Multi-modal Factorized High-order Pooling for Visual Question Answering

Team **HDU-USYD-UNCC** with members

Zhou Yu[1], Jun Yu[1], Chenchao Xiang[1], Dalu Guo[2], Jianping Fan[3] and Dacheng Tao[2]

[1]Hangzhou Dianzi University, China

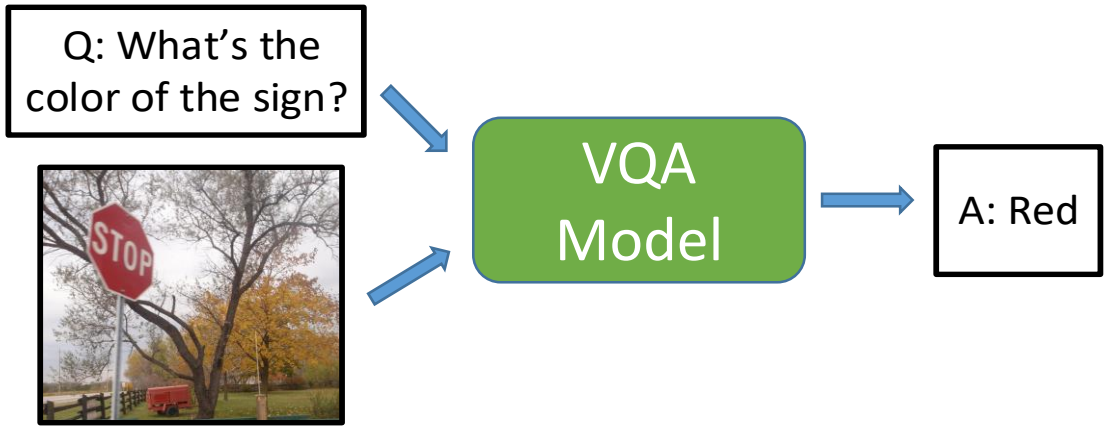
[2]The University of Sydney, Australia

[3]University of North Carolina at Charlotte, USA

26th July @ Honolulu, Hawaii

The VQA Problem

- The Problem
  - Given an image and a free question(in free text) about the image, output a textual answer.



- The Core Components
  - Multi-modal feature fusion
  - Co-Attention Learning

Multi-modal feature fusion

- Common-used first-order linear pooling model
  - Concatenation
  - Summation
- Second-order bilinear pooling
  - MCB[1]: the champion of VQA-2016, very effective ☺ and converge fast ☺, but need high-dimensional output feature ☹ to guarantee good performance.
  - MLB[2]: slightly better performance than MCB ☺ with compact output feature ☺ but converge slowly ☹.
  - MFB (ours): much better performance than MCB and MFB ☺ , enjoy the both the merits of fast convergence ☺ and compact output feature ☺ simultaneously.
- High-order pooling
  - We extend the bilinear MFB to a high-order pooling model MFH with cascading several MFB blocks

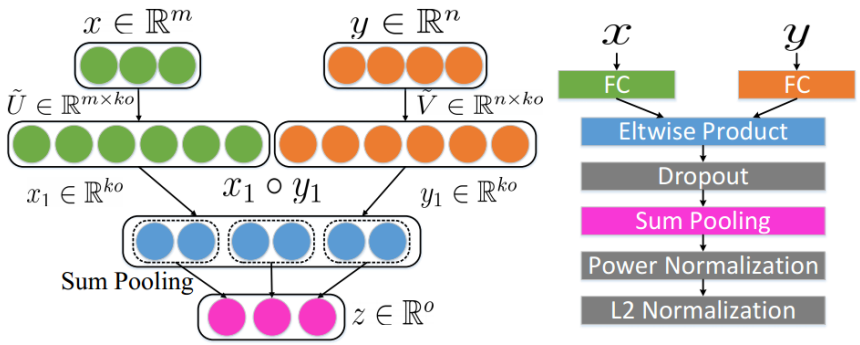
Multi-modal Factorized Bilinear Pooling (MFB)

- Formulation

$$z_i = \mathrm{MFB}(x, y) = (x^T U_i V_i^T y) = \sum_{d=1}^{k} x^T u_d v_d^T y$$
$$= \mathbb{1}^T (U_i^T x \circ V_i^T y)$$

where $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ are the multi-modal features, $z_i \in \mathbb{R}$ is *i*-th output neuron. $U_i \in \mathbb{R}^{m \times k}, V_i \in \mathbb{R}^{n \times k}$ are the **factorized** low-rank weight matrices. *k* is the rank or the factor number. To output $z \in \mathbb{R}^o$, three-order tensors $U = [U_1, \ldots, U_o] \in \mathbb{R}^{m \times k \times o}, V = [V_1, \ldots, V_o] \in \mathbb{R}^{n \times k \times o}$ are to be learned.

