DMQA Open Seminar

Multimodal Representation Learning: How to Narrow Heterogeneity gap

2022.07.29

Data Mining & Quality Analytics Lab.

고은성

발표자 소개

❖ 고은성

- 고려대학교 산업경영공학과
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2021.03 ~)

❖ Research Interest

- Deep Learning
- Self-Supervised Learning
- Multimodal Learning

❖ Contact

- E-mail : esko328@korea.ac.kr

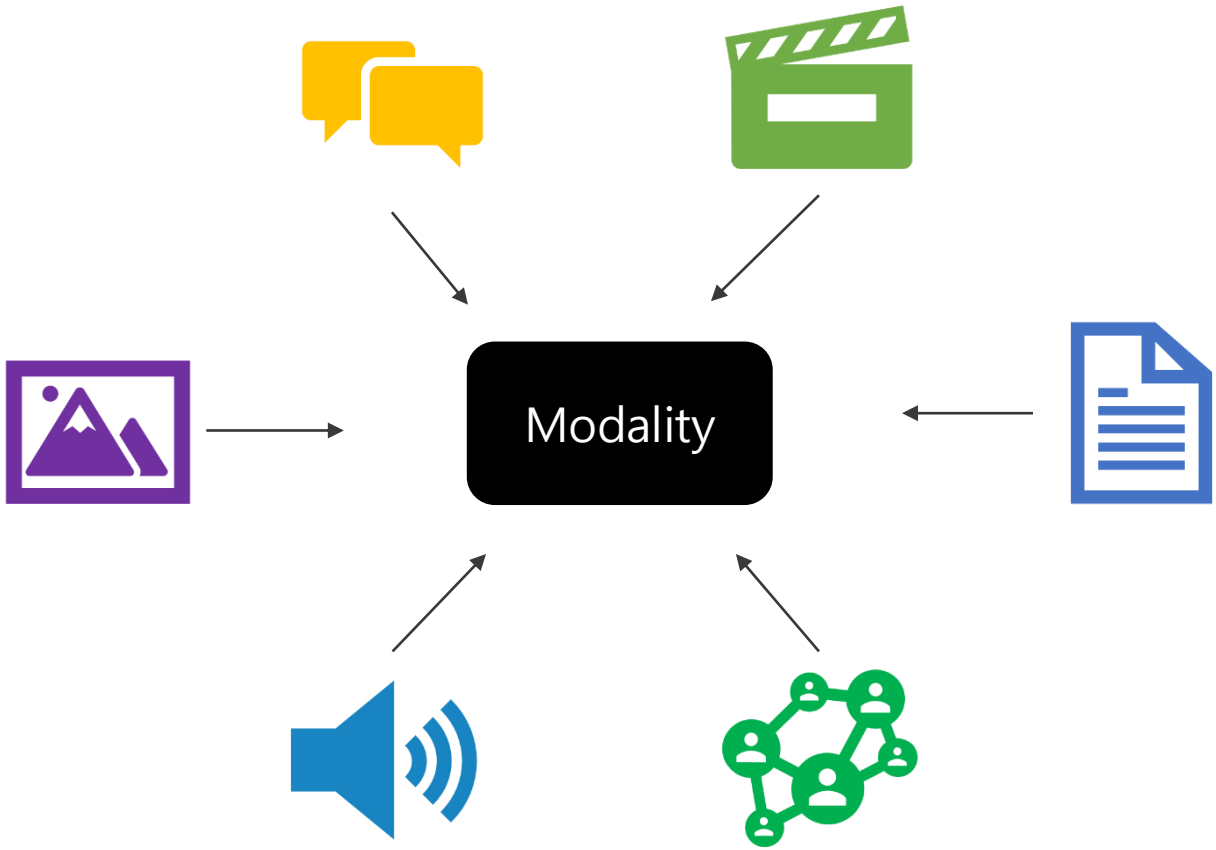
목차

Multimodal Representation Learning ?

Introduction

❖ Modality

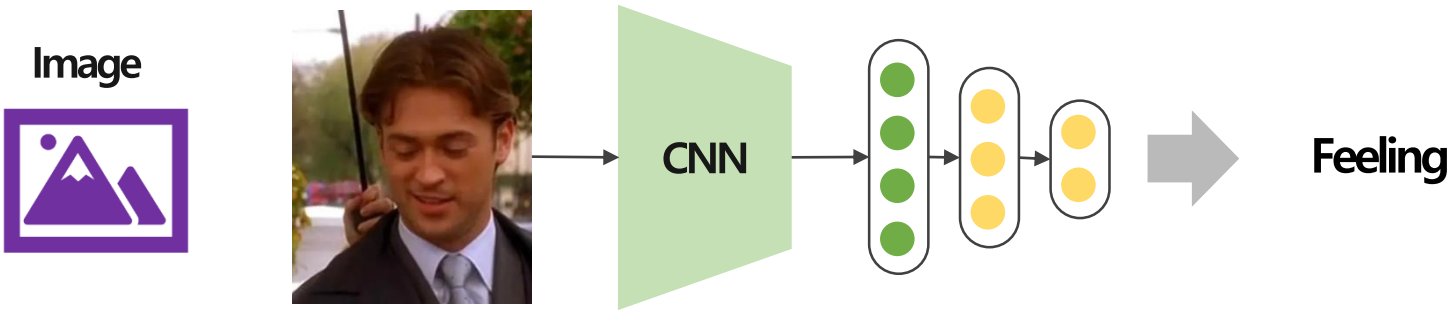
- 정보를 인코딩하는 특정 방식(양식)



Introduction

❖ Unimodal vs Multimodal

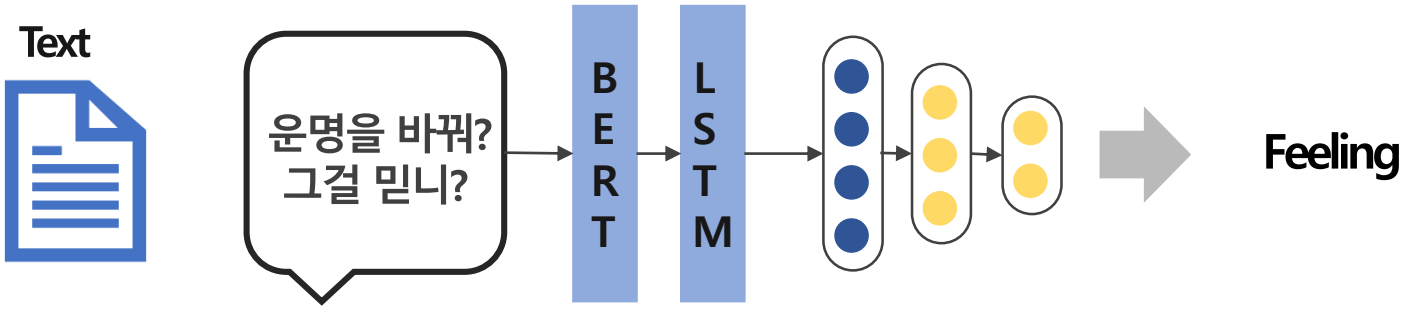
- Unimodal model : 하나의 modality만 활용하여 풀고자 하는 문제를 해결하는 모델
- Multimodal model : 두개 이상의 modality를 활용하여 풀고자 하는 문제를 해결하는 모델



Introduction

❖ Unimodal vs Multimodal

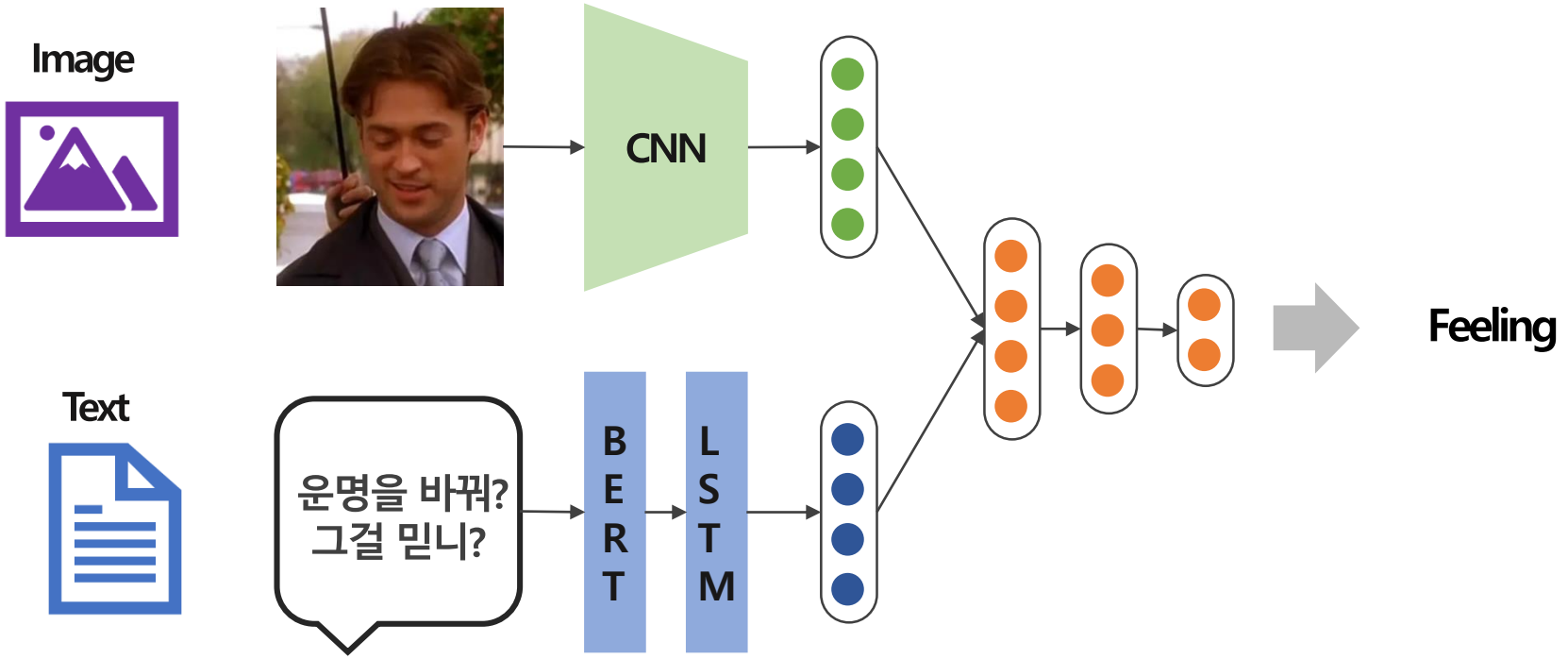
- Unimodal model : 하나의 modality만 활용하여 풀고자 하는 문제를 해결하는 모델
- Multimodal model : 두개 이상의 modality를 활용하여 풀고자 하는 문제를 해결하는 모델



Introduction

❖ Unimodal vs Multimodal

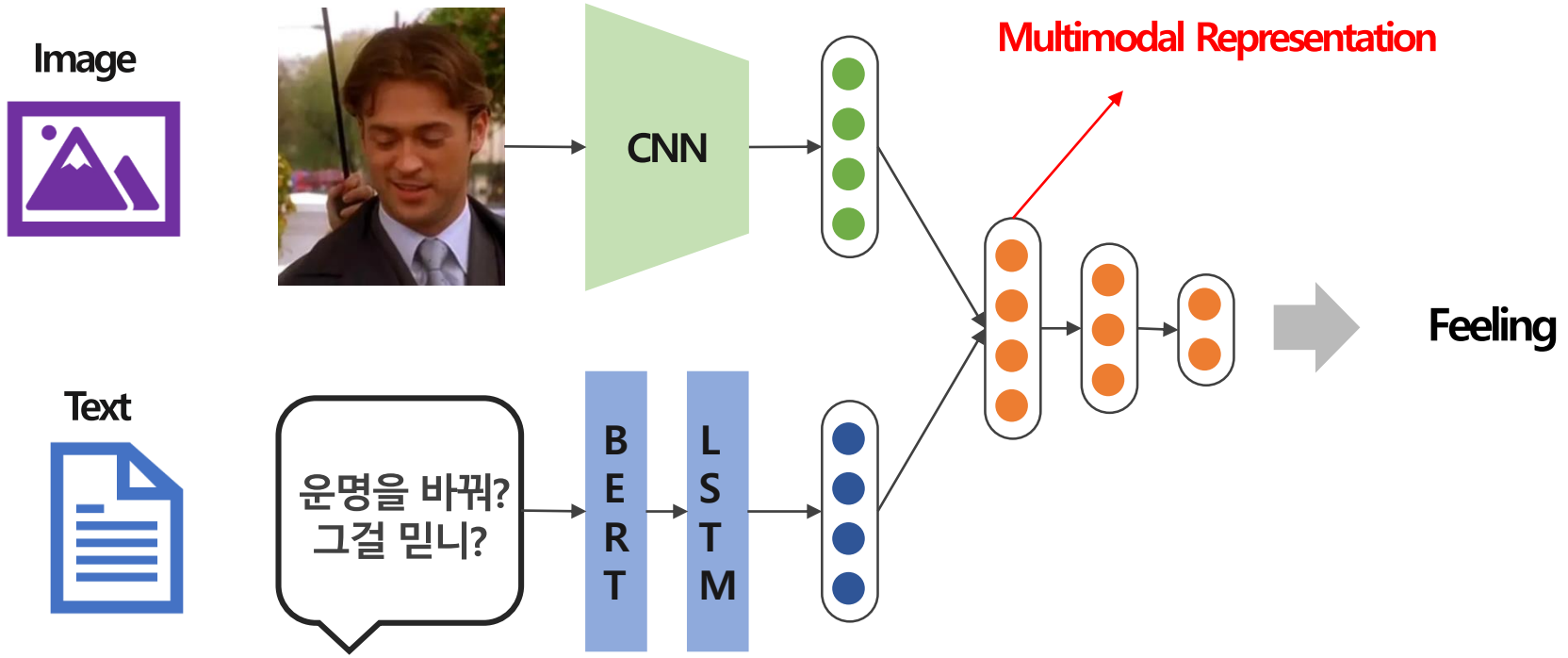
- Unimodal model : 하나의 modality만 활용하여 풀고자 하는 문제를 해결하는 모델
- Multimodal model : 두개 이상의 modality를 활용하여 풀고자 하는 문제를 해결하는 모델



Introduction

❖ Multimodal Representation

- Multimodal Representation : 여러 modality 정보가 결합되어있는 representation vector를 의미



Why Multimodal?

Introduction

❖ Unimodal

- 애매모호~

Image

행복한 미소?

냉소적인 웃음?

Text

운명을 바꿔?
그걸 믿니?

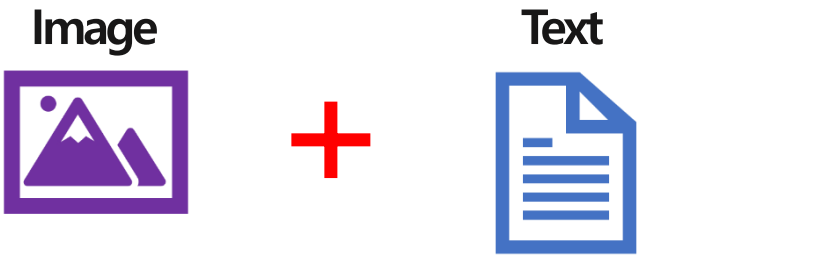
순수한 궁금증?

냉소적 질문?

