Open-Flamingo를 활용한 Multimodal Task 재현 및 성능 평가

CV 2팀 | 백성은, 강지윤

KUBIG
DATA SCIENCE & AI

CONTENTS

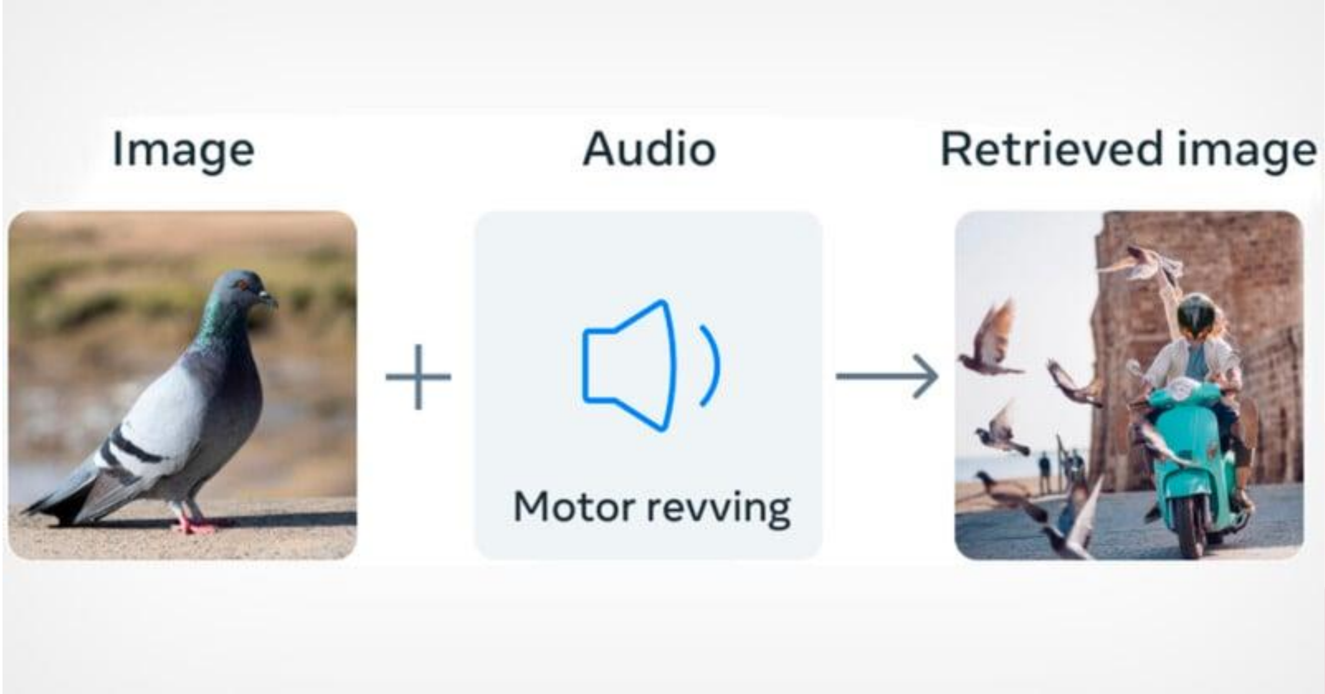
01. Introduction

01. Introduction

스터디 주제 : 'Multimodal' AI

= Multi (여러 개의) + Modal (modality, 양식 = 데이터 형식)

= 텍스트, 이미지, 영상, 음성 등 다양한 데이터 모달리티를 함께 고려하여 서로의 관계성을 학습 및 표현하는 AI



02. 모델 소개 및 논문 요약

02. 모델 소개 및 논문 요약

Flamingo를 선정한 이유 :

(1) Multimodal (Visual Language Model)

(2) 다양한 Open-ended VL tasks에 대해 few-shot learning을 통해 광범위하게 적용 가능한 범용성

→ 오픈 소스로 공개되어 있는 Open-Flamingo로 experiment 진행

DeepMind 28-04-2022

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac[*,‡], Jeff Donahue[*], Pauline Luc[*], Antoine Miech[*], Iain Barr[†], Yana Hasson[†], Karel Lenc[†], Arthur Mensch[†], Katie Millican[†], Malcolm Reynolds[†], Roman Ring[†], Eliza Rutherford[†], Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan[*,‡]

[*]Equal contributions, ordered alphabetically, [†]Equal contributions, ordered alphabetically, [‡]Equal senior contributions

Building models that can be rapidly adapted to numerous tasks using only a handful of annotated examples is an open challenge for multimodal machine learning research. We introduce Flamingo, a family of Visual Language Models (VLM) with this ability. Flamingo models include key architectural innovations to: (i) bridge powerful pretrained vision-only and language-only models, (ii) handle sequences of arbitrarily interleaved visual and textual data, and (iii) seamlessly ingest images or videos as inputs. Thanks to their flexibility, Flamingo models can be trained on large-scale multimodal web corpora containing arbitrarily interleaved text and images, which is key to endow them with in-context few-shot learning capabilities. We perform a thorough evaluation of the proposed Flamingo models, exploring and measuring their ability to rapidly adapt to a variety of image and video understanding benchmarks. These include open-ended tasks such as visual question-answering, where the model is prompted with

README | MIT license

OpenFlamingo

pypi package 2.0.1

Paper | Blog posts: 1, 2 | Demo

Welcome to our open source implementation of DeepMind's Flamingo!

