ACM: Adaptive Cross-Modal Graph Convolutional Neural Networks for RGB-D Scene Recognition

Yuan Yuan, Zhitong Xiong, Qi Wang
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
{y.yuan1.ieee, xiongzhitong, crabwq}@gmail.com

Abstract

RGB image classification has achieved significant performance improvement with the resurge of deep convolutional neural networks. However, mono-modal deep models for RGB image still have several limitations when applied to RGB-D scene recognition. 1) Images for scene classification usually contain more than one typical object with flexible spatial distribution, so the object-level local features should also be considered in addition to global scene representation. 2) Multi-modal features in RGB-D scene classification are still underutilized. Simply combining these modal-specific features suffers from the semantic gaps between different modalities. 3) Most existing methods neglect the complex relationships among multiple modality features. Considering these limitations, this paper proposes an adaptive cross-modal (ACM) feature learning framework based on graph convolutional neural networks for RGB-D scene recognition. In order to make better use of the modal-specific cues, this approach mines the intra-modality relationships among the selected local features from one modality. To leverage the multi-modal knowledge more effectively, the proposed approach models the inter-modality relationships between two modalities through the cross-modal graph (CMG). We evaluate the proposed method on two public RGB-D scene classification datasets: SUN-RGBD and NYUD V2, and the proposed method achieves state-of-the-art performance.

Image classification has been researched for years, and excellent performance has been obtained with the development of deep learning methods. With the advent of large scale image dataset ImageNet (Russakovsky et al., 2015) and advanced graphics processing unit (GPU), more and more excellent deep learning architectures have been proposed, such as VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016) and DenseNet (Huang et al., 2017). Image classification is one of the basic tasks in computer vision research, and the main difficulty is how to learn effective image representations. As the learned image features can be applied to other high-level image analysis tasks, other computer vision research such as scene classification can benefit from the improvement of image recognition(Wang et al., 2018b). However, there is significant difference between scene classification and general image classification. For general image classification, the category of each image is highly related to the object in the image, as shown in Fig. 1. However, the category of the scene image is related to several typical objects and the spatial layout of the scene. Object-centric image usually contains one object, so the difference between images mostly relies on the objects. However, in order to classify a scene from another, we need recognize all the key objects in the scene and consider the relationships between them. Thus deep learning methods for common image classification are not suitable for scene classification. Considering this, to learn more robust and effective scene image representation, we propose to extract local object-level features and model their relationships for scene classification.

Figure 1: The difference between image classification and scene classification task. Images for scene classification are usually not object-centric.

Recently, depth sensors have been widely used in our daily life such as Microsoft Kinect. As RGB-D image can provide additional robust geometric cues which is not sensitive to illumination variability, more and more research focuses on the RGB-D image scene classification. Conventional RGB images only provide the appearance texture information and have difficulty in understanding the spatial layouts of complex scenes without the depth information. With the extra geometric information in RGB-D images, the performance of scene recognition can be improved promisingly. Although the depth information provided by RGB-D images can help to extract more discriminative features, how to make the extracted depth cues complementary to the RGB information is still a hard problem.

Exacting complementary information from RGB and depth modality effectively is a typical multi-modal feature learning problem. The method of (Wang et al., 2015b) proposes to minimize the distance of the RGB embedding with the depth representation using a correlation term on the loss function. However, merely enforcing the RGB and depth features to be correlated will make the model ignore the modal-specific information. Thus this method can not learn the modal complementary cues well. To get rid of this problem, the method proposed in (Li et al., 2018) constructs a fusion network for learning both distinctive and correlative information between two modalities. However, these methods neglect the relationships between the RGB features and depth features. The approach of (Song, Chen, and Jiang, 2017) proposes a framework to represent scene images with object-to-object representation for mining the relations and object co-occurrences in the scene. Nevertheless, it is a two-stage scene classification framework, so the scene classification accuracy is dominated by the object detection performance. Moreover, the computational complexity is also increased. To make better use of the complementary cues of the multiple modalities, this work designs the cross-modal graph (CMG) and exploits graph convolution to model the relationships between the RGB and depth modal features.

Considering these problems depicted above, this paper proposes a new RGB-D scene classification framework based on cross-modal graph convolutional neural networks. This framework exploits a two-stream CNN for modal-specific features extracting. Then cross-modal features are learned by the cross-modal graph convolutional neural networks (GCNs). Additionally, global RGB-specific and depth-specific scene features are also concatenated with the learned cross-modal representations for the final RGB-D scene classification. The contributions of this work can be summarized as follows.

