CV_Final_G16

by Shivram Dubey

Submission date: 01-May-2018 08:30AM (UTC+0530)

Submission ID: 956728214

File name: Group16.pdf (1.09M)

Word count: 3699

Character count: 17981

Visual Question Answering Using Deep Learning

Ravi Prakash   Vaishnav Murali   Yash Srivastava
Indian Institute of Information Technology, Sricity
Chittoor, Andhra Pradesh, India
{ravi.p15,murali.v15,srivastava.y15}@iiits.in

Abstract

Visual Question Answering is a new research task in the area of Computer Vision and Deep Learning. It combines both Computer Vision and Natural Language Processing problems together to answer basic 'common sense' questions on query images. In this report, we discuss about the pre-existing work in this research area as well as the techniques used in the Ask Your Neurons model, an approach proposed by Malinowski et al. and published in ICCV 2015. We also show our results on the reproduced work of their paper and finally discuss the scope and future that lies ahead in this field.

1. Introduction

Rise of deep learning has helped to solve many real world human vision challenges like classification, object detection etc. Visual Question Answering [1] is a task that has emerged in the last few years and has been getting a lot of attention from the machine learning community. The task typically involves showing an image to a computer and asking a question about that image which the computer must answer. The answer could be in any of the following forms: a word, a phrase, a yes/no answer, choosing out of several possible answers, or a fill in the blank answer.

Visual question answering is an important and appealing task because it combines the fields of computer vision and natural language processing. Computer Vision techniques are used to understand the image and NLP techniques are used to understand the question. Moreover, both must be combined to effectively answer the question in context of the image. This is challenging because historically both these fields have used distinct methods and models to solve their respective tasks.

The VQA problem helps to achieve many day-to-day problems. The most visible problem is using VQA as a device to help to blind and visually impaired individuals, enabling them to get information about images on devices as well as scenes in the real world. For example, as a blind user scrolls through Facebook, a captioning system can describe the image and then a VQA system is used to query the image to get more details about the scene. VQA can also be used to improve human-computer interaction as a natural way to query visual content. A VQA system can also be used for image retrieval, without using image meta-data or tags.

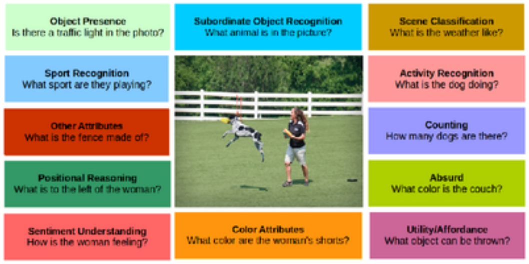
Object Presence
Is there a traffic light in the photo?
Subordinate Object Recognition
What animal is in the picture?
Scene Classification
What is the weather like?
Sport Recognition
What sport are they playing?
Activity Recognition
What is the dog doing?
Other Attributes
What is the fence made of?
Counting
How many dogs are there?
Positional Reasoning
What is to the left of the woman?
Absurd
What color is the couch?
Sentiment Understanding
How is the woman feeling?
Color Attributes
What color are the woman's shorts?
Utility/Affordance
What object can be thrown?

Figure 1. Variety of questions that can be asked with respect to an image

In this paper we present the reproduced work by Malinowski *et al.* [9], where they have proposed a Ask Your Neuron model, which combines both, a Convolution Neural Network (CNN) and Long-Short Term Memory (LSTM), to obtain a representation vector which gives us the scores for each probable answer. We report the experiments performed and the accuracy obtained over the VQA dataset [3].

In the work, we assume that the answers consist of only a single word, which allows us to treat the problem as a classification problem. This also makes the evaluation of the models easier and more robust, avoiding the prickly evaluation issues that plague multi-word generation problems.

