Semantic Multinomial Representation for Scene Images Using CNN-Based Pseudo-concepts and Concept Neural Network

Deepak Kumar Pradhan[1], Shikha Gupta[2], Veena Thenkanidiyoor[1(✉)], and Dileep Aroor Dinesh[2]

[1] Department of Computer Science and Engineering, National Institute of Technology Goa, Ponda 401403, Goa, India
eb.deepakpradhan@gmail.com, veenat@nitgoa.ac.in
[2] School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Mandi 175001, H.P., India
shikha_g@students.iitmandi.ac.in, addileep@iitmandi.ac.in

Abstract. For challenging visual recognition tasks such as scene classification and object detection there is a need to bridge the semantic gap between low-level features and the semantic concept descriptors. This requires mapping a scene image onto a semantic representation. Semantic multinomial (SMN) representation is a semantic representation of an image that corresponds to a vector of posterior probabilities of concepts. In this work we propose to build a concept neural network (CoNN) to obtain the SMN representation for a scene image. An important issue in building a CoNN is that it requires the availability of ground truth concept labels. In this work we propose to use pseudo-concepts obtained from feature maps of higher level layers of convolutional neural network. The effectiveness of the proposed approaches are studied using standard datasets.

1 Introduction

Scene image understanding is important for cognitive task such as scene classification. Early efforts in scene understanding shown that the rich semantic content comprising of multiple concepts can be effectively represented using set of local feature vectors comprising of SIFT [1] and HoG [2] that are low-level features. Our brain identifies a scene image from the composition of semantic content rather than colors or edges of the scene over small patches. For effective understanding of scene images, it is necessary to map them onto a suitable semantic representation. Some of the semantic scene representations are holistic scene representation [3], bag-of-visual-words representation [4], topic space representation [5], and concept occurrence vector representation [6] which represent the semantic content of an scene image. In this work, we explore semantic multinomial (SMN) representation for scene images.

R. Rameshan et al. (Eds.): NCVPRIPG 2017, CCIS 841, pp. 400–409, 2018.
https://doi.org/10.1007/978-981-13-0020-2_35

Semantic multinomial representation (SMN) is a semantic representation of a scene. SMN representation of a scene image is a vector of posterior probabilities corresponding to semantic concepts [7]. The value of a concept posterior probability can be computed by first building a suitable model for the concept. In [7], a GMM-based approach to obtain the SMN representation for a scene image is proposed. This method involves building a GMM for every semantic concept using the local feature vectors from all the images having that concept label in their ground truth. In [8], SVM based approaches to obtain SMN representation were explored. In this work, we propose to build a concept neural network (CoNN) to obtain the SMN representation for scene images. An important issue in building the concept models is that it requires the availability of ground truth concept labels for images. In [9], an approach to build concept models using pseudo-concepts in the absence of ground truth concept labels is proposed.

In this work we propose to use the pseudo-concepts obtained from feature maps of the higher level layers of convolutional neural network (CNN) [10]. We also propose to build a CoNN using the features extracted form CNN. Major contributions of this work are as follows: (1) Obtaining pseudo-concepts from CNN (2) Building CoNN to obtain SMN representation. The effectiveness of the approaches presented in this work are studied using standard datasets.

2 An Overview of Semantic Scene Representations

Semantic multinomial representation (SMN) is an intermediate scene representation. Several intermediate scene representations such as holistic scene representation [3], bag-of-visual-words representation [4], topic space representation [5], and concept occurrence vector representation [6] exist which represent the semantic content of an scene image. A holistic scene representation is be made up of structure or shape of a scene image using a set of spatial envelope properties such as degree of naturalness, degree of openness etc. [3]. The bag-of-visual-words [4] representation is a vector of frequencies of occurrence of visual words in the image. A visual word represents a specific local semantic pattern shared by a group of local feature vectors. The topic space (TS) representation of a scene image is a latent space representation that is a vector of posterior probabilities corresponding to the topics. Here, a topic corresponds to a 'theme' or a latent aspect that is related to a semantic concept as discussed in [5]. It is a generative process. The concept occurrence vector (COV) representation of a scene image corresponds to a vector of frequencies of occurrence of local semantic concepts in a scene image [6,11]. The COV representation of a scene image is obtained by assigning a concept label to every local feature vector of a scene image and the frequency of occurrence of every concept label is determined. In this method, it is necessary to build a suitable concept classifier to assign a local feature vector of an image to a semantic concept. For building a classifier to obtain a COV representation, we essentially need a data set where the concept label for each local feature vector of the image is available. It is difficult to have such a data set

