Multimodal Compact Bilinear Pooling for VQA

Akira Fukui[1,2], Dong Huk Park[1], Daylen Yang[1],
Anna Rohrbach[1,3], Trevor Darrell[1], Marcus Rohrbach[1]

[1]UC Berkeley EECS, CA, United States
[2]Sony Corp., Tokyo, Japan
[3]Max Planck Institute for Informatics, Saarbrücken, Germany

Berkeley
UNIVERSITY OF CALIFORNIA

Multimodal language and visual understanding

Description

A table full of food for a feast

Multimodal language and visual understanding

Grounding

The bowl with the brown souce

Multimodal language and visual understanding

Visual Question Answering

What is the brown souce?

Gravy

How to Combine Image Representation and Question Representation?

CNN

spoon
plate
bowl
table
food
corn
…
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
…

Yes

- ☐ **All elements can interact**
- ☐ **Multiplicative interaction**

How to Combine Image Representation and Question Representation?

CNN

spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...

Concatenate

FC

RELU

FC

Yes

☑ **All elements can interact**

☐ **Multiplicative interaction**

- Difficult to learn output classification

How to Combine Image Representation and Question Representation?

CNN

spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...

Elementwise Multiplication

FC

RELU

FC

Yes

☐ **All elements can interact**

☑ **Multiplicative interaction**

- Difficult to learn input embedding

How to Combine Image Representation and Question Representation?

CNN

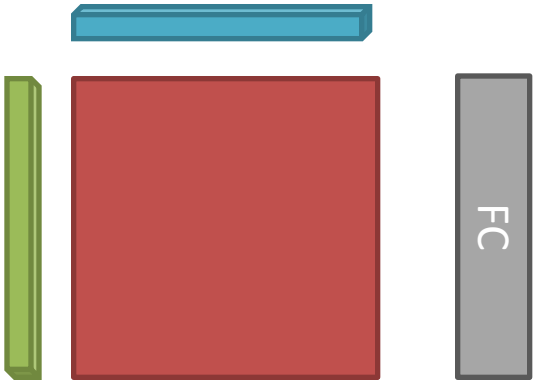
spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...

Outer Product / Bilinear Pooling [Lin ICCV 2015]



☑ **All elements can interact**

☑ **Multiplicative interaction**

[Lin ICCV 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. ICCV 2015

How to Combine Image Representation and Question Representation?

CNN

spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...

Outer Product / Bilinear Pooling

[Lin ICCV 2015]

2048

2048

4 million

FC

Yes

4 million x 1000

- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☐ **High #activations & computation**
- ☐ **High #parameters**

[Lin ICCV 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. ICCV 2015

Multimodal Compact Bilinear Pooling

CNN

spoon
plate
bowl
table
food
corn
...
person

Is this going to be a feast?

LSTM

Is?
feast
going to be
...

Compact Bilinear Pooling [Gao CVPR 16]

2048

2048

MCB

FC

Yes

16k x 1000

- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☑ **Low #activations & computation**
- ☑ **Low #parameters**

[ICLR Workshops 2016] Fine-grained pose prediction, normalization, and recognition N Zhang, E Shelhamer, Y Gao, T Darrell

[Gao CVPR 16] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CVPR 2016

Multimodal Compact Bilinear Pooling

CNN

Is this going to be a feast?

LSTM

Random Projection: Countsketch Ψ

Pham & Pagh (2013):

$$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$$

FC

Yes

16k x 1000

- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☐ **Low #activations & computation**
- ☑ **Low #parameters**

[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.

[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013

Multimodal Compact Bilinear Pooling

CNN

Is this going to be a feast?

LSTM

Ψ

Ψ

Convolution

∗

Pham & Pagh (2013):

$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$

FC

Yes

16k x 1000

- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☑ **Low #activations & computation**
- ☑ **Low #parameters**

[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.
[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013

Multimodal Compact Bilinear Pooling

CNN

Is this going to be a feast?

LSTM

Ψ

Ψ

Convolution

FFT

FFT

FFT-1

Pham & Pagh (2013):

$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$

FC

Yes

16k x 1000

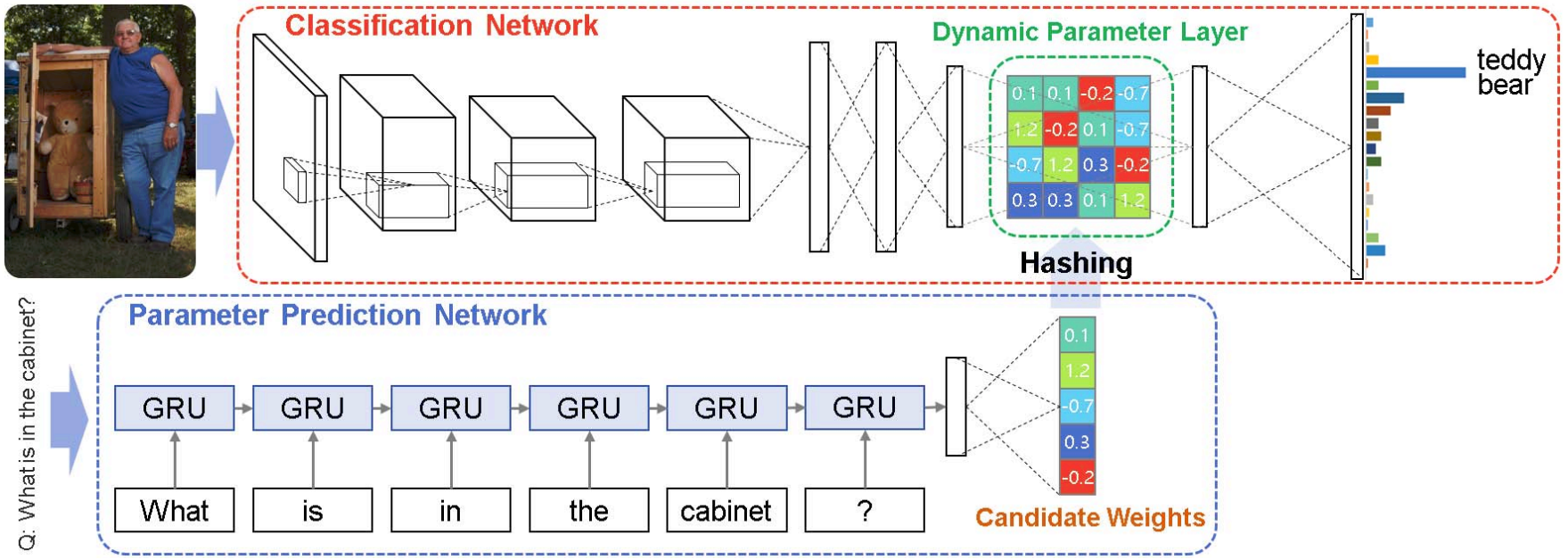
- ☑ **All elements can interact**
- ☑ **Multiplicative interaction**
- ☑ **Low #activations & computation**
- ☑ **Low #parameters**

