Learning to Recognize Visual Concepts for Visual Question Answering With Structural Label Space

Difei Gao, Ruiping Wang, *Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*—Solving visual question answering (VQA) task requires recognizing many diverse visual concepts as the answer. These visual concepts contain rich structural semantic meanings, e.g., some concepts in VQA are highly related (e.g., red & blue), some of them are less relevant (e.g., red & standing). It is very natural for humans to efficiently learn concepts by utilizing their semantic meanings to concentrate on distinguishing relevant concepts and eliminate the disturbance of irrelevant concepts. However, previous works usually use a simple MLP to output visual concept as the answer in a flat label space that treats all labels equally, causing limitations in representing and using the semantic meanings of labels. To address this issue, we propose a novel visual recognition module named Dynamic Concept Recognizer (DCR), which is easy to be plugged in an attention-based VQA model, to utilize the semantics of the labels in answer prediction. Concretely, we introduce two key features in DCR: 1) a novel structural label space to depict the *difference* of semantics between concepts, where the labels in new label space are assigned to different groups according to their meanings. This type of semantic information helps decompose the visual recognizer in VQA into multiple specialized sub-recognizers to improve the capacity and efficiency of the recognizer. 2) A feature attention mechanism to capture the *similarity* between relevant groups of concepts, e.g., human-related group "chef, waiter" is more related to "swimming, running, etc." than scene related group "sunny, rainy, etc.". This type of semantic information helps sub-recognizers for relevant groups to adaptively share part of modules and to share the knowledge between relevant sub-recognizers to facilitate the learning procedure. Extensive experiments on several datasets have shown that the proposed structural label space and DCR module can efficiently learn the visual concept recognition and benefit the performance of the VQA model.**

***Index Terms*—Visual question answering, visual concept recognition, structural label space.**

I. Introduction

VISUAL Question Answering (VQA) [1], [5], [18], [29] is a widely studied task that enables users to obtain useful visual information by querying a VQA system about an image. Unlike other computer vision tasks which require the system to understand some concepts limited to a specific field, an advanced VQA system should be able to recognize a large variety of visual concepts to handle a wide range of questions, as shown in Fig. 1(a). The labels in current VQA tasks usually cover the most commonly used types of visual concepts, e.g., objects, attributes, actions, scenes, etc. Therefore, labels in VQA inherently contain rich and diverse semantic meanings, and learning various visual concepts becomes one unique challenge in the VQA task.

Previous VQA approaches mainly focus on designing sophisticated attention mechanisms, while just using a simple flat label space to represent each label, as shown in Fig. 1(b), (c). The flat label space treats each concept as an isolated sign, so it is difficult to distinguish one concept from others and depict the meanings of concepts. However, humans are very good at utilizing semantic meanings to learn new concepts efficiently. For example,

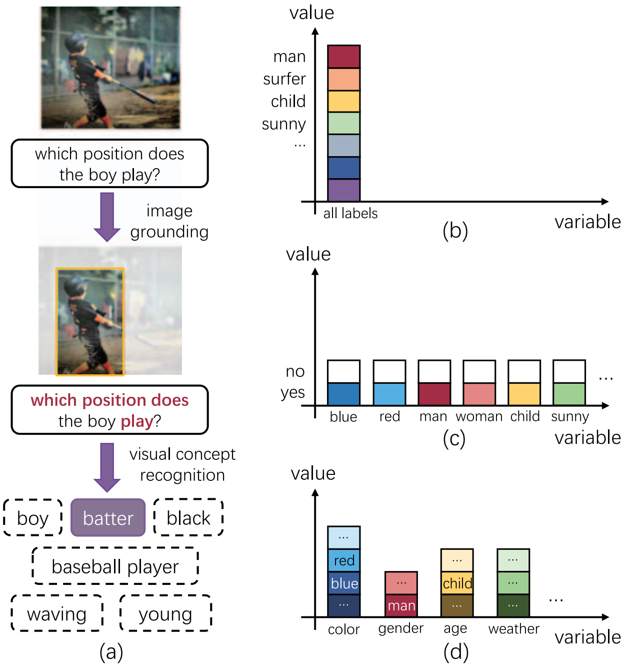

Fig. 1. (a) Common VQA models usually have two parts: image grounding and visual concept recognition. An advanced VQA system should be able to recognize many different types of visual concepts to fit the need of a specific question. (b) (c) (d) is three kinds of label space, i.e., the single-label classification model (b), the multi-label binary classification model (c), and our designed structural label space model (d). (See Section III-A for more details.)

Manuscript received September 15, 2019; revised January 30, 2020; accepted March 6, 2020. Date of publication April 22, 2020; date of current version June 24, 2020. This work was supported by Natural Science Foundation of China under Contracts U19B2036, 61922080, 61772500 and in part by CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Chao Zhang. *(Corresponding author: Xilin Chen.)*

The authors are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, CAS, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: difei.gao@vipl.ict.ac.cn; wangruiping@ict.ac.cn; sgshan@ict.ac.cn; xlchen@ict.ac.cn).

Digital Object Identifier 10.1109/JSTSP.2020.2989701

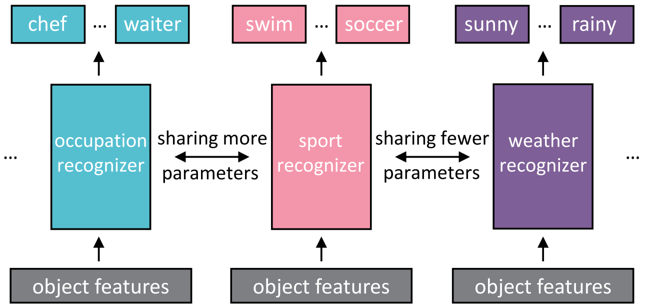

Fig. 2. Basic idea of our Dynamic Concept Recognizer (DCR). The DCR contains a set of sub-recognizers, where each sub-recognizer is responsible for classifying one group of concepts. When answering a question, the DCR will pick one sub-recognizer to output the answer. To capture the similarity between different groups, we allow recognizers to adaptively share part of parameters with their relevant recognizers.

when learning a concept (e.g. "pitcher"), we usually subconsciously learn the new concept with relevant concepts together (e.g. "batter, pitcher, catcher") and eliminate the disturbance of irrelevant concepts (e.g. "red, sunny"). Motivated by this idea, we propose a structural label space to represent the semantic meanings of the answers, as shown in Fig. 1(d). Our structural label space organizes the labels into many groups, where the concepts in one group have relevant semantic meanings. More concretely, the concepts in one group classify the things from the same perspective, e.g. "pitcher, batter" classify the people based on their baseball positions, while "red, blue, etc." classify the objects based on their colors. In our label space, one concept label is determined by two values: *which group* and *which concept*. After obtaining this type of semantics, we design a novel classification module that outputs the labels in our structural label space, named Dynamic Concept Recognizer (DCR). The DCR contains multiple specialized sub-recognizers, where each sub-recognizer is responsible for classifying within a group of concepts, as shown in Fig. 2. Compared to traditional MLP, DCR contains larger capacity for visual recognition, and it is easier to concentrate on learning each of the sub-recognizers related to the question.

While the structural label space mainly aims to distinguish the concepts belonging to different groups, it is also essential to capture the similarity between different groups of concepts. For example, occupation-related group "chef, waiter, etc." is more relevant with "swim, soccer, etc." than "sunny, rainy, etc.", because distinguishing the concepts in the first two groups may both utilize clothing information. A sophisticated model should be able to learn the relevance between different groups to transfer or share knowledge between classifying different groups of concepts. To achieve this goal, we allow the sub-recognizers to adaptively learn to share some parts of parameters with their relevant recognizers during the training procedure, as shown in Fig. 2 (see more details in Fig. 3(b)). Therefore, the relevant sub-recognizers can share the knowledge between each other, which can facilitate the training of some sub-recognizers containing relatively few samples.

