Categorizing Concepts with Basic Level for Vision-to-Language

Hanzhang Wang Hanli Wang * Kaisheng Xu
Department of Computer Science and Technology, Tongji University, Shanghai, P. R. China
Key Laboratory of Embedded System and Service Computing, Ministry of Education,
Tongji University, Shanghai, P. R. China

Abstract

Vision-to-language tasks require a unified semantic understanding of visual content. However, the information contained in image/video is essentially ambiguous on two perspectives manifested on the diverse understanding among different persons and the various understanding grains even for the same person. Inspired by the basic level in early cognition, a Basic Concept (BaC) category is proposed in this work that contains both consensus and proper level of visual content to help neural network tackle the above problems. Specifically, a salient concept category is firstly generated by intersecting the labels of ImageNet and the vocabulary of MSCOCO dataset. Then, according to the observation from human early cognition that children make fewer mistakes on the basic level, the salient category is further refined by clustering concepts with a defined confusion degree which measures the difficulty for convolutional neural network to distinguish class pairs. Finally, a pre-trained model based on GoogLeNet is produced with the proposed BaC category of 1,372 concept classes. To verify the effectiveness of the proposed categorizing method for vision-to-language tasks, two kinds of experiments are performed including image captioning and visual question answering with the benchmark datasets of MSCOCO, Flickr30k and COCO-QA. The experimental results demonstrate that the representations derived from the cognition-inspired BaC category promote representation learning of neural networks on vision-to-language tasks, and a performance improvement is gained without modifying standard models.

1. Introduction

For human being, basic level [39] is an abstraction that plays a primary role in the form of early cognition. It contains several objects that are firstly perceived [39] and faster being named by children, then other subordinate concepts are learned afterwards [45]. In the taxonomic hierarchy, basic level is located on the middle which represents mid-level generality (*e.g.*, bird, chair), oppose to more general concepts (*e.g.*, vertebrate, furniture) and more specific concepts (*e.g.*, sparrow, recliner). It also has been found that children make fewer mistakes at basic level on matching and sorting tasks [41, 4] as well as simple decision making on adults. The annotation of ESP dataset [48] is accomplished through an online game which demands two players independently propose labels to one image for matching as many words as possible in a certain time limit. It is demonstrated that humans tend to label visual objects at an easily accessible semantic level termed as the basic level [3]. Similar observations can also be obtained on other human annotated datasets such as the image captioning dataset MSCOCO [24] which limits each description in 16 words. This limitation on time and length will impel people to use simple but cognitive useful words for expression.

Vision-to-Language (V2L) includes high-level semantic tasks which demand automatically transferring visual information into human language. Recently, a number of approaches [7, 14, 47] treat V2L as translation tasks using Convolutional Neural Networks (CNN) as encoder while employing Long Short Term Memory (LSTM) [9] as decoder to directly translate an image to a sentence. In the predominant neural network based approaches, it seems that each partial network has an explicit objective function to learn, however the real evaluation of the generated results is highly subjective which makes V2L challenging both on realization and evaluation.

The main problems which V2L faces to are two folds. First, the understanding of one image varies for different people. The diversity of thinking habit, attention and even expression leads to different explanations of visual content. In the MSCOCO dataset, each image is associated with 5

[*] Corresponding author (H. Wang, E-mail: hanliwang@tongji.edu.cn).

This work was supported in part by National Natural Science Foundation of China under Grants 61622115 and 61472281, Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. GZ2015005), Shanghai Engineering Research Center of Industrial Vision Perception & Intelligent Computing (17DZ2251600), and IBM Shared University Research Awards Program.

captions which are all correct on content but various on expression. This problem makes it difficult to reach a cognitive consensus for same visual content among different persons. Second, for certain person, the understanding of one image can be varied on multiple levels which means it can be very specific (dense caption [13], fine-grained classification [16]) or very general (scene classification [2, 51, 22]). However, for most V2L tasks, especially for image captioning, it only needs a middle level overview of the content. Thus, a proper level of understanding is vital for efficient representation in V2L.

For human being, these two problems can be largely eliminated by the basic level. Consequently, we are wondering, for neural networks, whether there exists a level which resembles the basic level to possess the ubiquity and generality of visual representations? Furthermore, could the V2L performance be improved by adopting such cognition-inspired mechanisms? Specifically, a salient concept (*SaC*) category is firstly proposed that contains candidate basic level concepts by matching ImageNet [40] classes with image captions from MSCOCO. This operation provides a set of cognitive words which are located in the middle layers of the ImageNet/WordNet [32] hierarchy. Then, these salient concepts are refined by clustering according to the observation from human early cognition that children make fewer mistakes on the basic level. A confusion degree is defined to measure the difficulty for CNN to distinguish class pairs, and the concepts with large confusion degrees are merged to minimize the CNN classification error. After that, the Basic Concept (BaC) category is generated and visualized by examples as compared to the basic level in human cognition. Finally, the BaC level is used as a semantic representation between vision model and language model in neural network based approaches to test its efficiency on image captioning and Visual Question Answering (VQA) tasks.

