Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction

by Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han.[1]

Presented : Badri Patro[1]

[1]Computer Vision Reading Group
Indian Institute of Technology Kanpur.

June 24, 2016

Plagiarism

- All the figures, equations and text are taken from above paper and reference papers.
- This presentation is meant for educational purpose only.

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

Introduction

Problem Statement:
Given an image,machine will automatically answer questions posed by humans in natural language query.

Introduction

Problem Statement:
Given an image,machine will automatically answer questions posed by humans in natural language query.

- Training on a set of triplets (image, question, answer).
- Answers can be single word or multiple word, depending on dataset.

Introduction

Type of Question

- Fine-grained recognition (e.g., "What kind of cheese is on the pizza?").

Introduction

Type of Question

- Fine-grained recognition (e.g.,"What kind of cheese is on the pizza?").
- Object detection(e.g., "How many bikes are there?").

Introduction

Type of Question

- Fine-grained recognition (e.g.,“What kind of cheese is on the pizza?”).
- Object detection(e.g., “How many bikes are there?”).
- Activity recognition (e.g., “Is this man crying?”).

Introduction

Type of Question

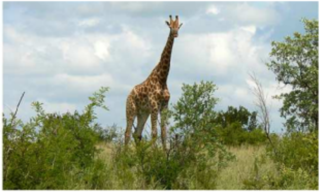
- Fine-grained recognition (e.g.,“What kind of cheese is on the pizza?”).
- Object detection(e.g., “How many bikes are there?”).
- Activity recognition (e.g., “Is this man crying?”).
- Knowledge base reasoning(e.g., “Why did Katappa killed Bahubali?”).

Introduction

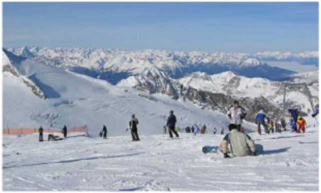
Type of Question

- Fine-grained recognition (e.g.,“What kind of cheese is on the pizza?”).
- Object detection(e.g., “How many bikes are there?”).
- Activity recognition (e.g., “Is this man crying?”).
- Knowledge base reasoning(e.g., “Why did Katappa killed Bahubali?”).
- Commonsense reasoning (e.g., “Does this person have 20/20 vision?”, “Is this person expecting company?”)..

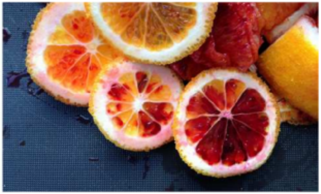
Introduction



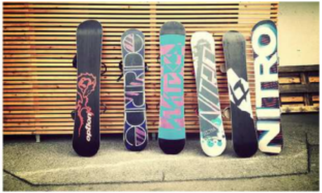
Q: What type of animal is this?
Q: Is this animal alone?



Q: Is it snowing?
Q: Is this picture taken during the day?

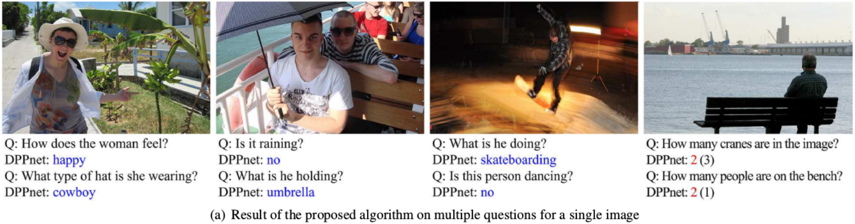


Q: What kind of oranges are these?
Q: Is the fruit sliced?



Q: What is leaning on the wall?
Q: How many boards are there?

