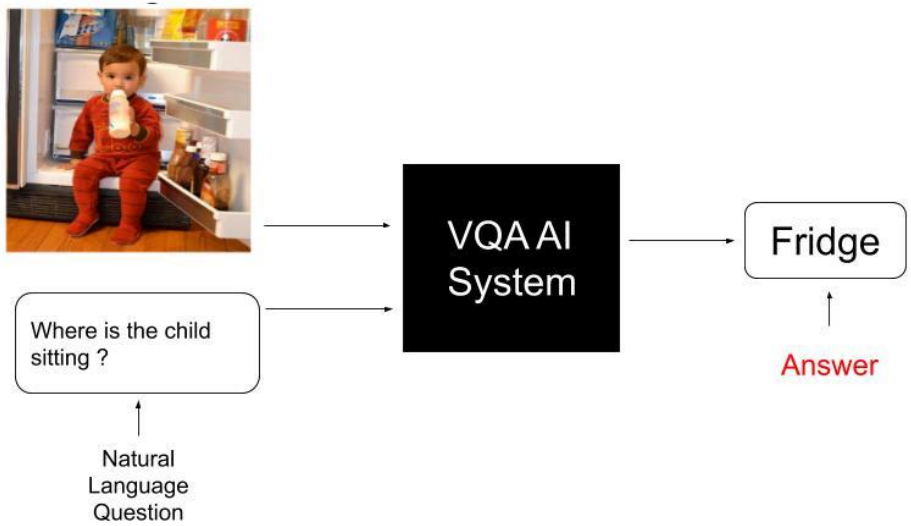
Multi-stage Attention based Visual Question Answering

Aakansha Mishra, Ashish Anand, Prithwijit Guha

Problem Statement

- Given an image (**I**) and a question (**q**), the objective of VQA system is to predict the most probable answer (**a**)
- Most of literature solved VQA as classification problem



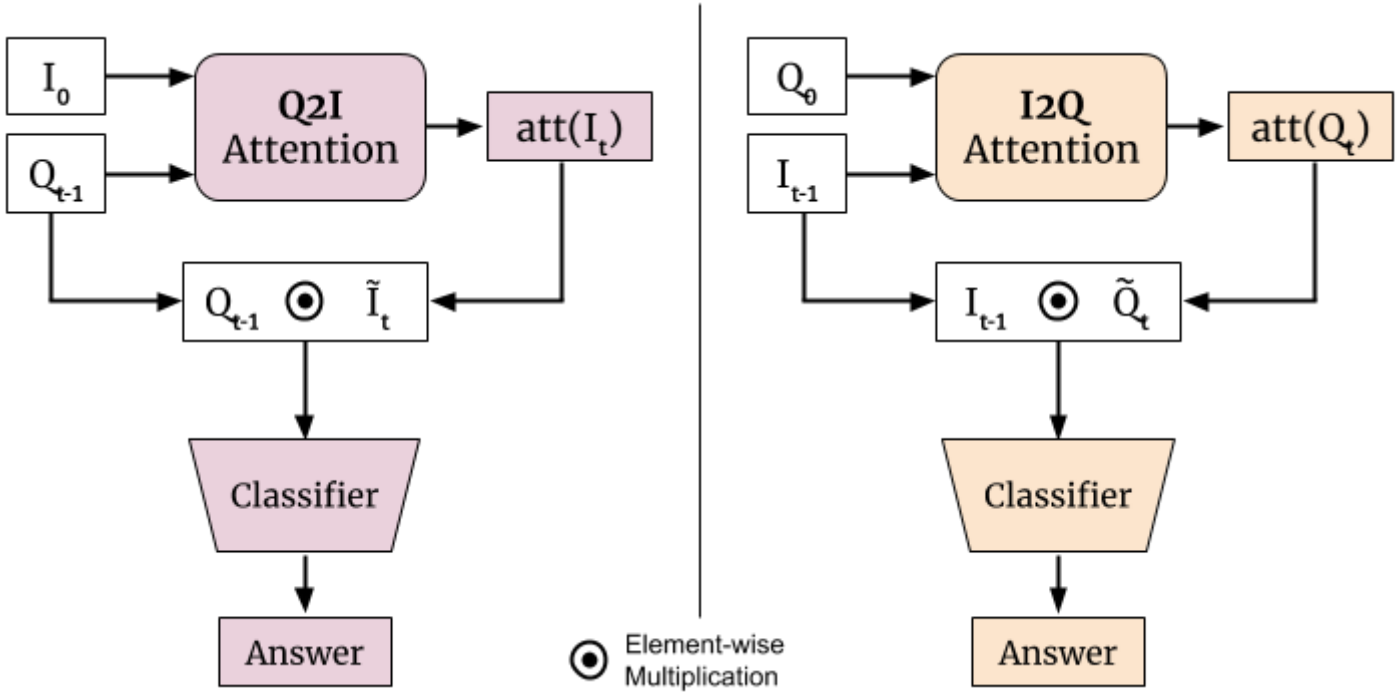
Motivation

- In general, it is always not possible to understand input (image, question, video) in one go.
- Specifically for VQA, it is always not possible to understand image in one go
- Sometimes it is required to look into image multiple times in context of question and to understand question in context of image.
- With this intuition our proposed model works.