- To embed the intra-modality object relations into the deep features, we propose a graph based CNN framework to mine the relations between local image features at different locations.
- To better learn the complementary features between the two modalities, the cross-modal graph GCNs is introduced to mine the inter-modality relations of RGB and depth modality considering the spectral and spatial aspects simultaneously.
- To learn more robust and discriminative representations for scene images, we fuse two kinds of global modal-specific features with the learned local cross-modal features together. Multi-task learning is used to simultaneously minimize three softmax loss functions.

Related Work

We review the related works in this section from three aspects: RGB-D scene classification methods, RGB-D multi-modal feature learning methods and neural networks on graphs.

RGB-D Scene Classification

Considerable efforts have been paid to scene classification research, and significant performance improvements have been obtained with the advent of large scale scene classification datasets and deep learning methods. Before the surge of deep convolutional neural networks (CNNs), handcrafted feature extraction is the mainstream method. Gupta et al. (2015) detect contours on depth images and extract local features from the segmentation outputs for scene classification. Bag of Words (BoW) features with spatial information is proposed by (Lazebnik, Schmid, and Ponce, 2006) for scene classification task. The work in Banica and Sminchisescu (2015) extracts local image features with second oder pooling for scene segmentation and classification.

As CNNs have achieved remarkable success in large scale image classification task (Krizhevsky, Sutskever, and Hinton, 2012), more and more recent scene recognition methods are based on CNNs. However, the dataset scale of RGB-D images is still not comparable to mono-modal RGB image datasets. Thus most deep learning based RGB-D scene classification methods rely on transferring pre-trained model weights on large scene dataset to relative small dataset. The work of (Zhou et al., 2014) finds that better performance can be achieved by training models on large scene dataset such as Places dataset (Zhou et al., 2017) than directly using models pre-trained on the object-centric dataset (ImageNet). To learn better CNN features, a multi-scale CNN framework is introduced in (Gong et al., 2014), which aggregates multiple scale features via vector of locally aggregated descriptors (VLAD).

More recently, many methods try to employ the semantic parts, i.e., features of objects or object parts in scene images as higher level representations for scene images. Dixit et al. (2015) propose to encode the scene images by combining features extracted from different locations and scales. A two-step learning framework is employed in (Song, Herranz, and Jiang, 2017), and the first step is weakly-supervised training on depth image patches. However, these image patches for feature encoding may contain noises, which affect the performance. Other methods employ object detection to extract higher-level semantic features. Wang et al. (2016) propose to encode the features of detected object proposals via fisher vector to learn component aware representations. The work of (Song, Chen, and Jiang, 2017) further models the object co-occurrences of scene images to gain better spatial layout information via object-to-object representation. However, these two-stage pipeline methods rely on the performance of object detection task. The error accumulation and computational complexity are problems remaining to be resolved.

RGB-D Multi-Modal Feature Learning

RGB-D scene classification is a typical multi-modal feature learning task. To fuse multiple modality features, considerable methods have been investigated(Wang et al., 2018a). Couprie et al. (2013) combine RGB and depth images by constructing the RGBD Laplacian pyramid, which fuses the two modalities at the image level. Other methods like (Song, Lichtenberg, and Xiao, 2015), employ two-stream CNN ar-

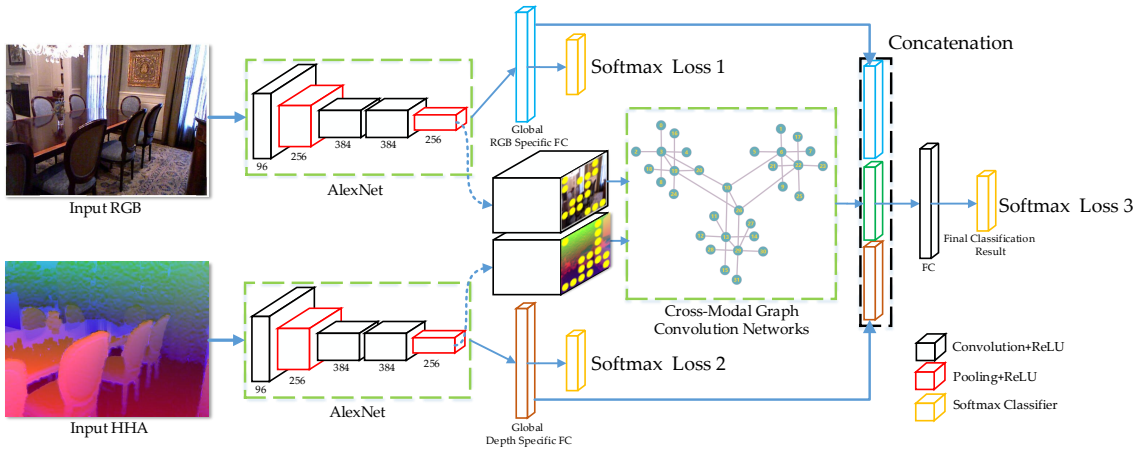

Figure 2: The whole network architecture of the proposed framework. The two-stream CNNs are exploited for feature learning. Global modal-specific features and local cross-modal features are concatenated together for scene classification.

