IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 24, 2019, accepted August 13, 2019, date of publication August 21, 2019, date of current version September 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936616

Deep Attention Neural Network for Multi-Label Classification in Unmanned Aerial Vehicle Imagery

AALIYAH ALSHEHRI, (Student Member, IEEE), YAKOUB BAZI, (Senior Member, IEEE), NASSIM AMMOUR, (Member, IEEE), HAIDAR ALMUBARAK, (Member, IEEE), AND NAIF ALAJLAN, (Senior Member, IEEE)

Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Yakoub Bazi (ybazi@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University through the Local Research Group Program under Project RG-1435-055.

ABSTRACT The multi-label classification problem in Unmanned Aerial Vehicle (UAV) images is particularly challenging compared to single-label classification due to its combinatorial nature. To tackle this issue, we propose in this paper a deep learning approach based on encoder-decoder neural network architecture with channel and spatial attention mechanisms. Specifically, the encoder module which is based on a pre-trained convolutional neural network (CNN) has the task to transform the input image to a set of feature maps using an opportune feature combination. To improve the feature representation further, this module incorporates a squeeze excitation (SE) layer for modelling the interdependencies between the channels of the feature maps. The decoder module which is based on a long short terms memory (LSTM) network has the task of generating, in a sequential way, the classes present in the image. At each time step, it predicts the next class-label by aligning its hidden state to the corresponding region in the image by means of an adaptive spatial attention mechanism. The experiments carried out on two UAV datasets with a spatial resolution of 2-cm show that our method is promising in predicting the labels present in the image while attending the relevant objects in the image. Additionally, it is able to provide better classification results compared to state-of-the-art methods.

INDEX TERMS UAV imagery, deep learning, attention neural network, multi-label image classification.

I. INTRODUCTION

The increase adoption of unmanned aerial vehicles (UAVs), commonly known as drones have proven their effectiveness in collecting images with extremely high spatial details over inaccessible areas and limited coverage zones due to their small size and fast deployment. The availability of this type of imagery has opened the door for several methodological developments such as classification, object detection and more recently multi-label classification [1], [2].

Multi-label image classification aims to assign multiple class labels from predefined a set of objects. It has a wide range of applications, such as visual object recognition [3]–[5], image content annotation [6], [7], and content-based image retrieval [8]–[10]. The multi-label classification task is particularly challenging compared to the single-label classification due to its combinatorial nature.

The general literature of computer vision convoys several approaches to solve the multi-label classification problem. The existing methods can be divided into transformation or adaptation methods. In the first group, the baseline method is to decompose a multi-label task into a set of binary classification problems [11]. The idea of the binary classification is to independently learn one binary classifier for each label [12]. This method becomes costly when the number of classes is high. Other methods cast the task of multi-label classification into a multiclass problem, where each multiclass problem represents one or more labels [13]. Other work uses label ranking techniques which ranks the related labels before the unrelated ones [14]. The second group, the adaptation

The associate editor coordinating the review of this article and approving it for publication was Tony Thomas.

methods, includes adaptive boosting [16], and lazy learning [17]. For more details, we refer the reader to Huang *et al*. work [15].

Recently, deep learning strategies have been introduced as a promising solution to improve further the representation aspect [18]. For instance, Gong *et al.* [19] proposed a convolutional neural network (CNN) with a similar structure to AlexNet [20] and then uses various multi-label loss functions for training the network. In particular, they generalize the standard softmax loss function widely used for single-label classification to handle to multi-label classification scenarios. Wei *et al.* [21] introduced a novel Hypotheses-CNN-Pooling (HCP) where a set number of object segment hypotheses are taken as the inputs. Then a shared CNN is connected with each hypothesis, and finally the CNN output resulting from different hypotheses are fused with max pooling to generate the multi-label predictions. In another work, the authors propose to better exploit the correlation information between labels [22] by maximizing the score of labels present in the image over the absent ones based on a predefined margin in addition to the correlation between the extracted features and their corresponding labels through a learned semantic space. Zhu *et al.* [24], introduced a spatial regularization network (SRN) based on attention maps to capture both semantic and spatial relations of the multiple labels present in the image. The SRN generates attention maps for all labels and captures the underlying relations between them via learnable convolutions.

Furthermore, recurrent neural networks (RNNs) were also used to discover the correlation between labels in a multi-classification problem. The authors in [23] combined RNNs with CNNs to learn a joint image-label embedding to characterize the semantic label dependency as well as the image-label relevance. The CNN is employed as the image representation, while the recurrent layer captures the information of the previously predicted labels. Then the output label probability is computed according to the image representation and the output of the recurrent layer. Wang *et al.* [25] proposed a spatial transformer layer with a long short-terms memory (LSTM) network. The spatial transformer layer aims to locate regions from the convolutional feature maps in a region-proposal-free way while the LSTM network predicts the semantic labeling scores.

In the context of remote sensing imagery, few contributions have been reported to solve the problem of multi-label classification compared to the general literature of computer vision. For instance, the authors in [1] proposed to exploit the spatial contextual information besides label cross-correlation between adjacent tiles in UAV images through a multi-label conditional random field (CRF) method. In a first step, the UAV image is subdivided into a grid of tiles, which are then, processed using a bag of word (BOW) model followed by an encoder network for generating the feature representation. Then the output of this module is fed to another neural network for providing the tile-wise multi-label prediction probabilities. In the second phase, the multi-label CRF model is applied by integrating the spatial correlation between adjacent tiles and the correlation between labels within the same tile to improve iteratively the multi-label classification map. In another work, Zeggada *et al.* [2] proposed to combine radial basis function neural networks (RBFNN) and a customized thresholding layer for label detection. For such purpose, the authors use the outputs of the RBFNN as indicators of presence/absence of the corresponding object. During the prediction phase, the thresholding layer is used for deciding on the presence/absence of classes instead of an intuitive decision mechanism which uses the simple rule ''the object is present if the output is greater than 0.5, otherwise it is absent.''

