MID-AMERICA
TRANSPORTATION CENTER

Report # MATC-MS&T: 137-1

Final Report
WBS:25-1121-0005-137-1

Spatial Attention Mechanism for Weakly Supervised Fire and Traffic Accident Scene Classification

Zhaozheng Yin, PhD
Associate Professor
Computer Science
Missouri University of Science and Technology

Ruwen Qin, PhD
Associate Professor
Engineering Management and Systems Engineering

Md Moniruzzaman
PhD Student
Computer Science

2019

A Cooperative Research Project sponsored by
U.S. Department of Transportation- Office of the Assistant Secretary for Research and Technology

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Spatial Attention Mechanism for Weakly Supervised Fire and Traffic Accident Scene Classification

Zhaozheng Yin, PhD
Associate Professor
Computer Science
Missouri University of Science and Technology

Ruwen Qin, PhD
Associate Professor
Engineering Management and Systems Engineering
Missouri University of Science and Technology

Md Moniruzzaman
PhD Student
Computer Science
Missouri University of Science and Technology

A Report on Research Sponsored by

Mid-America Transportation Center

University of Nebraska–Lincoln

June 2019

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 25-1121-0005-137-1	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle Spatial Attention Mechanism for Weakly Supervised Fire and Traffic Accident Scene Classification	**5. Report Date** June 30, 2019
	6. Performing Organization Code
7. Author(s) Zhaozheng Yin, PhD https://orcid.org/0000-0002-9602-6488; Ruwen Qin, PhD https://orcid.org/0000-0003-2656-8705; and Md Moniruzzaman.	**8. Performing Organization Report No.** 25-1121-0005-137-1
9. Performing Organization Name and Address Missouri University of Science and Technology Parker Hall, 106, 300 W 13th St, Rolla, MO 65409	**10. Work Unit No.**
	11. Contract or Grant No. 69A3551747107
12. Sponsoring Agency Name and Address Mid-America Transportation Center 2200 Vine St. PO Box 830851 Lincoln, NE 68583-0851	**13. Type of Report and Period Covered** Final Report (August 2017-June 2019)
	14. Sponsoring Agency Code

15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract

During the past ten years, on average there were nearly 16.5 thousands of hazardous materials (hazmat) transport incidents per year resulting in $82 millions of damages. Prompt, accurate, objective assessment on hazmat incidents is important for the first-responders to take appropriate actions timely, which will reduce the damage of hazmat incidents and protect the safety of people and the environment. Therefore, one of the most important steps is to automatically detect transport incidents, such as fire and traffic accidents. In this work, we introduce a simple and yet effective framework that integrates the convolutional feature maps of deep Convolutional Neural Network with a spatial attention mechanism for fire and traffic accident scene classification. Our spatial attention model learns to highlight the most discriminative convolutional features, which is related to the regions of interest in the input image. We train our network in a weakly supervised way. In other words, without the requirement of precise bounding box annotating the exact location of fire or traffic accidents in the image, our network can be learned from the only image-level label. In addition to the image-based traffic scene classification, the model is also applied on a set of collected videos for real-world applications. The proposed model, a simple end-to-end architecture, achieves promising performance on fire scene classification from images, and traffic accident scene classification from both images and videos.

17. Key Words Transportation of Hazardous Materials, Convolutional Neural Network, Spatial Attention, Weakly Supervised, Traffic Accidents	**18. Distribution Statement** No restrictions.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 22	**22. Price**

Reproduction of completed page authorized

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Mid-America Transportation Center (MATC)
Intelligent Systems Center (ISC)
Convolutional Neural Network (CNN)
Support Vector Machine (SVM)
Radial Basis Function (RBF)
Histogram of Oriented Gradient (HOG)
Artificial Neural Network (ANN)
Long Short-Term Memory (LSTM)

Acknowledgments

This work was supported by Mid-America Transportation Center (MATC), and the matching fund from the Intelligent Systems Center (ISC) at Missouri University of Science and Technology (Missouri S&T).