2. Related Work

Recent work has made significant progress using deep neural network models in both the fields of computer vision and natural language. For computer vision, methods based on Convolutional Neural Network (CNN) achieve the

	Number of Images	Number of Questions	Avg. questions per image	Average question length	Average answer length	Q/A generation
DAQUAR	1,449	12,468	8.60	11.5	1.2	Human
Visual7W	47,300	327,939	6.93	6.9	2.0	Human
Visual Madlibs	10,738	360,001	33.52	4.9	2.8	Human
COCO-QA	117,684	117,684	1.00	9.65	1.0	Automatic
FM-IQA	158,392	316,193	1.99	7.38 (Chinese)	3.82 (Chinese)	Human
VQA (COCO)	204,721	614,163	3.00	6.2	1.1	Human
VQA (Abstract)	50,000	150,000	3.00	6.2	1.1	Human

Table 1. VQA Datasets

state-of-the-art performance in various tasks, such as object classification, detection and segmentation. For natural language, the Recurrent Neural Network (RNN) and the Long Short-Term Memory network (LSTM) are also widely used in machine translation and speech recognition.

There has been a tremendous amount of work being done in the field of Visual Question Answering, in a short span of time as this area emerged recently in late 2014. The advancements that are being made include creating new datasets for robust and unbiased results plus working on different models with both deep learning and non-deep learning approaches.

2.1. Datasets

Multiple datasets have come up for Visual QA task. Old datasets have annotations for each pixel to newer datasets having synthetic images and ground truths with absolute and plausible answers. Some of the datasets cited in top papers are mentioned in Table 1.

Five major datasets for VQA with natural images have been published in the span of four years (2014-present): DAQUAR [8], COCO-QA [11], FM-IQA [12], The VQA Dataset [1], and Visual7W [15]. Here, we refer to the portion of The VQA Dataset containing natural images as COCO-VQA. Detailed dataset reviews can be found in [5].

An ideal VQA dataset needs to be large enough to capture the variability within questions, images, and problems that occur in real world scenarios. It should also have a fair and bias-free evaluation scheme that is difficult to trick and obtaining a great performance on it is indicative that an algorithm can answer a large variety of question types about images that have definitive answers. If a dataset contains easy biases in the collection of the questions or answers, it may be possible for an algorithm to perform well on the dataset without really solving the VQA problem.

We are using the VQA 2.0 [3] dataset introduced in 2017. In VQA 2.0, the same question is asked for two different images and annotators are instructed to give opposite answers, which helped reduce language bias. Hence, this dataset is more robust to biases and we are using the same for our experimentation. The author produced results on VQA 1.0, which released in 2015.

What color are her eyes?
What is the mustache made of?
How many slices of pizza are there?
Is this a vegetarian pizza?
Is this person expecting company?
What is just under the tree?
Does t appear to be rainy?
Does this person have 20/20 vision?

Figure 2. Some images and questions from VQA dataset

2.2. Models

The VQA task came after deep learning approaches were already performing producing state-of-the-art performance on various vision and NLP tasks. As a result most of the work on VQA currently involves deep learning approaches, as opposed to more classical approaches like graphical models. There are a couple of models which use a non-neural approach as well.

Some of the models with non-deep learning approaches include the Answer Type Prediction (ATP) [4] which proposes a Bayesian Model for VQA in which first the answer types are predicted for a question and then these are used to generate the answer. Other one is the Multi-World QA [8] which models the probability of an answer given a question and an image as:

$$P(A = a|Q, W) = \Sigma_T P(A = a|T, W)P(T|Q) \qquad (1)$$

Here T is a latent variable corresponding to semantic tree obtained from a semantic parser run on the question. W is the world, which is a representation of the image.

Some of the deep learning models for VQA task include iBOWIMG [14] which uses GoogleNet CNN and FullCNN model [7], which use three different CNNs: an image CNN to encode the image, a question CNN to encode the question, and a join CNN to combine the image and question encoding together and produce a joint representation.

