MUCH: MUtual Coupling enHancement of scene recognition and dense captioning

ID: 254

ABSTRACT

Due to the abstraction of scenes, comprehensive scene understanding requires semantic modeling in both global and local aspects. Scene recognition is usually researched from a global point of view, while dense captioning is typically studied for local regions. Previous works separately research on the modeling of scene recognition and dense captioning. In contrast, we propose a joint learning framework that benefits from the mutual coupling of scene recognition and dense captioning models. Generally, these two tasks are coupled through two steps, 1) fusing the supervision by considering the contexts between scene labels and local captions, and 2) jointly optimizing semantically symmetric LSTM models. Particularly, in order to balance bias between dense captioning and scene recognition, a scene adaptive non-maximum suppression (NMS) method is proposed to emphasize the scene related regions in region proposal procedure, and a region-wise and category-wise weighted pooling method is proposed to avoid over attention on particular regions in local to global pooling procedure. For the model training and evaluation, scene labels are manually annotated for Visual Genome database. The experimental results on Visual Genome show the effectiveness of the proposed method. Moreover, the proposed method also can improve previous CNN based works on public scene databases, such as MIT67 and SUN397.

KEYWORDS

Scene recognition, dense captioning, joint model, Visual Genome, annotation

ACM Reference Format:

. 2019. MUCH: MUtual Coupling enHancement of scene recognition and dense captioning: ID: 254. In *Proceedings of ACM Multimedia conference (Conference'19).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/****

1 INTRODUCTION

Typically, the scenes (e.g., coast, bedroom, and street) are abstract entities that consist of many less abstract regions such as objects (e.g., tables, chairs, and cars) or themes (e.g., water, and rock). The representations of scenes are essentially studied with local features. Classical approaches [16, 30, 34] exploit the handcrafted features (e.g., SIFT [21]), while recent works [5, 6, 11] exploit local (multi-scale) features of convolutional neural networks (CNN). In addition, some studies [13, 22] have shown that the object presences can be used as representations (such as object banks[18], topic models[8, 19], and discriminative parts [14]), which are particularly effective when the objects consistently appear in the related scenes. However, due to the limited category number of the common objects, object co-occurrences in different scenes are unavoidable in real world. Representing scenes with object presence may easily lead to the inter-class similarity to scenes with object co-occurrences, suffering the limitation of lacking discrimination. Thus in contrast to object presence, learning image representations with less information co-occurring between scenes is an important research to scene recognition.

On one hand, representing images with local captions provides the desired local information for (global) scene recognition. In addition to objects, local captions contain detailed contexts, such as themes, relatively relations, and attributes, which are more helpful to distinguish those scenes that are confused by the object co-occurrences. For instance in Fig. 1 (c-d), not only object ("train") is

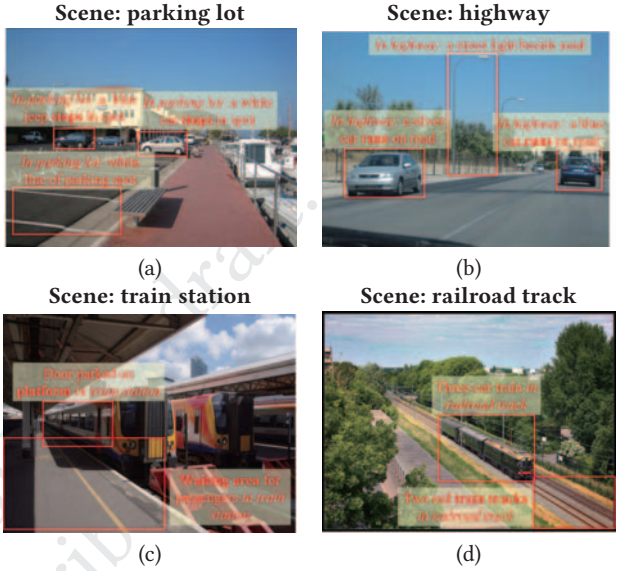


Figure 1: Some examples of understanding scene image with both scene labels and local annotations. Figure (a-b) illustrate that how scene labels can guide dense captioning, for instance cars usually stop in *parking lot*, but run on *highway*. And Figure (c-d) show that some particular local captions may be helpful to distinguish similar scenes, such as *train station* and *railroad track*, with co-occurrence of trains.

Conference'19, October 2019, Nice, France

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/****

detected, but also attributes, such as "*Three* car train", "*Two red* train tracks" and "*Waiting* area for passengers", and relatively relation, such as "Door parked at platform" are detected. Although such two images (of different scenes) contain co-occurrence of "train", other components of local captions like "platform" (in Fig. 1 (c)) and "tracks" in (in Fig. 1 (d)) are helpful to distinguish the scenes of those images. Thus, integrating scene labels after local captions like that in Fig. 1 (c-d) can benefit scene recognition from taking local captions as richer context information.

However, current caption decoding models [12, 37] focus more on describing regional objects or themes, which may lead to bias to scene recognition. For instance the duplicate annotations (of local captions) of regions is one limitation for scene recognition. In contrast to recognition task where one region usually corresponds to one exact category without heavily overlapping with other regions, some (spatially) closed regions may be multiply annotated as local captions from different aspects, such as attributes, relations and actions. This is caused by the bias between the goals of recognition and captioning. The former requires discriminative representations and classifiers, while the latter is more friendly to human understanding. Particularly, this duplicate problem may limit scene recognition in the following aspects: 1) multiple captions of the closed regions with semantic gap may confuse the scene classifiers; 2) too many duplicates may lead to over attention on some particular regions rather than the whole images. These limitations motivate us to propose region-wise and category-wise pooling method to address such bias problem.