While these methods provide an interesting set of solutions to the problem; they are mainly based on the combination of several tools and are not trainable in an end-to-end manner. In this paper, we propose an alternative solution based on encoder-decoder neural network architecture with channel and spatial attention mechanisms. Specifically, the encoder module based on a CNN has the task to transform the UAV image to a set feature maps. To improve further the feature representation by modelling the interdependencies between the channels, this module incorporates a squeeze excitation (SE) layer. The decoder module based on LSTM network has the task of predicting in a sequential way the classes present in the image. At each time step, it predicts the next class-label by aligning its hidden state to the corresponding region in the image by utilizing a spatial attention mechanism. The main contributions of this paper can be summarized as follows:

1) Propose an end-to-end deep learning method for UAV image multi-labeling based on CNN-LSTM networks;
2) Incorporate channel and spatial attention mechanisms to improve the feature representation, and the detection of the regions corresponding to the classes present in the image;
3) Validate the method on two UAV datasets with a spatial resolution of 2-cm acquired over the cities of Trento and Civezzano (Italy) in 2011 and 2012, respectively.

The paper is organized into five sections. In Section 2, we review the inception-v3 network used as a pre-trained CNN in addition to the LSTM network. In Section 3, we describe the proposed method in detail. In Section 4, we present the experimental results. Finally, we provide concluding remarks and directions for future developments in Section 5.

II. BACKGROUND

A. INCEPTION-V3 NETWORK

The inception-v3 network [26] was proposed by a research team from Google. This network is composed of 42 layers and includes three types of inception modules composed of convolutions filters with sizes in the range of 5×5 to 1×1. The main architecture of this network is given in Figure 1. The inception modules aim to reduce the number of parameters owing to the factorization of larger convolution layers into smaller layers. They use convolution filters of

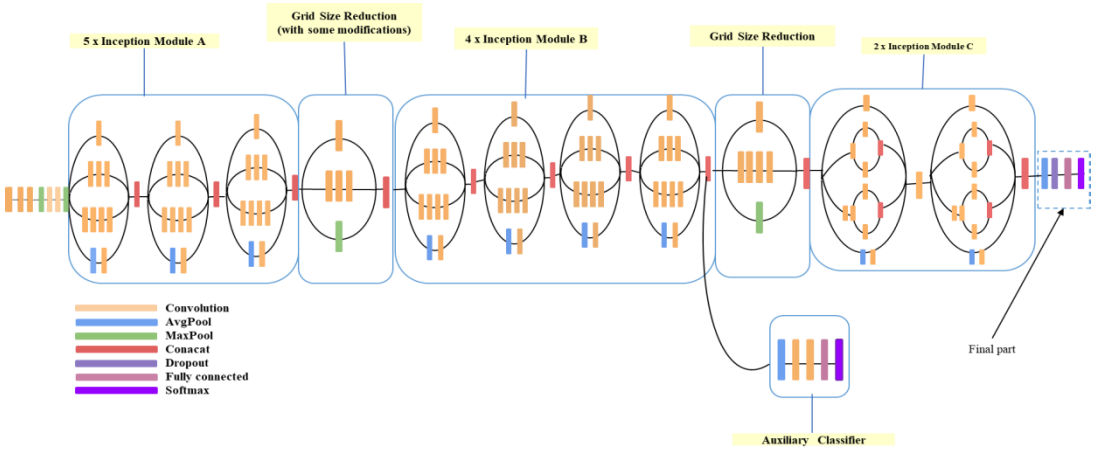

FIGURE 1. Inception-v3. Architecture [26].

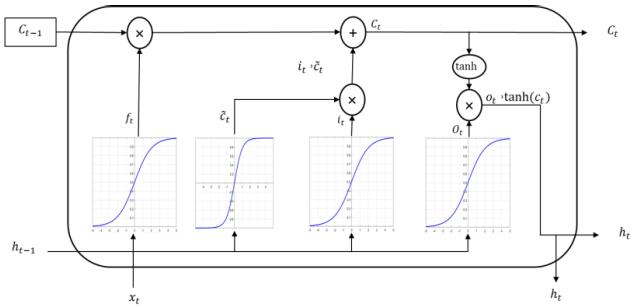

FIGURE 2. LSTM structure.

sizes 1×1 to reduce the number of input channels and then perform different parallel operations, which are then concatenated at the output. The 5×5 convolution filters of the original inception module are replaced by stacking two 3×3 convolutions with fewer parameters. As the network goes deeper, it uses high dimensional representations using two inception modules of type C referred as $2 \times$ Inception Module C (Figure 1). This network includes more improvements in the architecture compared to the original GoogLeNet network (inception-v1) which was the winner of the ILSVRC14 (ImageNet Large Scale Visual Recognition Competition). These improvements include; 1) the RMSProp optimizer, 2) Factorized 7×7 convolutions, 3) BatchNormalization in the auxiliary classifiers, and 4) Label Smoothing, which is a type of a regularizing component added to the loss formula that prevents the network from becoming too confident about a class and prevents overfitting.

