Knowledge-based Reasoning for Visual Question Answering

Sai Charan Tej Bandi, Raghuvamshi Joshi, Siddhartha Sriram, Sumanth Varma, Sumit Rawat
Arizona State University, University Drive, Tempe AZ 85281
{sbandi6, rjoshi20, svinjam, svitrout, srawat7}@asu.edu

Abstract

Through this project, we aim to analyze the existing model (4) which uses Knowledge-Based Reasoning to answer questions related to an image. Given an image and a question as input, the model uses features directly from the image and reasoning from a knowledge-base to provide appropriate answers. After analyzing this approach, we have developed a different *VQA* pipeline that incorporates the classical *CNN-LSTM* model and knowledge-base to answer common template questions discussed in the above paper. We also aim to introduce new templates that can be answered by our *VQA* pipeline and discuss possible future enhancements.

1 Introduction

There are many Visual Question Answering methods, but the novel part of this approach is that apart from answering just about the features of the image it answers various other questions with the help of the knowledge databases and provides the reason too. So, there are two parts of the process, the first is the extraction of the image features and the second is understanding the question (NLP part). As the questions are open-ended there are infinitely many possible patterns of the question. To reduce that, the model takes the question in the form of certain templates. The templates are not very rigid, they can be used to form many questions and also it is fairly easy to add new templates. As part of decoding the questions it maps the question elements into the pre-defined templates. Once the question is mapped into the templates it makes a query to the knowledge base.The query returns the answers along with the path of the concepts in the data which gives us the reasoning used to answer the questions.

2 Data set / Task Description

2.1 Data set

The different approaches to model Visual Question Answering solutions as discussed in this paper use two data sets i.e. The *VQA* data set and the COCO data set(**?**). The *VQA* data set is the largest available collection of data sets online that contains human-annotated questions on the Microsoft COCO data set. The latest *VQA* data set consists of a set of 443,757 Training questions, *214,354 Validation questions*, and *447,793 Test questions*. The MS COCO data set encompasses a collection of images representing everyday objects in their natural context. These images are used for object recognition, semantic scene labeling, and visual question answering. (**?**) uses the *COCO-QA* data set, which is a question-answer data set automatically extracted from captions in the MS COCO data set. The questions comprise entities like object, color, number, and location.

2.2 Task Description

Visual Question Answering deals with answering questions based on an image where the type or nature of the questions is not determined beforehand. The task to solve this problem, therefore, should involve a certain degree of image understanding, natural language processing. The simplest task in *VQA* could be answering questions that directly deal with the composition of the image e.g. identification of an object or a color. More complex tasks in-

volve finding the right attention for the model i.e. a model robust to the variety of irrelevant attributes that may lie in an object or a question (**?**). The other task associated with this problem seeks to explicitly map and reasoning of a question-answer based on existing knowledge sources and image understanding. In this paper, we attempt to understand and analyze different models used for Visual QA.

3 Methods / Implementation

3.1 Deeper LSTM + normalized CNN for Visual Question Answering

In the deeper LSTM + normalised CNN model(1) the question and the image are simultaneously fed to the network, where the question is fed to an LSTM model to get the final embedding of the entire question and the image is fed to a CNN model which extracts the feature vector of the entire image. These feature vectors are taken from the two models and a point-wise multiplication is applied to these to vectors to get a new vector which is then fed to a fully connected network which predicts the answer to the given question.

3.2 Hierarchical Question-Image Co-Attention for Visual Question Answering

In most of the Visual Question Answering (VQA) solutions presented in several papers, the primary focus was on the attention model of the image only. This means that there have been implementations of feature maps of the images that used attention mechanism to add weights to those pixels of the image which contained an important feature. This paper(**?**) debates that the attention models are important not only for images but also for the question asked. Therefore, in this model, the authors implemented a co-attention model where it simultaneously applies the attention to the question asked and extracts the words to be focused on and also applies the attention to the image to generate attention feature maps. It then selects a feature map according to the attention word with maximum score and predicts the output/answer to the asked question.

