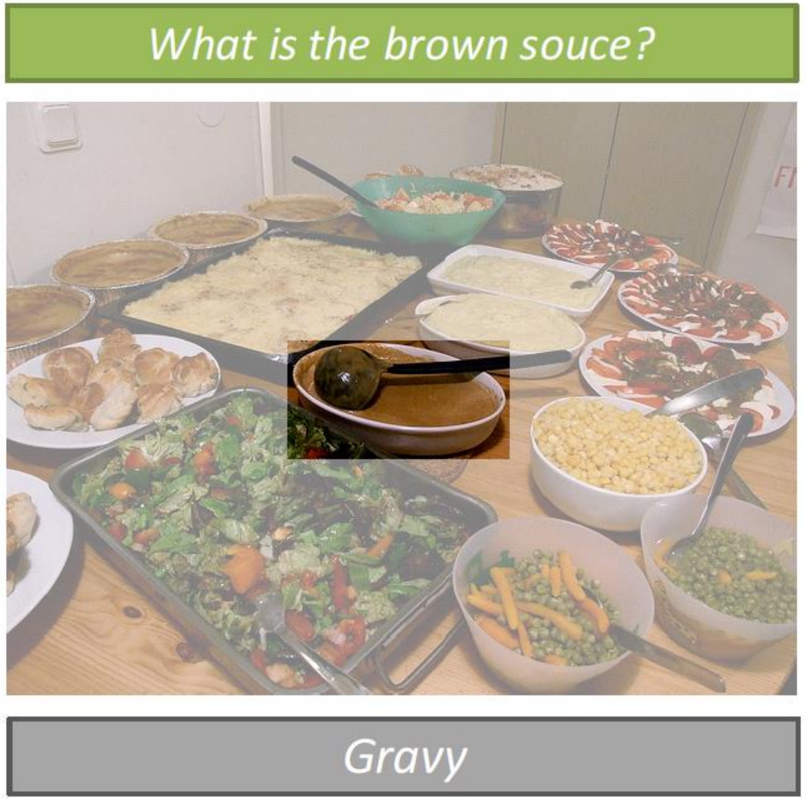
Multimodal Compact Bilinear Pooling for VQA

19.03.09
한상훈

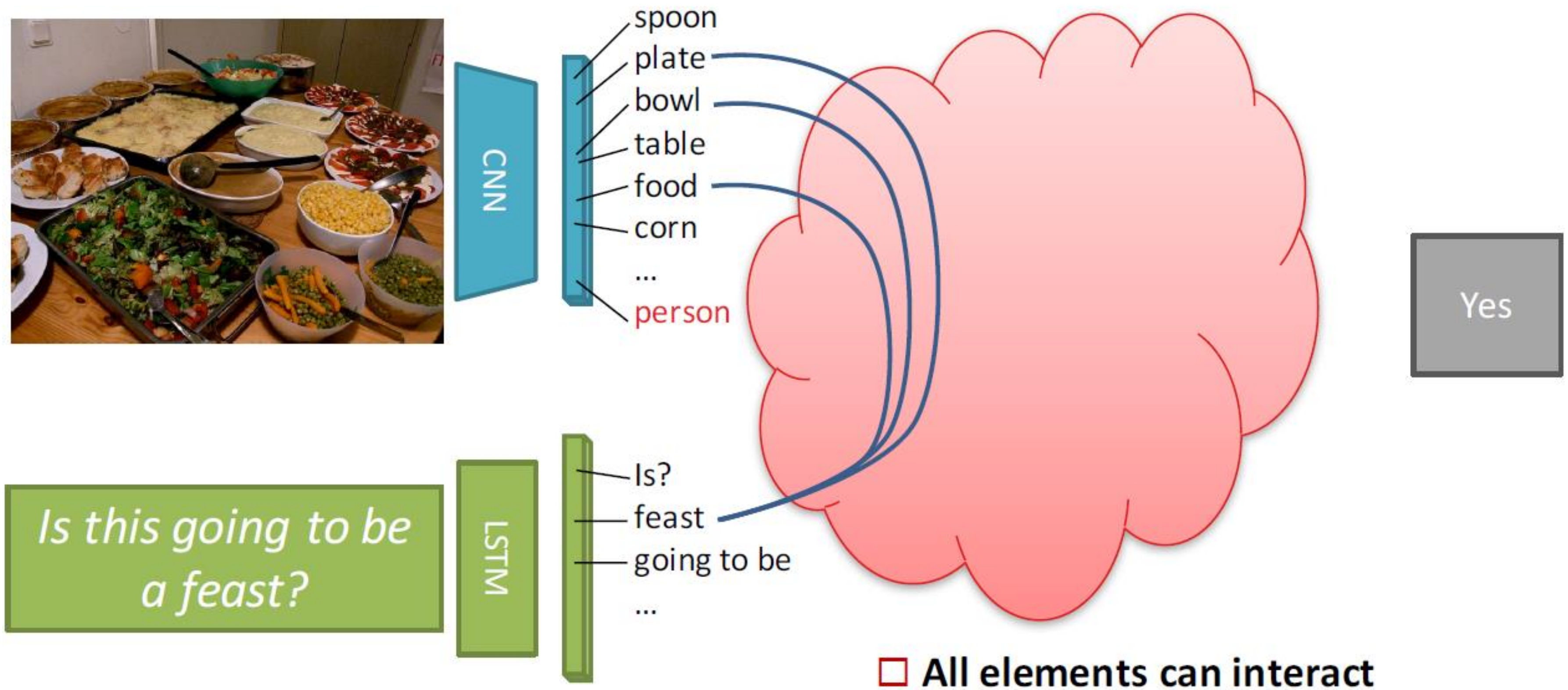
Visual Question Answering



What is the brown souce?

Gravy

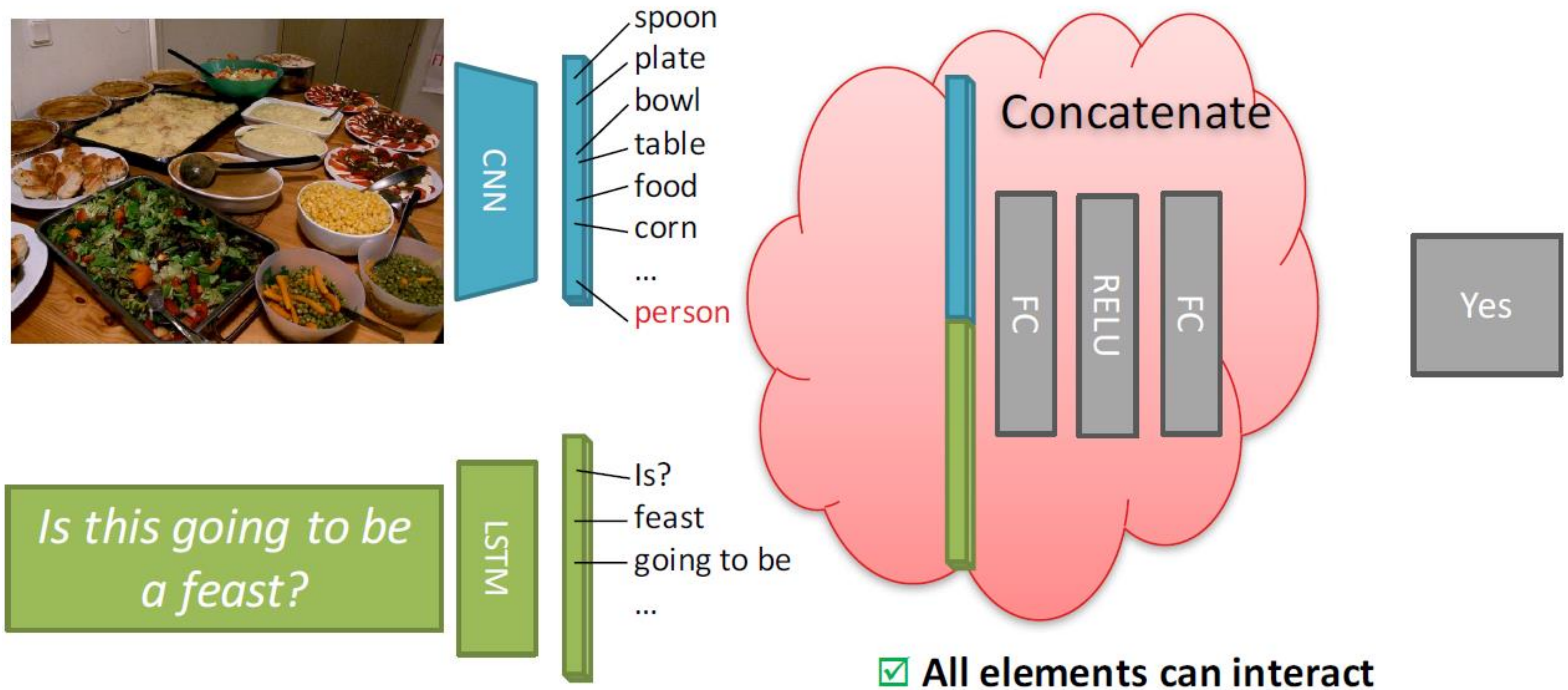
- Multimodal information
  - 이미지에서 미리 정해 놓은 질문에 대한 정보를 뽑을 수 있어야한다
  - Visual과 Language라는 서로 다른 정보를 잘 섞어야 함



□ **All elements can interact**

□ **Multiplicative interaction**

- Image(CNN) : plate, bowl, food, ….
- Question(LSTM) : feast

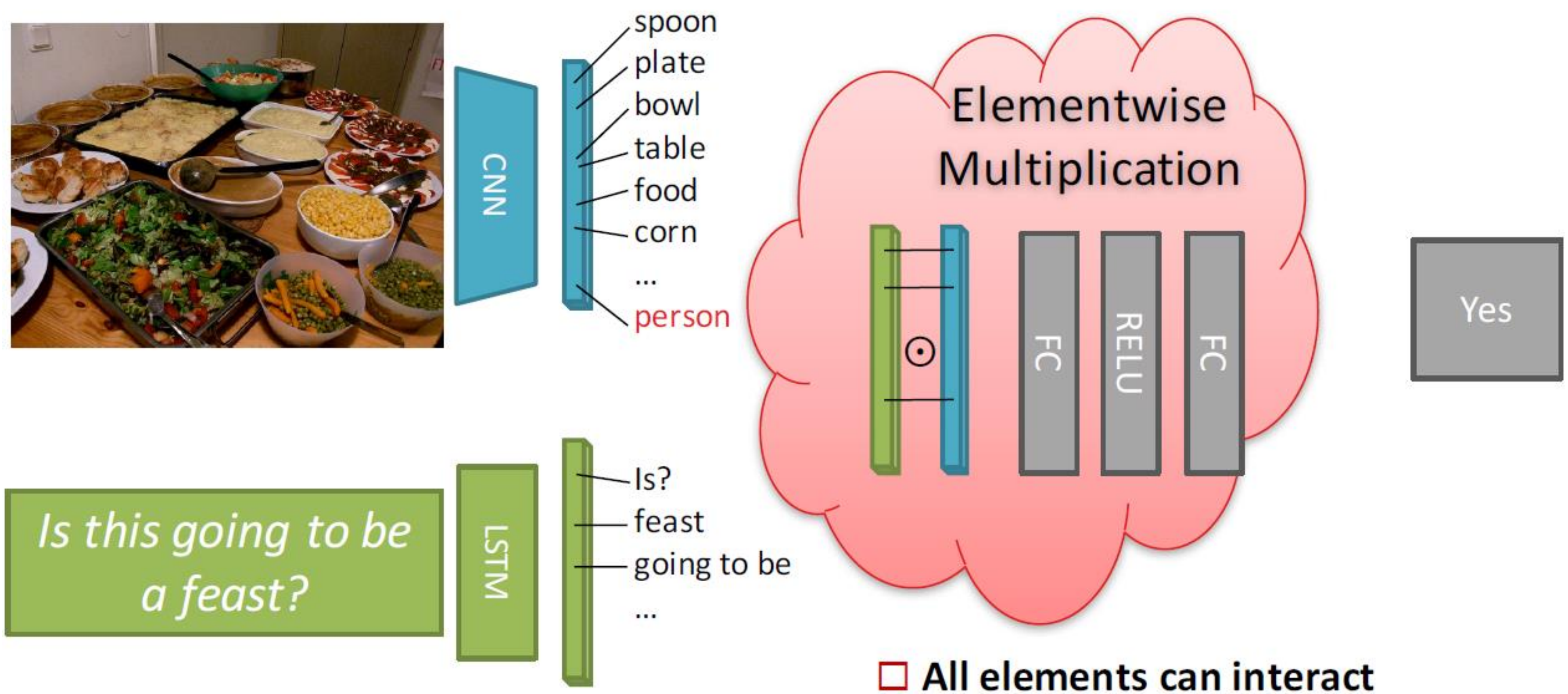


☑ **All elements can interact**

□ **Multiplicative interaction**

- Difficult to learn output classificatior

- 그저 합치는 방법!
- 하지만 상호작용이 없다.