Result of multiple questions for a single image



Q: How does the woman feel?
DPPnet: happy
Q: What type of hat is she wearing?
DPPnet: cowboy

Q: Is it raining?
DPPnet: no
Q: What is he holding?
DPPnet: umbrella

Q: What is he doing?
DPPnet: skateboarding
Q: Is this person dancing?
DPPnet: no

Q: How many cranes are in the image?
DPPnet: 2 (3)
Q: How many people are on the bench?
DPPnet: 2 (1)

(a) Result of the proposed algorithm on multiple questions for a single image

Figure : Result of the proposed algorithm on multiple questions for a single image

Introduction

- The critical challenge of this problem is that different questions require different types and levels of understanding of an image to find correct answers.
- For example, to answer the question like “how is the weather?” we need to perform classification on multiple choices related to weather, while we should decide between yes and no for the question like “is this picture taken during the day?”.
- For this reason, not only the performance on a single recognition task but also the capability to select a proper task is important to solve ImageQA problem.

Introduction

- Simple deep learning based approaches that perform classification on a combination of features extracted from image and question currently demonstrate the state-of-the-art accuracy.
- The existing approaches of VQA extract image features using a convolutional neu- ral network (CNN), and use CNN or bag-of-words to obtain feature descriptors from question.
- These methods can be interpreted as the answer is given by the co-occurrence of a particular combination of features extracted from an image and a question.

Introduction

- Contrary to the existing approaches, Author defines a different recognition task depending on a question.
- In order to realize this idea, Author proposed a deep CNN with a dynamic parameter layer whose weights are determined adaptively based on questions.
- This paper claims that a single deep CNN architecture can take care of various tasks by allowing adaptive weight assignment in the dynamic parameter layer

Introduction

Main contributions in this work are summarized below:

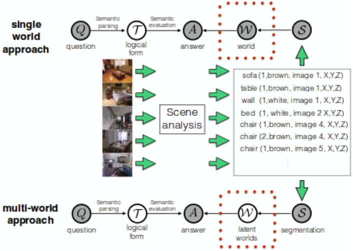
- successfully adopted a deep CNN with a dynamic parameter layer for ImageQA, which is a fully-connected layer whose parameters are determined dynamically based on a given question.
- To predict a large number of weights in the dynamic parameter layer, applyed hashing trick , which reduces the number of parameters significantly with little impact on network capacity.
- We fine-tune GRU pre-trained on a large-scale text corpus [14] to improve generalization performance of our network.
- This is the first work to report the results on all currently available benchmark datasets such as DAQUAR,COCO-QA and VQA.
- Our algorithm achieves the state-of-the-art performance on all the three datasets..

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

A multi-world approach to QA by[Malinowski et al.,2014] [6]

- It employs semantic image segmentation and symbolic question reasoning to solve ImageQA problem.
- However, this method depends on a pre-defined set of predicates,which makes it difficult to represent complex models required to understand input images.
- Deep learning based approaches demonstrate competitive performances in ImageQA.



sofa (1,brown, image 1, X,Y,Z)
table (1,brown, image 1,X,Y,Z)
wall (1,white, image 1, X,Y,Z)
bed (1, white, image 2 X,Y,Z)
chair (1,brown, image 4, X,Y,Z)
chair (2,brown, image 4, X,Y,Z)
chair (1,brown, image 5, X,Y,Z)
⋮

Neural Image-QA[Malinowski et al.,2015][5]

- Most of approaches based on deep learning, use CNNs to extract features from image while they use different strategies to handle question sentences.
- Some algorithms employ embedding of joint features based on image and ques- tion.
- Neural Image-QA model based approch, the image representation from CNN is fed to each hidden layers of LSTM.
- In this model, the answers is short, such as one single word ,i.e the object category, color, number, and so on.

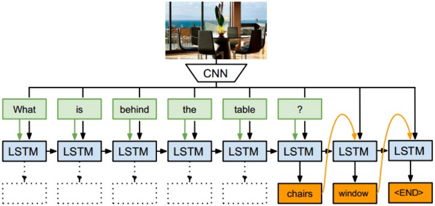


Figure : Neural Image-QA(Malinowski et al.,2015)[5]

VSE(VIS+LSTM) Model[Ren et al.,May,2015][4]

- In VSE(visual semantic embedding) Model, image QA task is formulated as a classification problem .
- This model also contains a single LSTM and a CNN.
- LSTM is employed to jointly model the image and question by treating the image as an independent word, and appending it to the question at the beginning or end.