Introduction

❖ Multimodal

- 상호보완적 정보 공유 (추가, 보충하는 정보 제공)
- 좀 더 정확한 특징 추출 가능

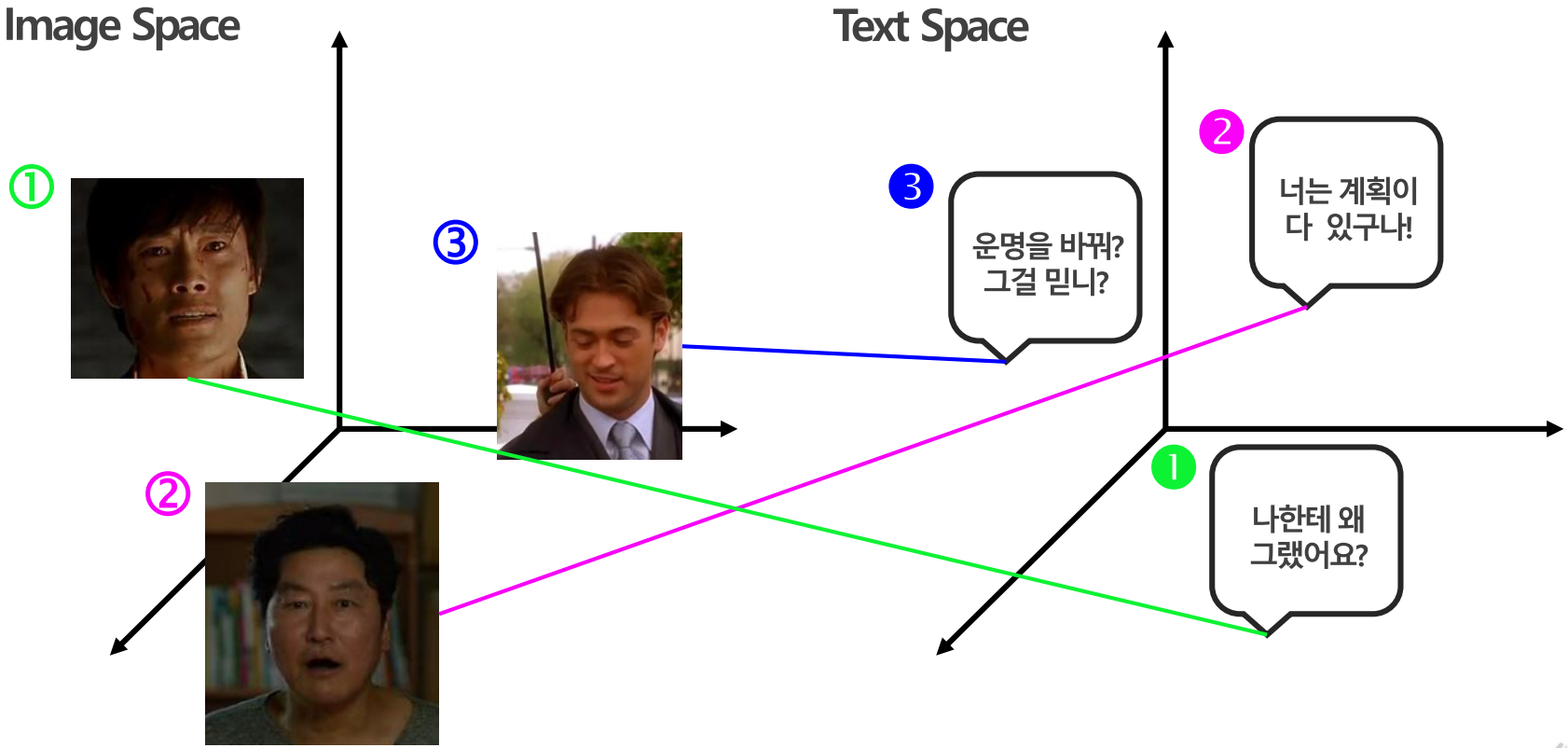


How to get multimodal representation?

Introduction

❖ Heterogeneity gap

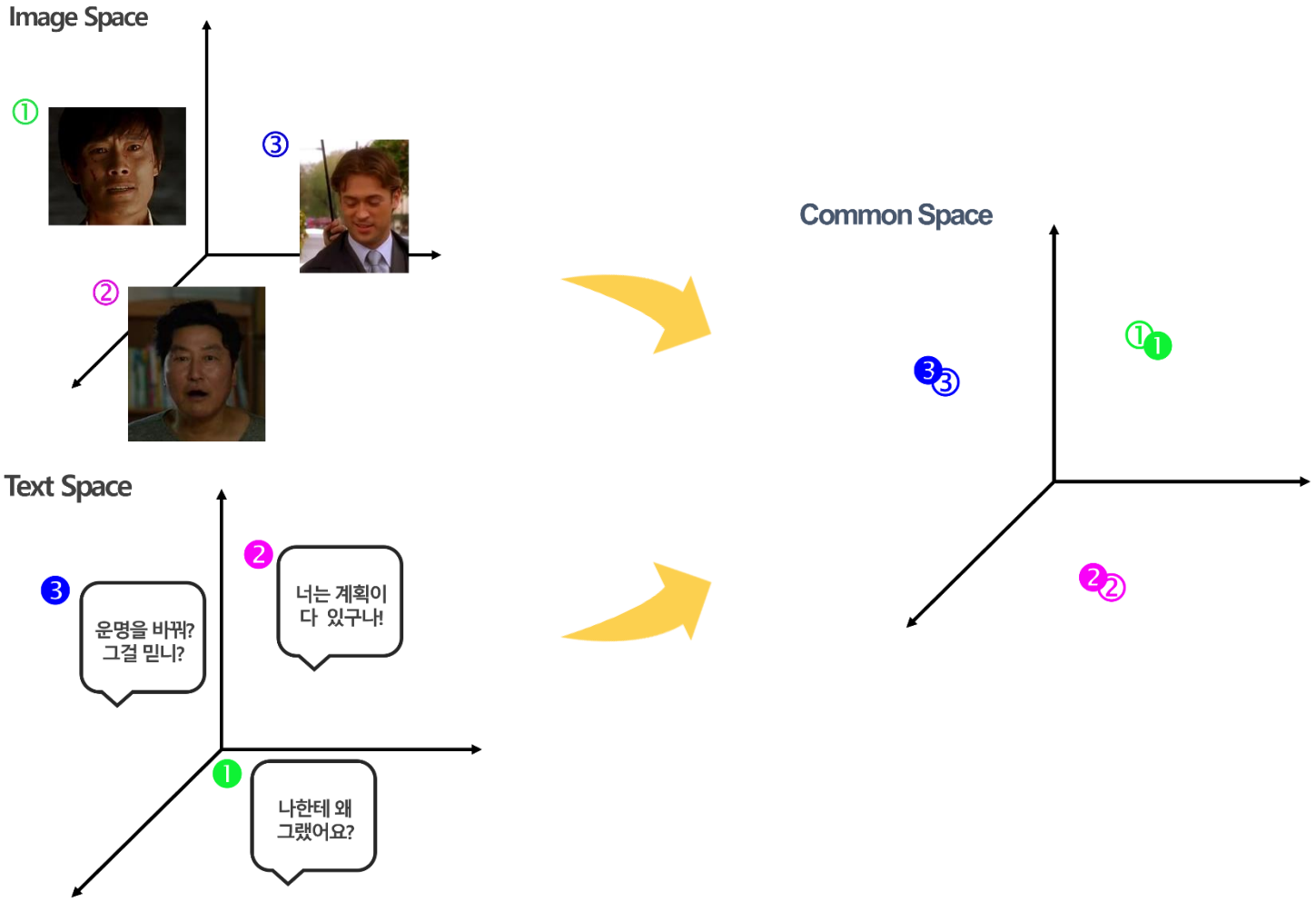
- 각 modality의 representation이 상이함
- 쌍을 이루는 데이터가 상이한 representation을 가짐
- multimodal 데이터가 포괄적으로 사용되는 것을 방해



Introduction

❖ Multimodal Representation

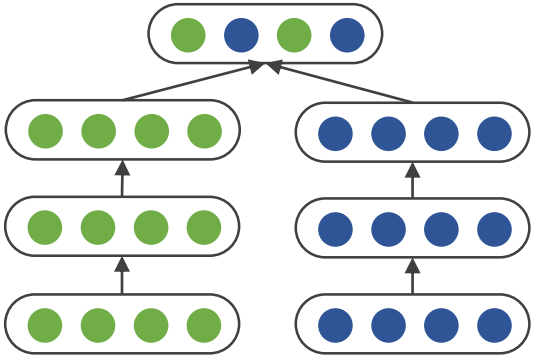
- 서로 다른 modality의 heterogeneity gap을 줄여 common subspace(공동 부분공간)에 representation vector를 매핑



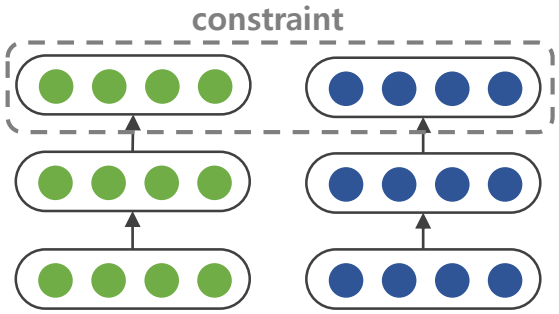
How to narrow heterogeneity gap?

How to narrow heterogeneity gap?

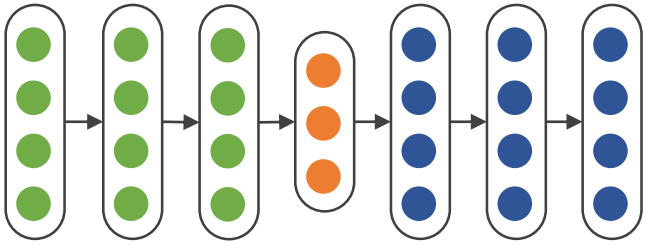
❖ Heterogeneity gap을 줄이는 3가지 방법



(a) Joint Representation



(b) Coordinated Representation



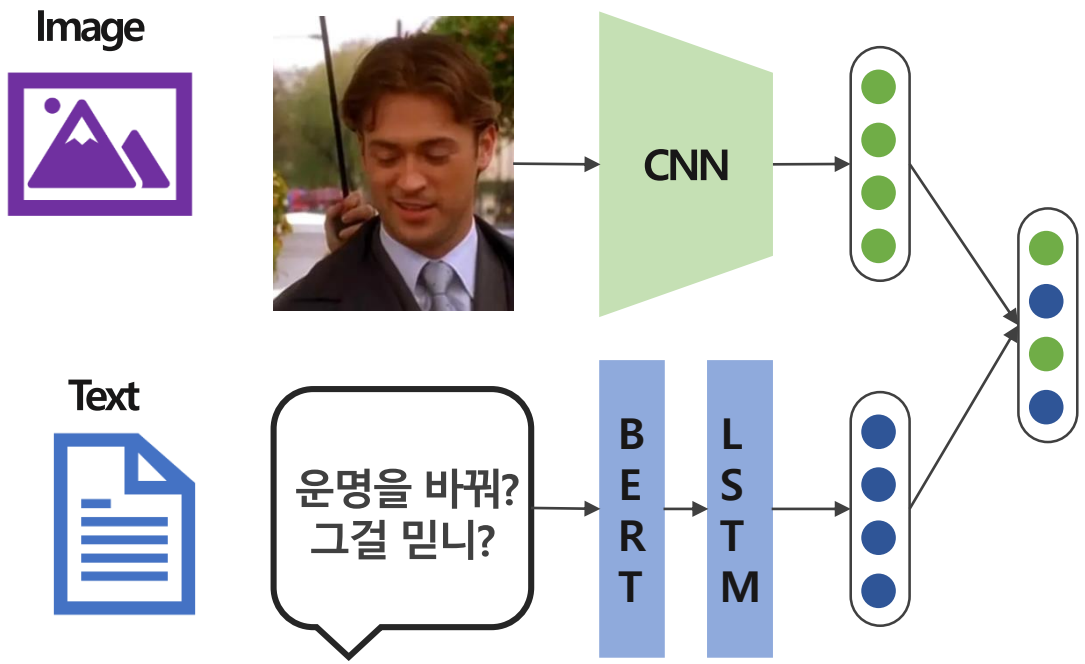
(c) Encoder-decoder

Joint Representation

How to narrow heterogeneity gap?

❖ Joint Representation

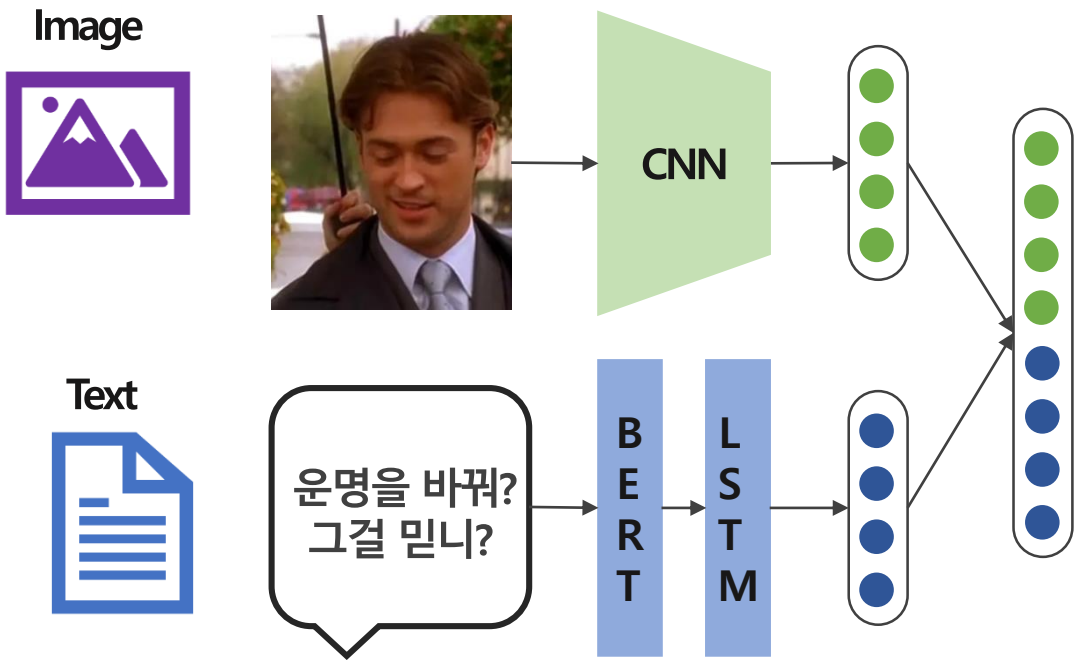
- 각 modality 별로 representation을 학습하고, 이를 결합하여 common subspace에 하나의 representation으로 매핑



How to narrow heterogeneity gap?

