Content Based Visual Question Answer

Badri Patro, Vinay Verma & Aatanu Samata

Group_30 :Project Proposal Presentation
CS676A:Computer Vision and Image Processing
IIT Kanpur

March 12, 2016

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?)..

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?)..

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?)..

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?)..

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?)..

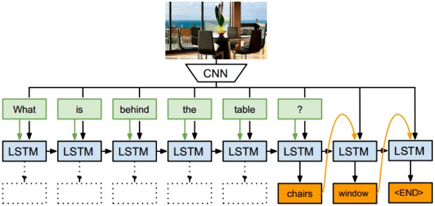
Problem Statement: Given an image, and content based question, find the correct answer and confidence of the answer.

- Training on a set of triplets (image, question, answer).
- Free-form and open-ended(*) questions.
- Answers can be single word or multiple word, depending on dataset.

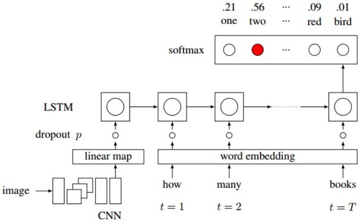
Problem Statement: Given an image, and content based question, find the correct answer and confidence of the answer.

- Training on a set of triplets (image, question, answer).
- Free-form and open-ended(*) questions.
- Answers can be single word or multiple word, depending on dataset.

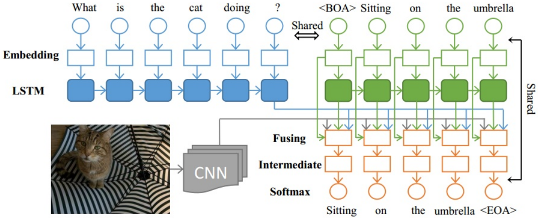
1 Neural Image-QA(Malinowski et al.,2015)[4]



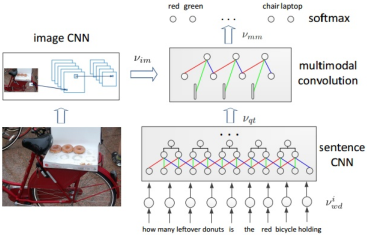
2 VSE model- VIS LSTM(Ren et al.,May,2015)[3]



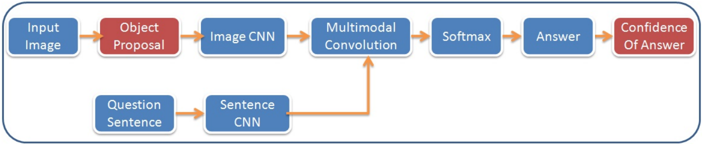
3 mQA(Gao et al.,Nov, 2015)[2]



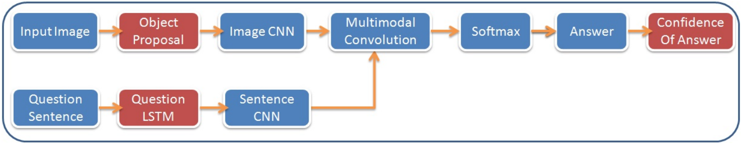
4 3CNN(Ma et al.,Nov,2015)[1]



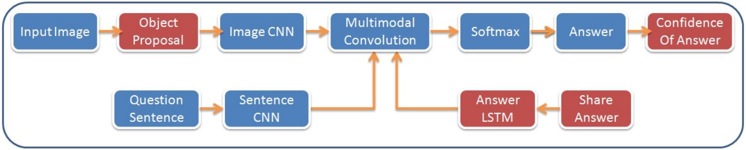
1. **Modified CNN,[Ma et al.]**



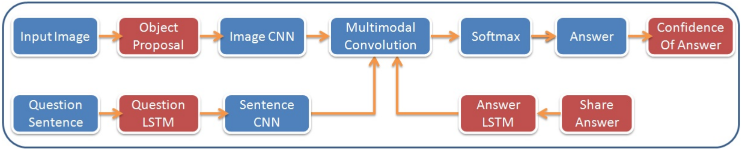
2. **Modified CNN with QuestionLSTM,[Ma et al.]**



3 Modified CNN with AnswerLSTM,[Geo et al.]



4 Final Approach,[Ma et al.] & [Geo et al.]



Component of Proposed Algorithm

- Object Proposal:Find Interest Object
- Image CNN:to extract the visual representation
- Question LSTM : extract the question representation
- Sentence CNN: question representation
- Multimodal Convolution: a fusing component to combine the information from the first three components and generate the answer.
- Answer LSTM:for storing the linguistic context in an answer

1. **Object Proposal[Cheng et al. [?]and Endres et al. [6]]**

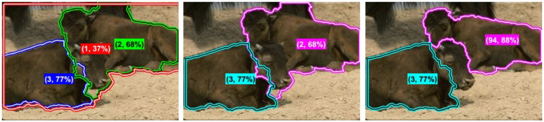


Figure: The left column shows the 3 highest ranked proposals, The center column shows the highest ranked proposal with 50% overlap for each object. The right column shows the same for a 75% threshold

2 Image CNN

$$v_{im} = \sigma(w_{im}(CNN_{im}(I)) + b_{im})$$

σ: Nonlinear activation function.
$w_{im}|dX4096$: Mapping matrix
CNN_{im}takes image as input and outputs a fixed length vector.
b_{im} constant

3 LSTM

- LSTM layer stores the context information in its memory cells and serves as the bridge among the words in a sequence (e.g. a question).
- It has three gate :
  - Input gate and Output gate : Regulate the read and write access to the LSTM memory cells.
  - Forget gate: Resets the memory cells when their contents are out of date.