chitecture and fuse the two modality features by concatenating two streams into one fully connected layer. Song, Jiang, and Herranz (2017) combine three stream features including RGB modality branch and two depth branches by element-wise summation. Different from previous work, Wang et al. (2015b), Zhu, Weibel, and Lu (2016) and Li et al. (2018) consider the relationships between two modalities. The work of (Wang et al., 2015a) enhances the modalities consistency by enforcing the network to learn common features between RGB and depth images. The method in (Li et al., 2018) aims to learn the correlative embeddings between the RGB and depth data by exploiting Canonical Correlation Analysis (CCA). However, different from these methods, the proposed method in this paper not only model the relationships of the intra-modalities and inter-modalities features together via graph convolution. As different modality contains unique features, we aim to learn the combination of the complementary cues from two modalities.

Graph Convolutional Neural Networks

Graph convolution is the generalization of traditional CNNs for non-structured data, such as the social networks, gen data and so on. As graph structure is a natural way for representing many kinds of signals, it is becoming a hotspot research area. The work of (Bruna et al., 2013) exploits spectral networks on graph for image classification, and (Defferrard, Bresson, and Vandergheynst, 2016) defined localized graph filters, which also decreases the computational complexity. Recent GCNs on graph structure data can be divided into two categories: spectral GCNs and spatial GCNs. Spectral GCNs mainly focus on spectral analysis, which define the convolution as linear transformation on the coefficients of Fourier basis, i.e., the eigenvectors of the Laplacian matrix. Spatial GCNs is more conceivable, as it provides localized filters similar to traditional CNNs. But spatial GCNs is harder to match local neighborhoods for each graph node, so it needs specific definition of neighborhoods (receptive field) and graph normalization. In this paper, as we mainly focus on mining the relationships of graph nodes, we opt to employ the spectral domain GCNs. Nevertheless, spatial information is also important for scene classification, inspired by the work of (Yan, Xiong, and Lin, 2018), we also take the spatial cues into consideration.

Methodology

The motivation of this work is that we human recognize scene categories mainly considering two aspects: 1) the global scene layout; 2) the key objects or object parts and their relationships. Inspired by this, in this work, we propose to learn the global modal-specific features and local object or object-parts level features simultaneously. Then the cross-modal graph is constructed to learn the relationships between the local multi-modal features by the GCNs. The whole framework is presented in Fig. 2. The RGB data and HHA encoded (Gupta et al., 2014) depth data are input to two stream CNNs for feature learning. As the final feature maps of each modality contain high-level semantic features, we adaptively select fixed number of feature vectors on high response locations for two modalities to construct the cross-modal graph. Meanwhile, each modality feature maps are connected to a modal-specific fully connected layer for global modal-specific feature learning. After the graph convolutions on the cross-modal graph, the learned cross-modal features and global modal-specific features are concatenated together for the final scene classification task. We will present the proposed method through the following three sections.

Single Modality Graph Construction

Observation reveals that the key objects or object parts are crucial for scene image representation. Considering this, we aim to encode the scene image with some important local component features. By selecting features of key objects and excluding the noise, the obtained image encoding can improve the classification accuracy. If we denote the input RGB data as x_{rgb} and the depth (HHA encoded) data as x_d. The weights of the two-stream CNNs can be represented by

$f_{W_{rgb}}$ and f_{W_d}. Then the final feature maps of two modalities is formulated as

$$F_{rgb} = f_{W_{rgb}}(x_{rgb}), \\ F_d = f_{W_d}(x_d), \tag{1}$$

where F_{rgb} and F_d are the final feature maps of RGB and depth modality through the two-stream CNNs. Then we construct the graph for each modality with these semantic features. We denote the graph of RGB and depth as G_{rgb} and G_d respectively. For the sake of simplicity, we introduce the graph construction for the RGB modality, which is similar to the depth modality.

The graph of RGB modality can be represented as $G_{rgb} = (V, E, A)$, where V denotes the nodes of the graph. The nodes of the graph are selected from F_{rgb}. E denotes the edges of the graph, and the adjacency matrix of the graph is A. The tensor shape of F_{rgb} is (N, C, H, W), where N is the batch size, C is the number of channels and H, W are the height and width of the feature maps respectively. In the scene images, only a small number of region proposals (key objects or object parts) contribute to the most discriminative features for scene classification. Thus we propose to adaptively select K highest response feature vectors for the graph nodes V.

