigma_i^2}{(n-1)^2}, \tag{2}$$

where σ_i^2 is the sample variance for all $1 \leq i \leq n$.

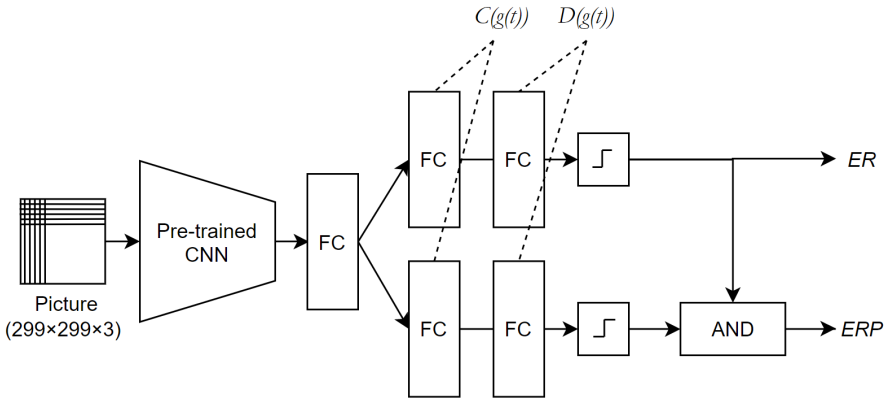


Figure 8. Double-ended classifier with compactness loss. The model is based on the Inception V3 network, replacing the last fully connected layer by a 1024 fully connected layer which extracts the image features and two parallel fully connected layers, one for each task. Two losses are trained simultaneously, a compactness loss to ensure low-intra class feature distance and a descriptiveness loss, to ensure a high-inter class feature distance.

In order to implement the backpropagation, we need to compute the gradient of l_c with respect to the input x_{ij}. In [55], the derivation of the backpropagation formula obtained from the gradient of l_C with respect to x_{ij} contains a mistake. Indeed, in Appendix A in [55], it is stated that the gradient is given by the following equation

$$\frac{\partial l_C}{\partial x_{ij}} = \frac{2}{(n-1)nk}\left[n \times \left(x_{ij} - m_{ij}\right) - \sum_{l=1}^{n}\left(x_{il} - m_{il}\right)\right]. \tag{3}$$

In Appendix A, we prove that the gradient is acutally given by the following equation:

$$\frac{\partial l_C}{\partial x_{ij}} = \frac{2}{(n-1)nk}\left[n \cdot \left(x_{ij} - m_{ij}\right) - \sum_{l=1}^{n}\left(x_{lj} - m_{lj}\right)\right].$$

4.3. Algorithm Based on Metadata and Visual Information

To combine the information from the metadata and the images, any of the previously proposed solutions for the image-only architecture can be combined with the metadata-only architecture by concatenating the features extracted from the bi-directional LSTM with the features extracted by the convolutional network, as seen in Figure 9.

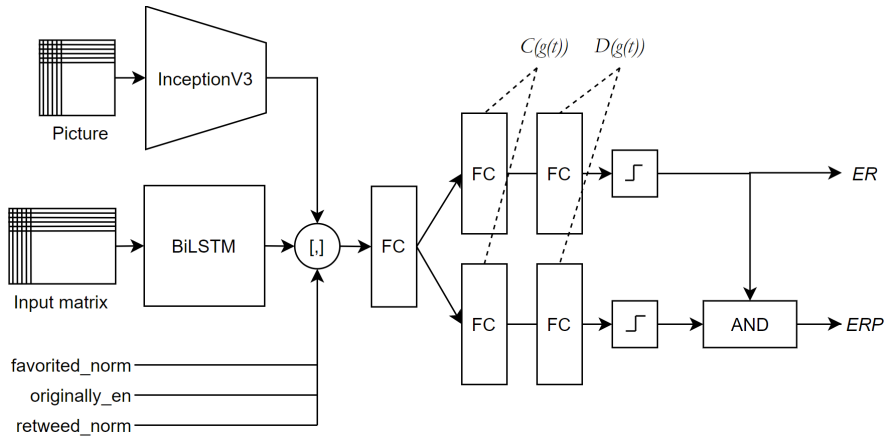


Figure 9. Combination of the *Double-ended classifier with compactness loss* and the metadata system.