On the other hand, recognizing scene labels provides global context for dense captioning.. For instance in Fig. 1 (a-b), when integrating the scene labels *parking lot* or *highway* before the local captions "A blue jeep stops in spot" or "a silver car runs on the road", the scene labels can be regarded as (globally) guiding information for training dense captioning model. Generally, in *parking lot*, a jeep probably stops there (see Fig. 1 (a)), but in *highway*, a car may run on road (see Fig. 1 (b)). Scene labels are the most abstract description of images, which may be used as global context to guide the selection of local captions (regions). Moreover, one limitation of current dense captioning is the lack of category label for region filtering, while object labels are included into the the process of non-maximum suppression (NMS) in object detection framework such as Faster R-CNN [28]. Thus, inspiring by category-wise NMS in Faster R-CNN and the semantically contextual relations between scene labels and local captions, a scene adaptive NMS is required for the dense captioning of scene images.

Scene recognition and dense captioning have potentials to complement each other. However, previous works separately model these two tasks, have not made attempts to simultaneously integrate them in a joint architecture. In this paper, we propose a joint framework with mutual coupling of both scene recognition and dense captioning (see Fig. 2), where two semantically symmetric long short term memory (LSTM) models are jointly trained with the supervision obtained by the contextual fusion of local captions and scene labels. Particularly, a scene adaptive NMS and a region- and category-wise weighted pooling are proposed to balance the bias between dense captioning and scene recognition. Scene recognition is to predict scene labels of the global images, requiring category discrimination to distinguish scenes, while dense captioning usually focuses on particular regions, mostly requiring understandable descriptions for regions. The scene adaptive NMS is proposed to reorder the regions proposals to emphasize the scene related regions, while filtering the unrelated and overlapped ones. Region-wise pooling is proposed to weighted regions with the area size, and category-wise pooling is proposed to measure category-wise union area, avoiding over attentions on the particular regions with duplicate proposals. For model training and evaluations, the scene labels are manually annotated for Visual Genome [15]. Besides Visual Genome, we also adapt our model to improve the performances of other public scene recognition databases, such as MIT67 [25] and SUN397 [36].

In addition, research on jointly training models of scene recognition and dense captioning requires annotations of both scene labels and local captions. Current databases either only contain scene labels, such as MIT67 indoor [25], SUN397 [36] and Places365 [39], or only contains annotation of local captions, such as Visual Genome [15]. Since scene labels are easier to be annotated, we organize the scene annotation on Visual Genome database, obtaining both scene labels and local captions with over 30K images, making it capable of training joint models of scene recognition and dense captioning. In addition to separately using two types of annotations, we particularly propose to integrate scene labels and local captions into more complete annotations. Particularly, scene labels and local captions are integrated in semantically symmetric ways. On one hand, for instance, the scene label such as *coast* is integrated into the start of local caption such as "An orange car on the sand", resulting in the new caption "*In coast* an orange car on the sand", where scene label can be used as (globally) guiding information for training dense captioning model. On the other hand, the scene label also can be integrated into the end of local caption, resulting in the new caption "An orange car on the sand *in coast*" , where local captions can be decoded as (locally) contextual features for training scene recognition model.

2 RELATED WORKS

2.1 Scene recognition

Conventional works [16, 30, 34] extract the handcrafted features (e.g., SIFT [21]) for scene recognition. Due to the semantic gap between the high-level (abstract) scenes and low-level visual features, a number of methods [4, 17, 18, 31, 33] try to use mid-level representations to shrink such gap. Vogel and Schiele [33] propose to use a vocabulary with nine mid-level categories to represent natural scenes. Object-bank [17, 38] is proposed as a type of mid-level representation that encodes the response of object classifiers at different positions in the images. Classemes [4, 31] are the intermediate representations that are extracted with a set of 2659 classifiers of mid-level categories. These methods require explicitly training corresponding mid-level classifiers with large scale preciously annotations of mid-level concepts. However, the low quality of annotations and classifiers limits the performance of those methods. More recent variants exploit discriminative parts, which are unknown and discovered during learning [2, 7, 14, 25]. The complexity of mining unknown patterns limits those methods from implementing on large scale databases.