Our proposed structural label space and Dynamic Concept Recognizer can be easily plugged on an attention-based VQA model. Concretely, our whole VQA framework, named Dynamic Answer Generator, is composed of three parts (see more details in Fig. 3(a)): The first part finds the image region most related to the question. Then, the second part predicts *which group* the answer belongs to. Finally, Dynamic Concept Recognizer activates one of its sub-recognizers that is asked by the questioner to predict the answer. Note that, besides better utilizing the semantics of labels which can benefit the performance, the whole framework also decomposes the VQA model into many refined modules, e.g. module for grounding and many sub-recognizers. Therefore, our framework inherits the advantages of compositional models on robustness and transparency. Especially for overcoming language priors (a.k.a. rely on the superficial correlation between the question and the answer to guess the answer), since our DCR learns a pure mapping between the image and the visual concepts. As shown in Fig. 2, the question information won't disturb the recognizing of a group of visual concepts as the answer.

To evaluate the proposed method, we conduct extensive experiments on four popular datasets: Visual Genome [20], GQA [15], VQA v2 [10] and VQA-CP v2 [1], and compare with the state-of-the-art methods. The results demonstrate both the effectiveness of our core module Dynamic Concept Recognizer with the structural label space and the effectiveness of the whole VQA model.

The rest of this paper is organized as follows: Section II briefly reviews the related works of the visual question answering and label spaces. Then, Section III introduces our proposed structural label space and new VQA model with Dynamic Concept Recognizer. In Section IV, we provide comprehensive evaluations of our whole VQA framework as well as the core recognition module. Finally, Section V concludes this paper.

II. Related Work

A. *Visual Question Answering*

Previous VQA methods mainly focus on how to ground the proper image region related to the question. These works can be approximately categorized into two branches by way of formulating the grounding procedure. The first branch [2], [9], [16], [21], [24], [32], [34], [38] proposes powerful attention mechanisms to compute the matching score between the question and the image regions. For example, [9] introduces compact bilinear pooling as a fusion technique to build the local relationship between two modalities. [2] proposes a bottom-up and top-down attention mechanism that attends the image at the level of objects and obtain better visual feature. The second branch [3], [4], [13], [17], [26], [37] formulates the image grounding procedure as a multi-step spatial reasoning problem. [4], [13], [17], [31] decompose a grounding model into several pre-defined modules, where each module is responsible for grounding one kind of concepts, e.g., finding "red" objects. These modules are dynamically assembled for a given question and are used to generate the answer. [14] proposes an alternative approach, Memory, Attention, and Composition (MAC) which also performs multi-step reasoning and can dynamically record and retrieve the visual information in its memory.

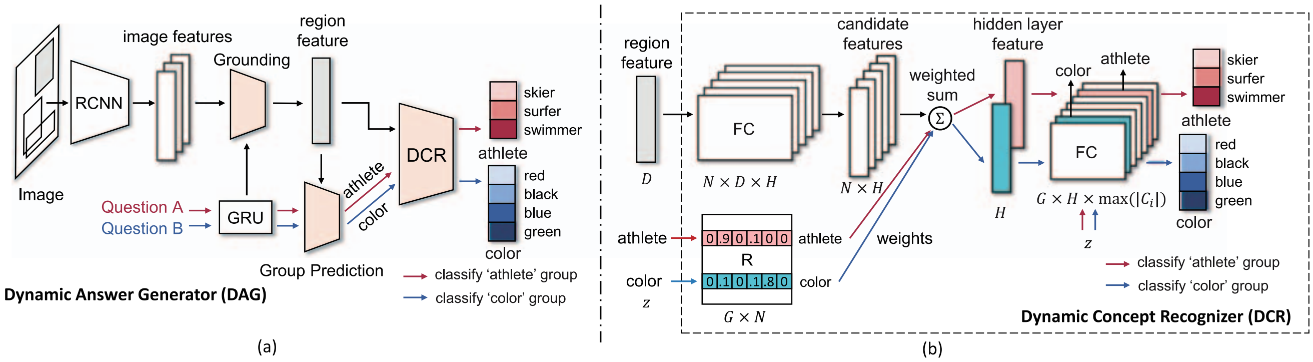

Fig. 3. The architectures of the whole VQA model Dynamic Answer Generator and the core module Dynamic Concept Recognizer. (a) Dynamic Answer Generator (DAG). Given a question and an image, DAG locates the image region related to the question and predicts the group index that determines the model to recognize which group of concepts. Then, the DCR generates the final answer based on the attention feature and the group index. (b) Dynamic Concept Recognizer (DCR). The DCR takes an image region feature and the group index as input. The group index indicates the DCR to activate a part of the module to output the visual concept in the image region. To better illustrate the DCR, we show two cases that activate different parts of the module.

In contrast, this paper focuses on how to correctly generate the concept asked by the question, after attending to the correct image region. Previous methods mainly implement an MLP with a flat label space to learn the mapping from the combination of the question and the image features to the visual concept. This type of methods has two drawbacks: 1) For performance, due to the limitation of the flat label space in representing the meanings of labels, these models cannot focus on distinguishing relevant concepts to efficiently learn recognition from the question-answer sample; 2) for robustness and transparency, these models combine the image and the questions features to answer the question, so they do not contain a pure visual concept recognizer mapping from visual features to labels. Thus, previous models have risks to overuse the questions to guess the visual concept answers. To address these problems, we first propose a structural label space to represent the semantic meanings of the labels. Then, we propose a Dynamic Concept Recognizer, which is an alternative method to generate the visual concept, to avoid directly combining the question and visual features.

B. Label Spaces

Many works in ML, CV, and NLP fields study how to design the label spaces to represent the relations between labels. The most common label space is the flat label space used in standard single-label classification and multi-label binary classification. [8], [19], [19], [28], [35], [39] propose a tree-structured label space to represent the class hierarchy of a large number of labels. They usually implement a top-down approach to classify the nodes from the root to the leaves. The tree-structured label space can be used in both single-label or multi-label classification. [36] represents the labels in a hypercube space to build the correlations between labels.

In this paper, we propose a label space, which organizes the labels into many groups, to represent the relations between visual concept labels. Our label space can be viewed as a simplified version of tree-structured label space, which only contains two layers, and all labels are leaf nodes. This simplified structure is more suitable for the relations of VQA labels (less hierarchical relations, more clustering relations). It allows our VQA model to focus on learning one group of concepts easily.

C. Clustering the Answers or Questions

Some of the existing methods [1], [33] propose to cluster the answers or questions and use answer type or question type information to improve the VQA model. [33] divides all questions into 12 different types (labeled in the TDIUC dataset) and introduces a question type guided attention mechanism that dynamically balances between bottom-up and top-down visual features based on question type information. [1] uses K-means to cluster the answers into 50 clusters. The model in [1] first generates all visual concepts in the grounded image region, then combines the visual concepts feature and answer type features to output the answers.

Previous methods usually formulate the answer or question type as a feature which provides additional information for some specific functions, such as, selecting different types of object features [33] or selecting concepts [1]. In contrast, we use the semantics of answer to construct a novel label space which allows the VQA model to utilize the semantic meanings of the labels to facilitate the training of the VQA model.

III. Methods

In this section, we first define the label space of our VQA model (in Section III-A). Then, we illustrate the main framework of the VQA model named Dynamic Answer Generator (in Section III-B).

A. Structural Label Space for Visual Concepts

To better understand our new label space, we first recap the definition of the label spaces in a single classification problem and multi-class binary classification problem.

In a standard single-label classification problem, the label space contains one discrete variable:

$$\mathcal{L} = \{Y\}, \qquad Y \in \{0, \ldots, C-1\} \tag{1}$$

where C is the number of classes. In this setting, all classes are assumed to be disjoint. In a multi-label binary classification problem, the label space contains a tuple of C discrete variables:

$$\mathcal{L} = \{Y_1, Y_2, \ldots, Y_C\}, \qquad Y_i \in \{0, 1\} \tag{2}$$

In this setting, all classes are considered as independent.