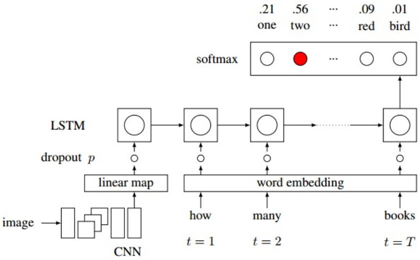
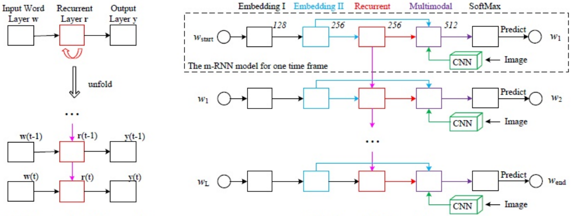


Figure : VSE model- VIS LSTM(Ren et al.,May,2015)[4]

mQA Model

The structure of mQA model is inspired by the m-RNN model [7] for the image captioning and image-sentence retrieval tasks. Block diagram of M-RNN as show in figure.

mQA adopts a deep CNN for computer vision and a RNN for language. mQA model is extended to handle the input of question and image pairs, and generate answers.



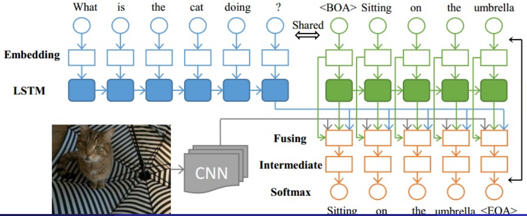
(a). The simple RNN model

(b). The m-RNN model

Figure : m_RNN model(Mao et al.,2015)[7]

mQA Model [Gao et al.,Nov, 2015][3]

- It contains four components: a Long Short-Term Memory (LSTM) to extract the question representation, a Convolutional Neural Network (CNN) to extract the visual representation, an LSTM for storing the linguistic context in an answer, and a fusing component to combine the information from the first three components and generate the answer.
- Different from [4][5], the image representation does not feed into the LSTM . Here, we use two separate LSTMs for questions and answers respectively in consideration of the different properties (e.g. grammar) of questions and answers.



ConvQA Model [Ma et al.,Nov,2015] [2]

In this approach uses 3 CNN's - one to extract sentence representation, one for image representation, and the third is a multimodal layer to fuse the two.

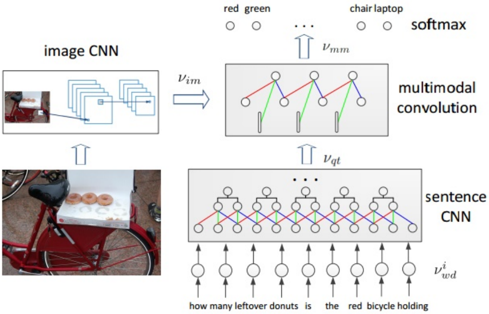


Figure : ConvQA(Ma et al.,Nov,2015)[2]

Table of Contents

1 Introduction

2 Related Work

3 Algorithm Overview
- Problem Formulation

4 Network Architecture
- Classification Network
- Parameter Prediction Network
- Parameter Hashing

5 Training Algorithm
- Training by Error Back-Propagation

6 Experiment
- Datasets
- Results

7 Conclusion

8 References

Introduction

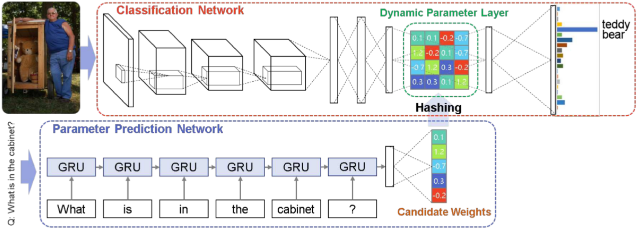


Figure : Overall architecture of the proposed Dynamic Parameter Prediction network (DPPnet), which is composed of the classification network and the parameter prediction network.