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Abstract

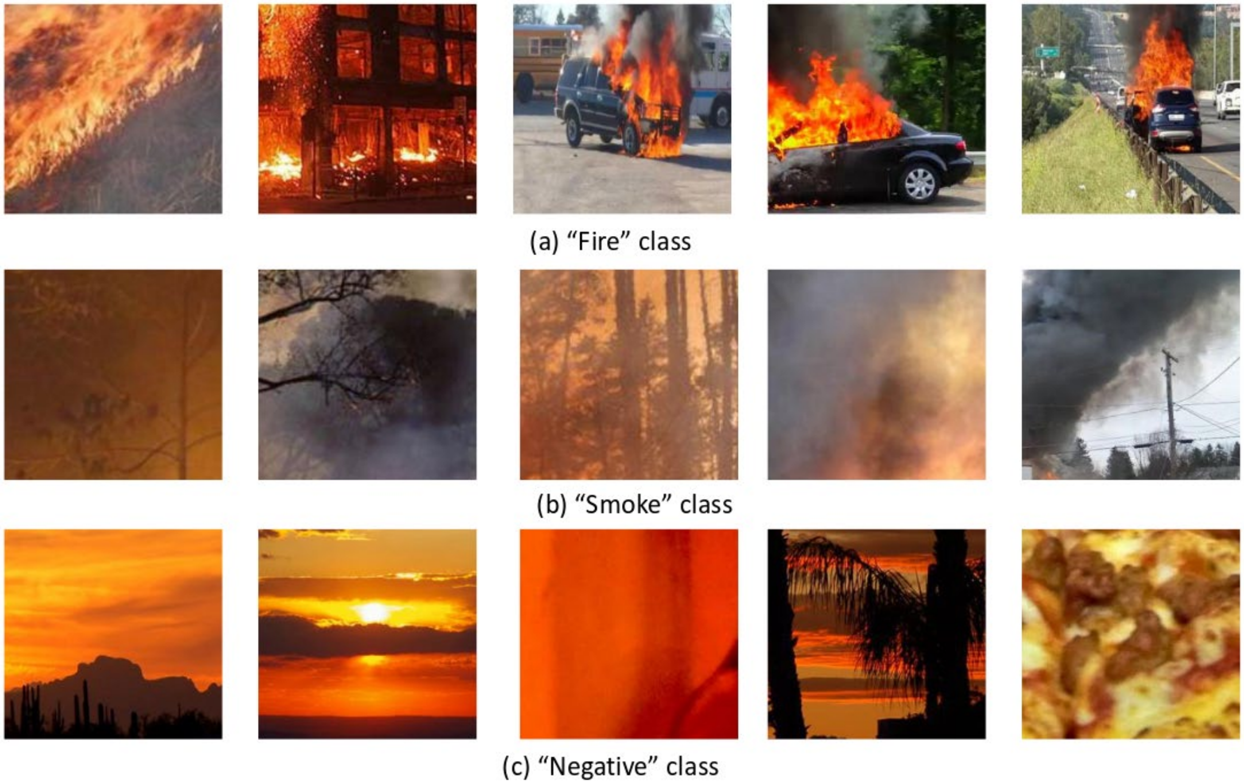
During the past ten years, on average there were nearly 16.5 thousands of hazardous materials (hazmat) transport incidents per year resulting in $82 millions of damages. Prompt, accurate, objective assessment on hazmat incidents is important for the first-responders to take appropriate actions timely, which will reduce the damage of hazmat incidents and protect the safety of people and the environment. Therefore, one of the most important steps is to automatically detect transport incidents, such as fire and traffic accidents. In this work, we introduce a simple and yet effective framework that integrates the convolutional feature maps of deep Convolutional Neural Network with a spatial attention mechanism for fire and traffic accident scene classification. Our spatial attention model learns to highlight the most discriminative convolutional features, which is related to the regions of interest in the input image. We train our network in a weakly supervised way. In other words, without the requirement of precise bounding box annotating the exact location of fire or traffic accidents in the image, our network can be learned from the only image-level label. In addition to the image-based traffic scene classification, the model is also applied on a set of collected videos for real-world applications. The proposed model, a simple end-to-end architecture, achieves promising performance on fire scene classification from images, and traffic accident scene classification from both images and videos.

Chapter 1 Spatial Attention Mechanism for Weakly Supervised Fire and Traffic Accident Scene Classification

1.1 Introduction

A substantial amount of hazardous materials (hazmat), such as flammable liquids and poisonous gases, need to be transported to locations of consumption or disposal. During the past ten years, on average there were nearly 16.5 thousands of hazmat transportation incidents per year resulting in $82 millions of damage [1]. When a transportation incident occurs (e.g., fire, traffic car accident), prompt and effective emergency response is critical to minimize the impact of the incident. For example, fire caused by hazmat accidents contains hazardous materials and has a dangerous influence on the environment, human health, and other valuable properties. Image-based fire detection (e.g., traffic surveillance cameras) are effective in large open areas. However, there are challenges in designing an automatic image classification algorithm to tell if an image contains a fire or not. Figure 1.1 shows some samples of the fire image dataset collected by us, from which we can see that some non-fire images have an appearance similar to the fire images. In this work, we utilize the deep Convolutional Neural Network to classify whether an image contains a fire or not.

In addition to fire detection in images, we also explore the general traffic accident scene classification in images and videos, as traffic accidents can cause serious injuries, which also require rapid assistance to reduce the additional rescue minute. Some samples of the traffic accident image dataset collected by us are shown in figure 1.2, and some samples of the traffic accident video dataset collected by us are shown in figure 1.3. The traffic accident images and videos were acquired by both traffic surveillance cameras and cameras on vehicles.



(a) “Fire” class

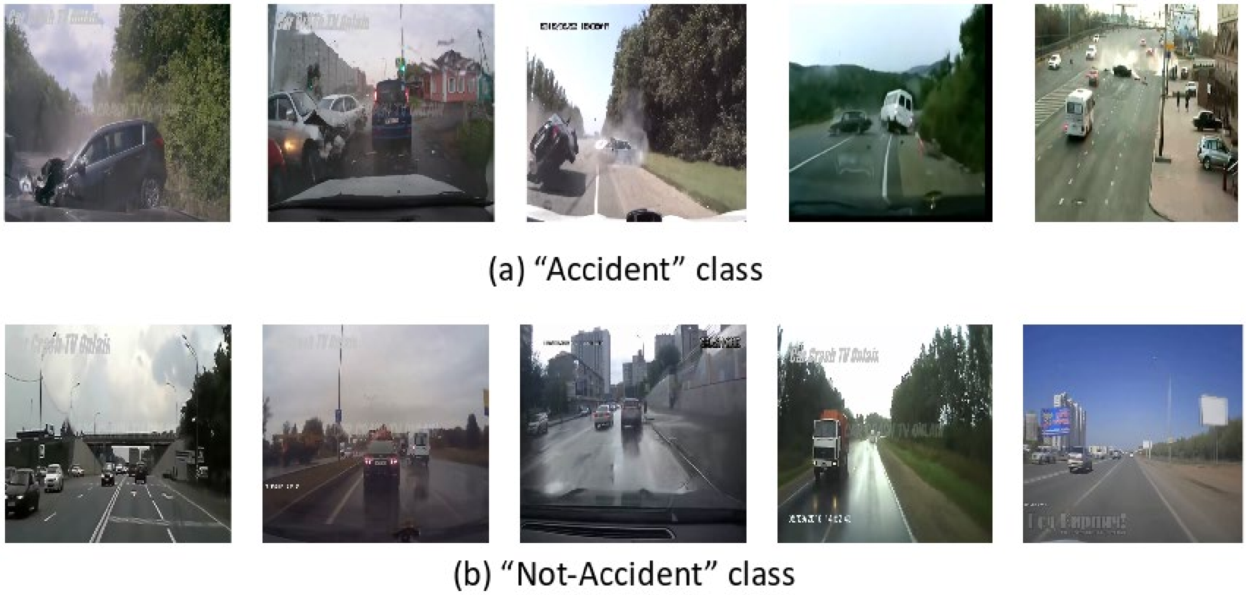
(b) “Smoke” class

(c) “Negative” class

Figure 1.1 Sample images from our fire dataset. The appearance of “smoke” and “negative” samples is very similar to the “fire” samples.