To select K highest response feature vectors from F_{rgb}. We first sum the feature maps F_{rgb} along the channel axis as:

$$F_{re} = \sum_{i=1}^{C} F_{rgb}(N, i, H, W). \tag{2}$$

Then the response map F_{re} is reshaped to $(N, H * W)$. As two-dimension image contains natural spatial order, and this order is important to the spatial layout of the scene. Thus we keep this order for the K selected features. This procedure can be described by the following algorithm:

Algorithm 1 Algorithm for selecting K feature vectors.

Input: The reshaped response map F_{re};
Output: The indexes of K selected feature vectors ind_{sel};

1: Sort the N response maps F_{re} with descending order, and select the first K indexes to ind_F:
2: **for** $i = 0 \rightarrow N$ **do**
3: $\quad index = sort(F_{re}(i))$;
4: $\quad ind_F(i) = index(1 : K)$;
5: **end for**
6: Sort the N indexes ind_F with ascending order to keep the original image spatial order;
7: **for** $i = 0 \rightarrow N$ **do**
8: $\quad ind_{sel}(i) = sort(ind_F(i))$;
9: **end for**
10: **return** ind_{sel};

The graph nodes $V = \{v_i | i = 1, ..., K\}$ are assigned with the selected features from K different locations of F_{rgb}. Through algorithm 1 we can get the index of selected features. The graph nodes assignment can be formulated as $V = \{v_i = F_{rgb}(N, C, i) | i = ind_{sel}(1), ..., ind_{sel}(K)\}$,

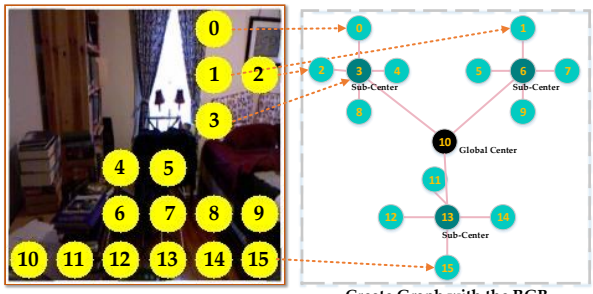

Figure 3: The selected important features for single-modality (RGB) graph constructing. It is a similar process for depth(HHA) graph construction.

where F_{rgb} is reshaped to $(N, C, H * W)$, and the shape of graph nodes V is (N, K, C).

As illustrated in Fig. 3, $K = 16$ nodes are used to construct the graph. To aggregate the local selected features progressively, the constructed graph is with one center node No.10 and three sub-center nodes No.3, No.6, and No.13. Every sub-center node is connected to its nearest four nodes, and the three sub-center nodes are connected to the global center node. Then the single modality graph is constructed for intra modality feature relationships learning.

Cross-Modal Graph Convolution

After the graph construction of the RGB G_{rgb} and the depth modality G_d, we further introduce the construction of the cross-modal graph G. As the cross-modal graph is the combination of two graphs, the vertexes of V is defined as:

$$V = V_{rgb} \cup V_d. \tag{3}$$

Thus the shape of V is $(N, C, 2K)$. As presented in Fig. 4, the yellow circles are the selected nodes of RGB and depth (HHA) for the same scene. We can see that the selected nodes for the two modalities are clearly different, which supports the assumption that features from different modalities provide specific cues for scene classification. To model the relationships between the two modalities, we connect the RGB graph nodes with the depth (HHA) graph nodes. As the center nodes and sub-center nodes aggregate the information of their adjacent nodes, connecting these high degree center nodes is efficient for different modalities information propagation. Considering this, we connect the RGB sub-centers node 3, node 6, and node 13 to the corresponding depth (HHA) sub-center nodes. The global center node of RGB (node 10) is also connected to the depth center node. We define the adjacency matrix of RGB modality as A_{rgb} and the depth adjacency matrix as A_d. The final adjacency matrix of the cross-modal graph should also include the cross-modal connections A_{cm}. Thus the final adjacency matrix is:

$$A = A_{rgb} + A_d + A_{cm}, \tag{4}$$

where $A \in \mathbb{R}^{(2K \times 2K)}$. Based on this definition, the cross-modal graph is constructed. Inspired by the work of (Yan,

Xiong, and Lin, 2018), we design our algorithm with spectral graph convolution and consider the spatial factor simultaneously.