B. LSTM NETWORK

The LSTM [28] is a special type of the traditional recurrent neural networks (RNN), characterized by its capability of learning over long-term dependencies. As shown in Figure 2, the LSTM has four types of gates at time step t in memory cell. These are the input gate $\boldsymbol{i}_t$, the update gate $\boldsymbol{c}_t$, the output gate $\boldsymbol{o}_t$, and the forget gate $\boldsymbol{f}_t$. At each time step, the gates receives as input the previous LSTM hidden state $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{y}_t$. The cell memory updates itself recursively based on the interaction of its previous values with the forget and update gates'.

The main working mechanism of the LSTM network is given as follows:

$$\boldsymbol{i}_t = sigmoid\left(\boldsymbol{W}_i.\left[\boldsymbol{h}_{t-1}, \boldsymbol{y}_t\right]\right) \quad (1)$$

$$\boldsymbol{f}_t = sigmoid\left(\boldsymbol{W}_f.\left[\boldsymbol{h}_{t-1}, \boldsymbol{y}_t\right]\right) \quad (2)$$

$$\tilde{\boldsymbol{c}}_t = tanh\left(\boldsymbol{W}_g\left[\boldsymbol{h}_{t-1}, \boldsymbol{y}_t\right]\right) \quad (3)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{c}}_t \quad (4)$$

$$\boldsymbol{o}_t = sigmoid\left(\boldsymbol{W}_o\left[\boldsymbol{h}_{t-1}, \boldsymbol{y}_t\right]\right) \quad (5)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * tanh\left(\boldsymbol{c}_t\right) \quad (6)$$

where $*$ denotes the Hadamard product and $\boldsymbol{W}_i$, $\boldsymbol{W}_f$, $\boldsymbol{W}_g$, and $\boldsymbol{W}_o$ are learnable weights. For simplicity, we can model the hidden state $\boldsymbol{h}_t$ as follows:

$$\boldsymbol{h}_t = LSTM\left(h_{t-1}, \boldsymbol{y}_t, \boldsymbol{r}_{t-1}\right) \quad (7)$$

where $\boldsymbol{r}_{t-1}$ is the memory cell vector at time $t-1$.

III. PROPOSED METHOD

Let us consider $D = \left\{\boldsymbol{X}_i, \boldsymbol{y}_i\right\}_{i=1}^{n}$ as a multi-label UAV dataset with n training images. Each training image X_i is associated with its corresponding ground-truth class labels $y_i = (\mathrm{y}_{1i}, \mathrm{y}_{2i}, \ldots, \mathrm{y}_{Ti})$. In a multi-label setting, the image $\boldsymbol{X}_i$ can be assigned to more than one object based on its content. Our goal is to learn a set of weights $\boldsymbol{W}$ for the encoder-decoder architecture depicted in Figure 3 that allows inferring in a sequential manner the objects contained in a test UAV image unseen during the training phase. Detailed descriptions of the method are provided in next sub-sections.

A. ENCODER MODULE WITH CHANNEL ATTENTION

In the image encoding module, we use the inception-v3 network described previously as a backbone. In particular, we fuse the outputs of the intermediate inception layers Mixed4 ($\boldsymbol{V}_4$) and Mixed7 ($\boldsymbol{V}_7$) of dimensions $28 \times 28 \times 288$ and $14 \times 14 \times 768$, respectively as shown in Figure 3-b.