The main contributions of this work are summarized as follows. First, the BaC category is designed to resemble the basic level in human cognition, which not only summarizes the consensus among different people but also provides a proper mid-level understanding for V2L. Meanwhile, the proposed categorization procedure provides a method of automatically generating the basic level for neural networks. Second, the proposed BaC level is applied as an optimized semantic representation that connects CNN and LSTM for V2L. The GoogLeNet [43] model is trained with the BaC category and the corresponding pre-trained model is provided. A significant performance improvement has been achieved by the proposed categorizing method as compared to the baselines. For image captioning, the performances in terms of CIDEr [46] and BLEU-4 [36] are improved by 7.7 and 2.1 on the MSCOCO dataset over the baseline. For VQA, the performances have been improved by 2% to 4% against the baseline models.

2. Related Work

V2L tasks attempt to bridge the gap between vision and natural language and thus enable universal artificial intelligence to some extent. As compared with other computer vision tasks, V2L tasks such as image captioning [14] and VQA [27] are more challenging because they require an integrated understanding of visual representation, semantics and natural language.

The purpose of image captioning is to automatically generate natural language descriptions that describe the main content for a given image. Image captioning has achieved significant successes in recent years and there are several solutions to this task. An intuitive technique is to use template based methods [17, 8], which detects objects, attributes, actions and scenes and then puts them into a fixed template. However, the generated captions are usually rigid and restricted on expression. Another solution is based on image retrieval [18, 19], which generates captions by transferring the corresponding descriptions from the retrieved images. As far as VQA is concerned, the task is to provide an accurate natural language answer given an image and a natural language question with free-form about the image content. As compared to image captioning, VQA needs a relatively detailed understanding rather than generic descriptions [1]. Early researches treat VQA as a Turing test proxy such as [27] which generates the answers by the combination of semantic parsing and image scene analysis in a Bayesian framework. More recently, similar to image captioning, end-to-end deep neural networks are employed for VQA [28, 37].

Currently, the combination of CNN and Recurrent Neural Network (RNN) is the predominant framework to achieve V2L. Generally, CNN is adopted as the encoder to extract visual features and RNN as the decoder to generate sentences [14, 47]. Between the encoder and the decoder, the CNN features are projected into the same representation space as word embedding to realize mapping from vision to language. In [47, 7], the features of the penultimate fully connected layer are employed for visual representation. Moreover, attention based models [50, 25] utilize the last convolutional layer to obtain spatial information so as to enable the model to adapt the focus on part of the image during caption generation. These researches prefer to bypassing the elaborate mapping from vision to language.

On the other hand, one class of studies focus on producing more human-like recognition results. The majority of these works are naming an image/object by the basic-level concept rather than specific variety in common classification tasks. In [5], high-level concepts are utilized to balance the accuracy and specificity of classification model if the accuracy is too low. Different from the proposed work, these high-level concepts are designed according to WordNet. Similarly, the WordNet structure is used in [30] to find appropriate basic words for image context instead of generating concept levels. In [33], the entry-level name is predicted which people are likely to call from the specific categories located at the leaf nodes in WordNet. Similar to [30], the entry-level name also takes WordNet structure, word frequency and image classification outputs into account. All these works rely on the WordNet structure. However, the limitation of WordNet is noticed in [34], and it is believed that visually similar objects are also linked in semantic scenarios and thus visual similarity is introduced into basic level extraction.

Furthermore, it is highlighted in [35] that the number of samples in each category can greatly affect the performance of pre-train models. Consequently, it suggests to uniform the sample number across classes. For a similar reason, the full ImageNet dataset is combined to 4K, 8K and 13K categories in [31] according to the sample number of each class in WordNet. Trained on this combined dataset, CNNs obtain a better performance than pre-trained models for V2L tasks. The human-categorization knowledge is introduced to CNN learning which benefits both classification and V2L tasks in [44].

3. Proposed Basic Concept Category

3.1. Categorizing Salient Concepts

Since basic level concepts are objects frequently appeared in daily lives so that children can easily access to them during their early learning. It is desired to narrow the range of basic concepts into a smaller semantic concept set which frequently appears in both vision and language. To this aim, the benchmark image captioning dataset MSCOCO [24] is employed, which contains 123,287 images and each image is associated with 5 reference sentences describing the corresponding image content. Given an object concept, its most common and visually sensitive substances are included.

Considering that MSCOCO provides only 91 stuffs and obviously it cannot cover as many salient concepts as needed, these concepts in image captions are aligned directly with the annotations in the large-scale image dataset of ImageNet [40]. ImageNet contains 21,841 classes and 14,197,122 images in total, and it is tree structured according to the WordNet organization, which ensures ImageNet has both visual and semantic hierarchical characteristics. Every ImageNet category is corresponding to a WordNet entry which provides rich form and synonyms.

First, the MSCOCO captions are split into words and the frequency of each word is counted. Then, the words with less than 5 occurrences are treated as insignificant and removed. After removing insignificant words, a vocabulary with 9,566 words is constructed and any word outside this vocabulary is replaced by a special "unknown keyword". Afterwards, these filtered 9,566 words are matched to ImageNet annotations and each match is considered as an extracted salient concept. Finally, a Salient Concept (*SaC*) Category is generated which contains 1,689 concepts. Therefore, a rough set of objects is extracted which contains salient concepts from human annotated datasets.