3.3 Tensorflow Implementation of Stacked Attention Networks for Image Question Answering

In the stacked attention networks for image question answering(**?**) approach, unlike the previous implementations where there were single attention mechanisms for image and question, they use multiple attention mechanisms to start with broad attention for the overall image and gradually narrowing down their area of focus in the image. In this model, the question is given to LSTM blocks and then to an attention layer to get the attention word or a CNN block which gives n-gram embeddings which are given to the attention layer to get the attention word. Now, instead of getting one feature map from the image, it extracts attention top attention words and gets the most relevant feature map for the least attention word in the selected ones. After this, it progressively selects attention words to narrow down the attention area in feature maps using several stacked attention layers, and finally, it predicts the answer to the question.

4 Existing model - KBVQA

Most of the existing VQA methods have developed tools and techniques to interactively answer questions about images without explicitly employing any reasoning. The method proposed by KBVQA approaches the VQA task from a reasoning standpoint, not usually explored by deeper neural models. This method extracts relevant objects and concepts from the image and links them to relevant parts of a Knowledge-Base. These visual concepts are stored in the form of RDF triples and linked to DBpedia entities. The method then answers the questions by parsing natural language queries into a set of regular expressions to extract slot phrases. These phrases are mapped to KB entities and appropriate queries are formed. By mapping visual concepts to concepts in KB, the method performs explicit reasoning about an image. This method performs better than other deep neural network approaches for questions requiring commonsense knowledge but does not address simpler

questions that rely on visual concepts and simple facts.

5 Proposed Enhancement - VQA Pipeline

During our analysis of the paper, we realized that the model handles some visual concepts poorly compared to LSTM based approaches (1). To strike a balance between deep neural-network methods and Knowledge-Base based methods, we propose a pipeline that involves the extraction of objects from a normalized LSTM model and links that to KB sources like wikidata to obtain reasoning based answers on the object extracted before. The specifics of the proposed method is illustrated below.

5.1 Concept Detection from Image

While the KBVQA approach uses different CNN models to obtain features of an image, our approach feeds question and image separately into a normalized LSTM with Glove word embeddings and a pretrained VGG-16 model. The question and image features obtained are combined using a perceptron. The question fed simply facilitates the detection of the relevant object or concept from the image.

5.2 Extracting facts from KB

Once the concept is extracted, the next step is to form questions that can be queried on these concepts with the help of a knowledge-base like DBpedia. Unlike KBVQA, we use SPARQLwrapper instead of quepy to build and execute queries that link to KB. While this restricts the ability to parse some questions, we can still test some templates used in KBVQA in addition to providing new templates of questions. These new templates listed below query different concepts or are variations of existing templates.

6 Findings

6.1 Concept Detection

As seen in Fig. 1, our *VQA* model can predict the object from the given image with a percentage confidence score. The confidence varies from about *50%* accuracy to as high as

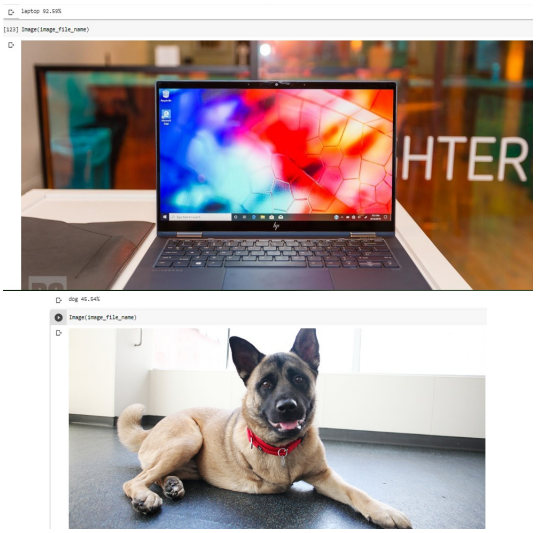


Figure 1: Detect object from image with percentage confidence score.

92% accuracy for some images. Using this as a base, *VQA* pipeline takes user query on the image and generates results as described in the next subsection.