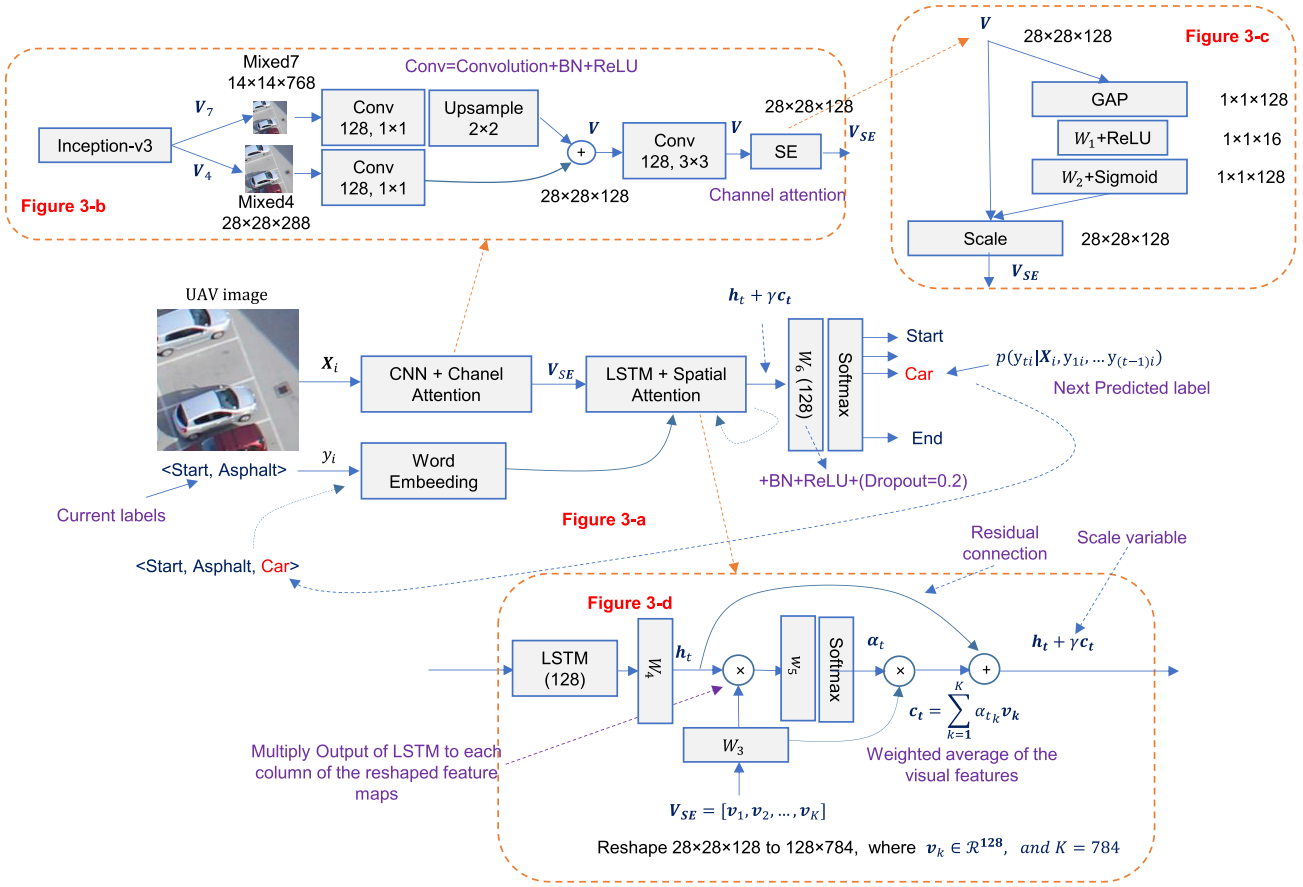

FIGURE 3. Proposed multi-labeling method with channel and spatial attention mechanisms (figure 3-a): The UAV image is fed to an encoder module based on a pre-trained CNN (inception_v3) for feature generation. The features are obtained by fusing the outputs of Mixed7 and Mixed4 using an opportune fusion mechanism (figure 3-b) followed by Squeeze excitation layer (SE) shown in figure 3-c for feature calibration. The resulting feature maps are fed to a decoder module composed on an LSTM with spatial attention (figure 3.d) for predicting in a sequential way the labels while attending the corresponding regions in the image. The complete architecture is learnable end-to-end with the backpropagation algorithm.

This choice is motivated by the fact that the Mixed4 layer has a better spatial details compared to the Mixed7 layer, but on the other side it is less discriminative. Thus the combination of both outputs aims to attend different regions in the image while using discriminative features. The combination is done by convolving both outputs with two filters of dimensions (1,1,128). Then the low spatial feature maps are up-sampled with a factor 2 and added to the feature maps with high spatial dimensions yielding features maps $\boldsymbol{V}$ of dimensions equal to (28,28,128).

To attend informative features and suppress less useful ones, we apply a squeeze (SE) layer to the resulting activation maps $\boldsymbol{V}$ for feature calibration [27]. The aim is to improve further the feature representation by modelling the interdependencies between the channels of the feature maps. As shown in Figure 3b. The feature maps are then squeezed through a global average pooling operation (GAP) to yield features of dimension (1, 1, 128), which are then fed to fully connected layer followed by ReLU activation function for dimensionality reduction with a reduction ratio equal to 8 (Figure 3-c). Then, we use another fully connected layer of dimension 128 to recover the original dimensions followed by a sigmoid activation function. The resulting feature vector $\boldsymbol{s}$ is used to modulate the channels of $\boldsymbol{V}$ through a simple channel-wise scaling operation. In brief, the SE layer operates as follows:

$$\boldsymbol{s} = Sigmoid\left(\boldsymbol{W}_2\left(ReLU\left(\boldsymbol{W}_1\left(\boldsymbol{V}\right)\right)\right)\right) \tag{8}$$

$$\boldsymbol{V}_{SE} = \boldsymbol{s} \odot \boldsymbol{V} \tag{9}$$

where $\boldsymbol{s}$ is the scaling vector and $\odot$ is the channel-wise multiplication operation. Thus, the outputs of this module are feature activation maps $\boldsymbol{V}_{SE} \in \mathcal{R}^{d \times w \times h}$, where $d = 128$ is the number of the activation maps, while and $w = h = 28$ represents their width and height. For computation convenience, we set $\boldsymbol{V_{SE}} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K] \in \mathcal{R}^{d \times K}$, where each feature vector $v_k \in \mathcal{R}^d$ and $K = w \times h$ is the total number of feature vectors.

B. DECODER MODULE WITH SPATIAL ATTENTION

This module (Figure 3-d) has of aligning the current hidden state of the LSTM to attend the corresponding region in the image by means of a spatial attention mechanism. Basically, the image features $\boldsymbol{V}_{SE}$ obtained from the encoder module

are combined with the hidden state $\boldsymbol{h}_t \in \mathcal{R}^d$ of the LSTM network through a Softmax layer to generate the attention weights:

$$\boldsymbol{\alpha}_t = Softmax\left(\boldsymbol{w}_5^T tanh(\boldsymbol{W}_3 \boldsymbol{V}_{SE} \odot \boldsymbol{W}_4 \boldsymbol{h}_t\right) \tag{10}$$

where $\boldsymbol{W}_3 \in \mathcal{R}^{K\times d}$, $\boldsymbol{W}_4 \in \mathcal{R}^{K\times d}$ and $\boldsymbol{w}_5^T \in \mathcal{R}^K$ are learnable weights, and $\boldsymbol{\alpha}_t \in \mathcal{R}^K$ is the attention weight vector with $\sum_{k=1}^{K} \boldsymbol{\alpha}_{tk} = 1$.

