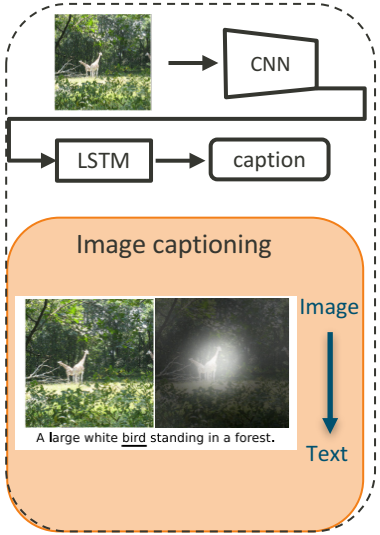
Answer-checking in Context: A Multi-modal Fully Attention Network for Visual Question Answering

MediaTek, Algorithm Department, Singapore
Huang Hantao, Tao Han, Wei Han, Deep Yap and Cheng-Ming Chiang
Dec, 2020

Multimodal Learning from Academy to Industry

Academy

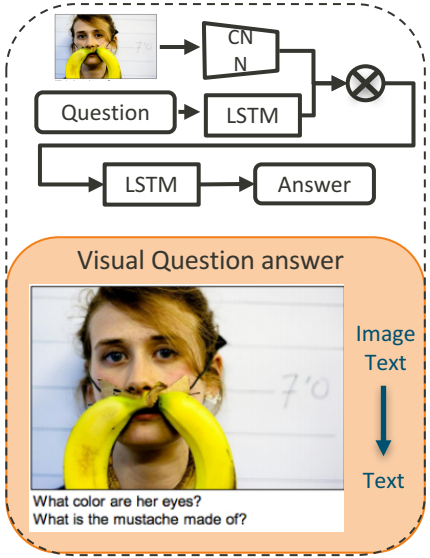
CNN+LSTM structure is widely used



2015

One modality to another [1]

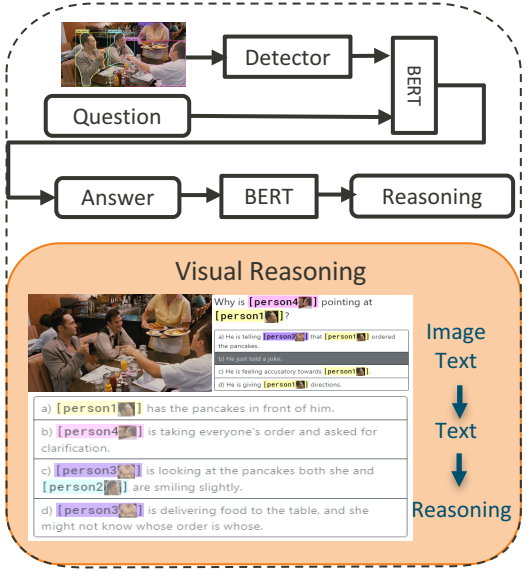
Attention-based fusion mechanism is popular



2016

Joint modality embedding learning [2]

BERT structure-based model dominates



2019

Learn the reasoning inside [3]

Industry Landing

Facebook is using AI to help blind people 'see' the photos [4]

Trend in V&L Multimodal

Motivation : We need a compact yet accurate model to land on edge devices.

Motivation and Contributions

Architecture for good accuracy

- A fully attention based VQA architecture, three attention based modules to mimic the human behavior (reading, answering and checking) to answer a question given an image

- A multi-modal answer related attention flow

Compact Model

- A layer-wise transfer learning and smaller yet more accurate.

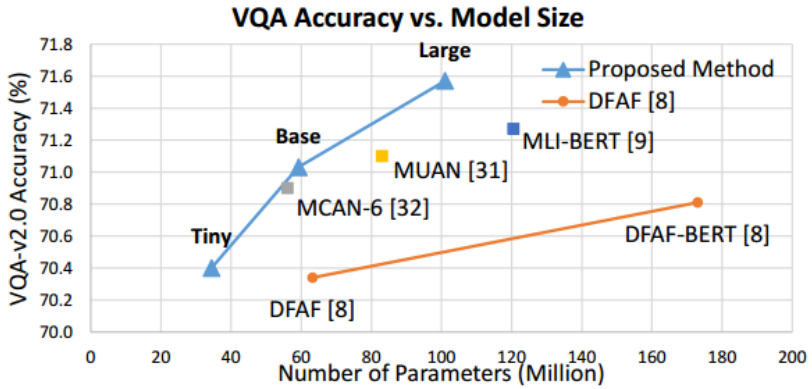


Figure 1. Proposed model performance (model size and accuracy) comparison with existing works. Our large model achieves better accuracy with smaller model size comparing to the current state-of-the-art model MLI-BERT [9] on VQA-v2.0 test-standard split.

