



# Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning

Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut

GOOGLE AI

2018/ACL

2024.02.13

이상민

# 목차

1. Abstract
2. Introduction
3. Conceptual Captions Dataset Creation
4. Image Captioning model
5. Experimental Results
6. Questions

# 1. Abstract

- Conceptual Captions dataset
  - 기존 MS-COCO dataset 보다 다양한 대규모 Conceptual Caption dataset 제안
- model architectures
  - Inception-ResNet-V2와 transformer를 사용한 image caption model 구조 제안

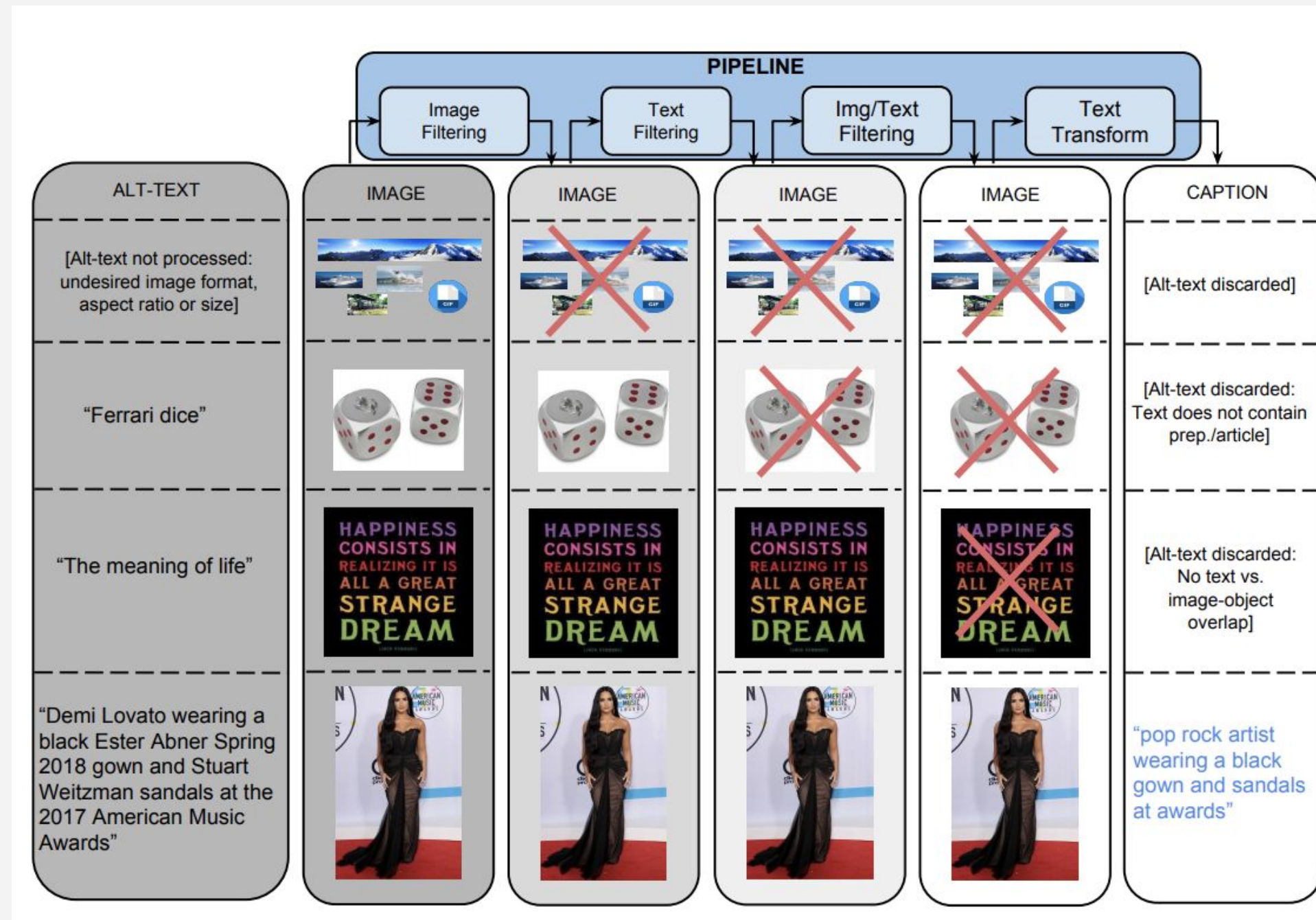
## 2. Introduction

- MS-COCO dataset
  - train data는 120,000개
  - 사람이 직접 신중하게 image를 고르고 caption을 묘사하기 때문에 image와 caption의 표현이 다양하지 않다.
- Conceptual Caption dataset
  - 3,300,000개의 <image, caption> 쌍으로 구성되어 있는 대규모 데이터셋
  - 수십억개의 web page에서 추출, 필터링 되었기 때문에 image와 caption의 표현이 다양하다.