These flat label spaces cannot represent the semantic meanings of the labels. Therefore, we create a new structured label space to represent the semantic relations among concepts. More specifically, we cluster the visual concepts as many groups of labels, where each group of labels depicts the visual world from one perspective, e.g., "male, female," "red, black, etc.". We consider concepts in one group are highly related, and concepts in different groups are independent. Formally, we let the new label space have a tuple of G discrete variables:

$$\mathcal{L} = \{Y_1, Y_2, \ldots, Y_G\}, \qquad Y_i \in \{0, \ldots, C_i - 1\} \tag{3}$$

where G indicates the number of groups and C_i indicates the number of concepts in the i-th group.

Practically, we cluster the concepts in the following two steps. First, the rough clustering result can be obtained by K-means clustering in Glove [25] embedding space. Second, we resort to WordNet [23] to manually fine-tune the clustering results. More specifically, we check if the concepts in one group belong to the same root in WordNet, if some groups omit some concepts, and if some concepts are mistakenly assigned to groups. Then, we manually assign misaligned concepts to proper groups. These steps make sure that the concepts in one group classify the visual world from the same perspective, e.g., all concepts are related to "color" or "material," etc. Note that, in this manually refining procedure, we cluster the concepts only considering their meanings without the restriction of group number. Besides, once we get the clustering result from one dataset, it is easy to generalize to a new dataset, because the meanings of concepts won't change much across different datasets. Therefore, to obtain the clustering result of answers in a new dataset, we only need to consider the additional concepts, e.g., merge additional concepts to existing groups, or create some new groups.

B. Dynamic Answer Generator

For Dynamic Answer Generator (DAG), the inputs to the model are the question and the image region features that are extracted by ResNet [12] or Faster R-CNN [30], and the goal is to generate the correct answer (the overview of the DAG is shown in Fig. 3(a)). Concretely, DAG is composed of three networks: 1) Image Grounding Network that uses the entire question to ground the image region related to the question, 2) Group Prediction Network that distills the question information to a group index which indicates the latter network to recognize a specific group of concepts and 3) Dynamic Concept Recognizer that uses the predicted group index and the grounded image feature to predict the answer.

Image Grounding Network: We first use Faster R-CNN [2] to extract a set of image region features $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}, \boldsymbol{x}_i \in \mathbb{R}^D$, where k is the number of image regions. In the meantime, every word in the question is encoded into a word vector by using a learned word embedding. Then the question feature $\boldsymbol{q}$ is obtained by feeding word vector sequence into a gated recurrent unit (GRU) [7].

We calculate the weights $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ attending to image regions, where $\Sigma_{i=1}^{k} \alpha_i = 1$. Concretely, we compute the normalized matching score α_i between every pair of $\boldsymbol{q}$ and $\boldsymbol{x}_i$. The matching score can be calculated by any sophisticated attention module, e.g. simple soft attention mechanism, compact bi-linear pooling, or multi-step attention in [14], etc. Here, we use simple soft attention mechanism as an example:

$$f_{att}(\boldsymbol{x}_i, \boldsymbol{q}) = \mathrm{ReLU}(\boldsymbol{W}_{\mathrm{v}} \boldsymbol{x}_{\mathrm{i}}) \cdot \mathrm{ReLU}(\boldsymbol{W}_{\mathrm{q}} \boldsymbol{q}) \tag{4}$$

$$\alpha_i = \frac{\exp(f_{att}(\boldsymbol{x}_i, \boldsymbol{q}))}{\Sigma_{j=1}^{k} \exp(f_{att}(\boldsymbol{x}_j, \boldsymbol{q}))} \tag{5}$$

where $\boldsymbol{W}_v$ and $\boldsymbol{W}_q$ are trainable parameters, which project $\boldsymbol{x}_i$ and $\boldsymbol{q}$ to the same dimension and $\cdot$ means dot product. Finally, the grounded image region is calculate as:

$$\boldsymbol{v} = \Sigma_{i=1}^{k} \alpha_i \boldsymbol{x}_i. \tag{6}$$

Group Prediction Network: GPN utilizes the image region feature $\boldsymbol{v}$ and the question feature $\boldsymbol{q}$ to distill the information that is used to guide the visual concept recognition (a.k.a the group index). Note that, the image feature is also used to predict the group index, because many questions contain not enough information to correctly predict the group indexes, e.g. for question "what is on the ground?", the answer could be in many concept groups, such as "cat, dog, etc", "man, woman". Thus, the distribution over all possible group indexes $p(\boldsymbol{z}|\boldsymbol{q}, \boldsymbol{v})$ can be formulated as

$$f_g(\boldsymbol{v}, \boldsymbol{q}) = \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{v}) \circ \mathrm{ReLU}(\boldsymbol{W}_2 \boldsymbol{q}) \tag{7}$$

$$p(\boldsymbol{z}|\boldsymbol{q}, \boldsymbol{v}) = \sigma(\boldsymbol{W}_g f_g(\boldsymbol{v}, \boldsymbol{q})) \tag{8}$$

$$\hat{z} = \underset{z \in \{1, \ldots, G\}}{\operatorname{argmax}} \; p(\boldsymbol{z}|\boldsymbol{q}, \boldsymbol{v}) \tag{9}$$

where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_g$ are trainable parameters, $\circ$ is the element-wise product, σ is the sigmoid function and $\hat{z}$ is the predicted group index. We use sigmoid function here to deal with the situation that the question is ambiguous and can be answered from different perspectives, because sigmoid function can output multiple labels as true.

Dynamic Concept Recognizer: The goal of Dynamic Concept Recognizer (DCR) is to classify the given group of concepts for the given image region. The predicted concept is used as the final answer. Concretely, the DCR is required to accomplish multiple tasks (classifying concepts in one group corresponds to one task), such as distinguishing colors, animals, etc. Besides, we hope DCR can automatically learn the relevance of different tasks from Question-Answer samples by let relevant tasks share part of the model architecture.

To achieve this goal, we propose a multi-task module based on two-layer MLP (as shown in Fig. 3(b).). The DCR first uses a set of fully connected (FC) layers to generate a set of candidate hidden features. Then, the DCR picks some candidate features for one specific task by implementing the soft-attention mechanism (supervision information of the answer implicitly encourages modules of relevant tasks to apply similar attention values on candidate features). Finally, the picked candidate features are fed into an FC layer corresponding to the current task to output the visual concept.

Formally, given the region feature $\boldsymbol{v}$ and the predicted group index $\hat{z}$, one part of DCR is dynamically activated and predicts the probability of visual concepts in the $\hat{z}$th group $p(\boldsymbol{y}_{\hat{z}}|\boldsymbol{v}, \hat{z})$ corresponding to the region feature, as illustrated in Fig. 3(b). DCR first uses a set of fully connected (FC) layers with parameters $\boldsymbol{W}_h = \{\boldsymbol{W}_h^1, \ldots, \boldsymbol{W}_h^N\} \in \mathbb{R}^{N \times D \times H}$, $\boldsymbol{b}_h = \{\boldsymbol{b}_h^1, \ldots, \boldsymbol{b}_h^N\} \in \mathbb{R}^{N \times H}$ to generate a set of candidate visual features $\boldsymbol{h} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\} \in \mathbb{R}^{N \times H}$ given the region feature $\boldsymbol{v}$, where N is the number of FC layers. The calculation of each $\boldsymbol{h}_i$ can be formulated as

$$\boldsymbol{h}_i = \tanh(\boldsymbol{W}_h^i \boldsymbol{v} + \boldsymbol{b}_h^i) \tag{10}$$

Then, the model picks $\boldsymbol{h}_i$ for the current task $\hat{z}$ as the hidden feature to output to the next layer.

$$\boldsymbol{h}_o = \Sigma_{i=1}^N r_{\hat{z},i} \boldsymbol{h}_i \tag{11}$$

where $\boldsymbol{R} = (r_{z,i})_{G \times N}$ are trainable parameters, and $r_{z,i}$ controls the z-th task attend how much on every $\boldsymbol{h}_i$ and $\Sigma_{i=1}^N r_{z,i} = 1$ for $z \in \{1, \ldots, G\}$.