Unified Answer-question-image Attention

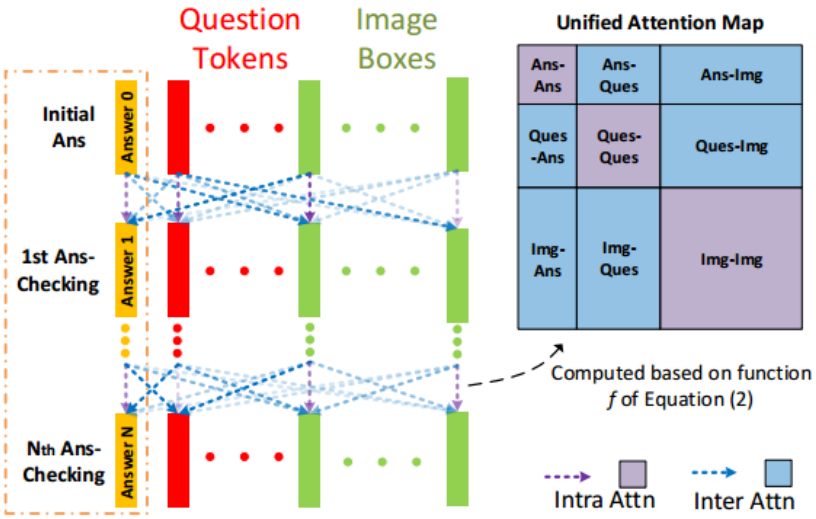
Previous works:

- Modular co-attention networks [5]
- Dynamic fusion with intra-and inter-modality attention flow [6]
  - Question-to-question/image attention flow
  - Image-to-question/image attention flow

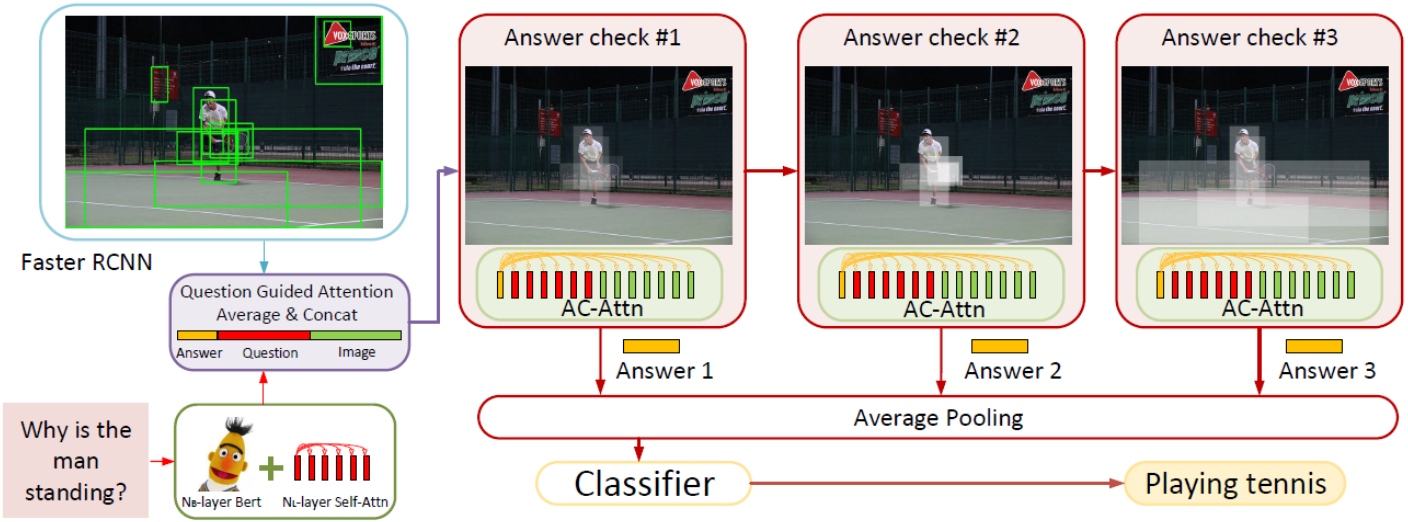
Our work: these attentions are performed at the same time, even with answer.

