Research on visual Question answering System Method based on deep learning

An Chang, Anna Wang

Abstract— **Visual Questions and Answers (VQA) is a new research task combining computer vision (CV) and natural language processing [1]. Given an image and a natural language problem, it needs to combine the information of the two modes, that is, according to the visual features of the image and the text features of the problem. And then combine the information of the two features to get the right answer. Firstly, the feature extraction method and feature fusion method of visual and text are introduced. Secondly, the effect of attention mechanism on visual question-answering model is introduced. The data set used to train and evaluate the visual question answering system was then reviewed. Finally, we discuss the future direction of the field**

Index Terms— **Computer vision,Visual questions and answers,Natural language processing, Attention mechanism**

I. INTRODUCTION

Visual Q&A is a task that combines computer vision and natural language processing to stimulate research and push the boundaries of the two fields. On the one hand, computer vision studies methods for acquiring, processing and understanding images [2]. Its aim is to teach machines how to see the world. Natural language processing, on the other hand, is the field that promotes computer and human interaction in natural language, that is, teaching machines how to read and understand language. Both fall under the category of artificial intelligence and can share research methods in machine learning. However, they are not in sync in their development. Both areas have made significant progress in achieving their respective goals. The combined explosion of visual and textual data is driving the convergence of efforts in both areas. For example, image subtitle generation [3]. That is, given a picture, and then generate a corresponding description of the picture.

One of the main differences between visual q&a and other tasks in computer vision is that the questions to be answered are not determined until runtime. In traditional questions, such as segmentation or object separation algorithms, the individual questions to be answered are predetermined, and only the input image changes. What form the questions will take in the visual q&A is unknown, as is the case with the opera. In this sense, it more closely reflects the challenges of understanding images in general. Visual q&A is related to text Q&A tasks, where the answers will be found in a specific text narrative (i.e. reading comprehension) or in a large knowledge base (i.e. information retrieval). Text question-and-answer has been studied for a long time in the field of natural language processing, and VQA is an extension of additional visual support information. The added challenge is important because images are much higher in dimension and are often noisier than plain text. In addition, the image lacks the structure and grammar rules of the language, which makes the visual q&a need more technical support to complete.

Manuscript received October 10, 2020

II. FEATURE EXTRACTION

A. IMAGE FEATURE EXTRACTION

In visual question-and-answer tasks, we generally need to extract the features of image data first [4]. There are many purposes for pre-extracting image features by deep learning, such as removing redundant information by reducing dimension of data, reserving key information that is useful for tasks, and reducing computation and parameter storage. In more general image processing tasks, the purpose of feature extraction is mostly to reduce the single calculation amount of deep learning model, so that the extracted features can be directly used for calculation in training [5].

Features obtained by traditional image feature extraction methods are called manual features, mainly because such methods are determined by human purpose. SIFT (scale invariant feature transformation) is commonly used to achieve affine transformation invariance of feature points, and HOG method is used to reduce the influence of lighting conditions on images. SIFT method is used to find the points in image feature space that will not change due to different light and direction, such as some corners, edges, dark points and bright points, etc. The method compares the gray value of the special point with that of the neighboring points, selects the maximum point and the minimum point, and leaves the stable point as the feature point to save its location information and scale information. The HOG method is to segment the image in a small region, calculate the gradient for each region segmented, and represent the gradient with histogram. The method is invariable to the illumination change and some geometric change of the image and is often used in human body detection.

One of the typical methods for deep learning image feature extraction is to adopt the region selection method, which was proposed by Girshick[6] in the target detection task and then widely used in the image target detection and segmentation task. The aim of this method is to extract the features of different objects in the image. As shown in formula 1 below:

$$feature = \mathrm{RoI}(\mathrm{CNN}(\mathrm{proposal}(\mathrm{I}))) \quad (1)$$

In formula 1, the Region selection method inputs picture I into the convolutional neural network (CNN) for feature extraction, and generates about 2000 regional proposal Windows. The features output by the convolutional layer of the network are mapped to the proposal Windows, and fixed-size features are generated through the RoI pooling layer.This method can get the boundary of the object and its

key features well, but the problem is that it ignores some information, such as the wall in the background, grass and so on. In a visual question-and-answer task, the question may involve not only the obvious object in the image but also the question about the insignificant information. Therefore, although this method can obtain clear object feature detection, it may omit the background feature and part of the object feature whose size is outside the preset marquee, which may be detrimental to the answer of some questions in the visual question-and-answer task.