3.2. Clustering Concepts by Confusion Degree

As mentioned above, the proposed Basic Concept (*BaC*) Category is defined as the superset of basic level concepts. In a semantic hierarchy, the width of basic level is one, then the width of BaC is equal to or greater than one, so BaC is a band that enfolds the basic level. Rather than choosing one strip in multiple layers, it is preferred to aggregate these concepts into a single one. Since basic level is a category with which children make fewer mistakes [41, 4], it inspires us to merge classes by the "difficulty" for CNNs to distinguish one from another. Specifically, a CNN model is firstly trained until its classification accuracy reaches a bottleneck, then the "difficulty" is measured according to the error rate of each class. Let $\mathcal{S}_{ij}$ be the number of images misclassified from class i to class j, for all classes the misclassification among each class-pair can be organized by a confusion matrix $\mathcal{S}$.

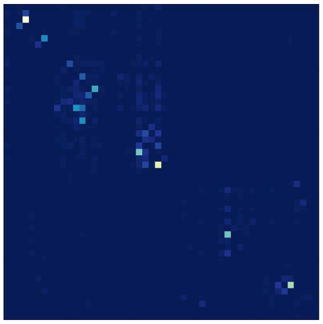


Figure 1. Illustration of confusion matrix $\mathcal{S}$. The brightness of element ij indicates the number of images belonging to class i being misclassified to class j. The diagonal is removed for better view.

The confusion matrix $\mathcal{S}$ is organized by each row and visualized in Fig. 1. The long horizontal light indicates the target class that is very prone to be misclassified while the long vertical light means that the corresponding class is visually similar to many other classes. As seen in Fig. 1, the following observations can be made. First, most confusing classes are gathered into groups rather than scattered separately. Visually similar objects are close to each other in semantic distance. Second, there are more vertical clusters than horizontal clusters. This phenomenon is due to the presence of the container class in the category. Container classes are usually located at the higher layer of ImageNet hierarchy and contain several sub-classes, *e.g.*, "cat" contains "Egyptian cat", "Persian cat" and "kitty", and subclasses are more confused than the corresponding container class. Third, there exist several clusters in the same row/column. This is because some classes that are far away in WordNet may share similar visual characteristics.

In order to measure how difficult a class-pair ij can be distinguished by a certain CNN model, the confusion degree ϕ_{ij} is defined as

$$\phi_{ij} = \frac{\mathcal{S}_{ij}\mathcal{S}_{ji} - \mathcal{S}_{ii}\mathcal{S}_{jj}}{\sqrt{(\mathcal{S}_{ii}+\mathcal{S}_{ij})(\mathcal{S}_{ij}+\mathcal{S}_{jj})(\mathcal{S}_{jj}+\mathcal{S}_{ji})(\mathcal{S}_{ji}+\mathcal{S}_{ii})}}. \tag{1}$$

The smaller the confusion degree is, the easier for CNN to distinguish a class-pair. This operation also transfers a asymmetric confusion matrix $\mathcal{S}$ to a symmetric similarity matrix Φ. It is common that the classes which are similar to each other have larger ϕ. It is a graph clustering problem to merge these classes. To achieve this, the graph-based distance d_{ij} between class i and j is defined as

$$d_{ij} = \max_{p_{ij}\in P_{ij}} \min_{e_{mn}\in p_{ij}} \phi_{mn}, \tag{2}$$

where P_{ij} indicates all paths from i to j, and e_{mn} is an edge in p_{ij} from m to n. The proposed BaC Category can be obtained by simply merging class i and class j once d_{ij} is larger than a given threshold ε.

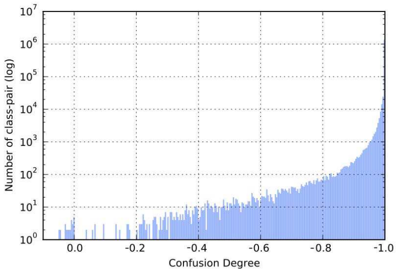
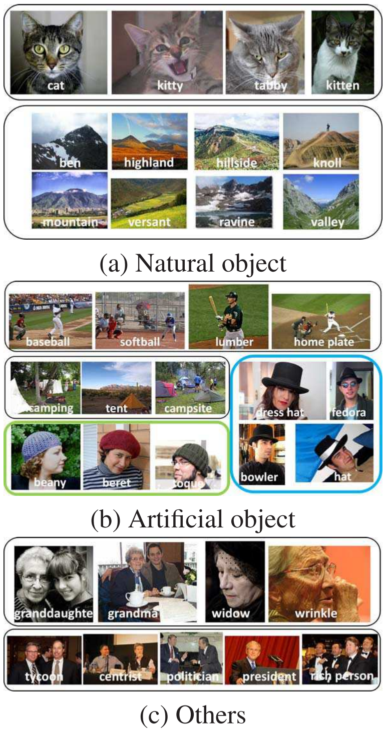


Figure 2. Histogram of confusion degree. It shows the number (in log) of confusion degree for each class pair in SaC.