❖ Joint Representation

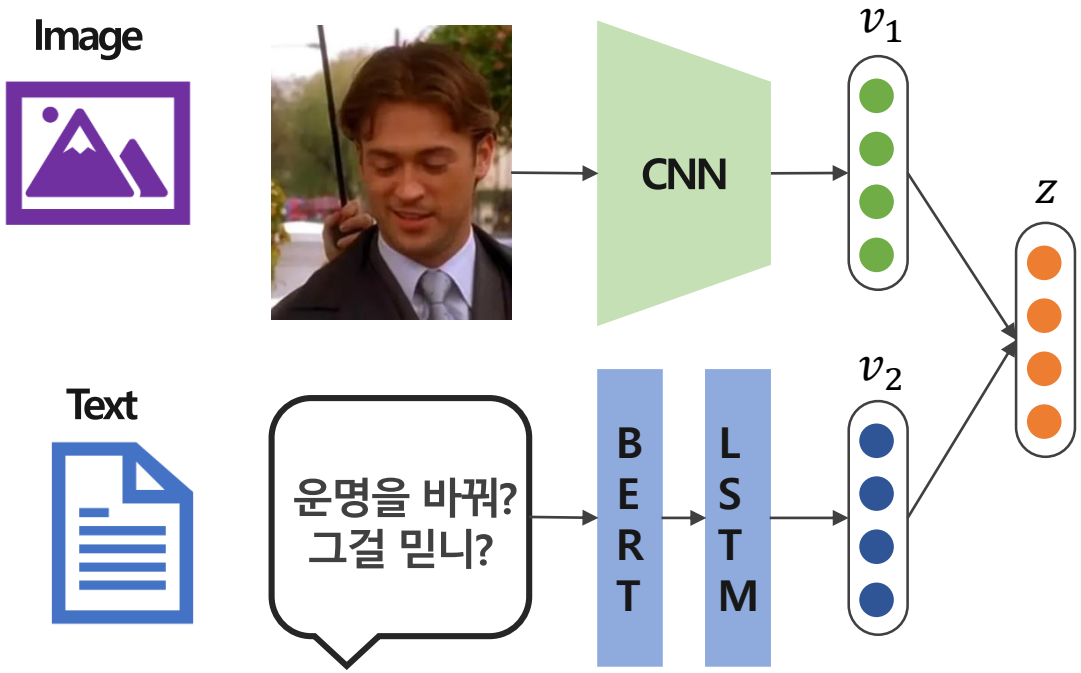
- Concatenate : 각 modality의 벡터를 이어 붙여 결합하는 방식



How to narrow heterogeneity gap?

❖ Joint Representation

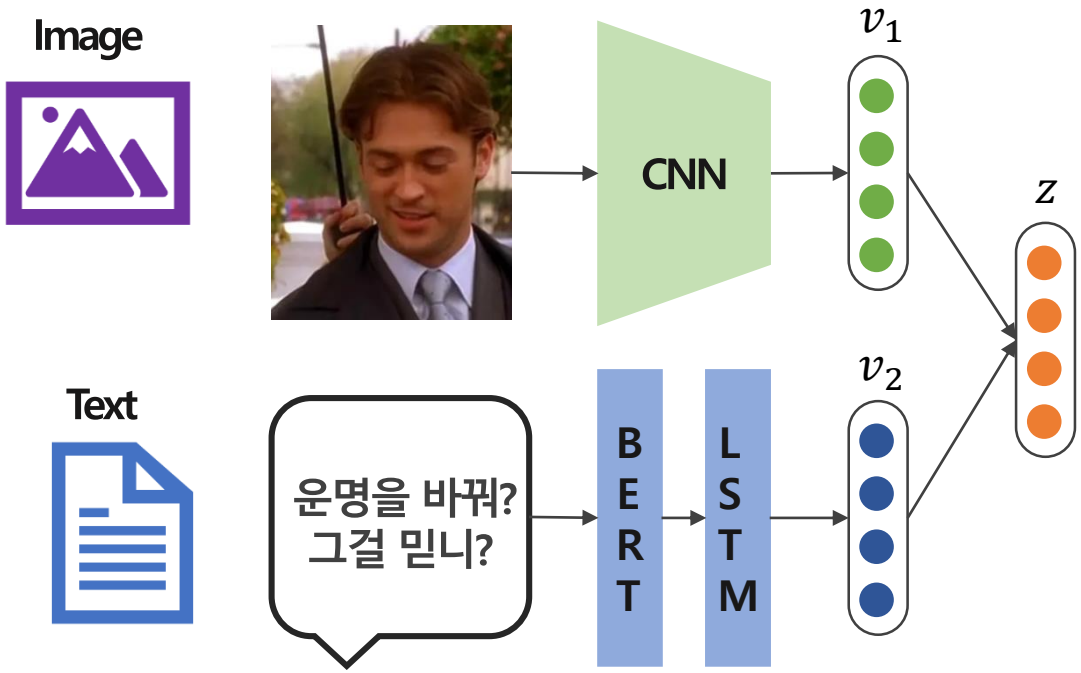
- Additive Approach : 각 modality representation의 가중 합을 하여 하나의 벡터로 결합
- $z = f(w_1^T v_1 + w_2^T v_2)$



How to narrow heterogeneity gap?

❖ Joint Representation

- Multiplicative Approach : 각 modality representation의 외적을 하여 하나의 벡터로 결합
- $z = W[v_1 \otimes v_2] = W[(v_1 v_2^T)]$



How to narrow heterogeneity gap?

❖ Joint Representation Application

- Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding
- Multimodal representation learning을 활용해 Visual Question Answering(VQA)과 Visual Grounding task 수행
- Joint Representation 방법 중 기존 외적을 통한 결합 방식의 장점을 활용하고, 단점을 개선한 방법론을 제안

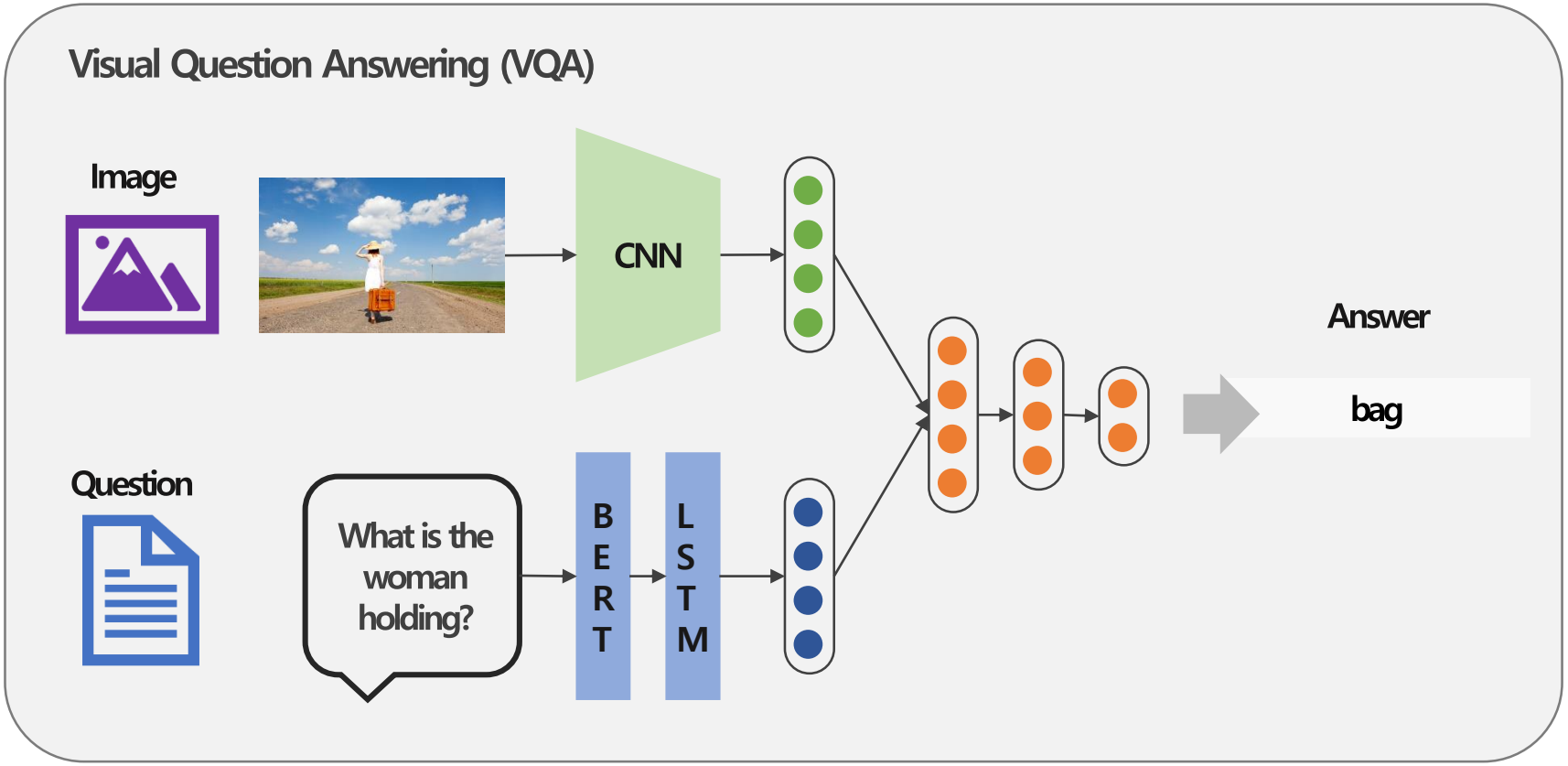
Multimodal Compact Bilinear Pooling
for Visual Question Answering and Visual Grounding

Akira Fukui[*1,2] **Dong Huk Park**[*1] **Daylen Yang**[*1]
Anna Rohrbach[*1,3] **Trevor Darrell**[1] **Marcus Rohrbach**[1]
[1]UC Berkeley EECS, CA, United States
[2]Sony Corp., Tokyo, Japan
[3]Max Planck Institute for Informatics, Saarbrücken, Germany

How to narrow heterogeneity gap?

❖ Joint Representation Application - Task

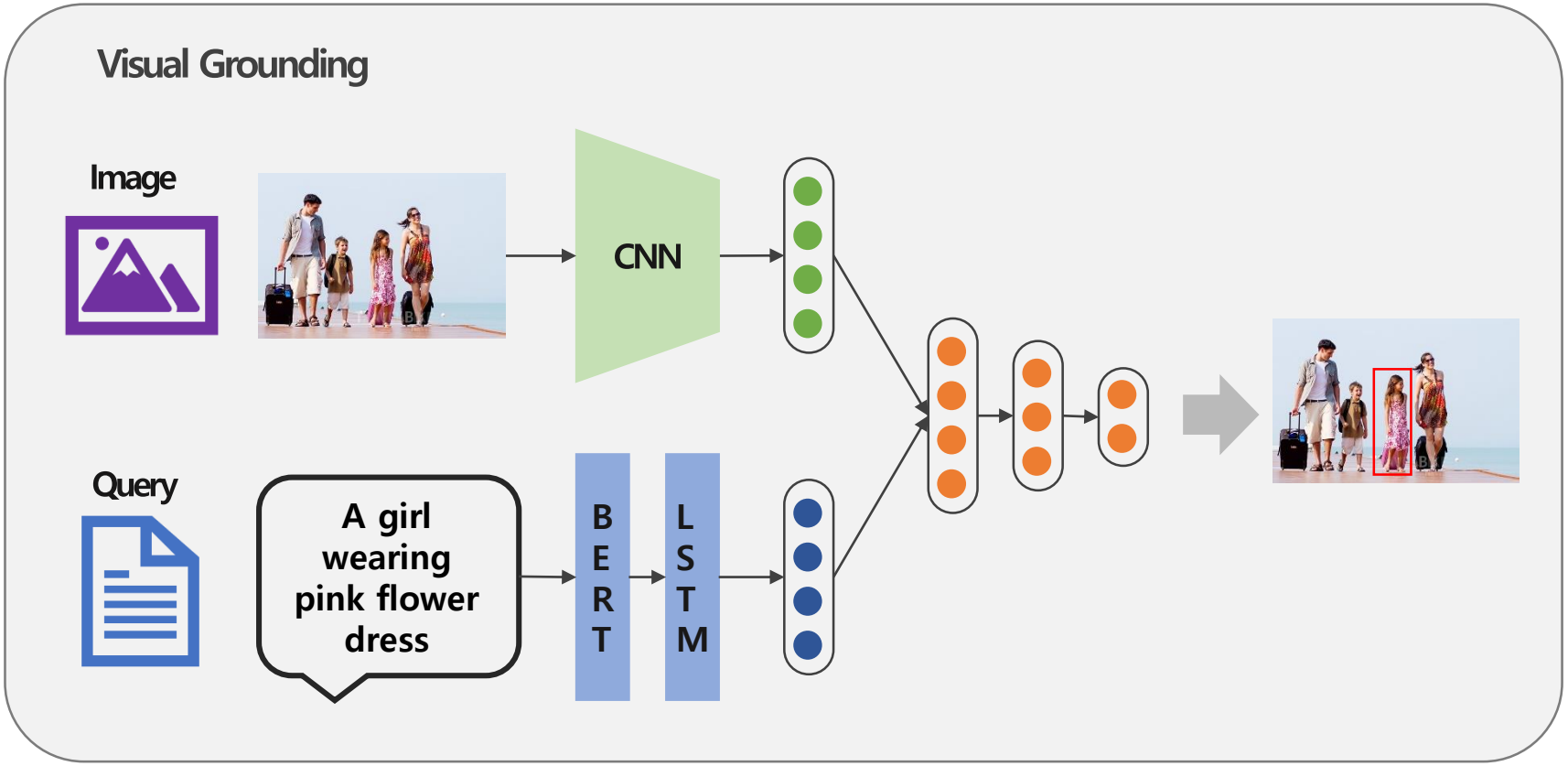
- Visual Question Answering (VQA) : 이미지와 이미지에 대한 질문(Question)이 주어졌을 때, 해당 질문에 맞는 올바른 답변(Answer)을 만들어내는 task



How to narrow heterogeneity gap?