In this repository, we provide a PyTorch implementation for training and evaluating OpenFlamingo models. If you have any questions, please feel free to open an issue. We also welcome contributions!

02. 모델 소개 및 논문 요약

Flamingo 모델이란?

- 2022년 Google DeepMind에서 나온 Visual Language Model (VLM)

→ Effective(few-shot), efficient(rapidly adapt), general-purpose model(various task)

= Few-shot learning(Task-specific 예시를 몇 개 학습시키는 것)으로도 Image나 Video understanding task를 단일 모델로 좋은 성능으로 수행 가능

등장 배경

(1) 기존 fine-tuning이 다수의 annotated dataset을 필요로 하고, Task별 hyper-parameter tuning을 다르게 해야 했기에 다양한 task에 대한 few-shot learning의 발전이 어려웠음.

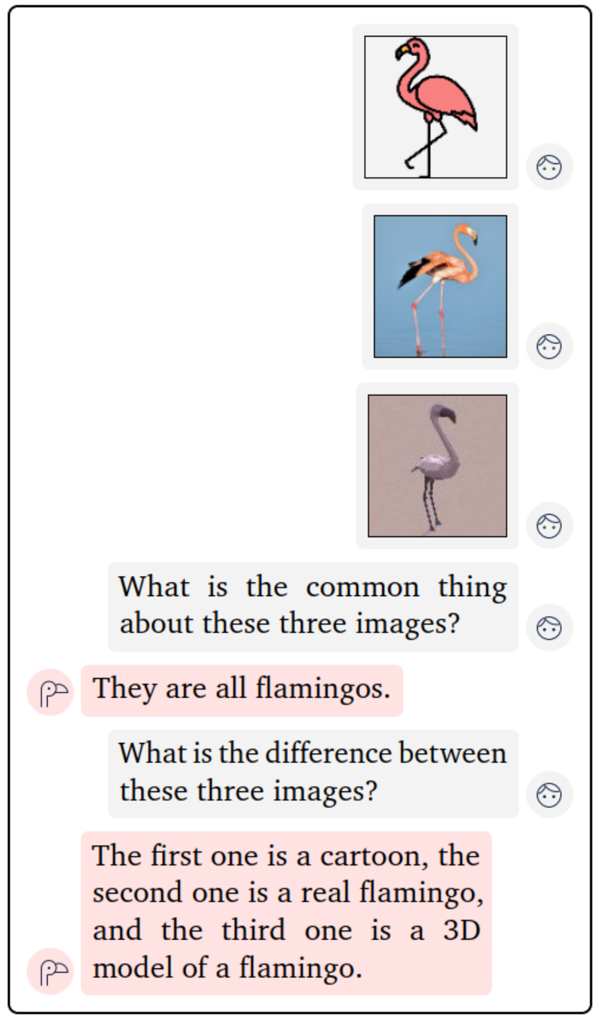
(2) 기존의 CLIP같은 모델은 새로운 task에 대해 뛰어난 zero-shot adaptation 능력을 보였으나, 이미지 분류 같은 문제에서만 효과적이고, 텍스트를 생성해내야 하는 open-ended tasks에선 취약했음.

https://deepmind.google/discover/blog/tackling-multiple-tasks-with-a-single-visual-language-model/

02. 모델 소개 및 논문 요약

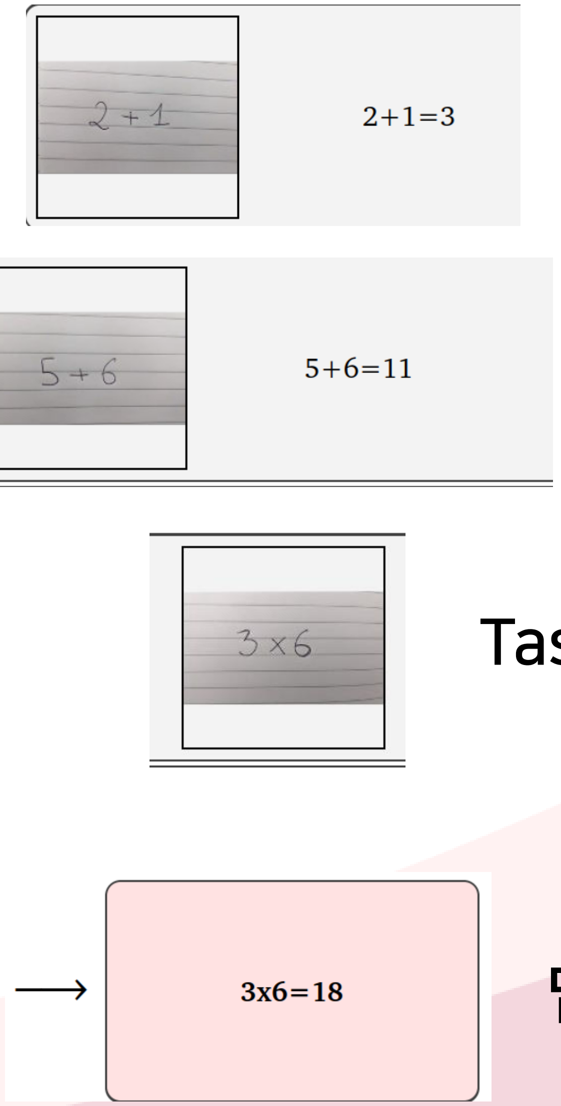
Flamingo 모델이 쓰이는 예시 : 이미지, 비디오, 텍스트 understanding Task