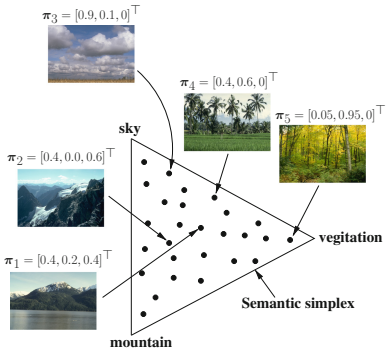


Fig. 1. Illustration of mapping of scene images onto points in a semantic simplex, where the number of concepts, $\mathcal{C} = 3$. Here, $\boldsymbol{\pi}$ for an image is the vector of posterior probabilities for the three concepts, 'sky', 'vegetation', and 'mountain' in that order.

because most of the datasets have the semantic information only at class level. Hence, it is better to have an intermediate scene representation such as SMN representation that can be obtained using the image level semantic information.

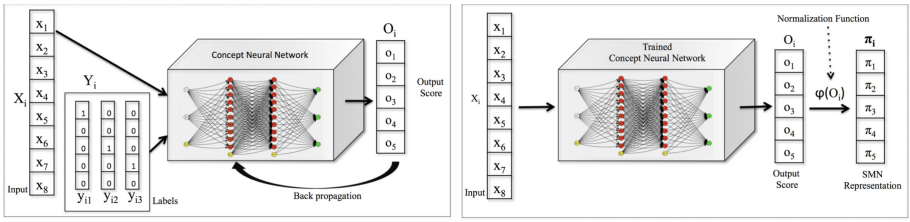
The semantic multinomial (SMN) representation of a scene image is a vector of posterior probabilities corresponding to semantic concepts [7]. An SMN representation $\boldsymbol{\pi}$ is given by $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_c]^\top$ where c corresponds to the number of semantic concepts and π_k is the posterior probability corresponding to kth concept. The SMN representation can be considered as a transformation that maps an image I onto a point in a C dimensional probability simplex as illustrated in Fig. 1. The SMN representation can be considered as a transformation that maps an image I onto a point in a C dimensional probability simplex, as illustrated in Fig. 1. To obtain the SMN representation of a scene image, it is necessary to build concept models. A concept model for kth concept can be built using all the images having that concept label. In [7], a GMM-based approach to obtain the SMN representation for a scene image is proposed. This method involves building a GMM for every semantic concept using the local feature vectors from all the images having that concept label in their ground truth. In [8], SVM based approaches to obtain SMN representation has been explored where dynamic kernels like intermediate matching kernel (IMK), GMM supervector kernel (GMMSVK) and GMM universal back-ground model mean interval kernel (GUMIK) for SVMs and used for obtaining SMN representation.

An important issue in obtaining SMN representation for an image by building concept models is that it requires ground truth concept labels for the images. All the datasets are not having the ground truth concept labels. In the absence of ground truth concept labels, alternative approaches to build the concept models may be explored. In [9], the usage of pseudo-concepts in the absence of ground truth concept labels was proposed. The paper proposed the usage of clusters of local feature vectors of all the images for getting pseudo-concept labels. The approach proposed in [9] cannot be used when the images are not represented

as sets of local feature vectors. In this work, we propose an alternative way of obtaining pseudo-concepts using the convolutional neural networks. This work also proposes to build a neural network based concept model to obtain the SMN representation.

3 Concept Neural Network for Generating SMN Representation

In this work, we propose to build a concept neural network (CoNN) to generate SMN representation. The CoNN is a feed forward neural network (FFNN) that comprises of an input layer, an output layer and one or more hidden layers. Let $\mathcal{D} = \{I_1, I_2, \ldots, I_i, \ldots, I_N\}$ be the set of all scene images in a dataset. Let $\{\omega_1, \omega_2, \ldots, \omega_k, \ldots, \omega_c\}$ be a set of c concept labels. Let $\mathbf{X}_i \in \mathcal{R}^d$ be a d-dimensional feature vector representation for an image I_i. Let there be p concept labels associated with I_i. The CoNN comprises of d number of neurons in input layer and c number of neurons in the output layer. The number of neurons in the hidden layers need to be chosen carefully for a dataset. Weights in a CoNN are estimated using error back propagation method. However, the training of a CoNN differs from that of a FFNN in the way in which examples are presented. In a conventional FFNN, every input $\mathbf{X}$ is associated with a target vector $\mathbf{y}$ which need to be presented together during training process. Typically $\mathbf{y}$ is a c-dimensional vector encoded in one-hot-notation. As we know that a scene image comprises of multiple semantic concepts. Hence there are multiple concept labels associated with one image. Suppose that $\mathbf{Y}_i = \{\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{ip}\}$ is a set of one-hot target vectors corresponding to p concept labels of an image I_i. Here $\mathbf{y}_{ij}$ is a c-dimensional one-hot vector. During training, a scene image presented to CoNN multiple times with different target vectors unlike as in a conventional FFNN where a scene image is presented only once. The instance of training a CoNN using an image represented as 8-dimensional vector and comprising of 3 concepts is illustrated in Fig. 2a. The image is presented to the network thrice, every time with a different target vector.



