

Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding

Akira Fukui et al., 2016

Mohammad Eshghi

January 31, 2020

Overview

1 Introduction

- Image encoding
- Text encoding
- Information fusion

2 Multimodal Compact Bilinear Pooling

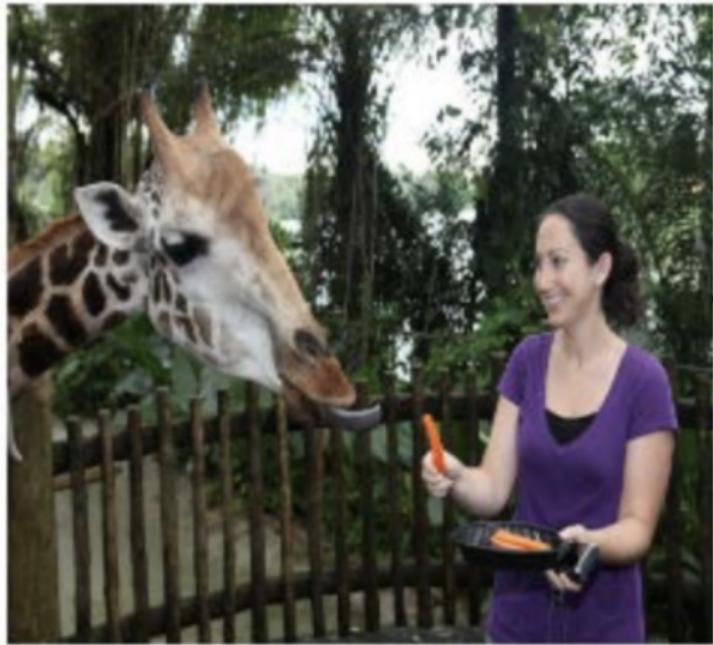
- Why MCB?
- What is MCB?
- How to MCB?
- Architecture of MCB

3 Experiments and Results

- Datasets
- VQA
- Visual Grounding

4 Conclusion

Introduction: VQA



Visual question answering: "What is the woman feeding the giraffe?" Correct answer: "Carrot"

Credit: [medium post](#)

Introduction: VG



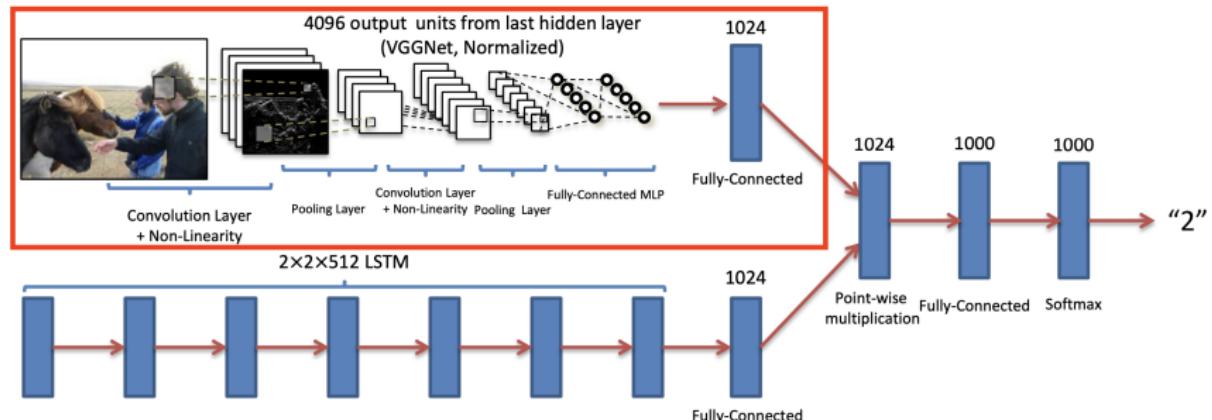
Visual grounding: "A tattooed woman (red) with a green dress (blue) and yellow backpack (green) holding a water bottle (pink) is walking across the street." The bounding boxes drawn would be excellent answers to this task.