Some of the attention based deep learning models are Where to Look [12] model, where they use VGGNet for

	DAQUAR (Reduced)			DAQUAR (All)			COCO-QA		
	Accuracy	**WUPS**	**WUPS**	**Accuracy**	**WUPS**	**WUPS**	**Accuracy**	**WUPS**	**WUPS**
	(%)	**at 0.9 (%)**	**at 0 (%)**	**(%)**	**at 0.9 (%)**	**at 0 (%)**	**(%)**	**at 0.9 (%)**	**at 0 (%)**
SWQA	9.69	14.73	48.57	7.86	11.86	38.79	-	-	-
MWQA	12.73	18.10	51.47	-	-	-	-	-	-
Vis+LSTM	34.41	46.05	82.23	-	-	-	53.31	63.91	88.25
AYN	34.68	40.76	79.54	21.67	27.99	65.11	-	-	-
2Vis+BLSTM	35.78	46.83	82.15	-	-	-	55.09	65.34	88.64
Full-CNN	42.76	47.58	82.60	23.40	29.59	62.95	54.95	65.36	88.58
DPPnet	44.48	49.56	83.95	28.98	34.80	67.81	61.19	70.84	90.61
ATP	45.17	49.74	**85.13**	28.96	34.74	67.33	63.18	73.14	91.32
SAN	**45.50**	**50.20**	83.60	**29.30**	**35.10**	**68.60**	61.60	71.60	90.90
CoAtt	-	-	-	-	-	-	65.40	75.10	92.00
AMA	-	-	-	-	-	-	**69.73**	**77.14**	**92.50**

Table 2. Results of various models on DAQUAR (reduced), DAQUAR (full), COCO-QA

encoding the image and concatenate the outputs of the last two layers to obtain image encoding. An attention vector is computed over the set of image features to decide which region in the image has to be given importance to. Another model is Hierarchical Co-attention (CoAtt) [6] where it uses two types of co-attention: 1) Parallel Co-attention and 2) Alternating co-attention.

A comparison of different models can be found in Table 2 where the performance is tested over DAQUAR and COCO-QA datasets.

The model we are using is based on non-attention deep learning model, known as Ask Your Neurons model [9], where we use VGGNet for CNN part and LSTM for word embeddings. Then the obtained vectors are multiplied together and class scores are obtained for multiple answers.

3. Ask Your Neurons

According to the paper [9], the problem of answering a question with respect to an image, can be visualized as parametric probability measure problem, which can be mathematically stated as:

$$\hat{\boldsymbol{a}} = \arg\max_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}; \boldsymbol{\theta}) \qquad (2)$$

In the above expression, $\boldsymbol{a}$ represents the answer, $\boldsymbol{x}$ image, $\boldsymbol{q}$ question and $\boldsymbol{\theta}$ represents a vector of all parameters to learn and A is a set of all answers. The question $\boldsymbol{q}$ is a sequence of words, as $\boldsymbol{q} = [q_1, \ldots, q_n]$, where each q_t is the t-th word question with $q_n = ?$ encoding the question mark - the end of the question.

As we are working on single worded responses to achieve better classification results and robustness, we would further discuss the Ask Your Neurons (AYN) model using Equation 2 for the parametric probability measure.

With respect to the AYN model, we need to discuss two important parts of the architecture: Long-Short Term Memory (LSTM) and Convolution Neural Network (CNN) whose outputs are taken to form vectors and vector operations give us the answer scores (as shown in Figure 3).

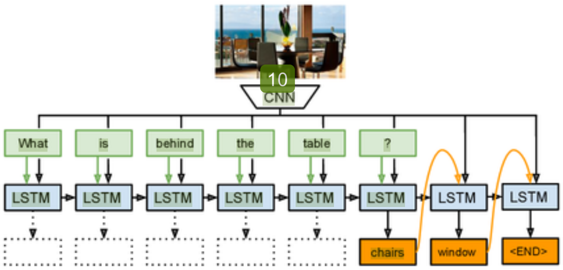

Figure 3. Ask Your Neurons

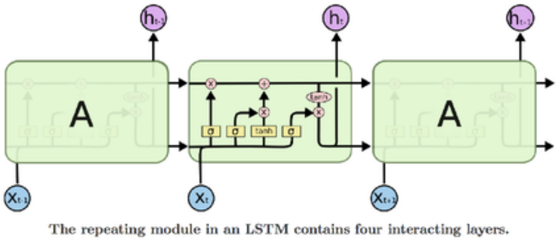
h_{t-1}
h_t
h_{t+1}
A
A
X_{t-1}
X_t
X_{t+1}
σ
σ
tanh
σ
tanh
The repeating module in an LSTM contains four interacting layers.

