Question Type Guided Attention in Visual Question Anwering

Yang Shi [1 2] Tommaso Furlanello[1 3] Sheng Zha [1] Animashree Anandkumar [1 4]

[1]Amazon AI [2]University of California Irvine [3]University of Southern California [4]California Institution of Technology

Keypoints

- Propose question type-guided attention(QTA) to balance between bottom-up and top-down visual features.
- Propose a multi-task extension that is trained to predict question types from the lexical inputs during training time that do not require ground truth labels during inference.
- QTA systematically improves the performance by more than 5% across multiple question type categories such as "Activity Recognition", "Utility" and "Counting" on TDIUC dataset.

VQA Task

- Provide a natural language answer given any image and any open-ended question.
- Require a joint representation of both visual and textual input.

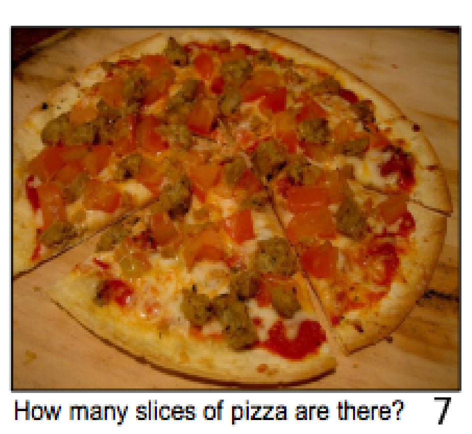

Figure 1: VQA task sample

VQA Dataset: TDIUC

- Total questions: 1653842, total images:179994.
- Categorized questions: Each question belongs to one of the 12 categories.
- Absurd questions: Questions that are totally irrelevant to the image.

Feature Extraction

Image and text features are extracted from pretrained/end-to-end neural networks.

Image pretrained model:

- ResNet
- Faster R-CNN

Question model:

- Word2Vec
- Skipthought
- GNMT encoder
- End-to-end LSTM

Question Type Guided Attention

Intuition: (1)Question type is very important in predicting the answer. (2)Explore different Image features: top-down and bottom-up features.

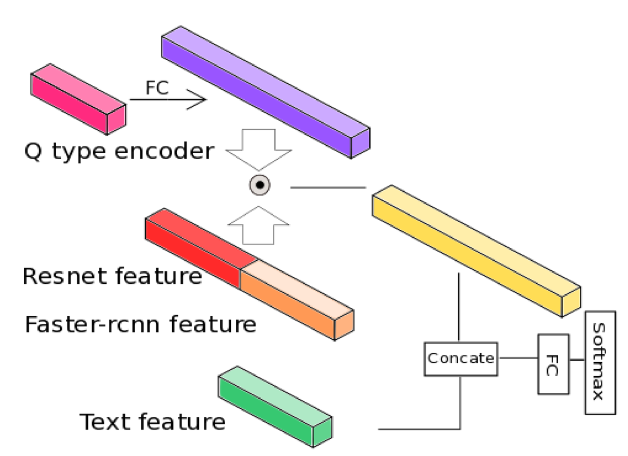

Figure 2: QTA structure

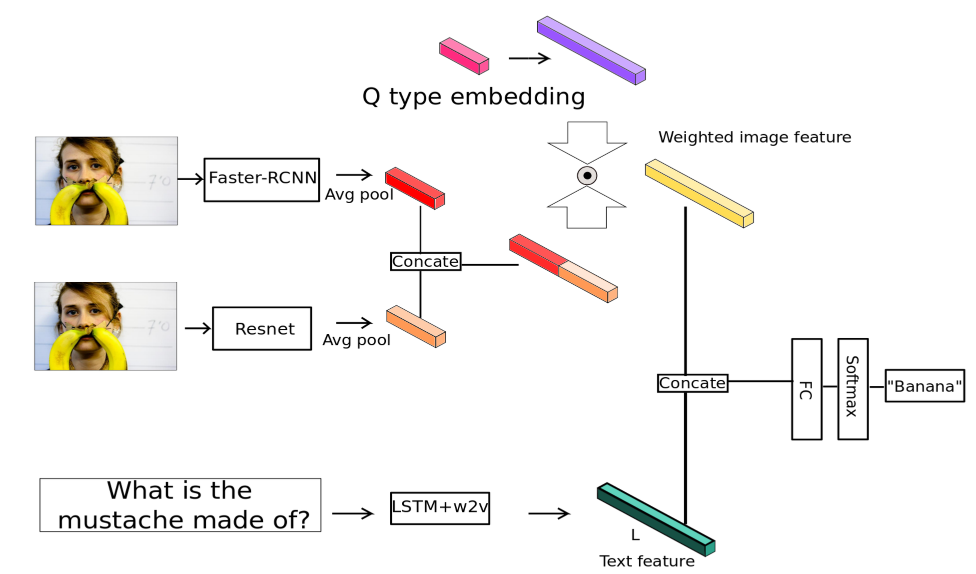

Figure 3: Concatenation model with QTA structure for VQA task(CATL-QTA^W)

