7th Iranian Joint Congress on Fuzzy and Intelligent Systems
18th conference on Fuzzy Systems and 16th conference on Intelligent Systems
29-31 Jan 2019
University of Bojnord

Hybrid Deep Learning Approach for Multi-label Image Classification

Reza Mohammadi Moqaddam[*†], Department of computer engineering, Bu-Ali Sina University, Hamedan, Iran

Hassan Khotanlou, Department of computer engineering, Bu-Ali Sina University, Hamedan, Iran

Yousef Rezaei, Department of civil engineering, Bu-Ali Sina University, Hamedan, Iran

Abstract: Multi-label image classification aims to predict multiple labels for a single image which consists of diverse contents. The main challenge in Multi-label classification task to achieve a decent performance is the lack of enough training data. Convolutional Neural Networks (CNN) has shown satisfying results in single-label image classification, but multi-label image classification is still an open field of research. In this paper an efficient hybrid method for multi-label image classification is proposed. The proposed model consists of multiple sub-networks. The experimental results obtained in this study demonstrate the plausible performance of the proposed method on "Pascal VOC 2012" and "Kaggle: Understanding the Amazon from space challenge" datasets.

Keywords: multi-label classification, deep learning, convolutional neural networks, satellite image classification

1 Introduction

Multi-label image classification is the task of mapping multiple tags/labels to an image. Multi-label image classification is more common and practical in real-world situations because it is more probable that more than one object exists in an image. Every multi-label image contains an area as background and objects from a set of known classes. Usually, the aim is to name the objects in the image ,but sometimes the background is also needed to be labeled.

For a long time, combining handcrafted features such as SIFT and SURF with common classifiers namely support vector machines(SVM) and random forests were the only approaches to solve image classification problems. Recently deep learning based methods have been replaced with traditional approaches.

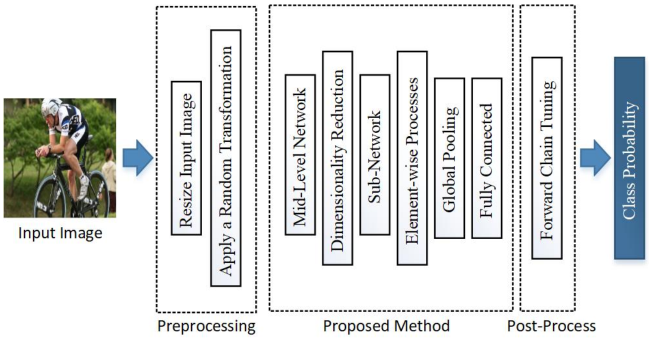

Figure 1: Data-flow on our proposed method.

The main problem of Convolutional Neural Networks (CNN) or generally deep learning approaches is that they require large amount of annotated data to prevent overfitting. One of the most common solution is using pretrained networks on large datasets such as imagenet, also known as transfer learning[14]. The other common solution is using data augmentation in training and testing phase.

The most influential challenge in the multi-label classification in compare to single-label classification is covering label space; the label space for each instance in single label classification is Y but the label space in multi-label classification becomes 2^{Y} for all possible label combinations for a new instance. Gathering labeled data for this label space is almost impossible.

In this paper a hybrid CNN network consists of several sub-network modules is proposed. The proposed method contains two state-of-the-art networks on single-

