Refinement of CNN Based Multi-label Image Annotation

Sangita Nemade[a] , Shefali Sonavane[b]

[a] Dept.of Computer Science & Engineering, Walchand College of Engineering, Sangli, India
[b] Dept. of Information Technology, Walchand College of Engineering, Sangli, India
email:[a]sangita.nemade@walchandsangli.ac.in,[b]shefali.sonavane@walchandsangli.ac.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: With the mushroom of technology, digital images are increasing rapidly, handling of these images has become an important research issue. Automatic image annotation (AIA) is a method for finding proper labels to an image in order to get a suitable way for searching and indexing the image data. AIA performs an influential role in image retrieval and its management. Exploiting the correlation among labels is a vital task in AIA for solving the semantic gap problems. Finding a contextual correlation among concepts can be helpful further to reduce this gap. To ensure effective capturing of this correlation, this paper presents co-occurrence patterns of labels along with random field methods for improving the performance of AIA. First, the DenseNet201 model is trained as a concept classifier for images and labels associated with images. Based on the training samples and concept vocabulary, co-occurrences of concepts are determined using association rule mining. The conditional random field (CRF) is used for refining the concept predicted by DenseNet201 CNN based on co-occurrence patterns. The experiment is carried out on the LableMe dataset. The performance analysis is carried out using the F1 score, recall and precision. From the obtained results, it is perceived that the proposed approach performs better than the DenseNet201 model.

Keywords: CNN, image annotation, co-occurrence pattern, CRF.

1. Introduction

Nowadays, digital photography plays a key role in humans' day-to-day life, which helps in sharing and remembering past events. Since the last decade, there is tremendous sharing of photographs over online social network websites such as Facebook, Twitter, Flicker, Instagram and Picasa, which enabled the posting and sharing of photos to end-users. This can be possible due to high dimensional and advanced photo capturing digital devices [1]. However, the handling of this large image collection is still not an easy task. Presently, content-based image retrieval (CBIR) is a useful method for handling large-scale images [2]. It has two types: keyword-based and instance-based methods. In the keyword-based method, images are searched based on keywords; therefore, database images need to be annotated. While in the case of an instance-based method rather than a keyword, the query image is given for searching and retrieving an image. Therefore, comparing these two techniques, the keyword-based method is more suitable for CBIR.

Image annotation's significance is valuable in CBIR [3]. The main aim of image annotation is to associate a set of related texts to the digital image so that the image's visual contents can be well described. The association between the content of the image and text labels enables the possibility of exploiting images with fast indexing and improving image retrieval performance. Image annotation is categorized into two types as handcrafted and automatic image annotation (AIA), respectively. Handcrafted techniques, where image labels are given manually to the image content, are not suitable for extensive image collection. The subjectivity of manual image annotation to the content of the image creates ambiguity. Due to these shortcomings of traditional image annotation, more research focuses on AIA. The primary aim of AIA is to reduce the semantic gap that generally persists in between the low-level visual content and high-level semantic description of an image [4]. AIA is usually utilized in areas such as image classification [5], image retrieval [6] and the medical domain [7].

The AIA problems are solved by five types of techniques from the last two decades: 1) Discriminative models, which are based on the prediction model using the classifiers such as SVM, ANN. 2) Generative models find labels based upon the combined probability of an image's labels and features from training data. It contains the topic models and mixture model techniques such as LDA, PLSA. 3) Nearest neighbor techniques such as k-nearest neighbor-based prediction, where it assumes that images with alike features are probable to have similar labels, such as JEC, Tagprop. 4) Sparse-coding models transferring reference image labels to the test image based on non-zero coefficients such as SELD, MLDL and SSRC 5) Recently, the deep learning model solved the AIA task through the representation of features based on deep learning. CNN is very popular for image annotation problems. The CNN's performance as a feature extractor is better than traditional handcrafted feature extraction methods for image annotation.

For image annotation, Wang et al. [8] have presented an architecture for ontology-based image annotation for finding contextual relevance between words that are automatically extracted from Google search. Linan Feng et al. [9] have introduced a method for reranking of concepts generated by generative and predictive model with the help of co-occurrence patterns and random walk. The multiple kernel learning (MKL) method for image annotation is presented by [10]. The multiple kernels refinement based on deep multi-layer networks is used and which is represented as the multi-layered combination of nonlinear activation methods. Every method is composed of many intermediate or elementary kernels, which lead to a positive semi-definite deep kernel. The different methods are introduced to learn network weights and plugged them into SVM for image annotation tasks. Discriminative feature mapping is done through the MIL scheme is presented in [11], where it explored both negative and positive correlations of concepts for the image annotation task.