[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.

[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013

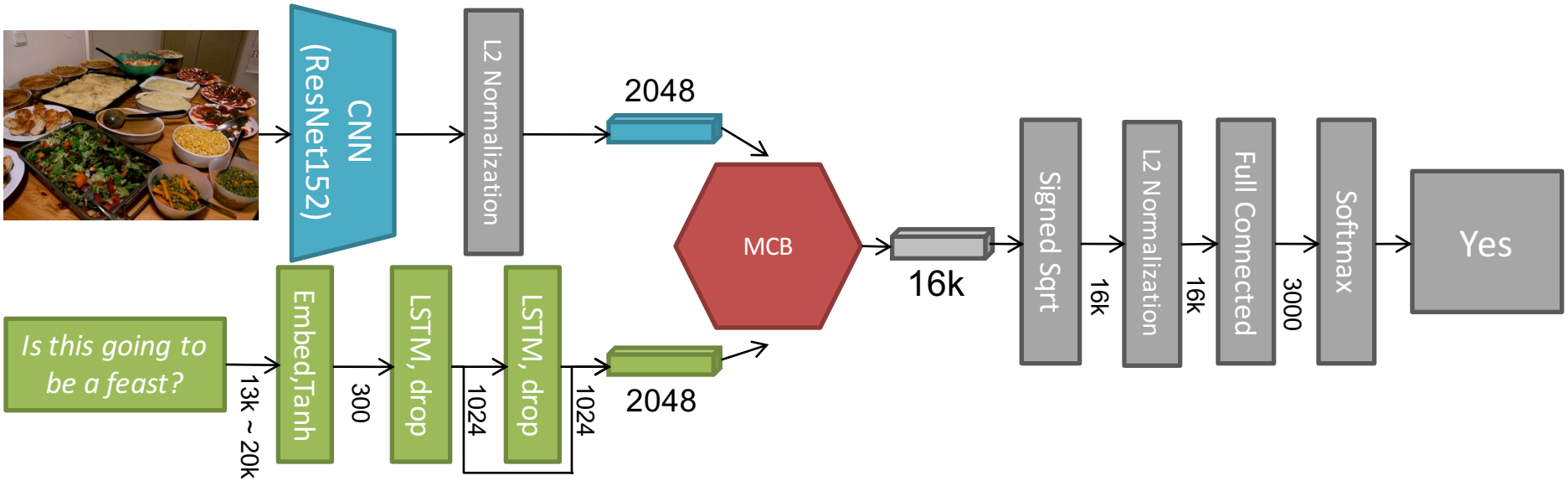
Related work

- Alternative approach to multiplicative interactions
  - DPP Net: Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. *Image question answering using convolutional neural network with dynamic parameter prediction*. CVPR 2016



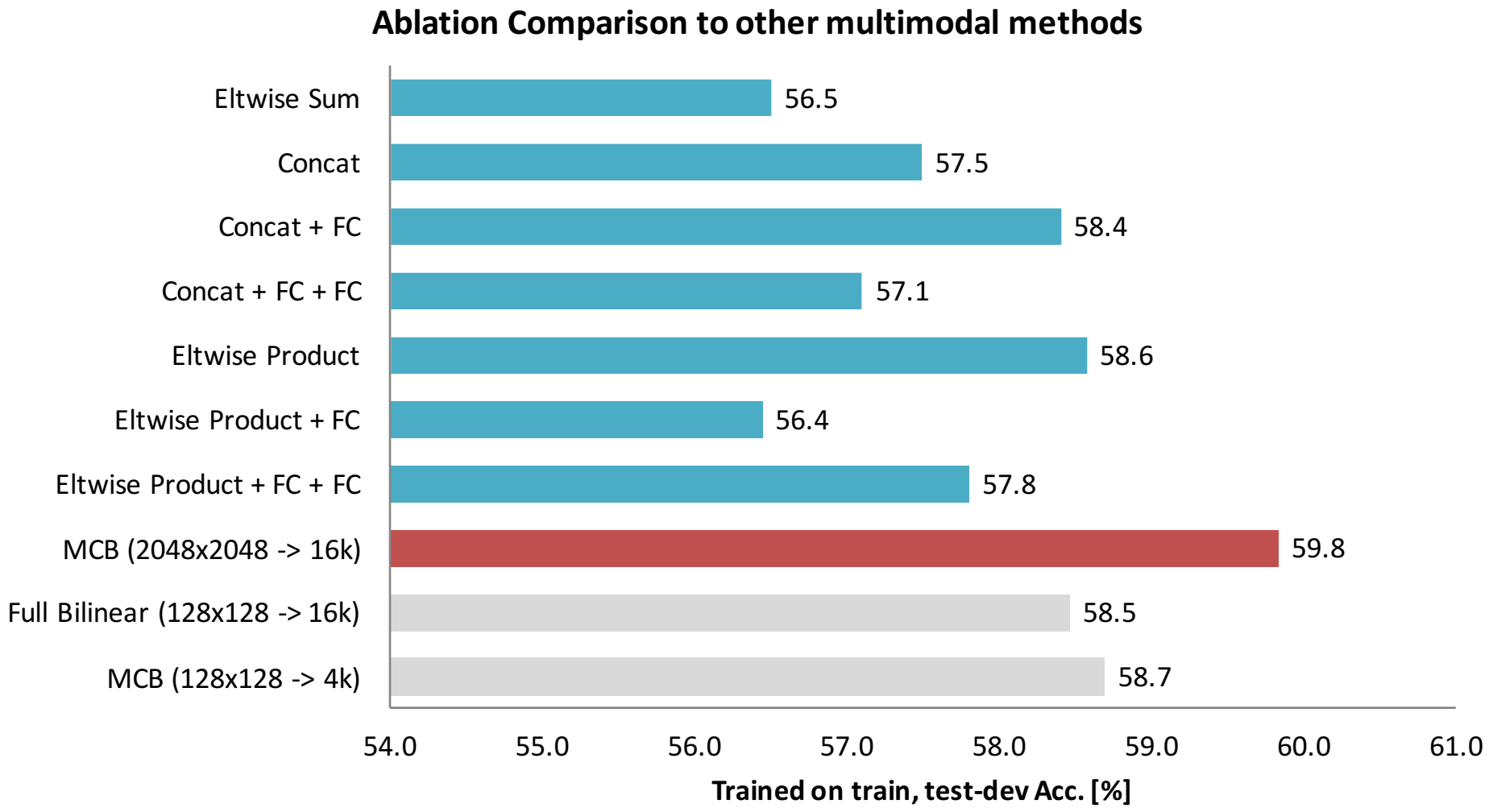
Experimental setup (without Attention)