Based on the attention distribution, the context vector c_t can be obtained by:

$$\boldsymbol{c}_t = \sum\nolimits_{k=1}^{K} \alpha_{tk} \boldsymbol{v}_k \tag{11}$$

where $\boldsymbol{c}_t$ and $\boldsymbol{h}_t$ are combined to predict the probability for the next label through a classification module composed of fully connected layer followed with a softmax classification layer as shown in Figure 3-a. The probability output of the current label conditioned by the previous predicted labels is given as follows:

$$p(\mathrm{y}_{ti}|X_i, \mathrm{y}_{1i}, \ldots \mathrm{y}_{(t-1)i}) = Softmax\left(\boldsymbol{W}_s\left(\boldsymbol{h}_t + \gamma \boldsymbol{c}_t\right)\right) \tag{12}$$

where $\gamma \in \mathcal{R}$ and $\boldsymbol{W}_s \in \mathcal{R}^{T\times d}$ are learnable parameters. The scaling variable γ allows controlling the contribution of the context vector. It is worth recalling that this spatial attention layer is similar to the one proposed in [29] mainly for image captioning but with some modifications. In particular, we multiply the outputs of the LSTM to each column of the reshaped feature maps instead of an addition (equation 10) and use a scaling parameters γ to weight the contribution of the context vector $\boldsymbol{c}_t$ (equation 12). In the experiments, we found that these modifications lead to a better alignment of the labels to their corresponding regions in the UAV image.

To learn the set of weights of the proposed encoder-decoder model $\boldsymbol{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \boldsymbol{W}_4, \boldsymbol{w}_5, \boldsymbol{W}_s, \gamma\}$, we propose to minimize the negative of the log-likelihood function:

$$L(D, \boldsymbol{W}) = -\frac{1}{n}\sum\nolimits_{i=1}^{n}\sum\nolimits_{t=1}^{T} log\left(p\left(y_{ti}|\boldsymbol{X}_i; y_{1i}, \ldots y_{(t-1)i}; \boldsymbol{W}\right)\right) \tag{13}$$

To optimize the cost function $L(D, \boldsymbol{W})$, we use the RMSProp optimization method, which is one of the most popular adaptive gradient algorithms introduced by Hinton to speed up the training deep neural networks. The RMSProp divides the gradient by a running average of its recent magnitude.

$$\mathrm{E}[\mathrm{g}^2]_\mathrm{t} = \beta \mathrm{E}[\mathrm{g}^2]_{\mathrm{t}-1} + (1-\beta)\left(\frac{\partial L}{\partial W}\right)^2 \tag{14}$$

$$\mathrm{W}_\mathrm{t} = \mathrm{W}_{\mathrm{t}-1} - \alpha\left(\frac{\partial L}{\partial W}\right)\frac{1}{\sqrt{\mathrm{E}[\mathrm{g}^2]_\mathrm{t}}} \tag{15}$$

where $\mathrm{E}[\mathrm{g}^2]_\mathrm{t}$ is the moving average of squared gradients at iteration t, $\frac{\partial L}{\partial W}$ is the gradients of the loss function with respect to the weights $\boldsymbol{W}$, while α is the learning rate and β is the moving average parameter. In the experiments, we set the parameter β to its default values ($\beta = 0.9$), while for the learning parameter α, we set it initially to 0.001 and decrease by a factor of 1/10 after 20 epochs.

FIGURE 4. UAV used for the acquisition of the images.

IV. EXPERIMENTS

A. DATASET DESCRIPTION

In the experiments, we evaluated the proposed attention network on two UAV datasets acquired over the faculty of science of the University of Trento (Italy) and near the city of Civezzano (Italy) on October 2011 and 2012 by means of a UAV equipped with imaging sensors spanning the visible range (Figure 4). All acquisitions are made using a Canon EOS 550D camera characterized by a CMOS APS-C sensor with 18 megapixels. Both datasets contain UAV images of dimension 256 × 256 × 3 pixels with a spatial resolution of approximately 2 cm. For the Trento dataset, 1000 images are used for training and 3000 images for testing. The dataset contains 13 classes named as: {'Asphalt', 'Grass', 'Tree', 'Vineyard', 'Pedestrian Crossing', 'Person', 'Car', 'Roof1', 'Roof2', 'Solar Panel', 'Building Façade', 'Soil', and 'Shadow'}. The Civezzano dataset, on the other hand, contains 1000 training images and 3105 testing images and it has 14 classes named as {'Asphalt', 'Grass', 'Tree', 'Vineyard', 'Low Vegetation', 'Car', 'Roof1', 'Roof2', 'Roof3, 'Solar Panel', 'Building Facade',' Soil', 'Gravel', and 'Rocks'}. Figure 5, shows sample images from these two datasets with the corresponding classes.

B. PARAMETER SETTING

To learn the set of weights $\boldsymbol{W}$ of the network composed of a total of 605845 learnable parameters including the convolution layers, the channel and spatial attention layers and the LSTM parameters, we used the RMSprop optimization method with a mini-batch size of 50 images. We set the parameter β to its default values ($\beta = 0.9$), while for the learning parameter α, we set it initially to 0.001 and decreased by a factor of 1/10 after 20 epochs.

