Visual Question Answering

Xiaotong Li [a] & Yuchen Yuan [b]

[a]Department of Electrical Engineering, [b]Department of Computer Engineering
Jack Baskin School of Engineering
University of California, Santa Cruz

March 23, 2018

Outline

Motivation

- A complex task in Computer Vsion and Natrual Language Processing
- Different as Image Caption, VQA needs deeper comprehension on image

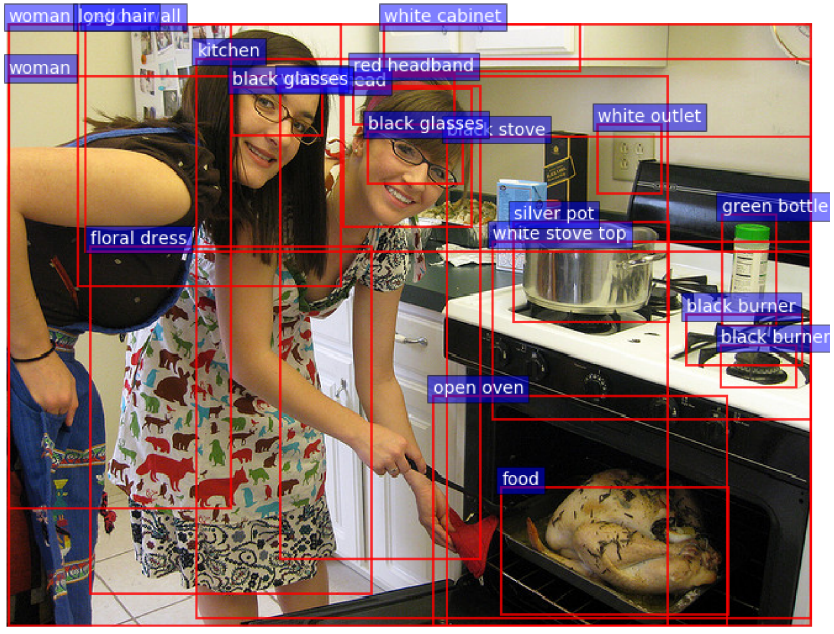


Figure 1: Example.

Can you park here?	no	no
	no	no
	no	yes
What color is the hydrant?	white and orange	red
	white and orange	red
	white and orange	yellow

Figure 2: VQA Example

Datasets

- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question

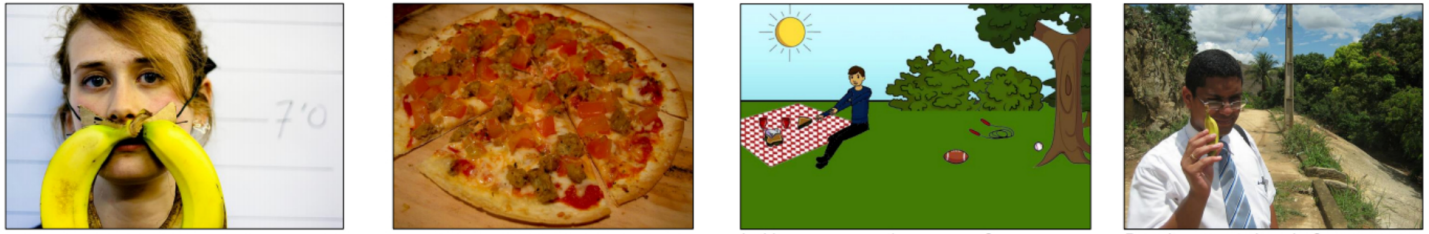


Figure 3: Dataset Example.

Feature Extraction

Image Feature

Different types of CNN:
VGG-16
ResNet-152
⋮

In the training process, the model is pre-trained CNN model. The reason is that the fine-tuning training of CNN is extremely time-consuming, and the fine-tuned results do not greatly improve the overall performance of the model.