□ **All elements can interact**

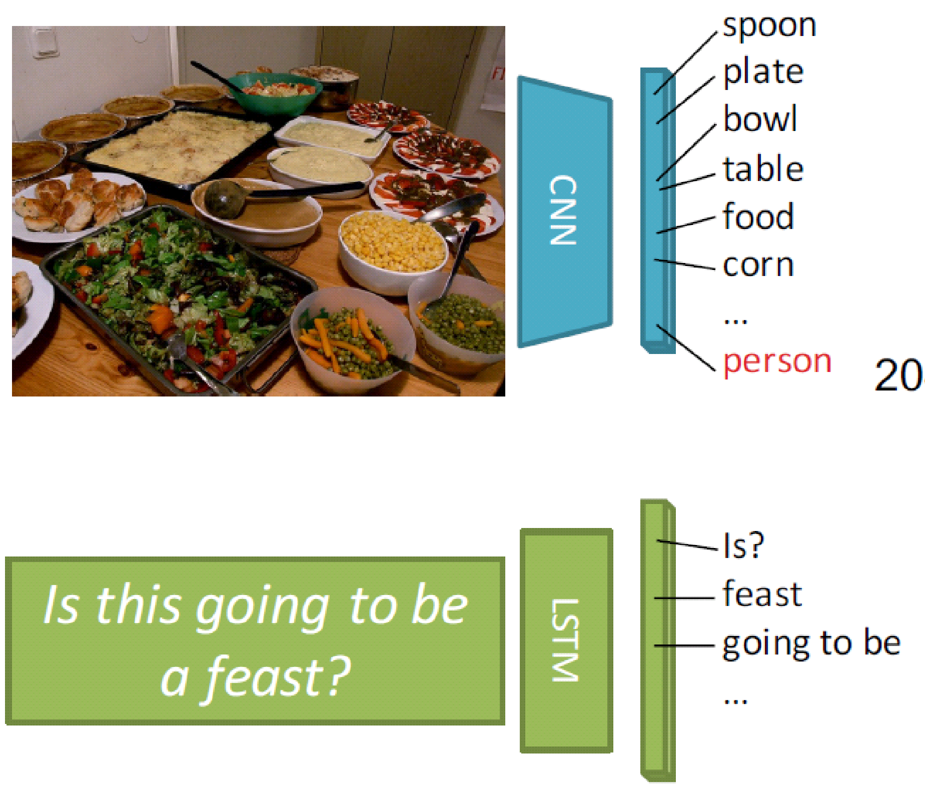
☑ **Multiplicative interaction**

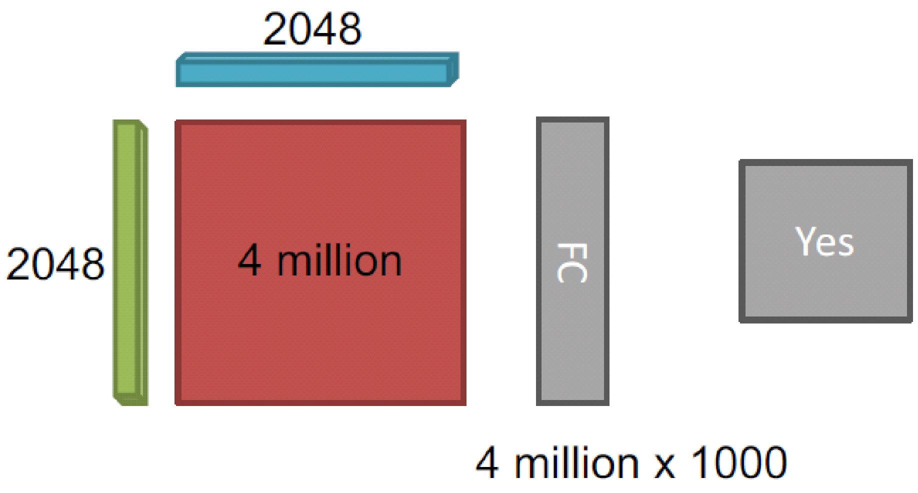
- Difficult to learn input embedding

- 같은 dimension vector가 있다면, 같은 위치를 곱한다.
- 각 위치가 같은 의미를 가질까 ??

Outer Product / Bilinear Pooling

[Lin ICCV 2015]





☑ **All elements can interact**

☑ **Multiplicative interaction**

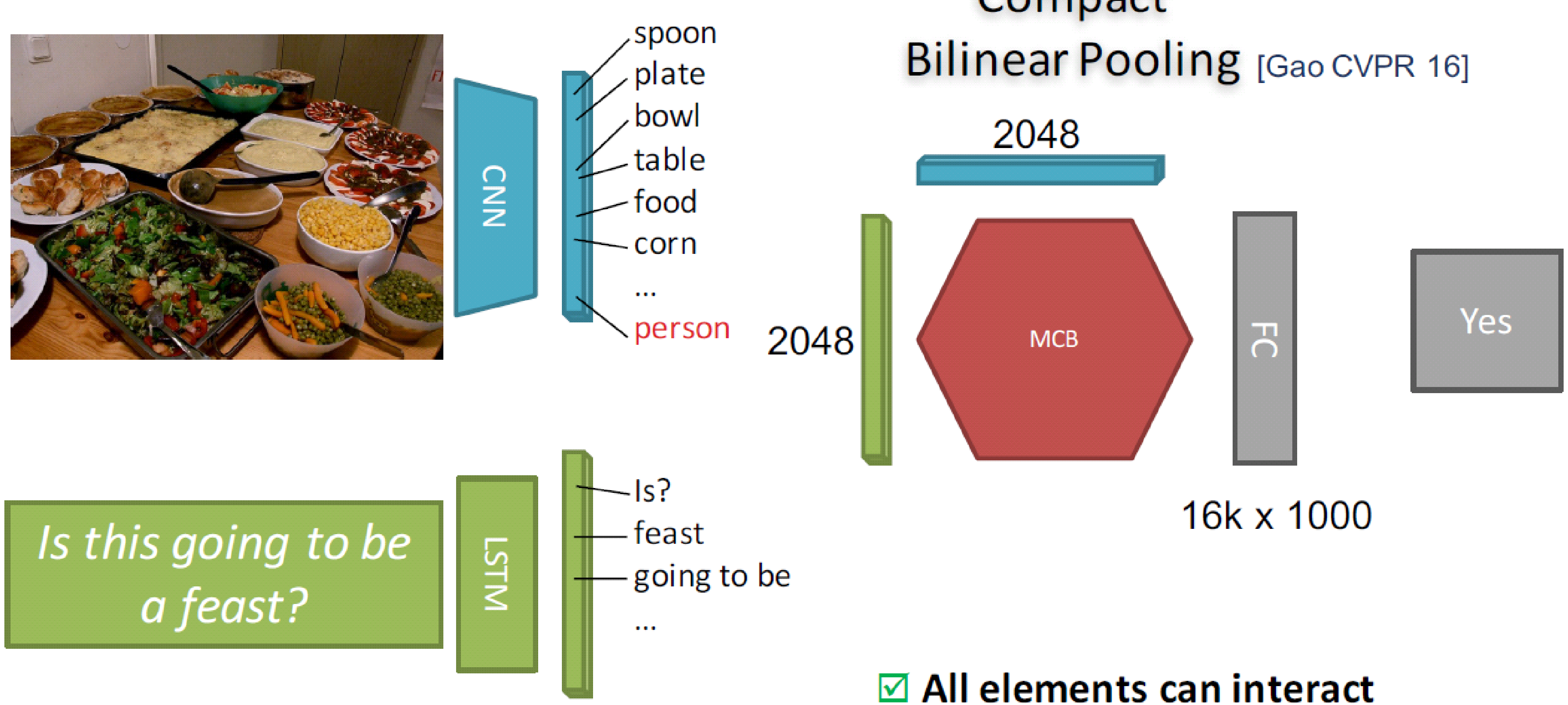
□ **High #activations & computation**

□ **High #parameters**

[Lin ICCV 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. ICCV 2015

- 두 벡터를 외적 시키는 방법
- Parameter수와 연산량이 너무 많아진다.

Compact Bilinear Pooling [Gao CVPR 16]



[ICLR Workshops 2016] Fine-grained pose prediction, normalization, and recognition N Zhang, E Shelhamer, Y Gao, T Darrell