For performance evaluation, we present the results in terms of *sensitivity* $\left(Se = \frac{TP}{TP+FN}\right)$, *specificity* $\left(Sp = \frac{TN}{TN+FP}\right)$, and *average accuracy* $\left(Ac = \frac{Se+Sp}{2}\right)$, where TP, FN, TN refer to the true positives, false negatives and true negatives, respectively. We use also the *Hamming loss* (HL) for measuring the fraction of incorrectly predicted labels; the *label ranking*

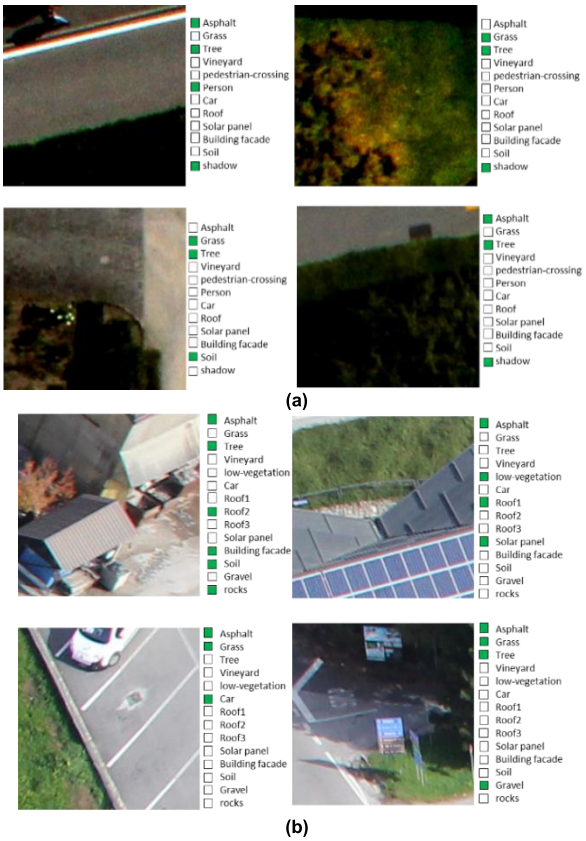

FIGURE 5. (a) Trento and (b) Civezzano datasets (green color indicates the labels associated with each image).

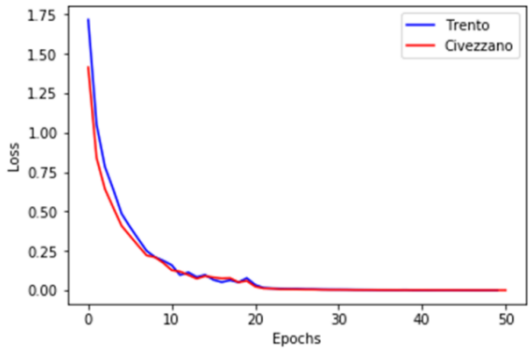

FIGURE 6. Loss versus the number of epochs obtained during the training phase for Trento (blue color) and Civezzano (red color).

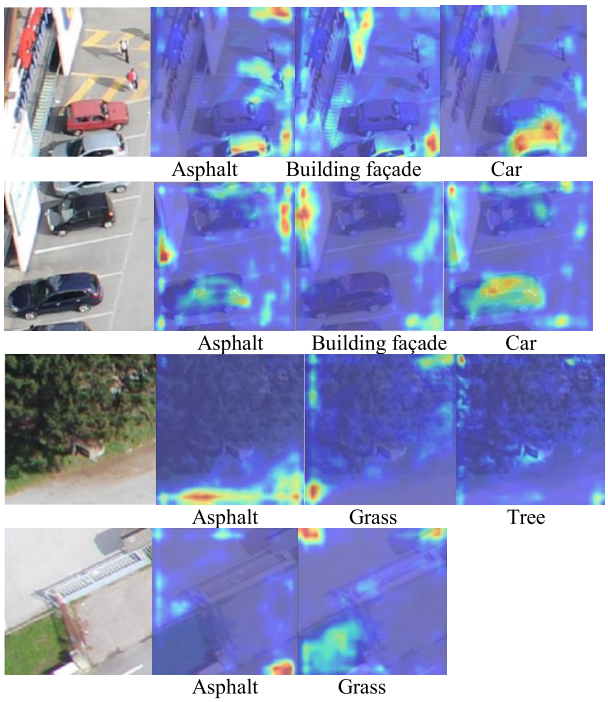

FIGURE 7. The class activation maps generated by the attention mechanism during the sequential prediction of the class labels detected in the image.

loss (RL) for computing the average number of label pairs that are incorrectly ordered; the *Precision score* $\left(Pr = \frac{TP}{TP+FP}\right)$; which is the proportion of labels correctly classified of the predicted positive labels; and the *mean average precision* (*mAP*), which refers to the average fraction of relevant labels ranked higher than the irrelevant ones.

We run all experiments on an HP Omen Station with the following characteristics: Central processing Unit (CPU)-Intel core (TM) i9-7920× CPU @ 2.9GHz with a RAM of 32 GB and an NVIDIA GeForce GTX 1080 Ti Graphical processing Unit (GPU) (with 11 GB GDDR5X memory). All code was implemented using Keras with TensorFlow backend, which is an open-source deep neural network library written in python.