### 3. Conceptual Captions Dataset Creation

- Pipeline

Webpage의 alt-text와 image로 부터 <image, caption> 쌍을 추출



# 3. Conceptual Captions Dataset Creation

## • *Image-based Filtering*

**pipeline**의 첫번째 필터로서 **image** 기반 필터링

- 이미지의 가로 혹은 세로 pixel이 400 pixel 이하인 이미지 제거
- 가로 세로 비율이 2:1 이상 차지하는 이미지 제거
- 자극적인 이미지 삭제

## • *Text-based Filtering*

두번째 필터링 구간으로 **webpage HTML**로 부터 **Alt-text**를 추출하는 구간

- 반복되는 토큰이 많은 텍스트 제거
- 대소문자 구별이 잘되지 않는 텍스트 제거
- alt-text의 모든 토큰이 wikipedia에 5번 미만 등장하는 텍스트 제거
- 자극적 혹은 폭력적인 내용이 있는 텍스트 제거

### 3. Conceptual Captions Dataset Creation

#### • *Image & Text-based Filtering*

- **Google cloud vision API**를 사용해서 image의 label을 예측
- 생성된 **label**과 **alt-text**의 어간(stem)을 비교한뒤 겹치지 않으면 제거

#### • *Text Transformation with Hypernymization*

Pipeline의 마지막 과정으로 image caption model의 학습 난이도를 낮추기 위한 단계

- 고유명사, 숫자, 단위 제거
- 날짜, 기간, 전치사 기반의 위치 (e.g., "in Los Angeles") 제거
- **knowledge-graph**를 이용해 named entity를 상위어로 교체 (e.g., 아이유 -> 가수)
- 위 과정을 거쳐 동일한 단어가 나오면 복수형으로 변경 (**actor and actor -> actors**)