5. Evaluation and Results

As we have already commented, part of this work has been carried out for a competition in the MediaEval 2018 workshop, in which a total of nine teams participated. In this section, we will compare not only the results of the different methods proposed in this paper but we will also compare them with the results of all the other participants of the workshop. However, since this is an extension of the work done for the competition and we do not have access to the ground truth of the test set, we had to divide the training set into training and test sets, as explained in Section 3. Therefore, the results given for our models will be tested on a different set as the ones for the competition. To make the comparison as fair as possible, we created a validation set from our training set, to validate the models and tune the hyperparameters. The test set was only used to provide the final results for the paper.

In addition, in order to have an understandable baseline to compare the results, we asked four people to perform the task on a subset of 50 images. These persons were external to the project but had knowledge about artificial intelligence and computer vision. They received a verbal explanation of the task in the lines of the explanation given by the challenge organizers and they were not given any examples before starting the annotation.

All the results provided in this section will be given in terms of F1-Score, the harmonic mean of precision and recall. For the human annotators, we will give the results as the average of their F1-Score. It is important to note that the second task, the classification between passable and not passable roads, depends on the first task since if an image has been classified as not containing evidence of road passability, it will not be considered for the second task. Therefore, a false negative detection in the first task (an image wrongly classified as not containing evidence of road passability) will also count as a false negative in the second task, regardless of its ground truth. At the same time, a false positive in the first task (an image wrongly classified as containing evidence of road passability) will also count as a false positive for the second task. Due to this error propagation, the performance of the second task cannot be higher than the performance of the first task.

We will evaluate the results of the proposed models in this paper in the same order that we have presented them in the previous section.

5.1. Results Using Metadata Only

In Table 3, the results of the model using metadata information only are provided. As it can be seen from the table, the performance in general is quite low, even in the case of human annotators. We believe that this is because a lot of the tweets have little information about the tasks in their metadata.

Table 3. F1-Scores on the challenge test set for both tasks using metadata information only. * Results given on our own test set. ** Results on a subset of 50 images.

Approach\Data	Evidence of Road [%]		Ev. of Road Passability [%]	
	Validation Set	Test Set	Validation Set	Test Set
Human annotation	51.48 **	-	18.18 **	-
Metadata only	59.93	62.56 *	56.82	57.05 *
Y. Feng et al. [20]	-	-	-	32.8
M. Hanif et al. [21]	-	58.30	-	31.15
Z. Zhao et al. [46]	-	32.60	-	12.86
A. Moumtzidou et al. [22]	-	-	-	30.17
A. Kirchknopf et al. [23]	-	-	-	20

5.2. Results Using Images Only

In the case of images where we have proposed several models, first we are going to comment and compare the results of the different models that we proposed in this article and then compare them with the algorithms proposed by the rest of the participants. Firstly, we have presented the "ensemble base image architecture" which was the algorithm presented for the competition. This algorithm is mainly

based on performing iterative cross-validation to train models and ensemble them, for which we proposed a new ensembling technique. Since for this paper we have carried out a new training and test split, to make the comparison as fair as possible, we have retrained this architecture on the new training set and tested on the new test set. In Figures 10 and 11, we show how the F1-Score evolves for both tasks as we ensemble more models and the difference between the different ensembling techniques. As it can be seen from the curves, both tasks benefit from the ensembling, particularly for the first networks, then the performance stabilizes. The evidence of road benefits more from this technique than the passability of road task, and the curves for the passability task are less stable. We believe that this is due to the difference in the difficulties of both tasks. As can be seen towards the end of the graphs, the results of both tasks begin to slowly worsen. That is because (i) we are adding different architectures and some of them yield better results on average than others; (ii) we have stacked the networks in order of the architectures' average performance and thus it gets a point in which adding more architectures starts degrading the results. Given the information from both graphs, the ensemble of more than 30 models (up to 90, in our test case) does not significantly improve the performance. The ensemble of 90 networks (45 per task) was the winning architecture presented in the MediaEval competition, and its performances are presented in Table 4. This is the only architecture for which we have results on both the challenge and our own test sets. The results of the MediaEval test set are very similar to the results obtained for our own test set, which indicates that the difficulty of both sets is quite similar, allowing us to make a fair comparison with the results of the other participants. Some differences might be due to the fact that we have had to retrain all the networks to fit them to the new training, validation and test set. As for the ensembling technique, voting seems to have slightly worsened the results in the evidence of road task while averaging led to worsening of the results in the road passability task. For this reason, we decided to use the ensembling technique that we proposed in this paper for the final results, since it is the technique that has the most stable results.