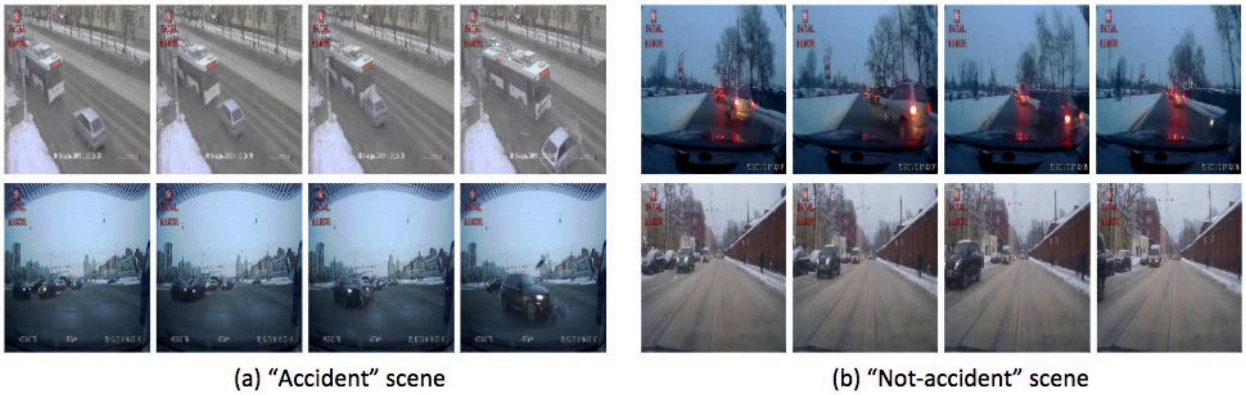
1.1.1 Deep Convolutional Neural Network

With the recent availability of powerful GPUs, effective optimization algorithms, and a large amount of human-annotated image data [13], Convolutional Neural Networks (CNN) [15,16,17,18,19,20] have achieved significant progress for the task of image classification. CNNs have the ability to learn meaningful feature representations from the large quantities of data for a wide range of tasks. In addition to image classification, CNNs pre-trained on ImageNet [13] contribute greatly in object detection [24,25,26], video classification [28], semantic segmentation [27], and many other tasks.


(a) “Accident” class

(b) “Not-Accident” class

Figure 1.2 Sample images collected in our traffic accident image dataset


(a) “Accident” scene

(b) “Not-accident” scene

Figure 1.3 Frames of sample videos collected in our traffic accident video dataset

1.1.2 Spatial Attention

Despite recent advances, image classification using deep CNN still has challenging research questions to address. Most of the state-of-the-art methods [15,16,17,18,19,20] employ CNN over the entire image region to compute the feature maps by convolution followed by standard pooling (average or max) operation or a fully-connected layer for the classification, without highlighting the features extracted from the most relevant spatial regions. But usually an object in an image does not occupy the entire spatial domain. Some of the pixels in the entire spatial domain are less, or not relevant to the target class. Therefore, the motivated research question is: "*from the convolutional feature maps, which features should get more importance to highlight the most discriminative regions of the input image?"* To address this challenge, we leverage the spatial attention mechanism on top of the convolutional feature maps to emphasize the most significant features. In other words, the spatial attention mechanism learns to focus on the most relevant parts of the input image.

1.1.3 Weakly-Supervised Learning

Most of the modern deep learning algorithms [21,22,23] are fully supervised, which rely on human-labeled annotations, such as the precise bounding box and the segmentation mask for training. But, in practice, collecting such accurate annotations are expensive and time-consuming. Building a training dataset with only image-level annotation is much easier than the bounding box or segmentation mask annotations. Therefore, the motivated research question arises: "*given a weakly labeled image dataset (i.e., each image in the training set has a label but which portion of the image contains the target class is unknown), how can we effectively train a deep learning algorithm?*" To address this challenge in this work, we use weakly-supervised learning that reduces the amount of human level intervention by using the image or video dataset that are partially labeled (e.g., "fire", "accident", etc.). In other words, without ever providing the network with information about the location of the target class, we train our network by utilizing only image-level labels.

1.2 Our Contribution

First, we utilize deep Convolutional Neural Networks with a spatial attention mechanism for fire classification. We introduce an end-to-end spatial attention model for weakly supervised fire classification from images using pre-trained CNN networks. Our method starts by adopting existing VGG [16] networks (e.g., VGG-16, VGG-19) pre-trained on ImageNet [13] data with only image-level supervision (no bounding box or segmentation mask annotating the precise region of interest in the image) for feature extraction. The extracted features from CNNs are passed through our spatial attention model to get the attentionally-pooled feature representation, which is then processed by a classification layer for the final image-level classification. In addition to the classification, our approach can also locate the fire regions in the image by our spatial attention model.

Second, we generalize our learned spatial attention model from the fire classification dataset to traffic accident classification. For this purpose, we simply transfer the learned attention weights of our spatial attention model to traffic accident image and video datasets.

