Multimodal Compact Bilinear Pooling
for Visual Question Answering and Visual Grounding
Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding

Akira Fukui*₁,₂   Dong Huk Park*₁   Daylen Yang*₁
Anna Rohrbach*₁,₃   Trevor Darrell₁   Marcus Rohrbach₁

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

Yang Gao\textsuperscript{1}, Oscar Beijbom\textsuperscript{1}, Ning Zhang\textsuperscript{2*}, Trevor Darrell\textsuperscript{1 \dagger}
\textsuperscript{1}EECS, UC Berkeley \quad \textsuperscript{2}Snapchat Inc.
{yg, obeijbom, trevor}@eecs.berkeley.edu \quad \textsuperscript{\textbullet} ning.zhang}@snapchat.com
• Background

• Multimodal Pooling

• Compact Bilinear Pooling

• Visual QA

• Architecture

• Results
Background
Fourier Transform

\[ \sum_{n=1}^{5} n \times \cos(n \times \omega \times t), \quad \omega = 10 \times 2\pi \]

Discrete Fourier Analysis
Hi, Dr. Elizabeth?
Yeah, uh... I accidentally took
the Fourier transform of my cat...

Meow!
Fourier Transform

- Transforms a signal into the frequency domain
- FFT processes a signal of length $n$ in $O(n \log(n))$
- Ubiquitous in signal processing
- But we don’t talk about it much in deep learning!
Convolution Theorem

- FT of a convolution is pointwise product of FTs

\[ \mathcal{F}\{f * g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\} \]

\[ f * g = \mathcal{F}^{-1}\{\mathcal{F}\{f\} \cdot \mathcal{F}\{g\}\} \]

- For two signals of length n, convolve in O(n log n)
- Speed-up from O(n^2) for equal-sized inputs
Count Sketch

• Fast, data structure related to Bloom Filter
• Function $\Psi(v)$ of $n$-dim input vectors
• Uses $d$-dim hashes, where $d \ll n$
Visual Question Answering (VQA)

What color are her eyes?  
What is the mustache made of?

How many slices of pizza are there?  
Is this a vegetarian pizza?

Is this person expecting company?  
What is just under the tree?

Does it appear to be rainy?  
Does this person have 20/20 vision?
• Background

• Multimodal Pooling

• Compact Bilinear Pooling

• Visual QA

• Architecture

• Results
Multimodal Pooling

- Image modality handled by CNN
- Language modality handled by RNN
- How do we merge/combine their outputs?
Is this going to be a feast?
Is this going to be a feast?

- All elements can interact
- Multiplicative interaction
  - Difficult to learn input embedding

Elementwise Multiplication

CNN

LSTM

spoon
plate
bowl
table
food
corn
... 

person

Yes
Is this going to be a feast?

All elements can interact
- Multiplicative interaction
  - Difficult to learn output classification
Is this going to be a feast?

Outer Product / Bilinear Pooling [Lin ICCV 2015]

All elements can interact
Multiplicative interaction

Bilinear CNN models for fine-grained visual recognition. ICCV 2015
Multimodal Pooling

- **Pointwise Addition**: Not expressive
- **Pointwise Multiplication**: Not expressive
- **Concatenation**: Interaction is not multiplicative
- **Outer Product**: Works great!
Is this going to be a feast?

Outer Product / Bilinear Pooling

- spoon
- plate
- bowl
- table
- food
- corn
- person

2048

4 million

Yes

4 million x 1000

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

Multimodal Pooling

- **Problem**: Outer Product is way too slow!
- **Solution**: Compact Bilinear Pooling
• Background
• Multimodal Pooling
• Compact Bilinear Pooling
• Visual QA
• Architecture
• Results
Is this going to be a feast?

Outer Product / Bilinear Pooling


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

4 million x 1000
Is this going to be a feast?

- 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

Compact bilinear pooling. CVPR 2016
Compact Bilinear Pooling

- Outer Product is what we want:

\[ B(\mathcal{X}) = \sum_{s \in S} x_s x_s^T \]  \hspace{1cm} (1)

- How do we approximate it?
We thus need a method that projects the outer product to a lower dimensional space and also avoids computing the outer product directly. As suggested by Gao et al. (2016) for a single modality, we rely on the Count Sketch projection function \( \Psi \) (Charikar et al., 2002), which projects a vector \( v \in \mathbb{R}^n \) to \( y \in \mathbb{R}^d \). We initialize two vectors \( s \in \{-1, 1\}^n \) and \( h \in \{1, \ldots, d\}^n \): \( s \) contains either 1 or \(-1\) for each index, and \( h \) maps each index \( i \) in the input \( v \) to an index \( j \) in the output \( y \). Both \( s \) and \( h \) are initialized randomly from a uniform distribution and remain constant for future invocations of count sketch. \( y \) is initialized as a zero vector. For every element \( v[i] \) its destination index \( j = h[i] \) is looked up using \( h \), and \( s[i] \cdot v[i] \) is added to \( y[j] \). See lines 1-9 and 12-16 in Algorithm 1.
Count Sketch