- Solver
  - Cross-entropy-loss, Adam, learning rate 0.0007
- Feature Extraction
  - ResNet 152, image: 448x448
- Answers
  - 3000 most frequent on train
  - Sampling with probability of answers
- Trained on train / validated on val / tested on test-dev



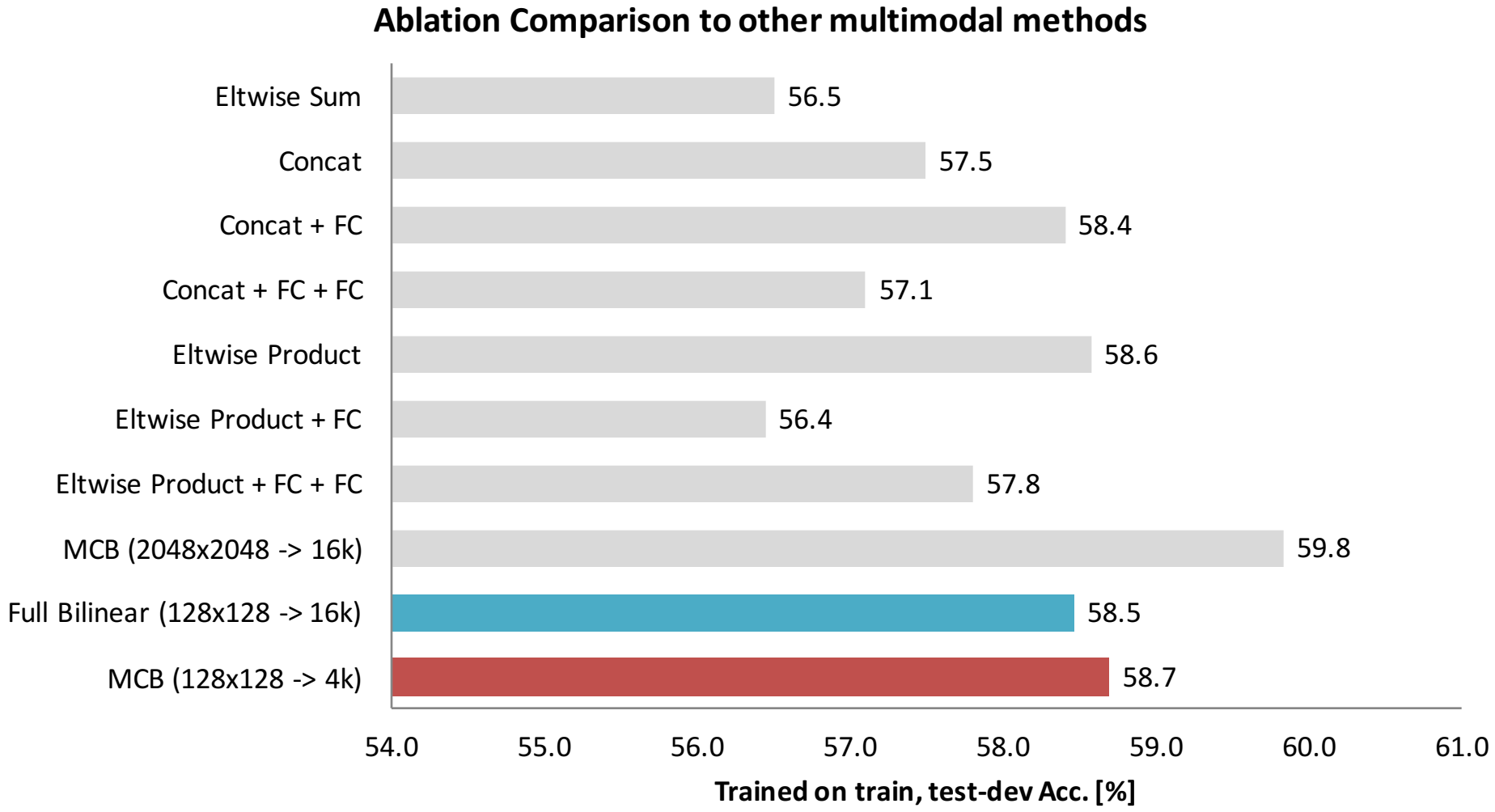
Ablation Comparison to other multimodal methods

- MCB achieves highest accuracy



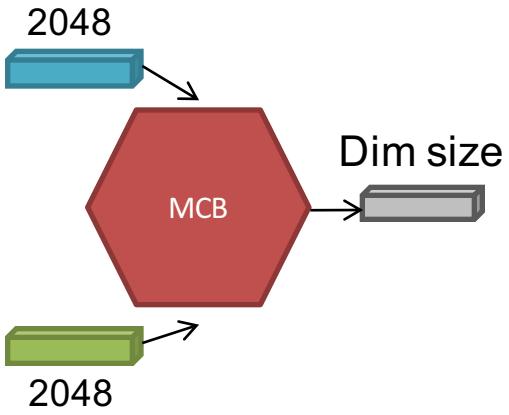
Ablation Comparison to other multimodal methods