Chapter 2 Related Works

2.1 Fire Detection and Classification

There are several works for vision-based fire detection and classification [2,3,4]. Healey et al. [2] used a purely color-based model for automatic fire detection. Spectral, spatial, and temporal models of fire regions were developed for vision-based fire detection [3]. Temporal wavelet features in addition to ordinary motion and color cues were utilized to detect fire and flame [4]. There are a limited number of machine-learning based fire classification and detection approaches [5,6]. Ko et al. [5] introduced a vision sensor-based fire detection method, which used wavelet coefficients as the input to the support vector machines (SVM) classifier with a radial basis function (RBF) for the fire-pixel verification. Zhang et al. [6] used temporal shape features as an input to the artificial neural networks (ANN) for real-time forest fire detection. Our approach differs from them, as we use deep convolutional features with an attention model for fire classification and localization.

2.2 Traffic Accident Classification

There are not many literature studies on traffic accident data for classification based on deep learning methods. Ess et al. [7] presented a segmentation-based method to recognize traffic scenes in front of moving vehicles with respect to the road topology and the existence of commonly encountered objects. Geiger et al. [8] proposed a probabilistic generative model for multi-object traffic scene understanding. Gupte et al. [9] presented a computer vision-based algorithm for detecting and classifying vehicles in monocular image sequences of the traffic scene. Lan et al. [10] used the histogram of oriented gradient (HOG) and support vector machine (SVM) for real-time automatic obstacle detection in urban traffic. Shiau et al. [11] developed a forecasting model based on data mining technology for road traffic accident classification.

Agarwal et al. [12] presented a hybrid model based on logistic regression with a wavelet-based feature extraction process for traffic incident detection.

2.3 Deep Learning for Image Classification

Deep Convolutional Neural Networks with the advance of architectures have become popular in large scale image classification. ImageNet [13] challenge played an important role to develop architectures from high-dimensional shallow SIFT features [14] to deep CNN [15]. Later a number of attempts [16,17,18,19,20] have been made to achieve better classification accuracy. VGGNet [16] steadily increased the depth of the network by adding more convolutional layers to design effective architectures. Residual connections [17] introduced the advantages of using additive merging of the signal with theoretical and practical evidence. The Inception CNN architecture was first introduced as GoogLeNet [18], which later came with different versions [19,20] by refining the architecture in various ways. These architectures are fully convolutional or fully connected and do not provide attention in the spatial domain.

2.4 Weakly-Supervised Learning

There are several works [24,25,26] that used weakly-supervised learning with CNN features in object detection and recognition. Oquab et al. [24] employed a pre-trained CNN to compute the mid-level feature representation for images of PASCAL VOC. Oquab et al. [25] also presented the weakly supervised learning with CNNs to localize object instances in images while predicting the label. Bilen et al. [26] introduced weakly supervised deep detection networks, which used pre-trained CNN features to recognize and detect the object without the requirement of bounding box annotations. In our work, we leverage the attention mechanism with deep pre-trained CNN to emphasize the most discriminative features for weakly supervised fire image classification as well as traffic accident classification from images and videos.

Chapter 3 Approach and Methodology

The workflow of our network is illustrated in figure 3.1. First, we obtain a pre-trained CNN network and extract the last convolutional feature maps by passing the image through the pre-trained CNN network. Second, we apply the spatial attention model on top of the last convolutional feature maps to get the attentionally-pooled feature vector. Third, we pass the attentionally-pooled feature vector through a classification module to get the final classification scores on the fire and the general traffic accident classification tasks.

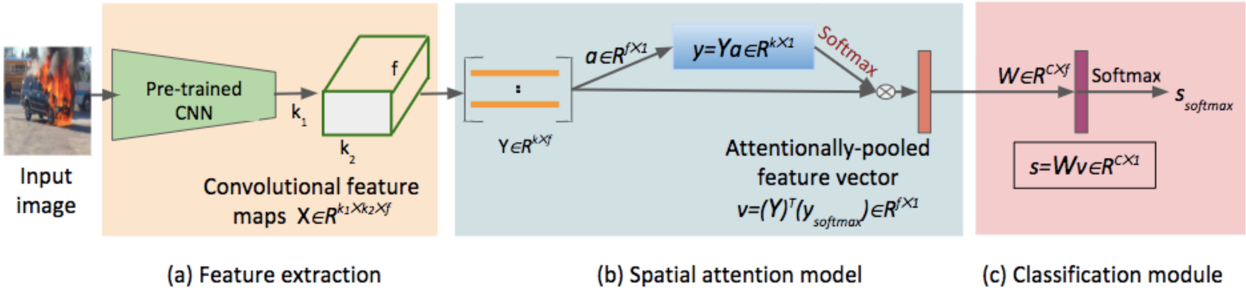

Input image
Pre-trained CNN
k_1
k_2
f
Convolutional feature maps $X \in R^{k_1 \times k_2 \times f}$
$Y \in R^{k \times f}$
$a \in R^{f \times 1}$
$y = Ya \in R^{k \times 1}$
Softmax
Attentionally-pooled feature vector $v = (Y)^T (y_{softmax}) \in R^{f \times 1}$
$W \in R^{C \times f}$
Softmax
$s_{softmax}$
$s = Wv \in R^{C \times 1}$
(a) Feature extraction
(b) Spatial attention model
(c) Classification module

Figure 3.1 The architecture of our approach. (a) Feature extraction, (b) Spatial attention network, and (c) Classification module.

3.1 Feature Extraction

An important component in our approach is feature extraction. In our approach, we choose two pre-trained networks, namely VGG-16 and VGG-19. These two pre-trained networks are similar, except there are more convolution and max-pooling layers in the VGG-19 network. We use the pre-trained 2D CNN models (VGG-16, VGG-19) trained on ImageNet dataset [13] with only image-level supervision to extract the last convolutional feature maps. The feature maps after the last convolutional layer preserve the spatial information of the input image, denoted as $X \in R^{k_1 \times k_2 \times f}$denotes the spatial dimension of the feature maps and f is the number of feature map, as shown in figure 3.1(a). These feature maps are utilized to describe the visual content of the input image and passed to the next layers for recognition.