Figure 4. The repeating module in a LSTM containing four interacting layers

3.1. Long-Short Term Memory (LSTM)

The LSTM module is used to obtain word embeddings (text encoding) using the question given as a sequence of words and get a fixed n word vector which is supplied further to the model for vector processing.

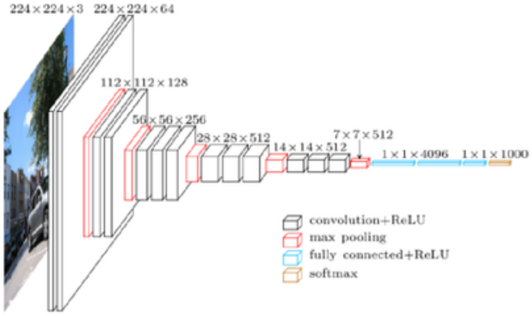

Figure 5. VGG 16 CNN Architecture

3.2. Convolution Neural Network (CNN)

Convolutional Neural Networks (CNNs) have become the state-of-the-art framework that provide features from images. Typically for object classification, the first step of using the visual models is to first pre-train them on the ImageNet dataset, a large scale object recognition dataset, and then use them as an input for the rest of the architecture. CNN architecture like AlexNet, GoogleNet, VGGNet (Figure 5) etc. have already shown great performance over object classification problems. Thus a CNN is used to obtain image features and form a fixed n length feature vector.

4. Methodology

The Ask Your Neurons (AYN) model is fed with image to the CNN architecture pre-trained on the Imagenet dataset and questions are fed to the LSTM network as a sequence of words.

Both question and answer words are represented with one-hot vector encoding (a binary vector with exactly one non-zero entry at the position indicating the index of the word in the vocabulary) and are embedded in a low dimensional space.

In the training phase, the question words sequence q is augmented with its corresponding ground truth answer words sequence a, i.e. $\hat{\boldsymbol{q}} := [\boldsymbol{q}, \boldsymbol{a}]$. While testing, in the prediction phase, at time step t, we augment q with previously predicted answer words i.e. $\hat{\boldsymbol{q}} := [\boldsymbol{q}, \boldsymbol{a}_{\hat{1..t-1}}]$. The above expression means the question q and the previous answer words are encoded implicitly in the hidden states of the LSTM, while the idle hidden representation is learnt.

The generated features (Image and Text) are combined using *point-wise multiplication*, which is further connected to Fully Connected (FC) layers, which have 1000 *"Most Possible Answers"* as classes/outputs.

There are two types of testing methodology that can be used evaluating the trained model: Open-Ended and Multiple-Choice. In Multiple-Choice, 18 candidate answers are possible which are divided into four types: Correct, Plausible, Popular and Random.

For Open-Ended task, the generated answers are evaluated using the following accuracy metric:

$$accuracy = \min\left(\frac{\#\text{humans given the same answer}}{3}, 1\right) \tag{3}$$

i.e., an answer is deemed 100% accurate if at least 3 workers provide the same exact answer.

The evaluation metric used in our experiments is the Open-Ended method.

5. Experiments

The AYN model that we used was a combination of VGG19 + LSTM architecture with their product fed into the stack of three fully-connected layers with last layer having 1000 classes as the output. The VGG19 model with pre-trained Imagenet weights was used for training and weights were froze for training.

Experiments were performed over VQA 2.0 which is more robust to bias as compared to VQA 1.0 as used by the author to report the accuracies. Data augmentation was used for the dataset, as the training images were made zero-centered and were normalized between -1 to 1. Other training constraints include usage of RMSProp optimizer with a learning rate of 0.0003, $\rho = 0.9$, and a weight decay of 0.99997592083, as mentioned in the original paper [9]. A batch size of 256 was used for training.