Visual Dialogue



Single Image + Text Prompt를 통한 Inference

Input:
Few example로 prompt



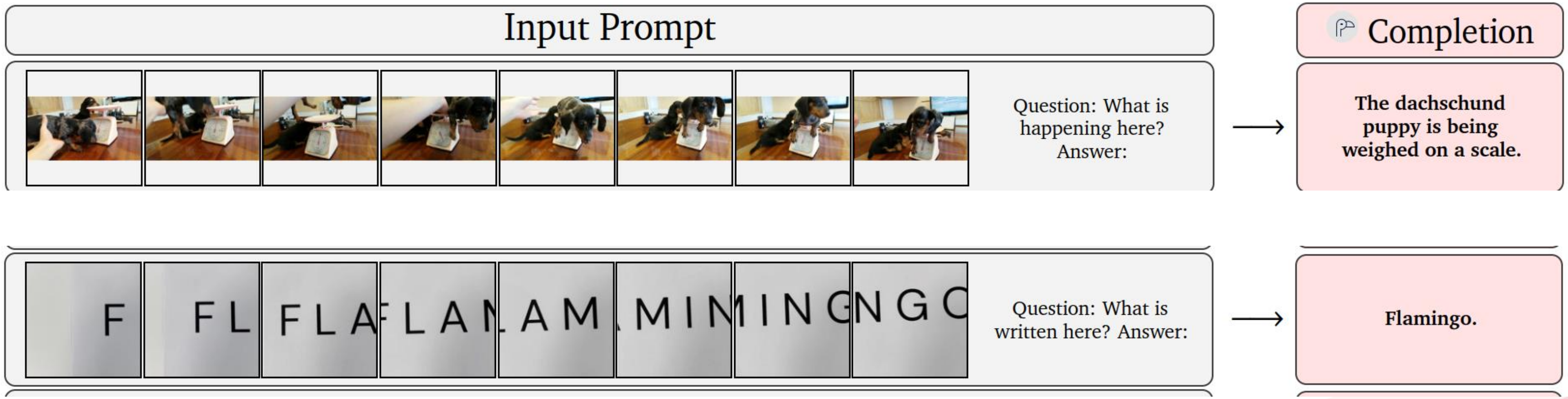
Task

답

02. 모델 소개 및 논문 요약

Flamingo 모델이 쓰이는 예시 : 이미지, 비디오, 텍스트 understanding Task

Video + Text Prompt를 통한 Inference



02. 모델 소개 및 논문 요약

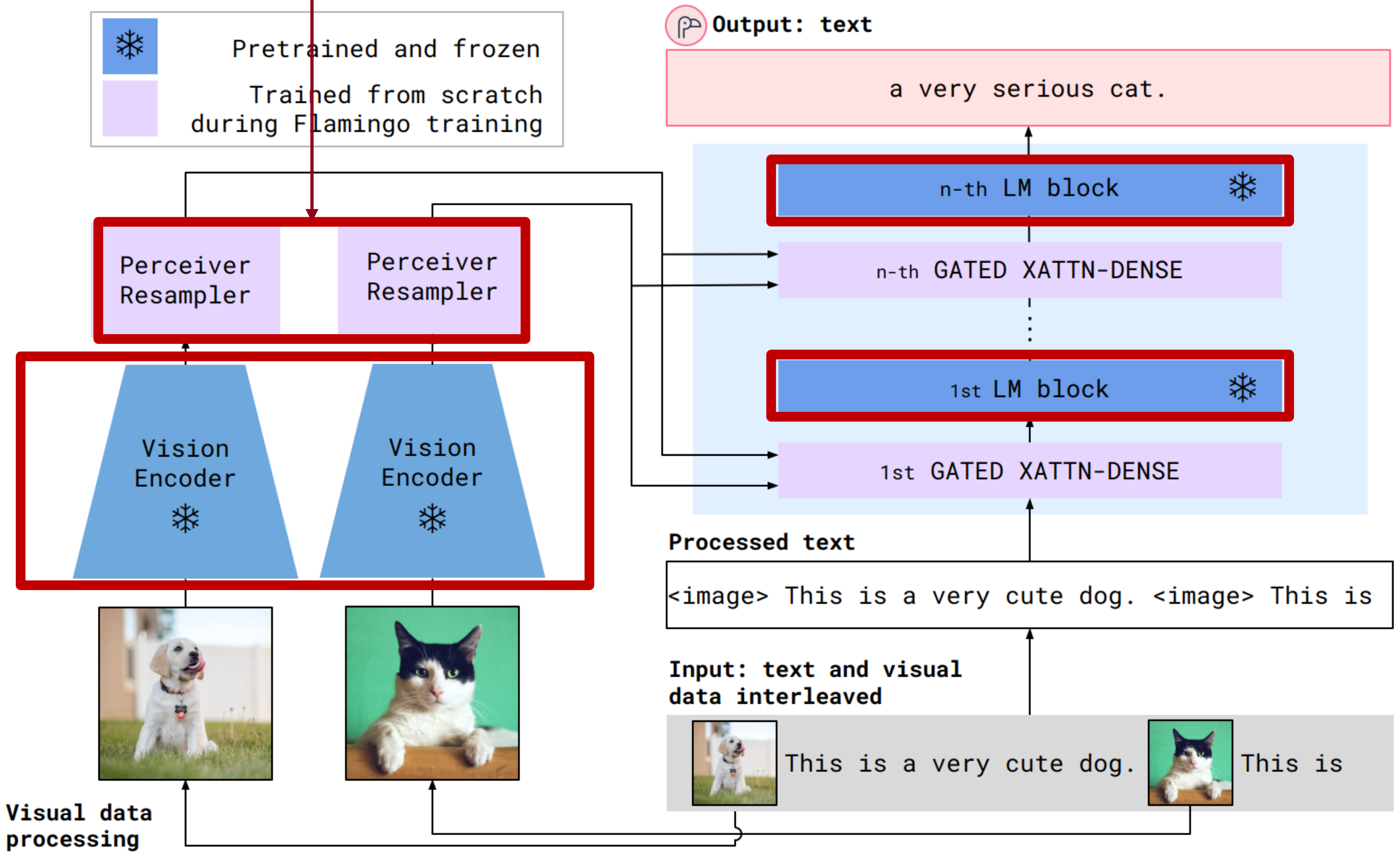
Flamingo 모델 아키텍처

Encoder-Decoder 구조 (Vision Encoder + Language Model)

Perceiver Resampler:
Vision encoder와 frozen LM을 이어주는 부분

- Vision feature들을 일정한 개수의 visual outputs으로 나오게 함
- text only LM에서 visual 정보도 처리할 수 있도록 함