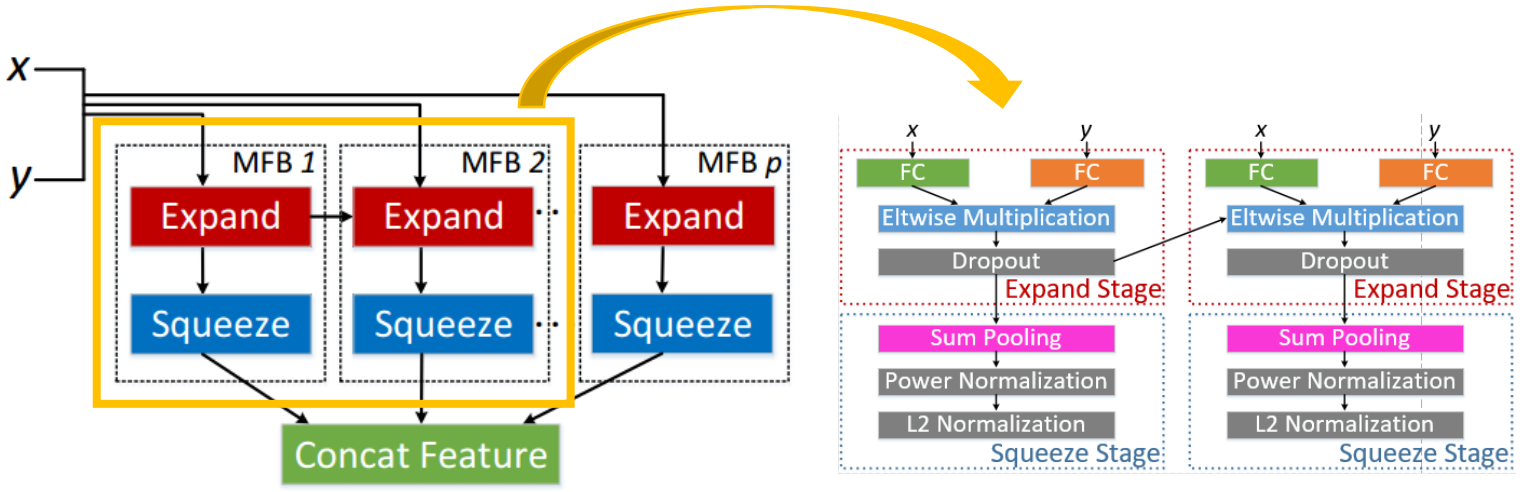
- Simple implementation with off-the-shelf layers
  - Fully-connected
  - Sum pooling (slightly modified from avg. pooling),
  - Elementwise-product
  - Feature normalizations (power & L2)