After calculating the hidden feature $\boldsymbol{h}_o$, there are G FC layers where each FC layer is responsible for classifying one group of concepts. DAG only activates the $\hat{z}$-th FC layer to output the probabilities of visual concepts in the given group.

$$p(\boldsymbol{y}_{\hat{z}}|\boldsymbol{v}, \hat{z}) = \mathrm{softmax}(\boldsymbol{W}_o^{\hat{z}} \boldsymbol{h}_o) \tag{12}$$

where $\boldsymbol{W}_o = \{\boldsymbol{W}_o^1, \ldots, \boldsymbol{W}_o^G\}$ represent the trainable parameters of G FC layers. To easily train DCR for batch samples, we need to keep the size of the probabilities to be consistent. Thus, we design every $\boldsymbol{W}_o^z$ to have the same size $H \times max(C_z)$ (it is easy for the network to know that the extra labels should not be the answer).

Notably, the module of DCR for classifying non-visual concepts "yes, no" (a.k.a. yes/no questions) is different from the aforementioned visual concept modules. This module uses the combination of image and question features to generate the answer, rather than only image features, because "yes/no" questions require the model to compare the content of the image and question. Concretely, for non-visual questions, we input the combination of the image feature and the question feature $f_g(\mathbf{v}, \mathbf{q})$ calculated by equation (7) to Dynamic Concept Recognizer, rather than image feature $\boldsymbol{v}$. In addition, the sub-recognizer of "yes/no" questions are independent with visual sub-recognizers (in practice, we fix the attention value for corresponding candidate feature, where only the attention value on corresponding candidate feature is 1, others are 0).

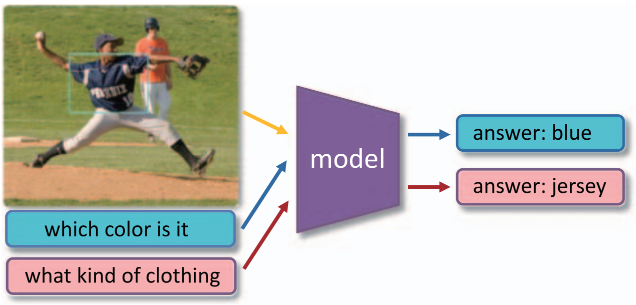

Fig. 4. An illustration of our toy benchmark. For one specific sample, the model takes the group index and an image region as input, and outputs the corresponding visual concept. The group index can be viewed as a parsed simple question, like "which color is it". In addition, providing the correct image region reduces the disturbance of image grounding model.

C. Training

For training our model, we minimize the combination of two losses $\mathcal{L} = \lambda \mathcal{L}_z + \mathcal{L}_y$, where $\mathcal{L}_z$ and $\mathcal{L}_y$ are the binary cross-entropy losses of predicting group indexes and the final answer respectively, and λ is a hyperparameter that balances the relative weights of training two modules. Note that, during training, the ground truth group index is fed into the DCR, instead of the predicted group index. Since if the predicted group index is wrong, there will be no loss to train the DCR.

IV. Experiments

In this section, we conduct experiments on widely studied datasets to show the effectiveness of our proposed label space, the core module Dynamic Concept Recognizer and the whole VQA model Dynamic Answer Generator.

A. Experiments on Visual Genome

This paper aims at improving one important sub-task of VQA, visual recognition, so that we want to firstly purely test the effectiveness of our proposed method on recognition, before conducting the experiments for seeing how the VQA model works out as a whole. To do so, we propose a simplified toy benchmark to purely test the ability of a VQA model for recognizing the visual concepts. Then, we compare the performance of our proposed model with baselines.

1) Experiment Settings: Task description: The visual concept recognition requires the model to classify many different kinds of concepts. The standard VQA benchmark is one way to test the performance of visual concept recognition, but many factors impact the results, such as whether the model attends on the correct region, whether the model correctly parses the question. Thus, we propose a toy benchmark to test the performance of visual concept recognition purely. In details, for one sample (as shown in Fig. 4), the inputs of the model are *one image*, *one ground truth bounding box* of an object and *one group index* that indicates the model need to classify one specific group of concepts. The group index can be viewed as a parsed simple question, e.g., "what color is it?". The output should be the corresponding concept appearing in the image. We use classification accuracy to evaluate a model.

TABLE I
SOME SAMPLES OF CONCEPT GROUPS IN CLUSTERING RESULTS

Group Name	color	sport	indoor scene	natural scene	vehicle part	body part	appliance
Concepts	white black gray red green blue yellow purple brown ...	racing baseball skiing skating surfing tennis riding soccer swimming ...	bedroom dinning room living room bedroom bathroom kitchen office room attic basement ...	mountain hill beach sky shore dock woods forest air ...	engine cargo steering wheel wheel tire propeller kickstand	head face neck shoulder arm hand lap tail foot ...	dishwasher microwave blender toaster oven refrigerator stove grill kettle ...

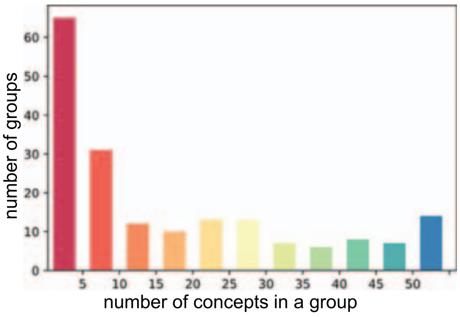

Fig. 5. Distribution of concept number among groups.

TABLE II
THE WEIGHTED MEAN ACCURACIES (%) OF CLASSIFYING GROUPS OF VISUAL CONCEPTS ON VISUAL GENOME

Method	Overall	Object	Attribute	Relation
S-L	52.67	54.09	50.95	51.28
M-L	55.23	56.51	54.50	53.41
DCR w/o Params-Sharing	59.17	60.33	57.24	58.09
DCR	60.89	61.64	59.05	59.42

Dataset: We test the performance of visual concept recognition on the Visual Genome dataset [20]. The dataset contains 108 K images annotated with bounding boxes and class names of objects, attributes, relationships. We unpack the relationship and attribute annotations of Visual Genome to <image, bounding box, concept> triplets (the bounding box of the relationship concept is the union of the bounding boxes of the object and the subject), and filter 1,000 most frequent visual concepts in the Visual Genome. The dataset is randomly split into train (70 K images), val (15 K images) and test (15 K images) sets.

Concept Clustering: We first collect the union set of 1) 1,000 most frequent visual concepts in the Visual Genome, 2) all answers in GQA dataset [15] and 3) 2,000 most frequent answers in VQA v2 dataset [10]. Then, we cluster these concepts by using the method illustrated in Section III-A. The experiments in Sections IV-A, IV-B, and IV-C filter their corresponding concepts in this clustering results to build their own concept groups. In Table I, we show 7 groups of concepts (185 groups in total) in visual concepts clustering results. We name every group of concepts according to their meanings to better record the results. In Fig. 5, we display the distribution of concept numbers among groups. We can see that many groups contain less than 5 concepts. This is because there are lots of paired attribute-type concepts in our language used for distinguishing objects from different perspectives, e.g., dark & bright, male & female. The groups in the middle scale (e.g., $5 <$ group number $<= 40$) are usually some complicated attributes, e.g., type of color, or fine-grained classes, e.g., type of bag. The groups in large size (group number > 40) are mainly related to some common objects, e.g., type of animal.

Implementation of DCR: We use the Dynamic Concept Recognizer to accomplish this task. We first use ResNet-101 [12] to extract the feature. Then, we implement ROI align [11] with ground truth bounding box to crop on the last layer conv feature and feed into the DCR. Besides, we set the number of FC layers N in the first hidden layer as 100, and the dimension of hidden layer feature H as 100 through cross-validation. The output size of each output FC layer $max(C_z)$ is 35.