The aforementioned models of AIA can be performed better than manual annotation of an image. However, results can be improved to a satisfactory level to overcome the semantic gap problem, which is a variation between the visual content and the semantics of an image. Image annotation refinement (IAR) is used to confront these problems. The purpose of using IAR is to find correlated tags for proper annotation of an image. Goodfellow et al. [12] have utilized CNN for object detection and classification with great achievement. However, the use of CNN in the literature can ignore some important principles that are used in object detection. In specific, computer vision researchers have revealed the effectiveness of semantic context in image annotation and object detection [13]. The CNN does not consider the contextual correlation between concepts, so the performance of AIA gets affected [14]. Thus, there is further scope for improvement of the performance of AIA. By taking this into account, this paper proposes AIA refinement based on the concept predicted by CNN and refined these concepts further based on the random field method using co-occurrence patterns, which is formed using association rule mining. Co-occurrence pattern [CP] is the association between concepts that gives some semantic clues for accurate prediction of concept. The semantic context in the form of the probable concept co-occurrence allows for predicting the correct label. For example, suppose the semantic concepts like "sky" and "building" form a co-occurrence pattern, then with strong confidence. In that case, the occurrence probability of “sky” can be anticipated by “building”, while due to weak co-occurrence, the prediction of “water” could be rejected. Therefore, an individual concept can be easily predicted using such a CP. Therefore, this CP can boost the detection accuracy of a concept/label. The paper's main contribution is to generate the concept signature using DenseNet201 and find the co-occurrence patterns of concepts from the training data. Finally, apply conditional random field (CRF) for refining the concept, which is the classifier's result.

The remaining paper is divided into various sections. Section 2 describes the architecture, CNN model, association rule mining, CRF and dataset. Experiments are conducted in Section 4, followed-by the conclusion given in Section 5.

2. Methodology

Generally, concepts are assigned randomly to the image. This limits the searching of the image based on labels. Therefore, in this paper, concept/label refinement is done. Here, the concept refinement model is proposed to automatically refine the concepts related to the test image. First, input images and vocabulary of concepts corresponding to an image are given as input to the DenseNet201 CNN model. This DenseNet201 model (inference model) generates concept signature. Concept signature is represented in a vector form where every entry denotes a row of concepts and its co-occurring probability score. The probability score in a vector (concept signature) is reranked by using a conditional random field (CRF) with the help of the CP.

2.1.CNN feature Extraction

In recent years, researchers have been using CNN everywhere for performing operations such as all kinds of image classification [15] and scene detection [16] because of their great achievement. CNN contains layers as convolutional, nonlinear transformation (that is ReLu), down-sampling (pooling) and fully-connected neuron layers. The convolutional layers create feature maps from the input layer and previous layers and deliver these feature maps to the ensuing layers. Generally, the ReLu layer is trailed after the convolutional layer for adding nonlinearity in the network. The down-sampling layers pick out the important features for the successive layers, which diminish the network's complexity. The image classification is accomplished using a fully-connected layer.

2.1.1 DenseNet201

In DenseNet201 [17], a simple connectivity pattern is applied in which each layer is directly connected to the layers which are succeeding to it. Thus, the full information flows in forward as well as backward pass computation. DenseNet consists of dense blocks. Every layer in the dense block gathers the feature map's information from its former layer and delivers the output (feature map) to its subsequent layers. The dense blocks are linked via a transition layer with batch normalization, a convolutional layer (1 x 1) and an average pooling layer (2 x 2). Each layer within a dense block includes batch normalization, nonlinear ReLu and 3 x 3 convolution layer. The network's computational efficiency is improved by adding a bottleneck layer of the convolutional layer

(1 x 1) before the convolutional layer (3 x 3). The global average pooling operation is performed on the final dense block and followed by the softmax classifier. This network gained an advantage over other networks like Alexnet, VGG, GoogleNet and ResNet in terms of improved accuracy, significantly reducing the training parameters, reusing features over the network with more compact learning, and relief from gradient vanishing problem which generally occurs during th