(a) Multi-modal Factorized Bilinear Pooling (b) MFB module

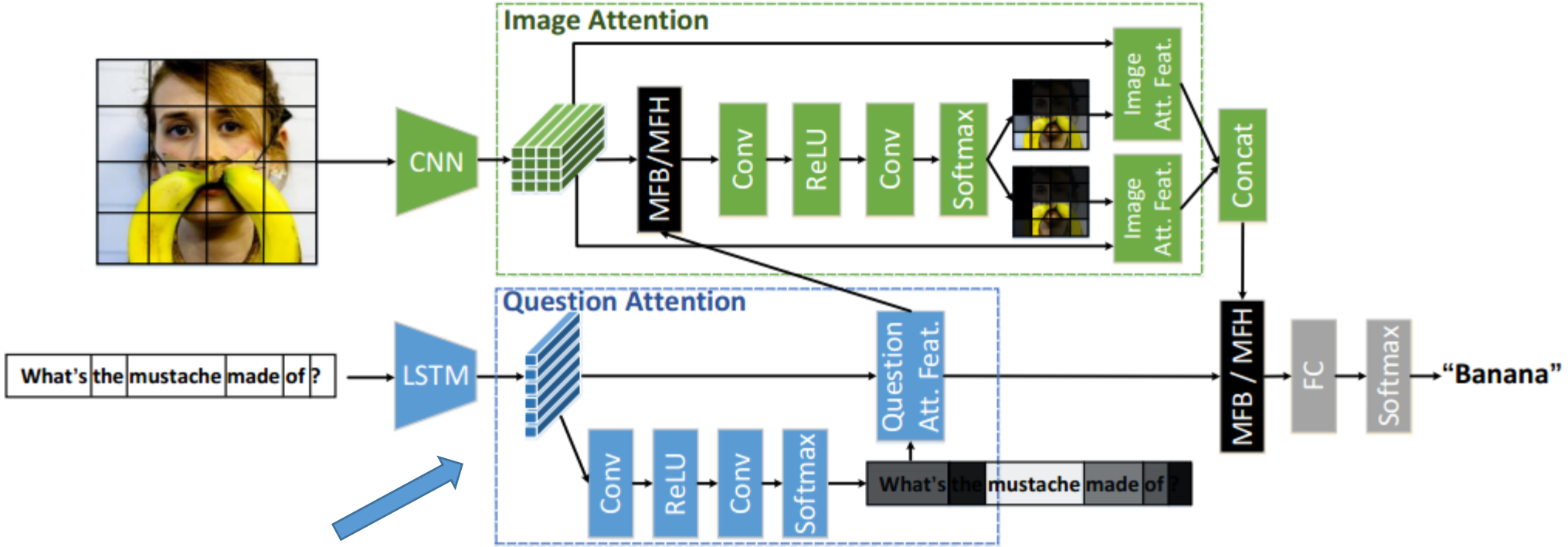
From Bilinear to High-order Pooling

- Motivation
  - Model more complex (high-order) interactions better capture the common semantic of multi-modal data.
- Multi-modal Factorized High-order Pooling (MFH)
  - MFB module is split into the *expand* and *squeeze* stages.
  - The expand stage is slightly modified to compose *p* MFB blocks (with individual parameters)
  - *p*=2 in our experiments



Network Architecture

- MFB/MFH with Co-Attention Learning



The **self-attentive** Question Attention module brings about 0.5~0.7 points improvement

Experimental Settings

- Image Features
  - *14x14x2048 res5c* feature extracted from pre-trained ResNet-152 model with input image resizing to 448x448
- Question Features
  - Single layer LSTM with 1024 hidden units.
- # of Image & Question Glimpses (Attention maps)
  - {1,2} glimpses for Question Attention (Q_{att}), {1,2,3} glimpses for Image Attention (I_{att}). The combinations different #. Q_{att} and #. I_{att} lead to different models with diversity.
- Training strategy
  - Adam solver with base learning rate 0.0007, decay every 4 epochs with exponential factor 0.25. Terminate training at 10 epochs (usually obtain the best result on 9th epoch).
  - Visual Genome dataset are used for training some models.

Results on VQA-1.0 and VQA-2.0 datasets

- Results on VQA-1.0 (test-standard) with model ensemble

Model	OE				MC
	All	Y/N	Num	Other	All
HieCoAtt [9]	62.1	80.0	38.2	52.0	66.1
RAU [35]	64.1	83.3	38.0	53.4	68.1
7 MCB models [7]	66.5	83.2	39.5	58.0	70.1
7 MLB models [12]	66.9	84.6	39.1	57.8	70.3
Human [5]	83.3	95.8	83.4	72.7	91.5
7 MFB models	68.4	85.6	41.0	59.8	72.5
7 MFH models	**69.2**	**86.2**	**41.7**	**60.8**	**73.4**

- Results on VQA-2.0 (VQA Challenge 2017)

