Multimodal Compact Bilinear Pooling for Visual Question Answering

Authors: Akira Fukui [1] [2] Dong Huk Park[1] Daylen Yang [1] Anna Rohrbach [1] [3] Trevor Darrell [1] Marcus Rohrbach [1] Presenter: Yang Shi [4] [5]

[1]UC Berkeley EECS [2]Sony Corp. [3]Max Planck Institute for Informatics [4]UC Irvine EECS [5]Amazon Inc.

Keypoints

- Introduce VQA task and dataset
- Feature extraction for VQA
- Introduce Multimodal Compact Bilinear Pooling
- Introduce attention scheme

VQA Task

Visual question answering(VQA) is a task focusing on providing a natural language answer given any image and any open-ended question. This task requires a deep understanding and reasoning of the image combined with the question: a joint representation of both visual and textual input.

Figure 1: VQA task sample

VQA Dataset

VQA dataset v1: First released VQA dataset.
VQA dataset v2: Due to the inherent structure in our world and bias in language, the learning process is biased in dataset v1. VQA dataset v2 prepares similar images with same questions but leads to different answers to avoid the biasness.
GuessWhat Dataset: Guess a target in a given image with a sequential questions and answers. This requires both visual question answering and spatial reasoning.
TDIUC(The Task Driven Image Understanding Challenge dataset): (1)Categorized questions: Each question belongs to one of the 12 categories. (2)Absurd questions: Questions that are totally irrelevant to the image. In this way, it force an algorithm to determine if a question is valid. This also helps to balance the dataset.

Feature Extraction

Usually people use pretrained models to extract features from the image and question.
Image pretrained model:

- VGG: Use many convolutional blocks with relatively narrow kernels, followed by a max-pooling step and to repeat this block multiple times.
- Resnet: Reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions.

Question pretrained model:

- Word2Vec: Takes a text corpus as input and produces the word vectors as output.
- Glove: Similar to Word2Vec, it is a word embedding dataset.
- Skipthought: Reconstruct the surrounding sentences of an encoded passage. Different from the previous two model, this is a sentence based model.
- GNMT encoder: We propose using the encoder of google neural machine translation system to extract the question features.

Pooling Methods

After we extract the features, information of each modality are then combined together, using concatenation, element-wise product or sum operations.
Pooling methods are widely used in visual tasks to combine information for various streams into one final feature representation. It inherently can be applied to VQA task. Common pooling methods are average pooling and bilinear pooling. Bilinear pooling requires taking the outer product between two features. However the high dimensional result makes it hard to be applied to huge real image dataset. Thus, different compact pooling methods are proposed to solve this problem.

Multimodal Compact Bilinear Pooling(MCB) [1]

The key idea in MCB is to avoid directly computing the outer product and to reduce the output dimension by using count sketch.
Given a vector $a \in \mathcal{R}^n$, random hash function $h \in \mathcal{R}^n$: $[n] \rightarrow [b]$ and binary variable $s \in \mathcal{R}^n$: $[n] \rightarrow \pm 1$, the **count sketch** operator $\psi(a, h, s) \in \mathcal{R}^b$ is:

$$\psi(a, h, s)[j] = \Sigma_{h[i]=j} s[i]a[i], \quad j \in 1, \cdots, b$$

Let x and y be two separate feature vectors, and their bilinear pooling feature be $x \otimes y$, then:

$$\psi(x \otimes y, h, s) = \psi(x, h, s) \star \psi(y, h, s)$$

where $\star$ is the convolution operator. This can further be simplified by using FFT properties: convolution in time domain equals to elementwise product in frequency domain. We show the procedure below:

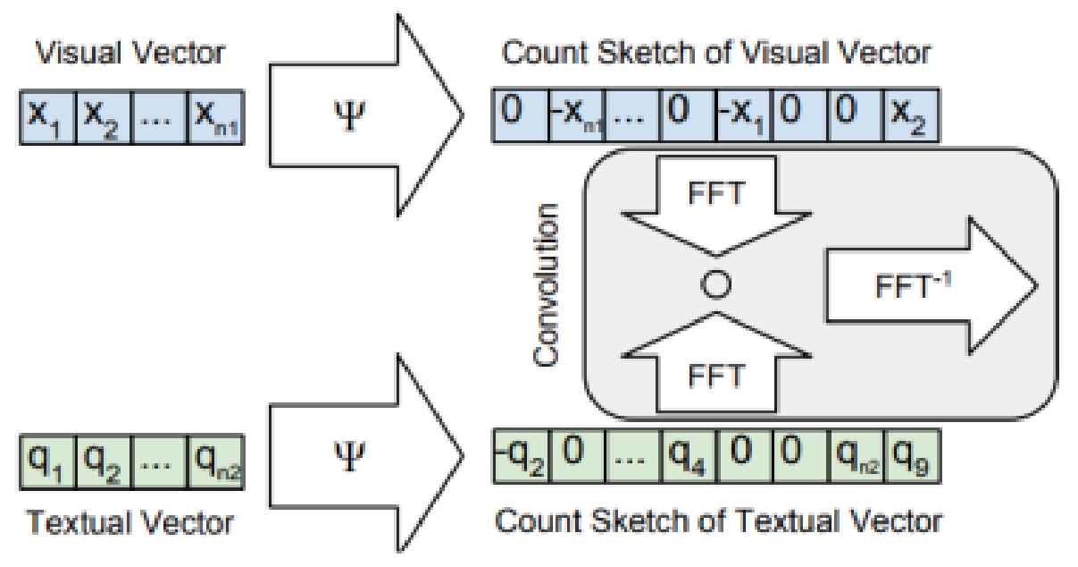
Visual Vector
x_1 x_2 ... x_{n1}
Ψ
Count Sketch of Visual Vector
0 $-x_{n1}$ ... 0 $-x_1$ 0 0 x_2
Convolution
FFT
FFT
FFT^{-1}
q_1 q_2 ... q_{n2}
Ψ
$-q_2$ 0 ... q_4 0 0 q_{n2} q_9
Textual Vector
Count Sketch of Textual Vector

Figure 2: Count Sketch(Ref: [1])

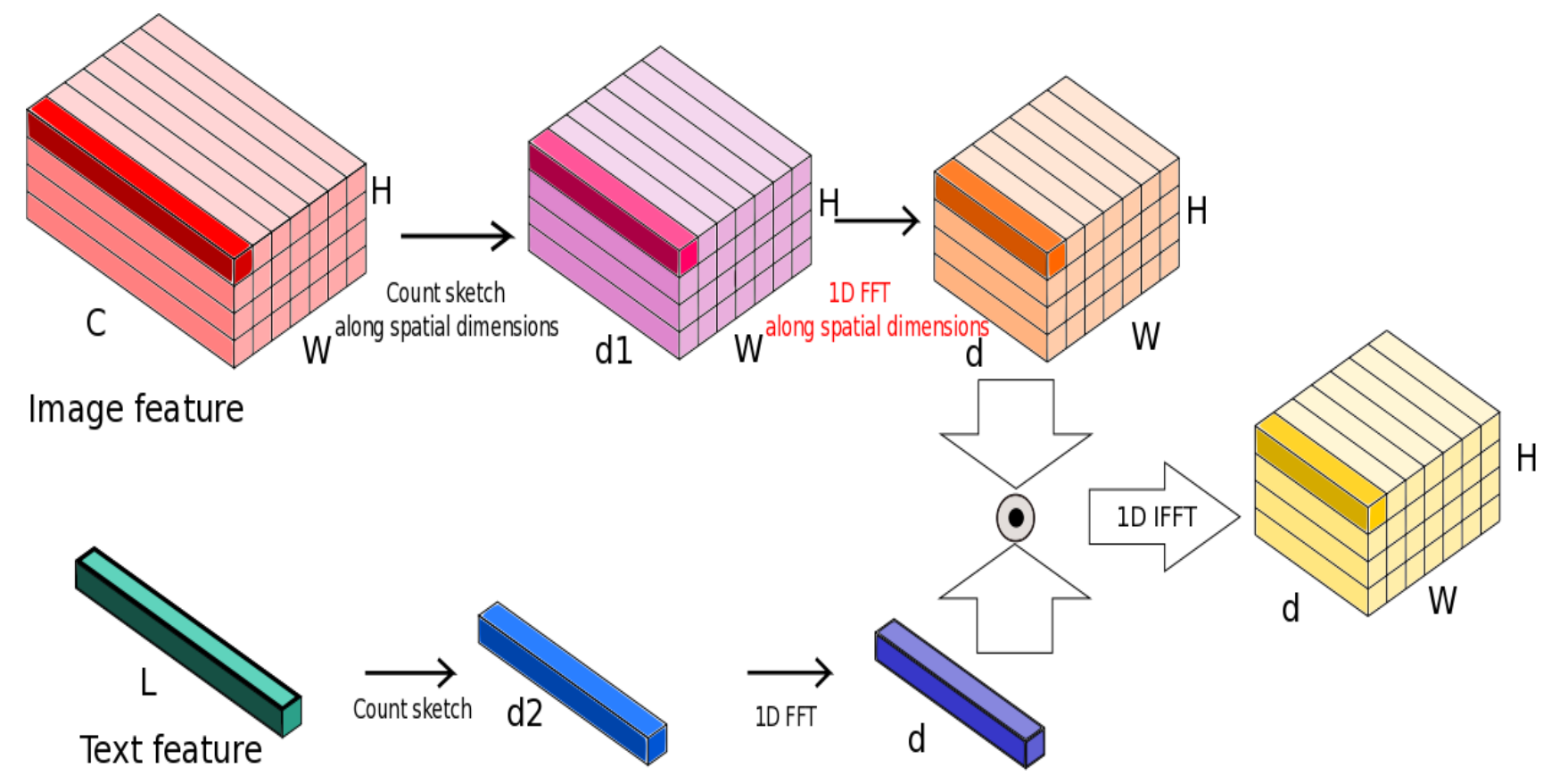
H
C
W
Image feature
Count sketch along spatial dimensions
d1
W
H
1D FFT along spatial dimensions
d
W
H
1D IFFT
H
d
W
L
Text feature
Count sketch
d2
1D FFT
d