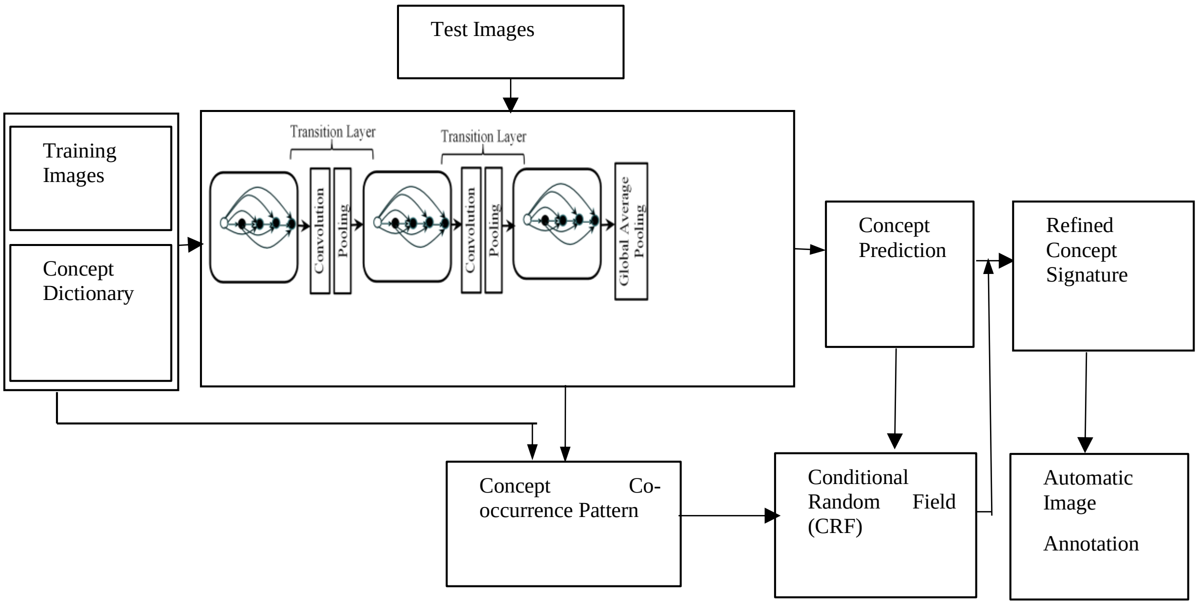


Fig. 1 Architecture of refined image annotation

2.1.2 Multilabel Image classification

Most of the deep learning models are created for single concept classification tasks. Therefore, in this research, the softmax loss layer of DenseNet201 is designed as per [18] for the multilabel classification problem. Using this softmax, the normalized prediction $p(Y_b|X_a)$ in the Image X_a in the bth concept Y_b is computed as

$$p_{a,b} = \frac{exp(q_b(X_a))}{\sum_{b=1}^{N_1} exp(q_b(X_a))} \qquad (1)$$

In the equation, the discrete probability distribution $q_b(X_a)$ of the image X_a for the bth class concept. $N_1$is the number of multi-concepts.

To lessen the KL divergence among ground-ruth and predicted probabilities, the softmax loss function is customized as per eq. (2).

$$\mathrm{f}\_{softmax} = -\frac{1}{N}\sum_{a=1}^{M_1}\sum_{b=1}^{N_1}\bar{p}_{a,b}log(p_{a,b}) \qquad (2)$$

(2
)

where, $\bar{p}_{a,b}$ is the ground truth for class b of image X_a that is if $\bar{p}_{a,b}$= 1 means the bth class concept of the image is present and if $\bar{p}_{a,b}$ = 0 means the bth class concept of the image is absent, M_1 is the number of images and f$_{softmax}$ is the cross-entropy loss.

2.2.Finding concepts co-occurrence patterns using association rule mining

Association rule mining algorithm [19] is used to identify the frequent occurrences of concepts over the training set. This technique operates between feature extraction and image classification. Based on identifying the frequent concept occurrences, association rules are formed. These rules suggest the presence of the concept with high confidence. The market-basket analysis is considered a standard example for the extraction of association rules where image labels are considered items. For example, items are described as kinds of stuff present in the market that any person can purchase and transactions are the varied items included in market baskets. In this study, the aim of discovering association rules is to identify the co-occurrences of concepts. Generating a proper set of association rules relies mostly on support and confidence. The calculations of support and confidence are given in eq. (3) and eq. (4).

$$Support(A \rightarrow B) = count(A \cup B)/N \qquad (3)$$

Where N denotes all images in the database.

$$Confidence(\mathrm{A} \rightarrow \mathrm{B}) = count(\mathrm{A} \cup \mathrm{B})/count(\mathrm{A}) \qquad (4)$$

In the rule, A → B defines how many labels B appear in other images that contain A is recognized as the confidence of the rule. Whereas, to determine how frequently a rule is utilized in the database is termed as the support of the rule. CP list is created based on confidence score in terms of $1 - Confidence$.