❖ Joint Representation Application - Task

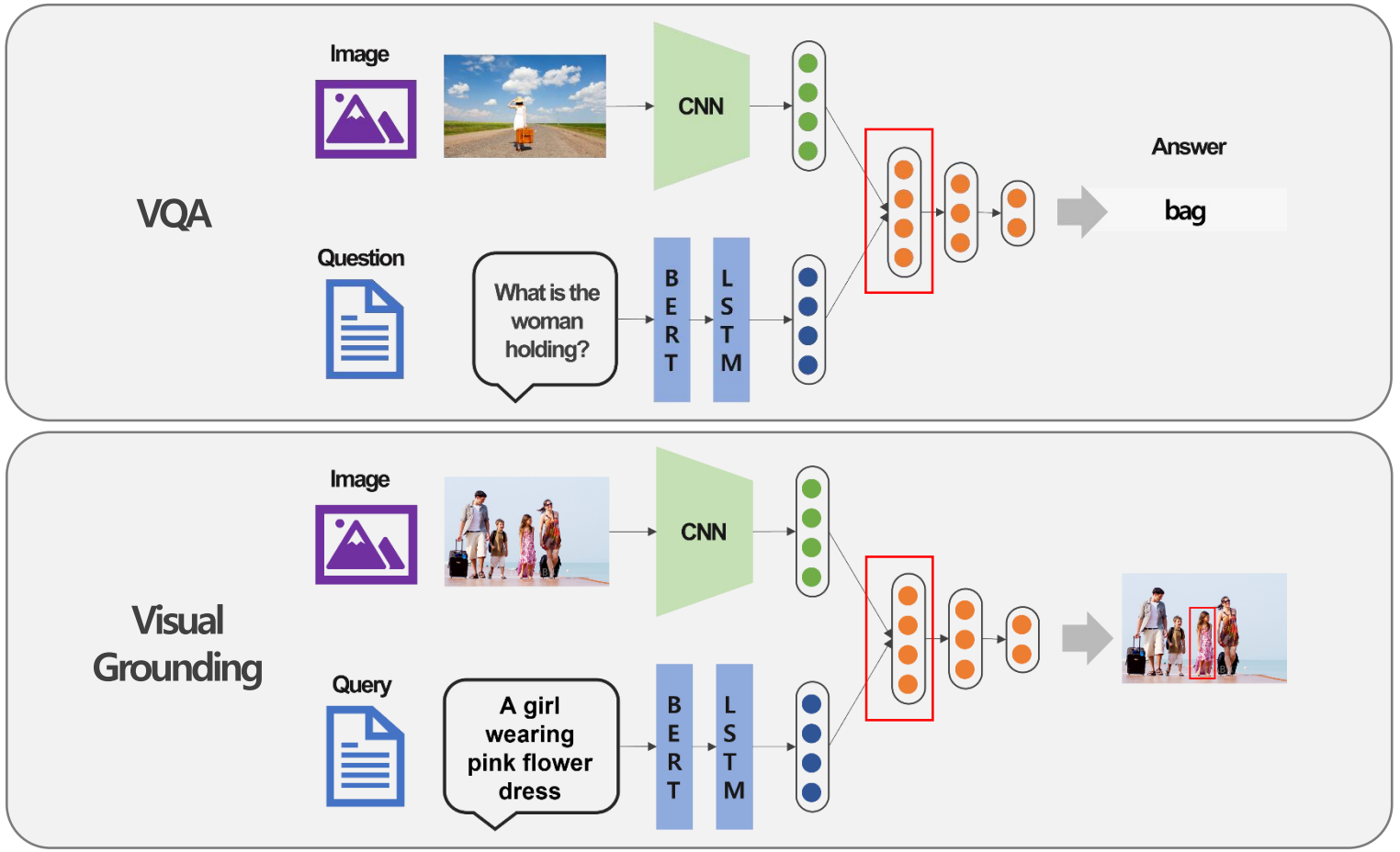
- Visual Grounding : query가 이미지 내에서 묘사하고 있는 부분을 bounding box로 표시하는 task



How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- VQA와 Visual Grounding 모두 image, text modality의 포괄적 정보를 담는 multimodal representation 추출이 중요
- Joint Representation 방법 중 multiplicative approach 방식으로 multimodal representation 생성



How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- Multimodal Compact Bilinear(MCB) : 기존의 외적을 이용해 결합하는 multiplicity approach 방식의 단점을 개선
- multiplicity approach 방식은 **모든 요소들 간의 상호작용을 가능**하게 하는 장점을 가짐
- 기존 외적을 통한 결합 방식은 **많은 메모리 소비와 계산 시간으로 현실적으로 적용이 어려움**

Outer Product : $W[x \otimes q] = W[(xq^T)] = z(x \in R^{n_1}, q \in R^{n_2}, z \in R^d)$

$$x = (x_1, x_2, \dots, x_{n_1})^T, q = (q_1, q_2, \dots, q_{n_2})^T$$

$$x \otimes q = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_1-1} \\ x_{n_1} \end{bmatrix} \begin{bmatrix} q_1 & q_2 & \cdots & q_{n_2-1} & q_{n_2} \end{bmatrix} = \begin{bmatrix} x_1q_1 & \cdots & x_1q_{n^2} \\ \vdots & \ddots & \vdots \\ x_{n_1}q_1 & \cdots & x_{n_1}q_{n^2} \end{bmatrix}$$

$n_1 \times 1$ $\quad$ $1 \times n_2$ $\quad$ $n_1 \times n_2$

How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- Multimodal Compact Bilinear(MCB) : 기존의 외적을 이용해 결합하는 multiplicity approach 방식의 단점을 개선
- Multiplicity approach 방식은 **모든 요소들 간의 상호작용을 가능**하게 하는 장점을 가짐
- 기존 외적을 통한 결합 방식은 **많은 메모리 소비와 계산 시간으로 현실적으로 적용이 어려움**

메모리 소비 : W 파라미터 개수

① $$x \otimes q = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_1-1} \\ x_{n_1} \end{bmatrix} \begin{bmatrix} q_1 & q_2 & \cdots & q_{n_2-1} & q_{n_2} \end{bmatrix} = \begin{bmatrix} x_1 q_1 & \cdots & x_1 q_{n^2} \\ \vdots & \ddots & \vdots \\ x_{n_1} q_1 & \cdots & x_{n_1} q_{n^2} \end{bmatrix}$$

$n_1 \times 1$ $\quad$ $1 \times n_2$ $\quad$ $n_1 \times n_2$

② $$Z = W[x \otimes q] = W \begin{bmatrix} x_1 q_1 \\ x_1 q_2 \\ \vdots \\ x_{n_1} q_{n^2-1} \\ x_{n_1} q_{n^2} \end{bmatrix}$$

<*W dimension*>

$z : d \times 1, [x \otimes q] : (n_1 \times n_2) \times 1$

$\therefore W : d \times (n_1 \times n_2)$

How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- Multimodal Compact Bilinear(MCB) : 기존의 외적을 이용해 결합하는 multiplicity approach 방식의 단점을 개선
- Multiplicity approach 방식은 **모든 요소들 간의 상호작용을 가능**하게 하는 장점을 가짐
- 기존 외적을 통한 결합 방식은 **많은 메모리 소비와 계산 시간으로 현실적으로 적용이 어려움**

계산 시간 : 계산량에 비례

① $$x \otimes q = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_1-1} \\ x_{n_1} \end{bmatrix} \begin{bmatrix} q_1 & q_2 & \cdots & q_{n_2-1} & q_{n_2} \end{bmatrix} = \begin{bmatrix} x_1 q_1 & \cdots & x_1 q_{n^2} \\ \vdots & \ddots & \vdots \\ x_{n_1} q_1 & \cdots & x_{n_1} q_{n^2} \end{bmatrix}$$

$n_1 \times 1$ $\quad$ $1 \times n_2$ $\quad$ $n_1 \times n_2$

② $$Z = W[x \otimes q] = W \begin{bmatrix} x_1 q_1 \\ x_1 q_2 \\ \vdots \\ x_{n_1} q_{n^2-1} \\ x_{n_1} q_{n^2} \end{bmatrix}$$

<Computational Cost>

① + ②

$= (n_1 \times n_2) + (d \times n_1 \times n_2)$

How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- Multimodal Compact Bilinear(MCB) : 기존의 외적을 이용해 결합하는 multiplicity approach 방식의 단점을 개선
- 외적에 Count Sketch projection function (Ψ)을 취해 메모리 사용과 계산 시간을 절감

$$\Psi(x \otimes q)$$

$$= \Psi(x) * \Psi(q) \; (*: \textit{Convolutional operator})$$

: Pham and pagh (2013) 에서 증명

$$= FFT^{-1}\left(FFT(\Psi(x)) \odot FFT(\Psi(q))\right) \; (\odot: \textit{element} - \textit{wise product})$$

How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- Count Sketch projection function(Ψ): 벡터의 차원을 변환해주는 함수 ($\boldsymbol{x} \in \boldsymbol{R^n} \rightarrow \boldsymbol{x'} \in \boldsymbol{R^k}$)

$\Psi(\boldsymbol{x})(\boldsymbol{x} \in \boldsymbol{R^n} \rightarrow \boldsymbol{x'} \in \boldsymbol{R^k})$ 구성 요소

- ✓ $\boldsymbol{s} \in \{-\mathbf{1}, \mathbf{1}\}^{\boldsymbol{n}}$: index마다 -1 or 1을 가지는 벡터 s
- ✓ $\boldsymbol{h} \in \{\mathbf{1}, \ldots, \boldsymbol{k}\}^{\boldsymbol{n}}$: index에 마다 1~k까지의 수 중 하나를 가지는 벡터 h

  ($\boldsymbol{s}$, $\boldsymbol{h}$ 는 uniform distribution에 따라 랜덤하게 초기화)
- ✓ $\boldsymbol{x'} = \{\mathbf{0}, \ldots, \mathbf{0}\}^{\boldsymbol{k}}$: index에 마다 0 값을 가지는 벡터 $\boldsymbol{x'}$

➡ Index마다 $\boldsymbol{x'}[\mathbf{h}[\mathbf{i}]] = \boldsymbol{x'}[\mathbf{h}[\boldsymbol{i}]] + \boldsymbol{x}[\boldsymbol{i}] \cdot \boldsymbol{s}[\boldsymbol{i}]$ 를 수행하여 $\boldsymbol{x} \rightarrow \boldsymbol{x'}$ 로 변환

How to narrow heterogeneity gap?

$\Psi(x)(x \in R^n \rightarrow x' \in R^k)$ 구성 요소

- ✓ $s \in \{-1, 1\}^n$: index마다 -1 or 1을 가지는 벡터 s
- ✓ $h \in \{1, \dots, k\}^n$: index에 마다 1~k까지의 수 중 하나를 가지는 벡터 h
  (s , h 는 uniform distribution에 따라 랜덤하게 초기화)
- ✓ $x' = \{0, \dots, 0\}^k$: index에 마다 0 값을 가지는 벡터 x'
- ➞ Index마다 $x'[h[i]] = x'[h[i]] + x[i] \cdot s[i]$ 를 수행하여 $x \rightarrow x'$ 로 변환

❖ Joint Representation Application - MCB

- Count Sketch projection function(Ψ): 벡터의 차원을 변환해주는 함수 ($x \in R^n \rightarrow x' \in R^k$)

$\Psi(x)(x \in R^n \rightarrow x' \in R^k)$: $n = 10, k = 5$ 라고 가정

i	1	2	3	4	5	6	7	8	9	10
x	2	3	9	7	5	6	3	1	0	9
s	1	1	1	-1	1	-1	-1	1	1	1
h	1	3	2	5	4	4	1	5	3	2
$x \cdot s$	2	3	9	-7	5	-6	-3	1	0	9

➞ $x'[h[1]] = x'[h[1]] + x[1] \cdot s[1] = x'[1] + 2 = 2$

How to narrow heterogeneity gap?

$\Psi(x)(x \in R^n \rightarrow x' \in R^k)$ 구성 요소

- ✓ $s \in \{-1, 1\}^n$: index마다 -1 or 1을 가지는 벡터 s
- ✓ $h \in \{1, \dots, k\}^n$: index에 마다 1~k까지의 수 중 하나를 가지는 벡터 h
  (s , h 는 uniform distribution에 따라 랜덤하게 초기화)
- ✓ $x' = \{0, \dots, 0\}^k$: index에 마다 0 값을 가지는 벡터 x'
- ➞ Index마다 $x'[h[i]] = x'[h[i]] + x[i] \cdot s[i]$ 를 수행하여 $x \rightarrow x'$ 로 변환

❖ Joint Representation Application - MCB

- Count Sketch projection function(Ψ): 벡터의 차원을 변환해주는 함수 ($x \in R^n \rightarrow x' \in R^k$)

$\Psi(x)(x \in R^n \rightarrow x' \in R^k)$: $n = 10, k = 5$라고 가정

i	1	2	3	4	5	6	7	8	9	10
x	2	3	9	7	5	6	3	1	0	9
s	1	1	1	-1	1	-1	-1	1	1	1
h	1	3	2	5	4	4	1	5	3	2
$x \cdot s$	2	3	9	-7	5	-6	-3	1	0	9

➞ x'

$= \{2 + (-3), 9 + 9, 3 + 0, 5 + (-1), (-7) + 1\}$

$= \{-1, 18, 3, 4, -6\}$

How to narrow heterogeneity gap?

❖ Joint Representation Application - MCB

- Count Sketch projection function(Ψ): 벡터의 차원을 변환해주는 함수 ($\boldsymbol{x} \in \boldsymbol{R^n} \rightarrow \boldsymbol{x'} \in \boldsymbol{R^k}$)

$\boldsymbol{\Psi(x \otimes q)}$

$= \boldsymbol{\Psi(x) * \Psi(q)}$ ($*$: $\boldsymbol{Convolutional\ operator}$)

: Pham and pagh (2013) 에서 증명

$= \boldsymbol{FFT^{-1}\left(FFT(\Psi(x)) \odot FFT(\Psi(q))\right)}$ ($\odot$: $\boldsymbol{element-wise\ product}$)

➡ $\boldsymbol{x \otimes q}$ 에 $\boldsymbol{\Psi}$ 를 취해 element-wise product 계산 식으로 변경

➡ $\boldsymbol{FFT}$차원 $= \boldsymbol{n_f}$ 일 때 계산 : $\boldsymbol{n_f \ll (n_1 \times n_2) + (d \times n_1 \times n_2)}$

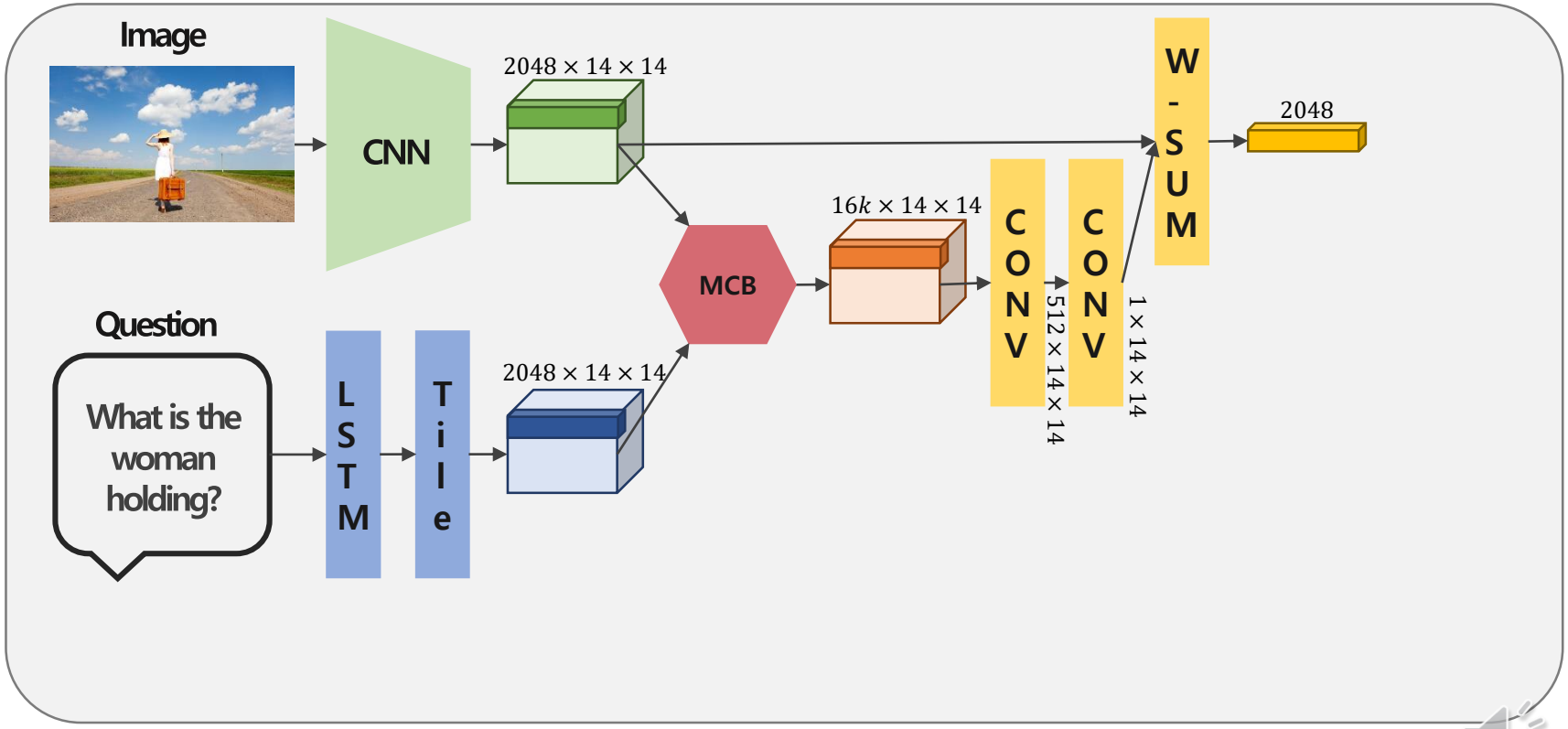
(같은 index 요소들 끼리만 곱해주는 element-wise product로 계산이 감소)

➡ $\boldsymbol{s, h}$에 대한 메모리 할당 : $\boldsymbol{2n_1 + 2n_2 \ll W: d \times (n_1 \times n_2)}$

How to narrow heterogeneity gap?