Baselines: The main novelty of our Dynamic Concept Recognizer is the use of new label space. Therefore, we evaluate the models with different label spaces to show the effectiveness of ours. We keep the extracted feature same and only change the classification model.

- *S-L:* The single-label classification model (S-L) treats the visual concepts as disjoint classes. The S-L model first converts the group index into a 10-dimensional feature and concatenates it on the image feature. The combined feature is fed into an MLP to predict the answer. Besides, the model is trained with softmax cross entropy loss.
- *M-L:* The multi-label binary classification model (M-L) treats the visual concepts as independent. The M-L model is similar to the S-L model, except the model is trained with binary cross entropy loss.
- *DCR w/o Params-Sharing:* The modules for all recognition tasks in DCR are independent, a.k.a the number of candidate features is equal to the number of tasks, and the module of each task uses its own candidate features.

2) Results and Analysis: Table II shows the results of the baselines and our method on Visual Genome. The results show

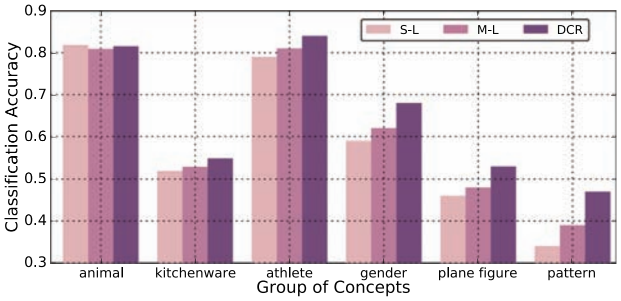

Fig. 6. Classification accuracies (%) for selected groups of concepts.

that our Dynamic Concept Recognizer can better learn to recognize the visual concepts. There are mainly two reasons. First, the structural label space let relatively dependent modules to recognize different groups of concepts. Thus, each model can focus on learning how to classify one specific group of highly related concepts without the disturbance of other unrelated concepts. This advantage mainly helps to recognize the visual concepts that are relatively difficult to classify. To demonstrate this idea, we select the results on classifying some groups of concepts, as shown in Fig. 6. For classifying “kitchenware,” “athlete,” “gender,” we can see that our method achieves better results. These groups are all relatively hard to classify, compared to “animal” group. Secondly, The correlation among the groups can help Dynamic Concept Recognizer transform the knowledge of classifying one group of concepts to classifying another relevant group. The transformed knowledge can facilitate classifying those concept groups with relatively few samples. For example, in Visual Genome, “plane figure” and “pattern” are two groups with relatively fewer training samples. From Fig. 6 in the paper, we can see that our method obtains better results on classifying these two groups of concepts. We also calculate the mean accuracy of the groups which contain less than 100 samples. The DCR w/o Params-Sharing achieves 47.1% accuracy, and DCR achieves 50.3% accuracy. These experiments show that correlation information indeed helps to learn with relatively few samples.

In summary, the semantic meaning of the concepts is informative. Using this information in designing recognizer and building structured label space helps the model better classify the visual concepts, compared to the model neglecting it.

B. Experiments on GQA

1) Experiment Settings: Datasets: The GQA [15] is a recent proposed large-scale visual question answering dataset that contains 113 K real images from Visual Genome, and 1.7 M balanced question. Their questions involve multi-step relational reasoning, grounding, and recognizing diverse visual concepts. In addition, GQA provides rich annotations: 1) scene graph annotations which contain the bounding boxes, classes, and attributes of the objects in the image, and pairwise relationships of objects; 2) functional programs which list the series of reasoning steps that have to be performed to arrive at the answer. The dataset is split into four splits, train, val, test-dev, and test splits. In our experiments, we train models on train split and test the performances on the other GQA splits. Note that there is a domain shift between val split and the other splits, which may cause a performance drop from val to test-dev and test.

Implementation: For visual concept clustering, we use the clustering results obtained in Section IV-A. For the attention function, we use the one in the state-of-the-art method on GQA, MAC model [14]. We use Glove [25] as initialization of the word embedding layer, and the embedded words are fed into a GRU with $512d$ hidden states. The visual features in our model are the object detection features provided by the GQA dataset with size $N \times 2048$ (where N is the number of detected objects) from a Faster R-CNN detector [30]. In Dynamic Concept Recognizer, the hyper-parameters of the module are as same as in Experiments IV-A. Our model is trained by using the rmsprop optimizer. We set the learning rate as 5e-5 with relative weight $\lambda = 0.1$.

Baselines: To demonstrate the effectiveness of our proposed DAG model, we compare our model with several baselines.

- *MAC:* MAC is the state-of-the-art method on GQA which implements multi-step reasoning on the image. For visual concept recognition, MAC combines the question and image features and uses an MLP to output the answer on a basic flat label space.
- *DAG w. S-L:* We replace the Dynamic Concept Recognizer in DAG with **S-L** model illustrated in Section IV-A.
- *DAG w. M-L:* We replace the Dynamic Concept Recognizer in DAG with **M-L** model illustrated in Section IV-A.
- *DAG w/o Params-Sharing:* We replace the Dynamic Concept Recognizer in DAG with **DCR w/o Params-Sharing** illustrated in Section IV-A.

2) Results and Analysis: Table III shows the performances of the state-of-the-art methods and our method on the test split of GQA. Table IV displays the performances of our model and its variations on validation and test-dev split.

Utilizing the semantic meaning of labels can facilitate the performance of VQA: By comparing MAC with DAG in Table III, it can be seen that our DAG brings 3% (absolute) improvement in accuracy, indicating that it is effective to represent the semantic meaning of labels and allow the model to utilize the semantic meanings while inferring. Moreover, comparing the results of MAC with DAG w/o Params-Sharing and DAG in Table IV, we find that the structured label space and dynamic concept recognizer both are effective, and the structured label space contributes more to overall accuracy.

The quality of the visual concept recognizer impacts VQA performance: Comparing different visual concept recognition modules, DAG, DAG w. S-L and DAG w. M-L, DAG achieves higher accuracies on both datasets. The results show that using different visual concept recognizers with the same grounding method and group prediction method dramatically impacts the VQA performance. Besides, for the DAG model evaluated on test-dev set, we show the performances of three modules in the DAG model: image grounding network, group prediction network, and dynamic concept recognizer. The outputs of three modules all have ground truth annotation: the object related to questions, the group related to questions, and the answers. Therefore, we use the accuracy to evaluate the performance of

TABLE III
ACCURACY (%) OF OUR SINGLE MODEL ON THE GQA TEST SET

Model	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accuracy
LSTM	61.90	22.69	68.68	87.30	96.39	17.93	41.07
LSTM-CNN	63.26	31.80	74.57	84.25	96.02	7.46	46.55
Bottom-Up [2]	66.64	34.83	78.71	84.57	96.18	5.98	49.70
MAC [14]	71.23	38.91	81.59	84.48	96.16	5.34	54.06
Ours: DAG	72.79	42.94	84.41	84.68	96.37	5.76	56.94

TABLE IV
PERFORMANCES (%) OF DIFFERENT VARIATIONS OF DAG MODEL ON VAL AND TEST-DEV SET OF THE GQA DATASET

Model	val Acc.	test-dev Acc.
MAC	57.50	53.26
DAG w. S-L	57.31	52.47
DAG w. M-L	59.62	53.34
DAG w/o Params-Sharing	62.37	55.25
DAG (Full model)	64.45	56.89