[Gao CVPR 16] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CVPR 2016

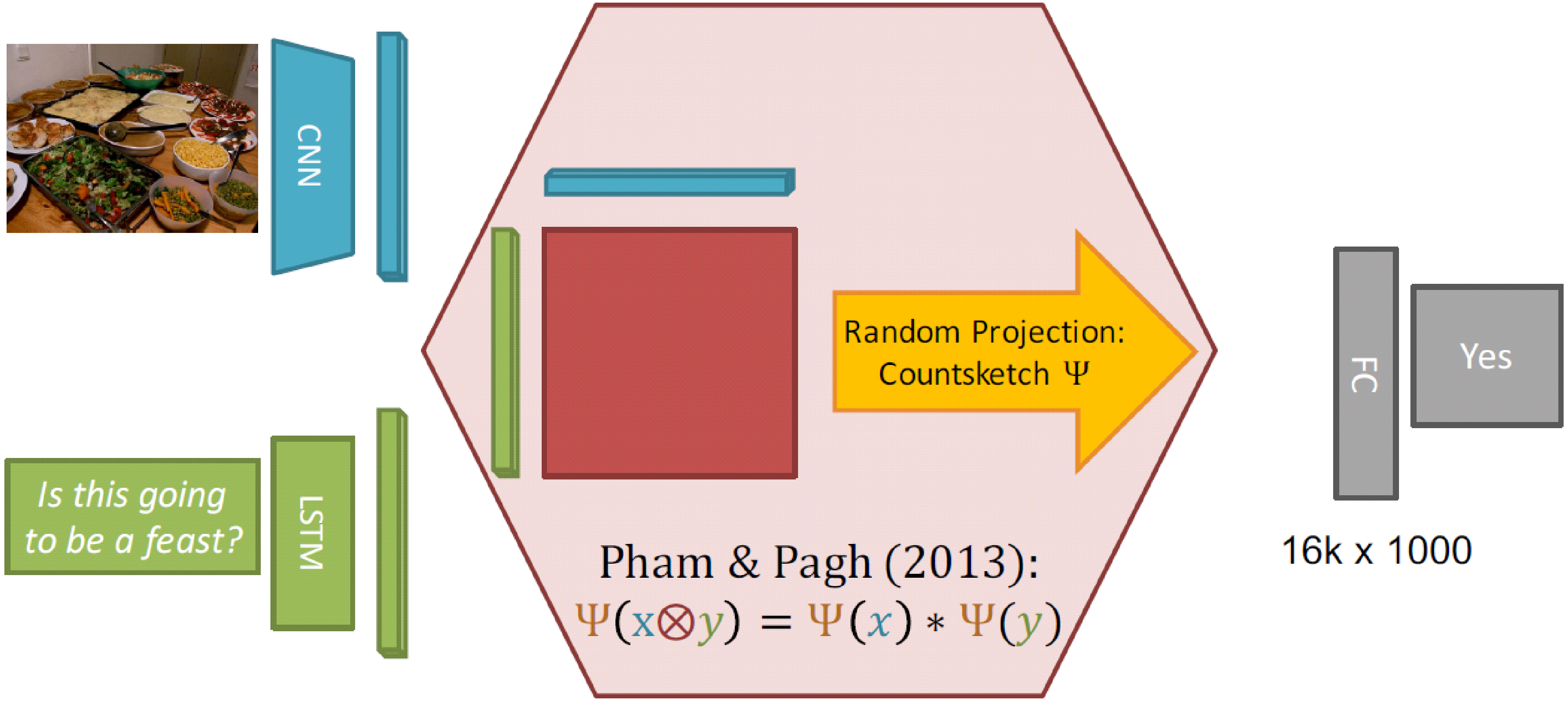
☑ **All elements can interact**

☑ **Multiplicative interaction**

☑ **Low #activations & computation**

☑ **Low #parameters**

- Multimodal Compact Bilinear Pooling
- Mapping을 통해 옮겨간 더 작은 vector를 이용해 분석을 해보자!


CNN
Is this going to be a feast?
LSTM
Random Projection: Countsketch Ψ
FC
Yes
16k x 1000

Pham & Pagh (2013):

$$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$$

☑ **All elements can interact**

☑ **Multiplicative interaction**

□ **Low #activations & computation**

☑ **Low #parameters**

[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.
[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013

두 벡터의 외적을 countsketch한 결과 -> Low parameters

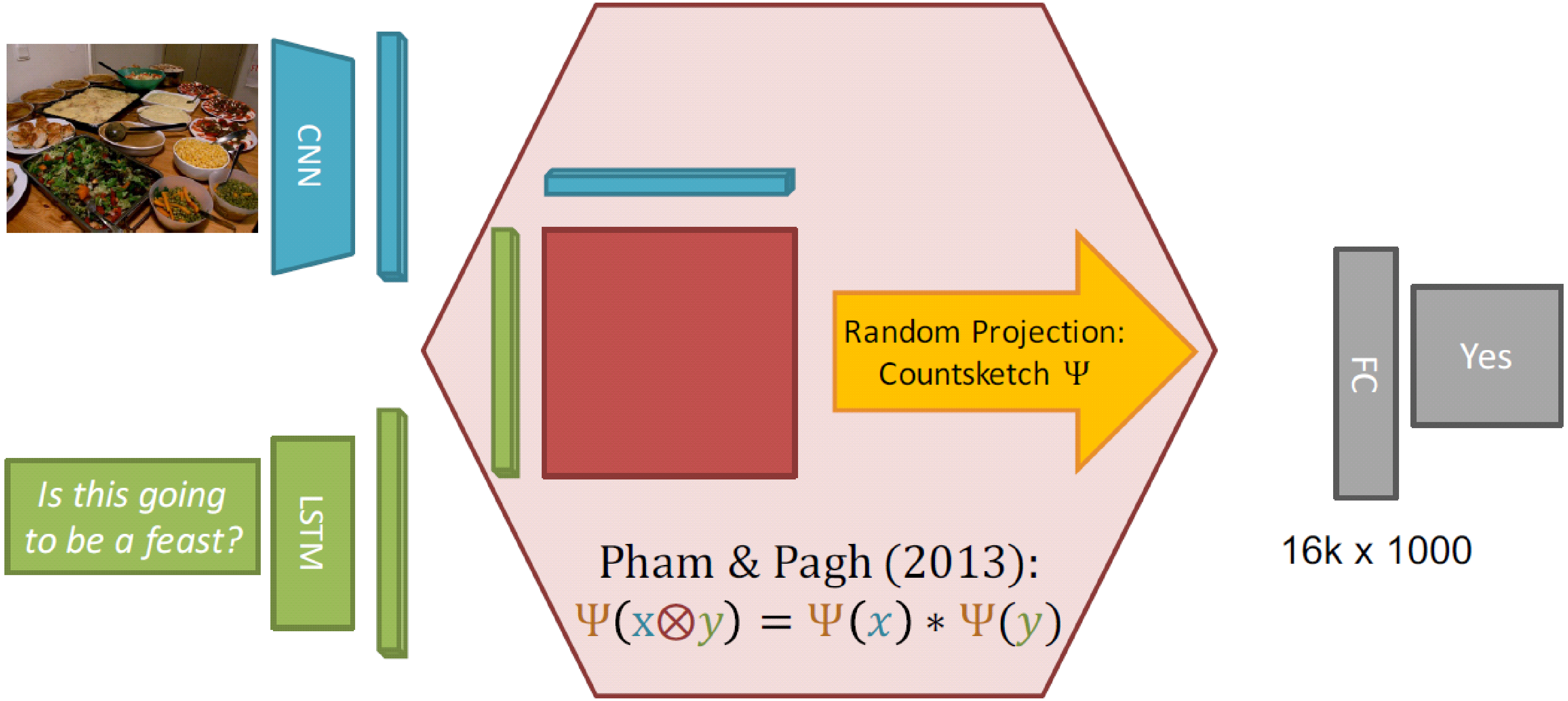
Ψ = Count sketch operation

Count sketch of $v = \begin{bmatrix}1\\5\\2\\3\\2\end{bmatrix}$ given $\boxed{s = \begin{bmatrix}1\\1\\-1\\1\\-1\end{bmatrix} \text{ and } h = \begin{bmatrix}1\\3\\2\\3\\2\end{bmatrix}}$ $(d = 3)$ can be computed by:

Randomly initialized !

$$y = \begin{bmatrix}1\\-2-2\\5+3\end{bmatrix} = \begin{bmatrix}1\\-4\\8\end{bmatrix}$$

s = {-1, 1}를, h = {1, …, d} (d는 변환 해줄 차원) 의 값을 랜덤하게 대입



Is this going to be a feast?

CNN

LSTM

Random Projection: Countsketch Ψ

FC

Yes

16k x 1000

Pham & Pagh (2013):

$$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$$

☑ **All elements can interact**

☑ **Multiplicative interaction**

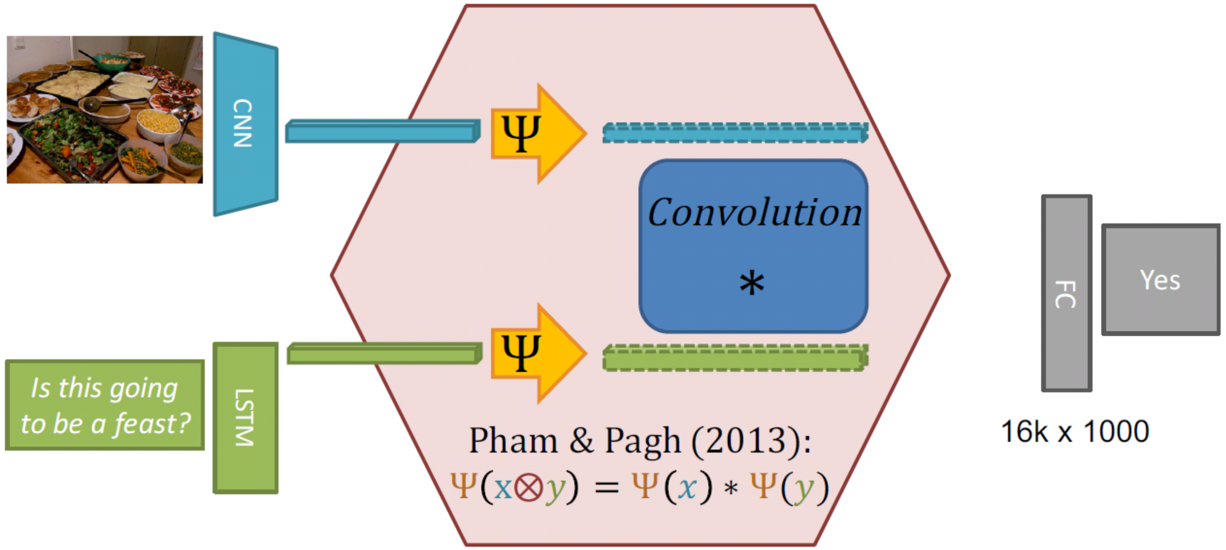
□ **Low #activations & computation**

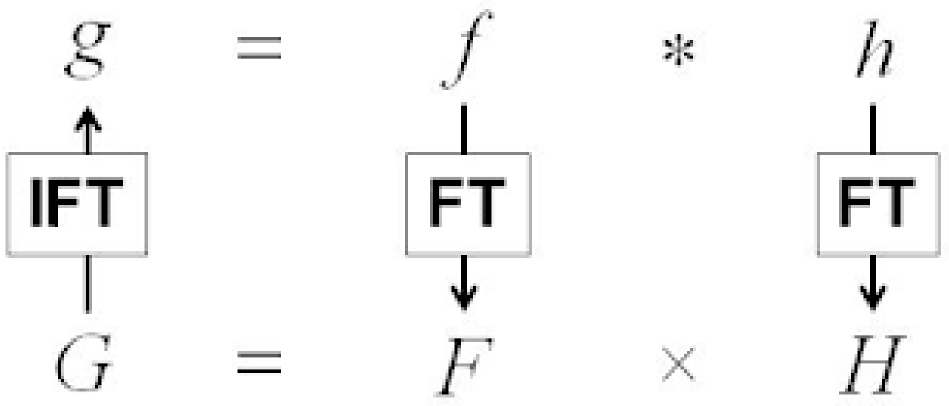
☑ **Low #parameters**

[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.