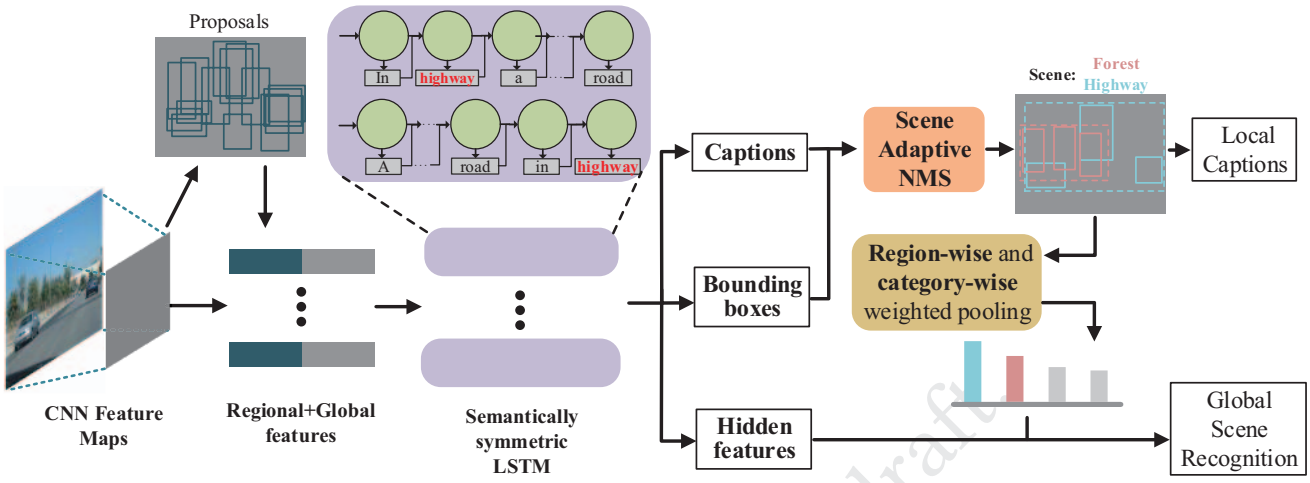


Figure 2: The framework of joint scene recognition and dense captioning.

With the exciting performance on object detection and recognition, CNN models are also exploited in recent works [5, 6, 11] to extract local (multi-scale) features. More recently, some works [3, 9, 35] also implement object detection techniques, including R-CNN [10] and Faster R-CNN [28] for scene recognition. George *et al.* [9] propose to represent scene images with the distributions of object presences (obtained by object detection technique), which are optimized to distinguish fine-grained scenes by semantic clustering. Bappy *et al.* [3] combine automatic object detection and manual annotation of objects in a framework of active learning, which is trained for scene recognition. Wang *et al.* [35] extract local features with R-CNN, and local features are embedded with Fisher Vector. Some studies [13, 22] have shown that consistent object presences are more helpful to facilitate scene recognition, however, which can hardly be guaranteed in real world.

2.2 Image captioning

The goal of image captioning is to generate descriptions for images. With the development of deep learning architectures, recent works of image captioning [23, 32] essentially follow the similar CNN-RNN pipeline, which is inspired from the frameworks of machine translation [1], where the sentence in the source language can be translated to a new sentence in the target language through a cascade of encoding and decoding processes. While in image captioning, the visual (CNN) features are particularly considered as the source language.

Recently, dense captioning models [12, 37] are proposed to generate local captions. The framework of dense captioning is inspired by object detection framework of Faster R-CNN [28], which first apply the region proposal network (RPN) to obtain dense proposals. Then a multi-loss layer, consisting of a loss layer of the coordinates of bounding boxes and a loss layer of object labels, is included for the further detection. In the framework of Densecap [12], the loss layer of object labels is replaced with the long short term memory (LSTM) model. Then Yang *et al.* propose to improve Densecap with

Table 1: Evaluation of NMS (of dense captioning model) on Visual Genome V1.4

NMS		mAP (%)
Order	IoU	
None	1	0.45
	0.4	1.12
Yes	1	5.39
	0.4	8.75

context features, which are the hidden CNN features extracted from the global image.

In addition to the implicitly semantic context of features, we propose to include the scene labels as the explicitly semantic context in this paper. And the proposed framework can also be used for scene recognition task.

3 JOINT MODEL OF SCENE RECOGNITION AND DENSE CAPTIONING

The proposed framework (see Fig. 2) is designed for simultaneously captioning regions of images and recognizing scenes, which is inspired by [37]. In addition to [37], the scene information is particularly included from different aspects in the proposed framework. First, the scene labels are manually integrated into the annotations of captions. Second, the proposed framework consists of two decoding (i.e., LSTM) models, one of them is guided by the scene information, and another is designed to predict local scenes by decoding the captions. Particularly for dense captioning, we also propose scene adaptive NMS to filter overlapped proposals.

3.1 Dense captioning guided by scenes

One branch of the decoder is designed for captioning regions with the guiding of scene information. Previously, captions of Visual Genome [15] are annotated to describe all salient components of

images without constraint nor guiding information. While with the context features as guiding information, the dense captioning model of [37] outperforms Densecap [12] in a large margin. In order to represent scenes, we propose to use explicit scene labels as guiding information, which are manually included at the beginning of each annotation of captions. For instance, "An orange car on the sand" is extended to "*In coast* an orange car on the sand ". Note that the first two words are constrained to be "*in scene*" with a separated vocabulary that is particularly designed for scenes.

Since LSTM model decodes features in a sequential order, including scene labels in the beginning of the annotations can maximize the influence of scenes, which is equivalent to guiding the process of decoding with scenes. For each image, guiding with such unified semantic information brings two types of benefits. On one hand, scenes can be regarded as the global context, which implicitly build the correlations between captions of each image, or captions of different images but annotated with the same scene category. On the other hand, scene labels increase the variety and richness of the local captions. For convenience, we denote this model as LSTM-C for the rest of paper.

3.2 Decoding captions for scene recognition

Another branch of the decoder is designed for scene recognition of regions. In contrast to the captioning branch, scene labels are included at the end of each annotations of captions. For instance, "An orange car on the sand" is extended to "An orange car on the sand *in coast*". Since the LSTM model decodes features to captions sequentially, the former captions (before the appearance of scene labels) are decoded as the obtained contexts for predicting scene labels (in the end). During training, all the annotated captions are integrated with "*in scene*". In order to emphasize the scenes, the last two words are constrained to be "*in scene*" by using a vocabulary that is separated from other captions. For convenience, we denote this model as LSTM-S for the rest of paper.