B. TEXT FEATURE EXTRACTION

Text information in visual Q&A database is unstructured information that is difficult for computers to process [8]. In order to enable computers to process text information, we need to transform it. So we need to use word embedding to map text to the feature space. There are mainly two models skip-gram and CBOW in word embedding. In skip-gram model, the model obtains the probability distribution of the context vector of the input word through neural network training according to the input word given and the window with preset size. CBOW, on the other hand, is just the opposite. It gives the context and trains the input words, and outputs the probability distribution of the predicted words based on the context given by the input under the window size.

As for the feature extraction of the sentences in question, since the convolutional neural network cannot process the sequence information well, the cyclic convolutional neural network is used to iterate the word features in the sentences to obtain the sentence features. However, the gradient in cyclic convolutional networks will disappear with the increase of iteration, so LSTM and GRU[9] are proposed. LSTM[10] introduced gating to control the time length of self-cyclic accumulation information, and applied SIGMOID activation function to limit the value to a limited range, so as to retain the memory of the past gradient within a certain period of time and prevent the gradient from exploding due to accumulation. The GRU simplifies the threshold and only USES the update gate and reset gate to achieve a similar effect to the LSTM.

III. FEATURE FUSION

What a visual task, the visual feature extraction and the extraction of text feature in their respective fields have already has many performance method, and the fusion of visual features and text is still a new field, the results of the two corruption affected the visual effect of question and answer tasks is good or bad, so many scholars put forward many fusion method according to the characteristics of the different areas.

A. FUSIONMETHODBA BASED on LSTM

In many methods, LSTM is used to extract the features of the question, and LSTM is used to generate the answer, and the embedded words and CNN features are input to generate the answer of the question. "I" refers to the extracted image features, and "Norm I" refers to L2 normalization of the semantic information vector (1024 dimensions) extracted by VGG. This part describes the extraction of image features. The semantic information contained in the LSTM word - by - word extraction problem is responsible for the extraction of problem characteristics. Finally, image features and question features are fused by means of dot product and sent into a multi-layer MLP to generate answer output according to Softmax. The algorithm is targeted at the data set in the form of open and multi-choice, and the answer consists of one or two words.

The VIS + LSTM[11] model addresses the use of a single word as the answer to a visual question, so that the visual question can be considered as a multi-categorization question and the answer can be measured using existing accuracy evaluation criteria. The backbone structure USES the VGG19 pre-training model to extract a vector with 4096 dimensions (the last hidden layer). In order to make the image feature match the dimension of the word vector, affine/linear change is used to transform the image feature vector into 300/500 dimensions. Figure 1 shows the feature fusion method based on LSTM:

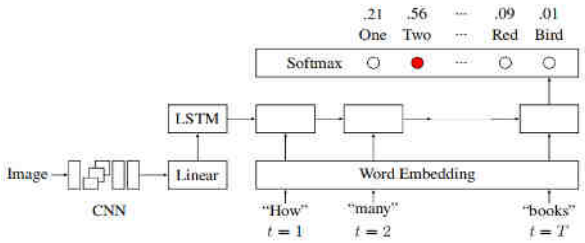

Figure 1: Feature fusion method based on LSTM

B. THE METHOD BASED ON GRID DIVISION

In the grid-based approach [13], the first step is naturally to project the grid on the image. By projecting a unified grid on the image, each grid contains different local features. The image and problem features are used to compute the correlation coefficients of the grid and problem to generate weights. After the grid is applied, the correlation of each region is determined by the specific problem. The whole process is shown in Figure 2:

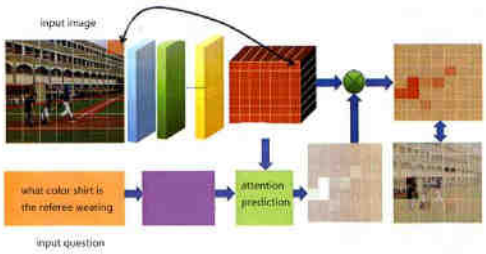

Figure 2: Feature fusion process based on grid generation

In the Stacked Attention Network(SAN)[14], the layer of Attention is assigned a single layer weight that USES the question and Softmax activation function to calculate the distribution of Attention throughout the image to the CNN feature map, which ACTS on the CNN feature map and then USES weights and combinations in the locations of spatial elements to provide some areas of enhanced Attention. Then the combined eigenvectors and the eigenvectors of the problem are sent to the Softmax classification layer to predict the answer. Finally, this approach is generalized to produce multiple stacked layers of attention so that the complex relationships of multiple targets in the system can be fully modeled.