Vision Encoder:
NFNET(2021), BERT와 CLIP Loss로 pre-training (Freeze)



Input: Free-form 이미지(비디오) + 텍스트 Sequence
Output: Free-form 텍스트

Language Model
: Chinchilla-70B (2022)
: Text만 처리 가능한 LM

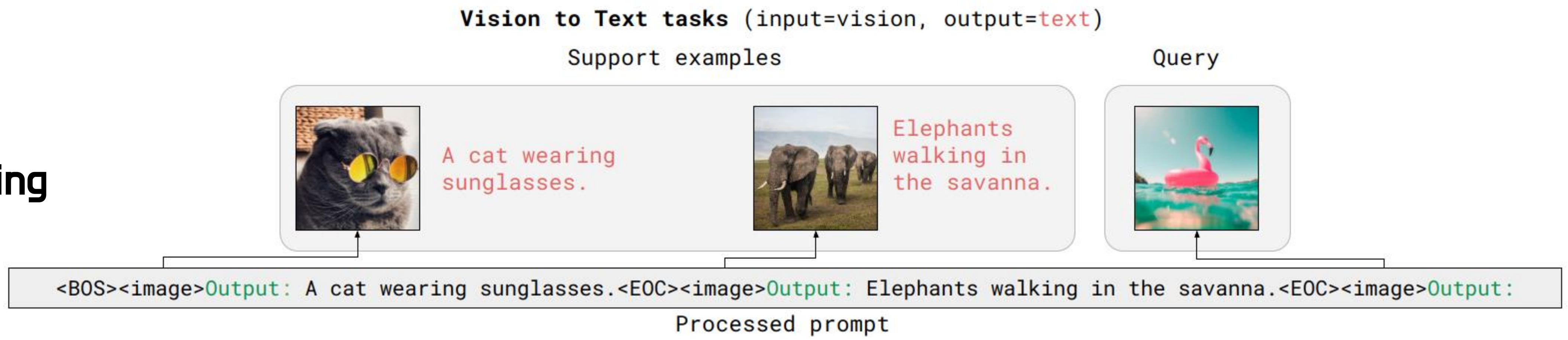
02. 모델 소개 및 논문 요약

Flamingo 모델의 Rapid Adaptation

(image,text) 형태의 example pair로 few-shot learning
→ 같은 prompt를 input하기만 하면 해당 task 수행

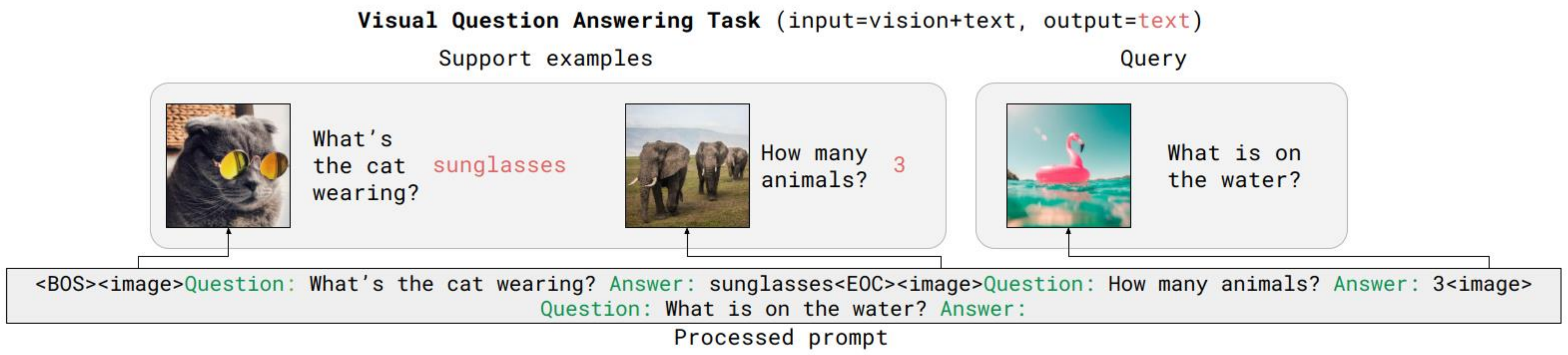
- In-context learning을 통해 새로운 Task에 빠르게 적용
- Flamingo를 few-shot을 통해 한 번 학습시키고, prompt를 condition해주기만 하면 새로운 task에도 적용할 수 있음