To decide the value of ε, the statistics of confusion degree are shown in Fig. 2. For most class pairs, their confusion degree range from -0.95 to -1, which demonstrates that CNN can easily distinguish them. However, for several class pairs, CNN can hardly perform well and the distribution of prone-confused classes is generally uniform. It is worth noting that when ε gets smaller its effect on cluster number will be weak since most class pairs own small confusion degrees. The selection of ε is quite robust from -0.8 to -0.95. In this work, ε is empirically set to -0.9 and finally the BaC category is obtained with 1,372 concept clusters.

3.3. Comparing BaC with Cognitive Basic Level

The effect of the proposed BaC Category is visualized in Fig. 3 with several toy examples divided into three clusters: natural object, artificial object and others. The examples of natural object show that different varieties belong to the same species with similar appearance. Even for the same scenario, various attention will make completely different understandings such as mountain and valley. The examples of artificial object, as the large proportion in the BaC category, mainly include activities and human created stuffs. The former contains several usually concurrent objects and actions whereas the latter is often various in shape but appears in same scenes. The last group represents the examples with social attributes which mainly contain high level semantic inference. All of these examples manifest that the objects inner one cluster share very general characteristics although their labels vary.



(a) Natural object

(b) Artificial object

(c) Others

Figure 3. Toy examples of the proposed BaC Category, including three clusters: (a) natural object, (b) artificial object, and (c) others with latent social attributes.

Moreover, some partial classes and the corresponding hierarchy related to head dress in WordNet/ImageNet are illustrated in Fig. 4, where the words in circle represent the classes in the SaC Category (1,689 classes in total) and the words in square indicate the nodes constructed by the hierarchy but not in the SaC Category. Besides, the words in the same color represent that the corresponding classes belong to one cluster in the BaC Category (1,372 classes in total). It is observed that beret, beany and toque have very similar appearances, however they are located on distinct depths and cross with other clusters in the hierarchy. The visual similarity among such classes can hardly be discovered only by the distance in WordNet. Apparently, visually similar classes may cross the lexical relations among categories. Grouping clusters cannot be accomplished by rolling classes between direct superclass and subclass, or simply selecting a certain layer where all deeper subclasses could be aggregated on. It is also mentioned in [34] that the encyclopedic knowledge in WordNet does not coincide with the expected organization of common sense knowledge.

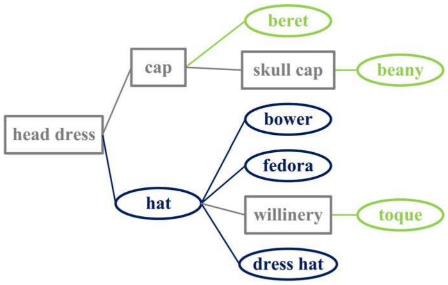


Figure 4. Hierarchical structure of head dress in WordNet. The words in circle represent the classes in the SaC Category and the words in square indicate the nodes constructed by the hierarchy but not in the SaC Category. The words in the same color represent that the corresponding classes belong to one cluster in the BaC Category.

Based on the aforementioned analysis, the following observations can be made. First, the natural objects in BaC are very similar to the basic level objects in human cognition. This infers that human categorizes natural objects essentially depending on shape and texture which are features that CNNs are good at learning. Second, the artifacts in BaC are located relatively deeper in the hierarchy compared with the basic level objects. In other words, it is much difficult for CNNs to summarize the generality among artificial objects as humans do, so it inclines to make more clusters on a more specific level. There are two possible explanations about this phenomenon. One is the insufficiency of training data, and another is that people categorize artifacts according to function rather than appearance. Consequently, it is not easy for CNN to infer the latent information for categorization.

4. Experimental Results

Since the BaC level is supposed as a proper level of cognitive consensus, it is employed as an optimized representation between CNN and LSTM to verify its efficiency for V2L tasks.

4.1. Image Captioning

Regarding image captioning, the benchmark datasets of MSCOCO [24] and Flickr30k [10] are employed. As mentioned above, MSCOCO contains 123,287 images, and each of them is annotated with 5 captions. The widely adopted split method [14] is applied with 5,000 images for both validation and testing. On the other hand, the dataset of Flickr30k contains 31,783 images and 5 captions for each image. For Flickr30k, 29,000 images are used for training, 1,000 images are used for test, and the rest are for validation.

As far as the evaluation metrics for image captioning are concerned, the following widely used evaluation metrics are used for performance presentation, including BLEU [36], METEOR [20], ROUGE_L [23] and CIDEr [46] which are denoted as 'B', 'M', 'R' and 'C' for short. About the CNN model for visual representation, GoogLeNet [43] is adopted to encode visual features. First, the GoogLeNet based CNN model is trained with the SaC/BaC Category by fine-tuning from the 4K pre-trained model [31] for convenience. Then, the model parameters are transferred to the caption task with the following two methods: (1) **OFC** (Original Fully Connected) method, in which the last fully connected layer (*i.e.*, classification layer) of the pre-trained model is extracted to be the input to the LSTM network. (2) **NIC** (Neural Image Captioning) method, which is a commonly used image captioning baseline model [47] with an image embedding layer between CNN and LSTM. The caption model follows the default setup of NIC which uses LSTM with 512 memory cells and 512 dimensions of word embedding. The training strategy is also the same as NIC with ADAM optimizer, and the learning rate is 4×10^{-4}, α is 0.8, β is 0.999, and ϵ is 10^{-8}.