The use of the ensemble models makes sense for a competition; however, it might not be suitable for a real life application, since it tends to be computationally expensive and time consuming. Therefore, we focused our analysis to compare the best available model, obtained without taking into account computational limitations, with a lightweight version, proposed in this work. We started using an "ensemble of one-model per task", that we named "single-network image base image architecture", and then we compared it with the "double ended architecture", presented in the previous section. To reduce the randomness associated with the training process, both architectures were initialized with the same weights and used the same hyperparameters and stopping criterion. As it can be seen from Table 4, the improvement is quite significant, particularly in the passability task. We believe that this is because the passability task has significantly fewer images to train when trained separately so it is more difficult for the model to generalize to new data and when trained together, the passability task can benefit from what the evidence task has learned. On top of obtaining better results than the single network base image architecture", the "double ended architecture" has fewer parameters, making it lighter and computationally less expensive, and it is an end-to-end architecture.

Then, we proposed to use the "compactness loss" to make the model more robust to unseen data. We retrained the previous model with the compactness loss on each branch, as shown in Figure 8. The results of both the validation and test set are in Table 4. Although we cannot extract direct evidence from this table that the compact loss improves the results, it does seem to generalize better to the test set since the results from validation to test are more similar than the ones without compactness loss.

Finally, by combining the improvements using the double-ended classifier and the compactness loss, we are able to reach almost the same performance as we had obtained using the ensemble of 30 models, meaning that we have a model that is almost 30 times lighter and faster with a comparable performance.

Table 4. F1-Scores on the challenge test set for both tasks using only the images from the tweets. * Results on a subset of 50 images.

Approach\Data	Evidence of Road [%]			Ev. of Road Passability [%]		
	Validation Set	Test Set (MediaEval)	Test Set (Own)	Validation Set	Test Set (MediaEval)	Test Set (Own)
Human annotation	87.32 *	-	-	47.71 *	-	-
Ensemble image base architecture (90)	90.14	87.79	90.17	64.33	68.38	65.91
Ensemble image base architecture (30)	88.91	-	89.45	70.18	-	65.28
Single network image base architecture	86.48	-	84.88	62.84	-	59.99
Double-ended architecture	88.73	-	85.00	67.51	-	67.91
Double-ended with compactness loss	87.78	-	86.42	67.49	-	68.53
Y. Feng et al. [20]	-	-	-	-	64.35	-
M. Hanif et al. [21]	-	74.58	-	-	45.04	-
Z. Zhao et al. [46]	-	87.58	-	-	63.13	-
A. Moumtzidou et al. [22]	-	-	-	-	66.65	-
A. Kirchknopf et al. [23]	-	-	-	-	24	-
N. Said et al. [48]	-	-	-	-	65.03	-
D. Dias [47]	-	-	-	-	64.81	-
B. Bischke [45]	-	87.70	-	-	66.48	-