3.3 Semantic adaptive NMS

In previous works [12, 37], the overlapped regions are filtered by the non-maximum suppression (NMS) algorithm, which is inspired by the object detection framework of Faster R-CNN [28]. In the process of NMS, all the proposals are first ranked by the confidence score. Then the proposals with higher confidence score are used as references to filter the overlapped proposals with lower scores. Particularly, there are two stages of NMS, including the first NMS after region proposal network (RPN) and the second NMS after semantic regression (such as the regression of object labels in Faster R-CNN, and the LSTM decoding in dense captioning). Since LSTM decoding does not provide any confidence score, previous dense captioning frameworks lack of semantic driven NMS.

Note that NMS is key to obtaining good performance of dense captioning, particularly, a reasonable order is also important. The comparisons in Table 1 can illustrate the effective of NMS in dense captioning framework. NMS with "none order" means that all the proposals are in a random order, NMS with order means ranking proposals by the confidence score of RPN. Intersection-over-Union (IoU) is used to drop overlapped proposals with IoU larger than a threshold. IoU = 1 means not dropping anyone. Both ordering and dropping overlapped impact the performances. It can be observed in Table 1, even without dropping any overlapped proposals, only ordering the proposals with the RPN scores obtains the mAP of 5.39%, outperforming the strategy that includes dropping with "none order" (obtaining the mAP of 1.12%) in a large margin. Thus, a suitable order is more necessary in NMS.

We propose scene adaptive NMS for dense captioning, where scene labels (predicted by scene classifiers) are used as guiding information. First, the proposals of LSTM-C and LSTM-S are merged together, then a mixed score function is defined for ordering. The score function is as follow: $\sigma(p_i) = R(p_i) + \lambda M(s,k)$, where $R(p_i)$is the RPN score of proposal p_i, $s = [w_1, w_2, ..., w_K]$ is the corresponding softmax output (probability distribution) of scene label of each proposal p_i, K is the number of scene categories, $M()$ is matching score of scenes, k is the scene label predicted by using global CNN features. $M(s,k) = \begin{cases} w_k, & w_k = \max(s) \\ 0, & w_k < \max(s) \end{cases}$. All proposals are first ranked with $\sigma(p_i)$, then the overlapped proposals with lower score are dropped during NMS.

4 REGION-WISE AND CATEGORY-WISE WEIGHTED POOLING

The outputs of LSTM-S and LSTM-C in the proposed framework can be used for scene recognition. However, due to the bias between the goal of dense captioning and scene recognition, it requires some methods to balance them. Dense captioning model usually focuses on some salient objects or themes, and generates descriptions of them in high density. While over focusing on some particular regions may mislead the scene classifiers, since scene classifiers finally require a global point of view of the scene. In this section we first analysis the feasibility of recognizing scenes with local captions. In order to exclude the factors out of captions, the analysis is first evaluated with the annotation (groundtruth) of local captions with a joint model of LSTM and Multi-layer Perceptron (MLP). Then the weighted pooling methods are proposed to address the problem caused by the duplicate bounding boxes, aiming to balance the bias between dense captioning and scene recognition. Also, a comparison of using different types of caption based features are discussed, such as decoding hidden states, encoding hidden states or softmax output of the corresponding scene labels.

4.1 Analysis of local caption

Since Visual Genome database is proposed to describe the rich information of images, many local captions are annotated for the salient objects or regions. Particularly, some salient regions are annotated with duplicated bounding boxes of various content of captions. However, the recognition of scene labels are more reasonable to be predicted from a balanced view of the global images. Thus, there is a semantic bias between the tasks of annotating local captions (basically for salient regions) and scene recognition (mainly for global images). For instance, we illustrate several examples from Visual Genome database in Fig. 3 (best view in color). Note that Visual Genome does not contain scene labels of images, we organize the annotation of scene labels for the images from Visual Genome, the details of the annotated database are introduced in following subsection.

(a) scene: **canal** / bridge (c) scene: **car_interior** / bakery (e) scene: **highway** / airfield

(b) scene: **campsite** / pizzeria (d) scene: **downtown** / tower (f) scene: **mountain** / parking_lot

Figure 3: Some examples of Visual Genome. The local captions are shown on the top of bounding boxes. The bounding boxes in red color are correctly predicted (for scene labels), while the blue ones are wrongly predicted (best view in color version). The annotated and predicted scene labels are shown below the images (the annotated labels are in left and red color, the predicted ones are in right and blue color). Prediction is implemented by the joint model of LSTM and scene classifier. We organize the annotation of scene labels on this database.

The illustrated images in Fig. 3 (a-c) are some of the failed samples, i.e., the predicted labels are different to the manually annotated ones. It can be observed that the failure of the prediction is mainly caused by some particular regions (with wrong predictions), suggesting that the local results can affect the global results. For instance, the regions of *bridge* in Fig. 3 (a) confuses the classifier, misleading the classifier to make wrong predictions to *bridge* (the groundtruth is *canal*). Recognizing scenes depends on the global view of images, which requires to avoid the over attention on some particular regions that may bring the ambiguity. One intuitive way is to weight the regions with the size of area (region-wise) during the pooling from local to global. While for the more difficult situations, e.g., the samples in Fig. 3 (b-c), which consist many duplicate bounding boxes of the captions about some particular objects. Particularly the objects are not that discriminative to the scenes. In such situations, only including the region-wise weight may be not enough to address this confusing problem. An alternative way is to include the category-wise weight, which is based on the union area (of regions) of each scene category. Calculating the union area can also avoid the impacts of the duplicate bounding boxes.