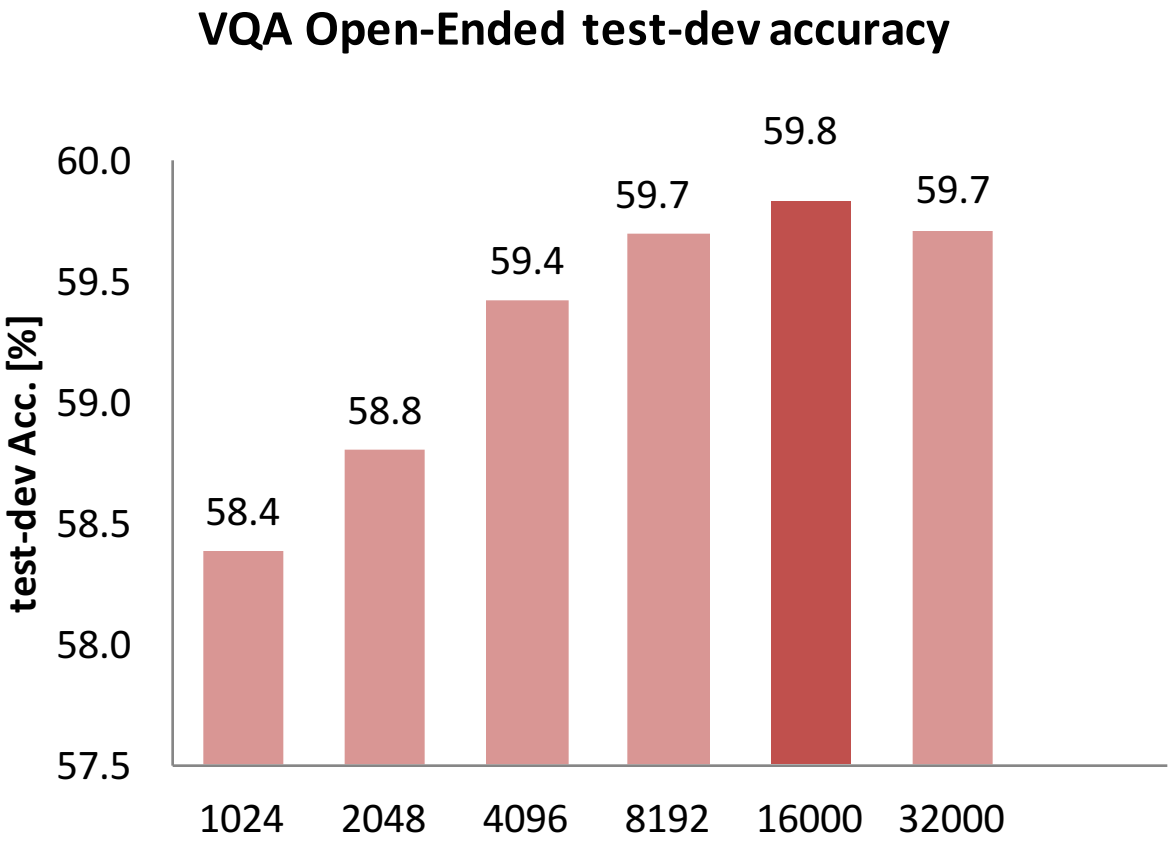
- MCB comparable to Full Bilinear



Dimensionality of MCB

- Dimensionality of MCB decides the performance of outer product approximation





Multimodal language and visual understanding

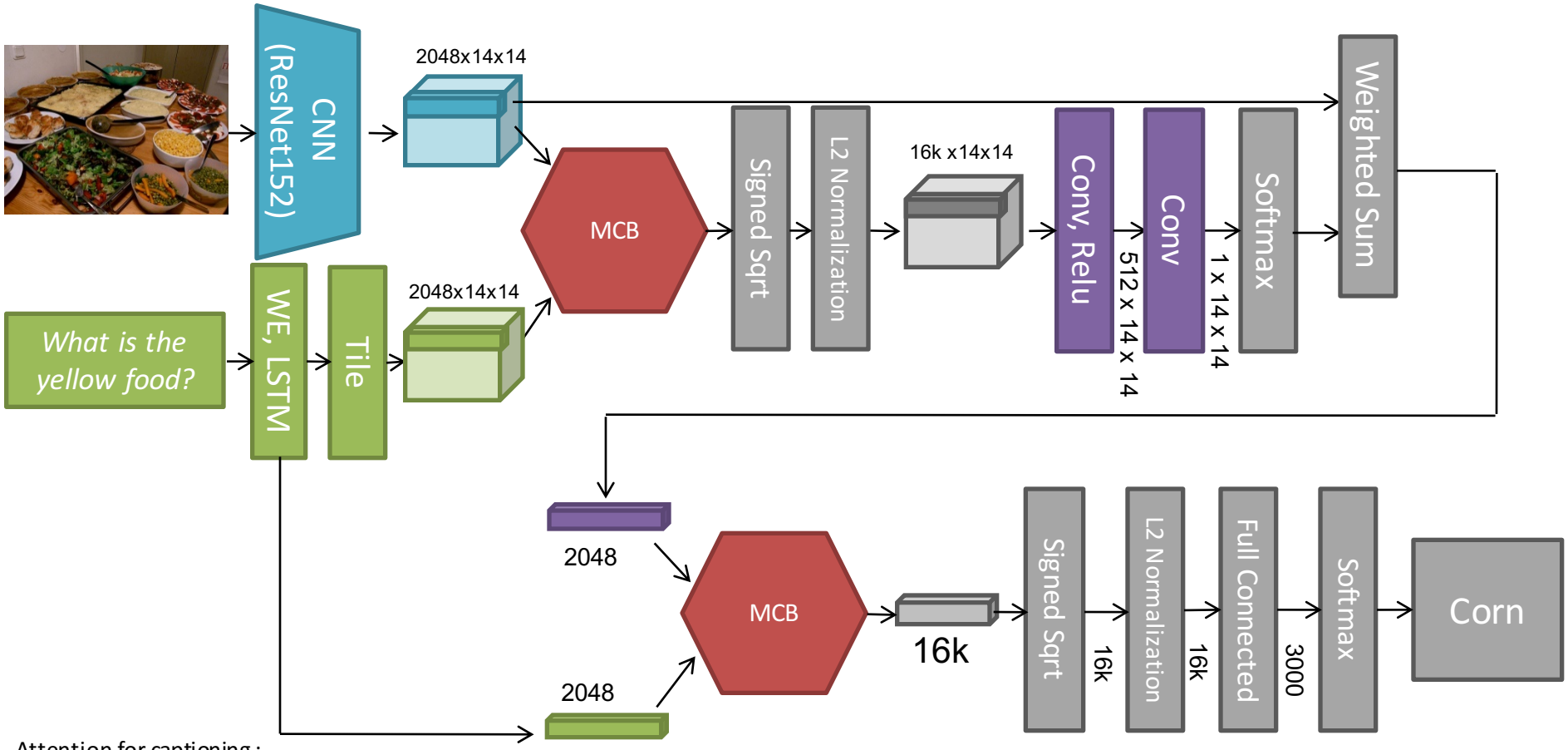
Visual Question Answering

What is the brown souce?

Gravy

MCB with Attention

- Predict spatial attentions with MCB