3.2 Spatial Attention Mechanism

The feature maps $X \in R^{k_1 \times k_2 \times f}$ treat every element equally, but some pixels in the image are not related to the target class. Thus, we propose a spatial attention mechanism on top of *X* to gain more attention on those discriminative regions in an image. The proposed spatial attention mechanism is a trainable layer, which attentionally pools the most discriminative features.

For this purpose, the feature maps $X \in R^{k_1 \times k_2 \times f}$ are converted to 2D matrix $Y \in R^{k \times f}$, where $k = (k_1 \times k_2)$, as shown in figure 3.1(b). Each row of matrix *Y* maps to different overlapping regions in the input space. Our spatial attention model learns to focus its attention on these k regions. Formally, our spatial attention model learns an attention weight vector $\boldsymbol{a} \in R^{f \times 1}$ and computes attention score vector $\boldsymbol{y}$, which indicates the feature importance from k regions:

$$\boldsymbol{y} = Y\boldsymbol{a}, \qquad \text{where} \qquad \boldsymbol{y} \in R^{k \times 1} \tag{3.1}$$

The attention score vector $\boldsymbol{y}$ is passed through a softmax layer to get the normalized attention scores-

$$y_{softmax}^{i} = \frac{exp(y^{(i)})}{\sum_{j=1}^{k} exp(y^{(j)})}, \quad \text{where} \quad i = 1, \ldots\ldots, k \tag{3.2}$$

where, y^i denotes the *i^th^* dimension of $\boldsymbol{y}$ and $y_{softmax} = [0, 1]^k$ denotes the normalized attention scores. After that, the attentionally-pooled feature vector $\boldsymbol{v}$, which is the feature representation for the classification module, is computed by

$$\boldsymbol{v} = (Y)^T y_{softmax}, \quad \text{where} \quad \boldsymbol{v} \in R^{f \times 1} \tag{3.3}$$

3.3 Classification Module

So far, the spatial attention mechanism has computed the attentionally-pooled feature vector $\boldsymbol{v} \in R^{f\times 1}$, which represents the most discriminative features of an image. Now, we aim to classify the image into the predefined class categories based on the attentionally-pooled feature vector $\boldsymbol{v}$, as shown in figure 3.1(c). We learn linear mapping $W \in R^{C\times f}$(C is the number of classes) and compute the C-dimensional score vector $\boldsymbol{s}$ from the attentionally-pooled feature vector:

$$\boldsymbol{s} = W\boldsymbol{v}, \qquad \text{where} \qquad \boldsymbol{s} \in R^{C\times 1} \tag{3.4}$$

Finally, the score vector $\boldsymbol{s}$ is passed through the softmax layer to get the normalized classification scores:

$$s_{sotmax}^{(i)} = \frac{\exp(s^{(i)})}{\sum_{j=1}^{C}\exp(s^{(j)})}, \quad \text{where} \quad i = 1, \ldots\ldots\ldots, C \tag{3.5}$$

where, $s^{(i)}$ denotes the i^{th} dimension of $\boldsymbol{s}$ and $s_{sotmax} = [0, 1]^{C}$ denotes the normalized classification scores. In other words, $\boldsymbol{s}$ denotes the original classification scores of an image, which encodes the raw class activation and its response is able to reflect the degree of containing a specific class, while $s_{softmax}$ represents the softmax classification scores, which performs the normalization operation, turning its sum into 1.

3.4 Attention Model for Video Recognition

Classifying videos instead of images adds a temporal dimension in addition to the visual appearance in individual frames. Therefore, in addition to the spatial attention model, we use Long Short-Term Memory (LSTM) [30] for video recognition, which is able to address variant-length input and capture the long-term temporal dynamics.

