AICS 2020 
Journal of Physics: Conference Series 1631 (2020) 012054 doi:10.1088/1742-6596/1631/1/012054

DB-Net: Dual Attention Network with Bilinear Pooling for Fire-Smoke Image Classification

Chen You\textsuperscript{1}, Zhen Li\textsuperscript{2}, Meng Li\textsuperscript{1}, Zongliang Gao\textsuperscript{1} and Wei Li\textsuperscript{3}

\textsuperscript{1}School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China
\textsuperscript{2}The Shenzhen Commercial Cipher Industry Association, Shenzhen, China
\textsuperscript{3}School of Information Science and Technology, North China University of Technology, Beijing, China
Email: shouhuniyc@163.com

Abstract. In recent years, with the help of deep learning, image classification has been significantly improved. More and more researchers are seeking fire alarm solutions based on these artificial intelligence algorithms. However, this raises new challenges in an effective method of classification for fire-smoke images. In this paper, we propose a Dual Attention Network with Bilinear Pooling of three steps to accomplish the task of fire-smoke classification. We first extract features of fire pictures or smoke ones by using ResNet-50 as a basic net. Then we attach two kinds of attention modules, channel attention and spatial attention, which are called dual attention, to extract key information ‘what’ and ‘where’ from pictures. Finally, we merge them by using the bilinear pooling module which has been shown to be effective at improving our classification rate. Results show that our most accurate model can reach 90.11% per-image accuracy, which is improved by 4.81% compared to the traditional ResNet-50.

1. Introduction

As is known to us all, the scenes of fire can be staged all over the world at any time, which can cause serious damage. With the development of technology, smart devices like video surveillance have made it easier to monitor fire areas. However, the recognition and classification of pictures of fire and smoke can be a tough job, requiring a lot of time. With the help of deep learning, effective classification of fire-smoke becomes more feasible.

Compared with simple object classification, fire-smoke classification is more challenging, because sometimes, it may be misjudged by people smoking, fog, or similar colors of fire-smoke to the environment. In addition, due to the movement of smart devices, many pictures may become blurred, which makes classification more difficult.

As usual, image classification can be performed with selected standards like shape, color, and texture features [1]. In some circumstances, people make an attempt to improve the result of image classification to further analyze where fire happens or the location extremely prone to fire, where smoke generates at first [2]. However, it is clear that a technique suitable for one set of images may not be suitable for another due to different conditions.

Recently, image classification based on deep learning has been widely used. For example, a residual learning framework is proposed in Ref. [3], which aims to simplify the training of networks that are deeper than previously used ones. Szegedy et al. present a deep convolutional neural network
architecture code named Inception, which achieves the new state-of-the-art for detection and classification [4]. Xie et al. focus on feature pooling based on a model called Task-Driven Pooling (TDP), which aims to achieve a better representation, combining classification tasks with representation learning [5]. Some researchers introduce bilinear pooling architecture to model local parts of objects, which have gotten promising results [6-8].

In this paper, we propose a Dual Attention Network with Bilinear Pooling (DB-Net) to accomplish the task of fire-smoke classification in three steps. First, we extract features of fire pictures or smoke ones by using ResNet-50 as a frame net. Then, we attach two kinds of attention modules, channel attention and spatial attention for deeper information, which is called dual attention. Last, we improve the classification rate by using the bilinear pooling module, which has been shown to be effective at image classification. Our experimental results show that our most accurate model can reach 90.11% per-image accuracy, which is improved by 4.81% compared to the traditional ResNet-50.

The rest of the paper is organized as follows. Section 2 we introduce the proposed method in details. Experimental results are presented in Section 3. Some concluding remarks are provided in Section 4.

2. Proposed Method

In general, we introduced a three-step method to improve the accuracy of classification of pictures for fire or smoke. The first step uses the ResNet-50 model for classification. The second step uses channel attention module to solve what happened in the picture (fire or smoke) and spatial attention module to explain where most likely to happen. The third step uses the bilinear pooling model to improve the classification accuracy. The overall frame structure is shown in figure 1.

![Figure 1. Framework of the proposed method.](image)

Fire or smoke scenes encountered in our daily lives may contain many complex information, with a lot of background noise. Therefore, in order to improve the accuracy of image classification, we first use the ResNet-50 network to initially classify the data input to the model. Then key information of the classified data is extracted through two different kinds of modules: channel attention and spatial attention. Finally, we use bilinear pooling to obtain final results. We use the deep transfer learning method in this paper, combining with the characteristics of Resnet network, attention module, and bilinear pooling, which achieves classification of fire-smoke images effectively.