❖ Joint Representation Application – 모델 구조

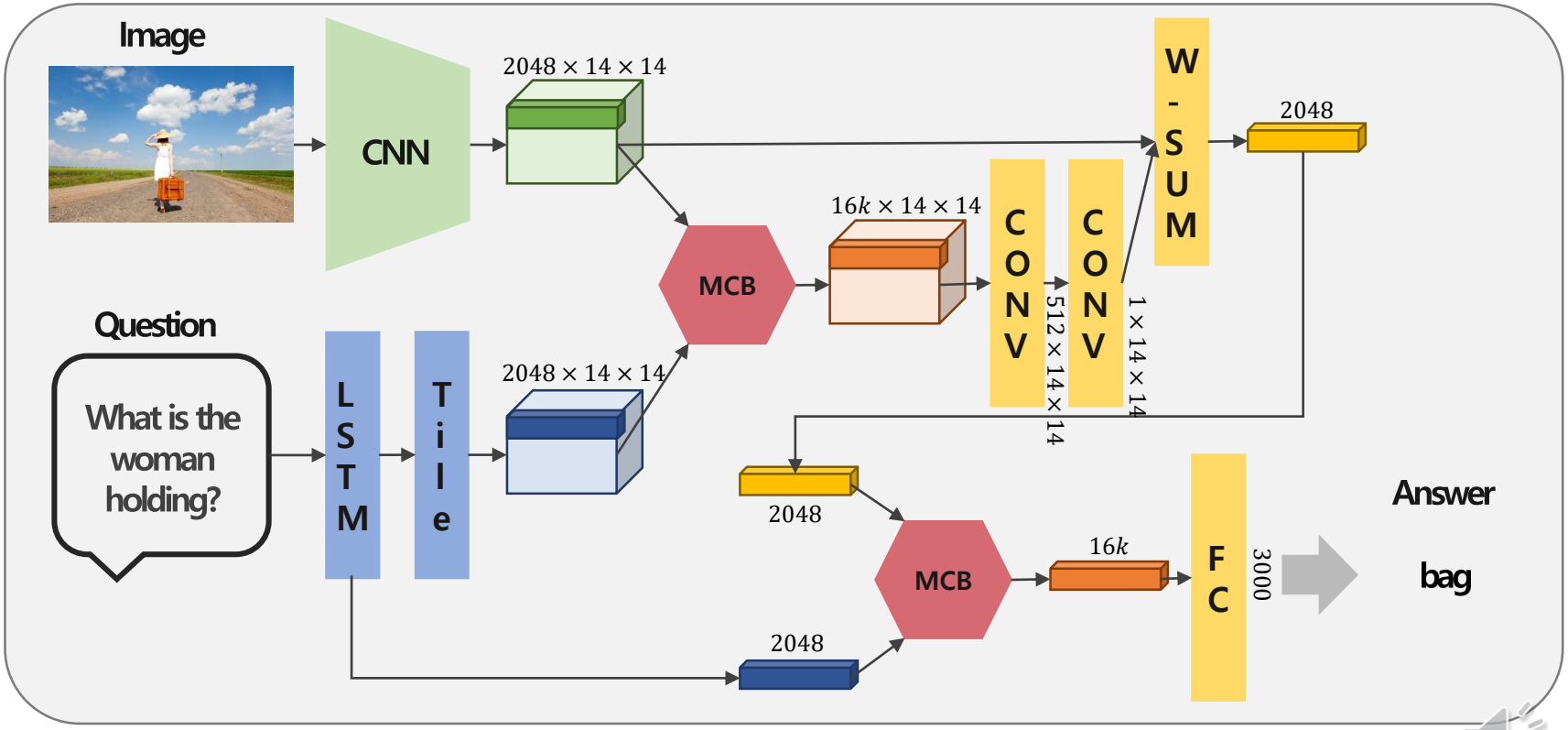
- Attention : MCB를 거쳐 이미지의 일부분과 문장의 연관성 정보를 담는 attention map을 추출 -> 이를 활용해 이미지 부분들에 대한 representation vector를 가중 합하여 이미지 representation vector 추출
- MCB를 거쳐 Attention mechanism을 통해 생성한 이미지 representation과 문제 representation을 결합하여 3000개의 class 중 하나로 정답을 할당하는 classification 수행



How to narrow heterogeneity gap?

❖ Joint Representation Application – 모델 구조

- Attention : MCB를 거쳐 이미지의 일부분과 문장의 연관성 정보를 담는 attention map을 추출 -> 이를 활용해 이미지 부분들에 대한 representation vector를 가중 합하여 이미지 representation vector 추출
- MCB를 거쳐 Attention mechanism을 통해 생성한 이미지 representation과 문제 representation을 결합하여 3000개의 class 중 하나로 정답을 할당하는 classification 수행



How to narrow heterogeneity gap?

❖ Joint Representation Application – 실험 결과

- MCB pooling이 다른 joint representation 방법론들에 비해 좋은 성능을 보이고 있음을 확인

VQA

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**

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

Visual Grounding

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**

Table 5: Grounding accuracy on Flickr30k Entities dataset.

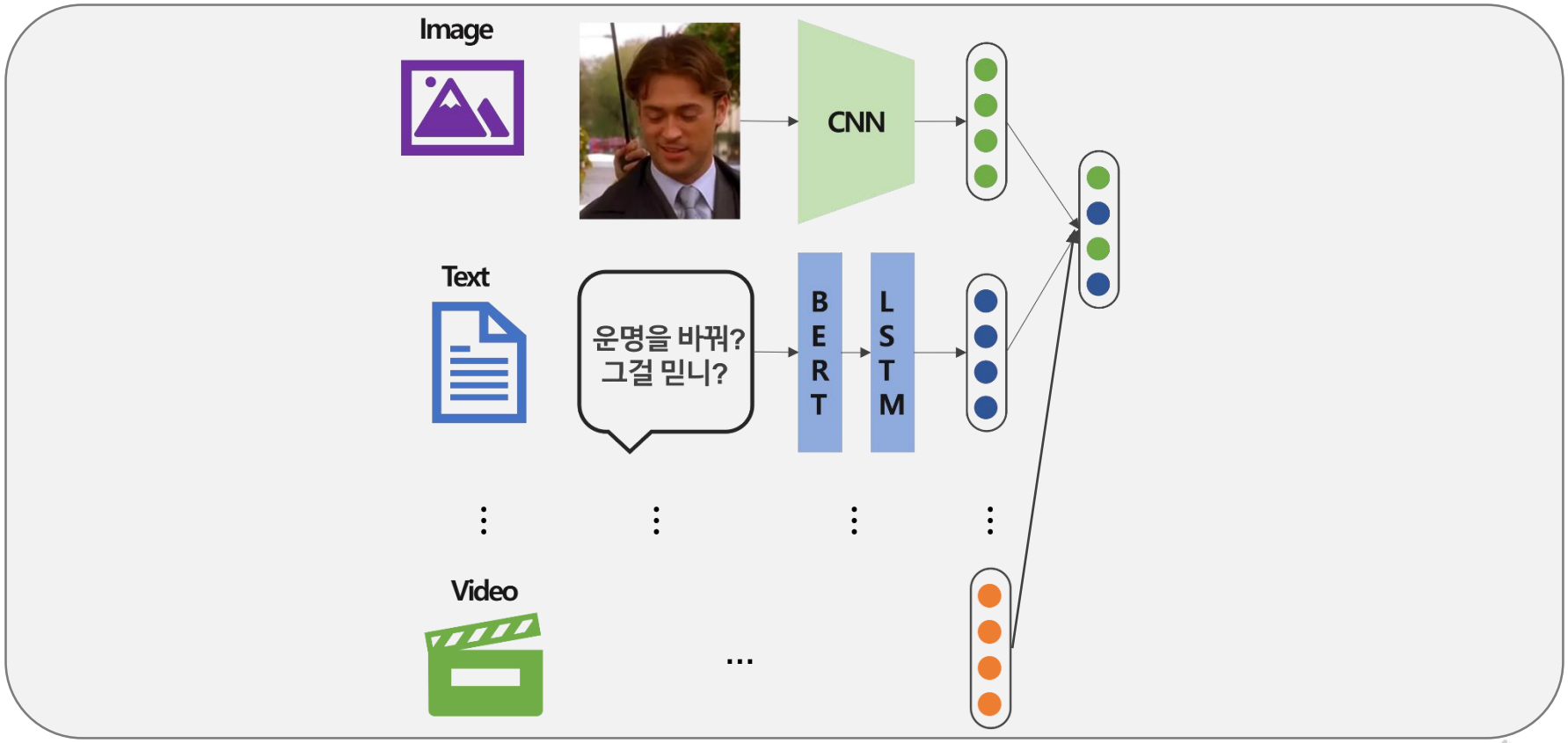
Method	Accuracy, %
Hu et al. (2016b)	17.93
Rohrbach et al. (2016)	26.93
Concatenation	25.48
Element-wise Product	27.80
Element-wise Product + Conv	27.98
MCB	**28.91**

Table 6: Grounding accuracy on ReferItGame dataset.

How to narrow heterogeneity gap?

❖ Joint Representation 장단점

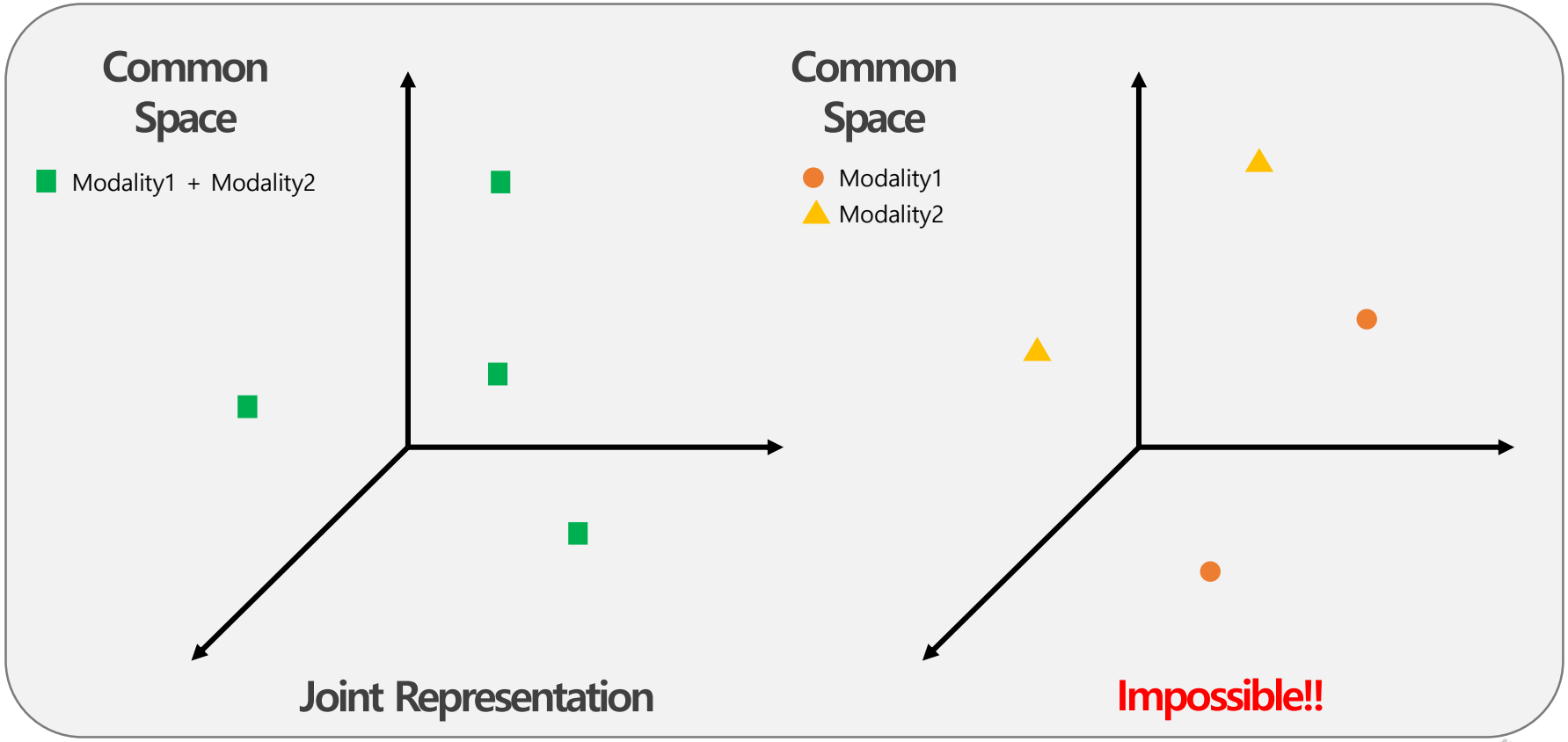
- 장점 : Modality 개수와 상관없이 정보 융합이 편리
- 단점 : Modality별로 분리된 representation을 얻을 수 없음 (일부 Modality 정보가 손실된 상황에 대한 추론 어려움)



How to narrow heterogeneity gap?

❖ Joint Representation 장단점

- 장점 : Modality 개수와 상관없이 정보 융합이 편리
- 단점 : Modality별로 분리된 representation을 얻을 수 없음 (일부 Modality 정보가 손실된 상황에 대한 추론 어려움)

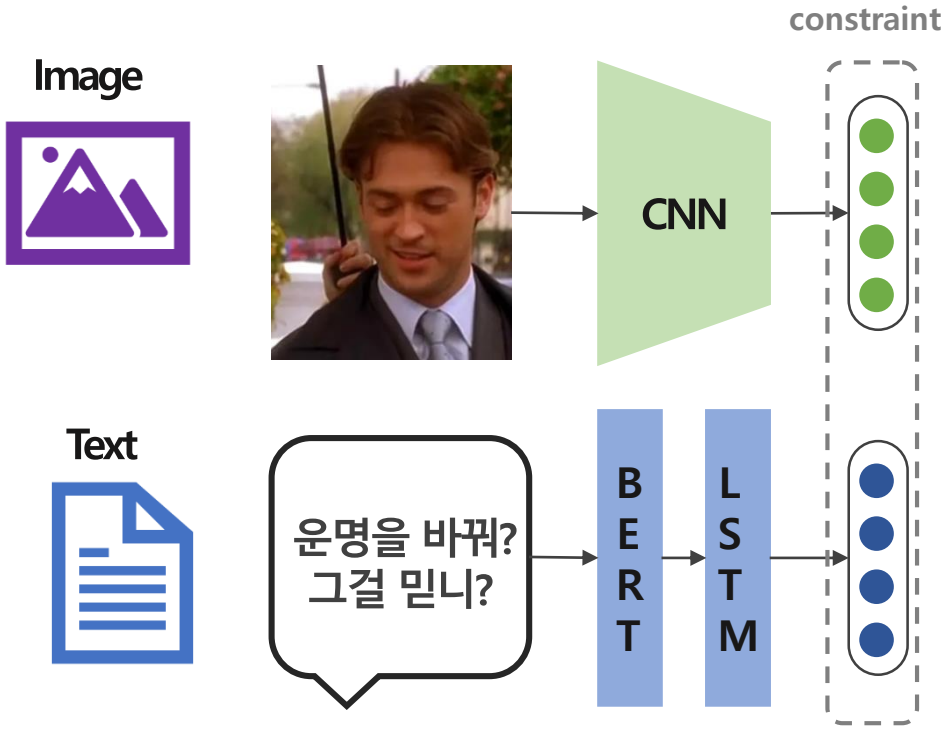


Coordinated Representation

How to narrow heterogeneity gap?

