Vision & Learning Lab

Multimodal Compact Bilinear Pooling

Sewon Min
2016. 07. 26

Contents

Topic

spoon source (brown)

corn (yellow)

bread

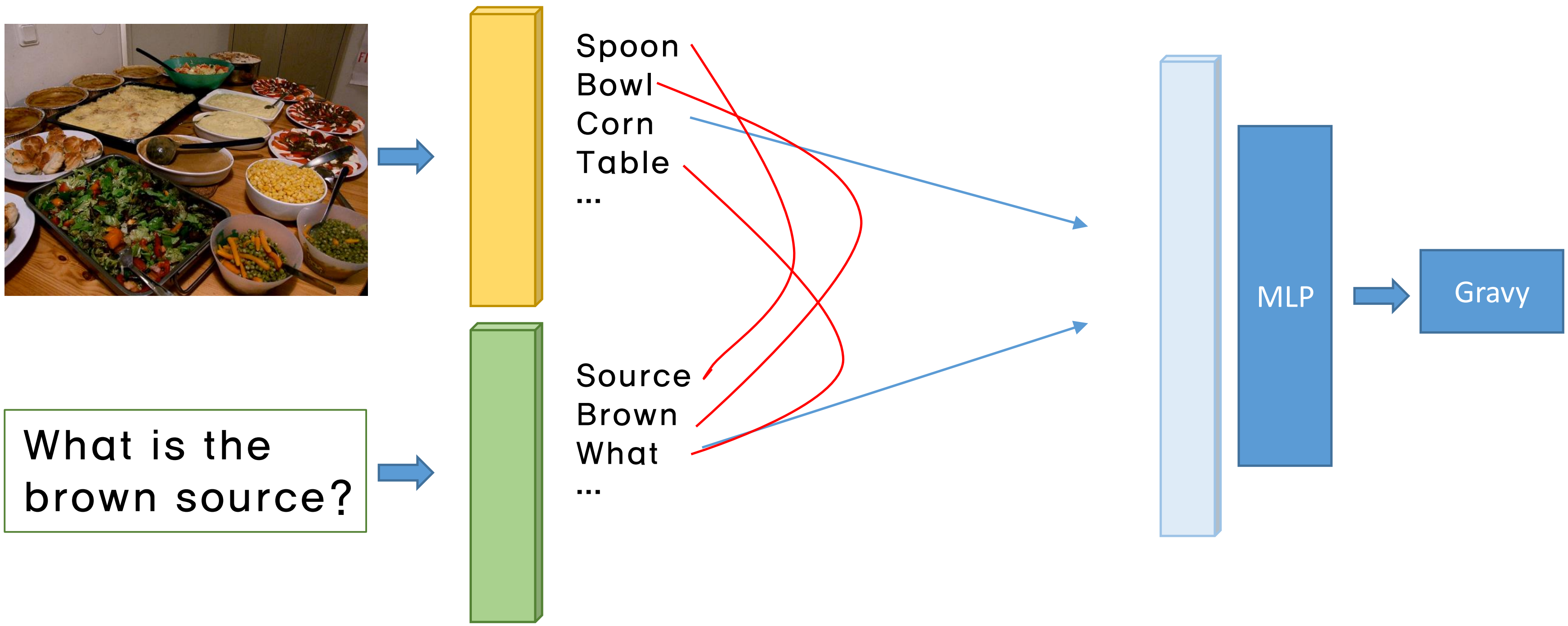
vegetable
(green)

table

What is the
brown source?

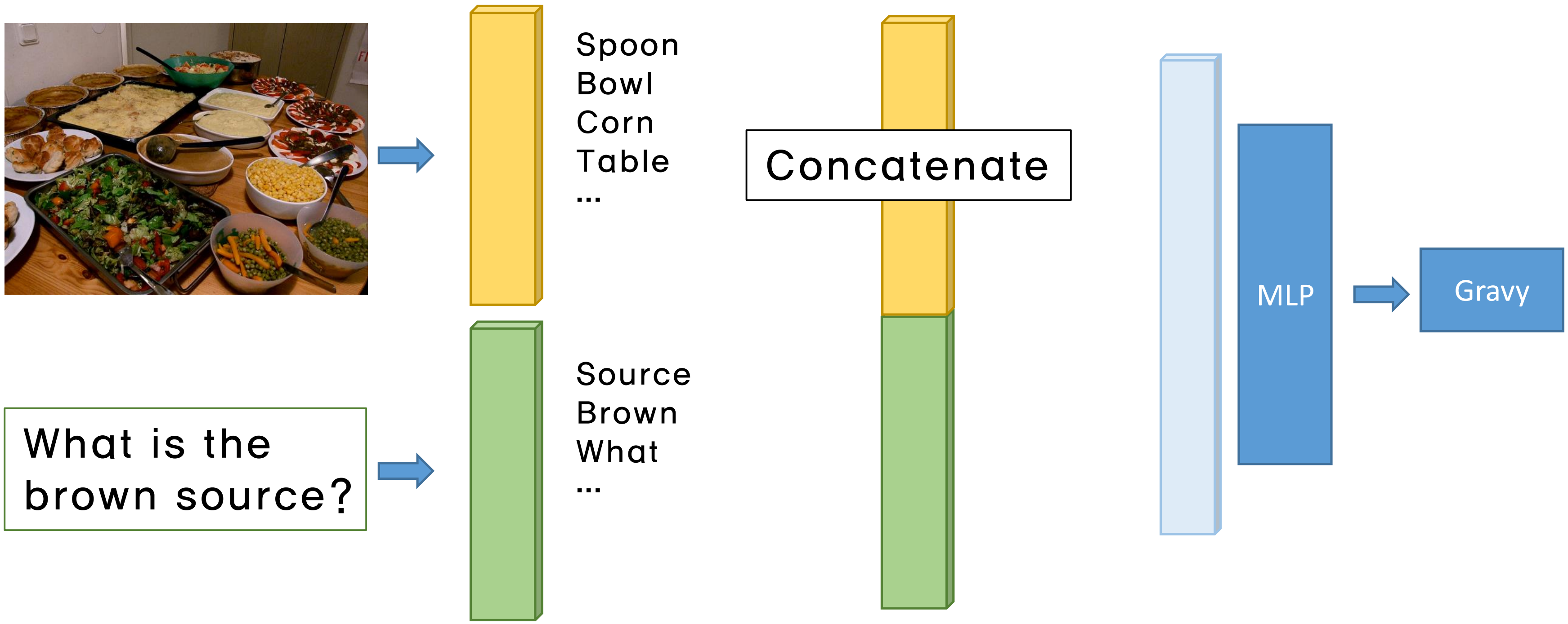
Gravy

Topic



Spoon
Bowl
Corn
Table
...
What is the brown source?
Source
Brown
What
...
MLP
Gravy