- The weights in the dynamic parameter layer are mapped by a hashing trick from the candidate weights obtained from the parameter prediction network.

Problem Formulation

ImageQA systems predict the best answer â given an image I and a question q.

Conventional approaches

It is a joint feature vector based on two inputs I and q and solve a classification problem for ImageQA giveb by following eq

$$\hat{a} = \underset{a \in \Omega}{\operatorname{argmax}}\ p(a|I, q; \boldsymbol{\theta})$$

Where ω is a set of all possible answers and θ is a vector for the parameters in the network

Problem Formulation

Proposed approaches

Here, Authur uses the question to predict weights in the classifier and solve the problem.

$$\hat{a} = \underset{a \in \Omega}{\operatorname{argmax}}\, p(a|I; \boldsymbol{\theta}_s, \boldsymbol{\theta}_d(q))$$

where θ_s and$\theta_d(q)$ denote static and dynamic parameters, respectively. Note that the values of $\theta_d(q)$ are determined by the question q

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

Classification Network

- The classification network is constructed based on VGG 16-layer net which is pre-trained on ImageNet.
- The last layer of Vgg net is removed and attached three fully-connected layers.
- The second last fully-connected layer is the dynamic parameter layer whose weights are determined by the parameter prediction network,
- The last fully-connected layer is the classification layer whose output dimensionality is equal to the number of possible answers.
- The probability for each answer is computed by applying a softmax function to the output vector of the final layer.

Classification Network

- Dynamic parameter layer in the second last fully-connected layer instead of the classification layer because it involves the smallest number of parameters.
- As the number of parameters in the classification layer increases in proportion to the number of possible answers, which may not be a good option to general ImageQA problems.

Dynamic Parameter Layer

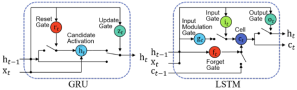
- Classification network has a dynamic parameter layer. That is, for an input vector of the dynamic parameter layer $f^i = [f^i_1, ..., f^i_N]$, its output vector denoted by $f^o = [f^o_1, ..., f^o_M]$ is given

$$\mathbf{f}^o = \mathbf{W}_d(q)\mathbf{f}^i + \mathbf{b}$$

- Where b denotes a bias and $W_d(q) \epsilon R^{M*N}$ denotes the weight matrix constructed dynamically using the parameter prediction network given the input question q.

Parameter Prediction Network(PPN)

- The parameter prediction network is composed of GRU cells [Chung. et al.,] followed by a fully-connected layer.
- fc layer of PPN produces the candidate weights to be used for the construction of weight matrix in the dynamic parameter layer within the classification network.
- fc layer of PPN produces the candidate weights to be used for the construction of weight matrix in the dynamic parameter layer within the classification network.



Parameter Prediction Network(PPN)

- Let w1, ..., wT be the words in a question q, where T is the number of words in the question.
- In each time step t, given the embedded vector $x_t$for a word w_t,GRU encoder updates its hidden h_t at time t is given by

$$
\begin{aligned}
\mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}) \\
\mathbf{z}_t &= \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}) \\
\bar{\mathbf{h}}_t &= \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h (\mathbf{r}_t \odot \mathbf{h}_{t-1})) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \bar{\mathbf{h}}_t
\end{aligned}
$$

- where r_t and z_t respectively denote the reset and update gates at time t,$\bar{h}_t$ is candidate activation at time t.
- Note that the coefficient matrices related to GRU such as $W_r, W_z, W_h, U_r, U_z, andU_h$ are learned by our training algorithm.

Parameter Hashing

- The final hash function is given by

$$w_{mn}^{d} = p_{\psi(m,n)} \cdot \xi(m,n)$$

- This function is useful to remove the bias of hashed inner product [3(original ref)].
- In our implementation of the hash function, we adopt an open-source implementation of xxHash.

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

Training by Error Back-Propagation

- The proposed network is trained end-to-end to minimize the error between the ground-truths and the estimated an- swers.
- The error is back-propagated to both the classification network and the parameter prediction network and jointly trained by a 1st order optimization method.