First, the comparisons of different categorizing methods with the same CNN model are demonstrated in Table 1, Table 2, Table 3 and Table 4. In these tables, the ILSVRC-1K method [40] is provided by the winner model GoogLeNet with 1,000 official categories of ILSVRC 2012. The methods of Shuffle-4K [31] and Shuffle-8K [31] reorganize the full ImageNet (*i.e.*, 21,841 classes) aiming to solve the data imbalance issue with a subset of classes. The MuCaLeNet [44] contains 480 basic categories from ImageNet-1K (*i.e.*, 1,000 classes) with human annotation. The comparisons with the OFC and NIC methods on MSCOCO are shown in Table 1 and Table 2, where it can be seen that the proposed BaC Category method outperforms all the baseline models in all of the evaluation metrics. When comparing Table 1 and Table 2, the performances achieved by Shuffle-4K and Shuffle-8K with NIC significantly surpass their performances with OFC, and the performances obtained by the proposed BaC Category method are stable with both OFC and NIC. The improvement of the SaC Category method is obvious with OFC as compared with the baseline methods, however, it becomes inconspicuous with NIC. The results imply that the SaC Category provides partial salient semantics for word embedding of language model but does not greatly promote the optimization of visual representations, while the BaC Category not only enables high-level semantic representation but also presents the generalization of visual features.

Table 1. Comparison of SaC and BaC to other taxonomies of ImageNet by the OFC model on the MSCOCO dataset.

Taxonomy	C	B-3	B-4	M	R
ILSVRC-1K [40]	90.7	38.7	28.1	24.2	51.9
Shuffle-4K [31]	89.5	38.5	27.6	23.9	51.8
Shuffle-8K [31]	85.9	37.7	26.5	23.3	51.3
MuCaLe-Net [44]	92.3	38.5	27.8	24.2	51.8
SaC Category	91.8	38.9	27.9	24.3	52.0
BaC Category	96.2	40.4	29.6	24.9	52.7

Table 2. Comparison of SaC and BaC to other taxonomies of ImageNet by the NIC model on the MSCOCO dataset.

Taxonomy	C	B-3	B-4	M	R
ILSVRC-1K [40]	90.1	37.9	27.6	24.0	51.7
Shuffle-4K [31]	92.9	39.8	29.1	24.5	52.4
Shuffle-8K [31]	92.8	39.0	28.2	24.4	51.9
MuCaLe-Net [44]	92.2	38.9	28.1	24.5	52.1
SaC Category	92.3	38.4	27.8	24.3	52.1
BaC Category	96.0	40.4	29.6	24.7	52.8

Similarly, the comparisons with the OFC and NIC methods on Flickr30k are shown in Table 3 and Table 4. From the results, it can be observed that both the SaC and BaC methods are better than the competing baseline methods when employing the OFC method, and the BaC Category method achieves the best. When considering the NIC method, the proposed BaC Category method achieves almost the same as the Shuffle-8K method but with less amount of model parameters.

Table 3. Comparison of SaC and BaC to other taxonomies of ImageNet by the OFC model on the Flickr30k dataset.

Taxonomy	C	B-4	M	R
ILSVRC-1K [40]	34.6	17.8	18.4	42.8
Shuffle-4K [31]	36.2	18.9	18.5	43.2
Shuffle-8K [31]	34.2	18.1	18.0	43.1
SaC Category	36.8	19.4	19.0	43.6
BaC Category	37.4	19.4	19.0	44.0

Table 4. Comparison of SaC and BaC to other taxonomies of ImageNet by the NIC model on the Flickr30k dataset.

Taxonomy	C	B-4	M	R
ILSVRC-1K [40]	33.1	17.8	18.3	42.6
Shuffle-4K [31]	36.8	18.7	18.8	43.2
Shuffle-8K [31]	36.9	19.0	18.8	43.7
SaC Category	35.2	19.1	18.5	43.6
BaC Category	36.7	19.3	18.8	43.7

Moreover, the comparisons with other state-of-the-art methods on MSCOCO and Flickr30k are shown in Table 5 and Table 6, respectively, where the performance of ROUGE_L is not given since almost all of these state-of-the-art methods do not present this result, except Scene+LSTM [12] on MSCOCO in which ROUGE_L = 50.9 which is lower than 53.8 achieved by the proposed BaC Category method. Note that the ROUGE_L performance achieved by the proposed BaC Category method on Flickr30k is 44.4. As shown in Table 5, the proposed BaC Category method achieves almost the best performance on the criteria of CIDEr and BLEU-4 with its METEOR performance a little bit lower than that of Att+CNN+LSTM [49]. A similar observation can also be made from the comparison result on the dataset of Flickr30k as shown in Table 6, where the proposed BaC Category method achieves the best in terms of CIDEr and BLEU-4 while its METEOR performance is a little bit lower than that of Soft-Attention [50] and Hard-Attention [50]. Note that the beamsearch technique is applied for the proposed BaC Category method with the beamsearch size of 3 which is marked as BS3 in Table 5 and Table 6.