C. RESULTS

Figure 6 shows the evolution of the loss function during the training phase of the proposed encoder-decoder network with channel and spatial attention mechanisms for both datasets. As can be seen, the training process converges after 30 iterations. In Table 1, we report the classification results obtained on the test sets. In terms of (Acc, and mAP) the network yields (83.59%, and 54.60%) and (86.93%, and 62.76%) for Trento and Civezzano, respectively. In Figure 7, we show the class activation maps obtained by the attention module during the generation of the classes present in the image. Although, no prior information is given about the class location during the training phase, the network exhibits an interesting behavior in providing different attention weights to each class in the image.

To analyze further the effect of the attention layers and the fusion trick of the activation maps proposed in Figure 3-b, we repeat the above experiments but with three different configurations of the network. In the first scenario, we remove the spatial attention layer; while in the second and third scenarios we apply the spatial attention layer to the activation maps $\boldsymbol{V}_4$ and $\boldsymbol{V}_7$ independently. For the first scenario without spatial attention, we obtain an (Acc, and mAP) of (79.00%, and 47.46%) and (82.00%, and 53.03%) for Trento and Civezzano, respectively. The application of the spatial attention to the activation maps $\boldsymbol{V}_4$ yields the worst results as these feature maps are less discriminative. On the other side, the utilization

TABLE 1. Classification results obtained for: (A) Trento, and (B) Civezzano datasets. ↑ (↓) Indicates that the larger (smaller) the value, the better the performance.

(A)

	Se↑	Sp↑	Acc↑	Pr↑	mAP↑	HL↓	RL↓
Without spatial attention	64.19	93.81	79.00	65.36	47.46	10.74	33.62
Spatial Attention V_4	62.43	93.08	77.75	62.14	44.58	11.63	35.96
Spatial Attention V_7	69.76	94.42	82.09	**69.47**	53.12	9.37	28.38
Proposed	**73.17**	94.01	**83.59**	68.98	**54.60**	**9.19**	**25.33**

(B)

	Se↑	Sp↑	Acc↑	Pr↑	mAP↑	HL↓	RL↓
Without spatial attention	69.40	94.60	82.00	69.72	53.03	09.22	28.93
Spatial Attention V_4	70.19	92.25	82.72	72.60	55.49	08.54	27.89
Spatial Attention V_7	73.80	95.57	84.68	74.89	59.25	07.73	24.85
Proposed	**78.30**	95.56	**86.93**	**75.94**	**62.76**	**07.06**	**21.18**

TABLE 2. Comparasion against state-of-the-art methods: (A) Trento, and (B) Civezzano datasets. ↑ (↓) Indicates that the larger (smaller) the value, the better the performance.

(A)

	Se↑	Sp↑	Acc↑	Pr↑	mAP↑	HL↓	RL↓
CNN-RBFNN [2]	68.60	92.60	80.60	----	----	----	----
CNN-MaxMargin loss [22]	61.32	95.44	78.38	71.02	49.50	09.80	35.72
CNN-Softmax loss [19]	69.49	92.04	80.77	61.37	47.34	11.42	29.89
LSTM-CNN [23]	64.19	93.81	79.00	65.36	47.46	10.74	33.62
LSTM-CNN+Spatial Attention [29]	66.49	93.86	80.17	66.43	49.27	10.30	31.69
Proposed	73.17	94.01	83.59	68.98	54.60	9.19	25.33

(B)

	Se↑	Sp↑	Acc↑	Pr↑	mAP↑	HL↓	RL↓
CNN-RBFNN [2]	68.60	92.60	80.60	----	----	----	----
CNN-MaxMargin loss [22]	70.97	95.70	83.33	74.71	57.44	08.53	26.51
CNN-Softmax loss [19]	70.37	95.22	82.79	72.52	55.53	08.54	27.13
LSTM-CNN [23]	69.40	94.60	82.00	69.72	53.03	09.22	28.93
LSTM-CNN Spatial Attention [29]	69.51	96.78	83.14	79.45	59.86	07.35	27.11
Proposed	**78.30**	**95.56**	**86.93**	**75.94**	**62.76**	**07.06**	**21.18**

of V_7 shows better improvements as it yields an (Acc, and mAP) of (82.09%, and 53.12%) and (84.68%, and 59.25%) for Trento and Civezzano, respectively. Yet, the combination of V_4 and V_7 besides the application of the channel and spatial attention layer provides better results confirming the effectiveness of the proposed network.

In Table 2, we compare our method to several state-of-the-art methods. The CNN-RBFNN method proposed in [2], which uses a customized thresholding layer for detecting the class labels. The CNN-MaxMargin method [22], which uses a max-margin loss to maximize the score of positive labels versus the scores of negative labels not present in the image. The CNN-softmax method [19], which uses a modified softmax loss function suitable for multi-label classification. Additionally, we compare our results against the standard CNN-LSTM method [23] in addition to another CNN-LSTM method [29] based on a spatial attention mechanism. The results reported in Table 2 confirm clearly the effectiveness of the proposed method. For instance, for Trento dataset, our method yields an accuracy of 83.59% and a mAP of 54.60%. For Civezzano dataset, it yields an accuracy of 86.93% and a mAP of 62.76%. The closest method based on spatial attention provides an accuracy of 80.17% and mAP of 49.27% for Trento and 83.14% and 59.86% for Civezzano.

V. CONCLUSION

In this paper, we have proposed an encoder-decoder network for UAV multi-labeling. This network incorporates a channel attention mechanism to model the interdependencies between the channels of the feature maps. It uses also a spatial attention layer to identify the regions corresponding to the labels available in the image. The experimental results obtained on two UAV datasets confirm the promising capability of the proposed method compared to state-of-the-art solutions. For future developments, we propose to use other pretrained-CNN models, and enhance the decoder layer by adding additional loss function that allows exploiting the image-to-label correlation and label-to-label correlation in a better way.