6.2 Query execution

Once an object or concept has been identified, we pass the user question to DBpedia using SPARQL query to obtain answers extracted from the Knowledge-Base. Some of the query execution results are illustrated in the appendix below.

7 Results

As described in previous sections, our VQA pipeline is able to detect an object from the image and extract the Knowledge-Base to answer related queries about the image. While the accuracy of the VQA object detection model could be a bottleneck for information retrieval for the latter parts, we believe this model can efficiently detect single object / concepts from the image as it uses a better deep learning architecture than the one used in *KBVQA*. We also believe that given the simplicity of the pipeline, we can extract multiple related entities from *DBpedia* for a given object or concept.

8 Contribution

Through the VQA pipeline that we have developed, we have successfully implemented

some templates of questions that the existing *KBVQA* model attempts to answer. In addition to this, the following new templates have been created and our VQA pipeline provides accurate and in some cases, verbose answers for the new question templates.

- **CommonThings** - What is common between the <object> and <object2>?
- **AnimalClass** - What is the class of the <object>
- **TellMeAbout** - tell me about the <object>?
- **ObjMadeOf** - what is the <object> made of?
- **ComponentsOf** - what are the ingredients of this food?
- **LengthOf** - what is the length of the <object>?
- **IntroDate** - when was the <object>% introduced?

9 Conclusion

In this paper, we attempt to compare and analyze existing and novel techniques used in Visual Question Answering. The paper explored three different models focused on determining answers based on features explicitly obtained from the recurrent networks e.g. attention, reasoning. *KBVQA* model in particular, addressed the task of adding explicit reasoning to *VQA* task. We implemented a basic pipeline to extend this model by incorporating new templates and separating the task of object detection and knowledge based query answering. This model answers fairly basic questions and visual concept questions not targeted by *KBVQA* model. The model has its own limitations. The object detection is limited to single objects and the robustness of the knowledge extraction need to be increased further. We believe that this work would provide insights into adding new template support for the existing *KBVQA* model.

10 Appendix

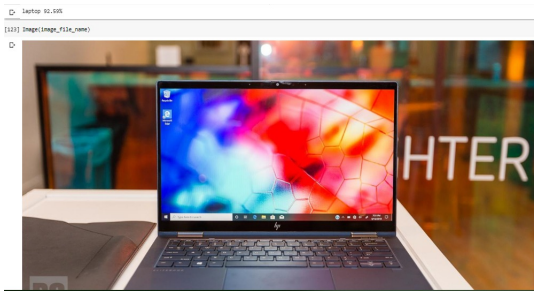
Q: What is common between the object and cat?
A: Mammal, Animal, BiologicalLivingObject
Template: CommonThings

Q: What is the class of the animal?
A: The following is / are the Animal class of dog : Canis
Template: AnimalClass

Q: Tell me about the object.
A: A train is a form of rail transport consisting of a series of that usually runs along a rail track to transport cargo or passengers.
Template: TellMeAbout

Q: What is the length of the object?
A: The average length of a train is 1.28 units.
Template: LengthOf

Q: When was the object introduced?
A: The train was introduced in 1862
Template: IntroDate



Q: What is the object made of?
A: The laptop is made of aluminum
Template: ObjectMadeOf

Q: What are the ingredients of this food?
A: The following is / are the ingredients of cake: Butter, Floor, Egg_(food), Sugar, Vegetable_oils_and_fats
Template: ComponentsOf

References

[1] Jiasen Lu, Xiao Lin, Dhruv Batra and Devi Parikh, 2015. Deeper LSTM and normalized CNN Visual Question Answering model.

[2] jiasenlu, jw2yang, dbatra and parikh, 2017. Hierarchical Question-Image Co-Attention for Visual Question Answering.

[3] Zichao Yang, Xiaodong He, Jianfeng Gao2, Li Deng, Alex Smola, 2016. Stacked Attention Networks for Image Question Answering.

[4] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel 2017. Explicit Knowledge-based Reasoning for Visual Question Answering.