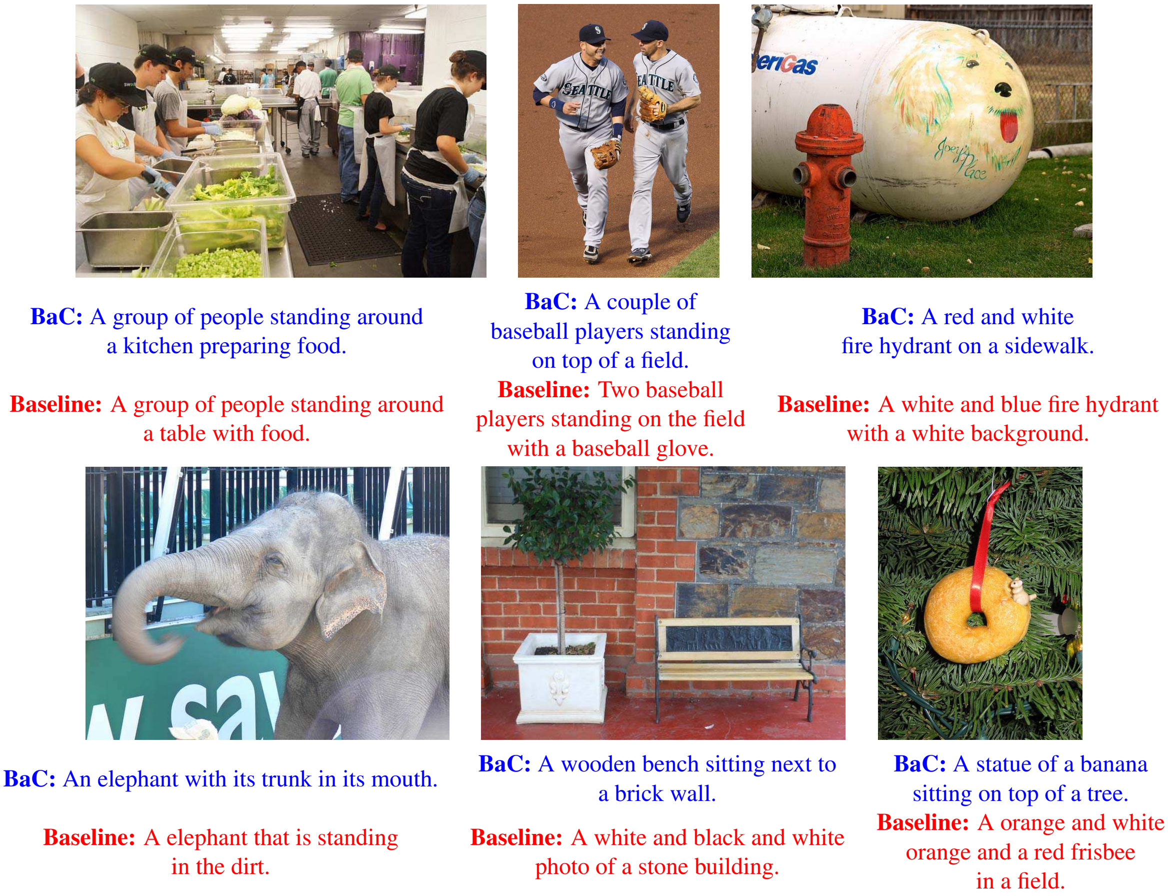
Table 5. Comparison of the proposed BaC by NIC model to state-of-the-art methods on the MSCOCO dataset.

Method	C	B-3	B-4	M
multimodal RNN [14]	66.0	32.1	23.0	19.5
Google NIC [47]	–	32.9	24.6	–
LRCN-CaffeNet [6]	–	30.4	21.0	–
m-RNN [29]	–	35.0	25.0	–
Soft-Attention [50]	–	34.4	24.3	23.9
Hard-Attention [50]	–	35.7	25.0	23.0
emb-gLSTM [11]	81.3	35.8	26.4	22.7
RA+SF [12]	83.8	38.1	28.2	23.5
Att+CNN+LSTM [49]	94.0	**42.0**	31.0	**26.0**
VN-Embed [38]	93.7	39.5	29.7	24.4
GLA [21]	96.4	41.7	31.2	24.9
BaC Category (BS3)	**99.7**	**42.0**	**32.0**	25.5

Table 6. Comparison of the proposed BaC by NIC model to state-of-the-art methods on the Flickr30k dataset.

Method	C	B-3	B-4	M
LogBilinear [15]	–	25.4	17.1	16.9
multimodal RNN [14]	–	24.0	15.7	15.3
Google NIC [47]	–	27.7	18.3	–
LRCN-CaffeNet [6]	–	25.1	16.5	–
m-RNN [29]	28.0	28.0	19.0	–
Soft-Attention [50]	–	28.8	19.1	**18.5**
Hard-Attention [50]	–	29.6	19.9	**18.5**
emb-gLSTM [11]	–	**30.5**	20.6	17.9
BaC Category (BS3)	**39.5**	**30.5**	**21.2**	18.4

Several examples of the generated captions obtained by the proposed BaC Category method are shown in Fig. 5, where the generated captions are given in the first row in



BaC: A group of people standing around a kitchen preparing food.