TABLE V
PERFORMANCES OF THREE MODULES IN DAG MODEL ON TEST-DEV SET ON THE GQA DATASET

Module	Image Grounding	Group Prediction	Concept Recognition
Accuracy	84%	94%	68%

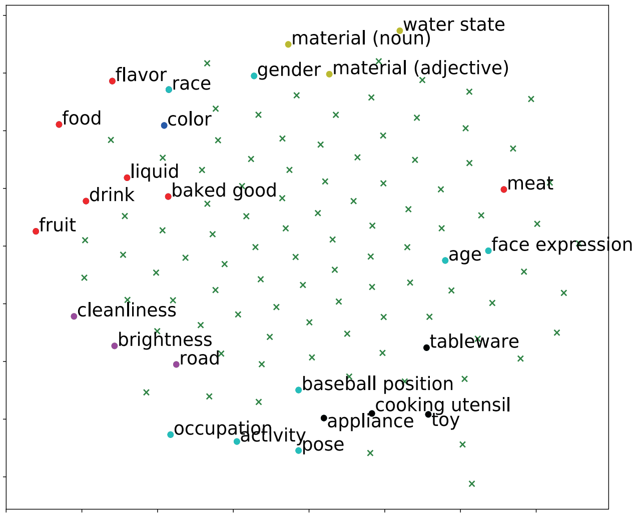

Fig. 7. The t-SNE [22] visualization of attention vectors on candidate features for recognizing each group of concepts.

each module, and the results are shown in Table V. The results demonstrate that the main challenge of the VQA model is the recognition of visual concepts, and how to design a better visual concept recognition module in the VQA procedure is worth study.

The effect of parameters sharing in Dynamic Concept Recognizer: Comparing the results of DAG and DAG w/o Params-Sharing, it can be seen that parameter sharing can improve the performance of the VQA model. To better know the effectiveness of parameters sharing on concept recognizer, we visualize the distribution of attention values on candidate features of each recognition sub-task in Fig. 7 (the points of some relevant sub-tasks are marked in the same color). First, we observe that the attention values are distributed uniformly in the space, because no constraint guides them to cluster. Then, it can be seen that many related groups are closed in attention value space, e.g. "*cleanliness* (clean, dirty), *brightness* (bright, dark), *road* (highway, pavement, etc.)" may all need to attention on the features capturing the brightness of the image (e.g., a highway is usually in a bright color, while a pavement is usually in a dark color), "*age*, *face expression*" may both requires attending on a face, "*material (noun)* and *material (adjective)*" are two similar groups of concepts in different forms, and their attention values are also closed. These observations demonstrate that DAG indeed improves performance by learning the similarity between different groups.

The influence of clustering results: Considering the complicated structure of concept semantic space, it is actually an open problem to find an exact or optimal clustering result. In this part, we evaluate four different clustering results with different group numbers to test the impact of different clustering results to VQA final accuracy. Concretely, we keep the clustering results of attributes and relationships related concepts be same among four clustering results, because these concepts classify things in a specific and clearly defined rule, e.g. "color," "material". Then, we merge or split object concepts to obtain other clustering results that group the concepts in different granularities. In addition, we propose another clustering result to support one concept relating to multiple groups.

- 153 Clusters: 153 groups of concepts obtained by using our clustering method.
- 100 Clusters-Random: 100 groups of concepts, where the object concepts are randomly clustered.
- 100 Clusters: 100 groups of concepts which is obtained by merging some groups of object concepts in 153.
- 200 Clusters: 200 groups of concepts which is obtained by splitting some groups of object concepts in 153.
- 153 Clusters-Multi-Groups: This clustering result allows a polysemous concept to relate to multiple groups. We split the polysemous concepts (4 concepts founded in the GQA dataset, *orange, light, short, old*) to specific meanings used

TABLE VI
PERFORMANCES (%) OF DIFFERENT VARIATIONS OF DAG MODEL WITH DIFFERENT CLUSTERING SCHEMES ON VAL AND TEST-DEV SET ON THE GQA DATASET

Clustering Results	val Acc.	test-dev Acc.
100 Clusters-Random	52.71	47.61
100 Clusters	63.78	56.07
153 Clusters	64.45	56.89
153 Clusters-Multi-Groups	64.51	56.82
200 Clusters	62.90	55.33

in datasets, then cluster the meanings of the concepts by using the WordNet and finally obtain the meaning-level groups. Then, we train the model with meaning-level answers obtained by utilizing the question-type and answer annotations. For example, for a polysemous answer "orange", if the question-type is query-object, then the "orange" represents an object; if the question-type is query-color, then the "orange" represents a type of color.

The validation and test-dev accuracies of DAG models with different clustering results are shown in Table VI. It can be seen that the VQA accuracy is relatively stable for different clustering results as long as the clustering results are meaningful. Besides, we find that the best accuracy is at 153 groups. It is because the growth of group numbers will increase the difficulty of group prediction, but eases the difficulty of visual concept recognition. The number of groups 153 is a balanced point for the current DAG model, which achieves relatively promising accuracies on both two sub-tasks. Besides, from the result of 153 Clusters-Multi-Groups, it shows that our method supports the case that some concepts are related to multiple groups. Moreover, using the 153 Clusters-Multi-Groups does not lead to obvious performance improvement compared to 153 Clusters. It might because, in the current dataset, limited samples belong to that case (91/ 12,578 $\approx$ 0.723% answers in test-dev split are in the above-mentioned 4 polysemous answers).

C. Experiments on VQA v2 and VQA-CP v2

1) Experiment Settings: Datasets: The VQA v2 [10] is a free-form open-ended visual question answering dataset that contains 200,000 MSCOCO images, 1,105,904 question-answer pairs. The VQA-CP v2 [1] (Visual Question Answering v2 under Changing Priors) is a new split of VQA v2 to test whether the model overuses the correlation between the questions and the answers. VQA-CP v2 re-organizes the train and val splits of VQA v2 to make the distribution of answer per question type to be different between two splits. In the VQA v2 and VQA-CP v2, the questions are separated into three categories: yes/no, number, other. Also, the results are evaluated by using the VQA evaluation metric [5].

Implementation: For visual concept clustering, we filter the 2,000 most frequent answers on VQA v2 dataset which are assigned to 175 groups in clustering results obtained in Section IV-A. We use Glove [25] as initialization of the word embedding layer, and the embedded words are fed into a GRU with $512d$ hidden states. We extract a fixed number of $k = 36$ Bottom-Up [2] features per image. In Dynamic Concept Recognizer, the hyper-parameters of the module are as same as in Experiments IV-A. Our model is trained by using the rmsprop optimizer. We set the learning rate as 1e-4 when separately training Group Prediction Network and Dynamic Concept Recognizer, and set the learning rate as 5e-5 when fine-tuning the whole module with relative weight $\lambda = 0.1$.

Baselines: To demonstrate the effectiveness of our proposed DAG model, we propose some variations of our method. First, we apply our method on two commonly used pre-training features **ResNet Feature** [12], **Bottom-Up Feature** [2]. Then, we test several baseline models:

- *SAN [38]:* SAN uses the ResNet feature to encode the image and performs two-hop question-based image attention to ground the image. Then, SAN uses an MLP to predict the answer on a flat label space.
- *GVQA [1]:* Similar to DAG, GVQA also clusters the answer into multiple groups, but it uses the clustering differently and still uses an FC layer to predict the answer on a flat label space. Concretely, GVQA first uses the attention mechanism to ground the image. Then, GVQA generates all the concepts in the grounded region. Finally, one concept will be selected based on the group index feature generated from question parser. GVQA formulates the group information as a feature to represent additional features, which is different from us which use clustering results to build label space.
- *Bottom-Up [2]:* Bottom-Up proposes an object-level image features. Then, Bottom-Up uses the question to calculate the soft attention value to attend an image region. Finally, Bottom-Up uses an MLP to predict the answer on a flat label space.
- *DAG w. Q:* This baseline aims to test the impact of the question information in visual concept recognition. Instead of only feeding visual feature to the Dynamic Concept Recognizer, we combine the image feature and question feature by using element-wise dot product and feed combined feature to it.