❖ Coordinated Representation

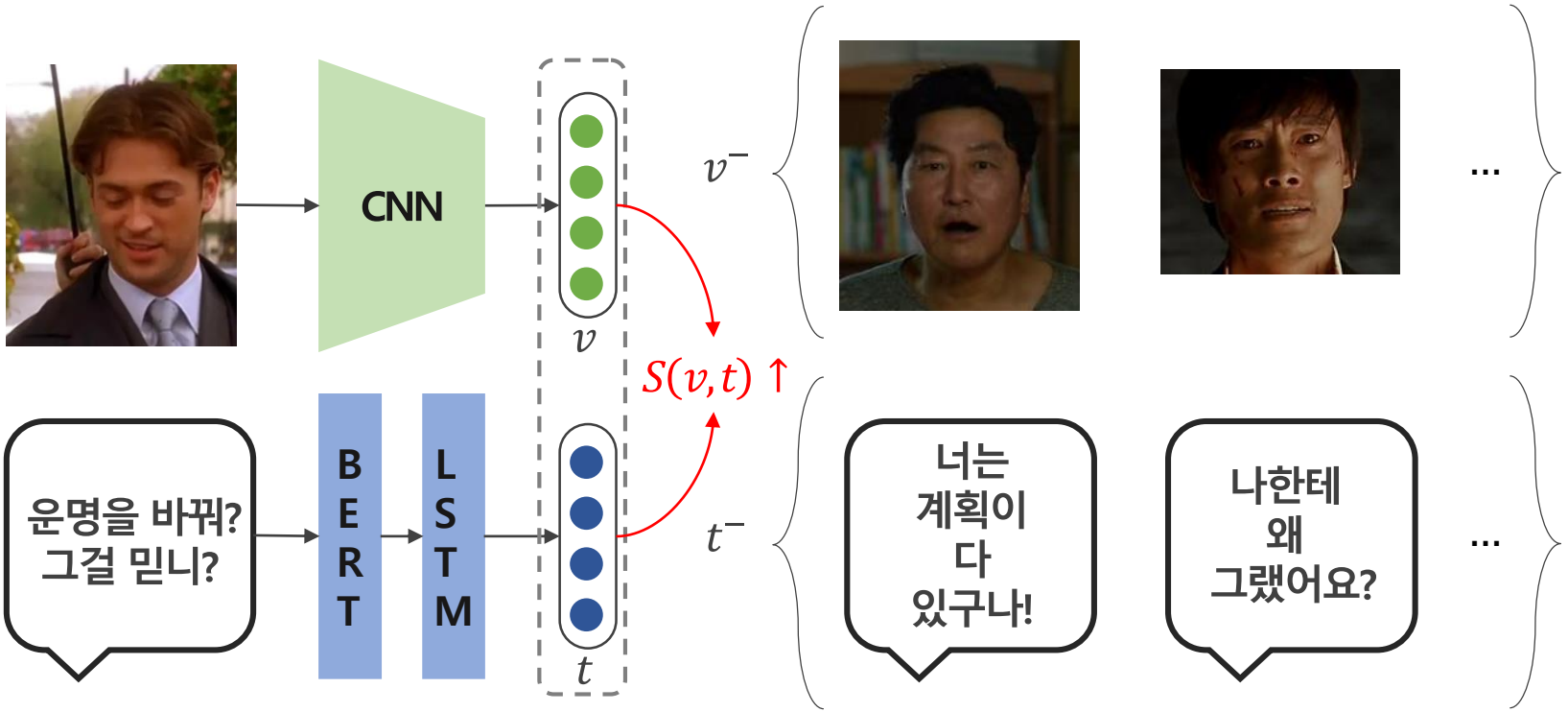
- 제약을 주어서 다른 modality 이더라도 같은 의미를 가지면 가깝게 매핑하는 common space 학습
- 각 modality representation을 분리해서 매핑



How to narrow heterogeneity gap?

❖ Coordinated Representation

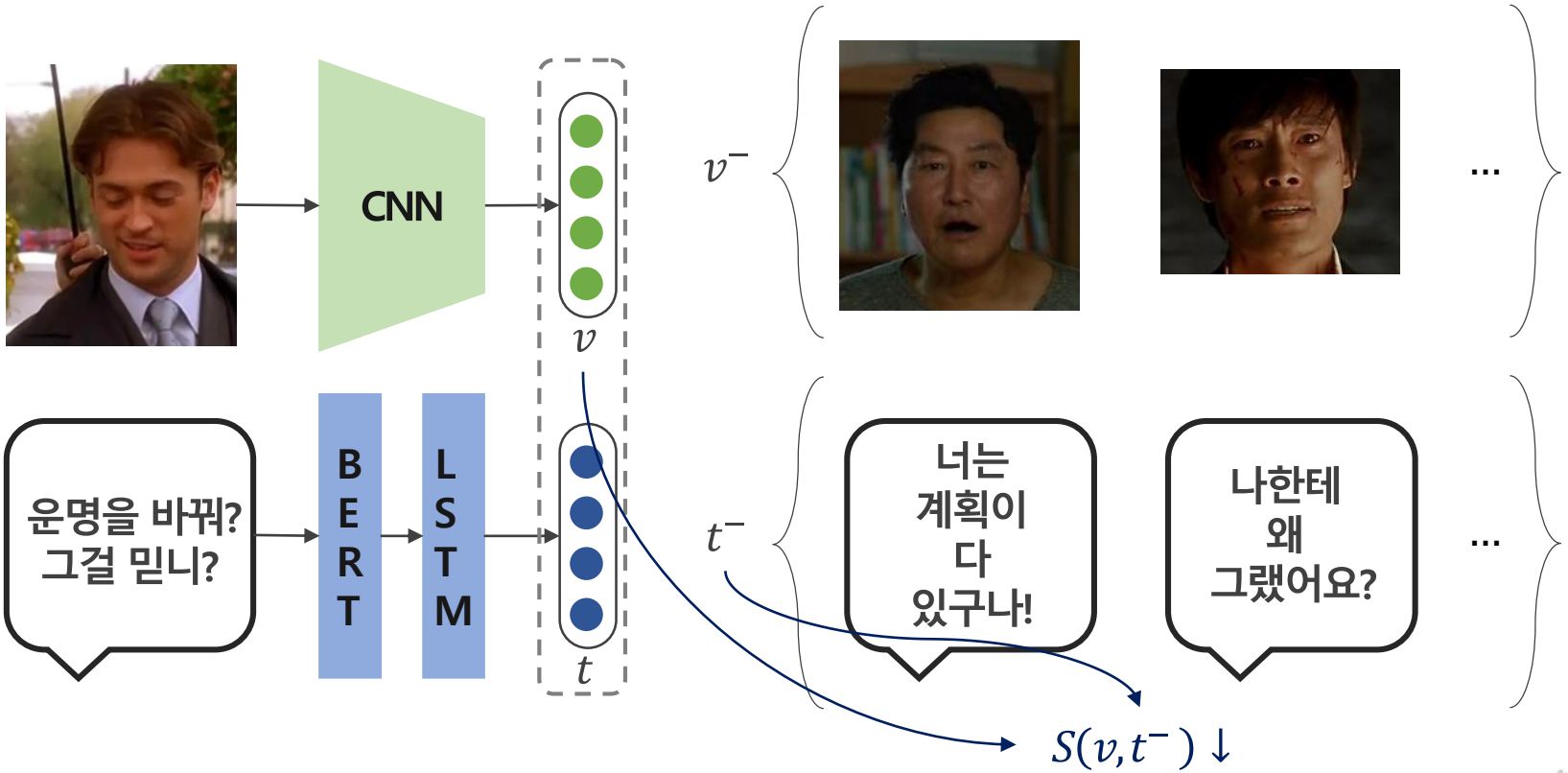
- Cross-modal ranking : 쌍을 이루는 다른 modality와의 similarity는 크도록, 쌍이 아닌 다른 modality와의 similarity는 작도록 제약
- rankLoss : $\sum_v \sum_{t^-} \max(0, \alpha - S(v,t) + S(v,t^-)) + \sum_t \sum_{v^-} \max(0, \alpha - S(t,v) + S(t,v^-))$



How to narrow heterogeneity gap?

❖ Coordinated Representation

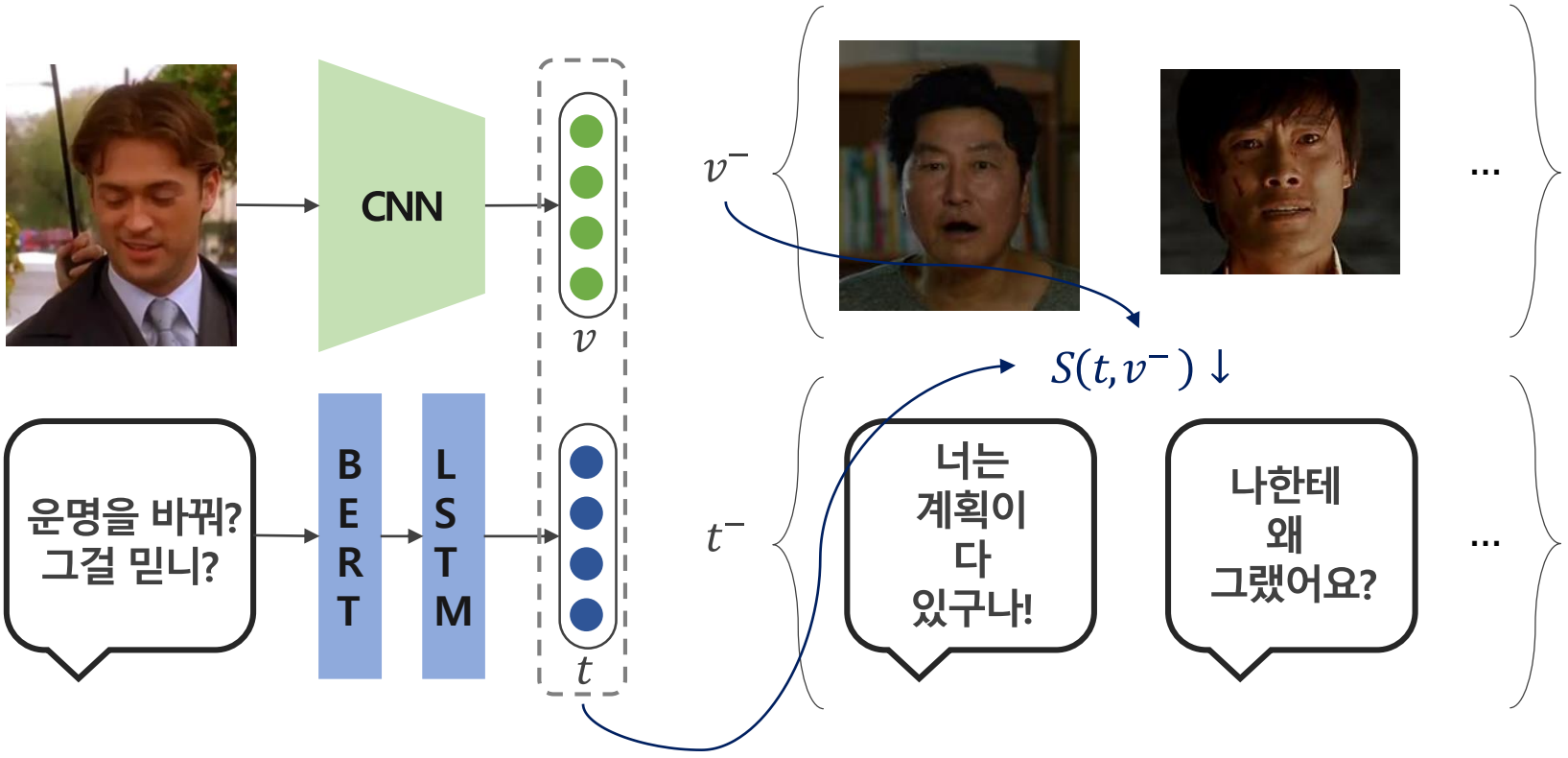
- Cross-modal ranking : 쌍을 이루는 다른 modality와의 similarity는 크도록, 쌍이 아닌 다른 modality와의 similarity는 작도록 제약
- rankLoss : $\sum_v \sum_{t^-} \max(0, \alpha - S(v,t) + S(v,t^-)) + \sum_t \sum_{v^-} \max(0, \alpha - S(t,v) + S(t,v^-))$



How to narrow heterogeneity gap?

❖ Coordinated Representation

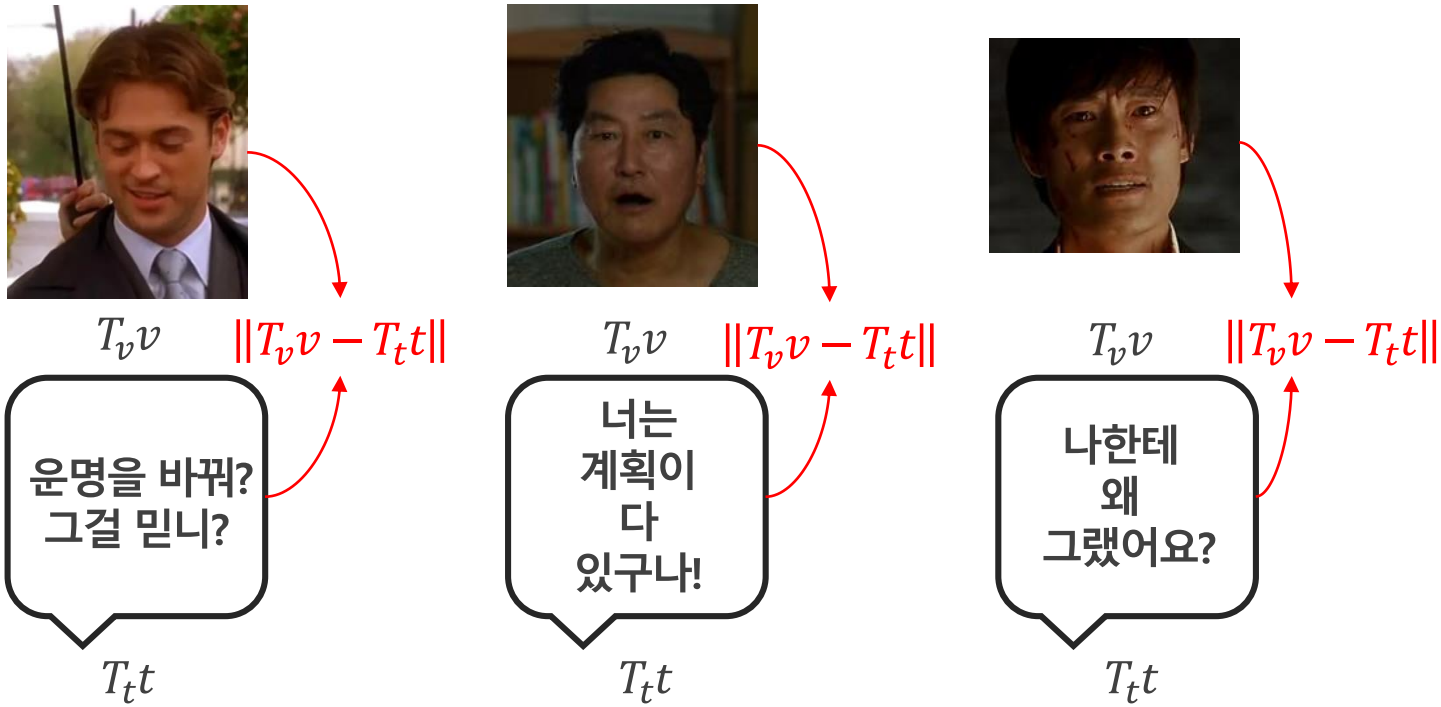
- Cross-modal ranking : 쌍을 이루는 다른 modality와의 similarity는 크도록, 쌍이 아닌 다른 modality와의 similarity는 작도록 제약
- rankLoss : $\sum_v \sum_{t^-} \max(0, \alpha - S(v,t) + S(v,t^-)) + \sum_t \sum_{v^-} \max(0, \alpha - S(t,v) + S(t,v^-))$



How to narrow heterogeneity gap?