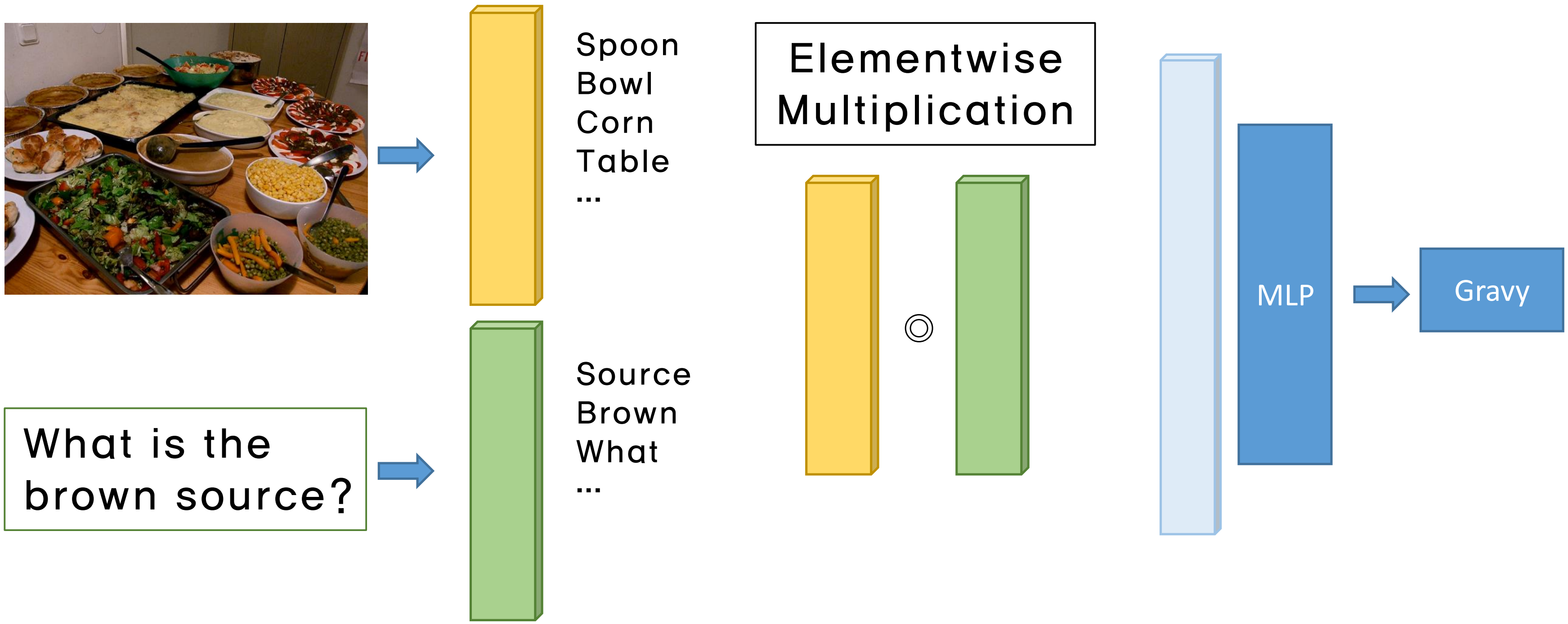
Bilinear Model



Hard to learn multiplicative interaction
between elements of two vectors

T.-Y. Lin et al. Bilinear CNN models for fine-grained visual recognition. 2015.

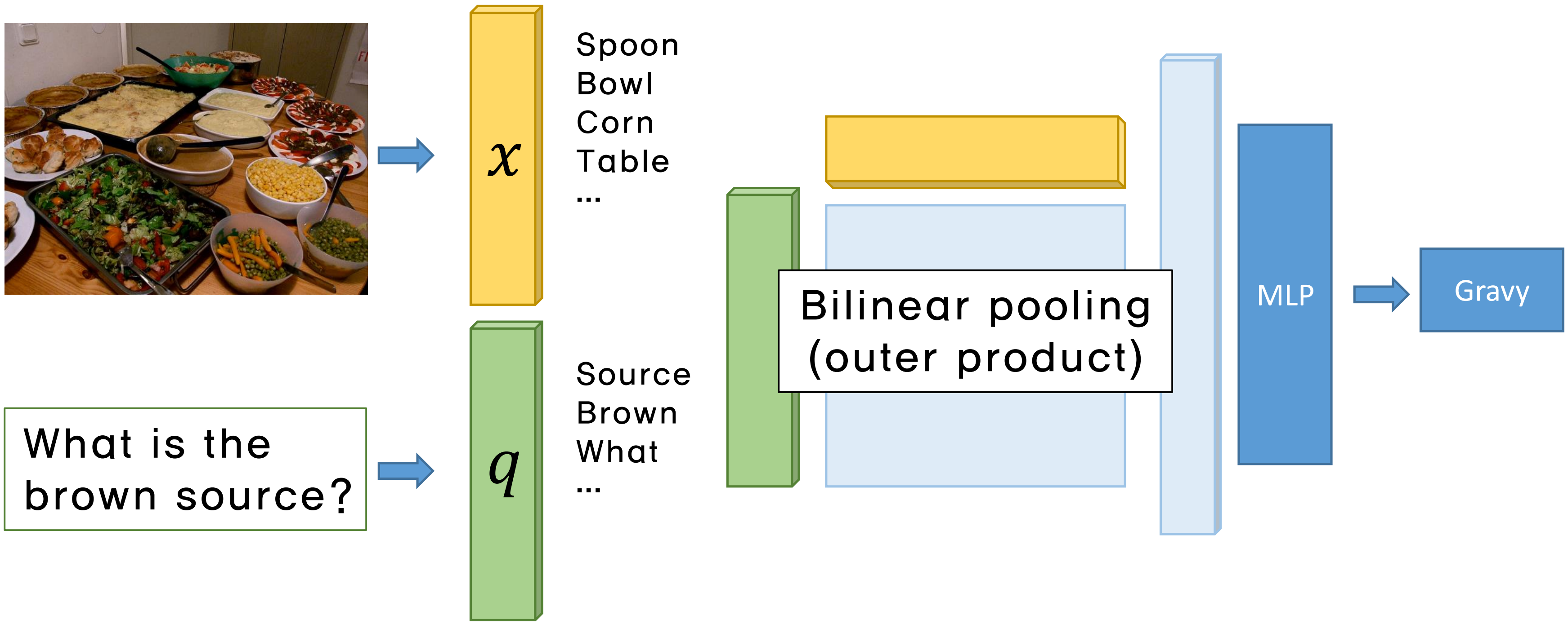
Bilinear Model



Hard to learn interaction between all elements

T.-Y. Lin et al. Bilinear CNN models for fine-grained visual recognition. 2015.

Bilinear Model

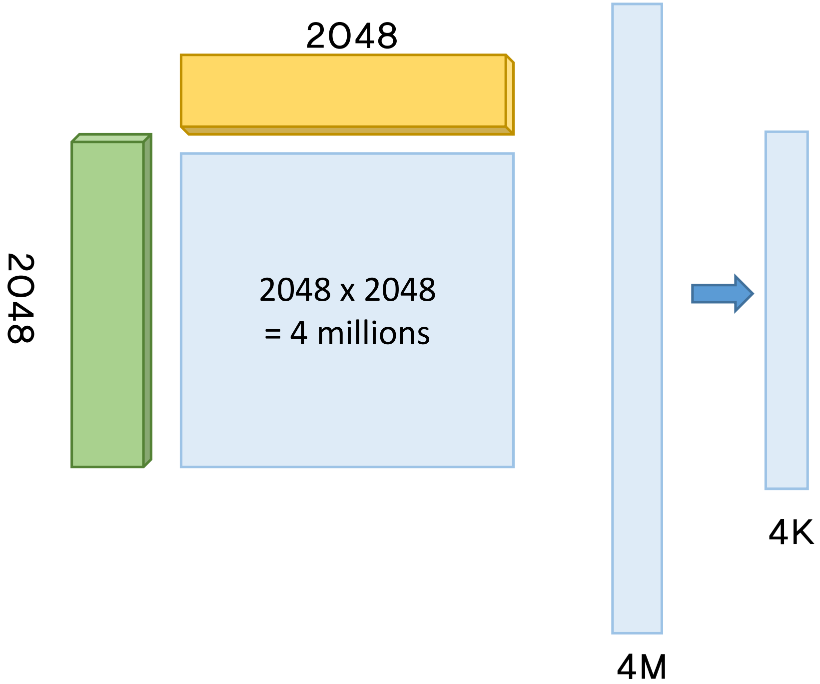


$$B(X) = \sum_{s \in S} x_s \cdot q_s{}^T = \sum_{s \in S} x_s \times q_s$$

(sum pooling)

T.-Y. Lin et al. Bilinear CNN models for fine-grained visual recognition. 2015.
J. Carreira et al. Semantic segmentation with second-order pooling. 2012.