2) Results and Analysis: Table VII shows the results of state-of-the-art methods and variations of our method. The results of state-of-the-art methods are cited from [1], [27]. We analyse the observations obtained from the results:

Dynamic Answer Generator effectively avoid overusing language priors: From the results of Bottom-Up vs. DAG (Bottom-Up feature) and SAN vs. DAG (ResNet feature) on VQA-CP v2, it demonstrates that DAG effectively prevents overusing the language priors for two image features. We also qualitatively evaluate our method on robustness. In the first two samples in Fig. 8(a), both question answering pairs are rare in the dataset; and the Bottom-Up model failed on these questions while our model correctly predicts the answers. The reason is that our visual concept recognizer learns a robust mapping between the image and the visual concepts. The question information used for grounding won't disturb the recognizing of a group of visual concepts.

TABLE VII
PERFORMANCE (%) ON VQA-CP V2 TEST SET AND VQA V2 VALIDATION SET

Model	Image Feature	VQA-CP v2 test				VQA v2 val			
		Yes/No	Number	Other	Overall	Yes/No	Number	Other	Overall
NMN [4]	ResNet	38.94	11.92	25.72	27.47	73.38	33.23	39.85	51.62
Bottom-Up + Q-Adv + DoE [27]	Bottom-Up	65.49	15.48	35.48	41.17	79.84	42.35	55.16	62.75
MuRel [6]	Bottom-Up	42.85	13.17	45.04	39.54	84.03	47.84	56.25	65.58
GVQA [1]	ResNet	**57.99**	13.68	22.14	**31.30**	**72.03**	31.17	34.65	48.24
SAN [38]		38.35	11.14	21.74	24.96	68.89	**34.55**	**43.80**	**52.02**
Ours: DAG		40.84	**13.86**	**29.58**	30.46	69.32	33.02	40.14	50.31
Bottom-Up [2]	Bottom-Up	41.56	12.19	43.29	38.04	81.18	42.14	**55.66**	**63.48**
Ours: DAG w. Q		41.05	11.32	40.28	36.09	**81.97**	**42.55**	54.42	63.21
Ours: DAG (*Full Model*)		**43.02**	**15.83**	**46.41**	**40.75**	81.05	42.47	54.53	62.91

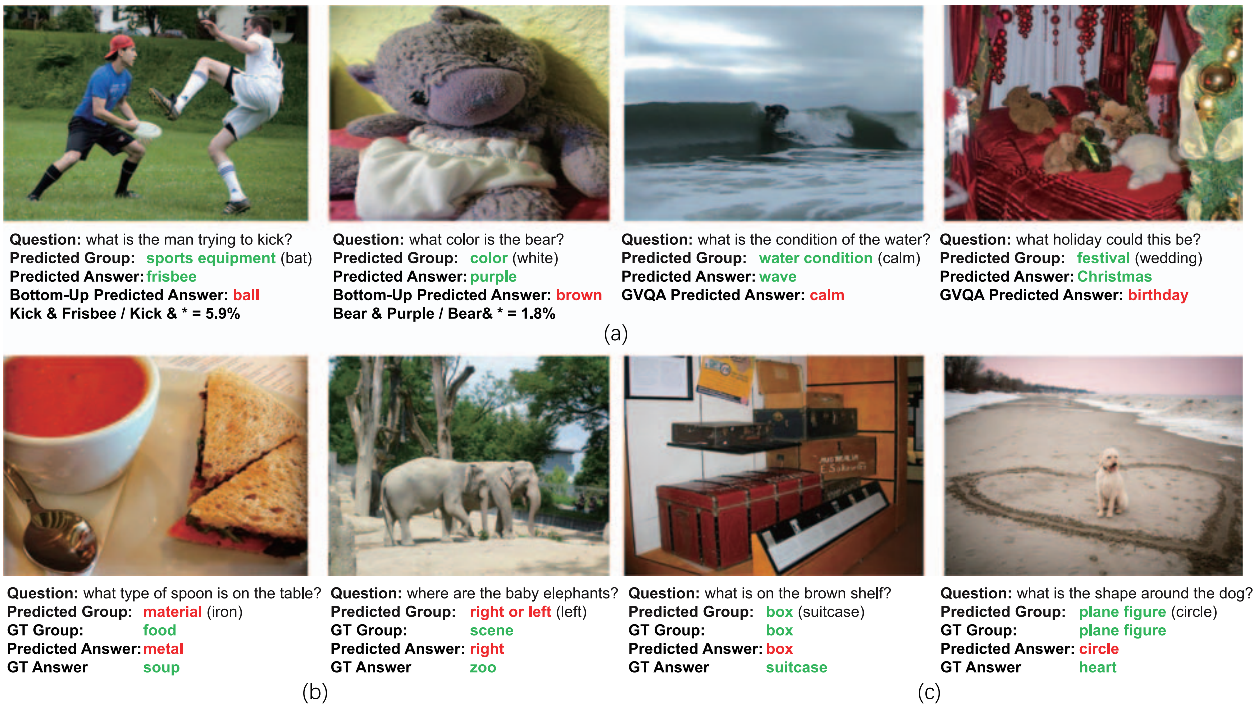
Question: what is the man trying to kick?
Predicted Group: sports equipment (bat)
Predicted Answer: frisbee
Bottom-Up Predicted Answer: ball
Kick & Frisbee / Kick & * = 5.9%
Question: what color is the bear?
Predicted Group: color (white)
Predicted Answer: purple
Bottom-Up Predicted Answer: brown
Bear & Purple / Bear& * = 1.8%
Question: what is the condition of the water?
Predicted Group: water condition (calm)
Predicted Answer: wave
GVQA Predicted Answer: calm
Question: what holiday could this be?
Predicted Group: festival (wedding)
Predicted Answer: Christmas
GVQA Predicted Answer: birthday
(a)
Question: what type of spoon is on the table?
Predicted Group: material (iron)
GT Group: food
Predicted Answer: metal
GT Answer soup
Question: where are the baby elephants?
Predicted Group: right or left (left)
GT Group: scene
Predicted Answer: right
GT Answer zoo
Question: what is on the brown shelf?
Predicted Group: box (suitcase)
GT Group: box
Predicted Answer: box
GT Answer suitcase
Question: what is the shape around the dog?
Predicted Group: plane figure (circle)
GT Group: plane figure
Predicted Answer: circle
GT Answer heart
(b)
(c)

Fig. 8. Qualitative results of our method. (a): The samples correctly answered by DAG model. The two samples on the left show that the DAG doesn't overuse the language priors and can recognize the relative rare situation. In addition, our model can better focus on learn the concepts which contains less samples and such as the "wave", "Christmas". (b) & (c): The samples mistakenly predicted by DAG model. DAG can provide intuitive information when the model predicts the wrong answer which is helpful for improving the model. Samples in (b) show that DAG predicts the wrong answer because of misunderstanding the question. Samples in (c) show that DAG is not capable of correctly recognizing the related concepts in the image. (In the figure above, the word in the brackets is one sample of the concepts in the given group.)

Besides, though our model neglects the language priors in VQA v2 which can improve the performance, the performance of our model doesn't drop much, compared to the baseline Bottom-Up and SAN. Moreover, since our method mainly focuses on the question related to recognizing visual concepts, our results on yes/no questions are similar to baselines.

Analysis of the components of DAG: We evaluate the performance of each component to comprehensively measure the performance of DAG. We check the accuracy of GPN to evaluate the ability of question parsing. In DAG (Bottom-Up feature) on VQA-CP v2, the accuracy of classifying all groups is 78.34%, where the accuracy of classifying three types of question is 99.23%, the accuracy of classifying the group index in other-type and number-type questions is 66.62%. It shows that understanding the human questions is more difficult, compared to template generated questions in GQA, where DAG achieves 93% accuracy. The qualitative results are shown in Fig. 8. In addition, to diagnose the difficulty of understanding human questions, we visualize part of the confusion matrix of the group prediction result, as shown in Fig. 9. We find that some groups (outdoor scene, left/right, outside/inside, etc.) often are mistakenly predicted. The concepts of these groups can all be used to answer the "where" question. The confusion matrix shows that our model overuses the left/right to answer the "where" questions rather than using other groups of concepts (the last column has a relatively large number). It shows that

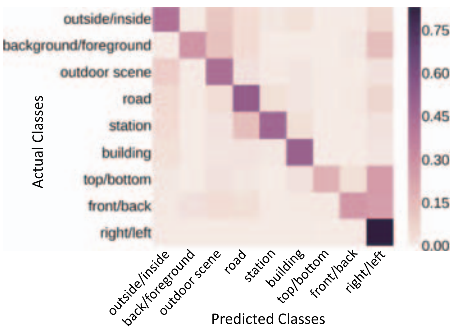

Fig. 9. Part of the confusion matrix on group prediction. The bigger the element on the diagonal, the better.