REFERENCES

[1] A. Zeggada, S. Benbraika, F. Melgani, and Z. Mokhtari, "Multilabel conditional random field classification for UAV images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 399–403, Mar. 2018.

[2] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

[3] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015.

[4] Y. Wu, J. Li, Y. Kong, and Y. Fu, "Deep convolutional neural network with independent softmax for large scale face recognition," in *Proc. ACM Multimedia Conf. (MM)*, Amsterdam, The Netherlands, 2016, pp. 1063–1067.

[5] Z. Ding, M. Shao, and Y. Fu, "Deep robust encoder through locality preserving low-rank dictionary," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 567–582.

[6] Q. Wang, B. Shen, S. Wang, L. Li, and L. Si, "Binary codes embedding for fast image tagging with incomplete labels," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2014, pp. 425–439.

[7] A. Tariq and H. Foroosh, "Feature-independent context estimation for automatic image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1958–1965.

[8] C. Wang, Z. Li, and C. Zhang, "A multi-instance multi-label learning framework of image retrieval," in *Intelligent Information Processing VII*. Berlin, Germany: Springer, 2014, pp. 239–248.
[9] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1556–1564.
[10] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2469–2479, Jun. 2016.
[11] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 191–202, Apr. 2018.
[12] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.
[13] G. Tsoumakas, I. Katakis, and L. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, Jul. 2011.
[14] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *J. Mach. Learn.*, vol. 73, no. 2, pp. 133–153, Nov. 2008.
[15] S.-J. Huang, W. Gao, and Z.-H. Zhou, "Fast multi-instance multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[16] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, Dec. 1999.
[17] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.
[18] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Columbus, OH, USA, vol. 2014, pp. 806–813.
[19] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," Dec. 2013, *arXiv:1312.4894*. [Online]. Available: https://arxiv.org/abs/1312.4894
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2012, pp. 1097–1105.
[21] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: Single-label to multi-label," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.
[22] W. Shi, Y. Gong, X. Tao, and N. Zheng, "Training DCNN by combining max-margin, max-correlation objectives, and correntropy loss for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2896–2908, Jul. 2018.
[23] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2285–2294.
[24] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5513–5522.
[25] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 464–472.
[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.
[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 375–383.

AALIYAH ALSHEHRI received the B.E. degree in networks and communications systems from Princess Nourah bint Abdulrahman University, Saudi Arabia, in 2014. She is currently pursuing the M.E. degree in computer science with King Saud University, Saudi Arabia. Her research interests include image processing and machine learning with applications to remote sensing image analysis.

YAKOUB BAZI (S'05–M'07–SM'10) received the State Engineer and M.Sc. degrees in electronics from the University of Batna, Batna, Algeria, in 1994 and 2000, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, in 2005. From 2000 to 2002, he was a Lecturer with the University of M'sila, M'sila, Algeria. From January to June 2006, he was a Postdoctoral Researcher with the University of Trento. From August 2006 to September 2009, he was an Assistant Professor with the College of Engineering, Al Jouf University, Al Jouf, Saudi Arabia. He is currently an Associate Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. His current research interests include remote sensing, signal/image medical analysis, and computer vision. He is also a Referee for several international journals. He is also an Associate Editor of the IEEE Geoscience and Remote Sensing Letters.

NASSIM AMMOUR (M'18) received the B.S. degree in electronics and the M.S. degree in robotics from Saad Dahlab University, Blida, Algeria, in 1988 and 1995, respectively, and the Ph.D. degree in computer vision and robotics from the École Supérieure Polytechnique, Algiers, Algeria, in 2009. He is currently an Associate Professor with the Advanced Lab for Intelligent Systems' Research (ALISR), College of Computer and Information Sciences, KSU, Saudi Arabia. His research interests include computer vision, pattern recognition, machine intelligence, robotics, and biometrics.

HAIDAR ALMUBARAK (M'03) was born in Al Ahsa, Saudi Arabia, in 1983. He received the B.S. degree in computer engineering from the King Fahd University of Petroleum and Minerals, Saudi Arabia, in 2005, the M.S degree in computer, information and network security from DePaul University, Chicago IL, USA, in 2011, and the Ph.D. degree in computer engineering from the Missouri University of Science and Technology, Rolla, MO, USA, in 2018. In 2005, he joined SAAD Hospital as a Computer Engineer, and in 2008, he moved to SABIC as a Systems Engineer before moving to USA to finish his M.S. and Ph.D. degrees. In 2019, he joined King Saud University as a Postdoctoral Fellow. His current research interests include computer vision, applied machine/deep learning, and medical image analysis.

NAIF ALAJLAN (M'11–SM'13) received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from King Saud University, Riyadh, Saudi Arabia, in 1998 and 2003, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2006. He was a Systems and Control Engineer with Saudi Basic Industries Company (SABIC), Riyadh, from 1998 to 2000. He then joined the Electrical Engineering Department, King Saud University, where he was a Lecturer, from 2000 to 2003, and an Assistant Professor, from 2007 to 2010. He is currently a Professor with the Computer Engineering Department, King Saud University. He is also the Founder and the Director of the Advanced Lab for Intelligent Systems Research, King Saud University. He authored or coauthored more than 50 journals articles (some with high impact factors) and 25 conference papers. His current research interests include shape retrieval, machine learning, pattern recognition, and remote sensing.