Training by Error Back-Propagation

- The proposed network is trained end-to-end to minimize the error between the ground-truths and the estimated an- swers.
- The error is back-propagated to both the classification network and the parameter prediction network and jointly trained by a 1st order optimization method.
- Let L denote the loss function. The partial derivatives of L with respect to the k th element in the input and output of the dynamic parameter layer are given respectively by

$$\delta_k^i \equiv \frac{\partial \mathcal{L}}{\partial f_k^i} \quad \text{and} \quad \delta_k^o \equiv \frac{\partial \mathcal{L}}{\partial f_k^o}.$$

Training by Error Back-Propagation

- The proposed network is trained end-to-end to minimize the error between the ground-truths and the estimated an- swers.
- The error is back-propagated to both the classification network and the parameter prediction network and jointly trained by a 1st order optimization method.
- Let L denote the loss function. The partial derivatives of L with respect to the k th element in the input and output of the dynamic parameter layer are given respectively by

$$\delta_k^i \equiv \frac{\partial \mathcal{L}}{\partial f_k^i} \quad \text{and} \quad \delta_k^o \equiv \frac{\partial \mathcal{L}}{\partial f_k^o}.$$

- The two derivatives have the following relation:

$$\delta_n^i = \sum_{m=1}^{M} w_{mn}^d \delta_m^o$$

Training by Error Back-Propagation

- Likewise, the derivative with respect to the assigned weights in the dynamic parameter layer is given by

$$\frac{\partial \mathcal{L}}{\partial w^d_{mn}} = f^i_n \delta^o_m.$$

- A single output value of the PPN is shared by multiple connections in the DPL.
- To compute the derivative with respect to an element in the output of the parameter prediction network(PPN) as follows

$$\frac{\partial \mathcal{L}}{\partial p_k} = \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial \mathcal{L}}{\partial w^d_{mn}} \frac{\partial w^d_{mn}}{\partial p_k}$$
$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\partial \mathcal{L}}{\partial w^d_{mn}} \xi(m,n) \mathbb{I}[\psi(m,n) = k],$$

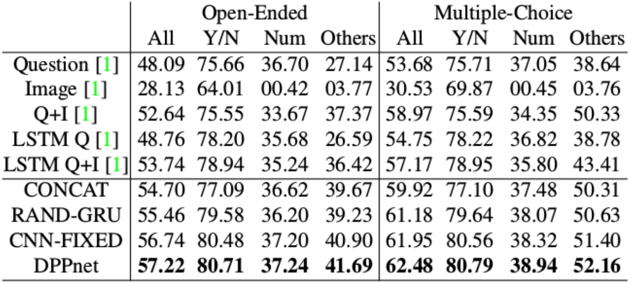
Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

- VQA
- MS-COCO
- DAQUAR-all and DAQUAR reduced.

Evaluation results on VQA

Table 1. Evaluation results on VQA test-dev in terms of Acc_{VQA}



	Open-Ended				Multiple-Choice			
	All	Y/N	Num	Others	All	Y/N	Num	Others
Question [1]	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image [1]	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
Q+I [1]	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q [1]	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q+I [1]	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
CONCAT	54.70	77.09	36.62	39.67	59.92	77.10	37.48	50.31
RAND-GRU	55.46	79.58	36.20	39.23	61.18	79.64	38.07	50.63
CNN-FIXED	56.74	80.48	37.20	40.90	61.95	80.56	38.32	51.40
DPPnet	**57.22**	**80.71**	**37.24**	**41.69**	**62.48**	**80.79**	**38.94**	**52.16**

Evaluation results on COCO-QA

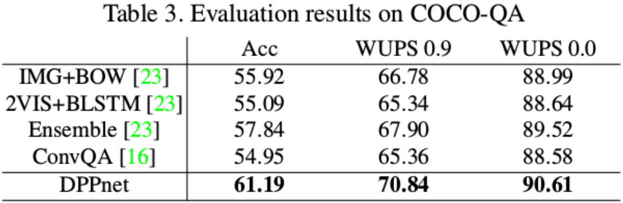


Table 3. Evaluation results on COCO-QA

	Acc	WUPS 0.9	WUPS 0.0
IMG+BOW [23]	55.92	66.78	88.99
2VIS+BLSTM [23]	55.09	65.34	88.64
Ensemble [23]	57.84	67.90	89.52
ConvQA [16]	54.95	65.36	88.58
DPPnet	**61.19**	**70.84**	**90.61**