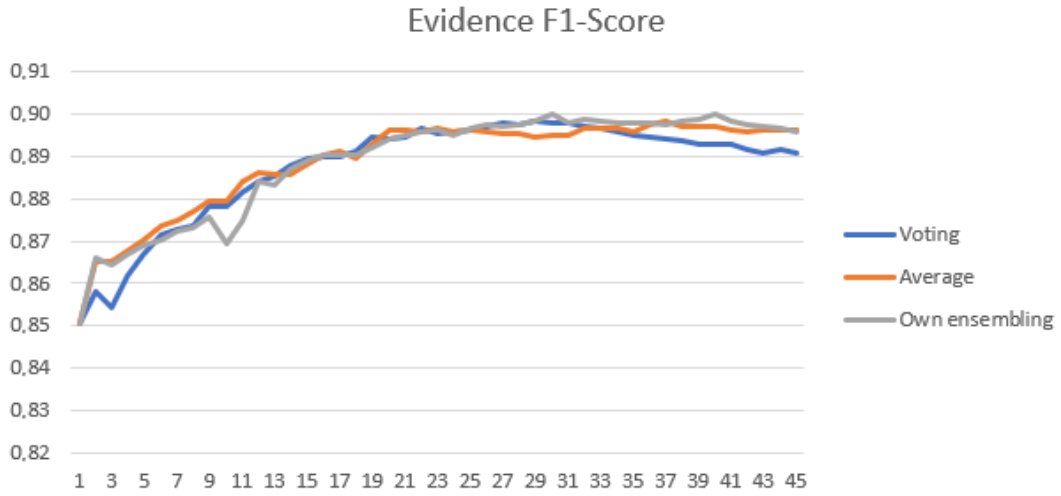


Figure 10. Evolution of F1-Score on the road evidence task as we ensemble more networks and make a comparison between the three different ensembling techniques.

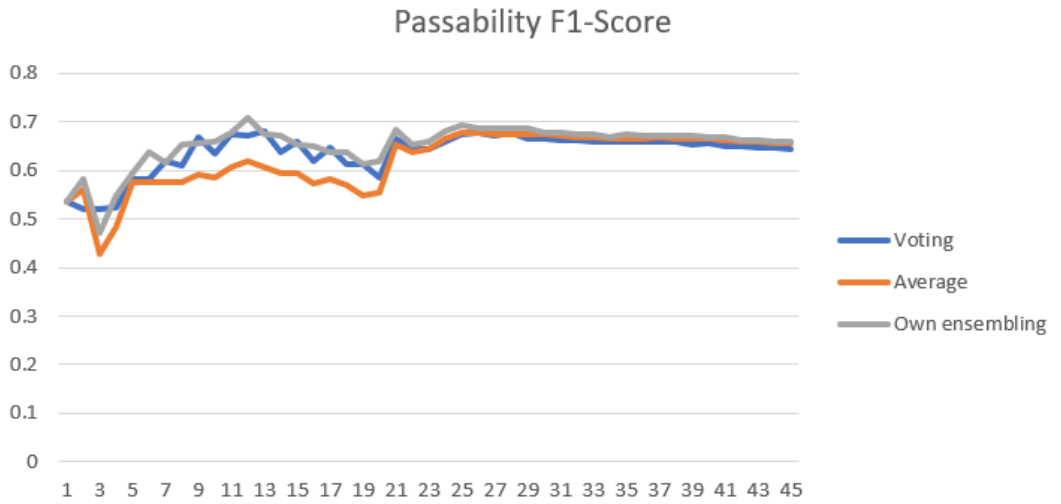


Figure 11. Evolution of F1-Score on the road passability task as we ensemble more networks and make a comparison between the three different ensembling techniques.

It is remarkable that the results using images are considerably better than the ones using metadata, not only in our case but also for humans or other participants. That is because the dataset has been

built and annotated using only the visual information, thus we know that the images should have enough information to solve the problem, but the metadata does not necessarily always have this distinctive information.