Credit: [medium post](#)

Introduction

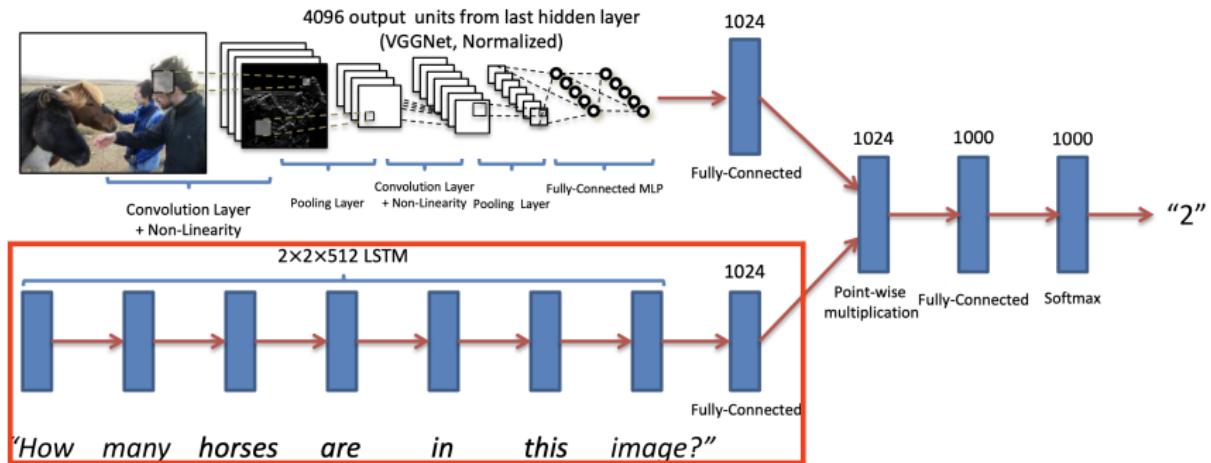
$$\hat{a} = \operatorname*{argmax}_{a \in A} p(a | \mathbf{x}, \mathbf{q}; \theta)$$

Image encoding



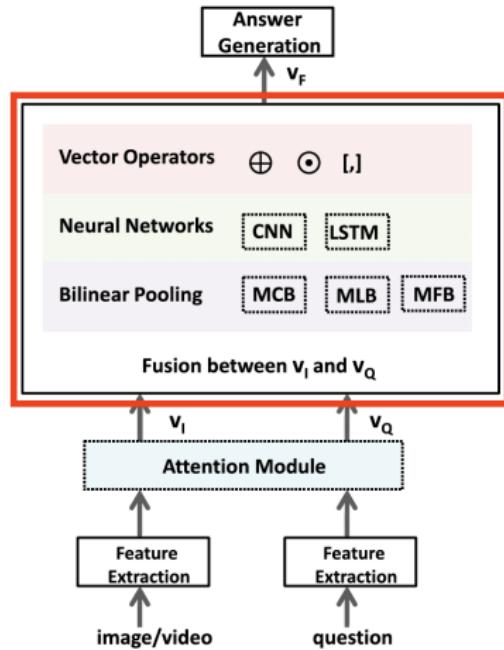
Credit: [Agrawal et al., 2019](#)

Text encoding



Credit: [Agrawal et al., 2019](#)

Information fusion



Credit: *Zhang et al., 2019*

Why MCB?

- Vector operations are too simplistic to fully capture the relationships between images and text
- So many different architectures that depend of the data set for non-linear neural networks to learn the interaction between the visual and textual features
- Research like [Zhang et al., 2019](#) confirm that Bilinear Pooling is the promising road to take for information fusion
- In bilinear pooling (aka a full outer product of vectors) the number of resulting parameters is too high to be practical. For example,
 - if the image and text vectors are of length 2048
 - the resulting matrix would have 2048^2 elements
 - and we'd need to fully connect that matrix to 3000 classes
 - it results in 12.5 billion learnable parameters
- Therefore, introducing MCB to capture the discriminating abilities of bilinear pooling with only a few thousand parameters (16k) while preserving backpropability

What is MCB?