Question Feature

LSTM
GRU
Bayesian network
⋮

Feature Fusion

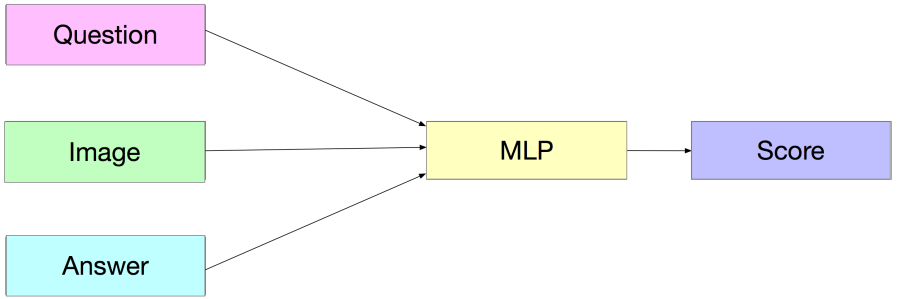


Figure 4: MLP model

Let x_i, x_q and x_a denote the image, question, and answer features, respectively. Image feature is a 2048-dim vector. Question feature is a 2048 dim vector.
Denoting the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$ and the concatenation operator $x_{iqa} = x_i \oplus x_q \oplus x_a$, we define the models as follows:

$$MLP : y = \sigma(\mathrm{W}_2 \max(0, \mathrm{W}_1 \mathrm{x}_{\mathrm{iqa}}) + \mathrm{b}). \quad (1)$$

Feature fusion - MLP

Each image-question-answer triple goes through the following MLP:

$$z_1 = W_1 x_{iqa} + b_1 \tag{2}$$

$$h_1 = max(0, z_1) \tag{3}$$

$$s = W_2 h_1 + b_2 \tag{4}$$

For each question, there are 3 incorrect answers and 1 correct answer, each of which is mapped to a score after going through the MLP. These four scores are then normalized using Softmax, and compared against ground-truth (boolean) labels to calculate cross-entropy loss.

MCB model

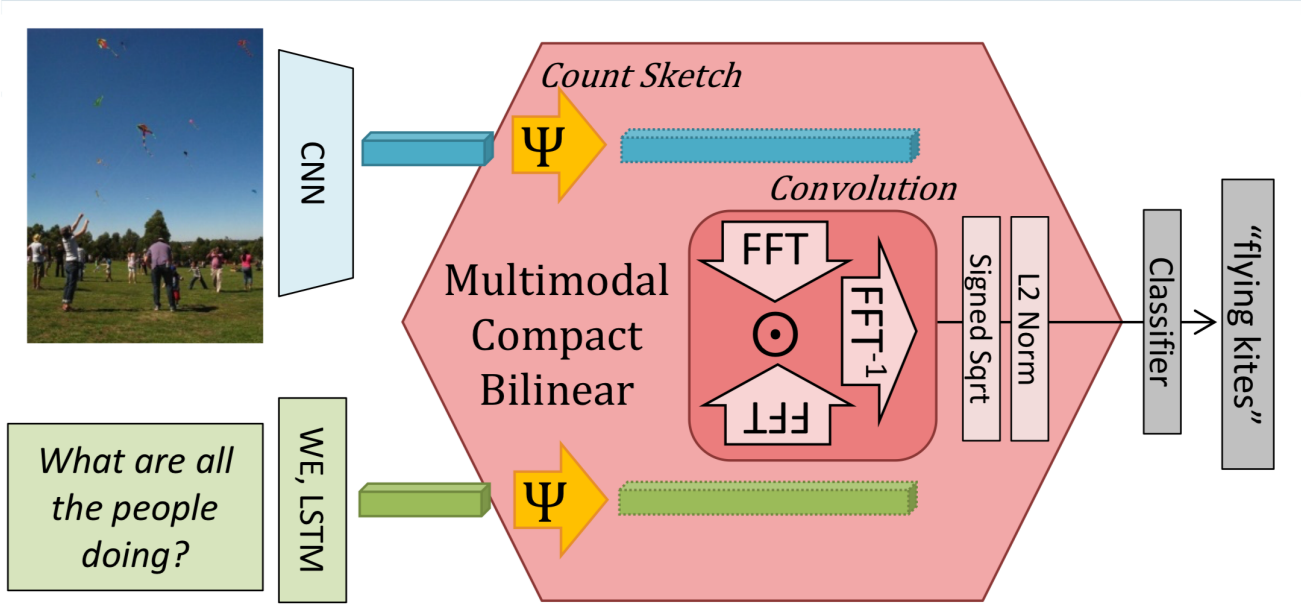


Figure 5: a sketch of MCB model

Extract the features of the Image and question respectively: use the pre-train CNN to extract the high-level features of the image, extract the word-embeding for the text, and decode the features through LSTM;