Compact Bilinear Pooling



$$B(X) = \sum_{s \in S} x_s \times q_s$$: c dimension

$$C(X) = ?$$: d dimension

$$(\; X = \{x_1, x_2, \dots, x_{|S|}, q_1, q_2, \dots, q_{|S|}\} \;)$$

Random Maclaurin

Tensor Sketch

Y. Gao et al. Compact bilinear pooling. 2016.

Compact Bilinear Pooling

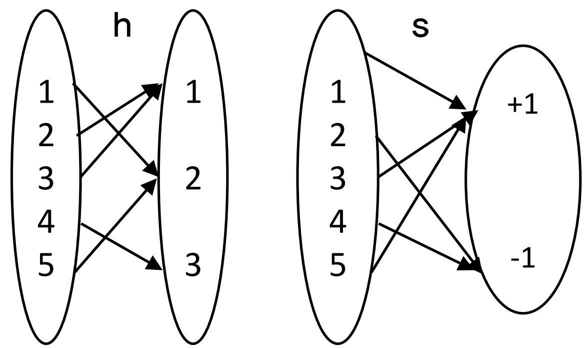
Count Sketch

Random projection

Given hash functions $h : [c] \rightarrow [d], s : [c] \rightarrow \{+1, -1\}$,
Count sketch of the point $x = \{x_1, x_2, \dots, x_c\} \in R^c$ is

$$\Psi(x, h, s) = \{y_1, y_2, \dots, y_d\} \in R^d,$$

where $y_j = \sum_{i:h(i)=j} s(i) x_i$



$$\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \Rightarrow \Psi(x, h, s) = \begin{bmatrix} -x_2 + x_3 \\ x_1 + x_5 \\ -x_4 \end{bmatrix}$$

M. Charikar et al. Finding frequent items in data streams. 2002.

Compact Bilinear Pooling

Count Sketch

1) $< u, v > \approx < \Psi(u, h, s), \Psi(v, h, s) >$

$E[< \Psi(u, h, s), \Psi(v, h, s) >] = < u, v >$

$Var[< \Psi(u, h, s), \Psi(v, h, s) >] = \frac{1}{d}\left(\sum_{i \neq j}({u_i}^2 {v_j}^2 + u_i v_i u_j v_j)\right)$

pf)

$$let < \psi(u, h, s), \psi(v, h, s) > \; = \; < u, v >_{\psi}$$

$$\begin{aligned}
\mathrm{E}_{\psi}[< u, v >_{\psi}] &= \mathrm{E}_{\boldsymbol{h}}[\mathrm{E}_{\boldsymbol{s}}[< u, v >_{\psi}]] = \mathrm{E}_{\boldsymbol{h}}[\mathrm{E}_{\boldsymbol{s}}[\sum_{i,j} s(i)s(j)u_i v_j \delta_{h(i),h(j)}]] \\
&= \sum_{i,j} u_i v_j \qquad (\because \; \mathrm{E}_{\boldsymbol{s}}[\sum_{i \neq j} s(i)s(j) = 0) \\
&= < u, v >
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\psi}}[< u, v >_{\psi}^2] &= \mathrm{E}_{\boldsymbol{\psi}}[\sum_{i,j,k,l} s(i)s(j)s(k)s(l)u_i v_j u_k v_l \delta_{h(i),h(j)} \delta_{h(k),h(l)}] \\
&= \sum_{i,k} u_i v_i u_k v_k + \sum_{i \neq j} u_i^2 v_j^2 \mathrm{E}_{\boldsymbol{h}}[\delta_{h(i),h(j)}] + \sum_{i \neq j} u_i v_i u_j v_j \mathrm{E}_{\boldsymbol{h}}[\delta_{h(i),h(j)}] \\
&= < u, v >^2 + \frac{1}{d}\left(\sum_{i \neq j} u_i^2 v_j^2 + u_i v_i u_j v_j\right)
\end{aligned}$$

K. Q. Weinberger et al. Feature hashing for large scale multitask learning. 2009.
N. Pham and R. Paph. Fast and scalable polynomial kernels via explicit feature maps. 2013.

Compact Bilinear Pooling

Count Sketch

1) $< u, v > \approx < \Psi(u, h, s), \Psi(v, h, s)>$

$E[< \Psi(u, h, s), \Psi(v, h, s)>] = < u, v >$
$Var[< \Psi(u, h, s), \Psi(v, h, s)>] = \frac{1}{d}(\sum_{i \neq j}(u_i{}^2 v_j{}^2 + u_i v_i u_j v_j))$

pf)

$$\mathrm{Var}_{\psi}[< u,v >_{\psi}] = \mathrm{E}_{\psi}[< u,v >_{\psi}^{2}] - \mathrm{E}_{\psi}[< u,v >]^2 = \frac{1}{d}\left(\sum_{i \neq j} u_i^2 v_j^2 + u_i v_i u_j v_j\right)$$

Relative error bound (Chebyshev's inequality)

$$\mathrm{P}\left[\left|\frac{< u,v >_{\psi} - < u,v >}{< u,v >}\right| \geq \epsilon\right] \leq \frac{\mathrm{Var}_{\psi}[< u,v >_{\psi}]}{\epsilon^2 \mathrm{E}_{\psi}[< u,v >]^2}$$
$$\leq \frac{2}{d\epsilon^2}\left(\frac{1}{\cos\theta_{xy}}\right)^2$$

K. Q. Weinberger et al. Feature hashing for large scale multitask learning. 2009.
N. Pham and R. Paph. Fast and scalable polynomial kernels via explicit feature maps. 2013.

Compact Bilinear Pooling

Count Sketch

2) Given $x, y \in R^c$, 2-wise independent hash functions h_1, h_2, s_1, s_2,

$$\Psi(x \times y, h, s) = FFT^{-1}(FFT(\Psi(x, h_1, s_1) \circledcirc FFT(\Psi(y, h_2, s_2)))$$

$$\equiv \Psi(x, h_1, s_1) * \Psi(y, h_2, s_2)$$

$$h(i, j) = h_1(i) + h_2(j), \mathrm{mod}\ d, s(i, j) = s_1(i) s_2(j)$$

Y. Gao et al. Compact bilinear pooling. 2016.
N. Pham and R. Paph. Fast and scalable polynomial kernels via explicit feature maps. 2013.

Compact Bilinear Pooling

$$\Psi(x \times y, h, s) = FFT^{-1}(FFT(\Psi(x, h_1, s_1) \circledcirc \mathrm{FFT}(\Psi(y, h_2, s_2)))$$

pf)

Count Sketch $\psi(x, h, s)$ of d dimension can be represented as a polynomial of d-1 dimension

$$P_x^{h,s}(\omega) = \sum_{i=1}^{c} s(i) x_i \omega^{h(i)}$$

(basis of d dimension : $[P_x^{h,s}(\omega^{*0}), P_x^{h,s}(\omega^{*1}), \ldots, P_x^{h,s}(\omega^{*d-1})]$ where $\omega^{*d} = 1$)

Then, $\psi(x, h_1, s_1) \rightarrow P_x^{h_1,s_1}(\omega) = \sum_{i=1}^{c} s_1(i) x_i \omega^{h_1(i)}$

$\psi(y, h_2, s_2) \rightarrow P_y^{h_2,s_2}(\omega) = \sum_{i=1}^{c} s_2(i) y_i \omega^{h_2(i)}$

$$\begin{aligned}
\psi(x \times y, h, s) \rightarrow P_{x \times y}^{h,s}(\omega) &= \sum_{i,j=1}^{c} s(i,j) x_i y_j \omega^{H(i,j)} \\
&= \sum_{i,j=1}^{c} s_1(i) s_2(j) x_i y_j \omega^{h_1(i)} \omega^{h_2(j)} \\
&= P_x^{h_1,s_1}(\omega) \cdot P_y^{h_2,s_2}(\omega) \\
&= FFT^{-1}(FFT(P_x^{h_1,s_1}(\omega)) \odot FFT(P_y^{h_2,s_2}(\omega)))
\end{aligned}$$

Y. Gao et al. Compact bilinear pooling. 2016.
N. Pham and R. Paph. Fast and scalable polynomial kernels via explicit feature maps. 2013.
R. Pagh. Compressed matrix multiplication. 2012.