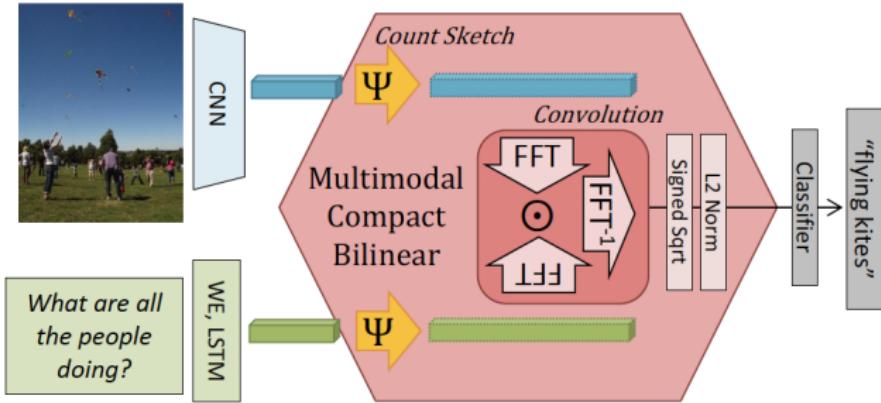
Bilinear Pooling simply means the outer product of the two input vectors; this is what happens when you multiply every element of one vector of length N by every element of the other vector of length M, resulting in a matrix of size NxM.

Compact bilinear pooling means the authors applied dimensionality reduction techniques to get almost the same level of power with way fewer parameters. **I think** it is something similar to principal components analysis.

Multimodal is a term that describes this specific use case but doesn't really mean anything different.

Compact Bilinear Pooling could be done on any two vectors, practically speaking it doesn't matter at all that one vector happens to represent one mode (images) while the other represents another (text).

How to MCB?

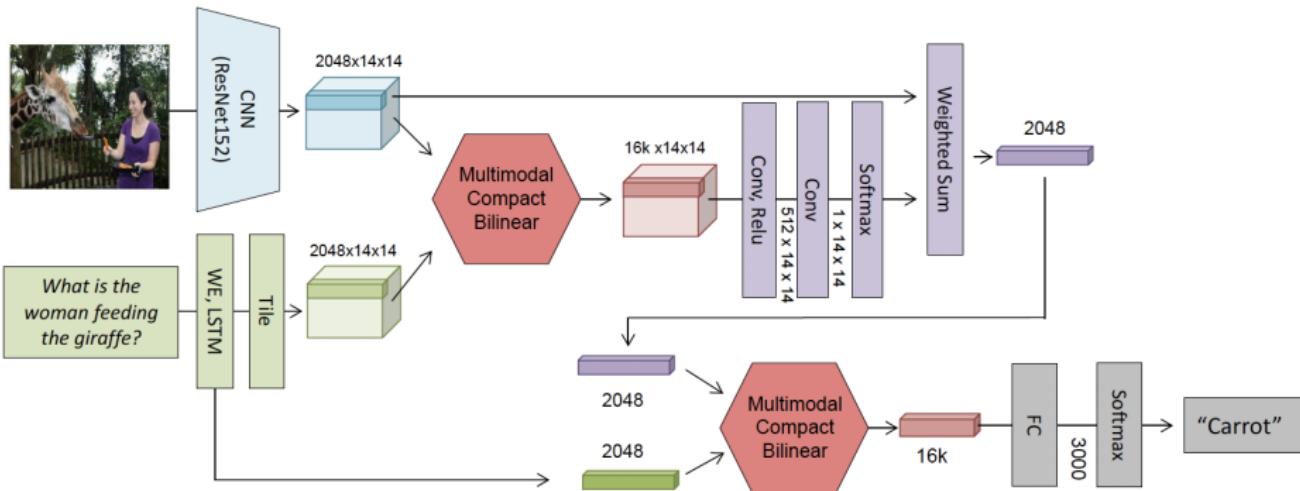


How to MCB?