Spectral graph convolution

The spectral graph convolution is operated in the Fourier domain. Given the cross-modal graph $G = (V, E, A)$, where V is the vertexes (multi-modal feature vectors), E represents edges of the graph and A is the adjacency matrix of the cross-modal graph. V is the graph nodes and $|V| = k$. An essential operator for spectral graph analysis is the Laplacian matrix L, which is defined by $L = D - A$, where $D \in \mathbb{R}^{k\times k}$ is the degree matrix, and $D_{ii} = \sum_j A_{ij}$. Then we can get the normalized Laplacian matrix by $L = I_k - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \in \mathbb{R}^{k\times k}$, where I_k is the identity matrix. The normalized Laplacian L is a symmetric positive semidefinite matrix, so its spectral decomposition can be represented as $L = U\Lambda U^T$. U is comprised of orthonormal eigenvectors $U = [u_1, ..., u_k] \in \mathbb{R}^{k\times k}$ and $\Lambda = diag([\lambda_1, ..., \lambda_k])$ is the combination of eigenvalues $\lambda \in \mathbb{R}^k$. Then the spectral convolution can be defined in the Fourier domain as:

$$y = \sigma(U g_\theta(\Lambda) U^T x), \tag{5}$$

where x,y are convolution input and output, $g\theta$ is the convolution filter and σ is the activation function. However, this spectral convolution has high computational complexity for large scale graph. Thus Hammond, Vandergheynst, and Gribonval (2011) propose to approximate the $g\theta(\Lambda)$ by m_{th} order Chebyshev polynomials $T_m(x)$ as:

$$\begin{aligned} g_\theta(\Lambda) &\approx \sum_{m=0}^{K} \theta_m T_m(\tilde{\Lambda}), \\ \tilde{\Lambda} &= \frac{2}{max(\lambda)}\Lambda - I_k. \end{aligned} \tag{6}$$

Kipf and Welling (2016) further limit the m to 1, and approximate the max eigen value to 2, i.e., $max(\lambda) = 2$. Then the simplified GCN can be expressed as:

$$Y = (D + I_k)^{-\frac{1}{2}}(A + I)(D + I_k)^{-\frac{1}{2}}X\Theta. \tag{7}$$

For graph convolution on the cross-modal graph G, the input $X \in \mathbb{R}^{k\times C}$ is the set of selected feature vectors with spatial order and the output $Y \in \mathbb{R}^{k\times F}$ is the learned features. Then the weights $\Theta \in \mathbb{R}^{C\times F}$ can be implemented by 2D convolution with the kernel size of 1×1 and output channels F. The normalized adjacency matrix A plus the self-connection is expressed by: $L_{norm} = (D + I_k)^{-\frac{1}{2}}(A + I)(D + I_k)^{-\frac{1}{2}}$. Thus GCN can be implemented by performing traditional 2D convolution on the input selected feature vectors and then multiplies L_{norm}.

As described above, we divide the nodes on the cross-modal graph into three types: the global center node, sub-center nodes and other nodes. We assume that the center nodes, sub-center nodes and other nodes are not equal important in the relationships modeling. Motivated by this, we partition all the nodes into three subsets and then the graph convolution in consideration of spatial factor can be defined as:

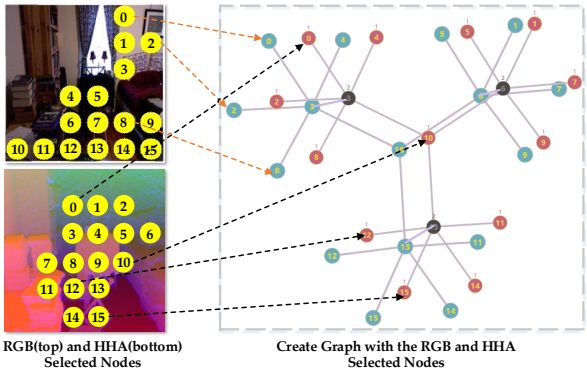

Figure 4: The cross-modal graph construction. Single modality graphs are constructed first by assigning selected features to graph nodes, then center and sub-center nodes are connected for two modalities.

$$Y = \sum_{j=1}^{3} L_{norm_j} X \Theta_j, \tag{8}$$

where $L_{norm\in\mathbb{R}^{k\times k}}$ is split into three matrices for the three groups of nodes connections. The first group contains the global center nodes, i.e., node 10 of two modalities, and their adjacency matrix is L_{norm_1}. The second group consists of the sub-center nodes, and the adjacency matrix is L_{norm_2}. The last split group is the collection of sub-center nodes and their neighboring nodes, whose adjacency matrix is L_{norm_3}. $\Theta_j \in \mathbb{R}^{C\times F}$ represents one of the three sets of weights for graph convolution.

Each Θ_j is implemented with the common convolution layer. Concretely, in this work we employ a 1×1 convolution layer with 256 channels, and then ReLU is applied as the activation function. Moreover, batch normalization and dropout with the keep probability of 0.5 are also utilized.