Compact Bilinear Pooling

Compact bilinear pooling using Tensor sketch

1) $< u, v > \approx < \Psi(u, h, s), \Psi(v, h, s) >$

2) $\Psi(x \times y, h, s) = \Psi(x, h_1, s_1) \ast \Psi(y, h_2, s_2)$

$B(X) = \sum_{s \in S} x_s \times q_s$ ($X = \{x_1, x_2, \dots, x_{|S|}, q_1, q_2, \dots, q_{|S|}\}$)

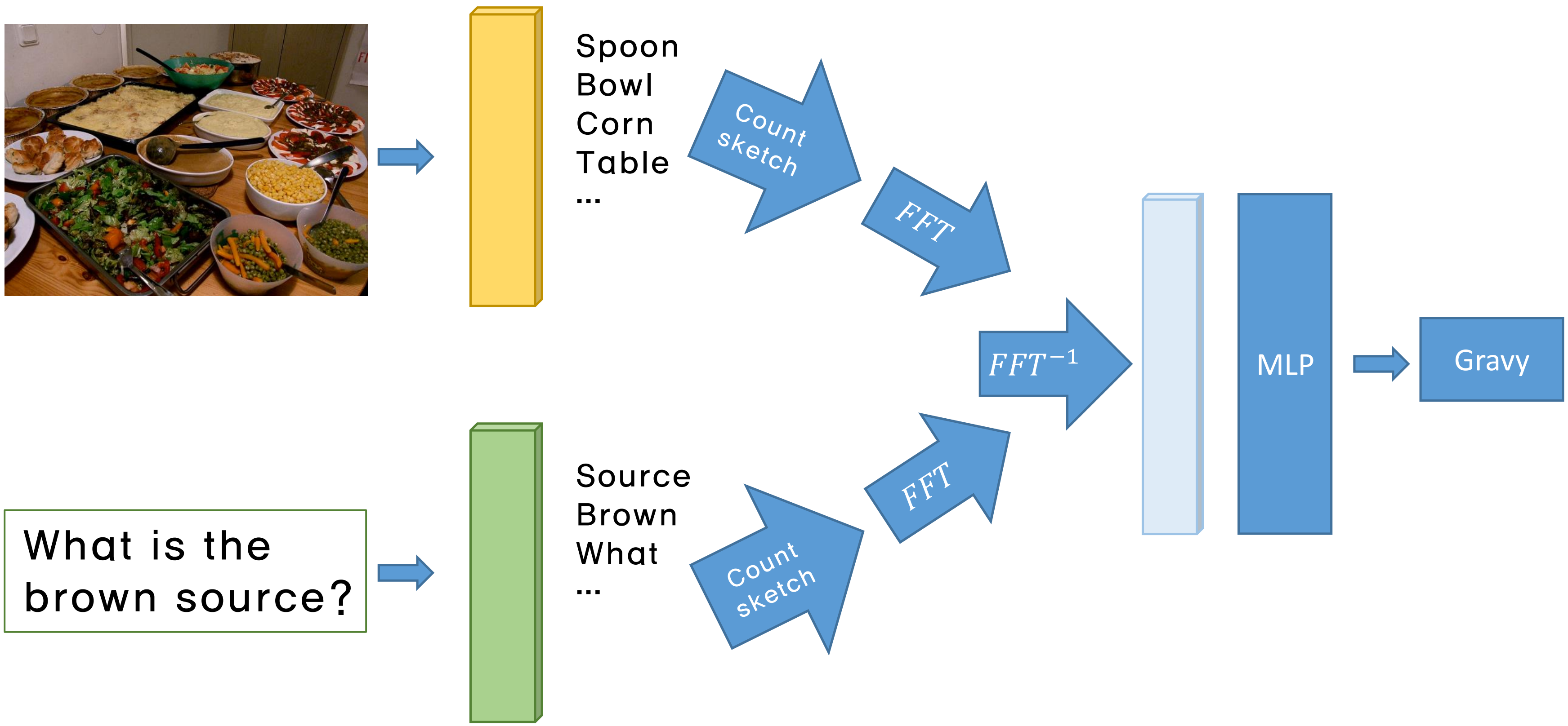
$C(X) = \sum_{s \in S} \Psi(x_s, h_1, s_1) \ast \Psi(q_s, h_2, s_2)$

Given $X = \{x, q\}, Y = \{y, r\}$,

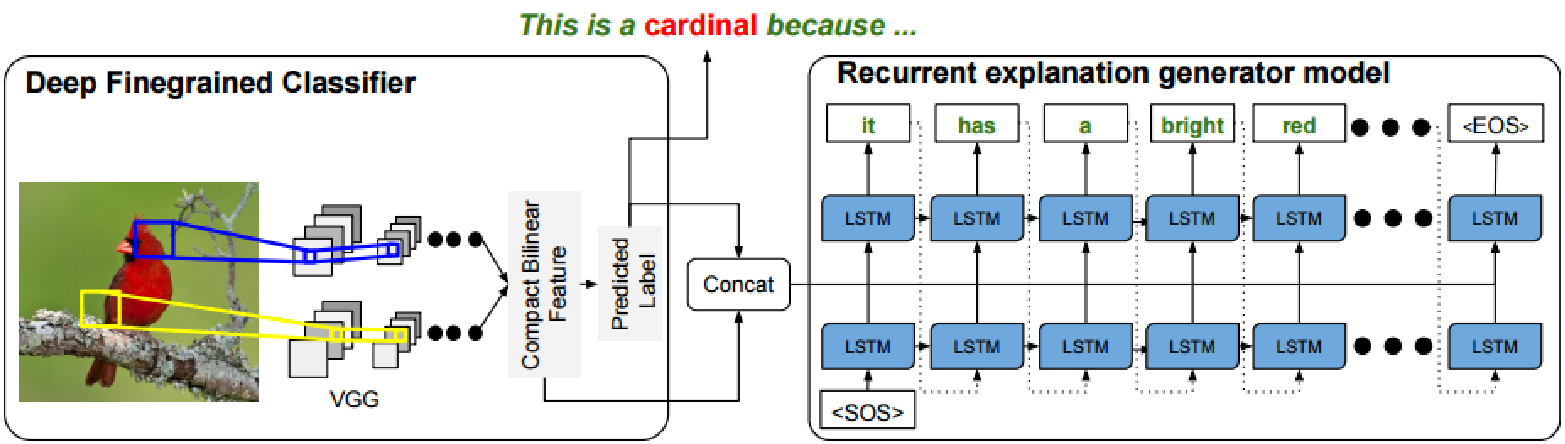
$$
\begin{aligned}
< B(X), B(Y) > &= \sum_{s \in S} \sum_{u \in U} < x_s \times q_s, y_u \times r_u > \\
&\approx \sum_{s \in S} \sum_{u \in U} < \Psi(x_s \times q_s, h, s), \Psi(y_u \times r_u, h, s) > \\
&= \sum_{s \in S} \sum_{u \in U} < \Psi(x_s, h_1, s_1) \ast \Psi(q_s, h_2, s_2), \Psi(y_u, h_1, s_1) \ast \Psi(r_u, h_2, s_2) > \\
&= < C(X), C(Y) >
\end{aligned}
$$