Baseline: A group of people standing around a table with food.

BaC: A couple of baseball players standing on top of a field.

Baseline: Two baseball players standing on the field with a baseball glove.

BaC: A red and white fire hydrant on a sidewalk.

Baseline: A white and blue fire hydrant with a white background.

BaC: An elephant with its trunk in its mouth.

Baseline: A elephant that is standing in the dirt.

BaC: A wooden bench sitting next to a brick wall.

Baseline: A white and black and white photo of a stone building.

BaC: A statue of a banana sitting on top of a tree.

Baseline: A orange and white orange and a red frisbee in a field.

Figure 5. Comparison of generated caption examples achieved by the proposed BaC Category method. The left row shows the examples without error, the middle row shows the examples with minor errors and inappropriate expressions, and the right row shows the examples with conspicuous mistakes.

blue and the captions generated by the ILSVRC-1K [40] method are also presented in the second row in red for comparison. From left to right, there are three columns indicating the good, fair and bad examples, respectively. The left column shows the captions that accurately describe the objects (*e.g.*, trunk) and achieve comprehensive understandings (*e.g.*, kitchen, preparing food rather than table, food). The middle column shows the captions without error but losing some points. As compared with the ILSVRC-1K [40] method, although the proposed method ignores some partial contents of the image (*e.g.*, glove), it produces less incorrect information. Similar conclusions can be derived from the examples with conspicuous errors in the right column. This comparison further verifies that although the language models are the same, the accuracy of visual representation greatly impacts the quality of the generated captions.

4.2. Visual Question Answering

In order to testify the effectiveness of the proposed BaC level on the VQA task, the dataset of Toronto COCO-QA [37] is applied, which belongs to the single-word answer question and contains four types of questions focusing on the object, number, color and location, respectively. The official training and test split methods are used, and 82,783 images are applied for training, 40,504 images for validation and 81,434 images for test, and each has 3 questions and 10 answers. The existing framework for VQA is mainly based on the VGG+LSTM model [28], which is also employed as the baseline to implement the proposed BaC Category method.

The accuracy comparison with a number of state-of-the-art methods on the COCO-QA dataset is given in Table 7, where the accuracy criteria of object, number, color and location are presented as “Obj”, “Num”, “Col” and “Loc”, respectively. Additionally, the weighted average accuracy is also listed as “Total” in Table 7. About the competing methods, GUESS is a very simple baseline that predicts the answer only by the question type (object, number, color and location). VGG+BoW [37] performs multinomial logistic regression based on the image feature without dimensional-

Table 7. Comparison of the proposed BaC to state-of-the-art methods on the COCO-QA dataset.

Method	Total	Obj	Num	Col	Loc
GUESS	6.7	2.1	35.8	13.9	8.9
VGG+BoW [37]	55.9	58.7	44.1	52.0	49.4
VIS+LSTM [37]	53.3	56.5	46.1	45.9	45.5
2VIS+BLSTM [37]	55.1	58.2	44.8	49.5	47.3
multimodal CNN [26]	54.9	–	–	–	–
ILSVRC-1K [40]	54.2	57.5	47.0	50.3	48.5
Shuffle-4K [31]	55.5	57.4	48.3	**52.7**	47.6
Shuffle-8K [31]	56.0	57.7	**48.7**	**52.7**	48.9
BaC Category	**57.2**	**59.3**	48.3	51.2	**51.5**

ity reduction, and a Bag-of-Word (BoW) vector is obtained by summing all the learned word vectors of the question. In [37], the method of VIS+LSTM uses the last hidden layer of VGG [42] for image embedding and treats it as one word of the question to be encoded at the start and the end of the sentences. The method of 2VIS+BLSTM employs two image features which are put into the language model at the start and the end of the sentences, respectively. On the other hand, the method of multimodal CNN [26] encodes both images and questions by CNN only. Moreover, the methods of ILSVRC-1K [40] and Shuffle-4K/8K [31] are also applied for comparison with the same configuration as the proposed method except the number of BaC categories.

As shown in Table 7, the proposed BaC Category method outperforms other state-of-the-art methods on the total accuracy, especially on the object accuracy with 59.3% and the location accuracy with 51.5%. The accuracy of number and color is slightly lower than some of the other competitors such as Shuffle-4K and Shuffle-8K. This may be due to the reason that the semantic concept classes are categorized mainly based on the object level while the learning of attributes is partly weakened.