5.3. Results Using Images and Metadata

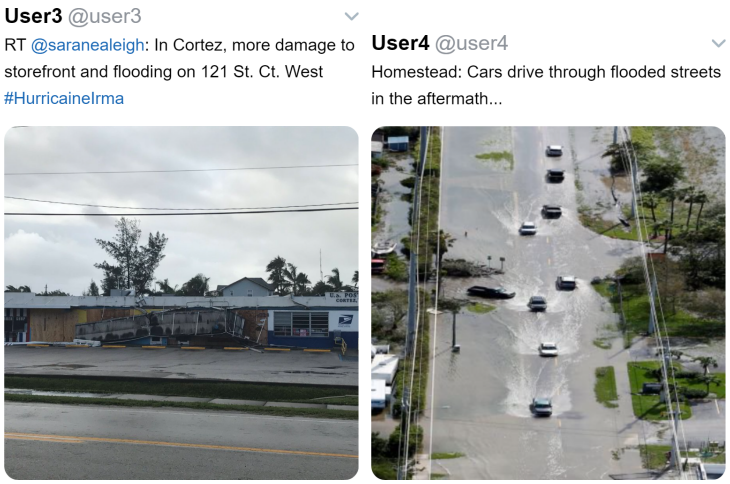
As a final step, we combined the previous best model with the metadata information. As it was not clear from the previous results if the compactness loss provided a significant boost in performance, we tried combining the metadata with the double-ended classifier with and without compactness loss. The results are given in Table 5. In this case, we can notice a considerable improvement in the model with compactness loss relative to the one without it. In fact, this model achieves only 3% below the best score in the evidence of road task, while it obtains almost a 10% improvement in evidence of road passability compared to the second best participant. Finally, it seems like adding the metadata information improves the road passability task. To understand how the metadata information can help to improve the results, in Figure 12, some tweet examples are given that were incorrectly classified by the image only model but correctly classified by the model which combined visual and metadata information. These tweets contain some very informative keywords, such as flooded street, stalled cars and drive through.

Table 5. F1-Scores on the challenge test set for both tasks using the metadata and image information. * Results given on our own test set.

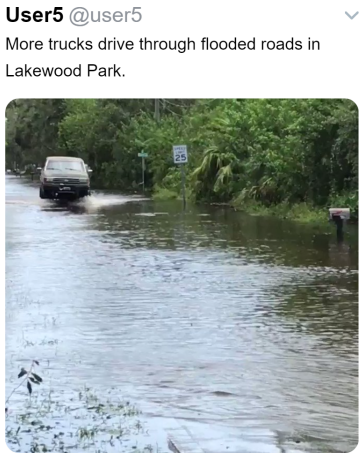
Approach\Data	Evidence of Road [%]		Ev. of Road Passability [%]	
	Validation Set	Test Set	Validation Set	Test Set
Double-ended architecture	78.96	86.99 *	61.06	62.96 *
Double-ended with compactness loss	77.85	84.56 *	73.61	75.93 *
Y. Feng et al. [20]	-	-	-	59.49
M. Hanif et al. [21]	-	76.61	-	45.56
Z. Zhao et al. [46]	-	87.58	-	63.88
A. Moumtzidou et al. [22]	-	-	-	66.43
A. Kirchknopf et al. [23]	-	-	-	35

(**a**) (**b**)

Figure 12. *Cont.*



(**c**) (**d**)



(**e**)

Figure 12. Examples of tweets contained within the dataset that allowed to disambiguate the visual content for a correct ERP prediction. In (**a**–**e**) the text of the tweets allowed to resolve ambiguous ERP images thanks to keywords such as "flooded" in relation to roads and "drive through" in reference to vehicles.

6. Conclusions

In this paper, we have extended the work conducted for the *Flood classification challenge* in MediaEval 2018 in which we had presented a winning architecture based on a fine-tuning and ensembling of 45 models for each task. In this paper, we evaluate the evolution of the performance of the algorithm as we ensemble more models. It is determined that by ensembling models in order of average performance, we can obtain some improvements in terms of F1-Score for both tasks. However, as we ensemble networks, at some point, adding more networks to the ensemble starts worsening the results. Thus, by reducing the number of ensembled models to 30, we can reduce the number of ensembled networks, making the solution faster and lighter while improving the results. Then, we proposed a double-ended architecture that trains both tasks simultaneously to further reduce the number of parameters of the solution and to let the network share knowledge between tasks. We also propose the usage of a compactness loss, a loss proposed to convert a binary classifier network into a one-class classification network. In the original paper, the derivation of this loss contained a mistake which is corrected in this paper. Through experiments, we conclude that, by combining the double-ended architecture and the compactness loss, we are able to obtain a single network that solves both problems, achieving comparable results for the evidence of road task and better results

for the road passability estimation compared to the ensemble models. Since this solution does not rely on ensemble networks, it is almost 90 times faster and also lighter than the originally proposed architecture, making it a viable solution for a real-life application.