Global and Local Feature Fusion

Beyond mining the relationships of the selected local cross-modal features, we also consider the global modal-specific features for the final scene image representation, as illustrated in Fig. 2. Two global modal-specific fully connected (FC) layers are employed for the learning of global representations of RGB and depth modality respectively. We connect the RGB modality FC layer to the final feature maps F_{rgb}, and a classification softmax loss $L_{softmax_{rgb}}$ is exploited for training. Similarly, the depth modality FC layer is connected to the F_d and another softmax loss $L_{softmax_d}$ is used for training. Meanwhile, the two learned global features H_{rgb} and H_d are concatenated together with the cross-modal features H_{cm} learned by GCNs on cross-modal graph.

$$H = concat(H_{rgb}, H_d, H_{cm}). \tag{9}$$

Then the final concatenated features are input to another fully connected layer and softmax layer with loss $L_{softmax_{cm}}$

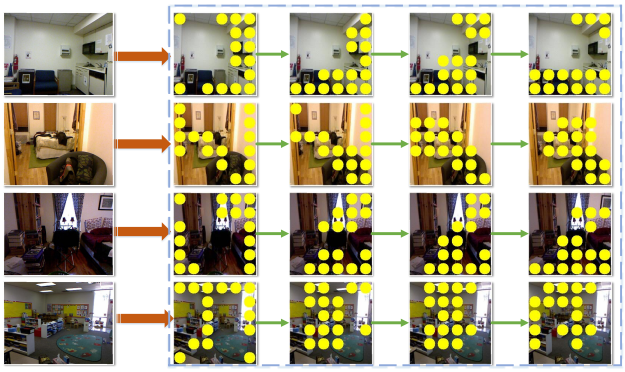

Figure 5: The visualization of selected important features for RGB modality. With more and more training iterations, the locations of selected features converge on objects or object parts.

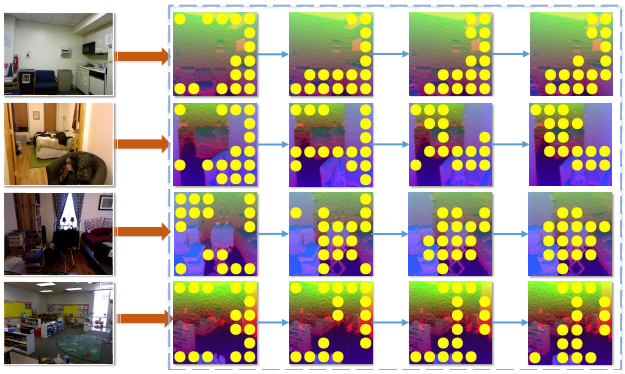

Figure 6: The visualization of selected important features for depth (HHA) modality. With more and more training iterations, the locations of selected features are different from the RGB selected features.

for scene classification. So the final loss L of this framework can be presented as:

$$L = \lambda_1 L_{softmax_{rgb}} + \lambda_2 L_{softmax_d} + \lambda_3 L_{softmax_{cm}}, \quad (10)$$

where λ_1, λ_2 and λ_3 are the balancing weights for the three loss components, and they are set to 1 in this work.

What is worth mentioning is that in the test phase, merely the final concatenated features are used to output the final classification result, as shown in Fig. 2.

Experiments

The proposed method is evaluated on two public RGB-D scene classification datasets: SUN RGB-D (Song, Lichtenberg, and Xiao, 2015) and NYU Depth Dataset version 2 (Nathan Silberman and Fergus, 2012). In this section, we will introduce the datasets and the parameters setup in detail. Moreover, we compare the proposed approach to other state-of-the-art methods and analyze the experiment results comprehensively.

Datasets

For RGB-D scene classification, there are mainly two popular datasets, one is the SUN RGB-D, and another is the NYU Depth Dataset version 2 (NYUD v2). The much larger dataset SUN RGB-D contains 10,355 RGB images with corresponding depth images captured from different camera sensors. To be correspondence with previous work, we only keep categories with more than 80 images. As with the experimental settings in (Song, Lichtenberg, and Xiao, 2015), there are 19 categories kept and 4,845 images for training, 4,659 images for testing.

NYUD v2 consists of 27 indoor categories and 1449 images in total, but many of the categories can not be presented well by these merely 1449 images. Thus Nathan Silberman and Fergus (2012) reorganized these 27 categories into 10 categories including 9 common indoor scene types and one "others" category. To compare our method with current state-of-the-art methods, we follow the dataset split settings in (Gupta, Arbelaez, and Malik, 2013). There are 795 images for training and 654 for testing.