- Answer representation is updated at the same time.

Unified Answer-Question-Image Self Attention

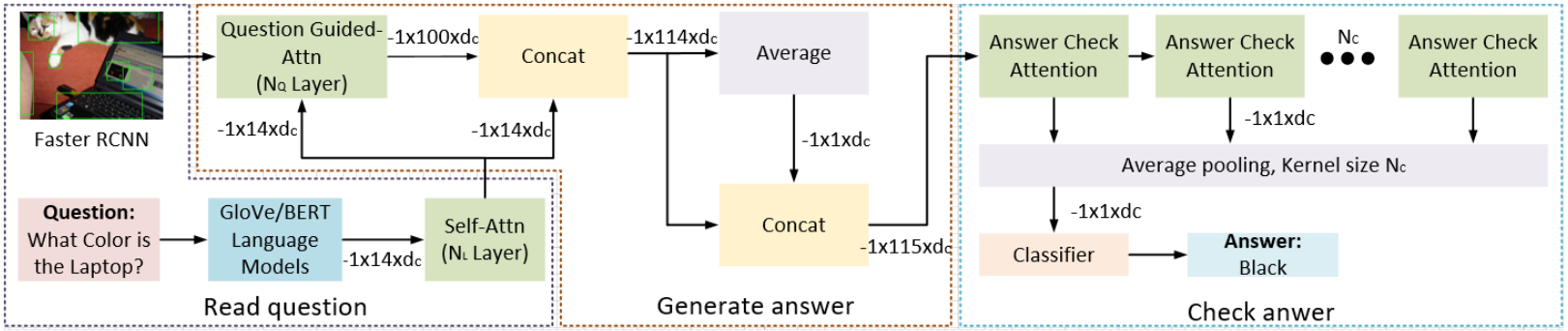


VQA Architecture General



- Proposed an **I**nterpretable **R**ead, **A**nswer and **C**heck architecture for VQA.
  - Read questions -> Generate Answer -> Check answer
- Transfer Learning to user 6 layer of BERT (total 12 layers)
- Achieve a state of the art results, 71.57% accuracy

VQA Architecture Detail



- Read question -> Generate Answer -> Check answer
  - Read question: language models/BERT
  - Generate Answer: question-guided attention
  - Check answer: answer-question-image self attention.

- Each module has its physical representation. This research work helps to improve *language+vision* applications.

VQA Performance Summary

STATE-OF-THE-ART ACCURACY COMPARISON ON A **SINGLE-MODEL** FOR THE TEST-DEV AND TEST-STANDARD SPLITS. THE RESULTS ARE COLLECTED FROM THE VQA-V2.0 COMPETITION SERVER. THE MODEL IS TRAINED ON VQA V2.0 TRAINING DATASET, VALIDATION DATASET AND VISUAL GENOME DATASET.

Model	test-dev				test-std
	Y/N	No.	Other	All	All
BUA[7]	81.8	44.2	56.1	65.32	65.67
BAN [13]	85.3	50.9	60.3	69.52	–
BAN-C [13]	85.4	54.0	60.5	70.04	70.35
DFAF [3]	86.1	53.3	60.5	70.22	70.34
DFAF-BERT [3]	86.7	52.9	61.0	70.59	70.81
MCAN [8]	86.8	53.3	60.7	70.63	70.90
MUAN [26]	86.8	**54.4**	60.9	70.82	71.10
MLI [6]	86.0	52.9	60.4	71.28	70.28
MLI-BERT [6]	87.1	53.4	60.5	71.09	71.27
Unified-VLP [33]	87.4	52.1	60.5	70.6	70.7
Visual-BERT [4]	–	–	–	70.8	71.0
VilBERT [5]	–	–	–	70.6	70.9
QBN [34]	87.1	52.93	60.8	70.8	71.0
ARAC-4-GloVe	85.9	52.5	60.5	70.06	70.40
ARAC-4-BERT-1	86.8	53.0	61.2	70.81	71.03
ARAC-5-BERT-6	**87.4**	54.1	**61.6**	**71.34**	**71.57**