Author Contributions: Conceptualization, L.L.-F., A.F. and M.Z.; methodology, L.L.-F., A.F. and M.Z.; software, L.L.-F., A.F. and M.Z.; validation, L.L.-F., A.F., M.Z., H.S. and P.G.; formal analysis, L.L.-F., A.F., M.Z., H.S. and P.G.; investigation, L.L.-F., A.F. and M.Z.; resources, L.L.-F., A.F. and M.Z.; data curation, L.L.-F., A.F. and M.Z.; writing—original draft preparation, L.L.-F., A.F., M.Z., H.S. and P.G.; writing—review and editing, H.S. and P.G.; visualization, L.L.-F., A.F. and M.Z.; supervision, H.S. and P.G.; funding acquisition, L.L.-F., A.F. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the Spanish Grant FEDER/Ministerio de Economía, Industria y Competitividad - AEI/TIN2016-75404-P and the European Commission H2020 FASTER project no. 833507. The APC was funded by University of the Balearic Islands.

Acknowledgments: Laura Lopez-Fuentes benefits from the NAERINGSPHD fellowship of the Norwegian Research Council under the collaboration agreement Ref.3114 with the UIB.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In Appendix A from [55], it is stated that the gradient is given by the following equation

$$\frac{\partial l_C}{\partial x_{ij}} = \frac{2}{(n-1)nk}\left[n \times (x_{ij} - m_{ij}) - \sum_{l=1}^{n}(x_{il} - m_{il})\right]. \tag{A1}$$

However, there are some mistakes in that equation. The first mistake is within the summation since the samples $\mathbf{x_i}$ have k components, not n. However, as we will prove, this is not a unique mistake.

Let us compute the gradient of l_C with respect to x_{ij}. Using the definition of the inner product, we have that $\mathbf{z_i}^T\mathbf{z_i} = \sum_{t=1}^{k} z_{it}^2$. Thus, l_C can be written as

$$l_C = \frac{1}{nk}\sum_{l=1}^{n}\sum_{t=1}^{k}(x_{lt} - m_{lt})^2.$$

Now, taking partial derivatives of l_C with respect to x_{ij} for all $1 \leq i \leq n$ and $1 \leq j \leq k$, we obtain

$$\frac{\partial l_C}{\partial x_{ij}} = \frac{2}{nk}\sum_{l=1}^{n}(x_{lj} - m_{lj}) \cdot \left(\frac{\partial (x_{lj} - m_{lj})}{\partial x_{ij}}\right).$$

This first step is already incorrect in [55]. The rest of the proof follows similarly. Let us check it. Note that

$$\frac{\partial (x_{lj} - m_{lj})}{\partial x_{ij}} = \begin{cases} 1 & \text{if } l = i, \\ -\dfrac{1}{n-1} & \text{otherwise.} \end{cases}$$

Thus, we obtain that

$$\begin{aligned} \frac{\partial l_C}{\partial x_{ij}} &= \frac{2}{nk}\left[x_{ij} - m_{ij} - \frac{1}{n-1}\sum_{\substack{l=1 \\ l\neq i}}^{n}\left(x_{lj} - m_{lj}\right)\right] \\ &= \frac{2}{nk}\left[\frac{n}{n-1}\cdot(x_{ij} - m_{ij}) - \frac{1}{n-1}\sum_{l=1}^{n}\left(x_{lj} - m_{lj}\right)\right] \\ &= \frac{2}{(n-1)nk}\left[n\cdot(x_{ij} - m_{ij}) - \sum_{l=1}^{n}\left(x_{lj} - m_{lj}\right)\right], \end{aligned}$$

retrieving finally

$$\frac{\partial l_C}{\partial x_{ij}} = \frac{2}{(n-1)nk}\left[n \cdot \left(x_{ij} - m_{ij}\right) - \sum_{l=1}^{n}\left(x_{lj} - m_{lj}\right)\right].$$