Image Captioning



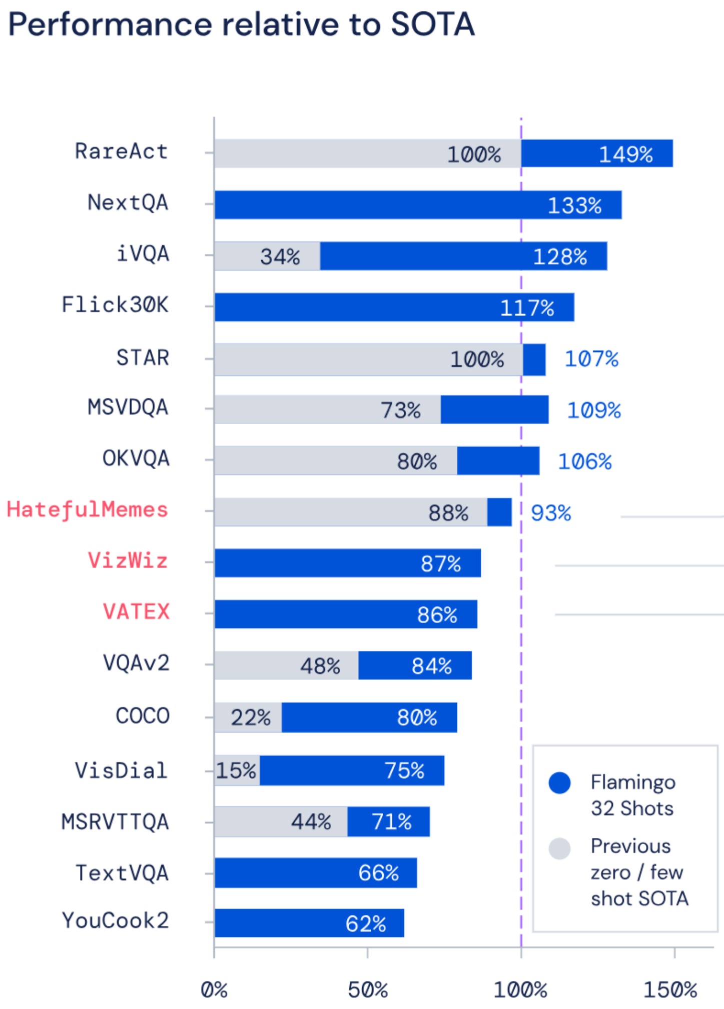
→ Prompt에 따라 다양한 task 수행 가능

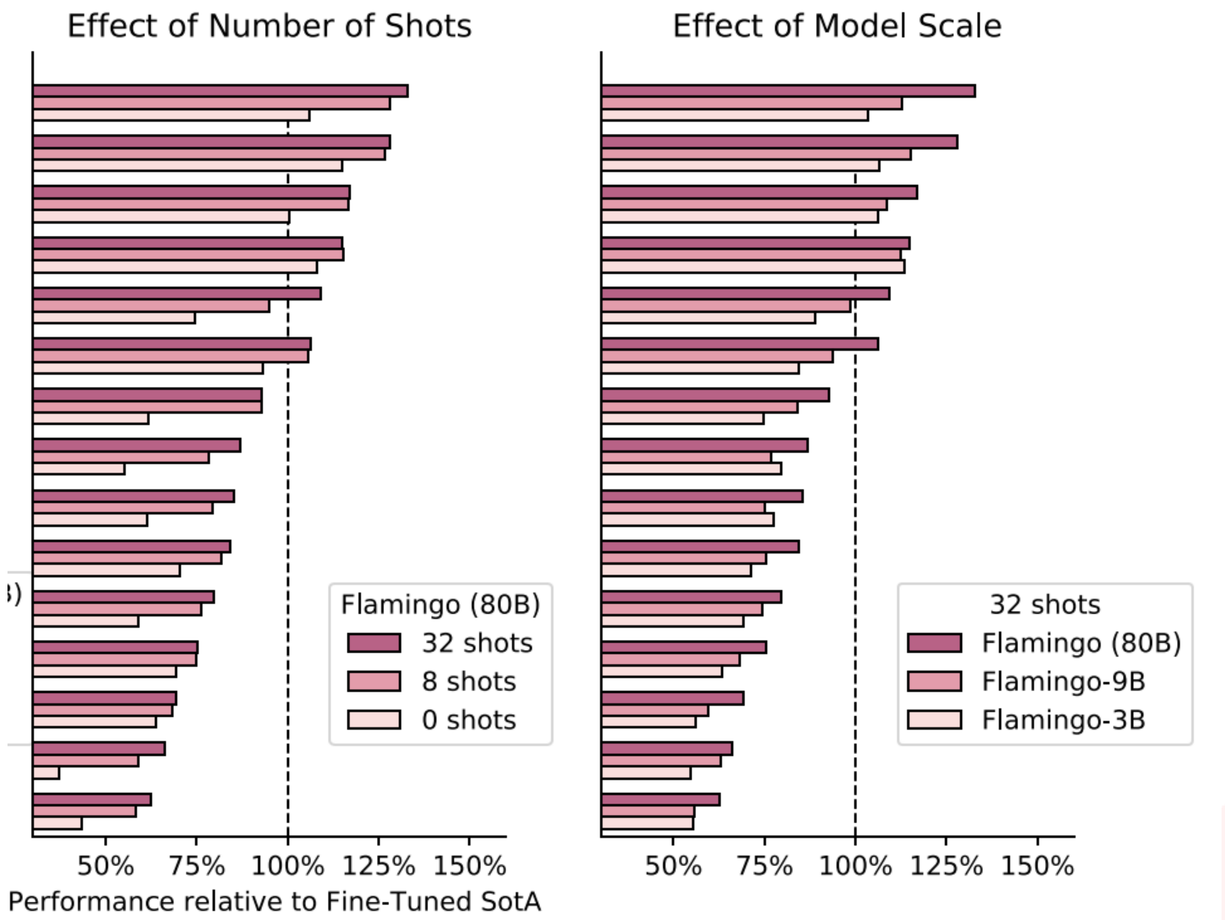
VQA



02. 모델 소개 및 논문 요약

Flamingo 모델의 성능





다양한 Task 수행 가능 + 각 Task에 대한 성능 좋은 평가

- 0 shots < 8 shots < 32 shots

03. 수행한 Task

03. Task List

- Image Classification
  - Hateful Memes
- Visual Question Answering (VQA)
  - Vizwiz
  - Textvqa
- Image Captioning (Qualitative)
  - COCO