❖ Coordinated Representation

- Euclid distance : 쌍을 이루는 서로 다른 modality의 representation vector간의 거리를 최소화 하는 제약
- distanceLoss : $\sum_{(v,s)\epsilon D} \|T_v v - T_t t\|$ $(T \rightarrow transform matrics)$



How to narrow heterogeneity gap?

❖ Coordinated Representation Application

- Deep Visual-Semantic Alignments for Generating Image Descriptions
- Multimodal representation embedding을 학습시키고, 이를 Image Descriptions에 활용
- Multimodal representation embedding에 이미지의 특정 부분과 문장의 단어들 사이의 관계를 고려하는 coordinated representation 방식을 적용

Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy Li Fei-Fei
Department of Computer Science, Stanford University
{karpathy,feifeili}@cs.stanford.edu

Abstract

We present a model that generates natural language descriptions of images and their regions. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Our alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. We demonstrate that our alignment model produces state of the art results in retrieval experiments on Flickr8K, Flickr30K and MSCOCO datasets. We then show that the generated descriptions significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations.

Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

How to narrow heterogeneity gap?

❖ Coordinated Representation Application – representation embedding

- 한 쌍으로 존재하는 image의 영역들과 이를 설명하는 문장 단어들을 각각 representation vector 생성
- Step 1. image representation vector는 이미 200개의 class에 대해 학습된 RCNN을 활용해 생성
- Step 2. 문장 내 단어들에 대한 representation vector는 BRNN을 활용해 생성

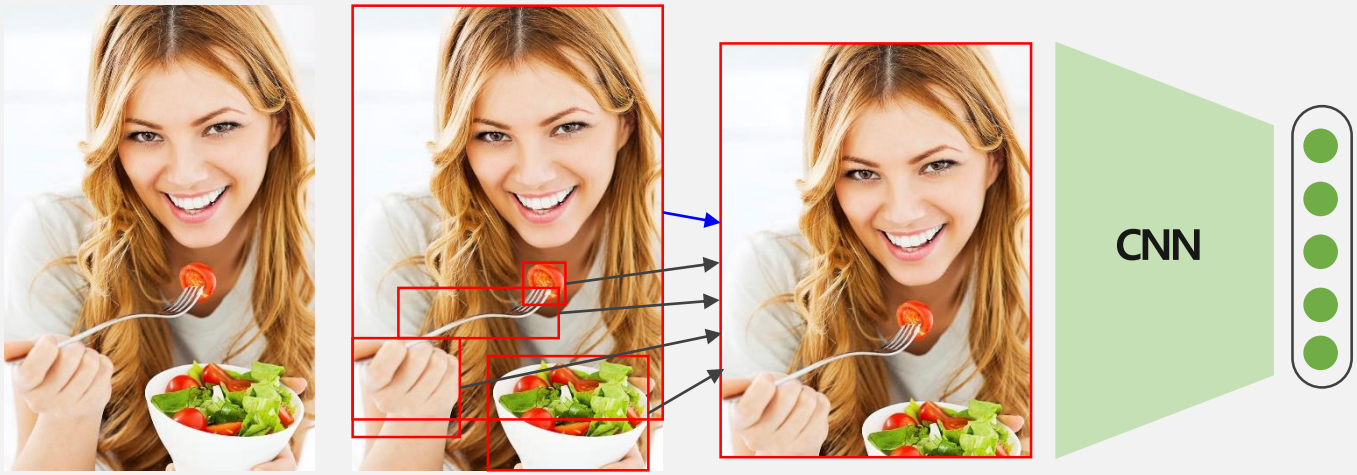
Step 1. 이미지 내에 객체가 존재할만한 bounding box 예측

How to narrow heterogeneity gap?

❖ Coordinated Representation Application – representation embedding

- 한 쌍으로 존재하는 image의 영역들과 이를 설명하는 문장 단어들을 각각 representation vector 생성
- Step 1. image representation vector는 이미 200개의 class에 대해 학습된 RCNN을 활용해 생성
- Step 2. 문장 내 단어들에 대한 representation vector는 BRNN을 활용해 생성

Step 1. 객체가 있을 확률이 높은 상위 19개의 bounding box, 전체 이미지에 대해 이미 학습된 CNN을 통해 4096 차원 벡터 생성



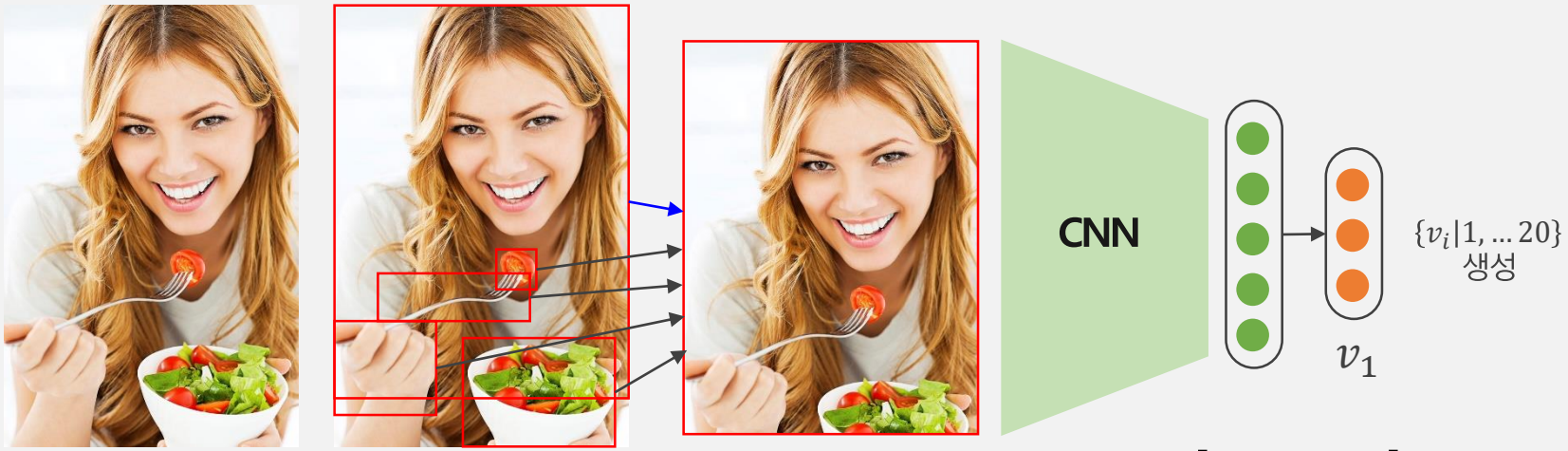
$[CNN_{\theta_c}(I_b)]$

How to narrow heterogeneity gap?

❖ Coordinated Representation Application – representation embedding

- 한 쌍으로 존재하는 image의 영역들과 이를 설명하는 문장 단어들을 각각 representation vector 생성
- Step 1. image representation vector는 이미 200개의 class에 대해 학습된 RCNN을 활용해 생성
- Step 2. 문장 내 단어들에 대한 representation vector는 BRNN을 활용해 생성

Step 1. 모든 bounding box와 전체 이미지에 대한 4096차원 벡터를 h 차원 representation 벡터로 임베딩



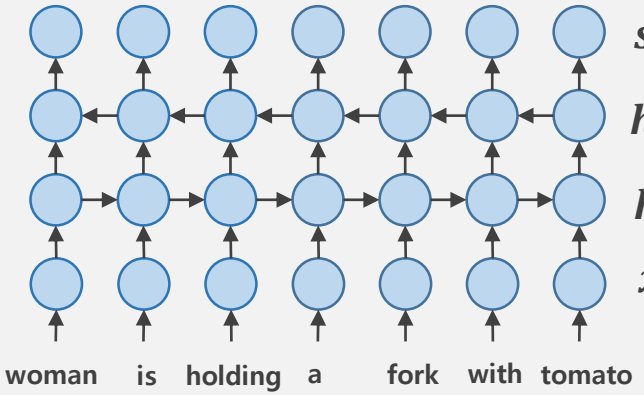
$$v = W_m[CNN_{\theta_c}(I_b)] + b_m$$

How to narrow heterogeneity gap?

❖ Coordinated Representation Application – multimodal representation embedding

- 한 쌍으로 존재하는 image의 영역들과 이를 설명하는 문장 단어들을 각각 representation vector 생성
- Step 1. image representation vector는 이미 200개의 class에 대해 학습된 RCNN을 활용해 생성
- Step 2. 문장 내 단어들에 대한 representation vector는 BRNN을 활용해 생성

Step 2. 문장 내 단어들에 대해 h 차원의 representation vector 생성



s_t : 양쪽에서 전달된 hidden state vector 정보 모두 결합

h_t^b : 왼쪽으로 hidden state vector 전달

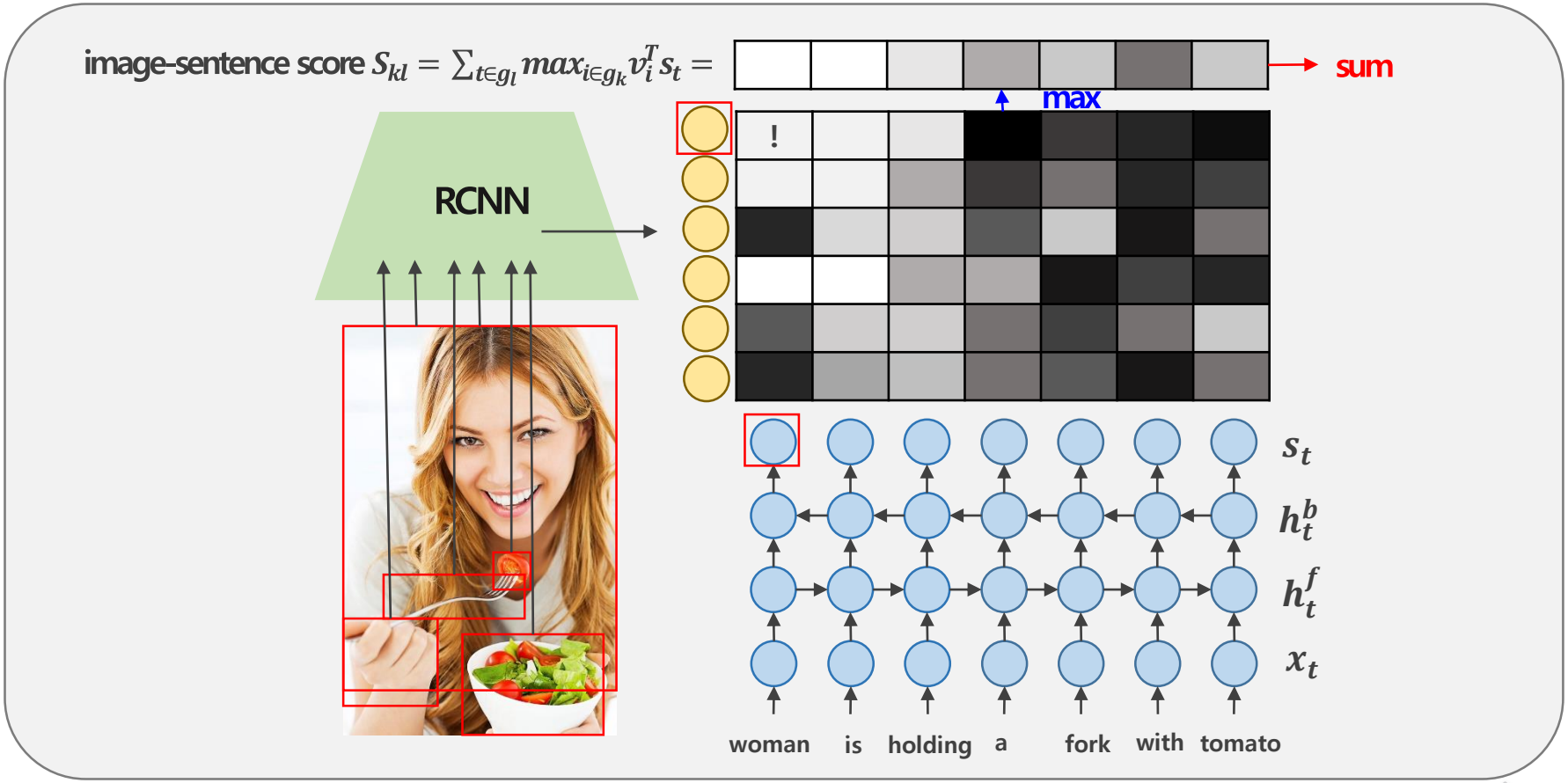
h_t^f : 오른쪽으로 hidden state vector 전달

x_t : word embedding

How to narrow heterogeneity gap?

❖ Coordinated Representation Application – coordinated representation 학습

- Step 1. image-sentence score를 구하기 : image의 특정 영역들과 문장의 단어들 사이의 관련 정도를 score로 부여
- Step 2. 쌍을 이루는 image와 문장간의 image-sentence score가 크도록 제약



How to narrow heterogeneity gap?

❖ Coordinated Representation Application – coordinated representation 학습

- Step 1. image-sentence score를 구하기 : image의 특정 영역들과 문장의 단어들 사이의 관련 정도를 score로 부여
- Step 2. 쌍을 이루는 image와 문장간의 image-sentence score가 크도록 제약

$$\text{Loss function} = \sum_k [\sum_l max(0, S_{kl} - S_{kk} + 1) + \sum_l max(0, S_{lk} - S_{kk} + 1)]$$

쌍을 이루는 image와 문장간의
image-sentence score가
클 것이라고 가정

↓

커질수록 loss function이
작아지도록 제약

↓

쌍을 이루는 image와 문
장간 similarity가 큰
common space 학습

How to narrow heterogeneity gap?