4.2 Multiple types of weights in pooling

4.2.1 Region-level semantic descriptor. Each caption is encoded to obtain a probability vector $s = [w_1, w_2, ...w_K]$ with a joint model of LSTM and scene classifier (MLP), where w_k is the probability of corresponding category, s is the softmax output of scene classifier. The probability vector s is also denoted as semantic multinomial (SMN) [26, 27]. Note that the feature encoding model can be LSTM, also can be conventional models, such as term frequency/ inverse document frequency (TF/IDF) and word to vector (Word2Vec).

4.2.2 Region-wise weights. Since scene recognition requires making prediction from the global view of the images, we take the area size of the bounding boxes (of local captions) as a factor, which is used as region-wise weights during pooling. With such weighted pooling, the local captions describing large regions (more closed to the global images in size) are more important than the small ones, which may over focus on some particularly local objects. For each image I, the region-wise weights are defined as $G^r = [g_1^r, g_2^r, ..., g_N^r]$, N is number of local regions, $g_n^r = \sqrt{a^n/a^I}$ is the region-wise weight. a^n, a^I are the areas of region n and image I, respectively.

4.2.3 Category-wise weights. In order to avoid the over attention on some particular regions with duplicate bounding boxes, we introduce the category-wise weights, which are decided by calculating the size of union area (of regions) of each scene category. In one image, we first predict a scene label $p_n \in [1, 2, ...K]$ for local caption cap_n with the SMNs $s_n = [w_1, w_2, ...w_K]$ obtained

Table 2: Comparisons of different adaptive methods of scene recognition (the local caption are from the ground truth) in accuracy (%) on VG-13K (defined in subsection 5.1)

Embedding	Pooling	Accuracy (%)	
		SMN	SVM
TF/IDF	Ave	18.4	27.2
Word2Vec.	Ave	22.7	35.4
LSTM	Ave	37.3	46.6
	Region-wise	39.3	47.8
	Category-wise	38.5	47.3
	Category-region-wise	**40.0**	**48.7**
	Weighted voting	29.6	38.2

Table 3: Comparisons of dense captioning on VG-108K and VG-13K in mAP (%), * indicates our re-implementation

	Method	mAP (%)	
		VG-108K	VG-13K
Baseline	[37]*	8.75	8.84
Proposed	LSTM-C	8.70	12.33
	LSTM-N	8.71	**13.06**

LSTM-N: LSTM-C with scene adaptive NMS

by softmax of scene classifier, i.e. $p_n = \arg\max_k [w_1, w_2, ...w_K]$. Then for each category $k \in [1, 2, ...K]$, the category-wise area a_k is calculated by taking the union of all the regions that are predicted to the corresponding scene categories. The category-wise weight $G^c = [g_1^c, g_2^c ... g_K^c]$ is represented as the normalization of area a_k, i.e., $g_k^c = a_k / \sum_k a_k$. The category-wise weight G^c can be used to re-weight the the local SMNs, i.e., $s_n^c = [w_1, w_2, ...w_K] \odot [g_1^c, g_2^c ... g_K^c]$. Also note that, this weight G can be used as the features to feed the classifier, which is equivalent to the weighted voting method. Note that the region-wise and category-wise weights can be used together, which is denoted as category-region-wise weights.

4.3 Comparisons of weighted pooling

With the manually annotated local captions of Visual Genome, we compare different types of encoding and weighted pooling methods. The results of comparison are illustrated in Table 2. The local features are obtained with different embedding methods, e.g. TF/IDF, Word2Vector [24] and LSTM. Then the local features are fed to scene classifier (MLP model) to obtain local SMNs. And the local SMNs are aggregated with different pooling methods, including average pooling, multiple types of weighted pooling, including region-wise, category-wise and regional-category-wise weighted pooling, and weighted voting. Compared to different embedding methods, the LSTM model outperforms TF/IDF and Word2Vector with a large margin. Comparing with different pooling methods, pooling with category-region-wise weights obtains the best performances of 40.0%/48.7% with SMN/SVM. We take the scene with maximum probability of SMNs for prediction. The results of SVM are obtained by training SVM with SMNs.

5 EXPERIMENTS

5.1 Databases

In order to conduct experiments of scene recognition with local captions, we organize the annotation of scene labels on Visual Genome database[1]. We invite more than 20 workers to take part in annotating all the 108077 images of Visual Genome, which costs more than 300 man-hours. The vocabulary of scene categories follows the 365 scene categories of Places365 [39], which contains relatively complete vocabulary of scenes for recognition. In our setting, 100 categories with (relatively) more images are selected, ensuring that all the categories contains enough images (at least 130). Note that about 20K images are not scene-centric, which are ignored in our task. However, the long tail (unbalance distribution of the amount of images) problem may heavily affect the evaluation. The distribution of the annotated images of each category are illustrated in Fig. 4, where top 100 categories with (relatively) more images are selected. In our setting, we randomly select 26 images as test and 104 images as training for each category, i.e., the training/test split is 10400/2600. Note that we exclude these test images to train the dense captioning model. In order to separate our annotated dataset from the original one, we denote Visual Genome as VG-108K, and denote our dataset with scene labels as VG-13K. Scene labels are appended to local captions in VG-13K.

We also evaluate the proposed method with the scene datasets MIT67 indoor [25] and SUN397 [36]. MIT67 contains 15620 images of 67 indoor scene classes. SUN397 consists of 397 categories, with 108762 images in total. In the case of MIT67 Indoor and SUN397, the training/test configurations are provided by the original authors.