- Fast, data structure related to Bloom Filter
- Function $\Psi(v)$ of n-dim input vectors
- Uses $d$-dim hashes, where $d << n$
Count Sketch

- Cool Property:

\[ \Psi(x \otimes q, h, s) = \Psi(x, h, s) \ast \Psi(q, h, s) \]

- Sketch of outer product is the convolution of sketches

- We know how to efficiently compute convolution!
Compact Bilinear Pooling

- Input: Two vectors each of length \( n \)
- Count Sketch to reduce dimensionality to \( d \)
- Convolve count sketches in \( O(d \log(d)) \) using FFT
- Has the same properties as outer product
Pham & Pagh (2013):
\[ \Psi(x \otimes y) = \Psi(x) \ast \Psi(y) \]

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

References:
Algorithm 1 Multimodal Compact Bilinear

1: input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
2: output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
3: procedure MCB($v_1, v_2, n_1, n_2, d$)
4: for $k \leftarrow 1 \ldots 2$
5: if $h_k, s_k$ not initialized then
6: for $i \leftarrow 1 \ldots n_k$
7: sample $h_k[i]$ from $\{1, \ldots, d\}$
8: sample $s_k[i]$ from $\{-1, 1\}$
9: $v'_k = \Psi(v_k, h_k, s_k, n_k)$
10: $\Phi = \text{FFT}^{-1}(\text{FFT}(v'_1) \odot \text{FFT}(v'_2))$
11: return $\Phi$
12: procedure $\Psi(v, h, s, n)$
13: $y = [0, \ldots, 0]$
14: for $i \leftarrow 1 \ldots n$
15: $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
16: return $y$
Figure 2: Multimodal Compact Bilinear Pooling (MCB)
• Background
• Multimodal Pooling
• Compact Bilinear Pooling
• Visual QA
• Architecture
• Results
Visual Question Answering

- Images with text questions and text answers
- Several related datasets:
  - Visual QA
  - Visual Genome
  - Visual7W
Visual Question Answering

**What is the brown sauce?**

Gravy
Visual Grounding

- Image, text description, set of region proposals
- Datasets:
  - Flickr30k Entities
  - ReferItGame
Grounding

The bowl with the brown sauce
Visual Question Answering (VQA)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?
Dataset

- >250K images
  - 200K from MS COCO
    - 80 train / 40 val / 80 test
  - 50K from Abstract
- QAs
  - 3 questions/image
  - 10 answers/question
    - +3 answers/question without showing the image
- >760K questions
- ~10M answers
  - will grow over the years

Stump a smart robot!
Ask a question that a human can answer, but a smart robot probably can’t!

- Mechanical Turk
- >10,000 Turkers
- >41,000 Human Hours

4.7 Human Years!
20.61 Person-Job-Years!

Slide credits: Dhruv Batra
Answers

Does this man have children?
- yes
- yes
- yes

Is this man crying?
- no
- no
- no

Has the pizza been baked?
- yes
- yes
- yes

What kind of cheese is topped on this pizza?
- feta
- feta
- ricotta

How many pickles are on the plate?
- 1
- 1
- 1

What is the shape of the plate?
- circle
- round
- round

How many glasses are on the table?
- 3
- 3
- 2

What is the woman reaching for?
- door handle
- glass
- wine
- fruit
- glass
- remote

Do you think the boy on the ground has broken legs?
- yes
- yes
- no

Why is the boy on the right freaking out?
- his friend is hurt
- other boy fell down
- someone fell

Are the kids in the room the grandchildren of the adults?
- probably
- yes
- yes

What is on the bookshelf?
- nothing
- nothing
- books

Slide credits: Dhruv Batra
Visual Genome

- **object detection**: Q: How many people are wearing a lettered, zip-up red jacket? A: Just one.
- **object attributes**: Q: What is the most valuable device in this room? A: The television.
- **object classification**: Q: What animal is the balloon modeled after? A: Blue whale.
- **scene classification**: Q: Where was the picture taken? A: At the beach.
- **fine-grained recognition**: Q: What kind of boat is the far left blue boat? A: Sail boat.
- **action recognition**: Q: What is the snowboarder doing? A: Jumping.