Given concatenated image feature $F = [F_1, F_2, \cdots F_k] \in \mathcal{R}^M$. Assume there are N different question types, QTA is defined as $F \circ WQ$, where $Q \in \mathcal{R}^N$ is the one-hot encoding of the question type and $W \in \mathcal{R}^{M \times N}$ is the hidden weight, $\circ$ is element-wise product.

Multi-task for QTA network

Limitation of QTA: Requires question type label.
Solution: Predict the question type from text, and use it as input to the QTA network.

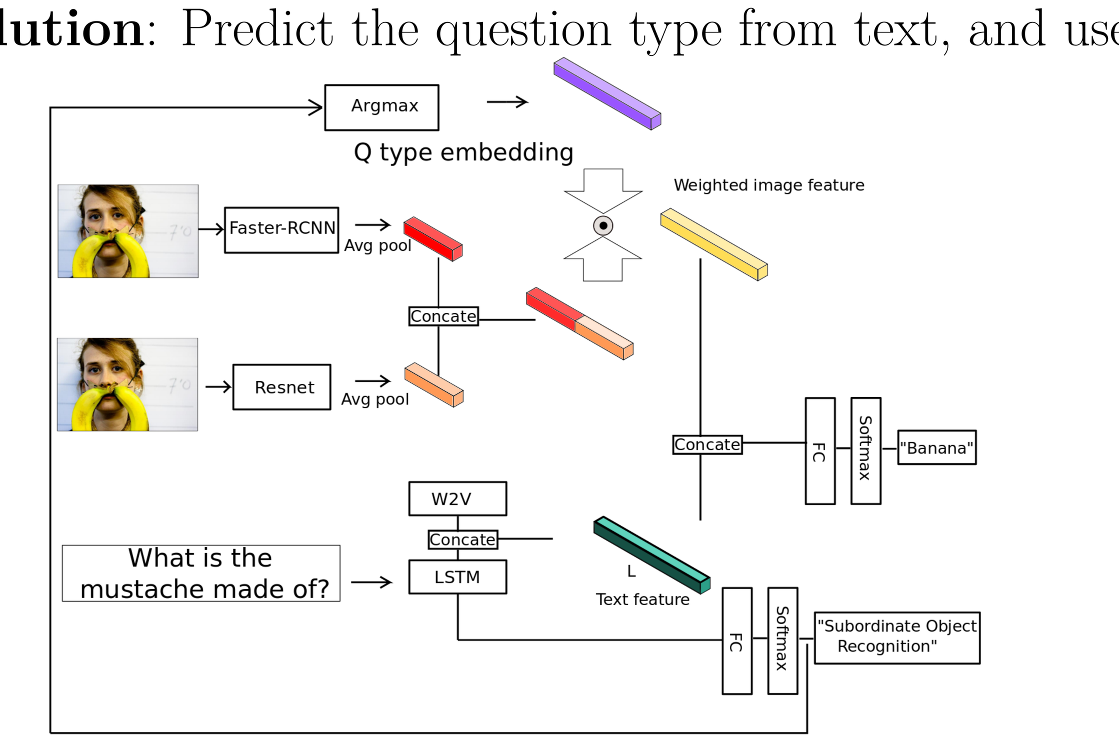

Figure 4: Concatenation model with QTA structure for multi-task

	Accuracy(%)
Element-wise Sum [1]	56.50
Concatenation [1]	57.49
Concatenation + FC [1]	58.40
Concatenation + FC + FC [1]	57.10
Element-wise Product [1]	58.57
Element-wise Product + FC [1]	56.44
Element-wise Product + FC + FC [1]	57.88
MCB(2048 × 2048 → 16K) [1]	59.83
CATL-QTA-M + FC	**60.32**

Table 1: Results of test-dev accuracy on VQA v1. Models are trained on the VQA v1 train split and tested on test-dev

Table 2: Results of QTA models on TDIUC dataset compared to state-of-art models. W denotes that additional Word2Vec embedding is concatenated to LSTM output