References

1. EM-DAT. *The International Disaster Database*; Centre of Research on the Epidemiology of Disasters—CRED: Brussels, Belgium, 2019.
2. EM-DAT. *The International Disaster Database—Data Access*; Centre of Research on the Epidemiology of Disasters—CRED: Brussels, Belgium, 2019.
3. EU-Commission. *Funding Opportunities to Support Disaster Risk Prevention in the Cohesion Policy 2014–2020 Period*; European Commission: Brussels, Belgium, 2014.
4. AnsuR Technologies AS. UN-ASIGN. 2019. App. Available online: https://play.google.com/store/apps/details?id=ansur.asign.un&hl=en_US (accessed on 4 November 2020).
5. AnsuR Technologies AS. UN-ASIGN. 2019. FP7 Project. Available online: https://cordis.europa.eu/project/rcn/94375/factsheet/en (accessed on 4 November 2020).
6. Istituto Superiore Mario Boella (ISMB). I-REACT. 2019. App. Available online: https://play.google.com/store/apps/details?id=it.ismb.iReact&hl=en_US (accessed on 4 November 2020).
7. Istituto Superiore Mario Boella (ISMB). I-REACT. 2019. H2020 Project. Available online: https://cordis.europa.eu/project/rcn/203294/factsheet/en (accessed on 4 November 2020).
8. Wukich, C. Social media use in emergency management. *J. Emerg. Manag.* **2015**, *13*, 281–294. [CrossRef] [PubMed]
9. Lopez-Fuentes, L.; Farasin, A.; Skinnemoen, H.; Garza, P. *Deep Learning Models for Passability Detection of Flooded Roads*; CEUR-WS: Aachen, Germany, 2018; p. 2283.
10. Saroj, A.; Pal, S. Use of social media in crisis management: A survey. *Int. J. Disaster Risk Reduct.* **2020**, *48*, 101584. [CrossRef]
11. Kankanamge, N.; Yigitcanlar, T.; Goonetilleke, A.; Kamruzzaman, M. Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets. *Int. J. Disaster Risk Reduct.* **2020**, *42*, 101360. [CrossRef]
12. Kankanamge, N.; Yigitcanlar, T.; Goonetilleke, A. How engaging are disaster management related social media channels? The case of Australian state emergency organisations. *Int. J. Disaster Risk Reduct.* **2020**, *48*, 101571. [CrossRef]
13. Ferner, C.; Havas, C.; Birnbacher, E.; Wegenkittl, S.; Resch, B. Automated Seeded Latent Dirichlet Allocation for Social Media Based Event Detection and Mapping. *Information* **2020**, *11*, 376. [CrossRef]
14. Kruspe, A.; Kersten, J.; Klan, F. Detection of informative tweets in crisis events. *Nat. Hazards Earth Syst. Sci. Discuss.* **2020**, 1–18. [CrossRef]
15. Lopez-Fuentes, L.; van de Weijer, J.; Bolanos, M.; Skinnemoen, H. Multi-modal Deep Learning Approach for Flood Detection. *MediaEval* **2017**, *17*, 13–15.
16. Zaffaroni, M.; Rossi, C. Water Segmentation with Deep Learning Models for Flood Detection and Monitoring. In Proceedings of the 17th ISCRAM Conference, Blacksburg, VA, USA, 24–27 May 2020; pp. 66–74.
17. Lopez-Fuentes, L.; Rossi, C.; Skinnemoen, H. River segmentation for flood monitoring. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 6–10 November 2017; pp. 3746–3749.
18. Bînă, D.; Vlad, G.A.; Onose, C.; Cercel, D.C. *Flood Severity Estimation in News Articles Using Deep Learning Approaches*; CEUR-WS: Aachen, Germany, 2019; p. 2670.
19. Zaffaroni, M.; Lopez-Fuentes, L.; Farasin, A.; Garza, P.; Skinnemoen, H. *AI-Based Flood Event Understanding and Quantification Using Online Media and Satellite Data*; CEUR-WS: Aachen, Germany, 2019; p. 2670.
20. Feng, Y.; Shebotnov, S.; Brenner, C.; Sester, M. Ensembled Convolutional Neural Network Models for Retrieving Flood Relevant Tweets. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.

