Multimodal Compact Bilinear Pooling for VQA

Akira Fukui\textsuperscript{1,2}, Dong Huk Park\textsuperscript{1}, Daylen Yang\textsuperscript{1},
Anna Rohrbach\textsuperscript{1,3}, Trevor Darrell\textsuperscript{1}, Marcus Rohrbach\textsuperscript{1}

\textsuperscript{1}UC Berkeley EECS, CA, United States
\textsuperscript{2}Sony Corp., Tokyo, Japan
\textsuperscript{3}Max Planck Institute for Informatics, Saarbrücken, Germany
A table full of food for a feast
Multimodal language and visual understanding

Grounding

The bowl with the brown souce
Visual Question Answering

What is the brown source?

Gravy
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

- All elements can interact
- Multiplicative interaction

Yes
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

- All elements can interact
- Multiplicative interaction
  - Difficult to learn output classification
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

Elementwise Multiplication

- All elements can interact
- Multiplicative interaction
- Difficult to learn input embedding
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

Outer Product / Bilinear Pooling [Lin ICCV 2015]

Yes

All elements can interact

Multiplicative interaction

[Bilinear CNN models for fine-grained visual recognition. ICCV 2015]

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji.
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

Outer Product / Bilinear Pooling

- spoon
- plate
- bowl
- table
- food
- corn
- ... (person)

2048

4 million

FC

4 million x 1000

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

Multimodal Compact Bilinear Pooling

Is this going to be a feast?

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


[Compact Bilinear Pooling] [Gao CVPR 16]

Yes

2048

2048

MCB

FC

16k x 1000

✔ All elements can interact
✔ Multiplicative interaction
✔ Low #activations & computation
✔ Low #parameters
Multimodal Compact Bilinear Pooling

Is this going to be a feast?

Pham & Pagh (2013):
\[ \Psi(x \otimes y) = \Psi(x) * \Psi(y) \]

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

Random Projection: Countsketch \( \Psi \)


Multimodal Compact Bilinear Pooling

Pham & Pagh (2013):
\[ \Psi(x \otimes y) = \Psi(x) \ast \Psi(y) \]

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

Multimodal Compact Bilinear Pooling

Is this going to be a feast?

Pham & Pagh (2013):
\[ \Psi(x \otimes y) = \Psi(x) \ast \Psi(y) \]

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

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**

**Diagram:**
- CNN (ResNet152) to L2 Normalization
- LSTM, drop
- L2 Normalization
- MCB
- 16k -> Signed Sqrt
- 16k -> L2 Normalization
- 16k -> Full Connected
- 3000 -> Softmax
- Yes

**Questions:**
- Is this going to be a feast?

- Embed/Tanh
- LSTM, drop
- LSTM, drop
- 13k ~ 20k
- 300
- 1024
- 2048
- 2048
Ablation Comparison to other multimodal methods

• MCB achieves highest accuracy

Trained on train, test-dev Acc. [%]
### Ablation Comparison to other multimodal methods

- **MCB comparable to Full Bilinear**

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Trained on train, test-dev Acc. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eltwise Sum</td>
<td>56.5</td>
</tr>
<tr>
<td>Concat</td>
<td>57.5</td>
</tr>
<tr>
<td>Concat + FC</td>
<td>58.4</td>
</tr>
<tr>
<td>Concat + FC + FC</td>
<td>57.1</td>
</tr>
<tr>
<td>Eltwise Product</td>
<td>58.6</td>
</tr>
<tr>
<td>Eltwise Product + FC</td>
<td>56.4</td>
</tr>
<tr>
<td>Eltwise Product + FC + FC</td>
<td>57.8</td>
</tr>
<tr>
<td>MCB (2048x2048 -&gt; 16k)</td>
<td>59.8</td>
</tr>
<tr>
<td>Full Bilinear (128x128 -&gt; 16k)</td>
<td>58.5</td>
</tr>
<tr>
<td>MCB (128x128 -&gt; 4k)</td>
<td>58.7</td>
</tr>
</tbody>
</table>
Dimensionality of MCB

- Dimensionality of MCB decides the performance of outer product approximation

<table>
<thead>
<tr>
<th>Dim size</th>
<th>test-dev Acc. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>58.4</td>
</tr>
<tr>
<td>2048</td>
<td>58.8</td>
</tr>
<tr>
<td>4096</td>
<td>59.4</td>
</tr>
<tr>
<td>8192</td>
<td>59.7</td>
</tr>
<tr>
<td>16000</td>
<td>59.8</td>
</tr>
<tr>
<td>32000</td>
<td>59.7</td>
</tr>
</tbody>
</table>

VQA Open-Ended test-dev accuracy
Visual Question Answering

What is the brown sauce?

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

Performance of Attention Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Test-dev Acc. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concat + FC</td>
<td>58.4</td>
</tr>
<tr>
<td>MCB</td>
<td>59.8</td>
</tr>
<tr>
<td>Concat + FC + Attention</td>
<td>58.4</td>
</tr>
<tr>
<td>MCB + Attention</td>
<td>62.5</td>
</tr>
</tbody>
</table>
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

VQA Open-Ended accuracy for genome and ensemble

Visual genome: Connecting language and vision using crowdsourced dense image annotations.
MCB on other Datasets and Tasks

- **Visual 7w (Multiple Choice)**
  - Visual 7W: Grounded Question Answering in Images
  - Accuracy on Visual7W
    - Zhu et al.: 54.3
    - Concat + Attention: 52.8
    - MCB + Attention: 62.2

- **Visual Grounding**
  - Grounding of textual phrases in images by reconstruction.
  - Accuracy on Flickr30k Entities
    - Plummer et al.: 43.8
    - Wang et al.: 43.9
    - Rohrbach et al.: 47.7
    - Concat: 46.5
    - Eltwise Prod: 47.4
    - Eltwise Prod + Conv: 47.9
    - MCB: 48.7

Our architecture for Visual 7w: MCB with Attention and Answer Encoding.
Examples for VQA
What is the woman feeding the giraffe?

Carrot

[Groundtruth: Carrot]
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
Is the man going to fall down?
No
[Groundtruth: No]
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