$$
\begin{aligned}
\langle B(X), B(Y)\rangle &= \langle \sum_{s\in S} x_s x_s^T , \sum_{u\in U} y_u y_u^T \rangle \\
&= \sum_{s\in S}\sum_{u\in U} \langle x_s x_s^T , y_u y_u^T \rangle \\
&= \sum_{s\in S}\sum_{u\in U} \langle x_s, y_u\rangle^2 \\
&\approx \sum_{s\in S}\sum_{u\in U} \langle \phi(x), \phi(y)\rangle
\end{aligned}
\tag{5}
$$

Algorithm 2 Tensor Sketch Projection

Input: $x \in \mathbb{R}^c$

Output: feature map $\phi_{TS}(x) \in \mathbb{R}^d$, such that $\langle \phi_{TS}(x), \phi_{TS}(y)\rangle \approx \langle x, y\rangle^2$

1. Generate random but fixed $h_k \in \mathbb{N}^c$ and $s_k \in \{+1, -1\}^c$ where $h_k(i)$ is uniformly drawn from $\{1, 2, \ldots, d\}$, $s_k(i)$ is uniformly drawn from $\{+1, -1\}$, and $k = 1, 2$.
2. Next, define sketch function $\Psi(x, h, s) = \{(Qx)_1, \ldots, (Qx)_d\}$, where $(Qx)_j = \sum_{t:h(t)=j} s(t)x_t$

3. Finally, define $\phi_{TS}(x) \equiv \text{FFT}^{-1}(\text{FFT}(\Psi(x, h_1, s_1)) \circ \text{FFT}(\Psi(x, h_2, s_2)))$, where the $\circ$ denotes element-wise multiplication.

Feature fusion - MCB

Using the two modal features obtained in the previous step, they are approximated (dimensionally reduced) using the Count Sketch method, respectively, to obtain the feature after dimension reduction;

$$\psi(x \otimes y, h, s) = \psi(x, h, s) \star \psi(y, h, s) \tag{6}$$

where ψ is the count sketch operator, x, y are the inputs, h, s are the hash tables, $\otimes$ defines outer product and $\star$ is the convolution operator. This can further be simplified by using FFT properties: convolution in time domain equals to elementwise product in frequency domain.