A similar approach is Spatial Memory Network, where the attention mechanism USES word-guided attention to predict attention distribution by evaluating correlation generation

between image blocks and problem single words. Then, the coding characteristics and weighted visual characteristics of the whole question are used to predict the answer[15], which is divided into two models, the one-hop model: the coding characteristics and weighted visual characteristics of the whole question are used to predict the answer. Two-hop model: Combinations of visual features and problem features are cyclically used to calculate the distribution of attention
Modified Dynamic Memory Network(DMN) consists of three parts, which are input module, episodic Memory module and response module respectively. The input module is responsible for feeding each word in the problem into an RNN to extract facts from the sentence, and then the positions in each spatial grid in the image are fed into the RNN like words to generate facts for the visual task. Subsets of these facts are passed multiple times by the episodic memory module, each time by learning to update the internal Memroy Representation. The answer module is responsible for predicting the answer using the final storage representation state and question input.

IV. ATTENTION

A. MONO-ATTENTION MECHANISM

In the study of attention mechanism, early researchers mainly considered the unidirectional problem attention mechanism, that is, the problem leading attention to the image area. As the idea expanded, in later studies, researchers began to consider another aspect of attention, that is, directing attention to the image regions associated with the question words. Subsequently, Lu et al. implemented the co-attention mechanism, which paid attention to both image features and problem features at the same time, and added the obtained common features into the model to influence the final feature generation [16]. Yu et al. also used the attention mechanism to extract the semantic concept information of spatial information and images, which effectively narrowed the gap between image and problem characteristics [17]. Z.yu et al. realized the use of attention mechanism to extract multi-modal features of images and problems, and the fusion of related feature representation to obtain multi-modal feature representation .AkiraFukui et al. 's improved algorithm based on MCB can fully demonstrate the basic structure of single-attention mechanism The purpose of the MCB model structure is to generate a concerned representation of the image based on the text. The joint embedding method based on MCB can effectively reduce the number of model parameters by reducing the dimension of image and text features through MCB processing. The main idea of MCB model is to influence the weight of image features by means of text representation through attention mechanism, so as to realize the extraction of image features related to problem information.

B. CO-ATTENTION MECHANISM

In addition to the visual attention, gradually the common attention of images and problems also attracted people's attention. Different from the single-attention mechanism, which only considers one-way attention mechanism, cooperative attention simultaneously considers mutual attention between images and problems. In addition, the cooperative attention mechanism can be created from the whole image and the whole sentence, and multiple cooperative attention machines can also be applied to pay attention to parts of the image and important words in the text. Based on the idea of joint attention mechanism, Duy-Kien et al proposed and implemented the mechanism of intensive common attention .
In his research, several common attention mechanisms are used to extract the important information of the problem and the important region information in the image. Firstly, the existing region selection algorithm is applied to generate the target region, and the joint concern mechanism is used to focus on the image region most relevant to the problem to generate the overall image features. In terms of the representation of the generation problem, the author input the text of the problem into the bidirectional LSTM network to obtain the initial vector representation of the problem. Then, by associating the picture content with the common concern mechanism, the author focused his attention on the problem words most relevant to the image to influence the final problem characteristics of model generation. The joint attention mechanism proposed by Duy-Kien et al can focus on the relationship between any image region and any question word, which will improve the accuracy of the answer to open questions and make it possible to solve the difficult problem of complex image-problem relationship. The calculation of the intensive common attention diagram and the representation of the image and problem participation are shown in Figure 3.

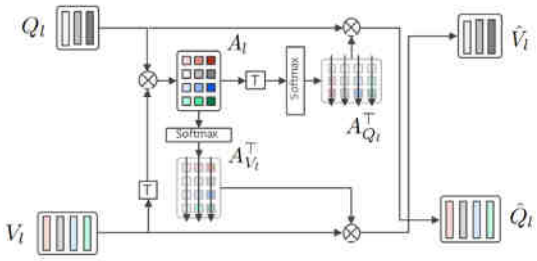

Figure 3: Calculation of intensive common attention diagrams and participative representations of images and problems

The construction method is as follows: the text of the problem is input into the two-way LSTM network to obtain the initial vector representation of the problem. Image feature extraction USES 152 layers ResNet, which has been pre-trained on the ImageNet data set, to generate the visual features of multiple image regions.
In general, the common concern mechanism will generate bidirectional related image-question feature pairs, which can more effectively combine the question information with the image content to answer the question.

V. DATASET

Several large data sets for visual question and answer tasks have emerged in recent years. Since most of the existing visual question-answering algorithms are based on data-driven training models, a good data set can help train a model with more generalization ability. Table 1 shows the summary contents of each large data set, which are introduced below.