Parameters Setup

The proposed method is implemented with the Pytorch (Paszke et al., 2017) deep learning framework. The HHA encodings are computed with the released code from (Gupta et al., 2014). For data augmentation, we resize the image pairs to 256×256 and random crop 224×224 as the input to the network. To compare with previous methods, we adopt AlexNet (Krizhevsky, Sutskever, and Hinton, 2012) as the back-bone network. Pre-trained models on Places scene classification dataset is used to initialize the network. For training parameters, the Adam (Kingma and Ba, 2014) optimizer is employed with initial learning rate 0.0001. The batch size for both datasets are set to 64 with shuffle.

Results and Comparisons

The results and analysis are presented for the two datasets in this section respectively.

SUN RGB-D dataset The comparing state-of-the-art methods on SUN RGB-D dataset including 6 methods. Song, Lichtenberg, and Xiao (2015) release the SUN RGB-D benchmark and use Places-CNN (Zhou et al., 2014) with RGB and HHA encoding as input for scene classification. Liao et al. (2016) employ a multi-task learning framework which combining scene classification and semantic segmentation tasks together. Zhu, Weibel, and Lu (2016) take the intra-class and inter-class correlations of image pairs for scene classification. Wang et al. (2016) propose component aware feature fusion framework by exploiting the region proposal component features. Similarly, Song, Chen, and Jiang (2017) further take the object-to-object relations into consideration. Li et al. (2018) present a discriminative fusion networks with structured loss. We use average precision over all scene classes for both datasets as evaluation metric.

From the results in Table 1, our approach achieves best accuracy 55.1% compared to other methods. Although the performance gain is not large compared to Li et al. (2018), our method do not employ the metric learning based training

Table 1: Comparison Results on SUN RGB-D Dataset

	Methods	Accuracy(%)
State-of-the-art	(Song, Lichtenberg, and Xiao, 2015)	39.0 %
	(Liao et al., 2016)	41.3%
	(Zhu, Weibel, and Lu, 2016)	41.5%
	(Wang et al., 2016)	48.1%
	(Song, Chen, and Jiang, 2017)	54.0%
	(Li et al., 2018)	54.6%
Proposed	Cross-Modal Graph (16 nodes)	**55.1%**

Table 2: Ablation Study on SUN RGB-D Dataset

Methods	Accuracy(%)
RGB	42.7%
Depth(HHA)	38.3%
RGB Graph	45.7%
RGB-D(HHA)	48.2%
RGB-D(HHA) Graph (16 nodes)	**55.1%**

loss used in (Li et al., 2018) and our method has no object detection stage as in (Song, Chen, and Jiang, 2017). Thus the proposed method has great potential for better performance. Additionally, we do ablation study for the proposed method as shown in Table 2. The performance of exploiting the single modality (RGB or depth) features is limited. As we can see from the results, the single modality graph modeling on RGB images named "RGB Graph" method improves 3.0% accuracy compared to the original "RGB" methods. By simply concatenating final layer features of RGB and HHA (RGB-D(HHA) method), the performance of scene classification can have a large improvement. At last, our cross-modal graph modeling on RGB-D images improves the baseline method "RGB-D(HHA)" by 6.9%.

NYUD v2 dataset We compare 5 state-of-the-art methods on NYUD v2 dataset. Some of the methods have been introduced in the comparison experiments on SUN RGB-D dataset. Gupta et al. (2015) propose to exploit both generic and class-specific features to encode the appearance and geometry of objects and used to classify scenes. Song, Herranz, and Jiang (2017) propose to learn depth features by combining local weakly supervised training from patches.

As shown in Table 3, the proposed method obtains state-of-the-art performance (mean class accuracy 67.2 %) on NYUD v2 dataset, outperforming existing methods. To better show what the proposed framework has learned, we visualize the locations of the selected feature vectors mapping to the RGB images and HHA images in Fig. 5 and Fig. 6.

Table 3: Comparison Results on NYUD v2 Dataset

	Methods	Accuracy(%)
State-of-the-art	(Gupta et al., 2015)	45.4 %
	(Wang et al., 2016)	63.9%
	(Li et al., 2018)	65.4%
	(Song, Herranz, and Jiang, 2017)	65.8%
	(Song, Chen, and Jiang, 2017)	66.9%
Proposed	Cross-Modal Ggraph (16 nodes)	**67.2%**

Table 4: Ablation Study on NYUD v2 Dataset

Methods	Accuracy(%)
RGB	53.2%
Depth(HHA)	51.1%
RGB Graph	55.4%
RGB-D(HHA)	61.1%
RGB-D(HHA) Graph (9 nodes)	**66.1%**
RGB-D(HHA) Graph (16 nodes)	**67.2%**
RGB-D(HHA) Graph (25 nodes)	**67.4%**