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  $k \leftarrow 1 \dots 2$  do
5:     if  $h_k, s_k$  not initialized then
6:       for  $i \leftarrow 1 \dots n_k$  do
7:         sample  $h_k[i]$  from  $\{1, \dots, d\}$ 
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:    $y = [0, \dots, 0]$ 
14:   for  $i \leftarrow 1 \dots n$  do
15:      $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$ 
16:   return  $y$ 
```

How to MCB?



How to MCB?

MCB is used at three different points where visual and textual representations need to be combined

- ① to predict spatial attention in the image
- ② to predict an answer to the question
- ③ to relate the encoded multiple-choice answers to the combined question+image space (this only applies to the multiple choice problem)

They also use attention maps and additional training data, which feels a bit like sacrificing experimental purity to achieve state-of-the-art results.

Architecture of MCB



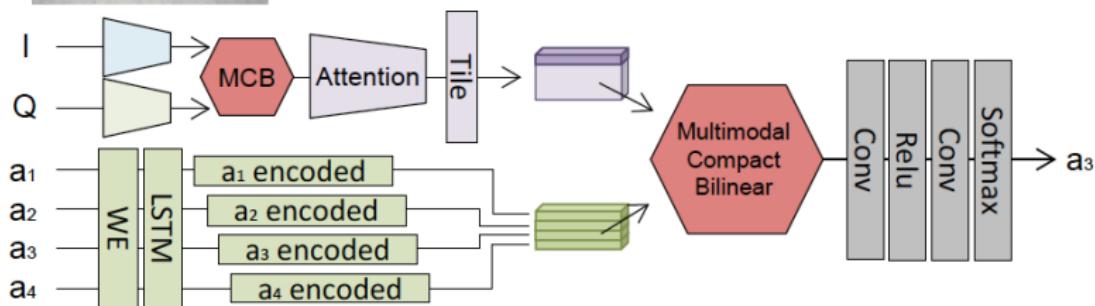
Q : "What do you see?" (Ground Truth : a_3)

a_1 : "A courtyard with flowers"

a_2 : "A restaurant kitchen"

a_3 : "A family with a stroller, tables for dining"

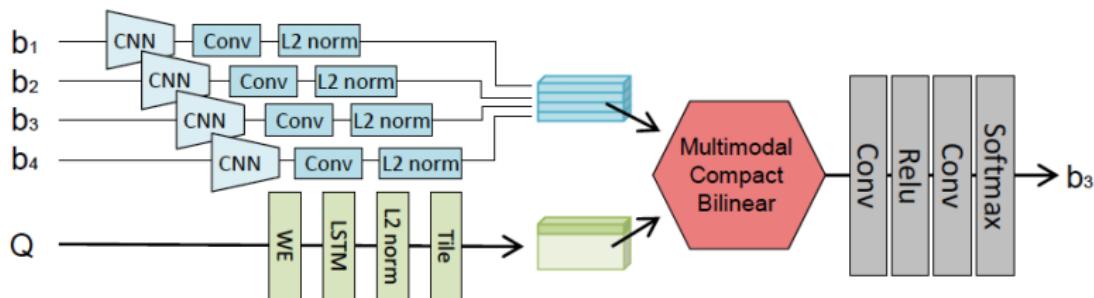
a_4 : "People waiting on a train"



Architecture of MCB



Q: "Person in blue checkered shirt"



Datasets

- ① **Visual Question Answering (VQA)**
- ② **Visual Genome**
- ③ **Visual7W**

Method	Accuracy
Element-wise Sum	56.50
Concatenation	57.49
Concatenation + FC	58.40
Concatenation + FC + FC	57.10
Element-wise Product	58.57
Element-wise Product + FC	56.44
Element-wise Product + FC + FC	57.88
MCB ($2048 \times 2048 \rightarrow 16K$)	59.83
Full Bilinear ($128 \times 128 \rightarrow 16K$)	58.46
MCB ($128 \times 128 \rightarrow 4K$)	58.69
Element-wise Product with VGG-19	55.97
MCB ($d = 16K$) with VGG-19	57.05
Concatenation + FC with Attention	58.36
MCB ($d = 16K$) with Attention	62.50

Compact Bilinear d	Accuracy
1024	58.38
2048	58.80
4096	59.42
8192	59.69
16000	59.83
32000	59.71

	Test-dev					Test-standard				
	Open Ended				MC	Open Ended				MC
	Y/N	No.	Other	All	All	Y/N	No.	Other	All	All
MCB	81.2	35.1	49.3	60.8	65.4	-	-	-	-	-
MCB + Genome	81.7	36.6	51.5	62.3	66.4	-	-	-	-	-
MCB + Att.	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
MCB + Att. + GloVe	82.5	37.6	55.6	64.7	69.1	-	-	-	-	-
MCB + Att. + Genome	81.7	38.2	57.0	65.1	69.5	-	-	-	-	-
MCB + Att. + GloVe + Genome	82.3	37.2	57.4	65.4	69.9	-	-	-	-	-
Ensemble of 7 Att. models	83.4	39.8	58.5	66.7	70.2	83.2	39.5	58.0	66.5	70.1
Naver Labs (challenge 2nd)	83.5	39.8	54.8	64.9	69.4	83.3	38.7	54.6	64.8	69.3