Table 1: Introduction to the visual Q&A dataset

Dataset	Number of Images	Number of Questions	Average Questions Per Image	Average Questions Length	Average Answer Length	Manually Annotated QA
DAQUAR	1449	12468	8.60	11.5	1.2	YES
Visual7W	47300	327939	6.93	6.9	2.0	YES
COCO-QA	117684	117684	1.00	9.65	1.0	NO
VQA	204721	614163	3.00	7.38	3.82	YES

A. DAQUAR[18]

DAQUAR published in 2015 is the first Dataset published for visual question and answer task. It is taken from NYU Depth Dataset V2 Dataset which contains images and their semantic segmentation. These images are indoor scenes and each image is tagged with multiple tags. The combination of image-based questions and answers is generated in two ways. The first is to define a number of question templates and automatically generate q&A pairs according to the image labels. The second is to use manual tagging, in which volunteers answer automatically generated questions.

B. Visual7W[19]

Visual7W is a data set generated by Microsoft COCO images that contains fill-in-the-blanks and multiple selection questions. The fill-in-the-blanks questions are automatically generated by the template based on the image tags and are answered by human hands. The answers can be words or phrases.

C. COCO-QA[20]

CCO-QA is made from images in Microsoft's COCO data set. The q&A pairs are automatically generated based on the image description and mainly consist of four types of themes: object, number, color and location. The data set is characterized by having only one question per image and only a single word answer.

D. VQA[21]

VQA data set is the most widely used data set in visual Q&A tasks. Its image content is rich, which is not only derived from Microsoft COCO data set based on real scenes, but also contains abstract clipping scene pictures created by human and animal models. The questions and answers in the dataset are generated manually.

ACKNOWLEDGMENT

Aiming at VQA task, this paper first analyzes the existing visual question-and-answer related algorithms and technologies in detail, and further discusses the deficiencies of the existing visual question-and-answer algorithm research, and finally analyzes the future research direction of visual question-and-answer task and the scientific problems to be solved. The current model can achieve many good results on the data, but most of them are due to the data bias, and the model only makes good use of this bias. In general, when data sets are more balanced, algorithms tend to be less effective. And the performance of VQA and the way to ask questions is also relatively large. Humans do a much better job of identifying these kinds of problems than algorithms do. It is difficult to evaluate open questions, and many of them are replaced by multiple choice questions. However, this may make it easier for the algorithm to take advantage of the data bias. Although the balance of data will make the algorithm more efficient, this is the road that must be taken and the difficulty that the algorithm must overcome. Therefore, future data sets are bound to move in a larger and more balanced direction. It is also expected that there will be more excellent data, more excellent algorithms and more excellent evaluation indicators in this direction.

REFERENCE

1) Agrawal A, Kembhavi A, Batra D, et al. C-VQA:A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset[J].2017.
2) Ghosh S, Burachas G, Ray A, et al. Generating Natural Language Explanations for Visual Question Answering using Scene Graphs and Visual Attention[J].2019.
3) Shrestha R, Kafle K, Kanan C. Answer Them All!Toward Universal Visual Question Answering Models[J].2019.
4) J Lu, J Yang, D Batra, et al. Hierarchical question-image co-attention for visual question answering.In International Conference on Neural Information Processing Systems (NIPS),2016.
5) Santoro A, Raposo D, Barrett D G T, et al. A simple neural network module for relational reasoning[J].2017.
6) Raposo D, Santoro A, Barrett D, et al. Discovering objects and their relations from entangled scene representations[J].2017.
7) Yujia Li, Daniel Tarlow, Marc Brockschmidt, et al. Gated graph sequence neural networks.ICLR,2016.
8) Battaglia P W, Pascanu R, Lai M, et al. Interaction Networks for Learning about Objects,Relations and Physics[J].2016.
9) Johnson J , Hariharan B , Maaten L V D , et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
10) Weston J, Bordes A, Chopra S, et al. Towards AI-Complete Question Answering:A Set of Prerequisite Toy Tasks[J].Computer Science,2015.
11) He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015.
12) He D, Zhao X, Huang J, et al. Read,Watch,and Move:Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos[J].2019.

13)Malinowski M, Fritz M. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input[J].2014.
14) Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J].2014.
15)Shih K J, Singh S, Hoiem D. Where To Look:Focus Regions for Visual Question Answering[J].2015.
16)Ilievski I, Yan S, Feng J. A Focused Dynamic Attention Model for Visual Question Answering[J].2016.
17)Teney D, Anderson P, He X, et al. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge[J].2017.
18) Zhu C, Zhao Y, Huang S, et al. Structured Attentions for Visual Question Answering[J].2017.
19)Yu D, Fu J, Mei T, et al. Multi-level Attention Networks for Visual Question Answering[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).IEEE,2017.
20)Yu Z, Yu J, Fan J, et al. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering[J].2017.
21)Zhou B, Tian Y, Sukhbaatar S, et al. Simple Baseline for Visual Question Answering[J].Computer Science,2015.