Evaluation results on DAQUAR

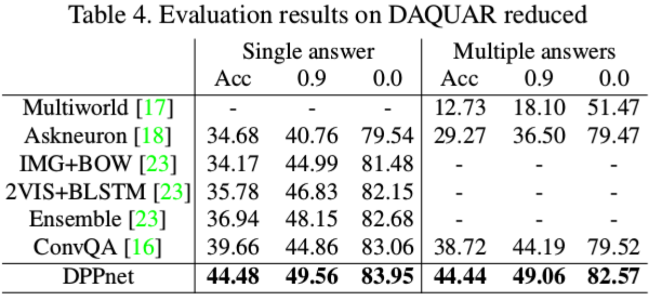


Table 4. Evaluation results on DAQUAR reduced

	Single answer			Multiple answers		
	Acc	0.9	0.0	Acc	0.9	0.0
Multiworld [17]	-	-	-	12.73	18.10	51.47
Askneuron [18]	34.68	40.76	79.54	29.27	36.50	79.47
IMG+BOW [23]	34.17	44.99	81.48	-	-	-
2VIS+BLSTM [23]	35.78	46.83	82.15	-	-	-
Ensemble [23]	36.94	48.15	82.68	-	-	-
ConvQA [16]	39.66	44.86	83.06	38.72	44.19	79.52
DPPnet	**44.48**	**49.56**	**83.95**	**44.44**	**49.06**	**82.57**

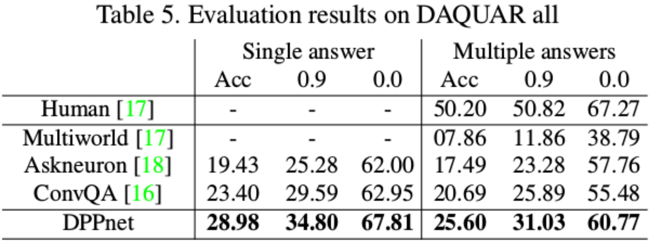
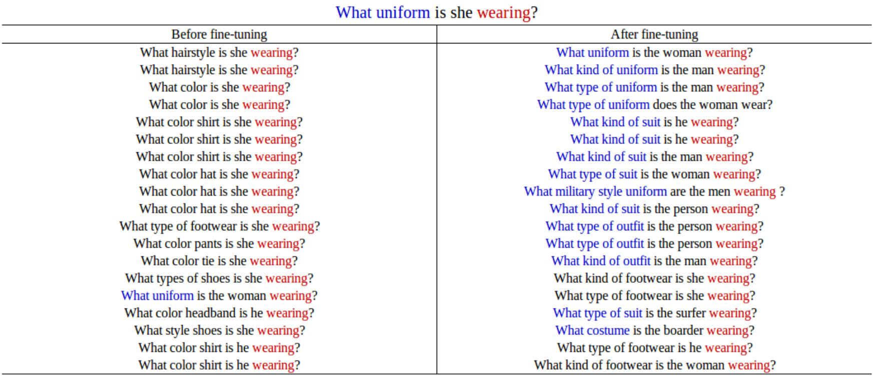


Table 5. Evaluation results on DAQUAR all

	Single answer			Multiple answers		
	Acc	0.9	0.0	Acc	0.9	0.0
Human [17]	-	-	-	50.20	50.82	67.27
Multiworld [17]	-	-	-	07.86	11.86	38.79
Askneuron [18]	19.43	25.28	62.00	17.49	23.28	57.76
ConvQA [16]	23.40	29.59	62.95	20.69	25.89	55.48
DPPnet	**28.98**	**34.80**	**67.81**	**25.60**	**31.03**	**60.77**

Sentences before and after fine-tuning GRU



What uniform is she wearing?