03. Image Classification

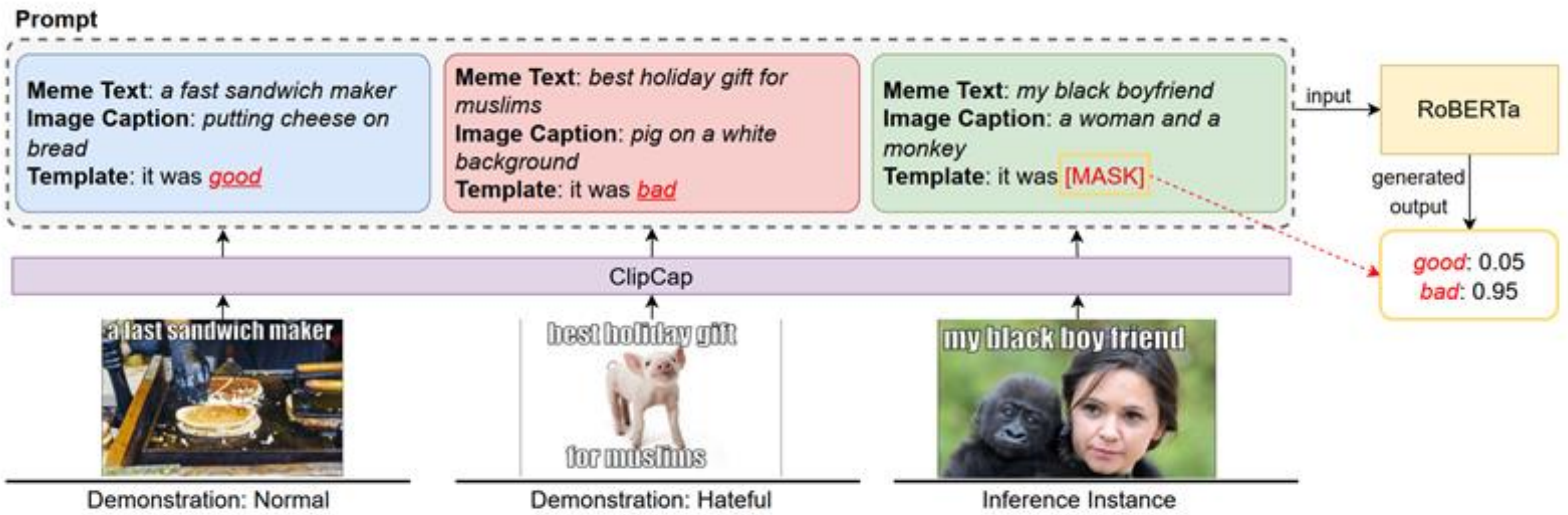
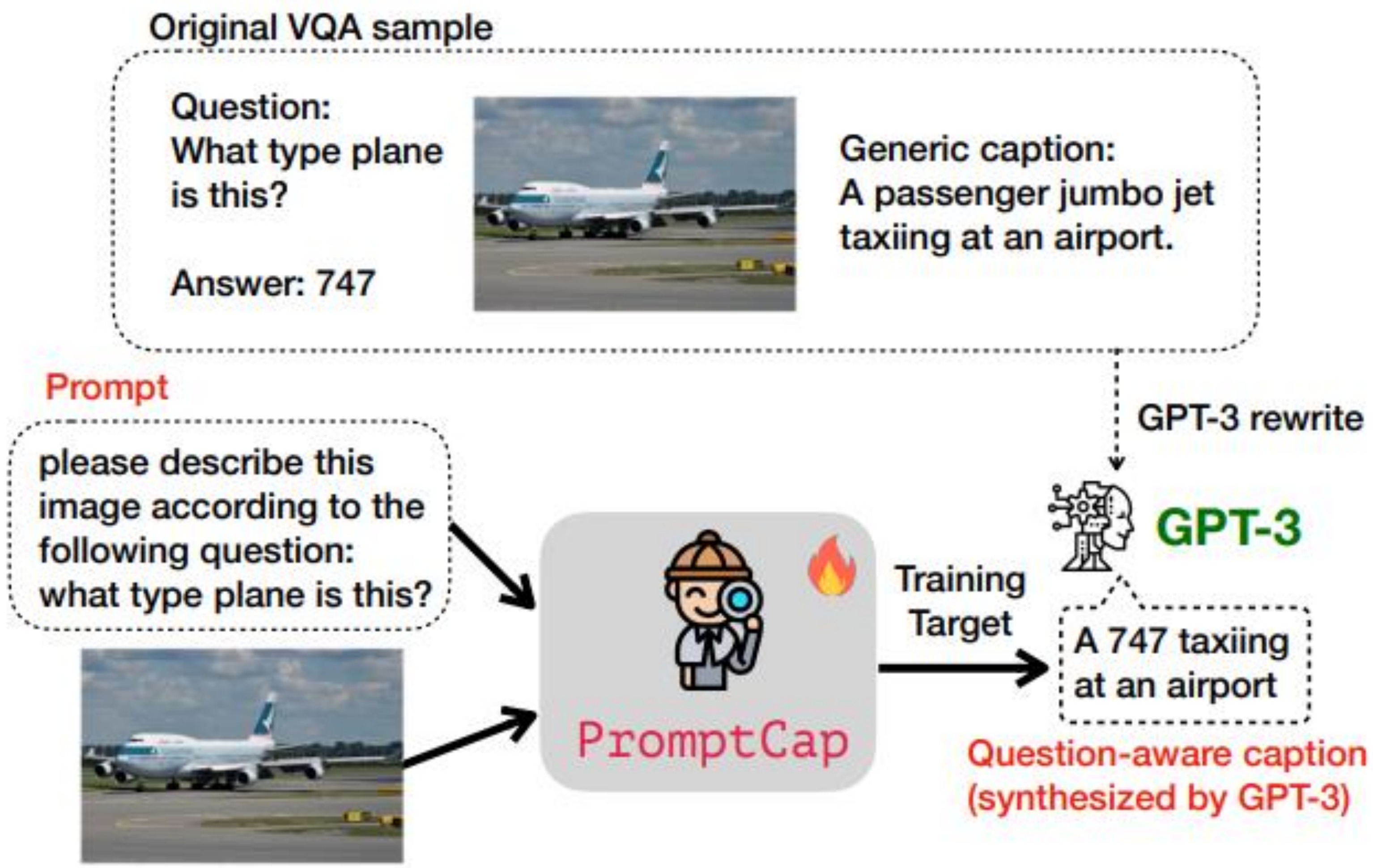