- **text detection**: Q: When was the bridge built? A: 1932.
- **spatial reasoning**: Q: Where is the American flag? A: Behind president Reagan.
- **event understanding**: Q: What holiday is being celebrated? A: Fourth of July.
- **common sense**: Q: Why is the man’s tie moving? A: The wind is blowing.
- **person identification**: Q: Who is this man? A: Derek Jeter.
- **facial expressions**: Q: What expression is on most people’s faces? A: They are smiling.
Visual7W

Where does this scene take place?
A) In the sea. ✓
B) In the desert.
C) In the forest.
D) On a lawn.

What is the dog doing?
A) Surfing. ✓
B) Sleeping.
C) Running.
D) Eating.

Why is there foam?
A) Because of a wave. ✓
B) Because of a boat.
C) Because of a fire.
D) Because of a leak.

What is the dog standing on?
A) On a surfboard. ✓
B) On a table.
C) On a garage.
D) On a ball.

Which paw is lifted?
Visual7W

Q: What endangered animal is featured on the truck?
A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?
A: Onto 24 ¾ Rd.
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?
A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?
A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

Q: Why was the hand of the woman over the left shoulder of the man?
A: They were together and engaging in affection.
A: The woman was trying to get the man’s attention.
A: The woman was trying to scare the man.
A: The woman was holding on to the man for balance.

Q: How many magnets are on the bottom of the fridge?
A: 5.
A: 2.
A: 3.
A: 4.

Q: Which pillow is farther from the window?

Q: Which step leads to the tub?

Q: Which is the small computer in the corner?

Q: Which item is used to cut items?

Q: Which doughnut has multicolored sprinkles?

Q: Which man is wearing the red tie?
Problem Formulation

• Select the most probable answer

\[ \hat{a} = \arg\max_{a \in A} p(a | \mathbf{x}, \mathbf{q}; \theta) \]  

(1)

• Softmax Classification among possible answers
Datasets

- Explain Visual QA and Visual Grounding
• Background
• Multimodal Pooling
• Compact Bilinear Pooling
• Visual QA
• Architecture
• Results
Visual QA Architecture

• Pre-trained CNN processes input image
• LSTM processes input questions
• Multimodal Compact Bilinear layer combines them
• Softmax output among possible answers
Visual QA Architecture
Attention. To incorporate spatial information, we use soft attention on our MCB pooling method. Explored by (Xu et al., 2015) for image captioning and by (Xu and Saenko, 2016) and (Yang et al., 2015) for VQA, the soft attention mechanism can be easily integrated in our model.
Visual QA Architecture
Grounding Architecture

Q: “Person in blue checkered shirt”

Figure 5: Our Architecture for Grounding with MCB
Architecture Details

• Resnet152 and GloVE pretrained models
• L2 normalization instead of batch norm
• Dataset filtered based on answer vocabulary
- Background
- Multimodal Pooling
- Bilinear Pooling
- Visual QA
- Architecture
- Results
What vegetable is the dog chewing on?  
**MCB:** carrot  
**GT:** carrot

What kind of dog is this?  
**MCB:** husky  
**GT:** husky

What kind of flooring does the room have?  
**MCB:** carpet  
**GT:** carpet

What color is the traffic light?  
**MCB:** green  
**GT:** green

Is this an urban area?  
**MCB:** yes  
**GT:** yes

Where are the buildings?  
**MCB:** in background  
**GT:** on left
<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element-wise Sum</td>
<td>56.50</td>
</tr>
<tr>
<td>Concatenation</td>
<td>57.49</td>
</tr>
<tr>
<td>Concatenation + FC</td>
<td>58.40</td>
</tr>
<tr>
<td>Concatenation + FC + FC</td>
<td>57.10</td>
</tr>
<tr>
<td>Element-wise Product</td>
<td>58.57</td>
</tr>
<tr>
<td>Element-wise Product + FC</td>
<td>56.44</td>
</tr>
<tr>
<td>Element-wise Product + FC + FC</td>
<td>57.88</td>
</tr>
<tr>
<td>MCB (2048 × 2048 → 16K)</td>
<td>59.83</td>
</tr>
<tr>
<td>Full Bilinear (128 × 128 → 16K)</td>
<td>58.46</td>
</tr>
<tr>
<td>MCB (128 × 128 → 4K)</td>
<td>58.69</td>
</tr>
<tr>
<td>Element-wise Product with VGG-19</td>
<td>55.97</td>
</tr>
<tr>
<td>MCB (d = 16K) with VGG-19</td>
<td>57.05</td>
</tr>
<tr>
<td>Concatenation + FC with Attention</td>
<td>58.36</td>
</tr>
<tr>
<td>MCB (d = 16K) with Attention</td>
<td>62.50</td>
</tr>
</tbody>
</table>