Figure 3: MCB model

The sketch is done independently between the 1-dimensional text feature vector and each image tensor fiber.

Attention

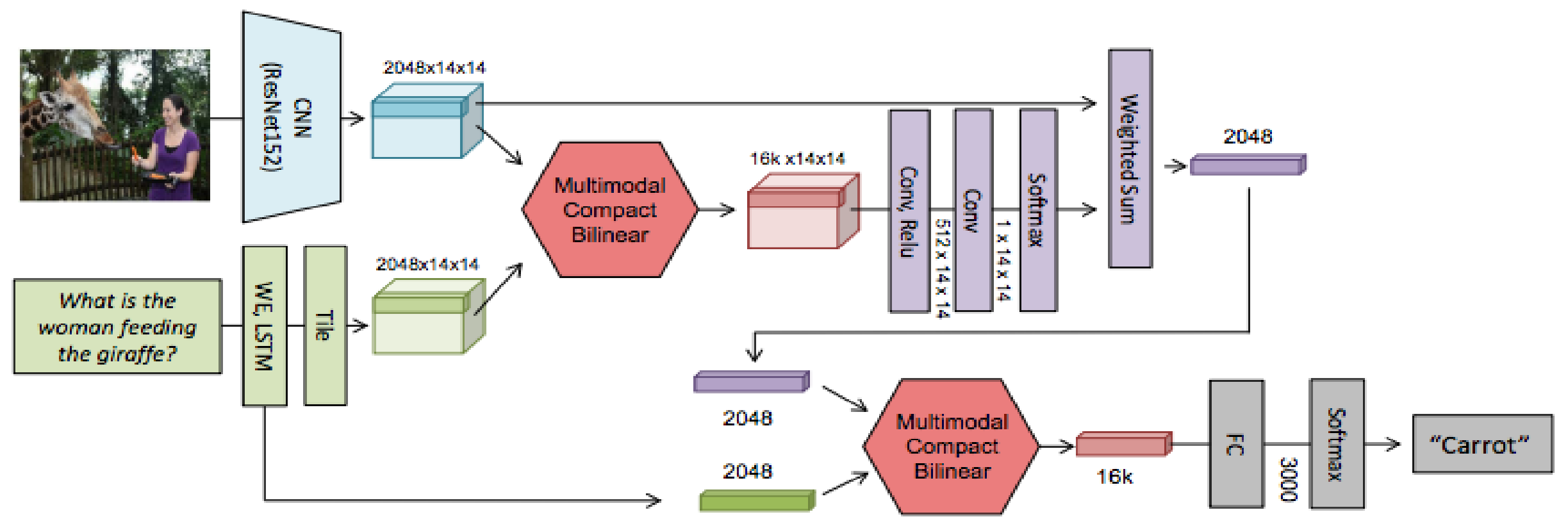

Figure 4: Multimodal compact tensor pooling (Ref: [1])

- Use MCB pooling to merge the slice of the visual feature with the language representation.
- Use two convolutional layers after the pooling to predict the attention weight for each grid location.
- Apply softmax to produce a normalized soft attention map.
- Take a weighted sum of the spatial vectors using the attention map to create the attended visual representation.

Experiemnt

Architecture	Y/N	No.	Other	All
MCB	81.2	35.1	49.3	60.8
MCB+Att.	82.2	37.7	54.8	64.2
MCB+Att.+GloVe+Genome	82.3	37.2	57.4	**65.4**
HieCoAtt	79.7	38.7	51.7	61.8
VQA team	80.5	36.8	43.1	57.8

Table 1: Results on VQA test-dev set(Ref: [1])

Reference

[1] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP 2016.*

Acknowledgement

We would like to thank the authors of the paper: Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell and Marcus Rohrbach.