The input questions are divided into tokens using Python's NLTK (Natural Language Toolkit) tokenizer, and each token is generated into 300-dimensional word embedding vector using Stanford's pre-trained GloVe model [10]. Average question length is made into 26 sequences, either zero padded or truncated.

The idea behind using a pre-trained VGG19 model on Imagenet is to reduce the time taken for training and training overhead by extracting training images features before training itself. The same applies for the question. Each word is preprocessed by tokenizing and converting it to word embedding.

The model was trained for 10 epochs on the whole dataset which had around 82,783 images and 443,757 questions for training and it was completed within the stipulated 16 hours slot provided for GPU access. The system configuration used has 4 NVIDIA K80 GPU's, each with 12GB of VRAM.

6. Observations and Analysis

The results that were obtained after the reported training time was around 42.81% using open-ended task evaluation scheme. This has been compared to the reported results in

input (224 × 224 RGB image)
conv3-64 x 2
maxpool
conv3-128 x 2
maxpool
conv3-256 x 4
maxpool
conv3-512 x 4
maxpool
conv3-512 x 4
maxpool
FC-4096
FC-4096
FC-1000
soft-max

Table 3. VGG 19 configuration

the Ask Your Neurons paper [9] which have used the same architecture as used in the experiment. Their best accuracy was obtained when used with ResNet + LSTM configuration. The current state of the art is achieved in Teney *et al*. [13] where there model is an attention based deep learning model. Some of the reported accuracies are listed below.

Model	Accuracy
VGG19 + LSTM (ours)	**42.81%**
VGG19 + LSTM [9]	54.29%
ResNet + LSTM [9]	55.52%
Teney *et al*.	**69.0%**

Table 4. Comparison of accuracies over open-ended task

We think that the accuracies achieved over 10 epochs is quite tremendous for the given task, as papers have mentioned about the usage of large number of epochs for their reported accuracies.

7. Visual Q&A Demo

To demonstrate the real-life applications of the VQA task, we demonstrate it using an interactive web application. The web application, made in Python based microframework named Flask, takes an input image and a question in text format and sends it to the backend server, where the trained model and preprocessing scripts are used to convert it into feasible form for prediction by the model. The query returns the top five answers based on the confidence scores, which is displayed back in the web application.

8. Future Work

As we observed from the results as seen in the paper [9] and from our own reproduced work, the AYN model seems to be very novel idea when pitched with other Deep Learning models like Full-CNN etc. Plus, the ever-evolving field of deep learning with numerous new architectures being published every year gives us a scope of advancement of this model by using newer architectures for both image and text features.

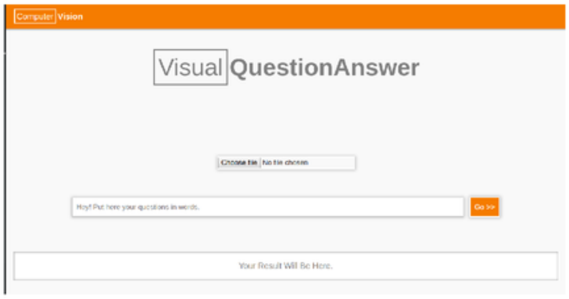
Computer Vision
Visual QuestionAnswer
Choose file No file chosen
Hey! Put here your questions in words.
Go >>
Your Result Will Be Here.

Figure 6. Visual Q&A web application

In our short span of time and limited availability of GPU access, we were able to train and test our model on LSTM + VGG configuration, which for today's standards is quite humble as the emergence of ResNet and SENet have shown residual and deeper networks perform better on image classification task and hence it would have been better for us to perform our experiments over these architecture to obtain better results as well as increase the number of fully connected layers to improve accuracy. Along with this, trying out new activations and loss functions would have given a greater scope for our project.

Along with testing with the architecture, we could analyze our network with different datasets like DAQUAR, COCO-QA which would have given us a report about the overall performance of the model.

Apart from deep learning models, we could experiment with attention based models, which might give us good results as discussed in the Section 2. Also, going with traditional non-deep learning models, it would have been an option, as it is observed that Answer Type Prediction (ATP) [4] model performs better than the deep learning models, which shows that simply using CNNS/RNN'S is not enough: identifying parts of the image that are relevant to the problem is also important.