HieCoAtt (Lu et al., 2016)	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
DMN+ (Xiong et al., 2016)	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
FDA (Ilievski et al., 2016)	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
D-MMN (Andreas et al., 2016a)	81.1	38.6	45.5	59.4	-	-	-	-	59.4	-
AMA (Wu et al., 2016)	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
SAN (Yang et al., 2015)	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
NMN (Andreas et al., 2016b)	81.2	38.0	44.0	58.6	-	81.2	37.7	44.0	58.7	-
AYN (Malinowski et al., 2016)	78.4	36.4	46.3	58.4	-	78.2	36.3	46.3	58.4	-
SMem (Xu and Saenko, 2016)	80.9	37.3	43.1	58.0	-	80.9	37.5	43.5	58.2	-
VQA team (Antol et al., 2015)	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
DPPnet (Noh et al., 2015)	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
iBOWIMG (Zhou et al., 2015)	76.5	35.0	42.6	55.7	-	76.8	35.0	42.6	55.9	62.0

Method	Accuracy, %
Plummer et al. (2015)	27.42
Hu et al. (2016b)	27.80
Plummer et al. (2016) ¹	43.84
Wang et al. (2016)	43.89
Rohrbach et al. (2016)	47.81
Concatenation	46.50
Element-wise Product	47.41
Element-wise Product + Conv	47.86
MCB	48.69

Method	Accuracy, %
Hu et al. (2016b)	17.93
Rohrbach et al. (2016)	26.93
Concatenation	25.48
Element-wise Product	27.80
Element-wise Product + Conv	27.98
MCB	28.91

Visual Grounding

Method	What	Where	When	Who	Why	How	Avg
Zhu et al.	51.5	57.0	75.0	59.5	55.5	49.8	54.3
Concat+Att.	47.8	56.9	74.1	62.3	52.7	51.2	52.8
MCB+Att.	60.3	70.4	79.5	69.2	58.2	51.1	62.2

Visual Grounding



What vegetable is the dog chewing on?

MCB: carrot
GT: carrot



What color is the traffic light?

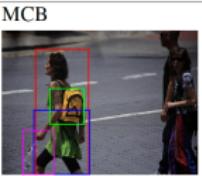
MCB: green
GT: green

Is this an urban area?

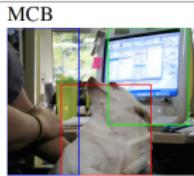
MCB: yes
GT: yes

Where are the buildings?

MCB: in background
GT: on left



A tattooed woman with a green dress and yellow backpack holding a water bottle is walking across the street.



A dog distracts his owner from working at her computer.

Bullet Points

- We propose the Multimodal Compact Bilinear Pooling (MCB) to combine visual and text representations. For visual question answering, our architecture with attention and multiple MCBs gives significant improvements on two VQA datasets compared to state-of-the-art. In the visual grounding task, introducing MCB pooling leads to improved phrase localization accuracy, indicating better interaction between query phrase representations and visual representations of proposal bounding boxes.
- The code is available in [github](#)

Thank You