2.3.Conditional random field (CRF)

The third component of refining IA is a conditional random field (CRF) model, a kind of Markov random field (MRF) of an undirected graphical model. In CRF, every node matches a random variable y (labels) and the dependency of random variables is shown by edges. Each random variable's distribution is conditioned on the x input sequence. In this study, concepts with powerful contextual co-relation (for example, building and sky) can promote each other, whereas weak contextual co-relation (for example, sky and table) can refute each other.

The conditional probability t given s is expressed as,

$$p(t|s) = \frac{e^{\psi(t,s;\Theta)}}{\sum_{t'} e^{\psi(t',s;\Theta)}} \qquad (5)$$

The potential function is expressed as,

$$\psi(t,s;\Theta) = \sum_1^p \sum_1^r \theta_r^1 \, f_r^1(t_p,p,s) + \sum_{1,1}^{p,q} \sum_1^h \theta_h^2 \, f_h^2(t_p,t_q,p,q,s) \qquad (6)$$

Where p and q are the indexes of the vertexes. $f_r^1(t_p,p,s)$ and $f_h^2(t_p,t_q,p,q,s)$ are the functions of node feature and edge feature respectively. $\Theta = \{\theta^1,\theta^2\}$ are the learning parameters of the model.

For decreasing the number of parameters required to learn the model by setting some parameters as prior information. Especially, the potential function given in equation (6) is modified as,

$$\psi(t,s;\Theta) = \alpha_1 * \sum_1^p \omega^1(t_p,p,s) \; + \; \alpha_2 * \sum_{1,1}^{p,q} \omega^2(t_p,t_q,p,s) \qquad (7)$$

Where ω^1 is the t_p state's local evidence, depending on the observation of the image s. The prior parameter ω^2 denotes the co-occurrence potential in between the states of $t_p \; and \; t_q$ variables. $\alpha_1 \; and \; \alpha_2$ are the learning parameters.

The local evidence is the probability score's logarithm generated by the classifier,

$$\omega^1\big(t_p = 1,p,s\big) = log p_{softmax}\big(c\big(t_p\big) = 1\big|s\big) \qquad (8)$$

$$\omega^1\big(t_p = 0,p,s\big) = log[1 - p_{softmax}\big(c\big(t_p\big) = 1\big|s\big)] \qquad (9)$$

Where $p_{softmax}$ is the probability score detected by the classifier specified in eq. 8 and 9, $c\big(t_p\big)$ represents the concepts of t_p. It is assumed that the potential (co-occurrence of concepts) is independent of input image s. Therefore, the co-occurrence potential is converted as,

$$\omega^2(t_p,t_q,p,q) \; = \begin{cases} 1 - (-log \; CP(c\big(t_p\big),c\big(t_q\big))) & if \; t_p = t_q = 1 \\ \quad 0 & otherwise \end{cases} \qquad (10)$$

The $CP(c\big(t_p\big),c\big(t_q\big))$ represent the function to determine the score of t_p and t_q based on co-occurrence patterns. The parameters are optimized using the gradient descent technique [20]. The refined annotation predicts the most likely state of an indicator variable through the marginalized probability, stated in eq. 11.

$$t_p^* = argmax \; p(t_p|s;\Theta)) \qquad t_p \in \{0,1\} \qquad (11)$$

2.4.Datasets

The LabelMe image annotation dataset [21] encompasses 72,852 images with more than 10,000 concepts such as sky, person books, building, rock, car, bus, cycle, etc. In this research, a subset of 10000 images with associated concepts is utilized. The raw images have varying resolutions. For the experimentation, images comprising a resolution of 1600 x 1200 are downloaded via the website's toolbox (of LabelMe dataset).