(a) Training of CoNN (b) SMN representation generation from CoNN

Fig. 2. Illustration of the training (a) and testing (b) phase of CoNN. Here we assume that the input feature vector is 8-dimensional and there are total $c = 5$ concepts available in the dataset. The illustration also assumes that the particular scene image that is being fed to CoNN has $p = 3$ concepts in it.

The process of testing CoNN is shown in Fig. 2b. A feature vector $\mathbf{X}_i$ corresponding to a scene image I_i is fed to the CoNN and a c-dimensional score vector $\mathbf{O}_i = [o_1, o_2, \ldots, o_k, \ldots, o_c]^\top$ is obtained at the output layer. Score o_k for kth concept label, ω_k is mapped onto a pseudo probability using a logistic function as follows:

$$\tilde{P}(\omega_k \mid \mathbf{X}_i) = \frac{1}{1 + \exp(-\alpha o_k)} \tag{1}$$

where α is a parameter that controls the slope of the logistic function. The pseudo probability is then normalized to obtain the posterior probability value corresponding to a concept k as follows:

$$\pi_k = P(\omega_k \mid \mathbf{X}_i) = \frac{\tilde{P}(\omega_k \mid \mathbf{X}_i)}{\sum_{j=1}^{c} \tilde{P}(\omega_j \mid \mathbf{X}_i)} \tag{2}$$

The semantic multinomial representation of a scene image is obtained as a vector of posterior probabilities corresponding to the semantic concepts as $\boldsymbol{\pi}_i = [\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_c]^\top$.

An important issue in building a CoNN is that it is necessary to have the ground truth (true) concept labels for the images. In the absence of ground truth concept labels, [9] proposed to obtain semantic scene representation using pseudo-concepts. In this work, we propose to obtain pseudo-concepts labels for images using convolutional neural networks.

4 Pseudo-concept Labels from Convolutional Neural Network

A convolutional neural network (CNN) is a deep neural network comprising of multiple convolutional layers and a fully connected layer [12]. The convolutional layers comprise of filters those when are convolved with the images extract certain specific features. It is important to note that in a CNN, filters are learnt from data. In the lower layers, the filters correspond to the primitive features like edge, texture etc. It is observed that the filters in higher level convolutional layers do contain semantic features [10]. We propose to consider these semantic features. The response of a filter to an image is known as feature map. All the images in a dataset are presented to a trained CNN. Feature maps in a layer for an image are considered. Every feature map is sum pooled and the filters corresponding to top c responses among all the images are considered as pseudo-concepts. Any image for which response of a particular filter chosen as pseudo-concept is above a threshold, that image is labelled with that pseudo-concept. While choosing c pseudo-concepts, it is necessary also to ensure that a particular pseudo-concept does not appear in a very few images in the dataset. This is because, with only a few example images, it becomes difficult to build concept models. It is also necessary to ensure that a particular pseudo-concept does not appear in almost all the images because then discriminating a pseudo-concept becomes difficult. We make sure that a particular pseudo-concept to be present in at least 'u' and at most 'v' number of scene images while considering total 'c' number of concepts.

5 Experimental Studies

The effectiveness of the proposed approaches is studied using MIT-8-scene [3] and Vogel-Schiele [13] scene image datasets. The MIT-8-scene dataset comprises of 2688 scene images belonging to 8 semantic classes. Every image in MIT-8-scene dataset is of 256×256 pixels in size. The results presented for this dataset correspond to 5-fold evaluation. In every fold, 100 scene images per class are used for training and the rest are used for testing. The Vogel-Schiele dataset comprises of 700 images belonging to 6 natural scene categories. The results presented for this dataset correspond to stratified 5-fold evaluation. Every image in Vogel-Schiele dataset is either 720×480 or 480×720 pixels in size. In this work, we propose to represent every scene image using features derived from a deep convolution neural network (CNN). We propose to use AlexNet [12] trained using Places 205 dataset [14]. We refer to this CNN as Places205-AlexNet. Though every layer in a CNN corresponds to a scene representation, we propose to consider convolutional layer 4 (Conv-4) and pooling layer 5 (Pool-5) features. The number of features in Conv-4 and Pool-5 layers correspond to 64896 and 9216 respectively.