Y. Gao et al. Compact bilinear pooling. 2016.

Compact Bilinear Pooling



Models

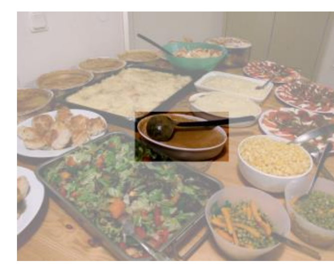


Visual QA

What is the brown source? → Gravy

Visual Grounding

The bowl with the brown source →



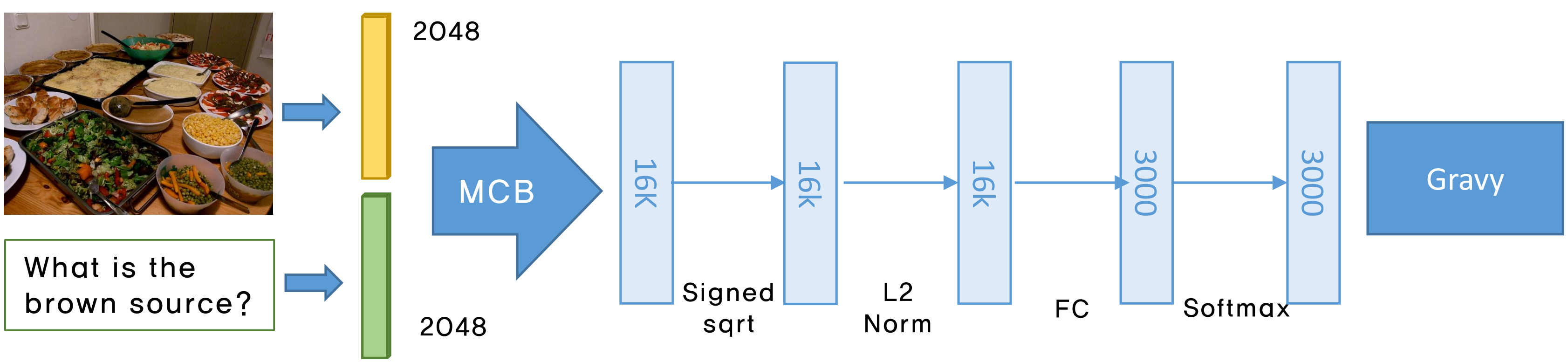
L. A. Hendricks et al. Generating Visual Explanations. 2016.
A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Models

1) Visual QA without Attention

Feature extraction
  image : ResNet 152 (Before FC)
  Question : 2-layer LSTM with output size 1024
  (embedding words 13k-20k, embedding size 300)

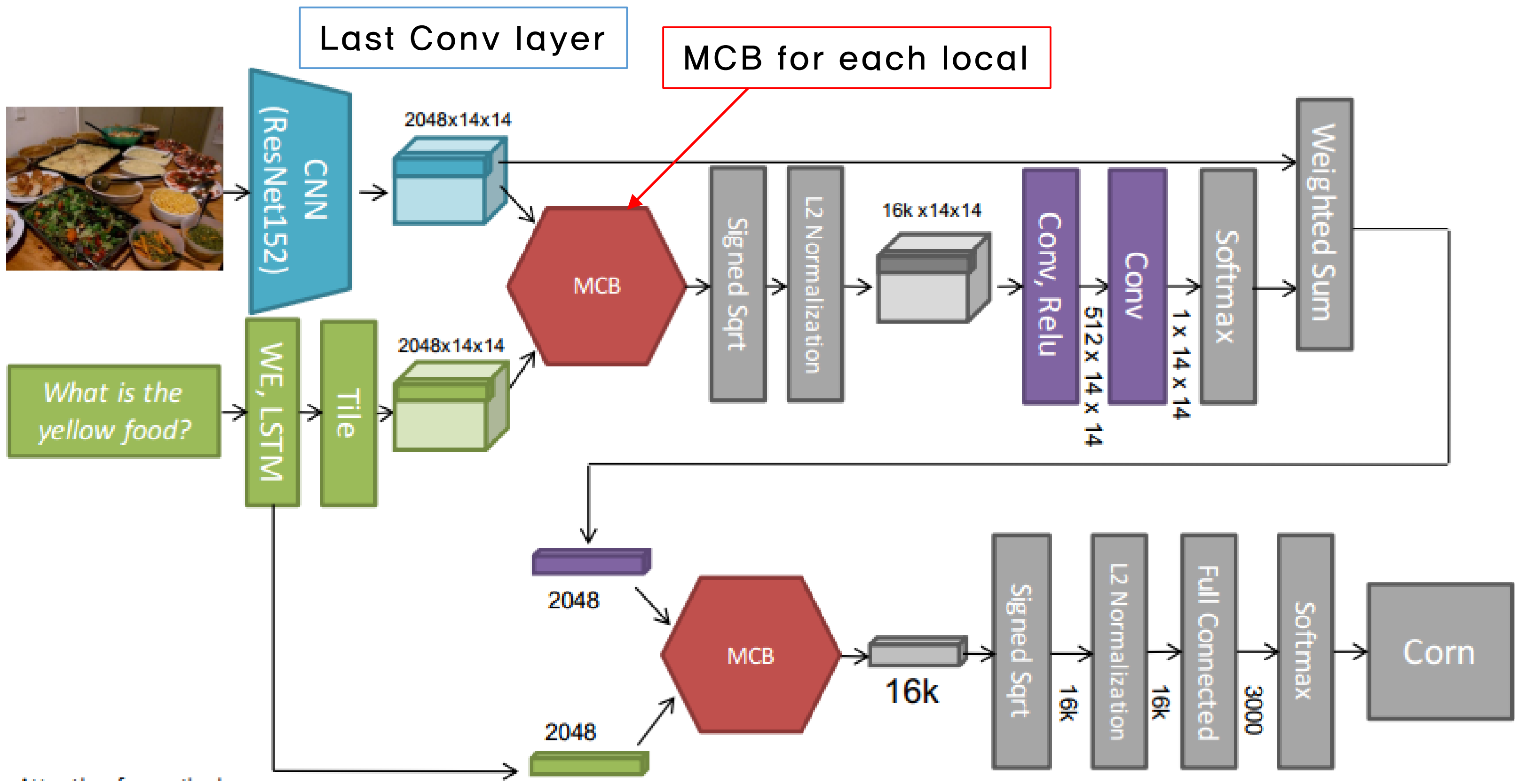
Answer decoding
  3000 most frequent answers on train