[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013

두 벡터의 외적을 countsketch한 결과 -> Low parameters
두 벡터의 countsketch를 convolution한 결과 -> Low computation & parameters


CNN
Is this going to be a feast?
LSTM
Ψ
Ψ
Convolution
*
Pham & Pagh (2013):
$\Psi(x \otimes y) = \Psi(x) * \Psi(y)$
FC
Yes
16k x 1000



☑ **All elements can interact**

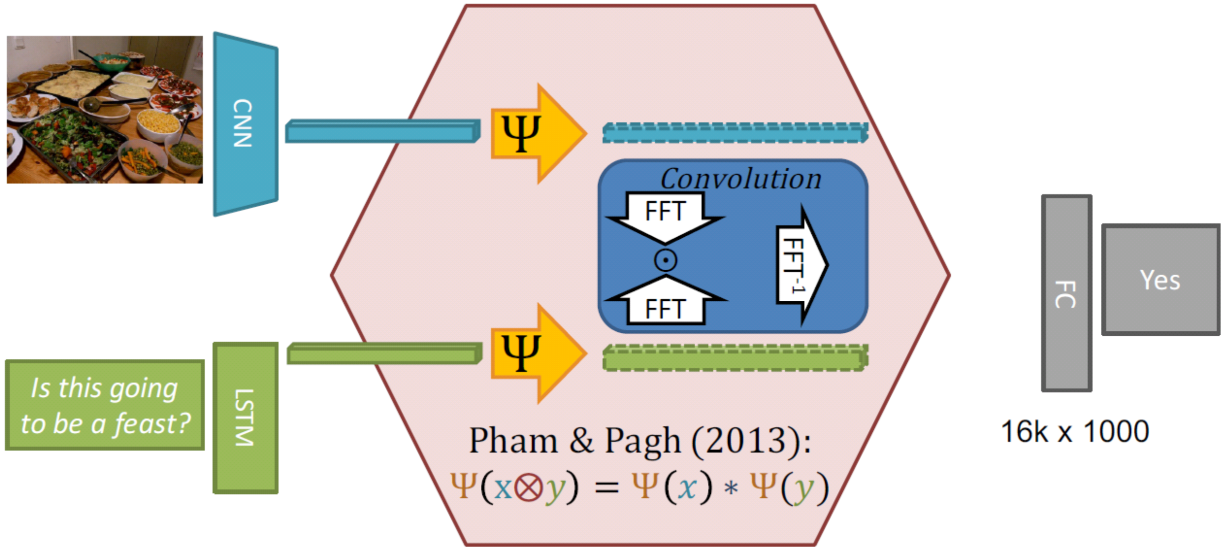
☑ **Multiplicative interaction**

☑ **Low #activations & computation**

☑ **Low #parameters**

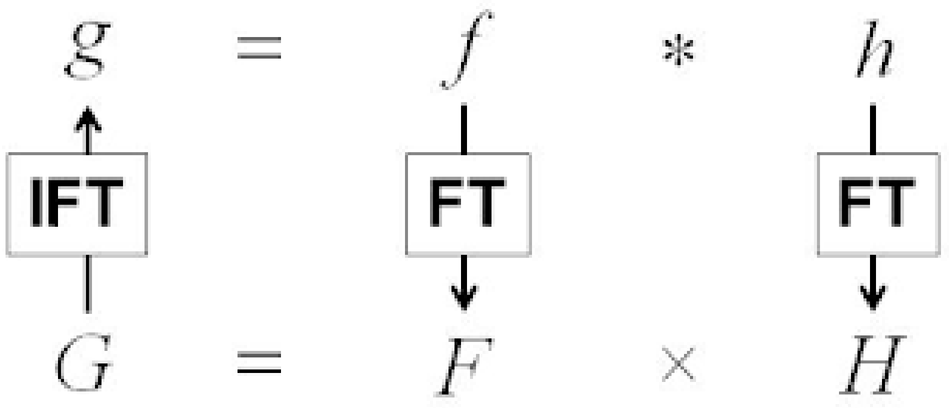
[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.
[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013

두 벡터(f, h)의 convolution은 두 벡터를 Fourier Transform한 뒤(F, H),
이 둘을 곱하고 Inverse Fourier Transform 한 결과와 같다.
연산을 더 빠르게 하기 위해 FFT를 사용



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

[Countsketch] M. Charikar, K. Chen, M. Farach-Colton. Finding frequent items in data streams. Automata, languages and programming 2002.
[Pham&Pagh 13] N. Pham, R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. KDD 2013



두 벡터(f, h)의 convolution은 두 벡터를 Fourier Transform한 뒤(F, H),
이 둘을 곱하고 Inverse Fourier Transform 한 결과와 같다.
연산을 더 빠르게 하기 위해 FFT를 사용

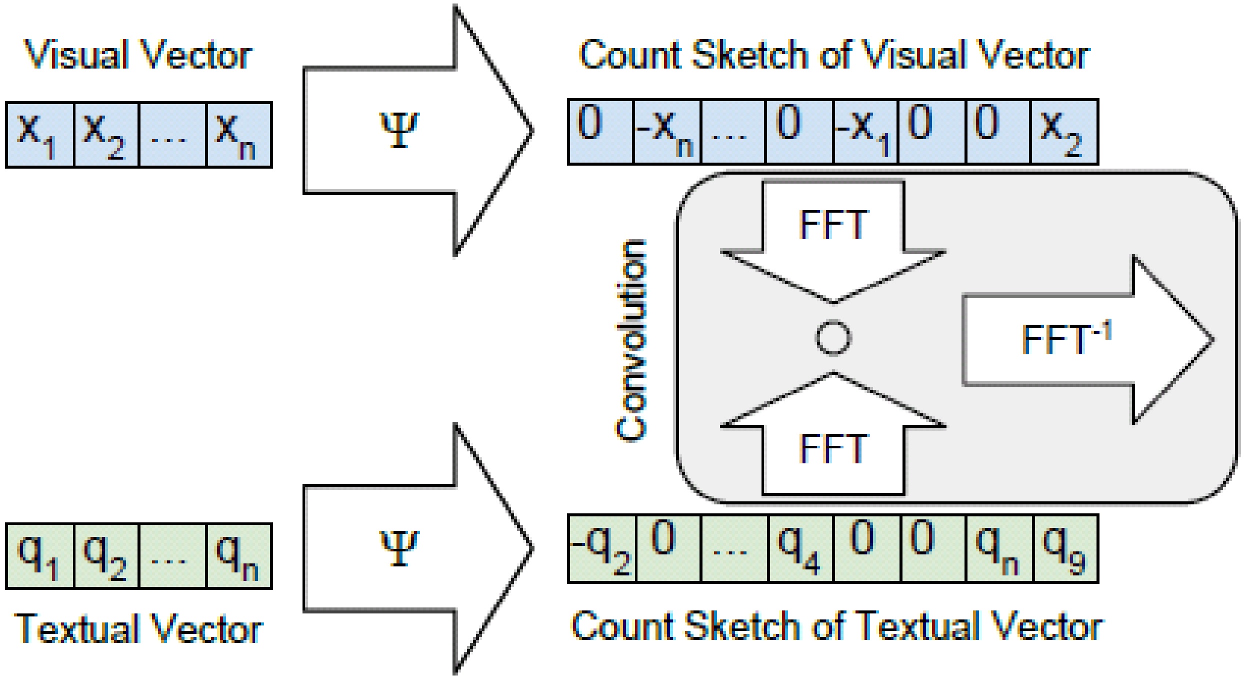
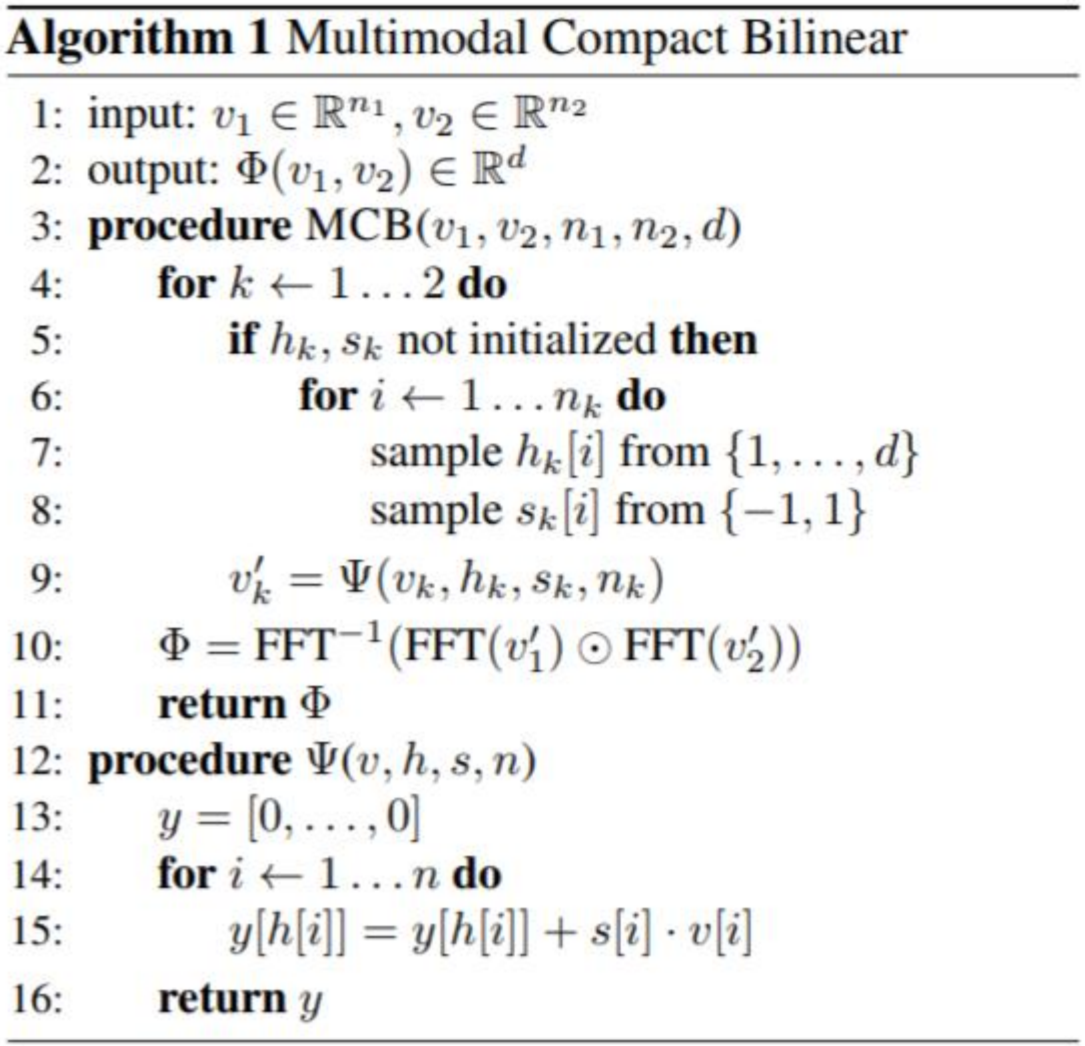

Visual Vector
x_1 x_2 ... x_n
Ψ
Count Sketch of Visual Vector
0 $-x_n$ ... 0 $-x_1$ 0 0 x_2
Convolution
FFT
○
FFT
FFT^{-1}
q_1 q_2 ... q_n
Textual Vector
Ψ
$-q_2$ 0 ... q_4 0 0 q_n q_9
Count Sketch of Textual Vector

Figure 2: Multimodal Compact Bilinear Pooling (MCB)