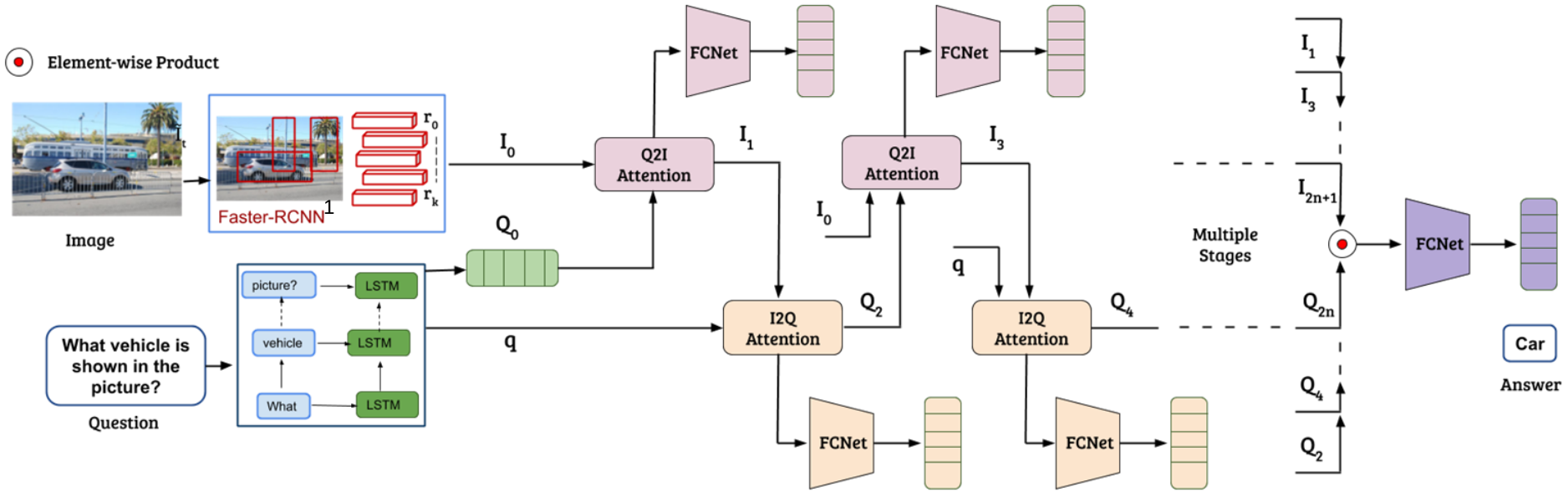
Key Contributions

- An alternate attention module that helps to learn better embedding for the image and question is proposed
- Multi-stage loss is proposed to overcome the gradient vanishing / explosion problem
- Cosine Normalization (cosine distance) based metric is a better distance measure in the joint embedding space and provides better performance.