5.2 Evaluation for dense captioning

Our LSTM-C is extended from the dense captioning framework [37], which is compared as the baseline in Table 3. On VG-13K, the proposed methods outperform the baseline [37] in a large margin, illustrating the effectiveness of integrating scene labels for dense captioning. Note that for a fairer comparison on VG-13K, scene labels are removed for the evaluation of [37] (obtains 8.84% in mAP), otherwise, this method obtains even lower mAP 6.97% (not listed in table). Comparing to LSTM-C, LSTM-N (LSTM-C with scene adaptive NMS ($\lambda = 0.2$)) obtains better result of 0.73% in mAP, which suggests that the proposed scene NMS is more helpful to select desired regions.

5.3 Some insights of scene recognition with local captions

In contrast to using the annotated local captions (groundtruth) in Table 2, we also evaluate using decoded captions with LSTM-S model on VG-13K. The amount of the local captions used for evaluations are first evaluated. The evaluation of the caption amount is illustrated in Fig. 5, where Fig. 5 (a-b) show the region level accuracy and image level accuracy that are obtained by different models. EC represents encoding features with local captions generated by using [37]. EC-C represents encoding features with local captions generated by our LSTM-C. In each figure, the x axis represents the amount (#) of generated local captions that are used for scene recognition,

[1]We will release the scene annotation after the reviewing period.

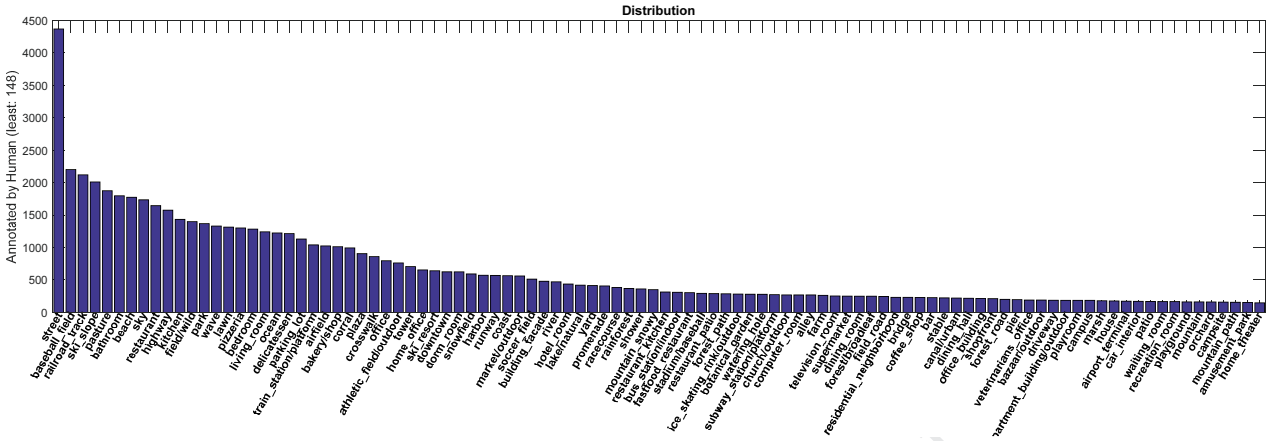
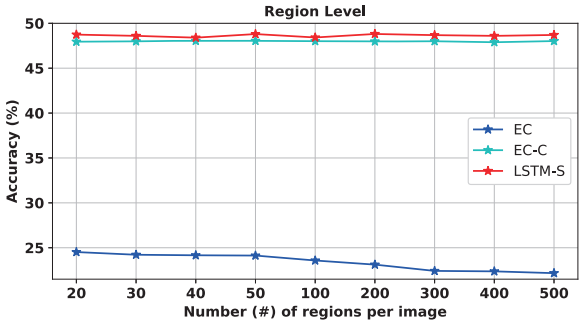
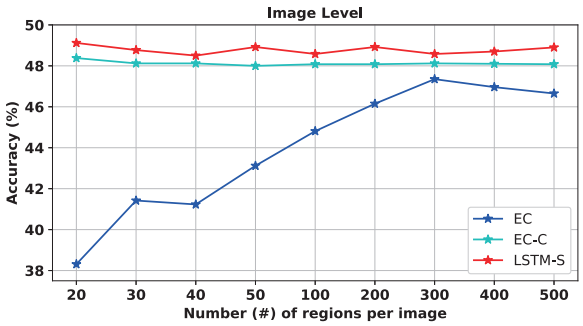


Figure 4: Number (#) of image of different categories.



(a)



(b)

Figure 5: Accuracy (%) of scene recognition with different amount of local captions on VG-13K. EC: encoded features of local captions generated by [37] (without scene labels), EC-C: encoded features of local captions generated by LSTM-C (with scene labels).

actually more than 500 local captions can be generated from each image. The resulted local captions (generated by the dense captioning model) are in order. Then the local regions, whose confidence score

Table 4: Accuracy (%) of scene recognition on VG-13K

Caption source	Method	Acc.
	Places(CNN)	58.2
Annotated	EC (Ave.)	46.6
	EC (CRW)	48.7
Generated	EC (Ave.)	47.3
	EC (CRW)	48.8
	LSTM-S (CRW)	49.6
	LSTM-S (CRW) + Places (CNN)	**60.2**

EC: encoded feature of local caption, Ave.: average pooling
CRW: category-region-wise weighted pooling

of captions are in top $L = [20, 30, 40, 50, 100, 200, 300, 400, 500]$ are selected for the evaluation of scene recognition.