Algorithm 1 Multimodal Compact Bilinear

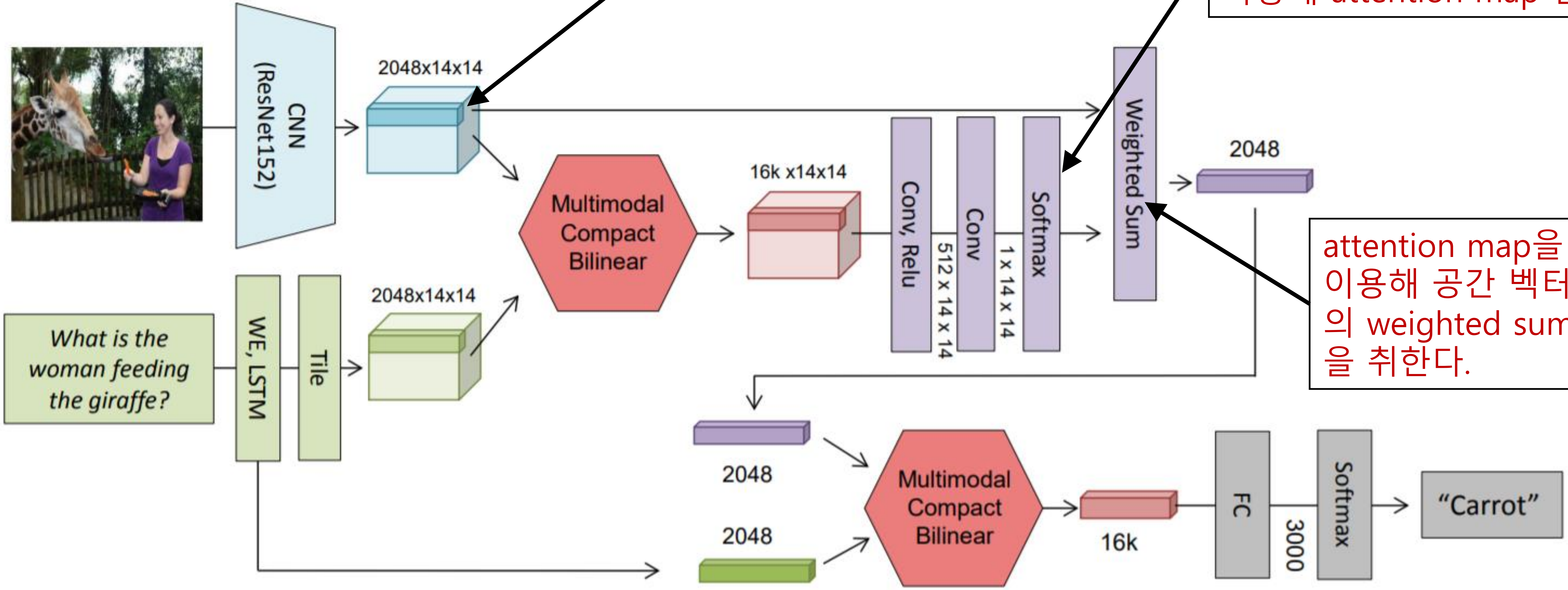
```
input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
procedure MCB($v_1, v_2, n_1, n_2, d$)
    for $k \leftarrow 1 \dots 2$ do
        if $h_k, s_k$ not initialized then
            for $i \leftarrow 1 \dots n_k$ do
                sample $h_k[i]$ from $\{1, \dots, d\}$
                sample $s_k[i]$ from $\{-1, 1\}$
        $v'_k = \Psi(v_k, h_k, s_k, n_k)$
    $\Phi = \mathrm{FFT}^{-1}(\mathrm{FFT}(v'_1) \odot \mathrm{FFT}(v'_2))$
    return $\Phi$
procedure $\Psi(v, h, s, n)$
    $y = [0, \dots, 0]$
    for $i \leftarrow 1 \dots n$ do
        $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
    return $y$
```

h[i] <- [1,...,d] (i=변환 전 차원 수, d=변환 후 차원 수)
s[i] <- [-1, 1]
v[i] <- x1, x2, ...

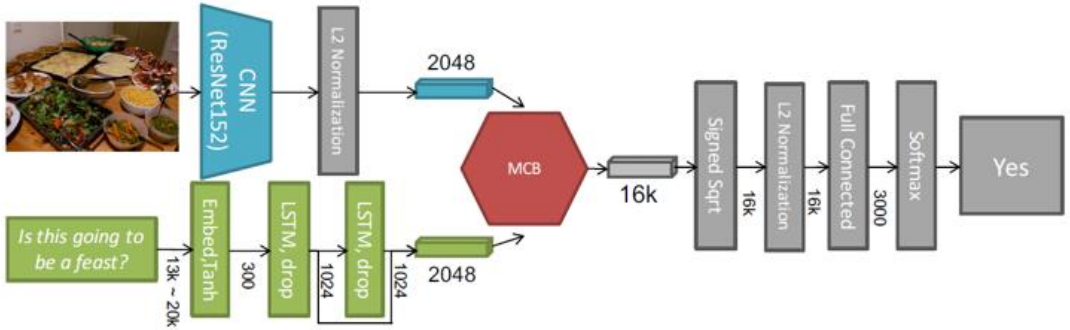
MCB with Attention

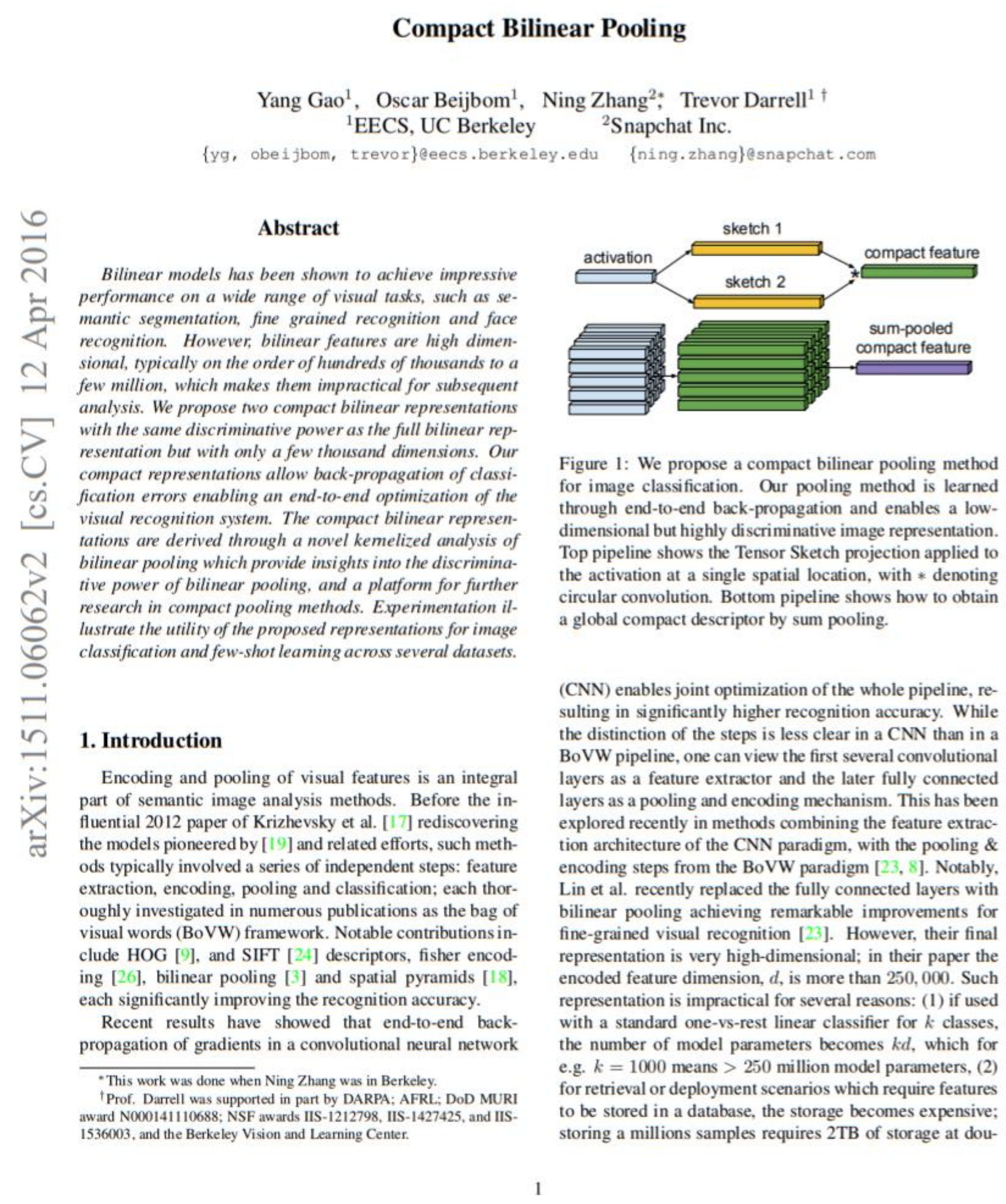
이미지를 14x14로 나눈 뒤, 각 grid의 feature를 2048개에 넣는다

각 grid의 attention weight를 예측하기 위해 softmax를 사용해 attention map 만듦.



attention map을 이용해 공간 벡터의 weighted sum을 취한다.





Compact Bilinear Pooling

Yang Gao[1], Oscar Beijbom[1], Ning Zhang[2]*, Trevor Darrell[1] †
[1]EECS, UC Berkeley [2]Snapchat Inc.
{yg, obeijbom, trevor}@eecs.berkeley.edu {ning.zhang}@snapchat.com

Abstract

Bilinear models has been shown to achieve impressive performance on a wide range of visual tasks, such as semantic segmentation, fine grained recognition and face recognition. However, bilinear features are high dimensional, typically on the order of hundreds of thousands to a few million, which makes them impractical for subsequent analysis. We propose two compact bilinear representations with the same discriminative power as the full bilinear representation but with only a few thousand dimensions. Our compact representations allow back-propagation of classification errors enabling an end-to-end optimization of the visual recognition system. The compact bilinear representations are derived through a novel kernelized analysis of bilinear pooling which provide insights into the discriminative power of bilinear pooling, and a platform for further research in compact pooling methods. Experimentation illustrate the utility of the proposed representations for image classification and few-shot learning across several datasets.

Figure 1: We propose a compact bilinear pooling method for image classification. Our pooling method is learned through end-to-end back-propagation and enables a low-dimensional but highly discriminative image representation. Top pipeline shows the Tensor Sketch projection applied to the activation at a single spatial location, with $*$ denoting circular convolution. Bottom pipeline shows how to obtain a global compact descriptor by sum pooling.