[*] Corresponding Author

[†] r.mohammadimoqaddam@eng.basu.ac.ir

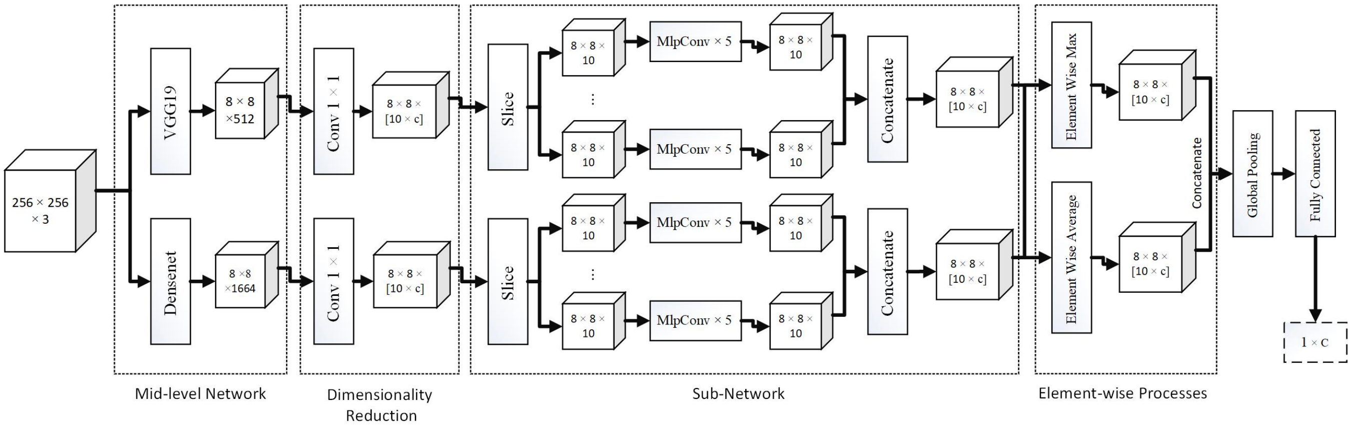

Figure 2: Overall Network Architecture. 'Mid-level Network' part consists of two pretrained networks. 'Dimentionality Reduction' and 'Sub-Network' parts are extracting more complex features while reducing the feature map channels. 'Element-wise processes' and 'Global Pooling' modules prepare feature maps, which are further sent to classifier. And the 'Fully Connected' part predict the class probabilities for each input.

label image classification followed by MlpConv layers.

2 Proposed Method

2.1 Background

The critical characteristic of the proposed model is that it extracts more elaborated and robust features in comparison to adopted state-of-the-art networks. The proposed method also uses MlpConv layers that enables it inherit the merits of network-in-network[10] architecture.

2.2 Model Architecture

Overall model architecture is demonstrated in Fig.2. First preprocessed image is passed through 'Mid-level Network' and the features are extracted. Extracted feature maps are then fed into 'Dimensionality Reduction' module. These new feature maps are fed into 'Sub-Network' followed by 'Element-wise operations'. The global-poolings are applied on final feature maps and finally, fully-connected layers predict class probabilities.

Mid-level Network As shown in fig 2 the 'Mid-level Network' contains two state-of-the-art networks in single-label classification. The proposed method follows the architecture of Densenet[7] and VGG19[13] networks in 'Mid-level Network' module. The openly available models pretrained on ILSVRC[12] dataset and the global pooling layers and fully-connected layers are eliminated from the last layers of the models. Output size of the models for Densenet will be [8, 8, 1664] and for VGG19 will be [8, 8, 512].

Dimensionality Reduction Despite the fact that 'Mid-level Network' module prepare favorable feature maps for classification, a convolution layer with filter size of [1 * 1] (conv 1*1) is applied in this module. This convolution layer is reducing the dimension of feature maps from [8, 8, 1664] and [8, 8, 512] to [8, 8, 10*c] (C is number of classes). It should be noticed that this layer also extracts more complex features from entered feature maps.

Sub-Network After dimensionality reduction by conv 1*1 all 10 * C channels are split to, 10 to 10 channels. Each of these slices will have [8, 8, 10] size. These layers are fed into 5 stacked MlpConv layers. This subnetwork network-in-network layers have 10 conv 1*1 in each stage followed by 2 mlp layers. After the last MlpConv layer each of these slices's shape will be [8, 8, 10] the same as the inputed shape. We will concatenate these slices; final shape of output layer of this module will be [8, 8, 10* C].

Element-wise Process Next module does element-wise max operation and element-wise average operation on outputted feature maps, came from each “Sub-Network” module. This layer extracts the compressed information from each inputted instance.

Global Pooling Global average pooling and global max pooling are done on output of ’Element-wise Mapping’ module that leads to [8,8] feature maps. Both of these feature maps are flattened and concatenated, so the final output shape will be [800,1]. Every one of these operations has its own benefits but aggregating them will lead to better performance which is the main motivation of using this section and our experiences prove this assumption. This layer will be fed into fully connected layer that plays the role of classifier.

Fully Connected Fully connected layer consists of 3 hidden layers. First layer consists of 512 neurons followed by a batch normalization layer followed by Leaky Relu activation function. Second stage is the same as first layer but it has 256 neurons. Final stage consists of number of classes neurons followed by sigmoid activation function so the final outputs will be class probability values.