Before fine-tuning	After fine-tuning
What hairstyle is she wearing?	What uniform is the woman wearing?
What hairstyle is she wearing?	What kind of uniform is the man wearing?
What color is she wearing?	What type of uniform is the man wearing?
What color is she wearing?	What type of uniform does the woman wear?
What color shirt is she wearing?	What kind of suit is he wearing?
What color shirt is she wearing?	What kind of suit is he wearing?
What color shirt is she wearing?	What kind of suit is the man wearing?
What color hat is she wearing?	What type of suit is the woman wearing?
What color hat is she wearing?	What military style uniform are the men wearing ?
What color hat is she wearing?	What kind of suit is the person wearing?
What type of footwear is she wearing?	What type of outfit is the person wearing?
What color pants is she wearing?	What type of outfit is the person wearing?
What color tie is she wearing?	What kind of outfit is the man wearing?
What types of shoes is she wearing?	What kind of footwear is she wearing?
What uniform is the woman wearing?	What type of footwear is she wearing?
What color headband is he wearing?	What type of suit is the surfer wearing?
What style shoes is she wearing?	What costume is the boarder wearing?
What color shirt is he wearing?	What type of footwear is he wearing?
What color shirt is he wearing?	What kind of footwear is the woman wearing?

Figure : Retrieved sentences before and after fine-tuning GRU

Sentences before and after fine-tuning GRU

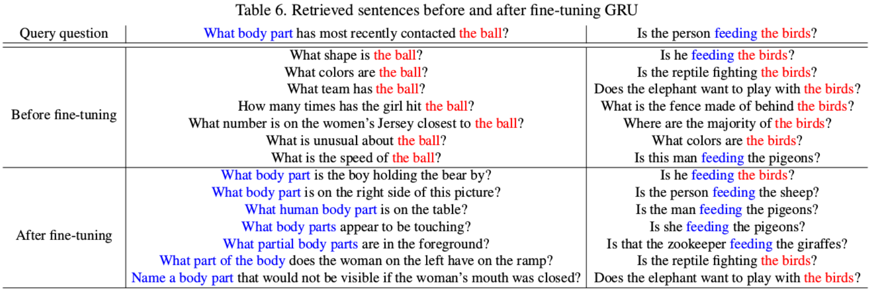
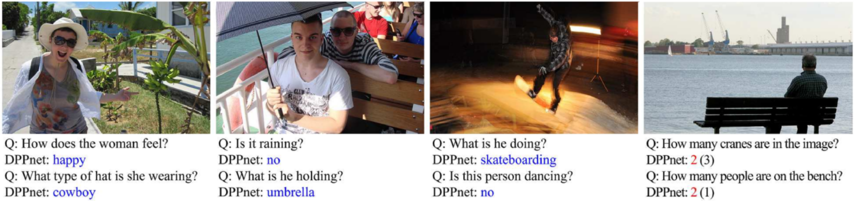


Table 6. Retrieved sentences before and after fine-tuning GRU

Query question	What body part has most recently contacted the ball?	Is the person feeding the birds?
Before fine-tuning	What shape is the ball?	Is he feeding the birds?
	What colors are the ball?	Is the reptile fighting the birds?
	What team has the ball?	Does the elephant want to play with the birds?
	How many times has the girl hit the ball?	What is the fence made of behind the birds?
	What number is on the women's Jersey closest to the ball?	Where are the majority of the birds?
	What is unusual about the ball?	What colors are the birds?
	What is the speed of the ball?	Is this man feeding the pigeons?
After fine-tuning	What body part is the boy holding the bear by?	Is he feeding the birds?
	What body part is on the right side of this picture?	Is the person feeding the sheep?
	What human body part is on the table?	Is the man feeding the pigeons?
	What body parts appear to be touching?	Is she feeding the pigeons?
	What partial body parts are in the foreground?	Is that the zookeeper feeding the giraffes?
	What part of the body does the woman on the left have on the ramp?	Is the reptile fighting the birds?
	Name a body part that would not be visible if the woman's mouth was closed?	Does the elephant want to play with the birds?