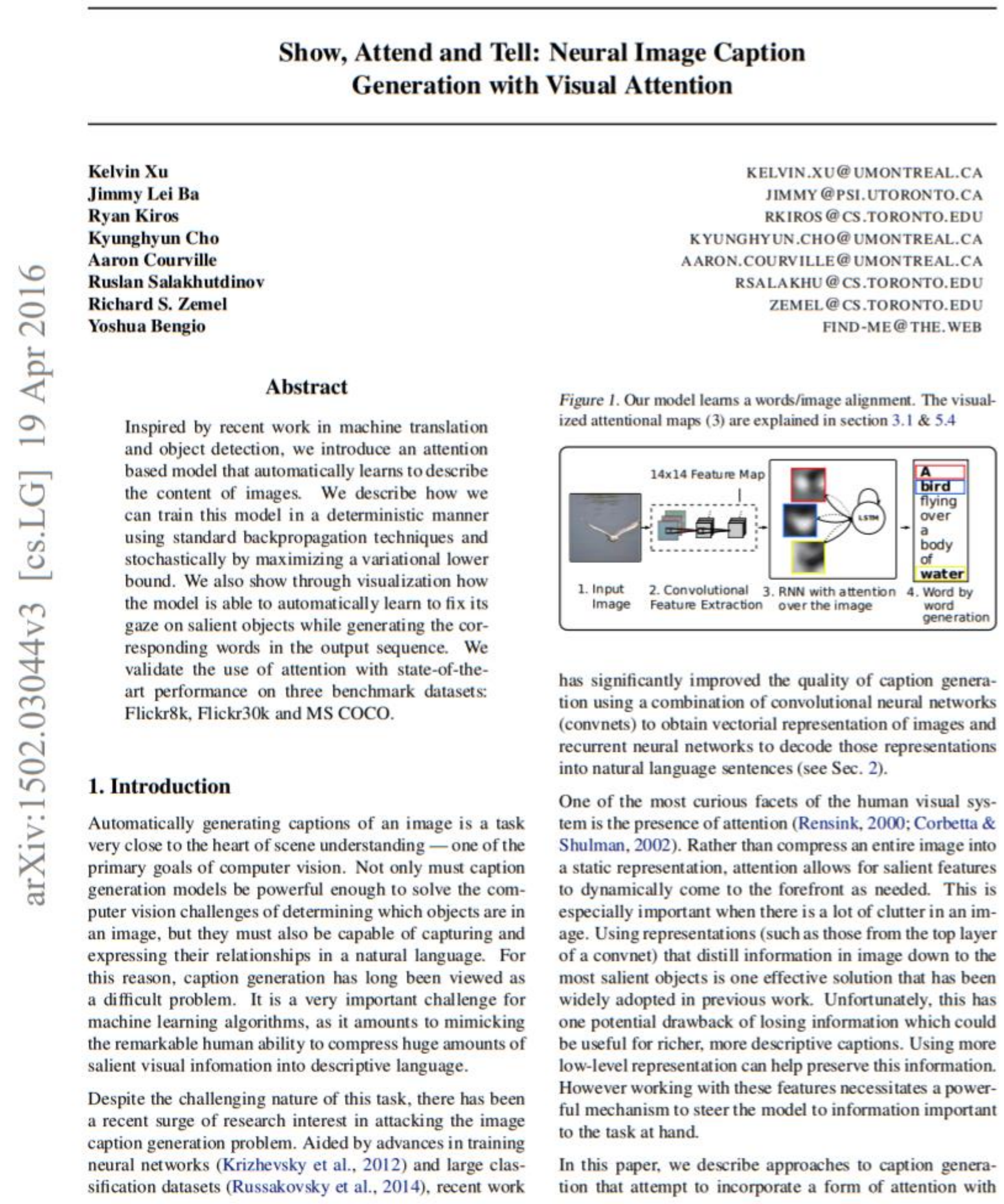
1. Introduction

Encoding and pooling of visual features is an integral part of semantic image analysis methods. Before the influential 2012 paper of Krizhevsky et al. [17] rediscovering the models pioneered by [19] and related efforts, such methods typically involved a series of independent steps: feature extraction, encoding, pooling and classification; each thoroughly investigated in numerous publications as the bag of visual words (BoVW) framework. Notable contributions include HOG [9], and SIFT [24] descriptors, fisher encoding [26], bilinear pooling [3] and spatial pyramids [18], each significantly improving the recognition accuracy.

Recent results have showed that end-to-end back-propagation of gradients in a convolutional neural network (CNN) enables joint optimization of the whole pipeline, resulting in significantly higher recognition accuracy. While the distinction of the steps is less clear in a CNN than in a BoVW pipeline, one can view the first several convolutional layers as a feature extractor and the later fully connected layers as a pooling and encoding mechanism. This has been explored recently in methods combining the feature extraction architecture of the CNN paradigm, with the pooling & encoding steps from the BoVW paradigm [23, 8]. Notably, Lin et al. recently replaced the fully connected layers with bilinear pooling achieving remarkable improvements for fine-grained visual recognition [23]. However, their final representation is very high-dimensional; in their paper the encoded feature dimension, d, is more than $250,000$. Such representation is impractical for several reasons: (1) if used with a standard one-vs-rest linear classifier for k classes, the number of model parameters becomes kd, which for e.g. $k = 1000$ means > 250 million model parameters, (2) for retrieval or deployment scenarios which require features to be stored in a database, the storage becomes expensive; storing a millions samples requires 2TB of storage at dou-

[*] This work was done when Ning Zhang was in Berkeley.

[†] Prof. Darrell was supported in part by DARPA; AFRL; DoD MURI award N000141110688; NSF awards IIS-1212798, IIS-1427425, and IIS-1536003, and the Berkeley Vision and Learning Center.



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba JIMMY@PSI.UTORONTO.CA
Ryan Kiros RKIROS@CS.TORONTO.EDU
Kyunghyun Cho KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
Yoshua Bengio FIND-ME@THE.WEB

Abstract

Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We validate the use of attention with state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO.

1. Introduction

Automatically generating captions of an image is a task very close to the heart of scene understanding — one of the primary goals of computer vision. Not only must caption generation models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, caption generation has long been viewed as a difficult problem. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual infomation into descriptive language.

Despite the challenging nature of this task, there has been a recent surge of research interest in attacking the image caption generation problem. Aided by advances in training neural networks (Krizhevsky et al., 2012) and large classification datasets (Russakovsky et al., 2014), recent work has significantly improved the quality of caption generation using a combination of convolutional neural networks (convnets) to obtain vectorial representation of images and recurrent neural networks to decode those representations into natural language sentences (see Sec. 2).

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in section 3.1 & 5.4

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

One of the most curious facets of the human visual system is the presence of attention (Rensink, 2000; Corbetta & Shulman, 2002). Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image. Using representations (such as those from the top layer of a convnet) that distill information in image down to the most salient objects is one effective solution that has been widely adopted in previous work. Unfortunately, this has one potential drawback of losing information which could be useful for richer, more descriptive captions. Using more low-level representation can help preserve this information. However working with these features necessitates a powerful mechanism to steer the model to information important to the task at hand.

In this paper, we describe approaches to caption generation that attempt to incorporate a form of attention with

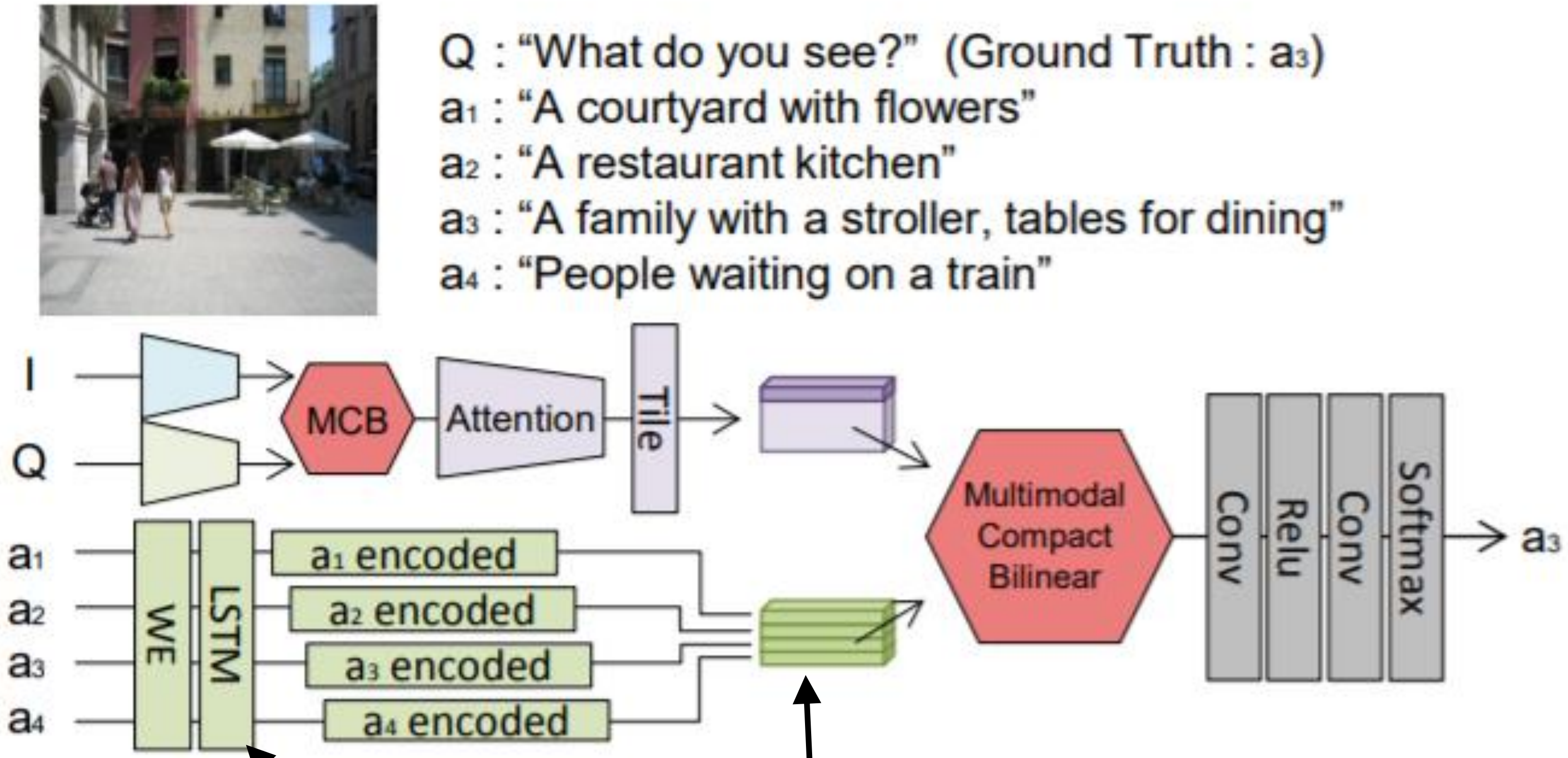

Q : "What do you see?" (Ground Truth : a3)
a1 : "A courtyard with flowers"
a2 : "A restaurant kitchen"
a3 : "A family with a stroller, tables for dining"
a4 : "People waiting on a train"
I
Q
MCB
Attention
Tile
a1
a2
a3
a4
WE
LSTM
a1 encoded
a2 encoded
a3 encoded
a4 encoded
Multimodal Compact Bilinear
Conv
Relu
Conv
Softmax
a3

Figure 4: Our architecture for VQA: MCB with Attention and Answer Encoding

다수의 다중 길이 답이 필요한 경우는, weight가 공유되는 WE, LSTM 사용

결과 embedding은 classification vector로 변하며, 이 벡터의 차원수는 답의 개수와 같다.

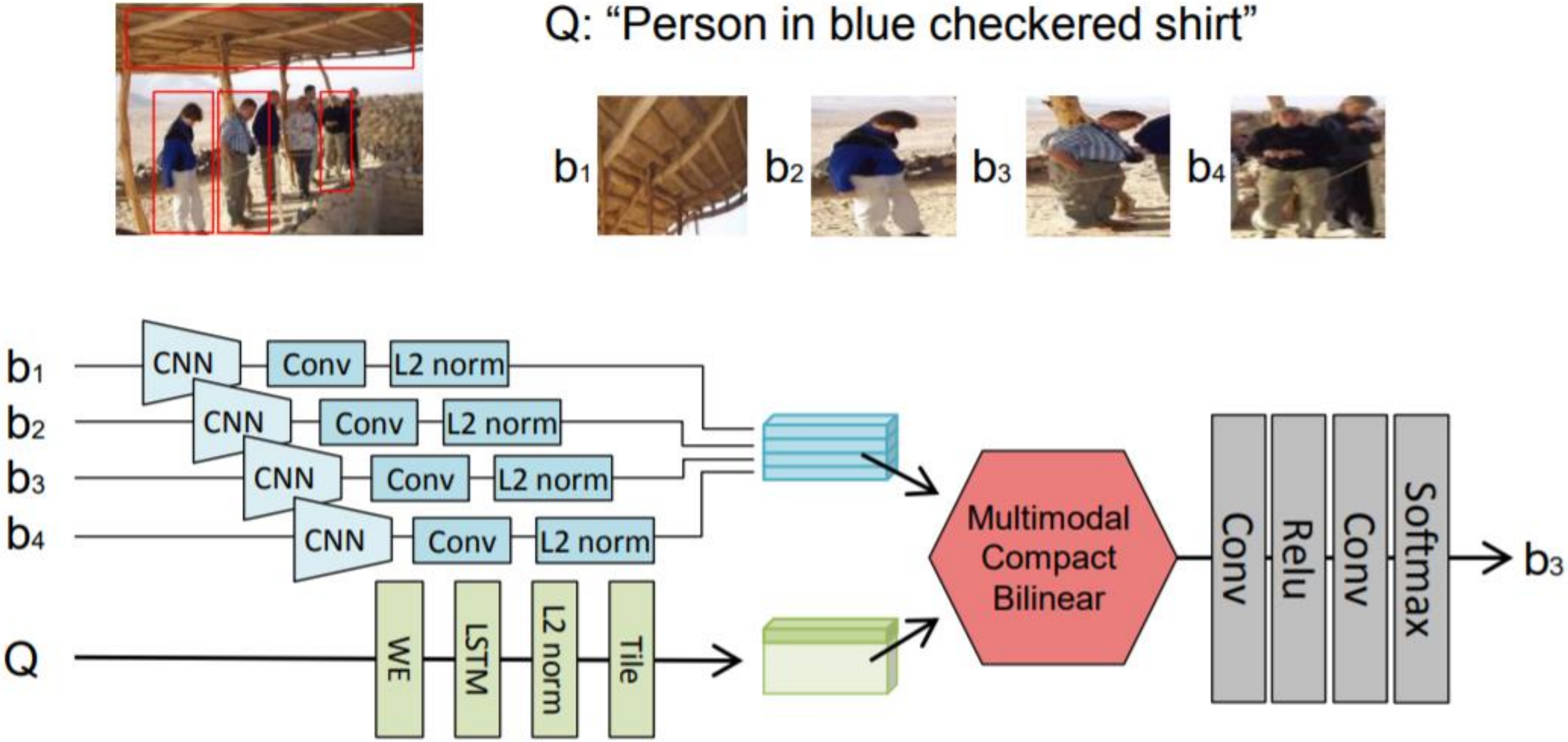


Figure 5: Our Architecture for Grounding with MCB (Sec. 3.3)