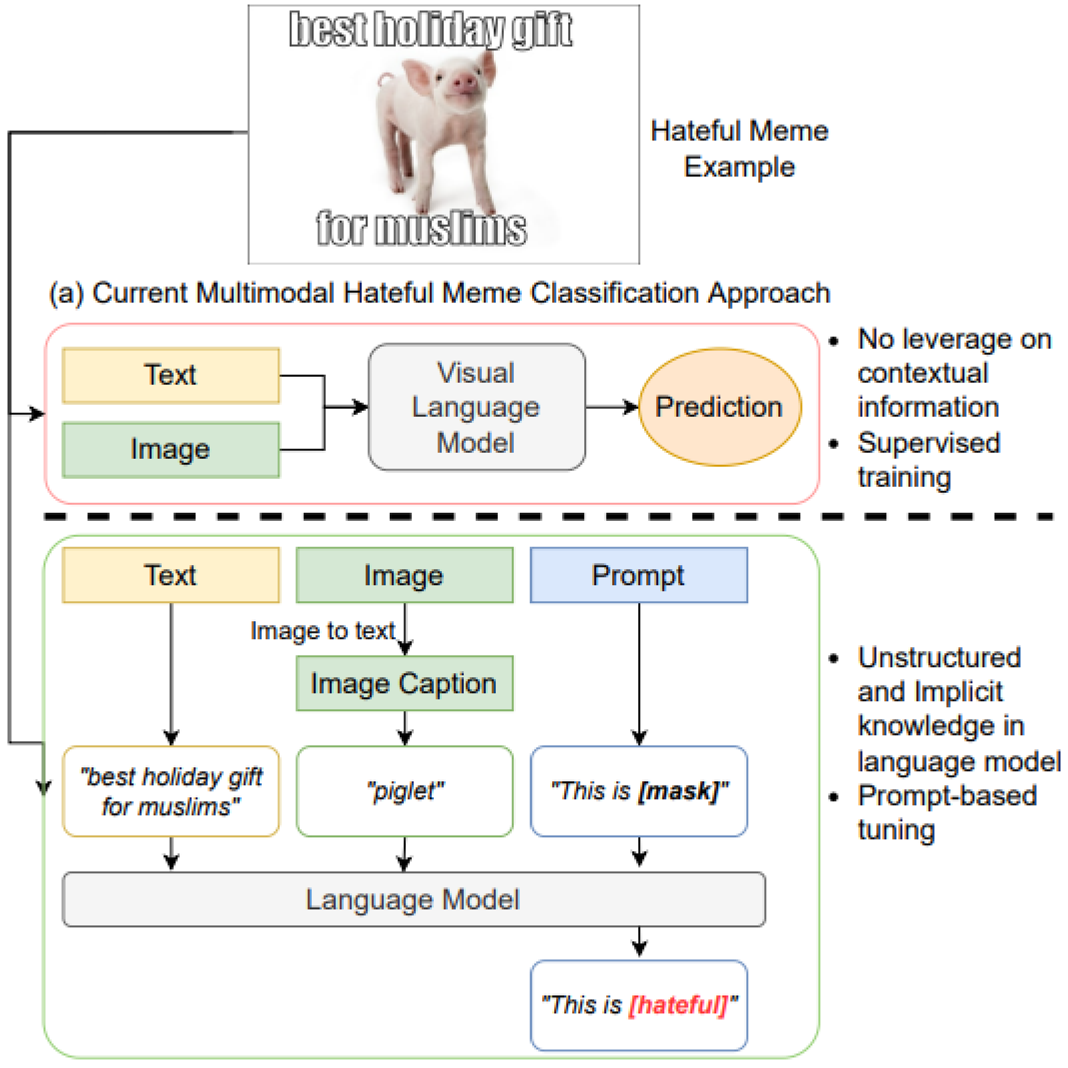


Figure 2: Overview of PromptHate Framework.