3 Experiments and Results

3.1 Datasets

Pascal VOC PASCAL Visual Object Classes Challenge (VOC) dataset [4] is used as the main dataset to examine the performance of the model. The ’VOC 2012’ dataset consists of 11540 samples for train/validation and 10991 samples for test. It has 20 categories and each sample contains one or more objects of these 20 classes. All images are in raw RGB format and they have different sizes. ’Pascal VOC’ challenge results are reported in mean of Average Precision (mAP) metric as described in eq.(1).Where $|Q|$ is total number of samples, R_{jk} is the class probability and m_j is total number of classes for each sample. Fig.3 represents the class distribution of Pascal VOC 2012 dataset.

$$mAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \qquad (1)$$

Kaggle: Understanding the Amazon from space challenge dataset The second database used

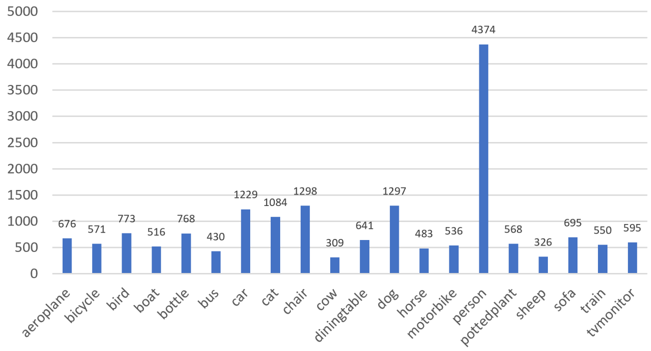

Figure 3: ’Pascal VOC’ dataset class distribution.

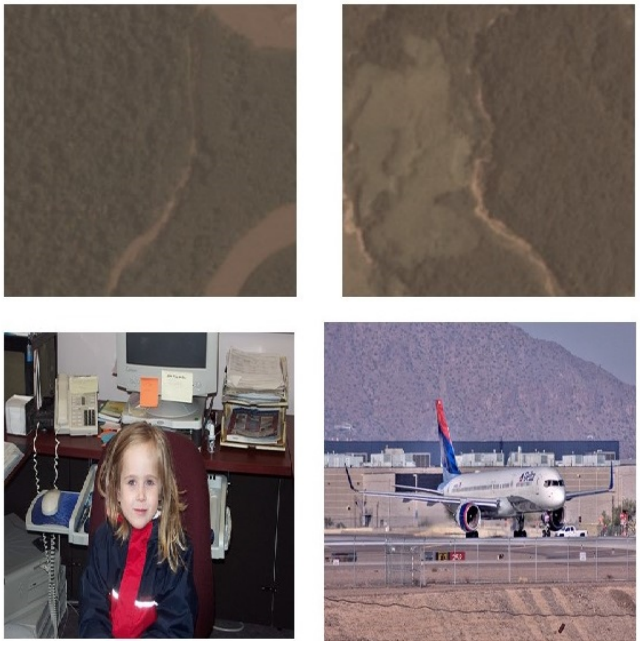

Figure 4: Example of some ’Kaggle: Understanding the Amazon from space challenge’ dataset images and ’Pascal VOC’ dataset images. Top-left image with ’clear primary water’ classes. Top-right image with ’agriculture clear primary water’ classes. Bottom-left image with ’chair person tvmonitor’ classes. Bottom-right image with ’aeroplane car’ classes.

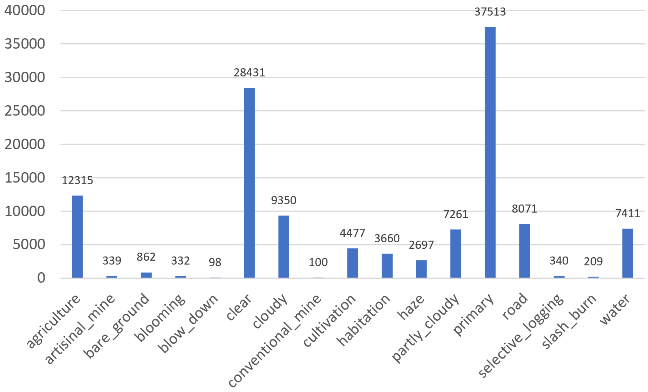

Figure 5: "Kaggle: Understanding the Amazon from space challenge" dataset class distribution.