Datasets

1. VQA

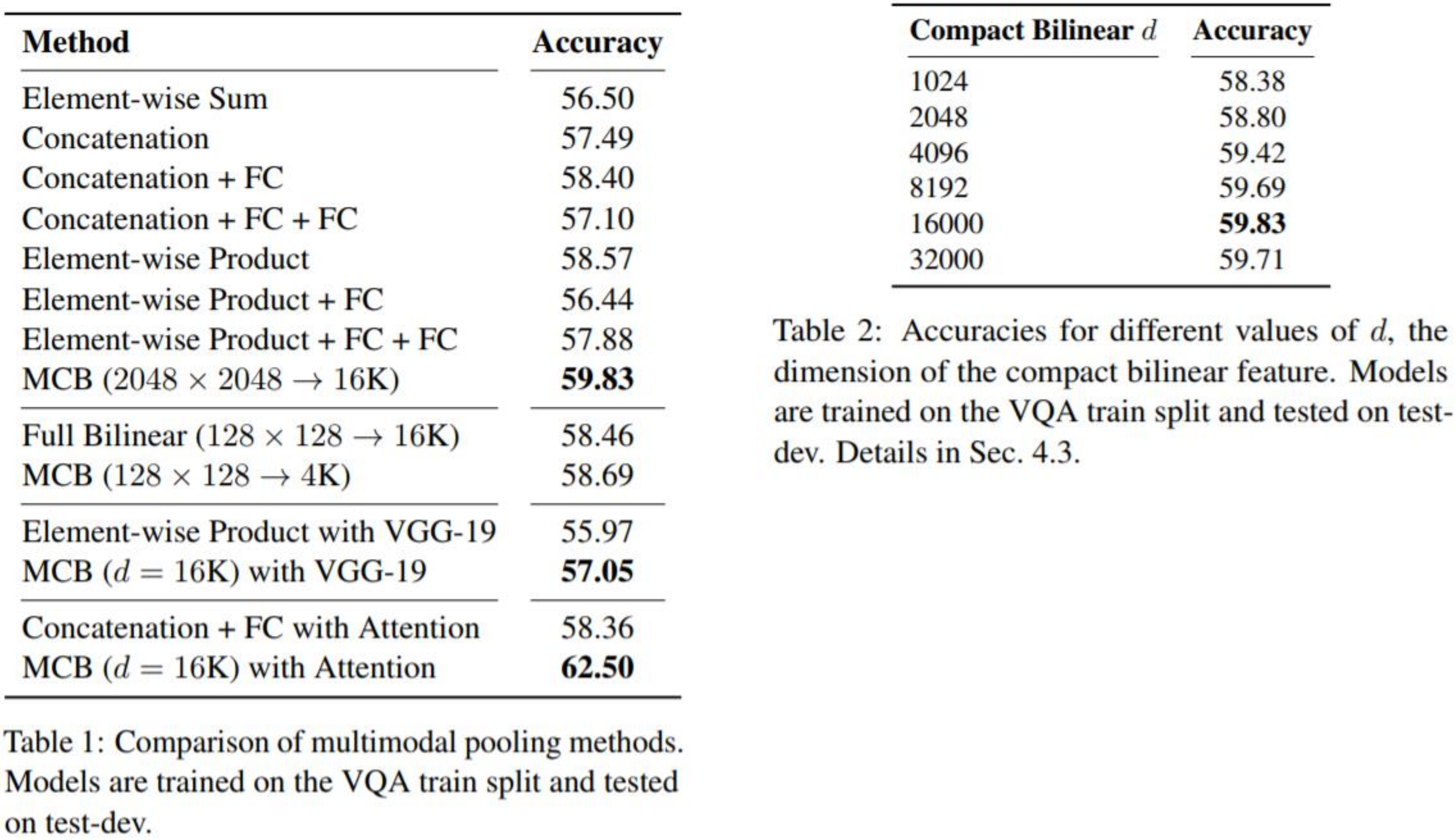
- 200,000 MSCOCO images
- 3 question per image and 10 answer per question
- Train 80K, Validation 40K, Test 80K

2. Visual Genome

- 108,249 images form YFCC100M + MSCOCO
- Average of 17 question-answer pairs per each image
- Consist of 6W question types (what, where, when, who, why, how)
- In this paper, they remove “a” “the” “it is” …

3. Visual7W

- 7W question types (6W + which)
- 47,300 images from MSCOCO / 139,868 QA pairs
- But, only evaluate 6W questions
- Multiple-choice format, each question comes with four answer candidates

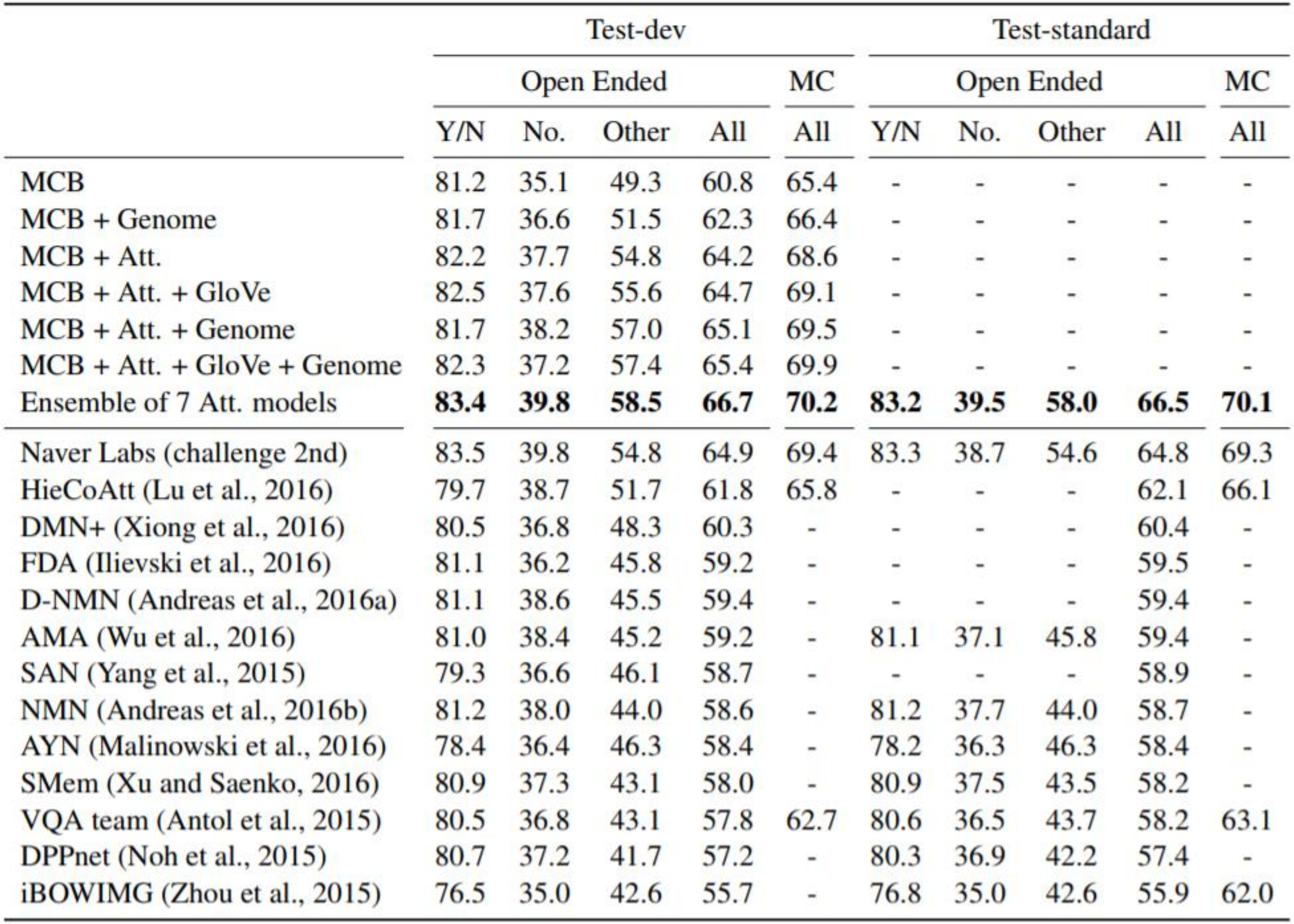


Method	**Accuracy**
Element-wise Sum	56.50
Concatenation	57.49
Concatenation + FC	58.40
Concatenation + FC + FC	57.10
Element-wise Product	58.57
Element-wise Product + FC	56.44
Element-wise Product + FC + FC	57.88
MCB (2048 × 2048 → 16K)	**59.83**
Full Bilinear (128 × 128 → 16K)	58.46
MCB (128 × 128 → 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**

Table 1: Comparison of multimodal pooling methods. Models are trained on the VQA train split and tested on test-dev.

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

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.