2.1. ResNet-50 Network Model

The structure of the ResNet-50 network is shown in figure 2. The network system mainly includes five stages from the input to the end, as is shown in the figure. ResNet-50 has 2 basic blocks which is shown in figure 3. One of them is identity block, where the input and output have the same dimensions.
and can be connected in series. The other is convolution block, where the input and output dimensions are different, which cannot be connected in series. It plays a role in changing the dimension of the feature vector originally.

![ResNet-50 Network model](image1)

**Figure 2. ResNet-50 Network model.**

In order to simplify the training of networks that are deeper than previously used ones, the Google team propose a residual learning framework, whose layers are restructured as the reference layer inputs to learn the residual functions instead of learning the unreferenced functions. This foolproof addition does not add extra parameters and calculation to the network, but it can greatly increase the training of the model. When the layer of the model deepens, this simple structure can deal with the problem of degradation well.

![Structure of two basic blocks](image2)

**Figure 3. Structure of two basic blocks.**

The three-layer residual structure has the same amount of layers, so it can reduce the number of parameters, which is extended to a deeper model. Therefore, a 50-layer ResNet was proposed, which not only did not have degradation problems, but also greatly reduced the error rate, while the computational complexity was also kept at a very low level.

### 2.2. Channel Attention

By using the relationship of features between inter-channels, a module called channel attention was produced. Since each channel of the feature map in the channel attention can be regarded as a feature detector, which focuses on “what” in an input image.
Given an input feature map \( f \in \mathbb{R}^{C \times H \times W} \), we infer a channel attention map \( m_c \in \mathbb{R}^{C \times 1 \times 1} \). The channel attention process can be summarized as:

\[
f_c = m_c(f) \ast f
\]

where \( \ast \) represents element-wise multiplication. During multiplication, channel attention values are broadcasted along the spatial dimension.

### 2.3. Spatial Attention
The spatial attention focuses on `where` in an input image, which is different from and complementary to the channel one.

Given an input feature map \( f \in \mathbb{R}^{C \times H \times W} \), we compute the channel attention map \( m_c \in \mathbb{R}^{C \times 1 \times 1} \). Similarly, the spatial attention process \( m_s \in \mathbb{R}^{1 \times H \times W} \) can be summarized as:

\[
f_s = m_s(f_c) \ast f_c
\]

where \( \ast \) represents element-wise multiplication. In the meantime, during multiplication, spatial attention values are broadcasted along the channel dimension.

### 2.4. Bilinear Pooling
In order to combine the feature maps \( f_c \) in equation (1) and \( f_s \) in equation (2) that we obtained effectively, we use a simple and effective architecture for classification named bilinear pooling, which is shown in figure 4, we refer to it as BP module.

![Figure 4. The Bilinear pooling module.](image)

As you can see, a BP module consists of a quadruple \( Q = (f_c, f_s, p, c) \), where \( f_c \) is a feature map that we obtained through channel attention module in equation (1), \( f_s \) is a feature map that we obtained through spatial attention module in (2), \( p \) is a function for pooling and what’s more, \( c \) is a function which can be applied for classification of different images. Given a feature map \( f : L \times I \rightarrow \mathbb{R}^{C \times D} \), we can take an input with an image \( i \in I \) which has a location \( l \in L \) and \( C \times D \) is the size of outputs with feature maps.

By exploiting the matrix outer product, the outputs with feature maps and the combination of bilinear pooling with \( f_c \) and \( f_s \) at a location \( l \) can be generally given by

\[
BP(l, i, f_c, f_s) = [f_c(l, i)]^T f_s(l, i)
\]

### 3. Experiment Validation
The main equipment and operating environment for our model training are as follows: Intel i7-7700 processor, 256 GB solid state drive, 8G memory, Nvidia GTX1050Ti graphics card, Ubuntu16.04 and
TensorFlow framework. In addition, the dataset we used contains 1200 pictures for each scenario (fire/smoke/neutral).

Due to the limited accuracy of traditional ResNet network for image classification, we produce a DB-Net to accomplish the purpose of classification accuracy enhancement. The data is divided into three groups, namely the training group, the verification group and the test set. The training set is used for training the convolutional neural network model and obtain the corresponding model. The validation set achieves the best results by training the validation model and fine-tuning the parameters. The test set is composed of pictures with complex background information that we collected and are difficult to classify. Images in this set are randomly selected from the 3 kinds of scenes data sets, fire or smoke or neutral ones, which can better simulate the actual application scenarios and test the classification effect of the model.