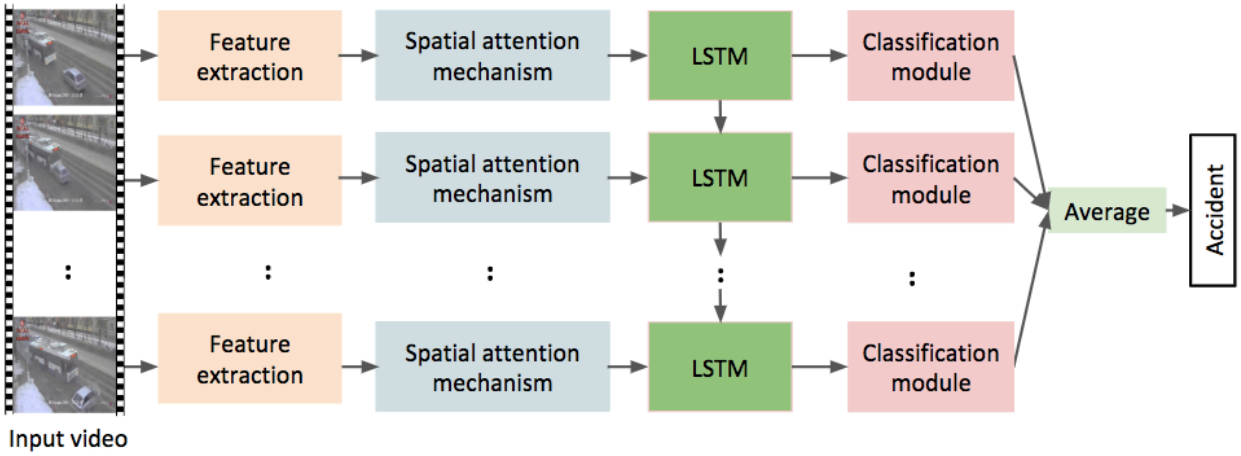


Figure 3.2 The overall process for video recognition

Since a video contains a sequence of frames, we extract the last convolutional feature maps obtained by pushing the video frames through the pre-trained network. Formally, given a video with the duration of *T* frames, at each time step *t*, we extract the last convolutional feature maps $X_t \in R^{k_1 \times k_2 \times f}$, which are passed through the spatial attention mechanism to get the attentionally -pooled feature vector $\boldsymbol{v_t} \in R^{f \times 1}$. The outputs of the spatial attention mechanism $\boldsymbol{v_t}$ are passed through the recurrent sequence learning module (Long Short-Term Memory (LSTM)). The weight parameters of LSTM map the input $\boldsymbol{v_t}$ and previous time step hidden state output to an output feature vector $\boldsymbol{o_t} \in R^{f \times 1}$., which is the feature representation for the classification module. The outputs of LSTM at each time step, are then fed into the classification module, which produces classification scores $s_{softmax}^{(t)}$ for each frame. Finally, the classification scores of each frame of a video are averaged to get the final video-level label prediction. The overall process for video recognition is shown in figure 3.2. It should be noted that we pass the outputs of LSTM to the classification module for video recognition, instead of directly passing the outputs of the spatial attention model to the classification module for image recognition.

Chapter 4 Experimental Results

4.1 Implementation Details

We use Keras (Tensorflow backend) python API to implement our network architecture. The input image is resized to $224 \times 224 \times 3$ pixels, which is passed through the pre-trained 2D CNN model (VGG-16, VGG-19). We extract the last convolutional layer of the VGG-16 or VGG-19 network, which produces $14 \times 14 \times 512$ feature maps. The feature maps are fed into our spatial attention model, which is a trainable layer. Our spatial attention model learns attention weights $\boldsymbol{a} \in R^{f \times 1}$ to get attentionally-pooled feature vector $\boldsymbol{v} \in R^{f \times 1}$, which is the feature representation for the classification module. The classification module learns a linear mapping $W \in R^{C \times f}$ to transform the feature representation $\boldsymbol{v}$ into a C-dimensional score vector. The loss is based on the standard cross-entropy loss between the ground truth and the prediction from our proposed model. The weights of spatial attention network, LSTM (for video recognition) and classification module are learned using Adam [29] optimizer with the minibatch size of 32 samples, where the optimization is stopped after 15 epochs.

4.2 Datasets

4.2.1 Fire Dataset

Our fire dataset contains 68979 images from 3 classes. These 3 classes are labeled as “fire”, “smoke”, and “negative”. The fire class has 21013 samples, the smoke class has 20818 samples and the negative class has 27148 samples. Evaluation is performed using the average classification accuracy.

4.2.2 Traffic Accident Image Dataset

Our traffic accident dataset consists of 1134 images collected from Google and frames of YouTube videos. The images are labeled with 2 classes: “accident” and “not-accident”, which have 570 and 564 samples, respectively. Evaluation is performed using the average classification accuracy.

4.2.3 Traffic Accident Video Dataset

Our traffic accident video dataset consists of 311 videos collected from YouTube. The videos are labeled with 2 classes: "accident" and "not-accident", which have 151 and 160 samples, respectively. Evaluation is performed using the average classification accuracy.

4.3 Quantitative Evaluation

4.3.1 Fire Classification

First, we evaluate the performance of our model on the problem of weakly-supervised image classification on the Fire dataset. We perform stratified random sampling on the Fire dataset four times and split the data 5/6 for training and 1/6 for testing for each class. Table 4.1 shows the comparison results of our approach with other existing support vector machine (SVM) [5] and artificial neural network (ANN) [6] based methods for the extracted features from the pre-trained VGG-16 network on the Fire dataset. As shown in table 4.1, our approach (86.19% average accuracy) outperforms the SVM (77.41% average accuracy) and ANN (82.87% average accuracy) based methods by a large margin.

Table 4.1 Comparison of our spatial attention model with other methods on the Fire dataset (accuracy) for VGG-16 features

Trial	SVM [5]	ANN [6]	Spatial attention model (Ours)
Trial-1	78.23	83.17	85.62
Trial-2	76.51	82.17	85.58
Trial-3	77.12	82.76	86.78
Trial-4	77.16	83.41	86.77
Average	77.41	82.87	86.19

Table 4.2 Comparison of our spatial attention model with other methods on the Fire dataset (accuracy) for VGG-19 features

Trial	SVM [5]	ANN [6]	Spatial attention model (Ours)
Trial-1	78.17	82.72	86.64
Trial-2	77.35	82.09	86.32
Trial-3	77.21	81.37	85.64
Trial-4	77.89	82.60	86.58
Average	77.66	82.19	86.29

Like VGG-16, we perform the same experiments with the extracted features from pre-trained VGG-19 network. The comparison results of our approach with other SVM and ANN based methods for the VGG-19 features on the Fire dataset is shown in table 4.2. As shown in table 4.2, our spatial attention model (86.29% average accuracy) achieves superior performance compared to SVM (77.66% average accuracy) and ANN (82.19% average accuracy) based methods.

To test the robustness of our approach, we use VGG-16 and VGG-19 pre-trained networks for our experiments. As shown in table 4.1 and table 4.2, we get consistent performance for both VGG-16 and VGG-19 networks. As VGG-19 network has a few additional convolutional and max-pooling layers compared to VGG-16 network, we get slightly better performance on average for VGG-19 network.

We also performed the ablation studies on our approach to see the accuracy on each individual class. Table 4.3 shows the performance of individual class accuracy for our spatial attention model with VGG-16, and VGG-19 features. As shown in table 4.3, the negative class classification accuracy is higher than fire and smoke classes, while most of the failure cases occur to classify the smoke class. Classifying the smoke class is hard, as sometimes fire itself creates smoke and some negative images also have the appearance similar to smoke images.