2 Sentence CNN

- For sequential input σ, convolution unit for feature map of type f on the l^{th} layer is

$$v^{i}_{(l,f)} \stackrel{def}{=} \sigma(w_{(l,f)} \vec{v}^{i}_{(l1)} + b_{(l,f)})$$

-

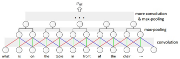
$$\vec{v}^{i}_{(l-1)} \stackrel{def}{=} v^{i}_{(l-1)} || v^{i+1}_{(l-1)} || v^{i+1}_{(l-1)}$$

-

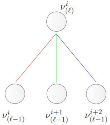
$$\vec{v}^{i}_{(0)} \stackrel{def}{=} v^{i}_{wd} || v^{i+1}_{wd} || v^{i+1}_{wd}$$

- Max-pooling after each convolution

$$v^{i}_{(l+1,f)} = \max(v^{2i}_{(l,f)}, v^{2i+1}_{(l,f)})$$



(a) High Level.



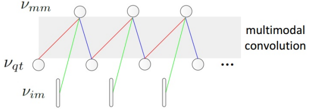
(b) Detailed Level.

2 Multimodal Convolution

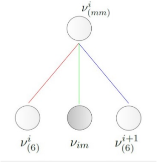
- input: $v_{qt} = [v^0_{(6)} \dots v^n_{(6)}]$
- Capturing the interaction between two multimodal inputs

$$\vec{v}^i_6 = v^i_6 || v_{im} || v^{i+1}_6 \tag{1}$$

$$v^i_{(mm,f)} = \sigma(w_{(mm,f)} \vec{v}^i_6 + b_{(mm,f)}) \tag{2}$$



(a) High Level.



(b) Detailed Level.

Figure: Multimodal Convolution[1].

2 VQA performances on DAQUAR-Reduced[1]

Table 2: Image QA performances on DAQUAR-Reduced.

	Accuracy	WUPS @0.9	WUPS @0.0
Multi-World Approach (Malinowski and Fritz 2014a)	12.73	18.10	51.47
Neural-Image-QA (Malinowski, Rohrbach, and Fritz 2015)			
-multiple words	29.27	36.50	79.47
-single word	34.68	40.76	79.54
Language Approach			
-multiple words	32.32	38.39	80.05
-single word	31.65	38.35	80.08
VSE (Ren, Kiros, and Zemel 2015)			
-single word			
GUESS	18.24	29.65	77.59
BOW	32.67	43.19	81.30
LSTM	32.73	43.50	81.62
IMG+BOW	34.17	44.99	81.48
VIS+LSTM	34.41	46.05	82.23
2-VIS+BLSTM	35.78	46.83	82.14
Proposed CNN			
-multiple words	**38.38**	**43.43**	**80.63**
-single word	**42.76**	**47.58**	**82.60**

(a) DAQUAR-Reduced.

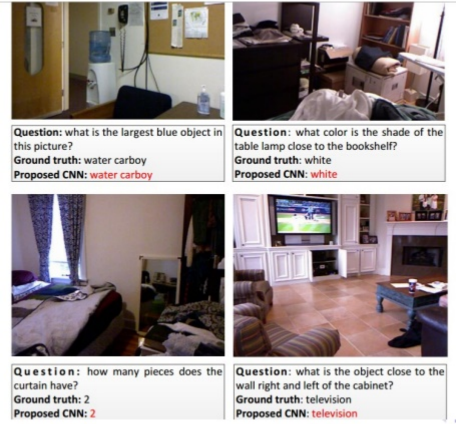
Table 3: Image QA performances on COCO-QA.

	Accuracy	WUPS @0.9	WUPS @0.0
VSE (Ren, Kiros, and Zemel 2015)			
GUESS	6.65	17.42	73.44
BOW	37.52	48.54	82.78
LSTM	36.76	47.58	82.34
IMG	43.02	58.64	85.85
IMG+BOW	55.92	66.78	88.99
VIS+LSTM	53.31	63.91	88.25
2-VIS+BLSTM	55.09	65.34	88.64
FULL	57.84	67.90	89.52
Proposed CNN without multimodal convolution layer	56.77	66.76	88.94
Proposed CNN without image representation	37.84	48.70	82.92
Proposed CNN	**58.40**	**68.50**	**89.67**

(b) COCO-QA.

Figure: VQA performances on DAQUAR-Reduced[1].

VQA Demo[1]


Question: what is the largest blue object in this picture?
Ground truth: water carboy
Proposed CNN: water carboy
Question: what color is the shade of the table lamp close to the bookshelf?
Ground truth: white
Proposed CNN: white
Question: how many pieces does the curtain have?
Ground truth: 2
Proposed CNN: 2
Question: what is the object close to the wall right and left of the cabinet?
Ground truth: television
Proposed CNN: television

Thank you

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

[2] 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.

[3] M. Ren, R. Kiros, and R. S. Zemel, ''Exploring models and data for image question answering'',arXiv 1505.02074, , 2015.

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

[5] A. Karpathy and L. Fei-Fei., ''Deep Visual-Semantic Alignments for Generating Image Descriptions.'',In CVPR, 2015.

[6] I. Endres, and D. Hoiem., ''Category-Independent Object Proposals With Diverse Ranking.'', http://vision.cs.uiuc.edu/proposals/,PAMI February, 2014.

[7] Cheng, Ming-Ming, et al.,''BING: Binarized normed gradients for objectness estimation at 300fps.'', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.