3.1. Architecture of the Whole Model

Our model is derived from deep transfer learning of the ResNet-50 network. First we retain all the parameters of the ResNet-50 pre-training model. Then, during the training process, we replace the last fully-connected layer in ResNet-50 to improve the efficiency of classification.

Figure 5 shows the architecture of the whole model. We set the data as input to the network, and then a 2048-dimensional feature vector is generated, followed by channel attention (as is shown in figure 6) and spatial attention (as is shown in figure 7) using the feature vector. Then we merge these two feature maps output from dual attention modules through bilinear pooling. Finally the SoftMax classifier is trained, and a model that can realize classification is obtained.

![Figure 5](image)

**Figure 5.** The architecture of the whole model.

![Figure 6](image)

**Figure 6.** Channel attention.
3.2. Experiment Results

3.2.1. Various Preprocessing. We first use the traditional ResNet-50 network architecture to process the data set. Then based on it, firstly we add the channel attention and the spatial attention modules (refer to it as add). Secondly, we contact the channel and the spatial attention modules (refer to it as contact). Finally, we use bilinear pooling to merge the channel and the spatial attention modules (refer to it as BP).

Then let learning rate = 0.01 and then use the processed images to evaluate the performance of image detection. At this stage, we take several statistics: TP as True Positive, TN as True Negative, FP as False Positive, and FN as False negative. Substitute calculation for evaluation, the main calculation rules are as follows:

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]
\[
\text{Precision} = \frac{TP}{(TP + FP)}
\]
\[
\text{Recall} = \frac{TP}{(TP + FN)}
\]
\[
\text{Misjudgment rate} = \frac{(FP + FN)}{(TP + TN + FP + FN)}
\]

Table 1 is the result of performance evaluation of various models. The experiment shows that compared to traditional ResNet-50 model or the other models based on it (just add dual attention or contact them), our DB-Net is more reliable and the classification error rate is only 8.55%.

| Model     | Accuracy | Precision | Recall | Misclassification Rate |
|-----------|----------|-----------|--------|------------------------|
| Resnet-50 | 87.20%   | 87.27%    | 88.09% | 12.80%                 |
| Add       | 88.32%   | 88.38%    | 89.23% | 11.68%                 |
| Contact   | 89.96%   | 90.01%    | 90.86% | 10.04%                 |
| BP(DB-Net)| 91.45%   | 91.51%    | 92.34% | 8.55%                  |

3.2.2. Fire-Smoke Classification. We use the fire/smoke/neutral images in the data set for training, verification, and testing according to different proportions. Figure 8 shows the accuracy of the experimental results of classification with different models.

Experiment results show that our model with better performance, the classification rate of fire-smoke images with traditional ResNet-50 is about 85.37%. By using DB-Net, the classification rate can be effectively improved. Compared with traditional ResNet-50, by using DB-Net, the classification accuracy is increased by 4.80% on average, and the highest reaches 8.12%.
4. Conclusion
The preprocessing method based on ResNet-50 network used in this paper is practical for fire-smoke classification. It can be used together with traditional CNN models that are used for the recognition of other images. Experiments show that with DB-Net, the accuracy of image classification can be improved dramatically.

Acknowledgement
This work is funded by General Research Project of Zhejiang Provincial Department of Education grant number [Y201942662].

References
[1] Li Y, Zhang Y and Zhu Z 2019 Learning deep networks under noisy labels for remote sensing image scene classification IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium (Yokohama, Japan) pp 3025-3028.
[2] Treebupachatsakul T and Poomrittigul S 2019 Bacteria classification using image processing and deep learning 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) (Jeju, South Korea) pp 1-3.
[3] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV) pp 770-778.
[4] Szegedy C, et al. 2015 Going deeper with convolutions 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston, MA) pp 1-9.
[5] Xie G, Zhang X, Shu X, Yan S and Liu C 2015 Task-driven feature pooling for image classification 2015 IEEE International Conference on Computer Vision (ICCV) (Santiago) pp 1179-1187.
[6] Cai S, Zuo W and Zhang L 2017 Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization 2017 IEEE International Conference on Computer Vision (ICCV) (Venice) pp 511-520.
[7] Gao Y, Beijbom O, Zhang N and Darrell T 2016 Compact bilinear pooling 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV) pp 317-326.
[8] Kong S and Fowlkes C 2017 Low-rank bilinear pooling for fine-grained classification 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI) pp 7025-7034.