Table 4.3 Ablation study to see the performance of individual class accuracy for our spatial attention model for VGG-16 and VGG-19 features

Class	VGG-16 features	VGG-19 features
Fire	87.64	87.65
Smoke	76.57	82.45
Negative	90.95	89.09

4.3.2 Image-Based Accident Scene Classification

We transfer the spatial attention model that we learned from the fire dataset to the Traffic Accident Image dataset. Table 4.4 shows the comparison results of the transfer learning approach with the baseline approach for the extracted features from pre-trained VGG-19 network (here, we only use pre-trained VGG-19 network, as we get better performance for VGG-19 features compared to VGG-16 features on the fire dataset) on the Traffic Accident Image dataset. For the baseline approach, we configure the network without attention pipeline. For this purpose, the extracted feature maps $X \in R^{k_1 \times k_2 \times f}$ are pooled and averaged to get *f*-dimensional feature vector and passed through the classification module for classification.

Table 4.4 Comparison of transfer learning performance (accuracy) with baseline approach on Traffic Accident Image dataset for VGG-19 features

Trial	Baseline approach	Spatial attention model (Ours)
Trial-1	89.12	94.54
Trial-2	89.72	95.63
Trial-3	88.37	92.89
Trial-4	88.60	91.80
Average	88.95	93.72

As shown in table 4.4, our approach (93.72% average accuracy) outperforms the baseline approach (88.95% average accuracy) on Traffic Accident Image dataset, which means our spatial attention model learns generalized features that can be effectively used in traffic accident and hazardous materials incident classification.

4.3.3 Video-Based Accident Scene Classification

We evaluate the performance of our video recognition framework (spatial attention model + LSTM) on Traffic Accident Video dataset for real-time traffic accident scene classification. As shown in table 4.5, we performed the ablation studies on our framework by comparing three configurations on Traffic Accident Video dataset. Over all three configurations, the combination of spatial attention model and LSTM achieves the best performance.

Table 4.5 Ablation study of different architectures on Traffic Accident Video dataset for VGG-19 features

Trial	Baseline approach	Spatial attention model (Ours)	Spatial attention model + LSTM (Ours)
Trial-1	78.84	81.39	82.69
Trial-2	79.16	84.67	86.53
Trial-3	77.44	80.10	82.69
Trial-4	77.81	81.18	84.61
Average	78.31	81.84	84.13

The first configuration in the second column of table 4.5 shows the results of the baseline approach without any attention pipeline, which achieves 78.31% average accuracy. The third column in table 4.5 shows the performance of spatial attention model, which achieves better performance (81.84% average accuracy) compared to the baseline approach. The last configuration, which combines the spatial attention model and LSTM, achieves the best performance (84.13% average accuracy), which indicates that our spatial attention model with LSTM can be effectively used in real-world applications of traffic accident scene classification.

4.4 Qualitative Evaluation

We visualize our spatial attention maps on some randomly selected test samples of fire and traffic accident datasets. As shown in figure 4.1, our spatial attention model can correctly focus on the fire and accident regions in the image, without the requirement of the bounding box and segmentation mask annotations.

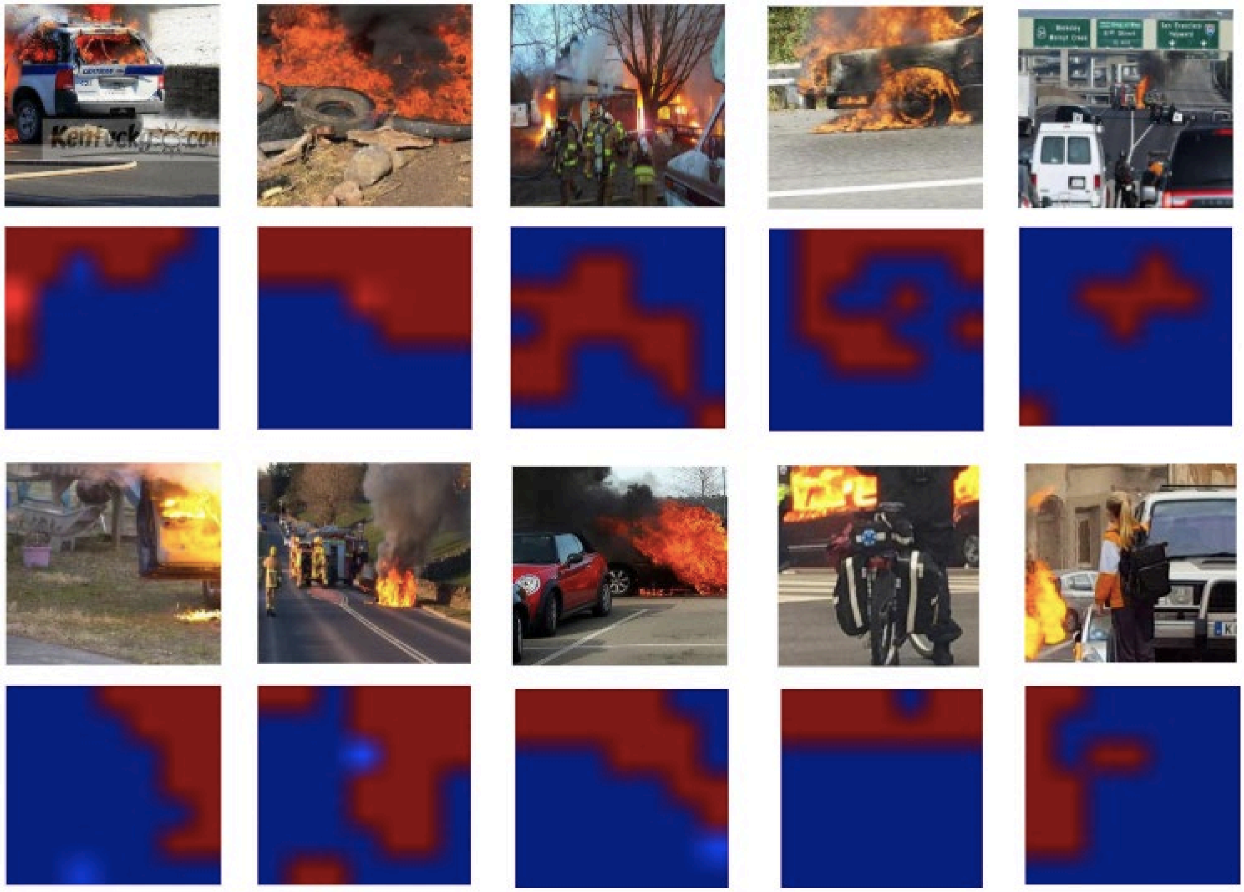


Figure 4.1 Visualization of our spatial attention map. Our spatial attention model learns to locate the fire and accident regions in the image.