Attention for captioning :
- K. Xu, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Attention for VQA :
- H. Xu, K. Saenko Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering
- J.Lu Hierarchal Question-Image Co-Attention for Visual Question Answering

Attention Visualizations

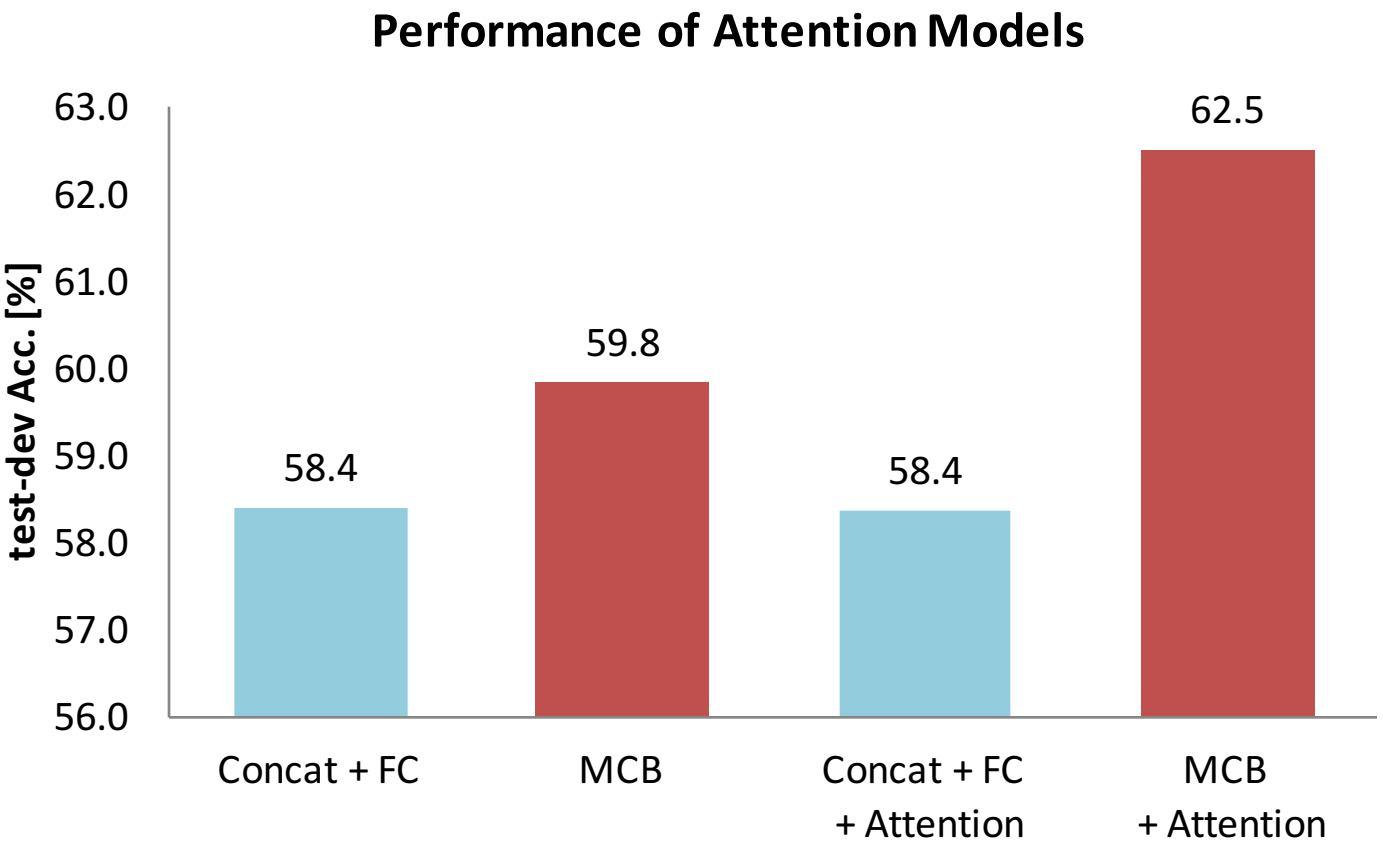
Is this person wearing a hat?