First we study the proposed approach for obtaining the pseudo-concepts using the Places205-AlexNet. In this work we consider the feature maps of Conv-4 or Pool-5 layers and sum pool every feature map. The feature maps whose responses are above a threshold are chosen as pseudo-concepts. The suitable value for the threshold need to be chosen empirically. In this work, we considered 1600 as a threshold value for the Conv-4 layer and 220 as a threshold value for the Pool-5 layer. The above thresholds were used for both MIT-8-Scene and Vogel-Schiele datasets. In addition to this the redundant pseudo-concepts and the pseudo-concepts that may cause bias on the semantic scene modelling are removed. For this we consider only those pseudo-concepts that are present in at least 40 scene images and those do not occur in more than 80 scene images. The total number of pseudo-concepts for each fold of the two datasets are given in Table 1.

Table 1. Number of pseudo-concepts in each fold corresponding to MIT-8-Scene and Vogel-Schiele datasets corresponding to Conv-4 and Pool-5 layers of Places205-AlexNet.

	Fold1	Fold2	Fold3	Fold4	Fold5
Conv-4 (MIT-8-scene)	91	92	89	94	94
Pool-5 (MIT-8-scene)	101	100	95	100	101
Conv-4 (Vogel-Schiele)	74	75	70	74	77
Pool-5 (Vogel-Schiele)	167	165	167	167	166

Next we study the proposed approach to obtain the SMN representation using concept neural network (CoNN). The number of neurons in the input layer of CoNN correspond to the number of features and the number of neurons in output layer correspond to the number of pseudo-concepts. The number of neurons in

Table 2. Comparison of classification accuracy (in %) of SVM-based classifiers on MIT-8-scene and Vogel-Schiele datasets using CoNN-based SMN representation. Here, linear kernel (LK), histogram intersection kernel (HIK) and χ^2 kernel (χ^2) are explored for building SVM-based classifiers.

Method to obtain SMN representation	LK	HIK	χ^2
CoNN using Conv-4 features (MIT-8-scene)	85.88	87.00	87.87
CoNN using Pool-5 features (MIT-8-scene)	91.25	91.17	**91.39**
CoNN using Conv-4 features (Vogel-Schiele)	85.88	87.00	87.87
CoNN using Pool-5 features (Vogel-Schiele)	91.25	91.17	**91.39**

Table 3. Comparison of classification accuracy (in %) of SVM-based classifier that uses the SMN representation proposed in this work and CNN-based classifier on MIT-8-scene (MIT) and Vogel-Schiele (VS) datasets.

Representation	Classifier	MIT	VS
Places205-AlexNet Pool-5	CNN	90.75	74.45
CoNN-based SMN	SVM-χ^2 kernel	**91.60**	**77.30**

the hidden layer is chosen empirically by varying it from 2 to 2048. The slope of sigmoidal activation function is also chosen empirically. The effectiveness of the proposed approach to obtain SMN representation is studied by using SVM-based scene classification that uses the SMN representation. Since the SMN representation is a non-negative vector (like histogram vector), we consider SVM-based classifier using histogram intersection (HI) kernel and χ^2 kernel apart from linear kernel (LK). We have used LIBSVM [15] tool to build the SVM-based classifiers. Table 2 gives the accuracy of SVM-based scene classifiers using the SMN representation obtained using the proposed method for MIT-8-scene and Vogel-Schiele datasets. It is seen from Table 2 that the χ^2 kernel based SVM on SMN representation obtained using the CoNN performs better than the SVM-based classifier using HIK and LK. It is also seen that the performance of scene classification for SMN representation obtained from Pool-5 features is better than that obtained using Conv-4 features. One possible reason for this may be the huge dimensionality of the Conv-4 features. Building a CoNN using Conv-4 features as input requires a lot of data due to the large number of parameters to be estimated. For the rest of the studies, we consider SMN representation obtained using the Pool-5 features as input to CoNN.