5. Conclusion

For the same image, diverse understandings among various persons and the multiple grained understanding bring great ambiguity to manage vision-to-language tasks. To tackle such problems, a basic concept category that resembles cognitive basic level is proposed to provide a consensus representation on a proper level. By intersecting classes in ImageNet and MSCOCO, a salient concept category is obtained. Then a confusion matrix based method is applied to refine the category thereby providing cognitively more accurate visual representation for language modeling. The comparative experimental results on the image captioning task and visual question answering task have verified that the proposed basic concept category method is able to significantly improve the prediction performance for vision-to-language tasks.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proc. ICCV*, pages 2425–2433, Dec. 2015. 2

[2] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004. 2

[3] M. Cole, N. Quinn, E. Rosch, J. GIick, and H. Sinclair. Principles of categorization. In *Cognition and categorization*, pages 189–206, 2002. 1

[4] M. W. Daehler, R. Lonardo, and D. Bukatko. Matching and equivalence judgments in very young children. *Child Development*, 50(1):170–179, 1979. 1, 3

[5] J. Deng, J. Krause, A. C. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Proc. CVPR*, pages 3450–3457, Jun. 2012. 2

[6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, pages 2625–2634, Jun. 2015. 6

[7] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proc. CVPR*, pages 1473–1482, Jun. 2015. 1, 2

[8] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proc. ECCV*, pages 15–29, Sep. 2010. 2

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 1

[10] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, Aug. 2013. 5

[11] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proc. CVPR*, pages 2407–2415, Jun. 2015. 6

[12] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. PAMI*, 39(12):2321–2334, Dec. 2017. 6

[13] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. CVPR*, pages 4565–4574, Jul. 2016. 2

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, pages 3128–3137, Jun. 2015. 1, 2, 5, 6

[15] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Multimodal neural language models. In *Proc. ICML*, volume 14, pages 595–603, Jun. 2014. 6

[16] J. Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *Proc. CVPR*, pages 5546–5555, Jun. 2015. 2

[17] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and

generating simple image descriptions. *IEEE Trans. PAMI*, 35(12):2891–2903, 2013. 2

[18] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *Proc. ACL*, volume 1, pages 359–368, Jul. 2012. 2

[19] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In *Proc. ACL*, pages 790–796, Aug. 2013. 2

[20] A. Lavie and M. J. Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115, Sep. 2009. 5

[21] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian. Image caption with global-local attention. In *Proc. AAAI*, pages 4133–4139, Feb. 2017. 6

[22] L.-J. Li, H. Su, E. P Xing, and L. Fei-fei. Object bank : A high-level image representation for scene classification & semantic feature sparsification. In *Proc. NIPS*, pages 1378–1386, Dec. 2010. 2

[23] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proc. ACL*, pages 605–612, Jul. 2004. 5

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755, Sep. 2014. 1, 3, 5

[25] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. pages 3242–3250, Jul. 2017. 2

[26] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *Proc. AAAI*, Feb. 2016. 8

[27] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proc. NIPS*, pages 1682–1690, Dec. 2014. 2

[28] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proc. ICCV*, pages 1–9, Dec. 2015. 2, 7

[29] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proc. ICLR*, pages II–595–II–603, Dec. 2015. 6

[30] A. Mathews, L. Xie, and X. He. Choosing basic-level concept names using visual and language context. In *Proc. WACV*, pages 595–602, Jan. 2015. 2, 3

[31] P. Mettes, D. C. Koelma, and C. G. Snoek. The imagenet shuffle: Reorganized pre-training for video event detection. In *Proc. ICMR*, pages 175–182, Jun. 2016. 3, 5, 6, 8

[32] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995. 2

[33] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *Proc. ICCV*, pages 2768–2775, Dec. 2013. 3

[34] V. Ordonez, W. Liu, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. Predicting entry-level categories. *IJCV*, 115(1):29–43, Oct. 2015. 3, 5

[35] W. Ouyang, X. Wang, C. Zhang, and X. Yang. Factors in finetuning deep model for object detection with long-tail. In *Proc. CVPR*, pages 864–873, Jun. 2016. 3

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, Jul. 2002. 2, 5

[37] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Proc. NIPS*, pages 2953–2961, Dec. 2015. 2, 7, 8

[38] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proc. CVPR*, pages 290–298, Jul. 2017. 6

[39] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976. 1

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *I-JCV*, 115(3):211–252, 2015. 2, 3, 5, 6, 7, 8

[41] L. Saxby and J. M. Anglin. Children's sorting of objects from categories of differing levels of generality. *Journal of Genetic Psychology*, 143(1):123–137, 1983. 1, 3

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, Apr. 2014. 8

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, pages 1–9, Jun. 2015. 2, 5

[44] Y. Tamaazousti, H. Le Borgne, and C. Hudelot. Mucale-net: Multi categorical-level networks to generate more discriminating features. In *Proc. CVPR*, pages 6711–6720, Jul. 2017. 3, 5, 6

[45] J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, 1991. 1

[46] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proc. CVPR*, pages 4566–4575, Jun. 2015. 2, 5

[47] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, pages 3156–3164, Jun. 2015. 1, 2, 5, 6

[48] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. IGCHI Conference on Human Factors in Computing Systems*, pages 319–326, 2004. 1

[49] Q. Wu, C. Shen, A. Liu, Lingqiaoand Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In *Proc. CVPR*, pages 203–212, Jun. 2016. 6

[50] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057, Jul. 2015. 2, 6

[51] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. E-valuating bag-of-visual-words representations in scene classification. In *Proc. MIR*, pages 197–206, 2007. 2