Yes

[Groundtruth: Yes]

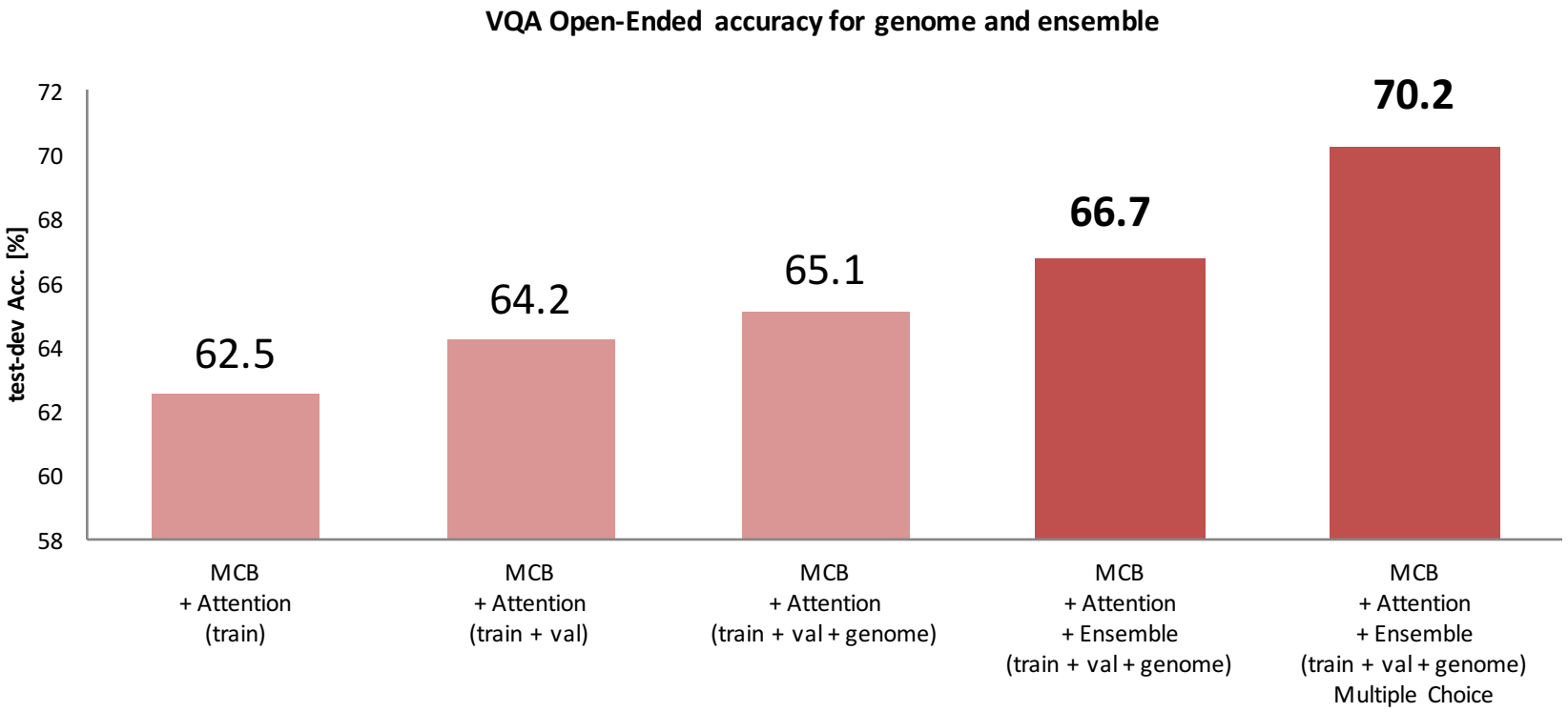
Results on MCB with Attention

- MCB performs well with Attention



Techniques to improve performance

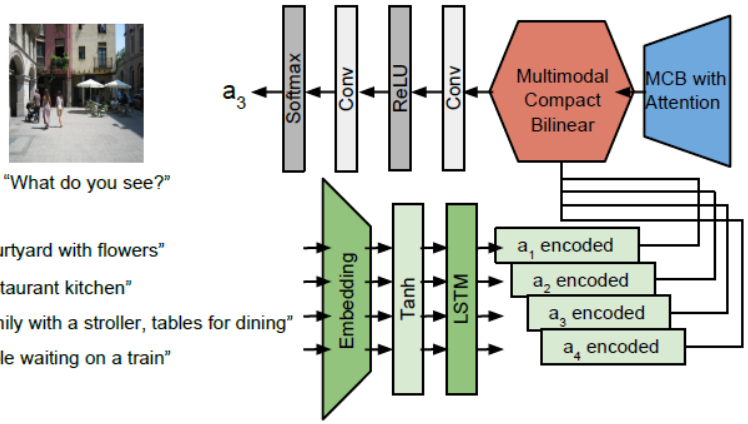
- Data Augmentation
  - VQA data from Visual Genome Dataset
    - Additional 1M Question and answer pairs
    - Removed articles, Single word answer
- Ensembles
  - Average the output of Softmax over models



Visual genome: Connecting language and vision using crowdsourced dense image annotations.

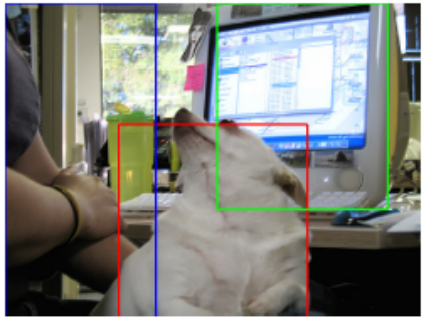
MCB on other Datasets and Tasks

- Visual 7w (Multiple Choice)



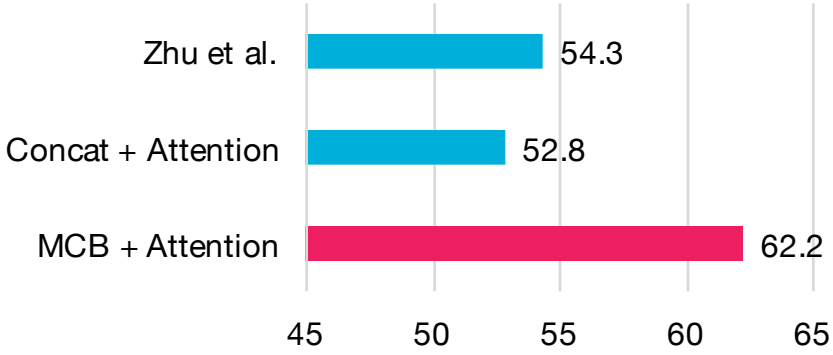
Our architecture for Visual 7w : MCB with Attention and Answer Encoding.