for evaluation of the model is 'Kaggle: Understanding the Amazon from space challenge' dataset. This dataset has 40479 images for train/validation and 61191 images for test. The Amazon dataset has 17 tags for all of its samples; 4 of these tags are weather labels and the rest belongs to land cover categories. Images are available in two different formats, first format is Tif files that contains RGB and near infrared bands of satellite images, second format is Jpg files that contains only RGB bands. In this research we used the second format that uses only RGB bands. The evaluation metric is mean F-Score beta as described in eq.(2). Where TP is Number of Positive instances that predicted truly, FN is number of negative instances that predicted mistaken, FP is positive instances that predicted mistaken and β^2 is the coefficient attaches β times as much importance to recall as precision.
Table 5 represents the data distribution of Kaggle: Understanding the Amazon from space challenge dataset.

$$F_\beta = \frac{(1+\beta^2).TP}{(1+\beta^2).TP + \beta^2.FN + FP} \tag{2}$$

Evaluation servers are used to appraisement the model for both of the datasets since the test data annotations are not available offline.

3.2 Preprocessing

All images are resized to fixed dimension as (256, 256, 3) for Densenet network and (128, 128, 3) for VGG19 network so the output feature maps of each network will have the same width and height size.
The augmentation technique includes 6 different transformations. These transformations vary from domain to domain for example a dog image cannot be flipped from 'Pascal VOC' dataset Horizontally because the transformation turns the image to an invalid image but an image from 'Kaggle: Understanding the Amazon from space challenge' dataset can surely be flipped because there is no difference in the content of satellite images when they are flipped. In the test phase, test time augmentation(TTA) is used. All of the transformations are used in testing phase for predicting the output labels and the probabilities are aggregated by averaging them for each class label.

3.3 Training Scheme

Loss Function Softmax Cross-entropy loss function is most common loss function in single-label classification problems. But in multi-label classification problems binary cross entropy loss is more common. We should consider that Regularization is also a very crucial technique in deep learning models to prevent overfitting; so L2 of model weights is also added to loss function.

$$\iota(\theta) = (\frac{1}{n}\sum_{i=1}^{n} y_i log(p_i) + (1-y_i)log(1-p_i)) + (\lambda\sum_{j=1}^{k} w_i^2) \tag{3}$$

Where θ is model learnable parameters, y_i is the ground truth label, p_i is the predicted class probabilities and λ is regularization coefficient.

Training The model trained with a 1080 Ti GPU on a computer possess a core i7 6700 CPU with 16GB of ram running Microsoft Windows 10 operating system.
Cross validation of 5 folds have been applied on each of our experiments. In each training procedure the whole network have been fine-tuned. During training phase each fold had two stop criteria: first if validation loss change value fell bellow 1e-4, second if the number of epochs exceeds 50. We initiated the learning rate value as 1e-4 and a decay algorithm that divides the learning rate by 10 if validation loss difference, between two epochs fell below a constant value is applied. Xavier initialization[5] technique is used for initializing the extra

parameters that are not included in pretrained models. Adam[8] is used for optimizing the model parameters.

3.4 Post-process

For 'Kaggle: Understanding the Amazon from space challenge' dataset, evaluation servers require exact labels instead of class probabilities for each class as standard submission protocol.

A post-processing algorithm is to find the best confidence thresholds of class probabilities. We implemented a forward chain parameter tuning algorithm that tunes the first class threshold with respect to F-score beta value of the validation data, next it fixes the first threshold and tunes the second class threshold value with respect to F-score beta of validation value and so on. Threshold starts at 0.01 value and added by 0.05 in each step until it reaches 1.00 value.

Raw predicted class probability values used for 'Pascal VOC' dataset since the evaluation server standard protocol calculates the performance of submitted output from reported class probability values for each test image.

3.5 Results

The results of the proposed method on two datasets were compared with some benchmarks and state-of-the-art methods. We should notice that there have been more researchs done on 'Pascal VOC' dataset in comparison to 'Kaggle: Understanding the Amazon from space challenge' dataset. Experimental results on 'Pascal VOC' dataset is compared with VGG16[13], DeepMIL[11], Resnet-101[6] and WELDON[3] methods. Table 1 shows the brief comparison on overall mAP measurement.