Next we compare the performance of an SVM-based classifier that uses SMN representation obtained using the proposed approach with that of a CNN-based classifier. For this, the fully connected layer of Places205-AlexNet adapted to the given dataset. The number of nodes in the hidden layers are empirically chosen after exploring various options. The performance of the CNN-based classifier is compared with that of the SVM-based classifier that uses SMN representation in Table 3. It is seen from Table 3 that the performance of SVM-based classifier that

Table 4. Comparison of scene classification accuracy (ca) (in %) using the SMN representation obtained using clustering based and CNN-based pseudo-concepts and set of local feature vectors representation of scene images for Vogel-Schiele and MIT-8-scene dataset.

Image representation		Classification models	MIT-8-scene	Vogel-Schiele
Set of HOG vectors		GMM	66.92	64.13
		CIGMMIMK-based SVM	79.86	71.33
		GMMSPMK-based SVM	81.84	71.65
		IFK-based SVM	81.28	72.34
Pool-5 feature of Alexnet-Places-205		CNN	90.75	74.45
Clustering based pseudo-concepts	SMN from GMM-based approach	χ^2-kernel SVM	79.43	68.23
	SMN from GMMSPMK-based SVM		81.45	71.21
CNN-based pseudo-concepts	SMN from CoNN		91.60	77.30

uses SMN representation is significantly better than the CNN-based classifier. To adapt a CNN to a given dataset needs to tune hidden layers in the fully connected layers where the number of neurons need to be chosen. The results in Table 3 show the effectiveness of using the SMN representation for scene classification.

In Table 4, we compare the performance of SVM-based scene classification using SMN representation obtained using proposed approach with that obtained using clustering based approach proposed in [9]. We also compare the performance of scene classification using SMN representation with that using the set of local feature vectors representation for images. Due to the limitation of space, the details of the scene classification using set of local feature vectors representation can not be presented here. The details can be seen in [9]. It is seen from Table 4 that the classification accuracy obtained for the SMN representation of scene images is better than that obtained for the set of local feature vectors representation of scene images. It is also seen that performance of CNN-based classifier is better than that of the SVM-based classifier that uses SMN representation obtained using the clustering based pseudo-concepts. This shows the effectiveness of CNN in tasks involving scene understanding such as scene classification.

6 Conclusion

An approach to obtain SMN representation of scene images using concept neural network (CoNN) is proposed in this paper. An SMN representation is a semantic

representation that is useful in cognitive tasks such as scene classification. Building a CoNN requires ground truth concept labels. In the absence of ground truth concept labels, the paper proposed to build CoNN using pseudo-concepts. An approach to obtain pseudo-concepts using the feature maps of higher level layers of convolutional neural network is proposed in this paper. The effectiveness of the proposed approach to obtain SMN representation is studied by building SVM-based scene classifiers that use the SMN representation. The SMN representation obtained using the proposed method is found to be very effective. The CoNN is found to be better in learning the semantic contents of an image. However the training time of CoNN is an issue to be addressed. In the future the proposed approaches need to be validated using larger datasets such as MIT-67-Indoor dataset and Scene Understanding (SUN) dataset.

References

1. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, p. II. IEEE (2004)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
3. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
4. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC, vol. 2, p. 8 (2011)
5. Rasiwasia, N., Vasconcelos, N.: Holistic context models for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 902–917 (2012)
6. Perina, A., Cristani, M., Murino, V.: Learning natural scene categories by selective multi-scale feature extraction. Image Vis. Comput. **28**(6), 927–939 (2010)
7. Rasiwasia, N., Moreno, P.J., Vasconcelos, N.: Bridging the gap: query by semantic example. IEEE Trans. Multimed. **9**(5), 923–938 (2007)
8. Thenkanidiyoor, V., Chandra Sekhar, C.: Dynamic kernels based approaches to analysis of varying length patterns in speech and image processing tasks. In: Pattern Recognition and Big Data, p. 407 (2016)
9. Gupta, S., Dileep, A.D., Thenkanidiyoor, V.: The semantic multinomial representation of images obtained using dynamic kernel based pseudo-concept SVMs. In: National Conference on Communication (2017)
10. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2528–2535. IEEE (2010)
11. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. Int. J. Comput. Vis. **72**(2), 133–157 (2007)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)

13. Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step. In: Enser, P., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A.F., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 207–215. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27814-6_27
14. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems, pp. 487–495 (2014)
15. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Tras. Intell. Syst. Technol. **2**(3), 27 (2011)