5.3.1 Consistent vs inconsistent trends of region and image level accuracy. Observing from both Fig. 5 (a) and Fig. 5 (b), the two accuracy lines (caption and image level accuracy) of EC are in inconsistent trend, while the ones of the proposed LSTM-S are in consistent trend. Region and image level accuracy are measured over all the predictions (of scene labels) of local captions and images, respectively. The region level accuracy of EC obviously drops with the increasing amount of local captions, since the front captions are more confident. However, the image level accuracy improves with the increasing amount of local captions, which suggests that EC requires more information to make a correct prediction. Compared to EC, the proposed LSTM-S and EC-C obtain consistent results in both region and image level accuracy, outperform EC in a large margin in region level accuracy, and outperform EC (in image level accuracy) even with smaller number of local captions.

5.3.2 Generated vs annotated local captions (with EC). The results of encoding generated local captions (with LSTM-S) are shown in Table 4. Compared to using annotated local captions, using generated ones even works better (with a gain of 0.9% in accuracy). In order to further emphasize the differences between using annotated and generated local captions, we include another detailed analysis

Table 5: Comparisons of scene recognition on MIT67 and SUN397 in accuracy (%)

	Method	Accuracy (%)	
		MIT67	SUN397
Baseline	Places (CNN)	81.3	67.6
	ImageNet (MS-CNN)	75.8	60.8
Proposed	EC (CRW)	63.8	44.8
	LSTM-S (CRW)	64.9	45.8
	LSTM-S (CRW) + Places (CNN)	82.8	68.4
	LSTM-S (CRW) + Places (CNN) + ImageNet (MS-CNN)	**87.4**	**72.6**
State-of-the-art	Semantic-FV[5]	79.0	61.7
	Hybrid-CNN [39]	79.5	61.8
	Places (CNN) + ImageNet (MS-CNN)[6]	87.2	71.1
	Dual CNN-DL [20]	86.4	70.1

EC: encoded features of local caption, CRW: category-region-wise weighted pooling
MS-CNN: multi-scale CNN

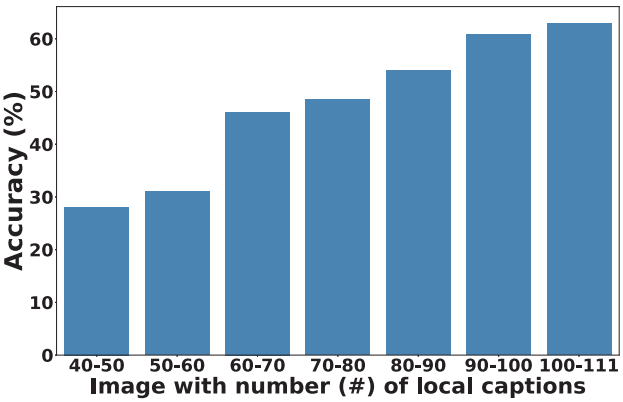


Figure 6: Accuracy (%, obtained by using EC) with annotated local captions VG-108K, distinguished by the number of local captions (e.g., the first bar represents the accuracy of images that contain 40 to 50 local captions).

of accuracy distribution (obtained by using EC) of using annotated captions that is categorized by the amount of local captions in Fig. 6. The results of Fig. 6 represent the accuracy of images that exactly contain corresponding number of local captions, e.g., the first bar represents the accuracy of images that contain 40 to 50 local captions. It can be observed that accuracy improves with the increasing of the number of local captions, where the accuracy of images with most local captions (last bar) is more than 30% higher the accuracy of images with least local captions. Thus, the lack of captions also limits the performance of using EC. Although the generated local captions seems not as confident as the annotated ones, after integrating scene labels into the annotations of local captions, the proposed LSTM-S model outperforms EC (with annotated captions). This may be caused by two reasons, one reason is that scene labels particularly adapt the training of LSTM-S to scene recognition (benefits LSMT-S); another reason is the lack of annotated local captions (in number) limits the performance of EC, compared to EC-C.

5.4 Comparison of the scene recognition

We also evaluate the proposed methods on the public scene databases MIT67 and SUN397, the results are illustrated in Table 5. By changing the scene vocabulary, LSTM-S can be directly used for those databases. The proposed LSTM-S is used to decode captions, where the hidden states are used as features to feed the scene classifiers. The proposed LSTM-S (CRW) outperforms EC (CRW) with about 1% in accuracy. Combining Places-CNN [40] features with the hidden state (features) of LSTM-S, i.e. LSTM-S (CRW) + Places (CNN), improves the Places-CNN models with 1.5%/0.8% on MIT67/SUN397. Note that state-of-the-art approaches [5, 6] also include the multi-scale features extracted from CNN pretrained with ImageNet [29] database. Combining the proposed features with Places (CNN) and ImageNet (MS-CNN), i.e., LSTM-S (CRW) + Places (CNN) + ImageNet (MS-CNN), outperforms [6] with 0.2% and 1.5% on MIT67/SUN397, and outperforms more recent works [20, 39] in larger margin.