Method	Mean AP
VGG16[13]	0.89
DeepMIL[11]	0.863
Resnet-101[6]	0.892
WELDON[3]	0.885
Proposed Method	0.916

Table 1: Classification performances (mAP) on Pascal VOC dataset.

For 'Kaggle: Understanding the Amazon from space challenge' dataset, our preference was to achieve the best performance among competitors in the leaderboard table. The leaderboard is the best way to evaluate the performance of the proposed model. The proposed method reached 14th position in the leaderboard table with F-measure beta score of 0.93204 that shows reliable performance on this competition. It is important to note that the winner of this competition used multiple test time augmentation(TTA) and used ensemble of models which is time consuming to predict the class labels followed by a post-porcessing algorithm. but we focused on creating a general method that can be used on various domains.

The proposed model is also compared with some other recent methods [1], [2] and [9] in table 2.

As the results shown in table 1 and 2, the proposed method achieves plausible results on Pascal VOC dataset and competitive results on kaggle dataset.

Method	Mean F-score beta
Amdahl et al.[1]	0.88941
Chen et al. [2]	0.93006
Kudli et al.[9]	0.9288
Competition Winner	0.93317
Proposed Method	0.93204

Table 2: Classification performances (mAP) on Pascal VOC dataset.

Discussion As evidenced by quantitative evaluations the proposed method have an admissible performance on mentioned datasets. We believe that using double CNN networks in 'Mid-level Network' module had an crucial effect on our proposed method performance. We also think extracting more complex features from obtained features from 'Mid-level Network' module was essential in multi-label classification problems. The 'Dimensionality Reduction' and 'Sub-Network' modules have done this task satisfactory.

One of the most potential part of CNNs for over-fitting is their top part that plays the role of classifier. Our experiments show that 'Element-wise Processes' module plays crucial role to prevent over-fitting in this part with applying pooling operations. The method is examined with ablation study on 'Pascal VOC 2012' dataset. First to authenticate the influence of using double networks in 'Mid-level Network' module, only VGG19 network was used and the performance dropped to 0.637 mAP, in the next step the 'Sub-Network' module was eliminated and the performance decreased to 0.905 mAP. Number of destination feature maps count in 'Dimentionality Reduction' module was also vary as we used 20 maps per class instead of 10 and the performance dropped to 0.901 mAP.

4 Conclusion

In this paper, a hybrid network structure that contains multiple sub-modules to address multi-label image classification problems was proposed. Pre-trained networks that trained on imagenet dataset can be used and various domains by using transfer learning technique. Recent works show transfer learning can achieve adorable results on different domains .The proposed method was evaluated on Pascal 'VOC 2012' and 'Kaggle: Understanding the Amazon from space challenge' datasets and shows admissible improvements.

The proposed architecture shows that using multiple pretrained networks as feature extractor with suitable engineered sub-modules leads to better performance in multilabel classification.

Future works include trying different networks in 'Midlevel Network' module.

References

[1] L. Amdahl-Culleton, M. Burkle and M. C. Horvitz, Understanding the Amazon from Space.

[2] Y. Chen, F. Dong, and C. Ruan, Understanding the Amazon from Space.

[3] T. Durand, N. Thome, and M. Cord, Weldon: Weakly supervised learning of deep convolutional neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 4743–4752.

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. J. I. j. o. c. v. Zisserman, The pascal visual object classes (voc) challenge. *International journal of computer vision*, **88(2)**(2010), 303–338.

[5] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, (2010), 249–256.

[6] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 770–778.

[7] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, Densely Connected Convolutional Networks, *CVPR*, **1(2)** (2017), 3.

[8] D. P. Kingma and J. J. a. p. a. Ba, Adam: A method for stochastic optimization, (2014).

[9] S. Kudli, S. Qian, and B. J. h. Pastel, Kaggle Competition: Understanding the Amazon from Space, **2913**, 749.

[10] M. Lin, Q. Chen, and S. J. a. p. a. Yan, Network in network, (2013).

[11] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, Is object localization for free?-weakly-supervised learning with convolutional neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 685–694.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and A.C. Berg. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, **115(3)**(2015), 211–252.

[13] K. Simonyan and A. J. a. p. a. Zisserman, Very deep convolutional networks for large-scale image recognition, (2014).

[14] L. Torrey and J. Shavlik, Transfer learning, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, (2010), 242–264.