TABLE VIII
THE ACCURACY (%) OF DCR WITH GROUND TRUTH GROUP INDEX ON RECOGNIZING SELECTED GROUPS OF CONCEPTS ON VISUAL GENOME AND VQA-CP V2

Group	VG	VQA-CP v2
animal	81.53	74.37
body part	51.28	27.59
plane figure	52.67	26.25
vehicle part	69.64	25.73

the DAG is struggling to speculate the intent of the ambiguous questions.

We also evaluate the performance of recognizing each group of concepts. To measure the performance of DCR without considering the group prediction, we feed the ground truth group index to the DCR to predict the answer. Table VIII shows the accuracy of recognizing some group of concepts (We also list the recognizing accuracy on Visual Genome (Experiment on Section IV-A) on Table VIII as references). We can find that some groups of concepts are difficult to classify, such as “plane figure”, “body part”. Besides, we observe that the accuracy of classifying “vehicle part” is relatively high on Visual Genome, but is relatively low on VQA-CP v2. This indirectly indicates that these concepts may be difficult to ground, because the main difference between two experiments is that one uses the ground truth region, and one uses the predicted region.

Question information and Overusing Language Priors: Comparing the results of DAG with DAG w. Q in Table VII, when we add question information on visual concept recognition procedure and keep anything else be same, DAG w. Q obtains lower accuracy on VQA-CP v2 and higher accuracy on VQA v2. This phenomenon demonstrates that if the question information takes part in visual recognition, the model will uncontrollably utilize the language priors, rather than truly understanding the image content. And the DCR module is an effective alternative to output the answer in the VQA model.

Different usages of concepts clustering results: Comparing the results of DAG with GVQA in Table VII, two models achieve similar performances on overall accuracy, but our model achieves much higher accuracy on other-type questions which most relates to visual concept recognition. The main reason of the performance gap is from the different usages of concept clustering results. GVQA formulates the group index as a feature, so that they can only use basic FC layer with flat label space to learn the visual concept. In contrast, since DAG uses concept clustering results to build a structural label space which allow us to design the Dynamic Concept Recognizer and brings two advantages: 1) expand the capacity of visual recognizer by decomposing itself to many sub-recognizers, and 2) easy to utilize the semantics of the label to concentrate on distinguishing relevant concepts. Note that, DAG doesn’t improve much on yes/no-type questions compared to GVQA, because DAG mainly focuses on improving visual concept recognition. The yes/no sub-recognizer in DAG can be easily replaced by another sophisticated module to improve accuracy.

V. CONCLUSION

In this paper, we propose a Dynamic Answer Generator (DAG) that can utilize the semantic meaning of labels in answer prediction. Specifically, instead of using a simple MLP to recognize the image content and output the label in a flat label space as the answer, DAG proposes a novel dynamic concept recognizer which contains many sub-recognizers for different visual sub-tasks to output the answer in a structural label space which represents the relations between labels. Experimentally, we show that our usage of the semantics of the labels improves the performance of visual recognition, facilitates the performance of the question answering, and reduces the risks on overusing language priors.

REFERENCES

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, “Don’t just assume; look and answer: Overcoming priors for visual question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4971–4980.

[2] P. Anderson *et al.*, “Bottom-up and top-down attention for image captioning and VQA,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6077–6086.

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Learning to compose neural networks for question answering,” in *Proc. NAACL-HLT*, 2016, pp. 1545–1554.

[4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, “Neural module networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 39–48.

[5] S. Antol *et al.*, “VQA: Visual question answering,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2425–2433.

[6] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, “Murel: Multimodal relational reasoning for visual question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1989–1998.

[7] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[8] S. Dumais and H. Chen, “Hierarchical classification of web content,” in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2000, pp. 256–263.

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *Proc. Conf. Empirical Methods Natural Language Process.*, 2016, pp. 457–468.

[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6904–6913.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[13] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 804–813.

[14] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–20.

[15] D. A. Hudson and C. D. Manning, "GQA: A new dataset for compositional question answering over real-world images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 6700–6709.

[16] I. Ilievski and J. Feng, "Multimodal learning and reasoning for visual question answering," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 551–562.

[17] J. Johnson *et al.*, "Inferring and executing programs for visual reasoning," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2989–2998.

[18] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1983–1991.

[19] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 170–178.

[20] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[21] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[22] L. V. D. Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[23] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[24] H. Noh, P. Hongsuck Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 30–38.

[25] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[26] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3942–3951.

[27] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 1541–1551.

[28] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6517–6525.

[29] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 2953–2961.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[31] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 4974–4983.

[32] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 3667–3677.

[33] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question type guided attention in visual question answering," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 151–166.

[34] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4613–4621.

[35] A. Sun and E.-P. Lim, "Hierarchical text classification and evaluation," in *Proc. IEEE Int. Conf. Data Mining*, 2001, pp. 521–528.

[36] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Computat.*, vol. 24, no. 9, pp. 2508–2542, 2012.

[37] D. Teney, L. Liu, and A. V. D. Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 804–813.

[38] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 21–29.

[39] B. Zhao, F. Li, and E. P. Xing, "Large-scale category structure aware image categorization," in *Proc. Advances Neural Inf. Process. Syst.*, 2011, pp. 1251–1259.

Difei Gao received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently pursuing a Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests mainly include computer vision, pattern recognition, machine learning, and, in particular, visual question answering and commonsense reasoning.

Ruiping Wang (Member, IEEE) received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2010. He has worked as a Post-Doctoral Researcher with Tsinghua University from 2010 to 2012, and a Research Associate with University of Maryland at College Park from 2010 to 2011. In 2012, he joined the Faculty of ICT, CAS, where he has been a Professor since 2017. His current research interests include computer vision, pattern recognition, and machine learning. He is currently an Associate Editor for Pattern Recognition, Neurocomputing, and The Visual Computer. He has served as Area Chair for IEEE WACV 2018-2020, ICME2019/2020, IJCB 2020, ICPR 2020, and CVPR 2021. He is a member of the IEEE.

Shiguang Shan (Senior Member, IEEE) received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. In 2002, he joined ICT, CAS, where he has been a Professor since 2010. He is currently the Deputy Director of the Key Laboratory of Intelligent Information Processing, CAS. He has authored over 200 papers in refereed journals and proceedings in computer vision and pattern recognition. His current research interests include computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics. Dr. Shan was a recipient of the China's State Natural Science Award in 2015 and the China's State S&T Progress Award in 2005 for his research work. He is an Associate Editor of several journals. He has served as the Area Chair for a number of international conferences.

Xilin Chen (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively.

He is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 300 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multi-modal interfaces. He is currently an Associate Editor of the IEEE Transactions on Multimedia, and a Senior Editor of the Journal of Visual Communication and Image Representation, a Leading Editor of the Journal of Computer Science and Technology, and an Associate Editor-in-Chief of the Chinese Journal of Computers, and the Chinese Journal of Pattern Recognition and Artificial Intelligence. He served as an organizing committee member for many conferences, including general Co-Chair of FG13/FG18, Program Co-Chair of ICMI 2010. He is/was an Area Chair of CVPR 2017/2019/2020, and ICCV 2019. He is a fellow of the ACM, IEEE, IAPR, and CCF.