Figure : Retrieved sentences before and after fine-tuning GRU

Result of multiple questions for a single image



Q: How does the woman feel?
DPPnet: happy
Q: What type of hat is she wearing?
DPPnet: cowboy

Q: Is it raining?
DPPnet: no
Q: What is he holding?
DPPnet: umbrella

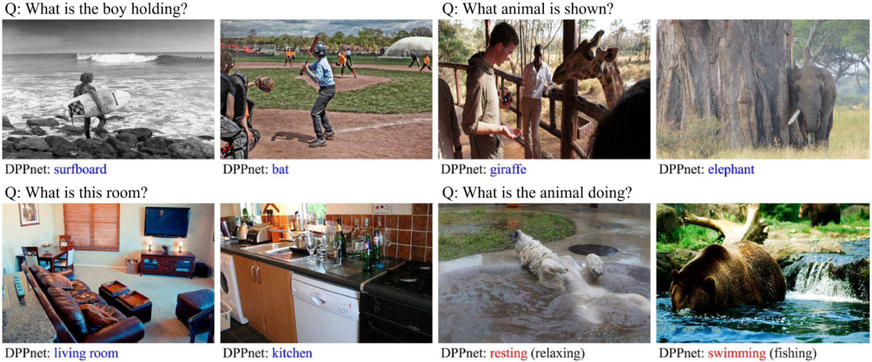
Q: What is he doing?
DPPnet: skateboarding
Q: Is this person dancing?
DPPnet: no

Q: How many cranes are in the image?
DPPnet: 2 (3)
Q: How many people are on the bench?
DPPnet: 2 (1)

(a) Result of the proposed algorithm on multiple questions for a single image

Figure : Result of the proposed algorithm on multiple questions for a single image

Results of single common question for multiple images



(b) Results of the proposed algorithm on a single common question for multiple images

Figure : Results of the proposed algorithm on a single common question for multiple images

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. **Conclusion**
8. References

Conclusion

- The effectiveness of the proposed architecture is supported by experimental results showing the state-of-the-art performances on three different dataset.
- Note that the proposed method achieved outstanding performance even without more complex recognition processes such as referencing objects.

Table of Contents

1. Introduction
2. Related Work
3. Algorithm Overview
   - Problem Formulation
4. Network Architecture
   - Classification Network
   - Parameter Prediction Network
   - Parameter Hashing
5. Training Algorithm
   - Training by Error Back-Propagation
6. Experiment
   - Datasets
   - Results
7. Conclusion
8. References

Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han. “Image question answering using convolutional neural network with dynamic parameter prediction.”,arXiv preprint arXiv:1511.05756 (2015).
””

L. Ma, Z. Lu, and H. Li., “Learning to Answer Questions From Image using Convolutional Neural Network”,CoRR abs/1506.00333, Nov, 2015.

H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang and W. Xu.,“Are you talking to a machine? dataset and methods for multilingual image question answering.”,arXiv 1505.05612v3, Nov, 2015.

M. Ren, R. Kiros, and R. S. Zemel, “Exploring models and data for image question answering”,arXiv 1505.02074, , 2015.

M. Malinowski, M. Rohrbach, and M. Fritz.,“Ask your neurons: A neural-based approach to answering questions about images.”,arXiv 1505.01121, Nov, 2015.

M. Malinowski and M. Fritz.,“Towards a visual turing challenge.”,In Learning Semantics (NIPS workshop), 2014.

J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille.,“Deep captioning with multimodal recurrent neural networks (m-rnn).”,In ICLR, 2015.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh.,“Vqa: Visual question answering.”, arXiv preprint arXiv:1505.00468, 2015.

- S. Hochreiter and J. Schmidhuber., "Long short-term memory.",Neural computation, 9(8):1735-1780, 1997.
- K. Simonyan and A. Zisserman., "Very deep convolutional networks for large-scale image recognition.", In ICLR, 2015.
- I. Sutskever, O. Vinyals, and Q. V. Le., "Sequence to sequence learning with neural networks.", In NIPS, pages 3104-3112, 2014.