THE OVERALL ACCURACIES ON THE TEST-DEV AND TEST-CHALLENGE SETS OF THE VQA-2.0 DATASET

Model	Test-Dev	Test-Challenge
vqateam-Prior	-	25.98
vqateam-Language	-	44.34
vqateam-LSTM-CNN	-	54.08
vqateam-MCB	-	62.33
Adelaide-ACRV-MSR (1st place)	-	**69.00**
DLAIT (2nd place)	-	68.07
LV_NUS (4th place)	-	67.62
1 MFB model	64.98	-
1 MFH model	65.80	-
7 MFB models	67.24	-
7 MFH models	67.96	-
9 MFH models (2nd place)	**68.02**	68.16

Observations:

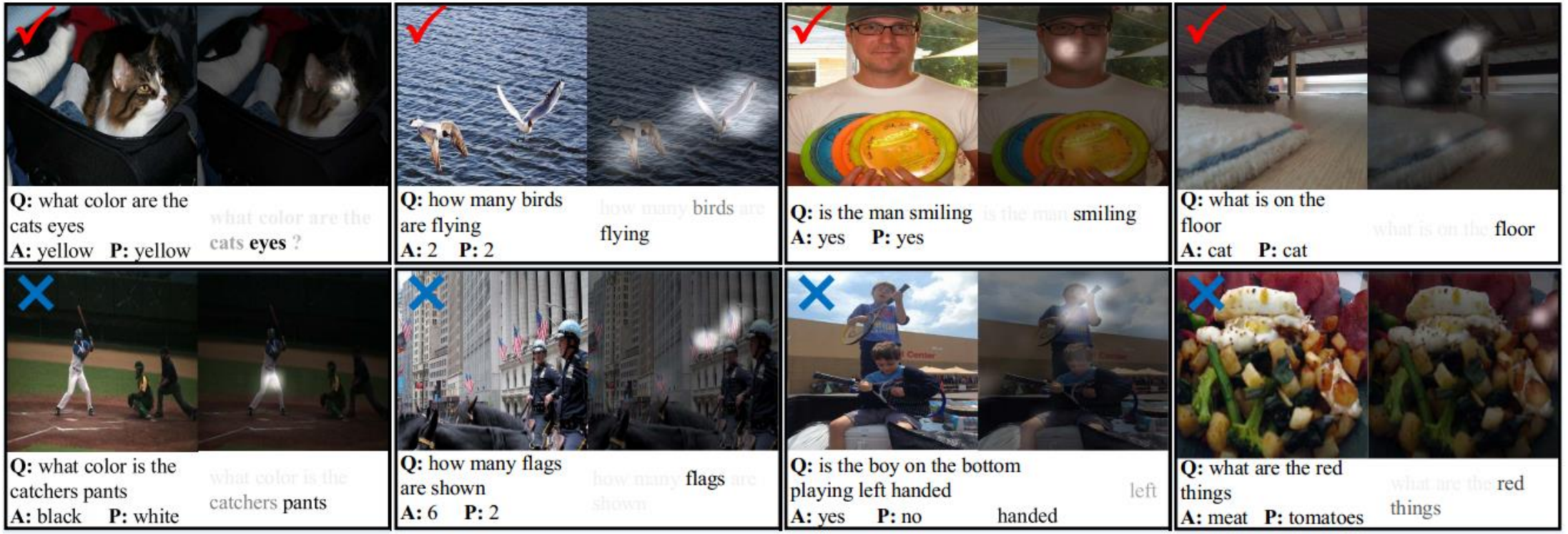
- MFB outperform the MCB and models with 1.5~2 points.
- MFH models are about 0.7~0.9 points higher than MFB models steadily.

With an ensemble of 9 models, we achieved the second place (tied with another team) on the Test-challenge set

Leaderboard:
http://visualqa.org/roe_2017.html

Effects of the Co-Attention Learning

- Image and question attentions of the MFB+CoAtt+GloVe model



Thanks for your attention!

- References

[1]. Fukui et al., Multimodal compact bilinear pooling for visual question answering and visual grounding, CVPR 2016

[2]. J. Kim et al., Hadamard product for low-rank bilinear pooling. ICLR 2017

- Code and pre-trained models for MFB and MFH are released at
  - https://github.com/yuzcccc/mfb
- Our Papers:
  - The MFB paper is *accepted by ICCV 2017: https://arxiv.org/abs/1708.01471*
  - The extended MFH paper is under review*: https://arxiv.org/abs/1708.03619*