21. Hanif, M.; Atif Tahir, M.; Rafi, M. Detection of passable roads using Ensemble of Global and Local Features. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
22. Moumtzidou, A.; Giannakeris, P.; Andreadis, S.; Mavropoulos, A.; Meditskos, G.; Gialampoukidis, I.; Avgerinakis, K.; Vrochidis, S.; Kompatsiaris, I. A multimodal approach in estimating road passability through a flooded area using social media and satellite images. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
23. Kirchknopf, A.; Slijepcevic, D.; Zeppelzauer, M.; Seidl, M. Detection of Road Passability from Social Media and Satellite Images. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
24. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
25. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations (ICLR 2013), Scottsdale, AZ, USA, 2–4 May 2013. Available online: https://arxiv.org/pdf/1301.3781.pdf (accessed on 4 November 2020).
26. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
27. Bramer, M. *Principles of Data Mining*; Springer: Berlin, Germany, 2007; Volume 180.
28. Kim, Y. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1746–1751. [CrossRef]
29. Chatzichristofis, S.A.; Boutalis, Y.S. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *International Conference on Computer Vision Systems*; Springer: Berlin, Germany, 2008; pp. 312–322.
30. Jalab, H.A. Image retrieval system based on color layout descriptor and Gabor filters. In Proceedings of the 2011 IEEE Conference on Open Systems, Langkawi, Malaysia, 25–28 September 2011; pp. 32–36.
31. Chatzichristofis, S.A.; Boutalis, Y.S. FCTH: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In Proceedings of the IEEE 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria, 7–9 May 2008; pp. 191–196.
32. Park, D.K.; Jeon, Y.S.; Won, C.S. Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM Workshops on Multimedia*; ACM: New York, NY, USA, 2000; pp. 51–54.
33. Zagoris, K.; Chatzichristofis, S.A.; Papamarkos, N.; Boutalis, Y.S. Automatic image annotation and retrieval using the joint composite descriptor. In Proceedings of the IEEE2010 14th Panhellenic Conference on Informatics, Tripoli, Greece, 10–12 September 2010; pp. 143–147.
34. Manjunath, B.S.; Ohm, J.R.; Vasudevan, V.V.; Yamada, A. Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Technol.* **2001**, *11*, 703–715. [CrossRef]
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
36. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
38. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-V4, Inception-ResNet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations. 2015. Available online: https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/simonyan15.pdf (accessed on 4 November 2020)

41. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
42. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
43. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [CrossRef] [PubMed]
44. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
45. Bischke, B.; Helber, P.; Dengel, A. Global-Local Feature Fusion for Image Classification of Flood Affected Roads from Social Multimedia. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
46. Zhao, Z.; Larson, M.; Oostdijk, N. Exploiting Local Semantic Concepts for Flooding-related Social Image Classification. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
47. Dias, D.; Dias, U. Flood detection from social multimedia and satellite images using ensemble and transfer learning with CNN architectures. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
48. Said, N.; Pogorelov, K.; Ahmad, K.; Riegler, M.; Ahmad, N.; Ostroukhova, O.; Halvorsen, P.; Conci, N. Deep learning approaches for flood classification and flood aftermath detection. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
49. MediaEval 2018 Multimedia Satellite Task. 2018. Available online: http://www.multimediaeval.org/mediaeval2018/multimediasatellite/ (accessed on 31 May 2018).
50. Bischke, B.; Helber, P.; Zhao, Z.; de Bruijn, J.; Borth, D. The Multimedia Satellite Task at MediaEval 2018: Emergency Response for Flooding Events. In Proceedings of the MediaEval 2018 Workshop, Sophia Antipolis, France, 29–31 October 2018.
51. Wang, P.; Qian, Y.; Soong, F.K.; He, L.; Zhao, H. Learning distributed word representations for Bidirectional LSTM recurrent neural network. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 527–533.
52. Lerman, P. Fitting segmented regression models by grid search. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1980**, *29*, 77–84. [CrossRef]
53. Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; Yan, J. Fots: Fast oriented text spotting with a unified network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5676–5685.
54. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
55. Perera, P.; Patel, V.M. Learning deep features for one-class classification. *IEEE Trans. Image Process.* **2019**, *28*, 5450–5463. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).