### 3. Conceptual Captions Dataset Creation

- *Conceptual Captions Quality*

	GOOD (out of 3)		
	1+	2+	3
Conceptual Captions	96.9%	90.3%	78.5%

Pipeline을 통해 만들어진 dataset에서 임의로 추출한 4,000개의 데이터를 사람이 평가한 점수

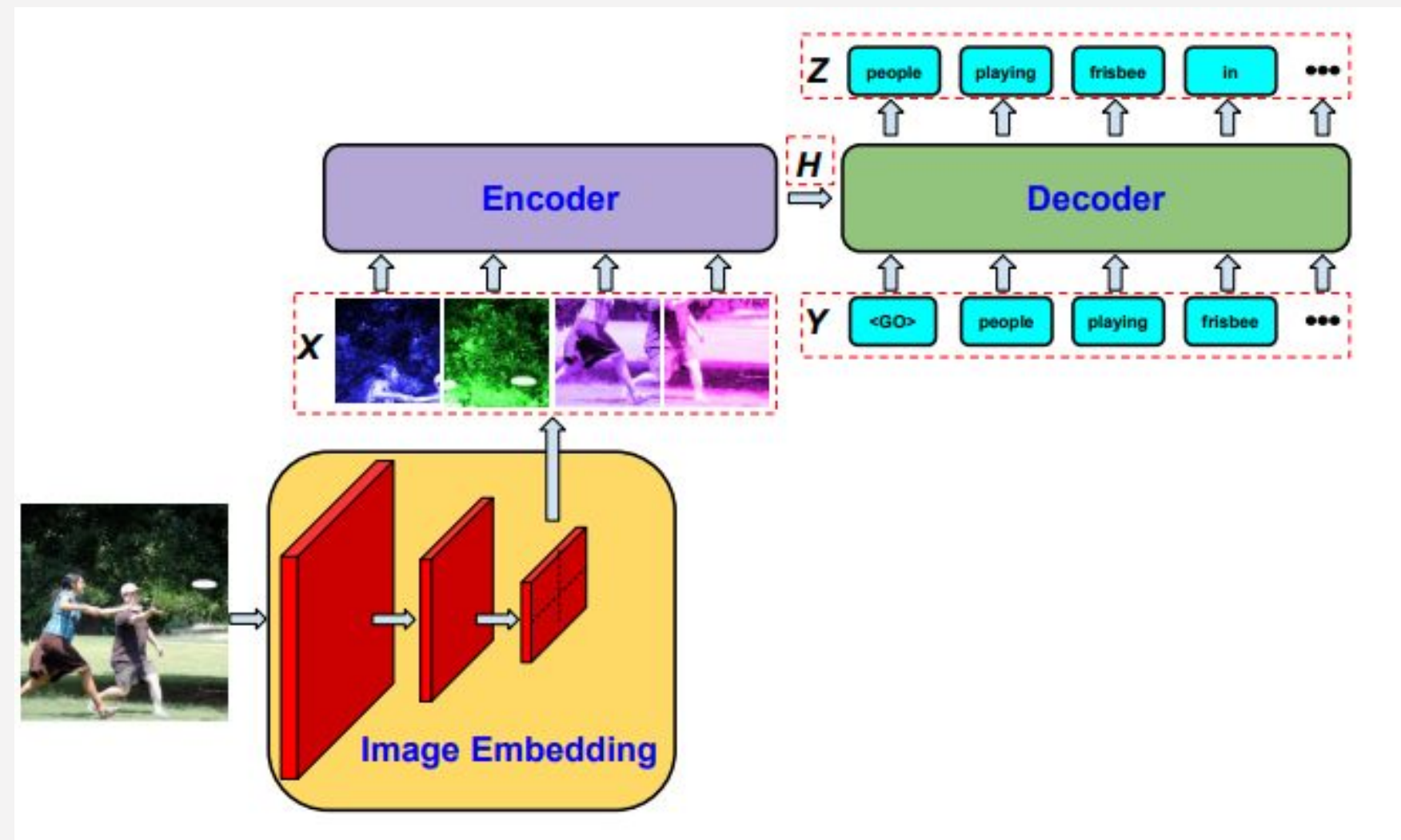
	Examples	Unique Tokens	Tokens/Caption		
			Mean	StdDev	Median
Train	3,318,333	51,201	10.3	4.5	9.0
Valid.	28,355	13,063	10.3	4.6	9.0
Test	22,530	11,731	10.1	4.5	9.0

Conceptual Captions의 train/valid/test 분할 분포



## 4. Image Captioning model

- main model architecture
  - 논문에서 사용한 image captioning model architecture
  - Inception-ResNet-V2, RNN, transformer 사용



## 4. Image Captioning model

- Image

- **Embedding**

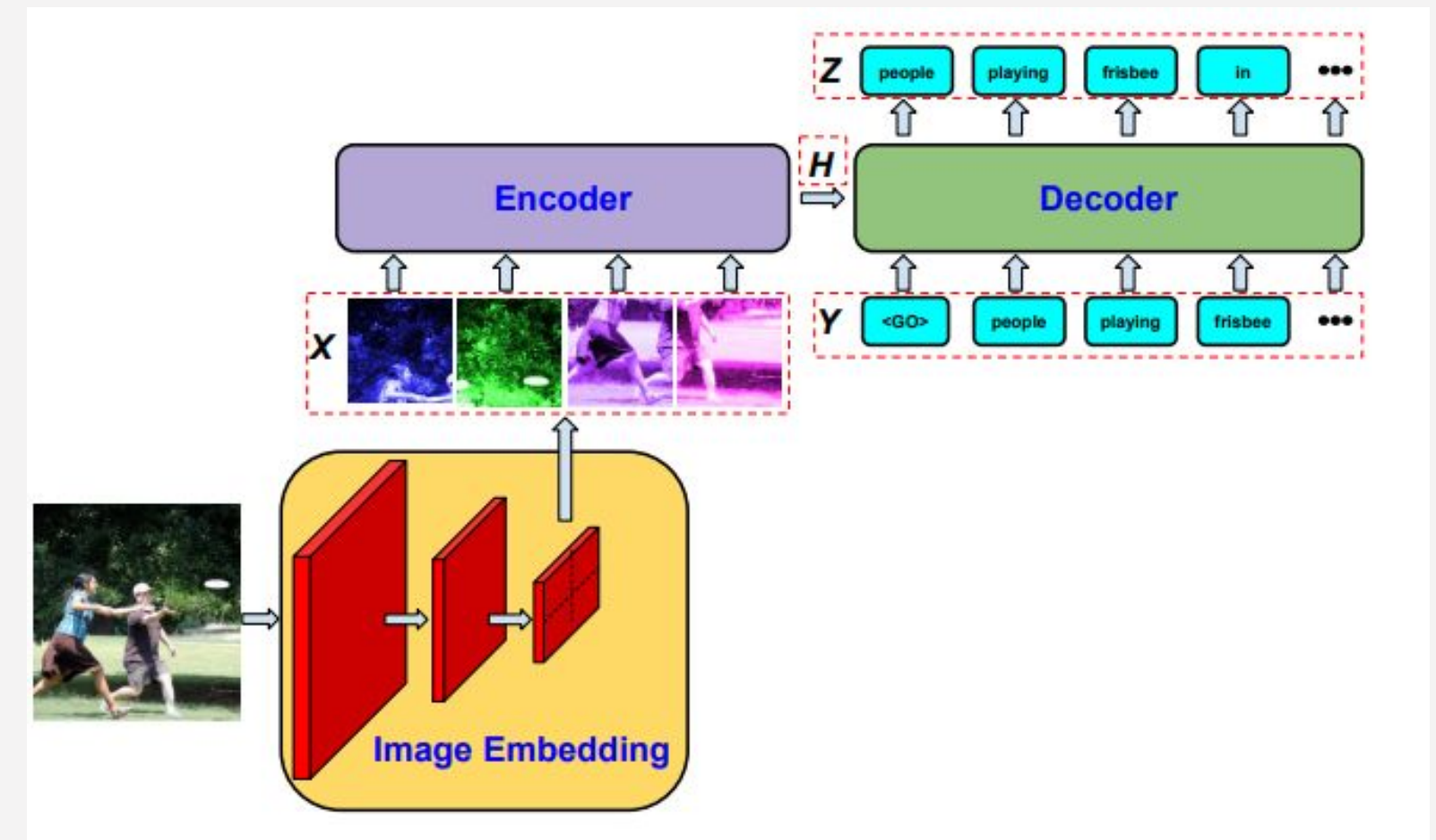
- 하나의 이미지를 64개(8x8)로 분할한 뒤 각 분할에 대해 image embedding을 추출