- Visual Grounding

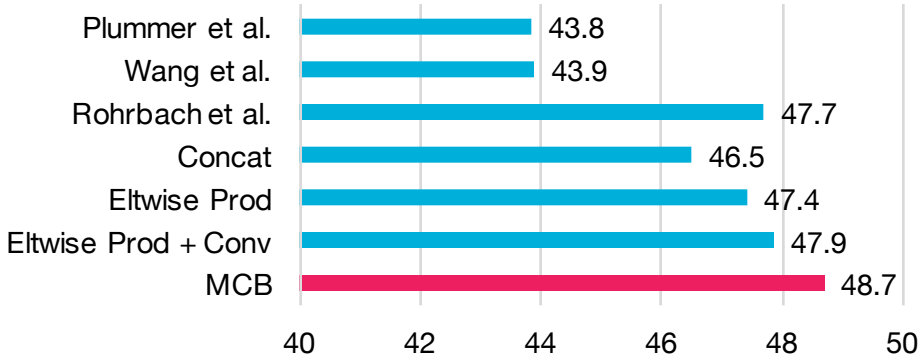


A dog distracts his owner from working at her computer.

Accuracy on Visual7W



Accuracy on Flickr30k Entities



Visual7W: Grounded Question Answering in Images

Grounding of textual phrases in images by reconstruction.

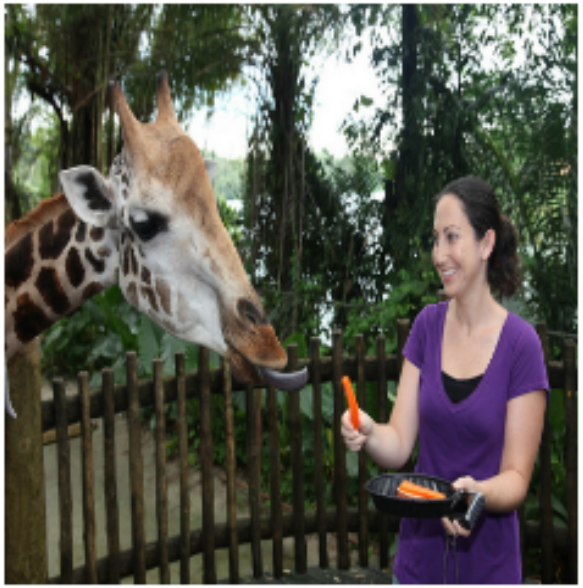
Examples for VQA

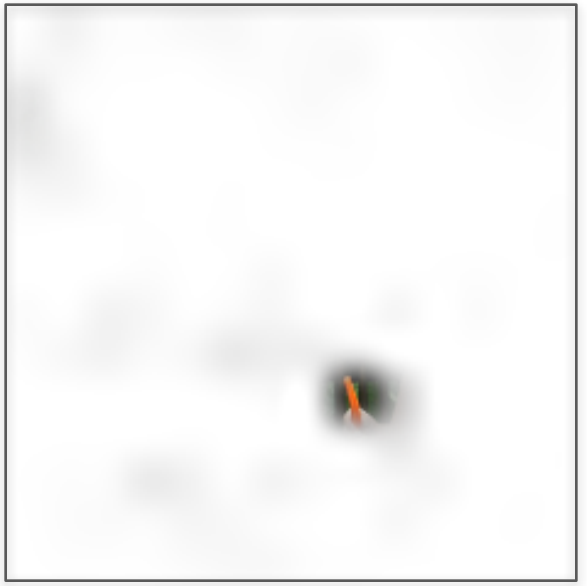
Attention Visualizations

What is the woman feeding the giraffe?

Carrot

[Groundtruth: Carrot]



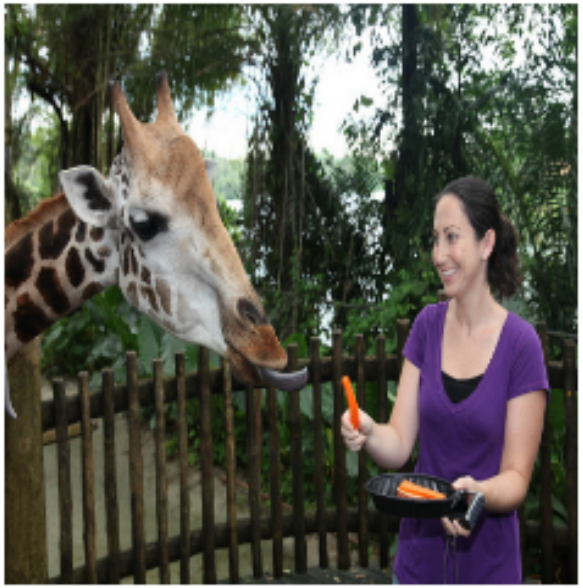


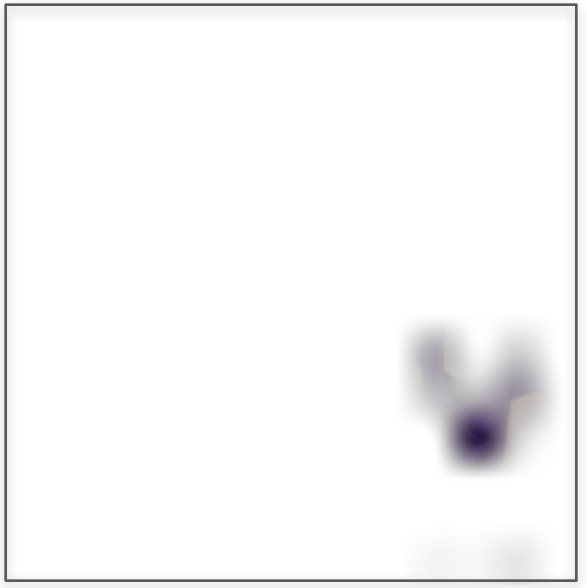
Attention Visualizations

What color is her shirt?