As shown in the Figures, the initial selected features are distributed randomly, but after a few iterations objects and object parts related locations are selected for graph convolution. Notably, there are clear differences in the final selected features for the corresponding RGB and depth (HHA) image pairs, which indicates that the features learned from two modalities are complementary for scene classification. Moreover, we also do ablation study for the proposed method, and the results are presented in Table 4. Similar to the results on SUN RGB-D dataset, the proposed approach enhances the baseline method "RGB-D(HHA)", which concatenates last layer features by 6.1% with $K = 16$. To evaluate the effect of parameter K, we also conduct experiments with $K = 9$ and $K = 25$ for comparison. As shown in Tab. 4, $K = 9$ performs worse than $K = 16$ by 1.1%. However, $K = 16$ and $K = 25$ achieve nearly the same performance, but $K = 25$ takes more computation cost. The reason of this phenomenon may be that 16 nodes are enough for describing the scene image. As for the computation cost, the average runtime of the feedforward is 0.0032 second with AlexNet and $K = 16$ on the Nvidia Titan X Pascal GPU.

Conclusion

In this paper, we introduce an adaptive cross-modal learning framework for RGB-D scene classification based on graph convolutional neural networks. This method adaptively selects important local features for each modality and constructs the cross-modal graph. Then graph convolution is exploited for local cross-modal feature relationships learning. Moreover, two global fully connected layers are employed for global modal-specific feature learning. Finally, the two global features and the learned cross-modal features are concatenated together for final scene classification. The experimental results on SUN RGB-D dataset and NYUD v2 have shown that the effectiveness of the proposed method.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant U1864204 and 61773316, State Key Program of National Natural Science Foundation of China under Grant 61632018, Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, Projects of Special Zone for National Defense Science and Technology Innovation, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and Open Research Fund of Key Laboratory of Spectral Imaging TechnologyChinese Academy of Sciences.

References

Banica, D., and Sminchisescu, C. 2015. Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3517–3526.

Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Couprie, C.; Farabet, C.; Najman, L.; and LeCun, Y. 2013. Indoor semantic using depth information. *arXiv preprint arXiv:1301.3572*.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, 3844–3852.

Dixit, M.; Chen, S.; Gao, D.; Rasiwasia, N.; and Vasconcelos, N. 2015. Scene classification with semantic fisher vectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2974–2983.

Gong, Y.; Wang, L.; Guo, R.; and Lazebnik, S. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, 392–407. Springer.

Gupta, S.; Arbelaez, P.; and Malik, J. 2013. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 564–571.

Gupta, S.; Girshick, R.; Arbeláez, P.; and Malik, J. 2014. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, 345–360. Springer.

Gupta, S.; Arbeláez, P.; Girshick, R.; and Malik, J. 2015. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision* 112(2):133–149.

Hammond, D. K.; Vandergheynst, P.; and Gribonval, R. 2011. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* 30(2):129–150.

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

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, volume 1, 3.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *null*, 2169–2178. IEEE.

Li, Y.; Zhang, J.; Cheng, Y.; Huang, K.; and Tan, T. 2018. Df^{2}net: Discriminative feature learning and fusion network for RGB-D indoor scene classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

Liao, Y.; Kodagoda, S.; Wang, Y.; Shi, L.; and Liu, Y. 2016. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, 2318–2325. IEEE.

Nathan Silberman, Derek Hoiem, P. K., and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, X.; Chen, C.; and Jiang, S. 2017. Rgb-d scene recognition with object-to-object relation. In *Proceedings of the 2017 ACM on Multimedia Conference*, 600–608. ACM.

Song, X.; Herranz, L.; and Jiang, S. 2017. Depth cnns for rgb-d scene recognition: Learning from scratch better than transferring from rgb-cnns. In *AAAI*, 4271–4277.

Song, X.; Jiang, S.; and Herranz, L. 2017. Combining models from multiple sources for rgb-d scene recognition. *IJCAI 2017, Melbourne, Australia* 4523–4529.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2015a. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 1125–1133.

Wang, A.; Lu, J.; Cai, J.; Cham, T.-J.; and Wang, G. 2015b. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Transactions on Multimedia* 17(11):1887–1898.

Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2016. Modality and component aware feature fusion for rgb-d scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5995–6004.

Wang, Q.; Chen, M.; Nie, F.; and Li, X. 2018a. Detecting coherent groups in crowd scenes by multiview clustering. *IEEE transactions on pattern analysis and machine intelligence*.

Wang, Q.; Liu, S.; Chanussot, J.; and Li, X. 2018b. Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* (99):1–13.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 487–495.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Zhu, H.; Weibel, J.-B.; and Lu, S. 2016. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2969–2976.