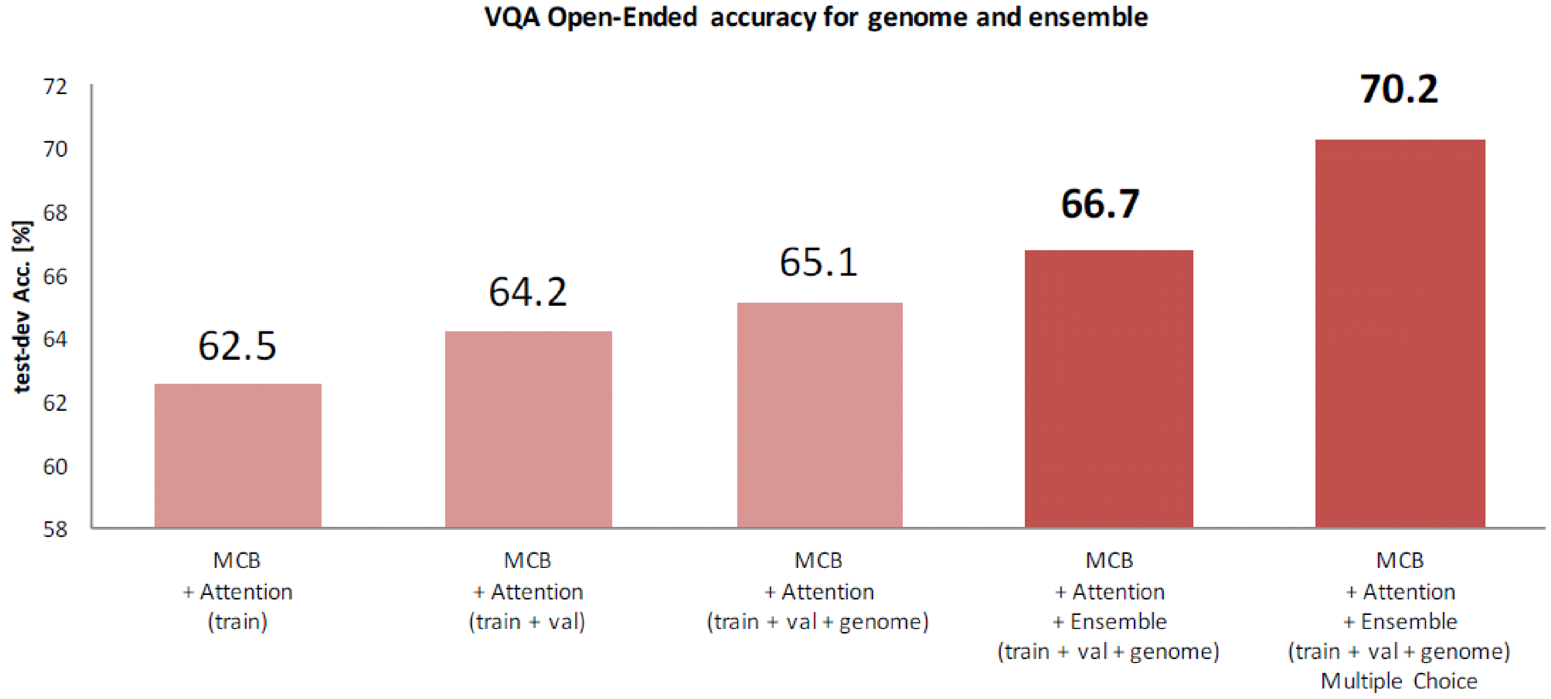


	Test-dev					Test-standard				
	Open Ended				MC	Open Ended				MC
	Y/N	No.	Other	All	All	Y/N	No.	Other	All	All
MCB	81.2	35.1	49.3	60.8	65.4	-	-	-	-	-
MCB + Genome	81.7	36.6	51.5	62.3	66.4	-	-	-	-	-
MCB + Att.	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
MCB + Att. + GloVe	82.5	37.6	55.6	64.7	69.1	-	-	-	-	-
MCB + 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-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, 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

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.

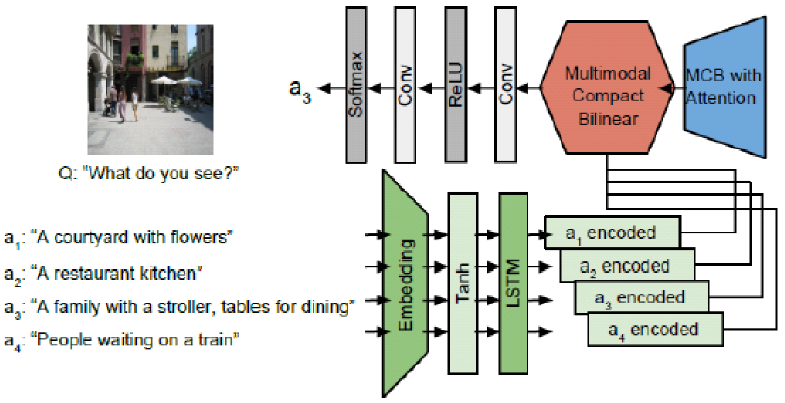
Results

- Data Augmentation
  - VQA data from Visual Genome Dataset
    - Additional 1M Question and answer pairs
    - Removed articles, Single word answer
- Ensembles
  - Average the output of Softmax over models



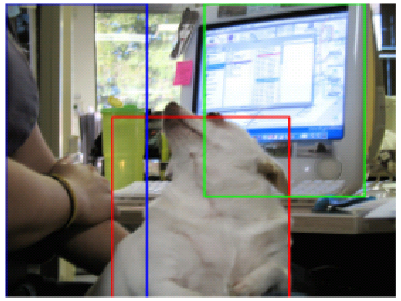
Results

- Visual 7w (Multiple Choice)


Q: "What do you see?"
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"
a_3
Softmax
Conv
ReLU
Conv
Multimodal Compact Bilinear
MCB with Attention
Embedding
Tanh
LSTM
a_1 encoded
a_2 encoded
a_3 encoded
a_4 encoded

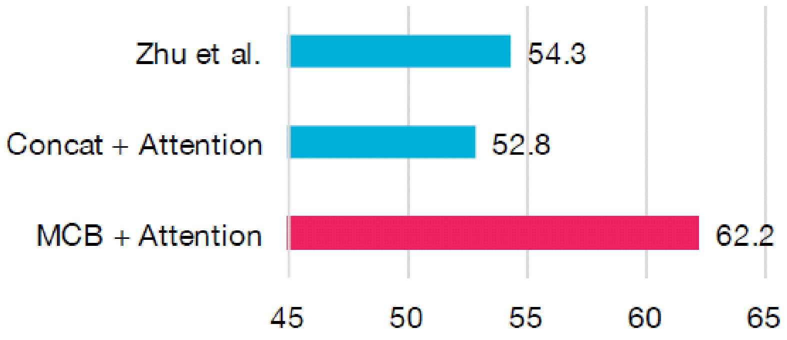
Our architecture for Visual 7w : MCB with Attention and Answer Encoding.

- Visual Grounding

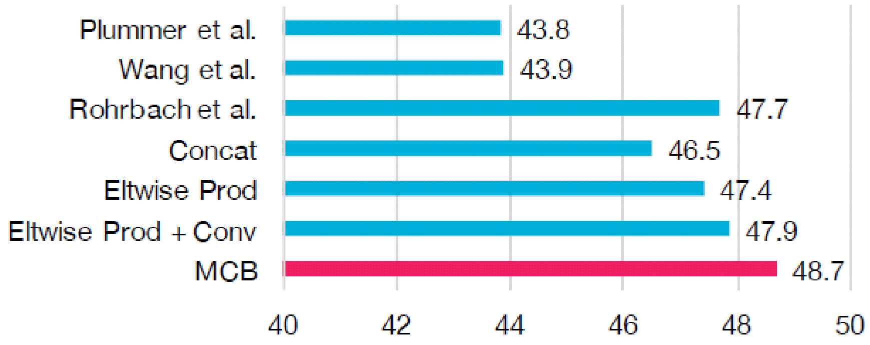


A dog distracts his owner from working at her computer.

Accuracy on Visual7W



Accuracy on Flickr30k Entities

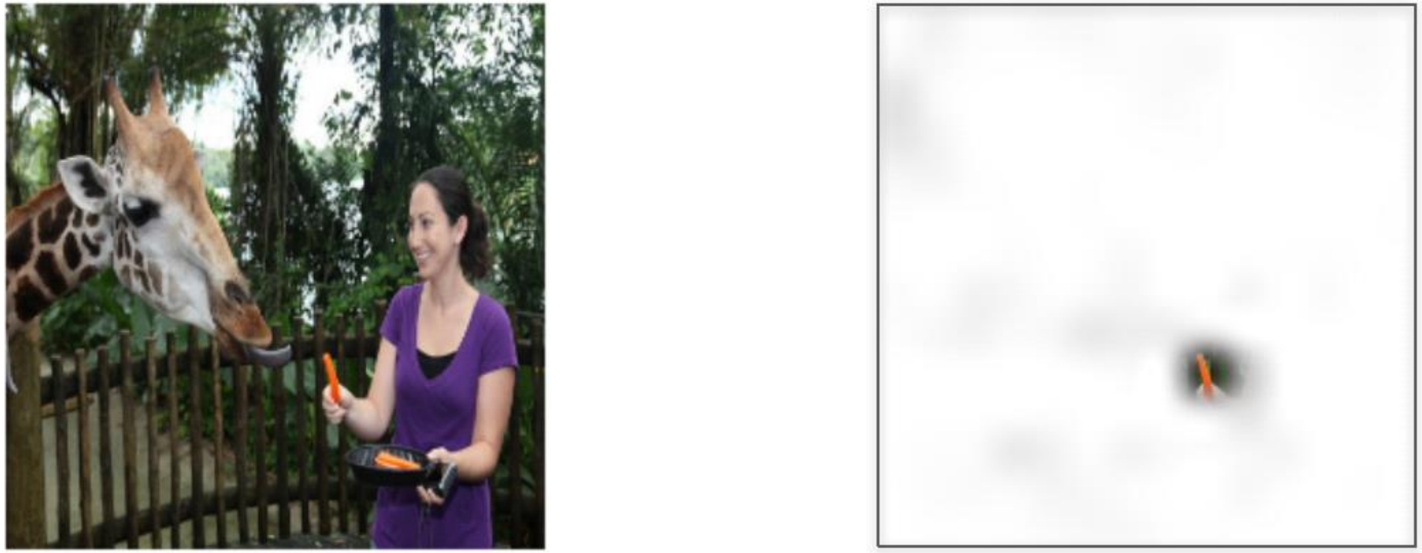


Results

What is the woman feeding the giraffe?

Carrot

[Groundtruth: Carrot]

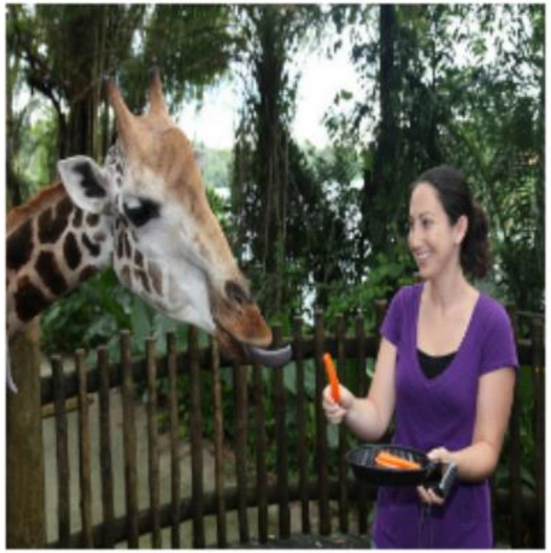


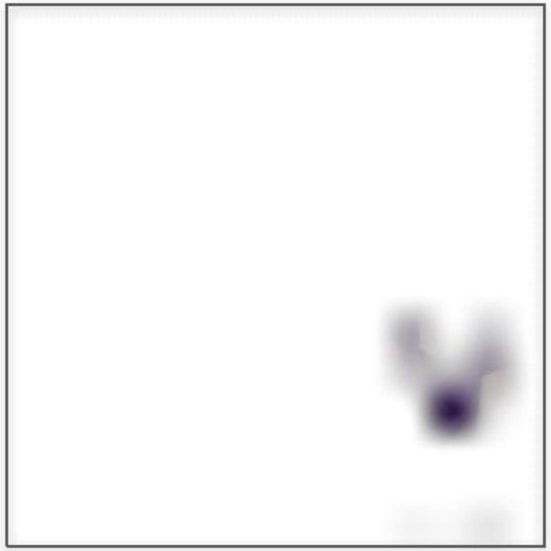
Results

What color is her shirt?

Purple

[Groundtruth: Purple]





Results

What is her hairstyle for the picture?

Ponytail

[Groundtruth: Ponytail]