We achieve a **state-of-the-art** performance **71.57%** accuracy

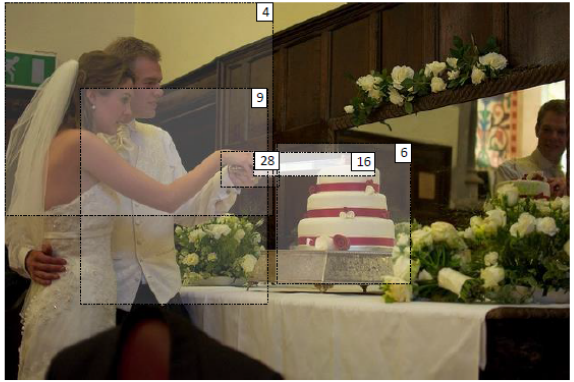
NUMBER OF MODEL PARAMETERS COMPARISON INCLUDING WORD EMBEDDING ON VQA-V2.0 VALIDATION DATASET

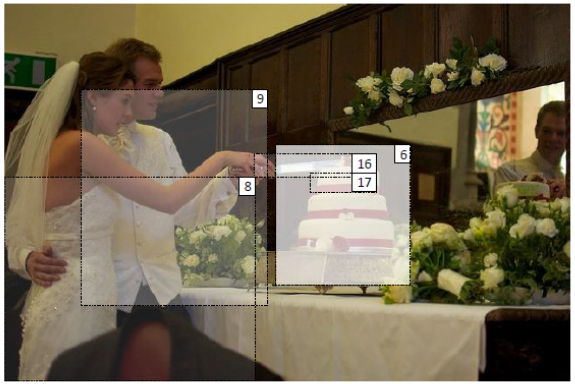
Model Name	Model Size	Size Ratio	Acc (%)
BAN-4 [13]	44.8M	0.76	65.81
MCAN-6[8]	56M	0.95	67.20
MUAN-768 [26]	83M	1.40	67.28
MUAN-1024 [26]	141.6	2.39	67.30
DFAF [3]	63.2M	1.07	66.66
DFAF-BERT [3]	173.2M	2.93	-
MLI-BERT [6]	120M	2.03	67.83
ARAC-4-GloVe	34.4M	0.58	66.89
ARAC-4-BERT-1	59.2M	1	67.48
ARAC-5-BERT-6	101.0M	1.71	**68.14**

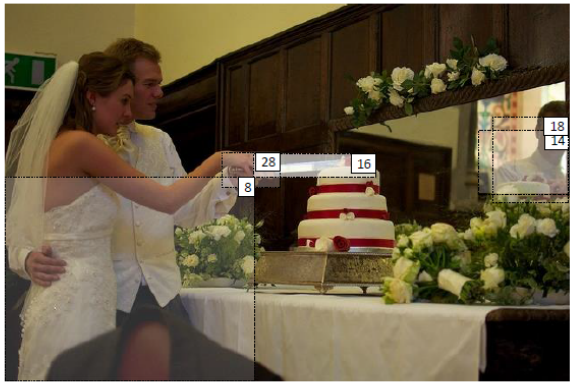
Provide a smaller model and better accuracy comparing to existing works.

Answer check figures: Visualization

Q: What is the woman holding in her right hand?
A: knife
P: knife







Reference

[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

[2] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

[3] Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[4] Facebook News, https://www.forbes.com/sites/amitchowdhry/2016/04/07/facebook-automatic-alternative-text/?sh=6d6426367c86

[5] Yu Z, Yu J, Cui Y, et al. Deep modular co-attention networks for visual question answering Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 6281-6290.

[6] Gao, Peng, et al. "Dynamic fusion with intra-and inter-modality attention flow for visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Note: some reference the Table could be found in our paper:

Huang, Hantao, et al. "Answer-checking in Context: A Multi-modal Fully Attention Network for Visual Question Answering." arXiv preprint arXiv:2010.08708 (2020).

MEDIATEK

everyday genius