A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

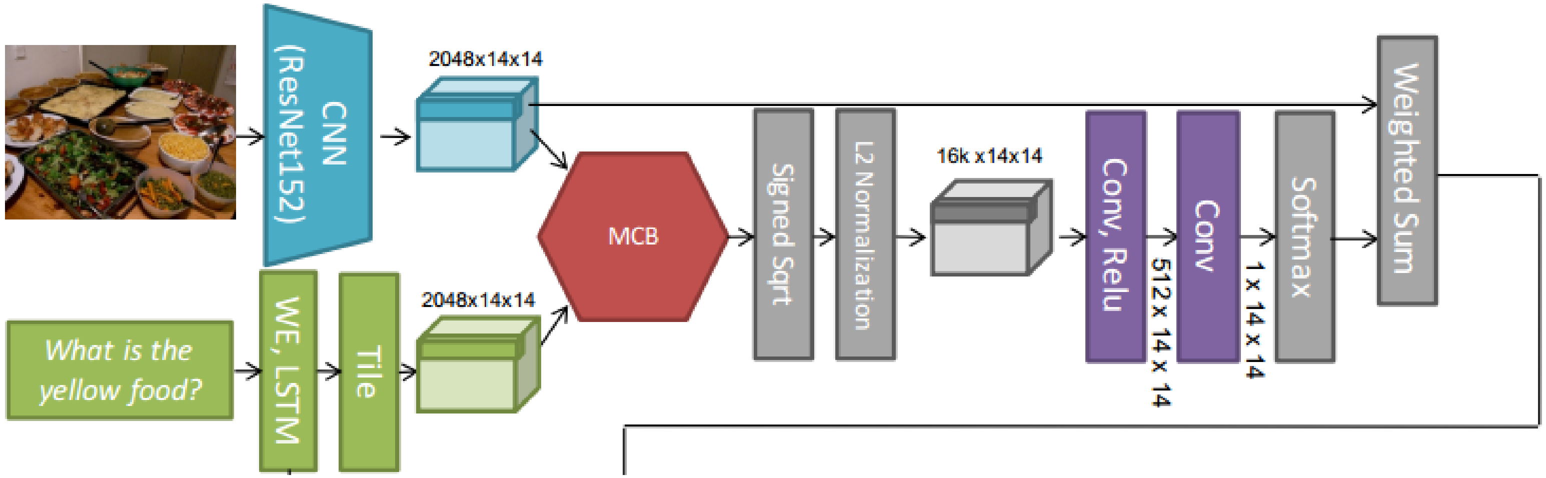
2) Visual QA with Attention

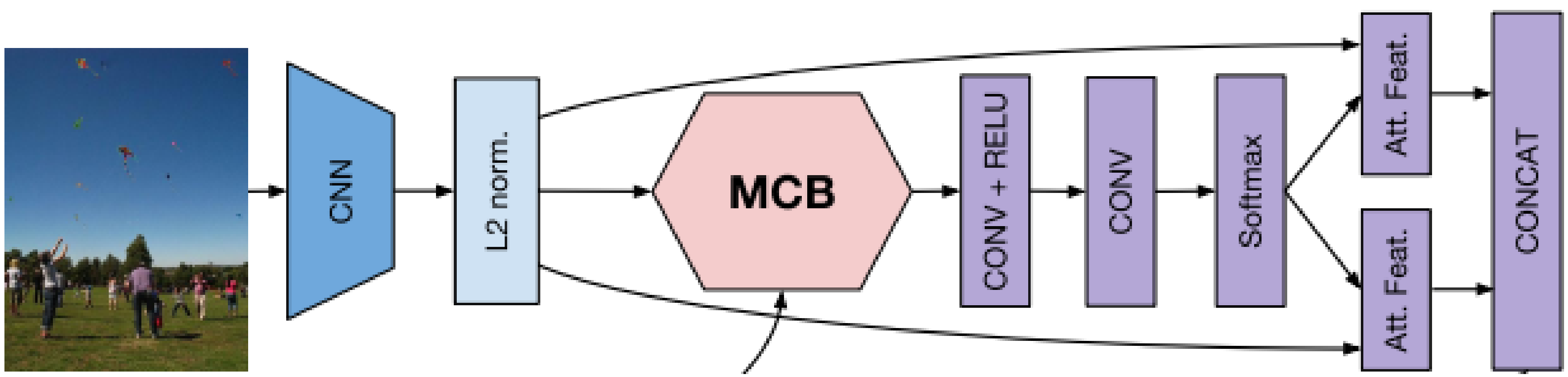
2 MCB, Multiple attentions



A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Models



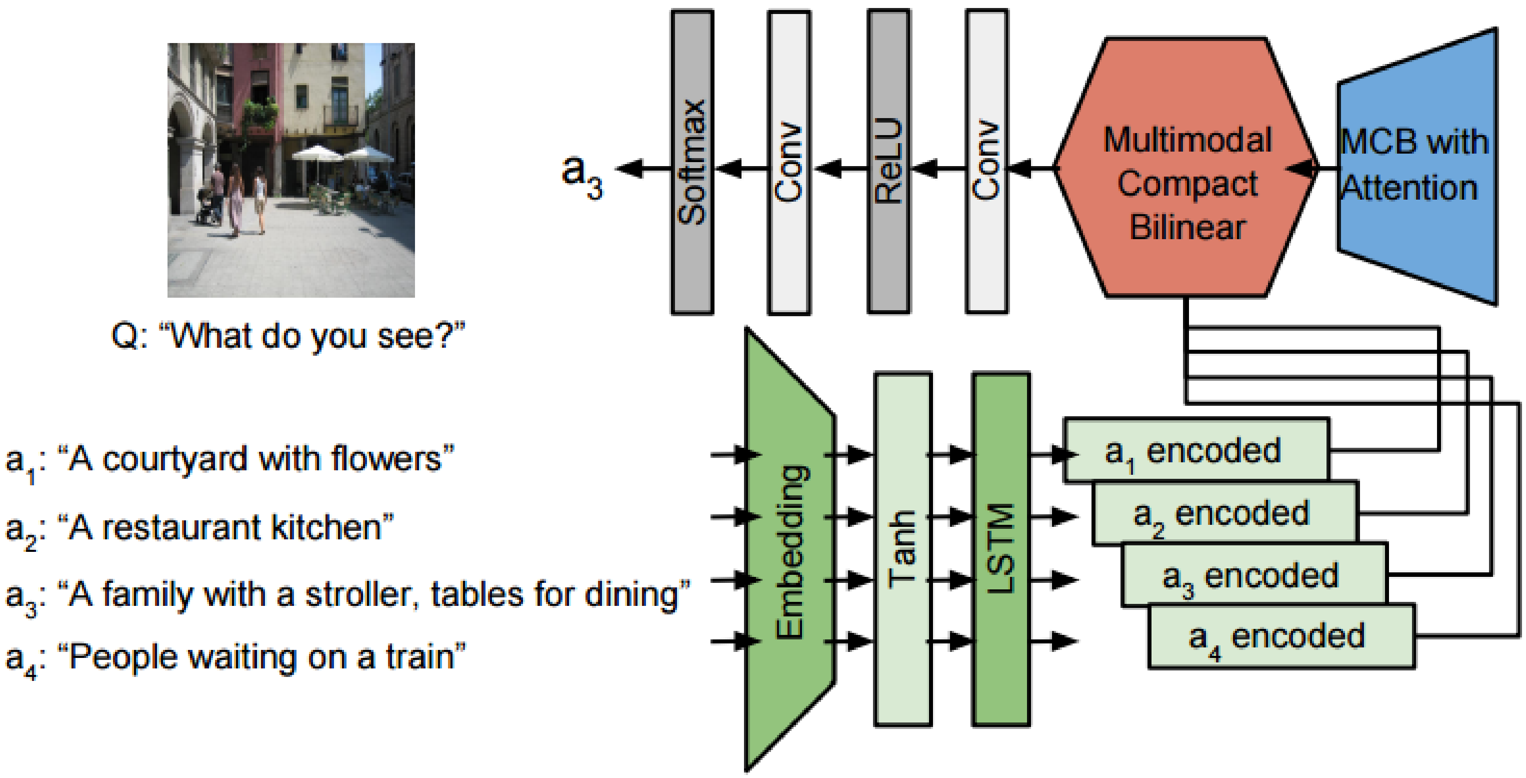


A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Models

3) Visual QA (Multiple choices)

Extra embedding for answer candidates (share parameters)



A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Datasets

1) Visual Question Answering real-image dataset

- Approximately 200K MSCOCO images (train 80K, valid 40K, test 80K)
- 3 questions/image, 10 answers/question
- Evaluation : accuracy = $\min(\frac{\#\text{ human provided that answer}}{3}, 1)$