Accuracy Comparision

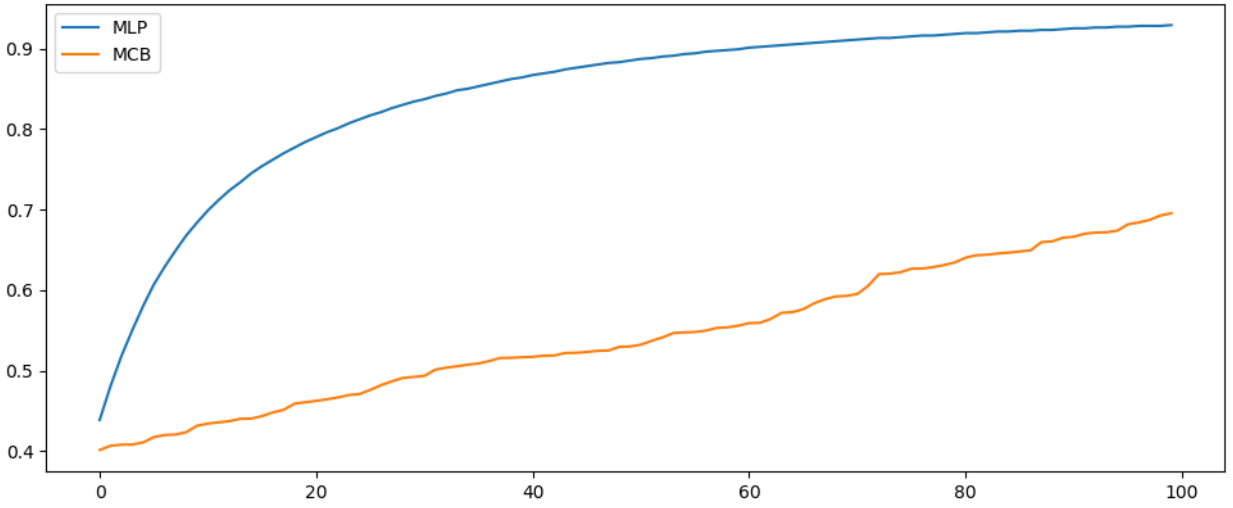
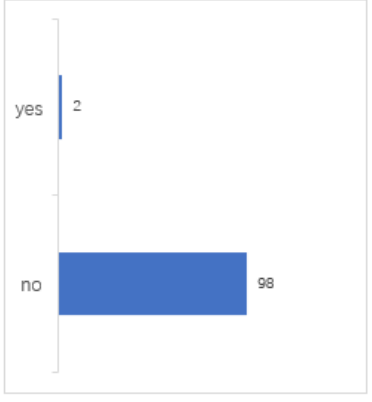


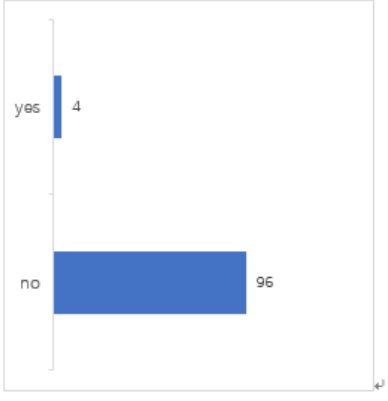
Figure 6: Accuracy Comparision

Results and Some Tricks

Are there any people at the train boarding station?



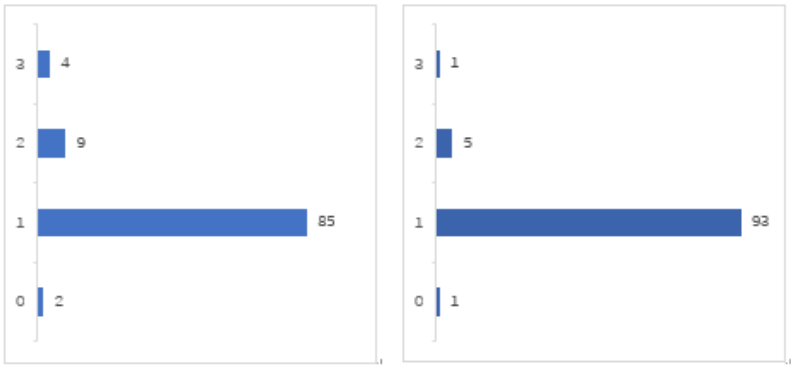
MCB



MLP

Results and Some Tricks

- How many wooden chairs are visible in this picture?



MLP MCB

Results and Some Tricks

Figure 7: MCB and MLP result comparision

Tricks

Attention

score the region embeddings according to the question vector, and compute a global visual vector as a sum pooling weighted by these scores.