9. Conclusion

The VQA task represents an important research topic both in the field of computer vision and natural language processing. It is neither restricted to the problems of object detection and classification or the text analysis and labeling. It requires to do much more than task-specific tasks, and an algorithm that can extract maximum information from images and text and provide the best results in representation of image in the form of text.

The Ask Your Neurons model shows a great way of combining these two not related features and get some factual information out of them. There still lies some scope of improvement in the algorithm in terms of optimizing architectures and obtaining a dataset which is robust to bias and helps to evaluate the important parts of a VQA algorithm, so that if a model performs well on that dataset then it indicates it is doing well on VQA in general. Also, the current evaluation metrics need to be improved because simply calculating accuracy is too naive for this problem.

In the end, we would like to conclude with that there exists a lot of scope as we can see from numerous publications every year and will grow by leaps and bounds in the next few years.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. 2015. International Conference on Computer Vision (ICCV).

[2] H. Gao, J. Mao, J. Zhou, Z. Huanga, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. 2015. NIPS.

[3] 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. 2017. Conference on Computer Vision and Pattern Recognition (CVPR).

[4] K. Kafle and C. Kanan. Answertype prediction for visual question answering. 2016. Computer Vision and Image Understanding.

[5] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. 2017. Computer Vision and Image Understanding.

[6] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image coattention for visual question answering. 2016. NIPS.

[7] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. 2015. arXiv preprint arXiv:1506.00333.

[8] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. 2014. NIPS Workshop on Learning Semantics.

[9] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A deep learning approach to visual question answering. 2016. arXiv preprint arXiv:1605.02697.

[10] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. 2014. Empirical Methods in Natural Language Processing (EMNLP).

[11] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. 2015. NIPS.

[12] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. 2016. CVPR.

[13] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. 2017. CoRR.

[14] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. 2015. arXiv preprint arXiv:1512.02167.

[15] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. 2016. CVPR.

CV_Final_G16

ORIGINALITY REPORT

22%	17%	18%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	6%
2	docplayer.net Internet Source	5%
3	papers.nips.cc Internet Source	3%
4	Kushal Kafle, Christopher Kanan. "An Analysis of Visual Question Answering Algorithms", 2017 IEEE International Conference on Computer Vision (ICCV), 2017 Publication	2%
5	link.springer.com Internet Source	1%
6	Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, Dhruv Batra. "VQA: Visual Question Answering", International Journal of Computer Vision, 2016 Publication	1%

7	Kushal Kafle, Christopher Kanan. "Visual question answering: Datasets, algorithms, and future challenges", Computer Vision and Image Understanding, 2017 Publication	1 %
8	Submitted to Rochester Institute of Technology Student Paper	<1 %
9	Damien Teney, Qi Wu, Anton van den Hengel. "Visual Question Answering: A Tutorial", IEEE Signal Processing Magazine, 2017 Publication	<1 %
10	www.cv-foundation.org Internet Source	<1 %
11	www.aclweb.org Internet Source	<1 %
12	www.cse.iitk.ac.in Internet Source	<1 %
13	Kushal Kafle, Christopher Kanan. "Answer-Type Prediction for Visual Question Answering", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 Publication	<1 %
14	personal.ie.cuhk.edu.hk Internet Source	<1 %

15	dbsj.org Internet Source	<1%
16	Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, Cornelia Fermüller. "Image Understanding using vision and reasoning through Scene Description Graph", Computer Vision and Image Understanding, 2017 Publication	<1%
17	Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao. "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering", 2017 IEEE International Conference on Computer Vision (ICCV), 2017 Publication	<1%
18	export.arxiv.org Internet Source	<1%
19	Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. "Visual question answering: A survey of methods and datasets", Computer Vision and Image Understanding, 2017 Publication	<1%
20	"Computer Vision – ECCV 2016", Springer Nature, 2016 Publication	<1%
21	Peng Wang, Qi Wu, Chunhua Shen, Anton van	

den Hengel. "The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

Publication

<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		