Purple

[Groundtruth: Purple]



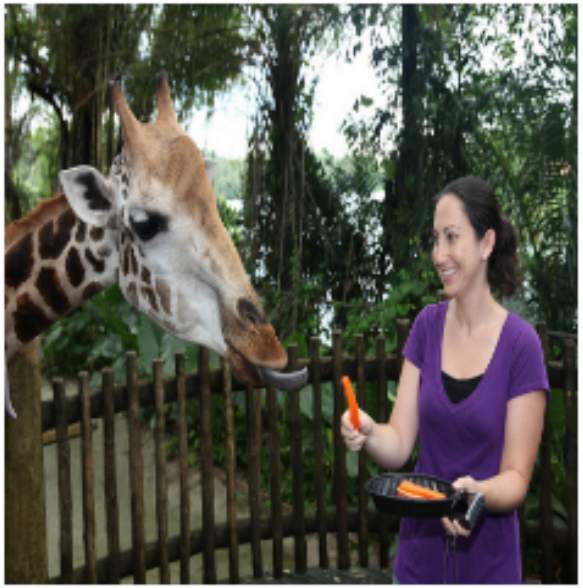


Attention Visualizations

What is her hairstyle for the picture?

Ponytail

[Groundtruth: Ponytail]



Attention Visualizations

What color is the chain on the red dress?

Pink

[Groundtruth: Gold]

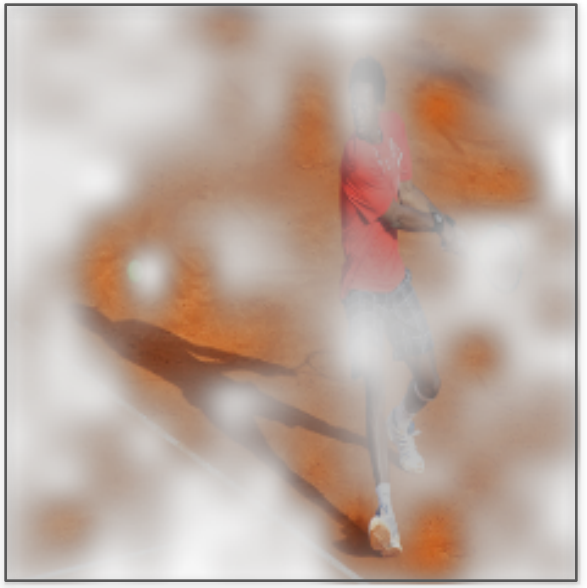
- Correct Attention, Incorrect Fine-grained Recognition

Attention Visualizations

Is the man going to fall down?

No

[Groundtruth: No]

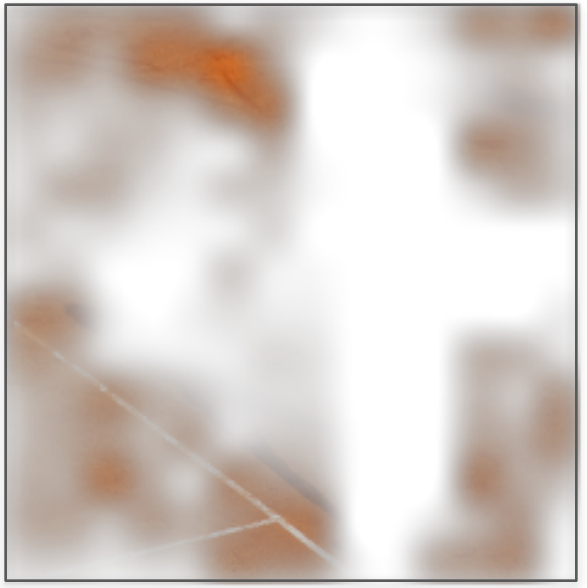


Attention Visualizations

What is the surface of the court made of?

Clay

[Groundtruth: Clay]

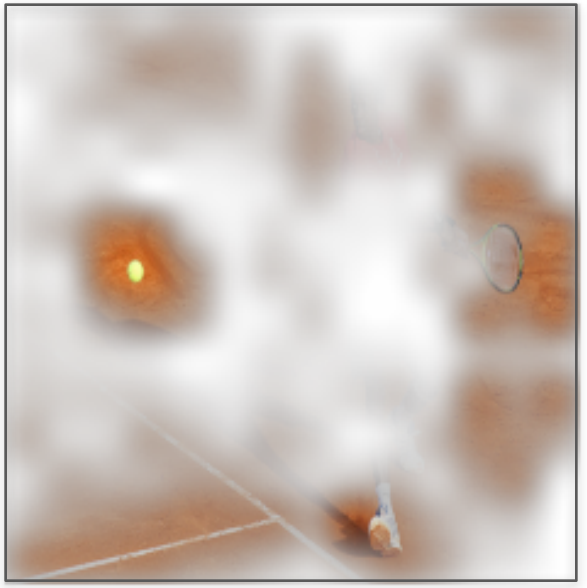


Attention Visualizations

What sport is being played?

Tennis

[Groundtruth: Tennis]



Attention Visualizations

What does the shop sell?

Clocks

[Groundtruth: Hot Dogs]

- Incorrect Attention

Attention Visualizations

What credit card company is on the banner in the background?

Budweiser

[Groundtruth: Mastercard]

- Correct Attention, Incorrect Concept Association

Conclusions

- Multimodal Compact Bilinear Pooling
  - All elements interact Multiplicatively
  - Compact and Efficient
- MCB with Attention
  - Successfully predict spatial attention
- Generalization Capability
  - Performance improvement in other vision and language tasks
    - Visual 7W, Visual Grounding
  - Compatible with other models
  - Applicable to general multimodal tasks, not only on vision and language

Thank you for your attention!

Demo : demo.berkeleyvision.org
Code: https://github.com/akirafukui/vqa-mcb/

Multimodal Compact Bilinear Pooling
for Visual Question Answering and Grounding

Akira Fukui, Dong Huk Park, Daylen Yang,
Anna Rohrbach, Trevor Darrell, Marcus Rohrbach

Arxiv 2016

Berkeley
UNIVERSITY OF CALIFORNIA