Table 1: Comparison of multimodal pooling methods. Models are trained on the VQA train split and tested on test-dev.
<table>
<thead>
<tr>
<th>Compact Bilinear $d$</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>58.38</td>
</tr>
<tr>
<td>2048</td>
<td>58.80</td>
</tr>
<tr>
<td>4096</td>
<td>59.42</td>
</tr>
<tr>
<td>8192</td>
<td>59.69</td>
</tr>
<tr>
<td>16000</td>
<td>59.83</td>
</tr>
<tr>
<td>32000</td>
<td>59.71</td>
</tr>
</tbody>
</table>

Table 2: Accuracies for different values of $d$, the dimension of the compact bilinear feature. Models are trained on the VQA train split and tested on test-dev. Details in Sec. 4.3.
<table>
<thead>
<tr>
<th>Method</th>
<th>What</th>
<th>Where</th>
<th>When</th>
<th>Who</th>
<th>Why</th>
<th>How</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu et al.</td>
<td>51.5</td>
<td>57.0</td>
<td>75.0</td>
<td>59.5</td>
<td>55.5</td>
<td>49.8</td>
<td>54.3</td>
</tr>
<tr>
<td>Concat+Att.</td>
<td>47.8</td>
<td>56.9</td>
<td>74.1</td>
<td>62.3</td>
<td>52.7</td>
<td>51.2</td>
<td>52.8</td>
</tr>
<tr>
<td>MCB+Att.</td>
<td><strong>60.3</strong></td>
<td><strong>70.4</strong></td>
<td><strong>79.5</strong></td>
<td><strong>69.2</strong></td>
<td><strong>58.2</strong></td>
<td>51.1</td>
<td><strong>62.2</strong></td>
</tr>
</tbody>
</table>

Table 3: Multiple-choice QA tasks accuracy (%) on Visual7W test set.
<table>
<thead>
<tr>
<th></th>
<th>Test-dev</th>
<th></th>
<th></th>
<th>Test-standard</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Open Ended</td>
<td></td>
<td></td>
<td>All</td>
<td>MC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Y/N</td>
<td>No.</td>
<td>Other</td>
<td>All</td>
<td>Y/N</td>
<td>No.</td>
</tr>
<tr>
<td>MCB</td>
<td>81.2</td>
<td>35.1</td>
<td>49.3</td>
<td>60.8</td>
<td>65.4</td>
<td>-</td>
</tr>
<tr>
<td>MCB + Genome</td>
<td>81.7</td>
<td>36.6</td>
<td>51.5</td>
<td>62.3</td>
<td>66.4</td>
<td>-</td>
</tr>
<tr>
<td>MCB + Att.</td>
<td>82.2</td>
<td>37.7</td>
<td>54.8</td>
<td>64.2</td>
<td>68.6</td>
<td>-</td>
</tr>
<tr>
<td>MCB + Att. + GloVe</td>
<td>82.5</td>
<td>37.6</td>
<td>55.6</td>
<td>64.7</td>
<td>69.1</td>
<td>-</td>
</tr>
<tr>
<td>MCB + Att. + Genome</td>
<td>81.7</td>
<td>38.2</td>
<td>57.0</td>
<td>65.1</td>
<td>69.5</td>
<td>-</td>
</tr>
<tr>
<td>MCB + Att. + GloVe + Genome</td>
<td>82.3</td>
<td>37.2</td>
<td>57.4</td>
<td>65.4</td>
<td>69.9</td>
<td>-</td>
</tr>
<tr>
<td>Ensemble of 7 Att. models</td>
<td>83.4</td>
<td>39.8</td>
<td>58.5</td>
<td>66.7</td>
<td>70.2</td>
<td>83.2</td>
</tr>
<tr>
<td>Naver Labs (challenge 2nd)</td>
<td>83.5</td>
<td>39.8</td>
<td>54.8</td>
<td>64.9</td>
<td>69.4</td>
<td>83.3</td>
</tr>
<tr>
<td>HieCoAtt (Lu et al., 2016)</td>
<td>79.7</td>
<td>38.7</td>
<td>51.7</td>
<td>61.8</td>
<td>65.8</td>
<td>-</td>
</tr>
<tr>
<td>DMN+ (Xiong et al., 2016)</td>
<td>80.5</td>
<td>36.8</td>
<td>48.3</td>
<td>60.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FDA (Ilievski et al., 2016)</td>
<td>81.1</td>
<td>36.2</td>
<td>45.8</td>
<td>59.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D-NMN (Andreas et al., 2016a)</td>
<td>81.1</td>
<td>38.6</td>
<td>45.5</td>
<td>59.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AMA (Wu et al., 2016)</td>
<td>81.0</td>
<td>38.4</td>
<td>45.2</td>
<td>59.2</td>
<td>81.1</td>
<td>37.1</td>
</tr>
<tr>
<td>SAN (Yang et al., 2015)</td>
<td>79.3</td>
<td>36.6</td>
<td>46.1</td>
<td>58.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NMN (Andreas et al., 2016b)</td>
<td>81.2</td>
<td>38.0</td>
<td>44.0</td>
<td>58.6</td>
<td>81.2</td>
<td>37.7</td>
</tr>
<tr>
<td>AYN (Malinowski et al., 2016)</td>
<td>78.4</td>
<td>36.4</td>
<td>46.3</td>
<td>58.4</td>
<td>78.2</td>
<td>36.3</td>
</tr>
<tr>
<td>SMem (Xu and Saenko, 2016)</td>
<td>80.9</td>
<td>37.3</td>
<td>43.1</td>
<td>58.0</td>
<td>80.9</td>
<td>37.5</td>
</tr>
<tr>
<td>VQA team (Antol et al., 2015)</td>
<td>80.5</td>
<td>36.8</td>
<td>43.1</td>
<td>57.8</td>
<td>80.6</td>
<td>36.5</td>
</tr>
<tr>
<td>DPPnet (Noh et al., 2015)</td>
<td>80.7</td>
<td>37.2</td>
<td>41.7</td>
<td>57.2</td>
<td>80.3</td>
<td>36.9</td>
</tr>
<tr>
<td>iBOWIMG (Zhou et al., 2015)</td>
<td>76.5</td>
<td>35.0</td>
<td>42.6</td>
<td>55.7</td>
<td>76.8</td>
<td>35.0</td>
</tr>
</tbody>
</table>