Accuracy(%)	**CATL**	**CATL-QTA**	**CATL^W**	**CATL-QTA^W**	**MCB-QTA**	**MCB-A** [2]	**RAU** [2]
Scene Recognition	93.18	93.45	93.31	93.80	93.56	93.06	**93.96**
Sport Recognition	94.69	95.45	94.96	95.55	**95.70**	92.77	93.47
Color Attributes	54.66	56.08	57.59	60.16	59.82	**68.54**	66.86
Other Attributes	48.52	50.30	52.25	54.36	54.06	**56.72**	56.49
Activity Recognition	53.36	58.43	54.59	60.10	**60.55**	52.35	51.60
Positional Reasoning	32.73	31.94	33.63	34.71	34.00	**35.40**	35.26
Sub. Object Recognition	86.56	86.76	86.52	86.98	**87.00**	85.54	86.11
Absurd	95.03	100.00	98.01	**100.00**	100.00	84.82	96.08
Utility and Affordances	29.01	23.46	29.01	31.48	**37.04**	35.09	31.58
Object Presence	93.34	93.48	94.13	**94.55**	94.34	93.64	94.38
Counting	50.08	49.93	52.97	53.25	**53.99**	51.01	48.43
Sentiment Understanding	56.23	56.87	62.62	64.38	65.65	**66.25**	60.09
Overall (Arithmetic MPT)	65.62	66.34	67.46	69.11	**69.69**	67.90	67.81
Overall (Harmonic MPT)	55.95	54.60	57.83	60.08	**61.56**	60.47	59.00
Overall Accuracy	82.23	83.62	83.92	**85.03**	84.97	81.86	84.26

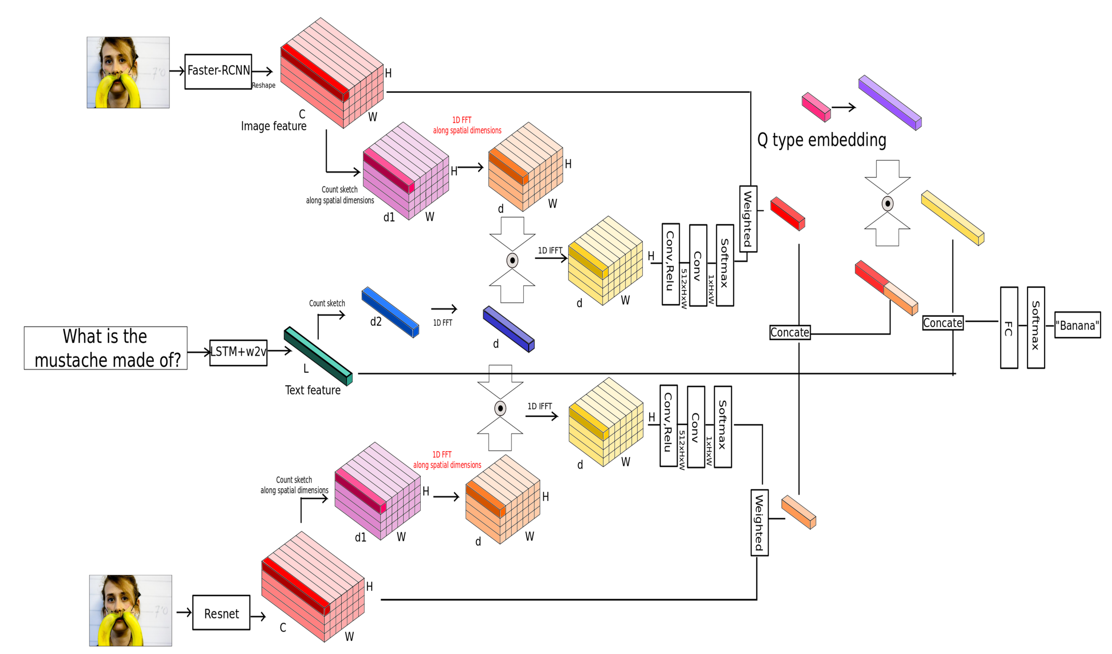

Figure 5: MCB model with QTA structure for VQA task(MCB-QTA)

Analysis

- **Top-down v.s. bottom-up visual features** Different question types need different visual features.
- **Pre-trained v.s. Jointly-trained text features** Jointly-trained text feature is better than pre-trained ones when corpus is large enough.
- **QTA v.s. Prior SOTA** QTA shows better performance than complicated deep networks such as RAU and MCB-A.
- **Multi-task** Apply to VQA v1.0 dataset that doesn't have question type information. Its performance is better than MCB's performance with approximately same number of parameters in the network.

Limitation of TDIUC

- **Bias** More than 60% of absurd questions start with "What color". Consequently, "Color" and "Absurd" question type predictions are most often

Reference

[1] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP 2016.*

[2] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017.