Chapter 5 Conclusions

In this work, we introduce a new weakly-supervised framework for fire classification from images, and accident scene classification from both images and videos. We use pre-trained deep CNN features and employ a spatial attention mechanism to address the challenge of highlighting the most discriminative features for fire classification. To see the effectiveness of our approach, we also transfer the learned weights of our spatial attention model to a generalized traffic accident dataset for classification. We performed extensive experimental evaluation and showed that our model performs better than the baseline approach, which did not use any attention pipeline. The proposed framework is also efficient and easy to implement.

References

1. US Department of Transportation Pipeline and Hazardous Materials Safety Administration (PHMSA). *10 Year Incident Summary Reports.* (https://www.phmsa.dot.gov/hazmat/library/data-stats/incidents)

2. Healey G., Slater D., Lin T., Drda B., and Goedeke A. D. 1993. "A system for real-time fire detection." *In Computer Vision and Pattern Recognition.*

3. Liu C. B., and Ahuja N. 2014. "Vision based fire detection". *In Pattern Recognition.*

4. Töreyin B. U., Dedeoğlu Y., Güdükbay U., and Cetin A. E. 2006. "Computer vision-based method for real-time fire and flame detection." *Pattern recognition letters.*

5. Ko B. C., Cheong K. H., and Nam J. Y. 2009. "Fire detection based on vision sensor and support vector machines." *Fire Safety Journal.*

6. Zhang D., Han S., Zhao J., Zhang Z., Qu C., Ke Y., and Chen X. 2009. "Image based forest fire detection using dynamic characteristics with artificial neural networks." *In Artificial Intelligence.*

7. Ess A., Müller T., Grabner H., and Van Gool L. J. 2009. "Segmentation-based urban traffic scene understanding." *In British Machine Vision Conference.*

8. Geiger A., Lauer M., Wojek C., Stiller C., and Urtasun R. 2014. "3d traffic scene understanding from movable platforms." *In Pattern Analysis and Machine Intelligence.*

9. Gupte S., Masoud O., Martin R. F., and Papanikolopoulos N. P. 2002. "Detection and classification of vehicles." *IEEE Transactions on intelligent transportation systems.*

10. Lan J., Jiang Y., Fan G., Yu D., and Zhang, Q. 2016. "Real-time automatic obstacle detection method for traffic surveillance in urban traffic." *Journal of Signal Processing Systems.*

11. Shiau Y. R., Tsai C. H., Hung Y. H., and Kuo Y. T. 2015. "The application of data mining technology to build a forecasting model for classification of road traffic accidents." *Mathematical Problems in Engineering.*

12. Agarwal S., Kachroo P., and Regentova E. 2016. "A hybrid model using logistic regression and wavelet transformation to detect traffic incidents." *IATSS Research.*

13. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., ... and Berg A. C. 2015. "Imagenet large scale visual recognition challenge." *International Journal of Computer Vision.*

14. Perronnin F., Sánchez J., and Mensink T. 2010. "Improving the fisher kernel for large-scale image classification." *In European conference on computer vision.*

15. Krizhevsky A. Sutskever I., and Hinton G. E. 2012. "Imagenet classification with deep convolutional neural networks." *In Neural Information Processing systems.*

16. Simonyan K., and Zisserman A. 2014. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556.*

17. He K., Zhang X., Ren S., and Sun J. 2016. "Deep residual learning for image recognition." *In Computer Vision and Pattern Recognition.*

18. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., and Rabinovich A. 2015. "Going deeper with convolutions." *In Computer Vision and Pattern Recognition.*

19. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., and Wojna, Z. 2016. "Rethinking the inception architecture for computer vision." *In Computer Vision and Pattern Recognition.*

20. Szegedy C., Ioffe S., Vanhoucke V., and Alemi A. A. 2017. "Inception-v4, inception-resnet and the impact of residual connections on learning." *In Association for the Advancement of Artificial Intelligence.*

21. Girshick R. 2015. "Fast r-cnn." *In international conference on computer vision.*

22. Ren S., He K., and Girshick R. 2015. "Faster r-cnn: Towards real-time object detection with region proposal networks." *In Advances in neural information processing systems.*

23. He K., Gkioxari G., Dollár P., and Girshick R. 2017. "Mask r-cnn." *In international conference on computer vision.*

24. Oquab M., Bottou L., Laptev I., and Sivic J. 2014. "Learning and transferring mid-level image representations using convolutional neural networks." *In Computer Vision and Pattern Recognition.*

25. Oquab M., Bottou L., Laptev I., and Sivic J. 2015. "Is object localization for free? weakly-supervised learning with convolutional neural networks." *In Computer Vision and Pattern Recognition.*

26. Bilen H., and Vedaldi A. 2016. "Weakly supervised deep detection networks." *In Computer Vision and Pattern Recognition.*

27. Hariharan B., Arbeláez P., Girshick R., and Malik J. 2014. “Simultaneous detection and segmentation.” *In European Conference on Computer Vision.*

28. Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., and Fei-Fei, L. 2014. “Large-scale video classification with convolutional neural networks.” *In Computer Vision and Pattern Recognition.*

29. Kingma D. P., and Ba, J. 2014. “Adam: A method for stochastic optimization.” *arXiv preprint arXiv:1412.6980.*

30. Hochreiter S., and Schmidhuber, J. 1997. “Long short-term memory.” *Neural computation.*