3.Experimental Results

Experiments are accomplished by the configuration of a 64-bit Intel Core i7 processor having 2.80 GHz operating-speed. The dataset is separated at random into the training (60%), validation (20%) and testing (20%) sets. The metrics of evaluations used for this experimentation are F1 score, recall and precision. Precision is a ratio of relevant labels (N1) among the total number of retrieved labels and recall is given as a proportion of correctly predicted labels with ground-truth labels. F1 score metric is a harmonic mean of recall and precision (eq. 12).

$$F1\ score = 2 * \frac{(Precision * Recall)}{Precision + Recall} \tag{12}$$

The learning rate, batch size, dropout rate and epoch are set to 0.0001, 32, 0.5 and 20, respectively, and the binary cross-entropy loss function is chosen. The initial value of α_1 and α_2 is set to 0.1.

The DenseNet201 model produces the labels with the score for every image. This probability score is reranked by using a conditional random field (CRF) using the co-occurrence pattern. The first ten concept signatures of DenseNet201 are selected for refinement purposes and rerank their concepts as per marginalized probability, given in eq. (11).

The experimental results of the DenseNet201 model and the proposed model are provided in Table 1. It contains the top-ranked ten individual concepts, which are taken based on the F1 score from 0.0 to 1.0. The performance is measured using top-ranked concepts from ground truth and refined Image annotation and compared these with the highest 60 concepts produced by DenseNet201.

Table 1. Average F1 score of top concepts

Sr. No.	**Individual concepts**	**F1 score of DenseNet201**	**F1 score of the proposed model**
1	building	0.9294	0.9573
2	sidewalk	0.9077	0.9364
3	sky	0.8942	0.9289
4	road	0.8545	0.904
5	car	0.805	0.8506
6	window	0.7457	0.8056
7	wheel	0.6758	0.7557
8	trees	0.6215	0.7049
9	door	0.5615	0.6356
10	person	0.5518	0.6124

The average precision values with the highest individual concepts is shown in Fig. 2. From Fig. 2, it is observed that the annotation performance is improved consistently using the proposed method after performing the concepts refinement task.

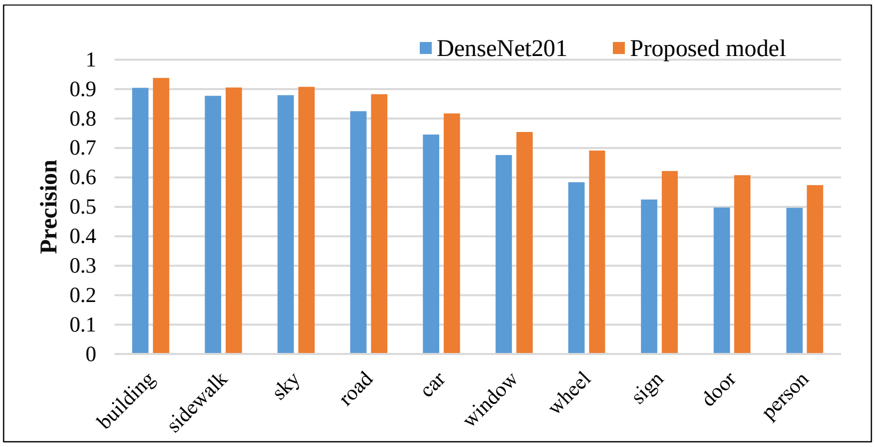


Fig. 2 Top concept’s average precisio

Test images	Ground truth	DenseNet201	Proposed model

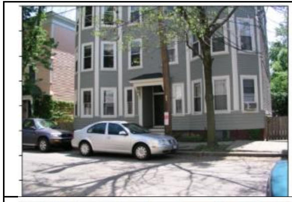
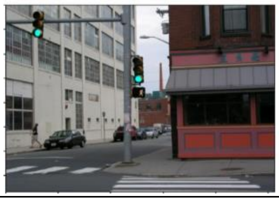
	building road car window sidewalk	building car door person road	building road car sidewalk window
	sidewalk road building sky car	sidewalk road building wheel car	sidewalk road building sky car

Fig. 3 Test images with their predicted concepts using DenseNet201 and proposed method.

Fig. 3 displays test images with their predicted concepts using the DenseNet201 model and the proposed method on the LabelMe dataset. As shown in Fig. 3, The top-k (k=5) annotation length is fixed for result comparison. For each image, k highest-ranked labels are assigned and then compared these labels with the ground-truth labels. Overall, when annotation (label) length is more than one then the proposed model attains the best performance, as the co-occurrence patterns information are further utilized for a greater number of annotations.