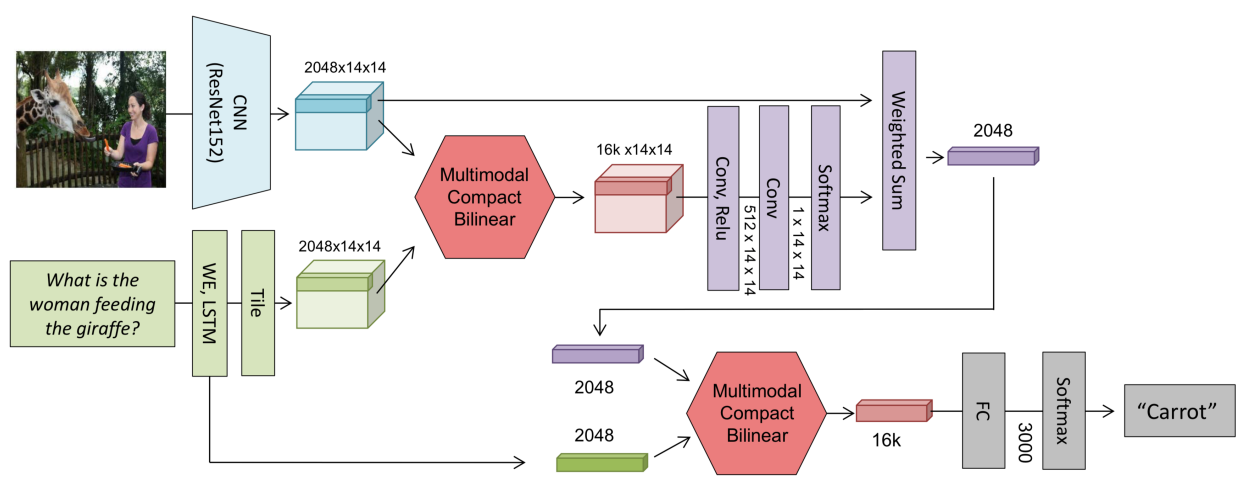


Figure 8: MCB model with attention

Visual Genomeo

Do an attention on object directly, instead of the attention of each block area in the picture.

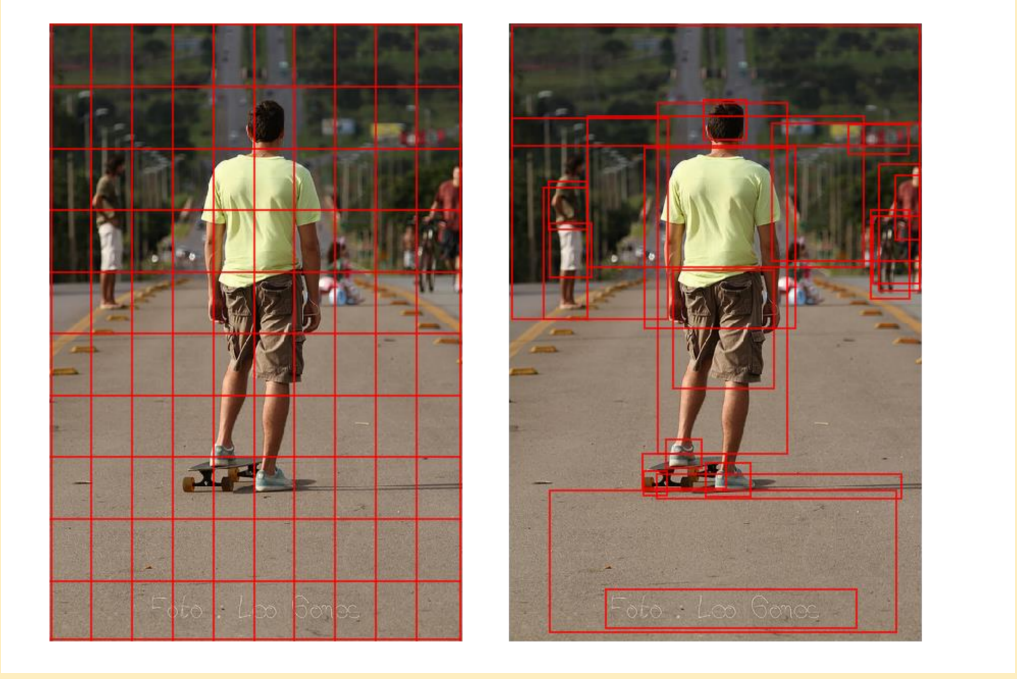


Figure 9: Image Attention

THANKS FOR LISTENING!

Refernece

[1]Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." *arXiv preprint arXiv*:1606.01847 (2016).

[2]Teney, Damien, et al. "Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge." *arXiv preprint arXiv*:1708.02711 (2017).

[3]Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.