Framework: Basic Modules

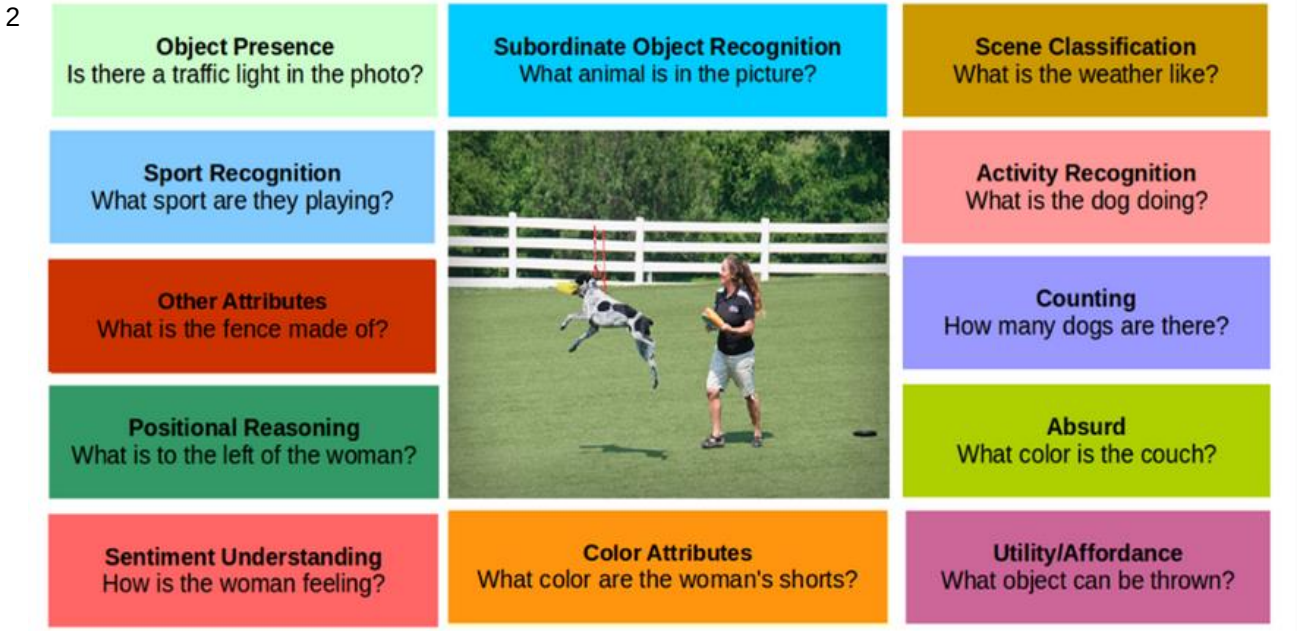


Overall Framework



[1] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015.

Dataset: Task Directed Image Understanding Challenge (TDIUC)



[2] Kafle et al, “An Analysis of Visual Question Answering Algorithms”, ICCV 2017.

Evaluation Metrics

- Overall Accuracy: Ratio of correctly predicted with that of total samples
- Mean-Per-Type: Due to skewed distributions of question types, if each test question is treated equally, then it is difficult to assess performance on rarer question-types
- To compensate, compute accuracy for each question-type separately
- Final unified accuracy is computed as Arithmetic and Harmonic Means across all per question-type accuracies, referred to as Arithmetic Mean-Per-Type (Arithmetic MPT) Accuracy and Harmonic Mean-Per-Type Accuracy (Harmonic MPT)

Result Comparison - I

Question Type	SAN	RAU	MCB	QTA	BAN	Ours
Scene Recognition	92.30	93.96	93.06	93.80	93.10	**94.64**
Sport Recognition	95.50	93.47	92.77	95.55	95.7	**95.90**
Color Attributes	60.90	66.86	68.54	60.16	67.50	**74.74**
Other Attributes	46.20	56.49	56.72	54.36	53.20	**60.62**
Activity Recognition	51.40	51.60	52.35	60.10	54.0	**61.26**
Positional Reasoning	27.90	35.26	35.40	34.71	27.9	**41.50**
Object Recognition	87.50	86.11	85.54	86.98	87.5	**88.87**
Absurd	87.51	93.40	84.82	**100.0**	94.47	94.99
Utility & Affordance	26.30	31.58	35.09	31.48	24.0	**39.77**
Object Presence	92.40	94.38	93.64	94.55	95.1	**95.78**
Counting	52.10	48.43	51.01	53.25	53.9	**57.61**
Sentiment Und.	53.60	60.09	66.25	64.38	58.70	**68.30**
Overall Accuracy	82.00	84.26	81.86	85.03	85.5	**86.90**
Arithmetic-MPT	65.00	67.81	67.90	69.11	67.4	**72.80**
Harmonic-MPT	53.70	59.00	60.47	60.08	54.9	**66.38**

Result Comparison - II

Model	Overall Accuracy
BTUP	82.91
QCG	82.05
RN	84.61
DFAF	85.55
RAMEN	86.86
MLIN⋆	**87.60**
Ours	86.90

Result Comparison - III (Without ‘Absurd’ Category)

Without Absurd				
Metrics	**MCB**	**QTA**	**BAN**	**Proposed**
Overall Accuracy	78.06	80.95	81.9	**84.58**
Arithmetic-MPT	66.07	66.88	64.6	**70.46**
Harmonic-MPT	55.43	58.82	52.8	**63.43**

Ablation Analysis – II (With Cosine Normalization)

Metrics	m = 1	m = 2	m = 3	m = 4	m = 5
Overall Accuracy	85.04	86.00	86.61	**86.90**	85.93
Arithmetic-MPT	71.92	72.16	72.02	**72.80**	70.37
Harmonic-MPT	65.83	65.89	64.73	**66.38**	61.32

Ablation Analysis - III (Without Cosine Normalization)

Metrics	m = 1	m = 2	m = 3	m = 4	m = 5
Overall Accuracy	85.02	85.51	86.12	**86.38**	85.64
Arithmetic-MPT	71.23	71.83	**72.10**	72.04	68.12
Harmonic-MPT	64.97	65.33	65.58	**65.72**	57.38

Conclusion

- An alternate attention scheme leveraging attention from textual to visual space and vice-versa to obtain enhanced feature representation in both textual and visual domains.
- Proposed bi-directional attention is applied multiple times, making it a multi-stage model.
- Multi-stage loss scheme is proposed to overcome the potential gradient vanishing problem.
- Comparative analysis indicates that the proposed bi-directional attention model outperforms existing methods on the TDIUC dataset.

Thank You !