2) Visual7W

- Part of the Visual Genome (6W + 7^{th} which question)
- 47300 images from MSCOCO, 139868 QA pairs
- Multiple choice, 4 answer candidates / question
- More balanced distribution of 6W question types, longer question and answers

A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Results

Non-bilinear vs bilinear
with same number of parameters

$4096^2 + 4096^2 + 4096 \times 3000 \approx 46\,million$

$16000 \times 3000 = 48\,million$

Full bilinear vs Compact bilinear

Better regardless of CNN

Better with attention

Method	Accuracy
Eltwise Sum	56.50
Concat	57.49
Concat + FC	58.40
Concat + FC + FC	57.10
Eltwise Product	58.57
Eltwise Product + FC	56.44
Eltwise 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
Eltwise Product with VGG-19	55.97
MCB ($d = 16K$) with VGG-19	**57.05**
Concat + FC with Attention	58.36
MCB ($d = 16K$) with Attention	**62.50**

A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Results

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

Dimension of compact bilinear

No. of attention maps	Accuracy
1	64.67
2	**65.08**
4	64.24

Number of attention maps

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**

Multiple-choice accuracy
on Visual7W

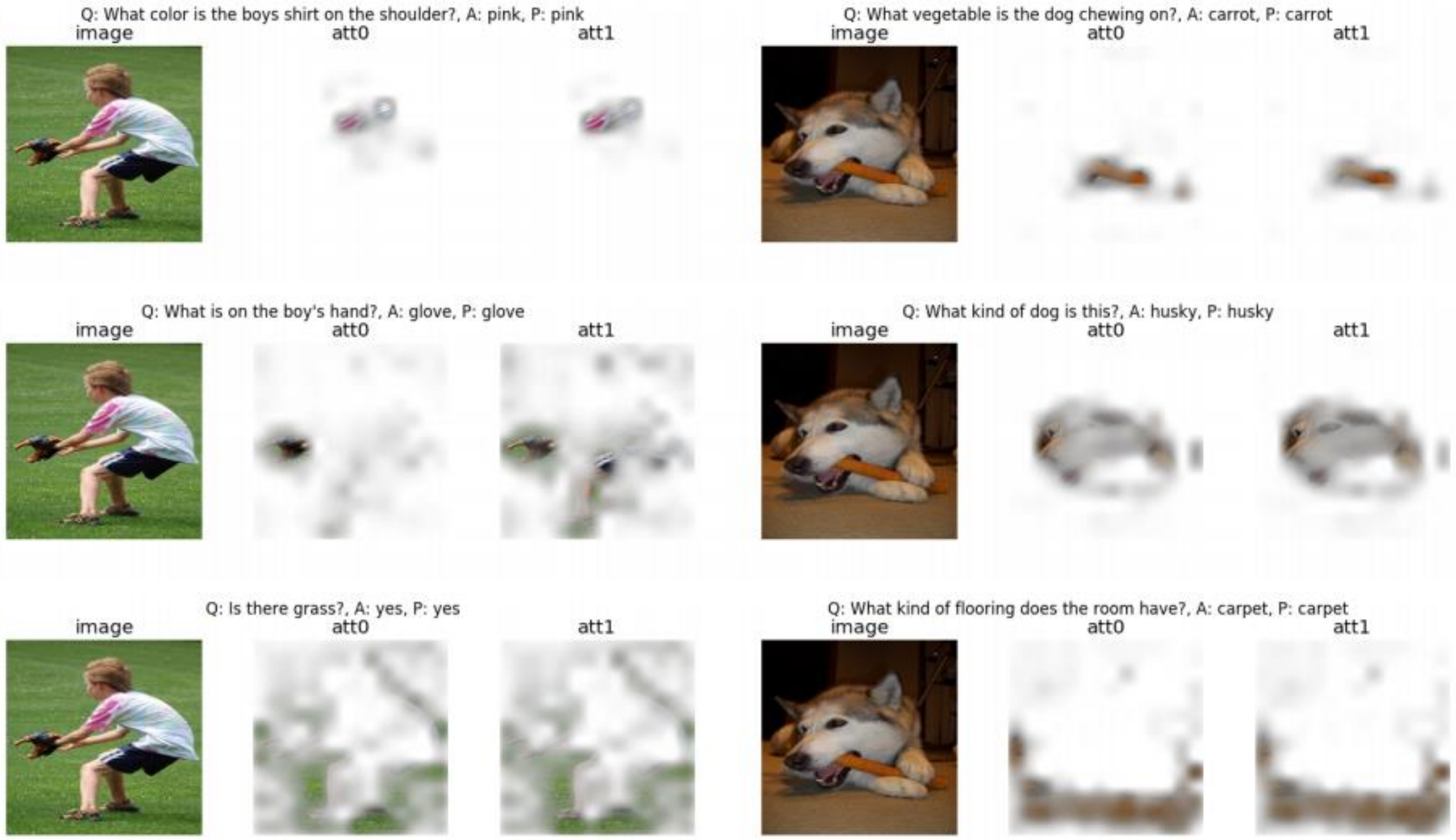
A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

QA Results

	Test-dev					Test-standard				
	Open Ended				MC	Open Ended				MC
	Y/N	No.	Other	All	All	Y/N	No.	Other	All	All
MCB	81.7	36.9	49.0	61.1	-	-	-	-	-	-
MCB + Genome	81.7	36.6	51.5	62.3	66.4	-	-	-	-	-
MCB + Att.	82.2	37.7	54.8	64.2	-	-	-	-	-	-
MCB + Genome + Att.	81.7	38.2	57.0	65.1	-	-	-	-	-	-
MCB + Genome + Att. + GloVe	82.3	37.2	57.4	65.4	-	-	-	-	-	-
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 (2nd best on server)	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-NMN (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, 2015)	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