Table 4: Open-ended and multiple-choice (MC) results on VQA test set (trained on train+val set) compared with state-of-the-art: accuracy in %. See Sec. 4.4.
Results: Visual Grounding

- Input: Text question, image, object proposals
- Output: Score for each object proposal
A tattooed woman with a green dress and yellow back-pack holding a water bottle is walking across the street.

A dog distracts his owner from working at her computer.

Figure 6: Top: predicted answers and attention maps from MCB model on VQA images. Bottom: predicted grounding from MCB model (left) and Eltwise Product + Conv model (right) on Flickr30k Entities images.
<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plummer et al. (2015)</td>
<td>27.42</td>
</tr>
<tr>
<td>Hu et al. (2016b)</td>
<td>27.80</td>
</tr>
<tr>
<td>Plummer et al. (2016)¹</td>
<td>43.84</td>
</tr>
<tr>
<td>Wang et al. (2016)</td>
<td>43.89</td>
</tr>
<tr>
<td>Rohrbach et al. (2016)</td>
<td>47.81</td>
</tr>
<tr>
<td>Concatenation</td>
<td>46.50</td>
</tr>
<tr>
<td>Element-wise Product</td>
<td>47.41</td>
</tr>
<tr>
<td>Element-wise Product + Conv</td>
<td>47.86</td>
</tr>
<tr>
<td>MCB</td>
<td><strong>48.69</strong></td>
</tr>
</tbody>
</table>

Table 5: Grounding accuracy on Flickr30k Entities dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu et al. (2016b)</td>
<td>17.93</td>
</tr>
<tr>
<td>Rohrbach et al. (2016)</td>
<td>26.93</td>
</tr>
<tr>
<td>Concatenation</td>
<td>25.48</td>
</tr>
<tr>
<td>Element-wise Product</td>
<td>27.80</td>
</tr>
<tr>
<td>Element-wise Product + Conv</td>
<td>27.98</td>
</tr>
<tr>
<td>MCB</td>
<td><strong>28.91</strong></td>
</tr>
</tbody>
</table>

Table 6: Grounding accuracy on ReferItGame dataset.
Summary

• Compact Bilinear Pooling is a useful tool
• It might get you that extra 1% accuracy
• Also useful: Attention, pretrained models
• Code is available on github
How many cats are there?

1 (0.97)
2 (0.03)
none (0.00)
3 (0.00)
0 (0.00)
Visual QA Demo

Interact with our state-of-the-art system for visual question answering.

How many dogs are there?

1 (0.97)
2 (0.02)
none (0.01)
0 (0.00)
3 (0.00)