4.Conclusion

This paper has presented the co-occurrences of concepts which acts as contextual cues for improvement of concept signature. The DenseNet201 CNN model produced the concept signature with the help of a modified softmax layer. CRF uses the probability score of concept signature and co-occurrence patterns produced by association rule mining to reranking the concepts to get the final concept set of the sample test image, enhancing the AIA performance and getting more ideal results. The experimentation is done on the large LabelMe dataset. This research's main objective is to examine the performance of DenseNet201 and the proposed model on the LableMe dataset. From the experimentation, it is perceived that the proposed model achieved good F1 score than DenseNet201. The maximum F1 score gained by the proposed model is 0.9573 for the concept building. In general, the overall performance is improved after utilizing the CP for concept signature refinement.

References

1. Duygulu, P. K., Barnard, J. F. G. de Freitas, D. A. Forsyth.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV, pp. 97–112 (2002).
2. Zou, F., Liu, Y., Wang, H. et al. Multi-view multilabel learning for image annotation. Mul-timed Tools Appl, 75, 12627–12644 (2016).
3. Yang, Y., Huang, Z., Yang, Y., Liu, J., Shen, H., Luo, J.: Local image tagging via graph regularized joint group sparsity. Pattern Recognit., vol. 46, no. 5, 1358–1368 (2013).
4. Zhang, L., Han, Y., Yang, Y., Song, M., Yan, S., Tian, Q.: Discovering discriminative graphlets for aerial image categories recognition. IEEE Trans. Image Process., vol. 22, no. 12, pp. 5071–5084 (2013).
5. Wang, C., Blei, D., Li, F.: Simultaneous image classification and annotation. In: IEEE Computer society conference on computer vision and pattern recognition workshops. CVPR Workshops, pp 1903–1910 (2009).
6. Ghosh, N., Agrawal, S., Motwani, M.: A survey of feature extraction for content-based image retrieval system. International Conference on Recent Advancement on Computer and Communication. Springer,Singapore, pp 305–313 ((2018).
7. Wang, S., Chang, XJ., Li, X., Long, G., Yao, L., Sheng, QZ.: Diagnosis code assignment us-ing sparsity based disease correlation embedding. IEEE Trans Knowl Data Eng vol. 28, no. 12, pp. 3191–3202 (2016).
8. Wang, Y. and Gong, S.: Refining image annotation using contextual relations between words. In Proceedings of the 6th ACM international conference on image and video retrieval, pp. 425-432 (2007).
9. Feng, L., Bhanu, B.: Semantic Concept Co-Occurrence Patterns for Image Annotation and Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 4, pp. 785-799 (2016).
10. Jiu, M., Sahbi, H.: Nonlinear Deep Kernel Learning for Image Annotation. IEEE Transactions on image processing, vol. 26, no. 4, pp. 1820-1832 (2017).

11. Hong, R., Wang, M., Gao, Y., Tao, D., Li, X., Wu, X.: Image Annotation by Multiple-Instance Learning With Discriminative Feature Mapping and Selection. IEEE Transactions on Cybernetics, vol. 44, no. 5, pp. 669-680 (2014).
12. Goodfellow, I., Pouget-Ab, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014).
13. Rabinoch, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Blongie, S.: Objects in context. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil pp. 14-20 (2007).
14. Xu, H.: Automatic Image Annotation Based on Hidden Markov Model and Convolutional Neural Network. Comput. Sci. Appl., vol. 08, no. 09, pp. 1309–1316 (2018).
15. Virnodkar, S., Pachghare, V., Patil, V., Jha, S.: CaneSat dataset to leverage convolutional neural networks for sugarcane classification from Sentinel-2. Journal of King Saud Univer-sity-Computer and Information Sciences, In Press, pp. 1-13 (2020).
16. Pan, X., Shi, J., Luo, P., Wang, X., Tang, X.: Spatial as deep: Spatial cnn for traffic scene understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7276-7283 (2018).
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016).
18. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe S.: Deep convolutional ranking for multilabel image annotation. In: International Conference on Learning Representations, pp. 1-9 (2014).
19. Agarwal, R., Srikant, R.: Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference pp. 487–499 (1994)
20. X. Li, X.: Preconditioned Stochastic Gradient Descent. IEEE Transactions on Neural Net-works and Learning Systems. vol. 29, no. 5, pp. 1454–1466 (2018).
21. Russell, C., Bryan, A., Murphy, K., William, T., Freeman.: LabelMe: a database and web-based tool for image annotation. International journal of computer vision, vol. 77, no. 3, pp. 157-173 (2008).