A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

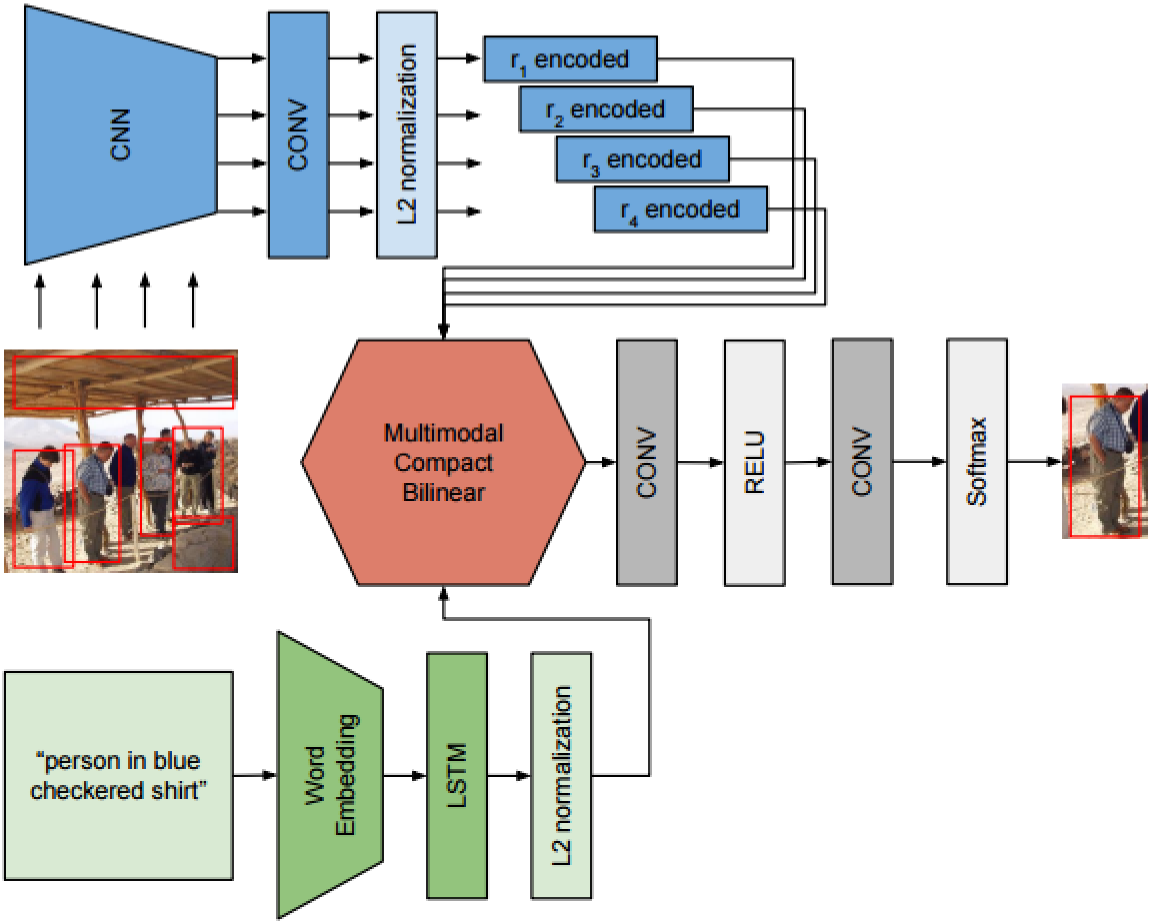
QA Results



A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

Visual Grounding Models

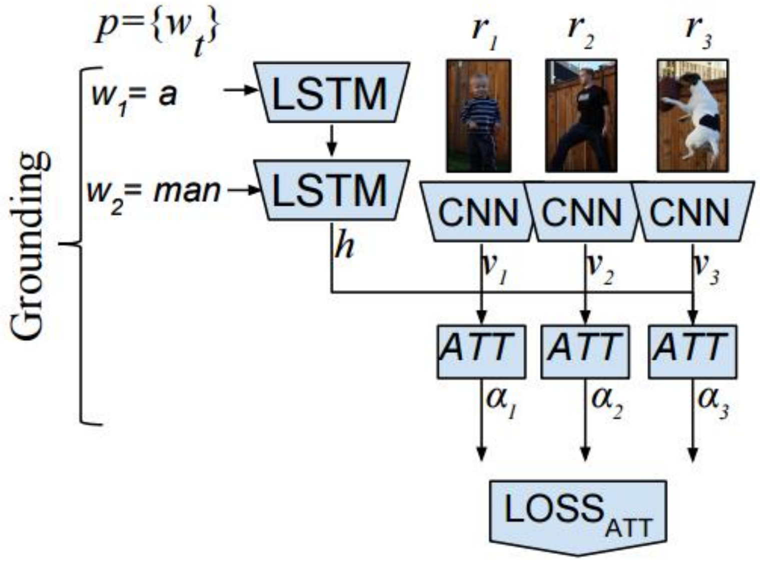
Visual Grounding



A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

GroundeR

Fully-supervised version of GroundeR (for comparing)



$$
\begin{aligned}
&Given\ I, p, r_i, \\
&h = f_{LSTM}(p) \\
&v_i = f_{CNN}(I, r_i) \\
&\alpha_i = f_{ATT}(p, r_i) = W\Phi(W_h h + W_v v_i + b_1) + b_2
\end{aligned}
$$

A. Rohrbach et al. Grounding of Textual Phrases in Images by Reconstruction. 2016.

Visual Grounding Datasets

1) Flickr30k Entities

- 31K images from Flickr30k with 244K phrases localized with bounding boxes
- follow the experimental setup from **Rohrbach et al.** (Selective Search, Fast R-CNN, fine-tuned VGG16 features)

2) ReferItGame

- 20K images from IAPR TC- 12 dataset, with segmented regions from SAIAPR-12 dataset, and 120K associated natural language referring expressions
- follow the experimental setup from **Hu et al.** (Edge Box object proposals and VGG16 combined with the spatial features)

A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

Visual Grounding Results

Method	Accuracy, %
Plummer et al. (2015)	25.30
Hu et al. (2016b)	27.80
Plummer et al. (2016)[1]	43.84
Wang et al. (2016)	43.89
Rohrbach et al. (2016)	47.70
Concat	46.50
Eltwise Product	47.41
Eltwise Product + Conv	47.86
MCB	**48.69**

Table 6: Grounding accuracy on Flickr30k Entities dataset

Method	Accuracy, %
Hu et al. (2016b)	17.93
Rohrbach et al. (2016)	26.93
Concat	25.48
Eltwise Product	27.80
Eltwise Product + Conv	27.98
MCB	**28.91**

Table 7: Grounding accuracy on ReferItGame dataset.

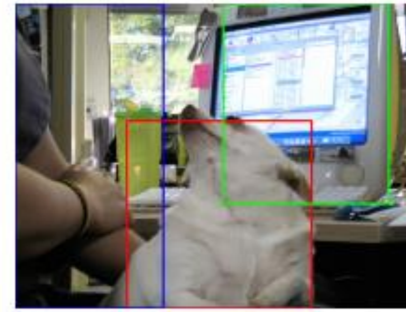
A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

Results (Visual QA + Visual Grounding)

What is the operating system on the laptop?
EP: apple
MCB: windows

How fast is this train?
EP: very
MCB: slow

What is parked next to the baskets?
EP: vegetables
MCB: motorcycle

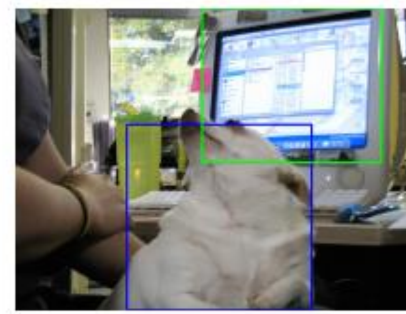


A dog distracts his owner from working at her computer.

What moves people to the top of the hill?
EP: snow
MCB: ski lift

What kind of vehicle is this?
EP: motorcycle
MCB: bicycle

What brand of tennis racket is that?
EP: adidas
MCB: wilson



A dog distracts his owner from working at her computer.

A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

Conclusion

- Bilinear pooling is able to learn multiple interaction between all elements of two vectors.
- Compact bilinear pooling using tensor sketch solve the problem of dimension.
- Multimodal compact bilinear pooling model for Visual QA achieve state-of-art with relatively simple structure.
- Multimodal compact bilinear pooling would be able to be used in any multimodal tasks.

Thank you!

Reference

[Examples and Models]
A. Fukui et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016.

[Bilinear pooling]
T.-Y. Lin et al. Bilinear CNN models for fine-grained visual recognition. 2015.
J. Carreira et al. Semantic segmentation with second-order pooling. 2012.

[Compact bilinear pooling]
Y. Gao et al. Compact bilinear pooling. 2016.
N. Pham and R. Paph. Fast and scalable polynomial kernels via explicit feature maps. 2013.
M. Charikar et al. Finding frequent items in data streams. 2002.
K. Q. Weinberger et al. Feature hashing for large scale multitask learning. 2009
R. Pagh. Compressed matrix multiplication. 2012.

[Models]
L. A. Hendricks et al. Generating Visual Explanations. 2016.
A. Rohrbach et al. Grounding of Textual Phrases in Images by Reconstruction. 2016.

More for MCB for Visual QA

VQA model slides : http://visualqa.org/static/slides/vqa_final.pdf
Demo : demo.berkeleyvision.org
Code: https://github.com/akirafukui/vqa-mcb/
VQA challenge : http://visualqa.org/challenge.html

A special thanks to Yunseok Jang, Hyungjin Ko, and Sangeon Park