❖ Coordinated Representation Application – embedding result

- 학습한 multimodal representation으로 test image들에 대해 Image-sentence score이 가장 높은 문장을 검색
- 각 bounding box 마다 가장 높은 score를 보인 단어를 표시

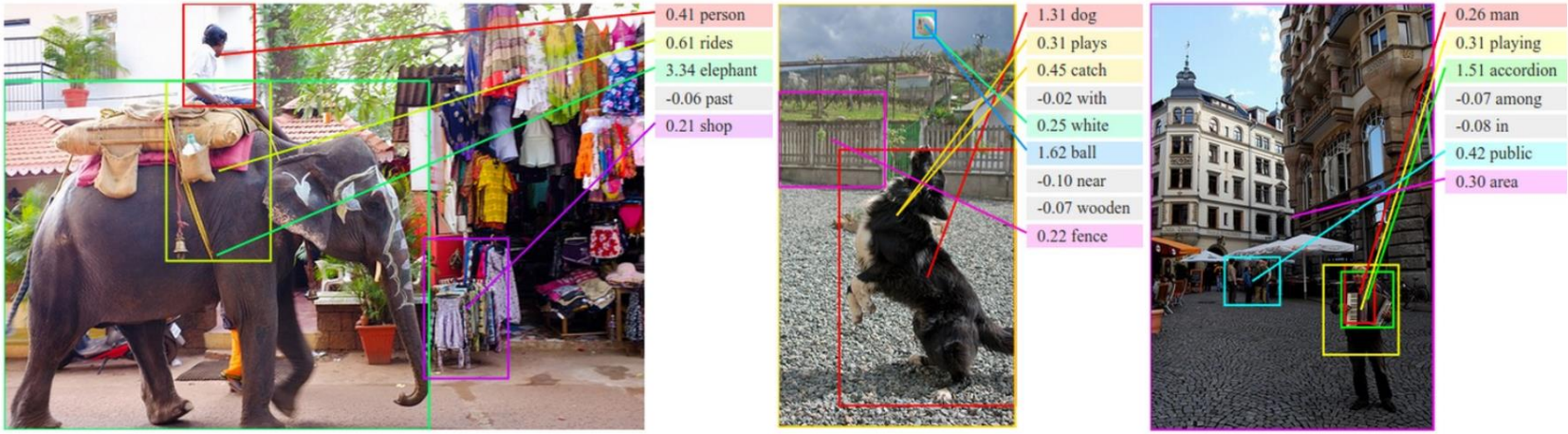
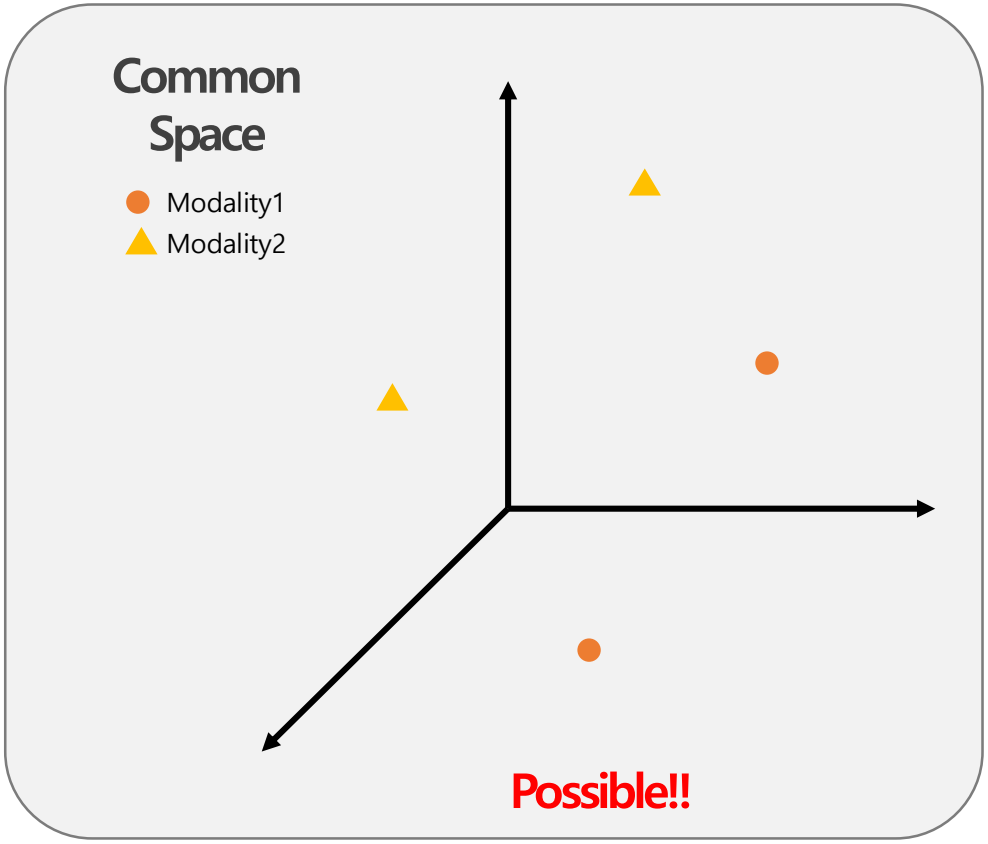


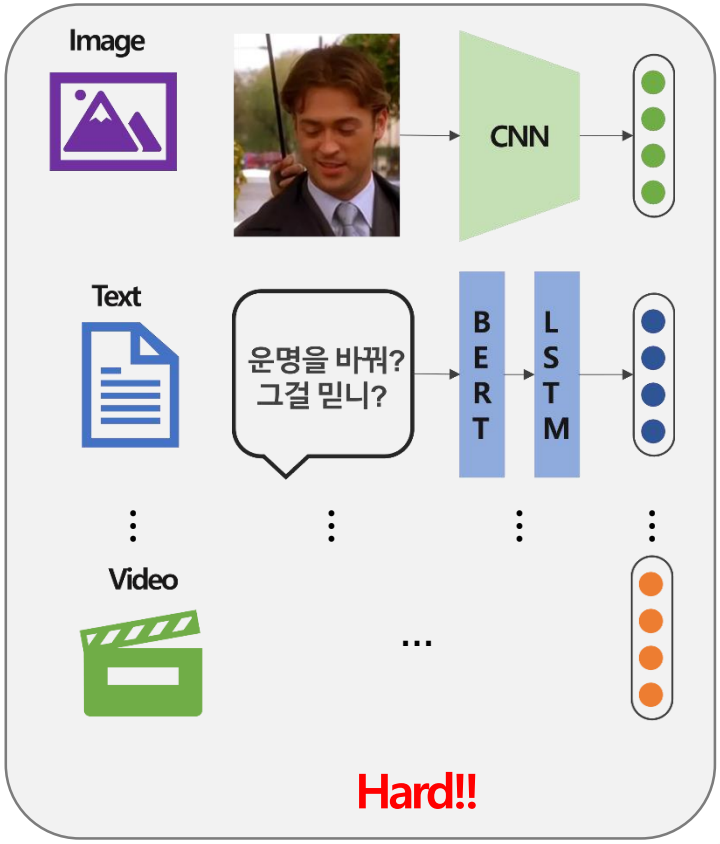
Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

How to narrow heterogeneity gap?

❖ Coordinated Representation 장단점

- 장점 : Modality-specific한 특징을 보존할 수 있다. (각 modality에 대해 분리된 representation을 학습 가능)
- 단점 : 2개 보다 많은 modality가 존재하는 경우 제약을 정의하는 것이 어려움



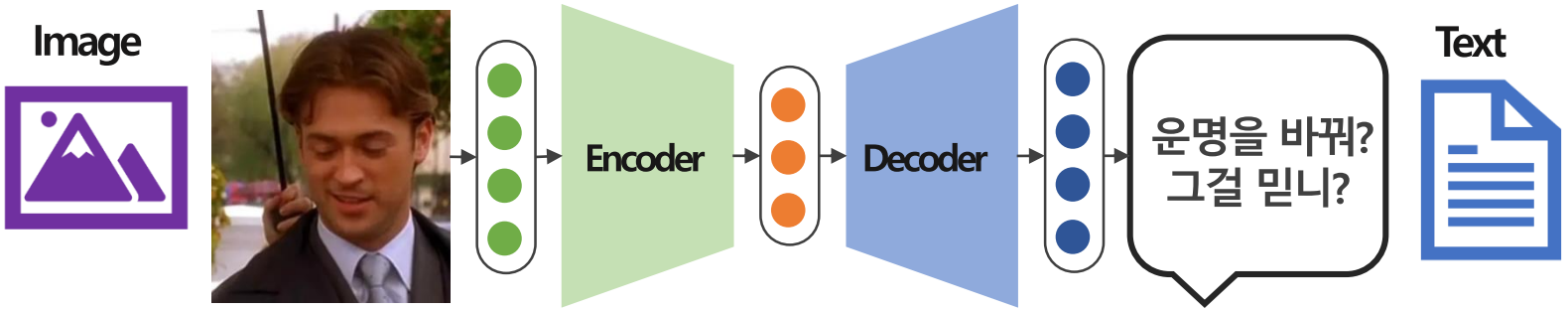


Encoder-decoder Representation

How to narrow heterogeneity gap?

❖ Encoder-decoder Representation

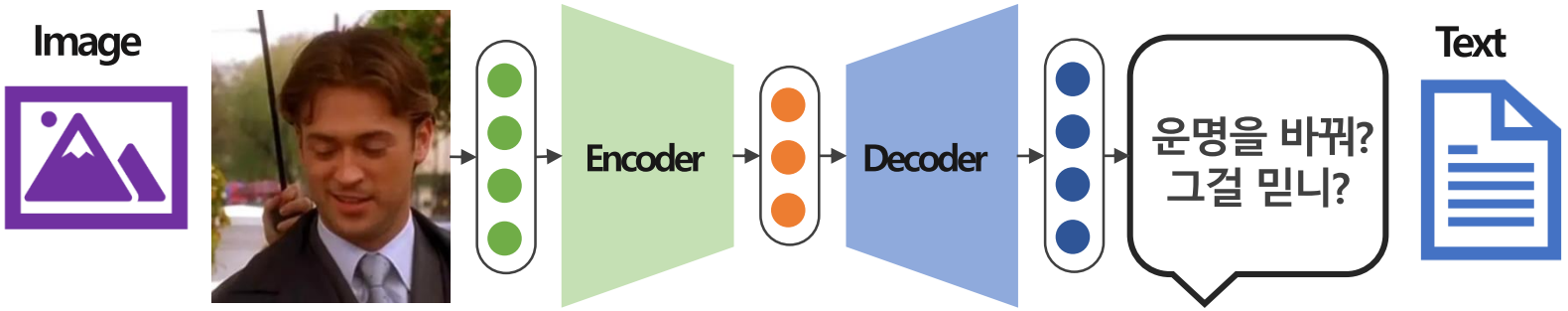
- 하나의 modality를 다른 modality의 representation space에 매핑
- Encoder : source modality -> latent vector
- Decoder : latent vector -> target modality



How to narrow heterogeneity gap?

❖ Encoder-decoder Representation

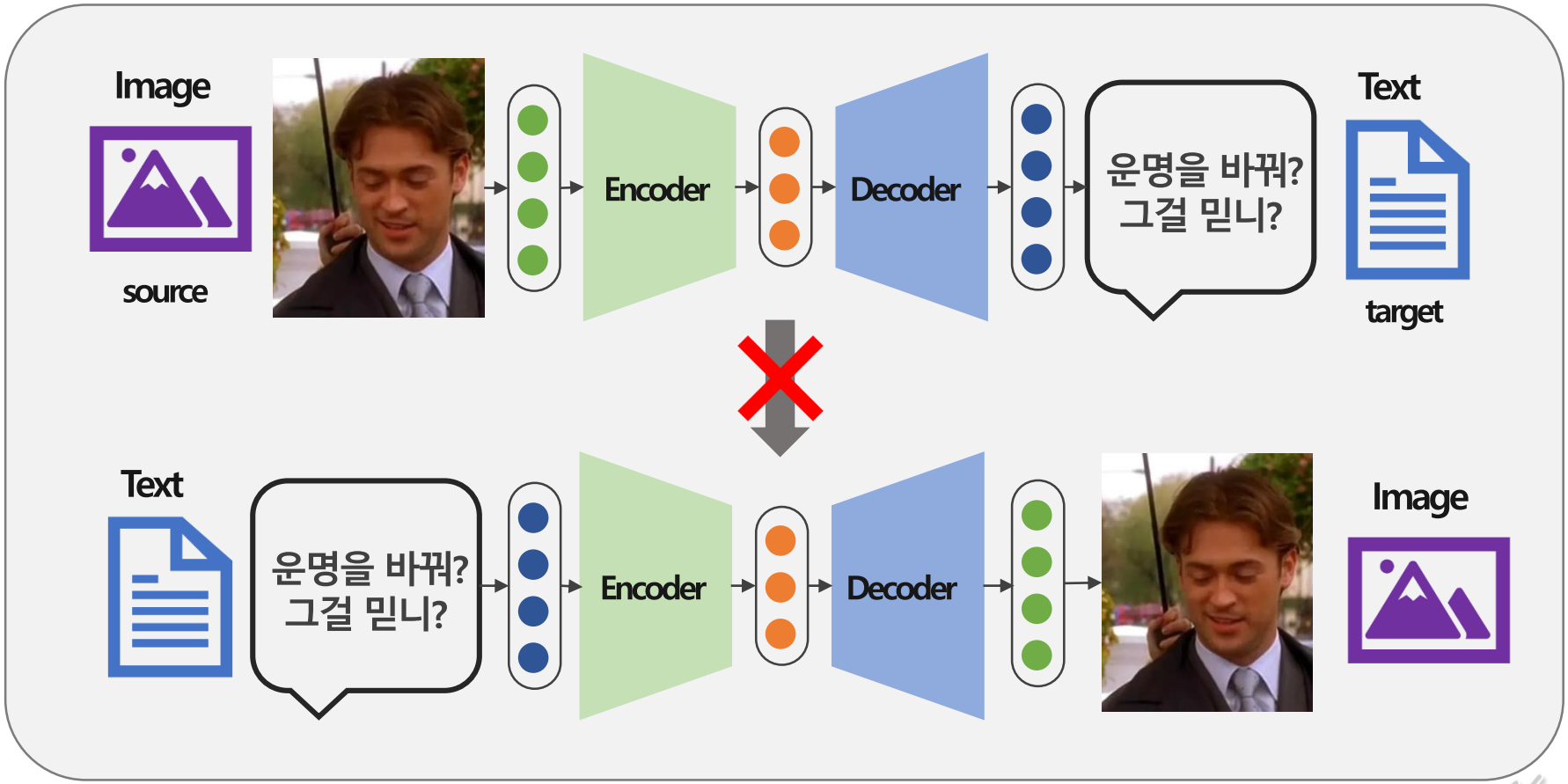
- $\theta^* = \underset{\theta}{argmax} \sum_{(v,s)} \log P(S|V;\theta)$
- 이미지로부터 문장을 생성한다면, 이미지가 주어졌을 때 해당하는 문장이 생성될 확률이 가장 커지도록 encoder와 decoder를 학습



How to narrow heterogeneity gap?

❖ Encoder-decoder Representation 장단점

- 장점 : source modality를 활용하여 target modality의 새로운 샘플 생성 가능
- 단점 : encoder와 decoder 모두 양식 중 하나만 인코딩 가능



Conclusion

Conclusion

❖ Conclusion

- 최근 modality간의 상호보완적 정보 공유를 가능하게 하는 multimodal representation vector를 생성하고 활용하는 연구가 많이 진행되고 있음
- multimodal representation vector를 생성하는 방법에는 크게 세가지가 있음
- Joint representation 방식, Coordinated representation 방식, Encoder-decoder 방식이 존재
- 세가지 방식의 장단점을 잘 고려하여 풀고자 하는 task에 적합한 방식을 활용하는 것이 중요
- 최근 여러 방식을 동시에 사용해서 representation vector를 생성하는 방법들도 생겨나고 있는데, 이처럼 앞으로도 더 좋은 multimodal representation vector를 생성하려는 연구가 활발하게 진행될 것으로 보임

Thank you