- **Encoder**

- image Embedding (8x8) 으로 부터 hidden state를 추출





- **Decoder**

- Encoder의 출력 결과를 이용해 다음에 존재할 토큰을 예측한다.
  - 이후에 이전 time step의 결과를 이용해 순차적으로 token을 예측



# 5. Experimental Results

• COCO-trained model VS Conceptual-trained model

				
<b>COCO-trained</b>				
RNN8x8	a group of men standing in front of a building	a couple of people walking down a walkway	a child sitting at a table with a cake on it	a close up of a stuffed animal on a table
T2T8x8	a group of men in uniform and ties are talking	a narrow hallway with a clock and two doors	a woman cutting a birthday cake at a party	a picture of a fish on the side of a car
<b>Conceptual-trained</b>				
RNN8x8	graduates line up for the commencement ceremony	a view of the nave	a child's drawing at a birthday party	a cartoon businessman thinking about something
T2T8x8	graduates line up to receive their diplomas	the cloister of the cathedral	learning about the arts and crafts	a cartoon businessman asking for help

- Conceptual-trained model이 COCO-trained model 보다 entity를 적합하게 표현한다.
- COCO-trained model의 경우 hallucination이 관찰됨

## 5. Experimental Results

- Human Evaluation Results

Model	Training	1+	2+	3+
RNN8x8	COCO	0.390	0.276	0.173
T2T8x8	COCO	0.478	0.362	0.275
RNN8x8	Conceptual	0.571	0.418	0.277
T2T8x8	Conceptual	0.659	0.506	0.355

**Table 4: Human eval results on Flickr 1K Test.**

사람이 평가한 평가 결과



## 5. Experimental Results

- Automatic Evaluation Results

Model	Training	CIDEr	ROUGE-L	SPICE
RNN1x1	COCO	0.340	0.414	0.101
RNN8x8	COCO	0.356	0.413	0.103
T2T1x1	COCO	0.341	0.404	0.101
T2T8x8	COCO	0.359	0.416	0.103
RNN1x1	Conceptual	0.269	0.310	0.076
RNN8x8	Conceptual	0.275	0.309	0.076
T2T1x1	Conceptual	0.226	0.280	0.068
T2T8x8	Conceptual	0.227	0.277	0.066

Table 7: Auto metrics on the Flickr 1K Test.

Flickr 1k Test에 대한  
평가지표

## 6. Questions

- 자동 평가지표의 한계점이 명확하지만 꾸준히 사용하는 이유?

감사합니다.