6 CONCLUSION

Scene recognition and dense captioning are two types of image understand tasks, focusing on different scale of understanding. Scene recognition predicts scene labels from a global view of images, yet it also depends on local features, such as local visual features or intermediate representations. While dense captioning generates local descriptions of image regions, which also requires global features as contextual information. In contrast to separately researching these two types of tasks, in this paper, we propose a joint model of scene recognition and dense captioning, where the mutual effects are beneficial to enhancing each task simultaneously. Two tasks are integrated in two aspects, including the contextual fusion of annotations and the semantically symmetric fusion of the models. For annotation fusion, scene labels are integrated in the start of the annotations of local captions as the globally guided information for the training of dense captioning LSTM model, in contrast, local caption are integrated before scene labels for scene recognition. In the aspect of model fusion, two LSTM models are jointly trained for both tasks, where scene adaptive NMS and region- and category-wise weighted pooling are proposed to balance the bias between dense captioning and scene recognition.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* (2015).

[2] X. Bai, C. Yao, and W. Liu. 2016. Strokelets: A Learned Multi-Scale Mid-Level Representation for Scene Text Recognition. *IEEE Transactions on Image Processing* 25, 6 (June 2016), 2789–2802. DOI: https://doi.org/10.1109/TIP.2016.2555080

[3] Jawadul H. Bappy, Sujoy Paul, and Amit K. Roy-Chowdhury. 2016. *Online Adaptation for Joint Scene and Object Classification.* Springer International Publishing, Cham, 227–243.

[4] Alessandro Bergamo and Lorenzo Torresani. 2014. Classemes and Other Classifier-based Features for Efficient Object Categorization. In *IEEE Trans. on Pattern Anal. and Mach. Intell.*

[5] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos. 2015. Scene Classification with Semantic Fisher Vectors. In *CVPR*.

[6] Mandar D Dixit and Nuno Vasconcelos. 2016. Object based Scene Representations using Fisher Scores of Local Subspace Projections. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2811–2819.

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2013. Mid-level Visual Element Discovery as Discriminative Mode Seeking. In *NIPS*. 494–502.

[8] L. Fei-Fei and P. Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR*.

[9] Marian George, Mandar Dixit, Gábor Zogg, and Nuno Vasconcelos. 2016. *Semantic Clustering for Robust Fine-Grained Scene Recognition.* Springer International Publishing, Cham, 783–798.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*.

[12] A. Karpathy J. Johnson and L. Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Olivier R. Joubert, Guillaume A. Rousselet, Denis Fize, and Michèle Fabre-Thorpe. 2007. Processing scene context: Fast categorization and object interference. *Vision Research* 47, 26 (dec 2007), 3286–3297.

[14] Mayank Juneja, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. 2013. Blocks that Shout: Distinctive Parts for Scene Classification. In *CVPR*.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (01 May 2017), 32–73. DOI: https://doi.org/10.1007/s11263-016-0981-7

[16] S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

[17] L.J. Li, H.Su, E.P. Xing, and L. Fei-Fei. 2010. Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. In *NIPS*.

[18] Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. 2014. Object Bank: An Object-Level Image Representation for High-Level Visual Recognition. *Int. J. Comput. Vision* 107, 1 (2014), 20–39. DOI: https://doi.org/10.1007/s11263-013-0660-x

[19] Xin Li and Yuhong Guo. 2014. Latent Semantic Representation Learning for Scene Classification. In *ICML*.

[20] Yang Liu, Qingchao Chen, Wei Chen, and Ian J. Wassell. 2018. Dictionary Learning Inspired Deep Network for Scene Recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018.* https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16947

[21] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* 60 (2004), 91–110. Issue 2.

[22] Michael Mack and Thomas J Palmeri. 2010. Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of vision* 10 (03 2010), 11.1–11.

[23] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). (2015).

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[25] A. Quattoni and A. Torralba. 2009. Recognizing indoor scenes. In *CVPR*.

[26] N. Rasiwasia and N. Vasconcelos. 2007. Bridging the Gap: Query by Semantic Example. *IEEE Trans. on Multimedia* 9, 5 (2007), 923–938.

[27] N. Rasiwasia and N. Vasconcelos. 2012. Holistic Context Models for Visual Recognition. *IEEE Trans. on Pattern Anal. and Mach. Intell.* 34, 5 (2012), 902–917.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, AlexanderC. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* (2015), 1–42. DOI: https://doi.org/10.1007/s11263-015-0816-y

[30] J. Sanchez and F. Perronnin. 2011. High-Dimensional Signature Compression for Large-Scale Image Classification. In *Neural Comput.*

[31] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. 2010. Efficient Object Category Recognition Using Classemes. In *ECCV*.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*. 3156–3164.

[33] Julia Vogel and Bernt Schiele. 2007. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. *Int. J. Comput. Vision* 72, 2 (April 2007), 133–157. DOI: https://doi.org/10.1007/s11263-006-8614-1

[34] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong. 2010. Locality-constrained Linear Coding for image classification. In *CVPR*.

[35] Qilong Wang, Peihua Li, Wangmeng Zuo, and Lei Zhang. 2016. RAID-G: Robust Estimation of Approximate Infinite Dimensional Gaussian With Application to Material Recognition. In *CVPR*.

[36] J. Xiao, J. Hayes, K. Ehringer, A. Olivia, and A. Torralba. 2010. SUN database: Largescale scene recognition from Abbey to Zoo. In *CVPR*.

[37] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li). 2017. Dense Captioning with Joint Inference and Visual Context). In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*.

[38] Lei Zhang, Xiantong Zhen, and Ling Shao. 2014. Learning Object-to-Class Kernels for Scene Classification. *IEEE Trans. on Image Process.* 23, 8 (Aug 2014), 3241–3253.

[39] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1452–1464. DOI: https://doi.org/10.1109/TPAMI.2017.2723009

[40] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger (Eds.). 487–495.