인종, 성별, 종교 등에 대한 부정적인 내용을 담고 있는 memes을 분류

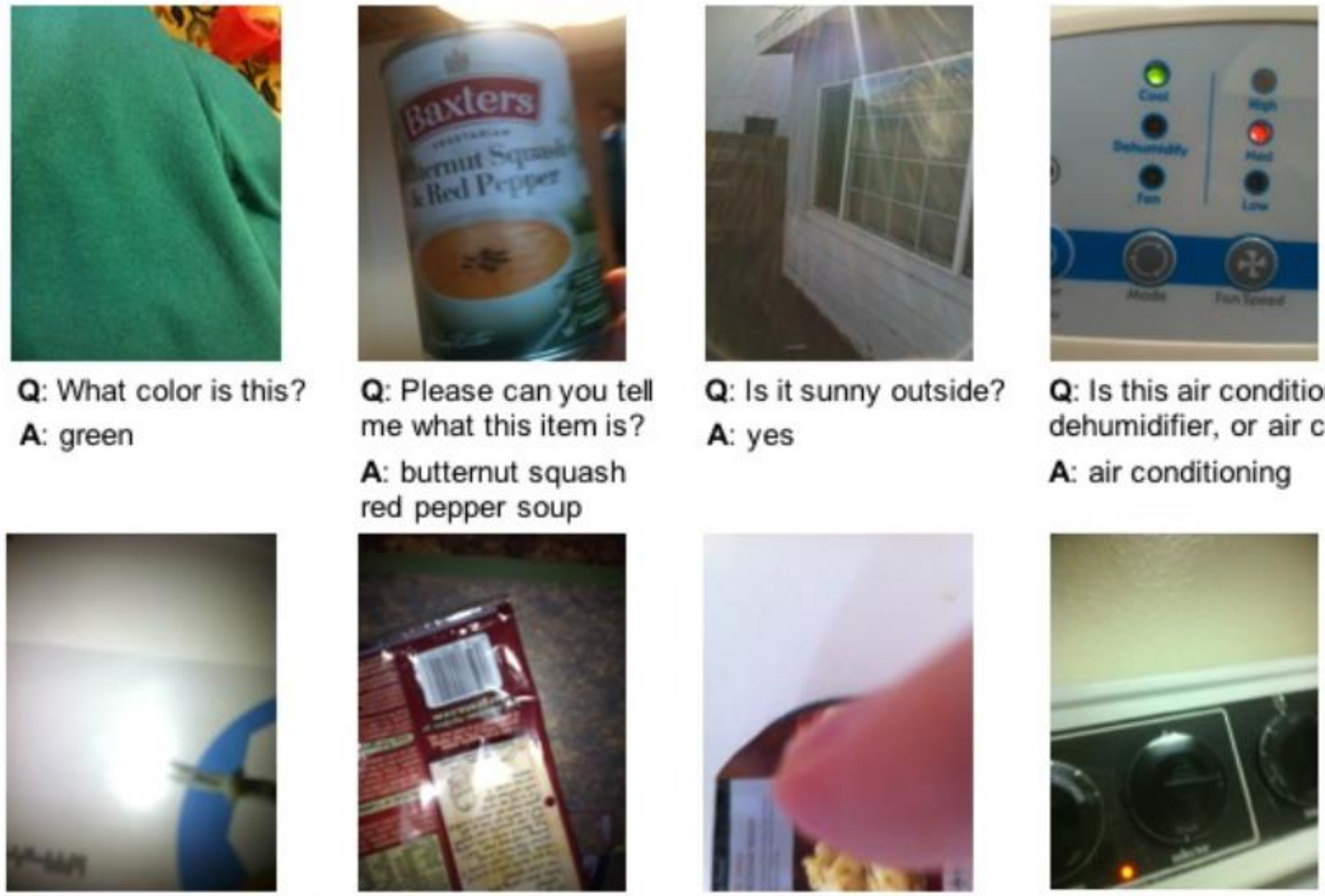
03. Image Classification (Prompt Engineering)





PromptCap + PromptHate를 활용한 prompt engineering 추가

03. Visual Question Answering



Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

Q: Is it sunny outside?
A: yes

Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning

Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable

Q: What is this?
A: unanswerable

Q: Can you please tell me what the oven temperature is set to?
A: unanswerable

03. Image Captioning

[Caption]

- A woman in yellow is hitting a tennis ball on a clay court.
- A tennis player prepares to return the ball.
- Etc.

04. Experiment Result

04. Experiment Result (Classification)

Shots	Mean	Std	Prompt Engineering
0	0.4985	0.02	X
0	0.502	0.017	O
4	0.4714	0.02	X

- Zero-shot에서 prompt engineering을 했을 때, 성능 향상
- Few-shot이 Zero-shot보다 성능이 낮다? → 충분한 example을 제시 X

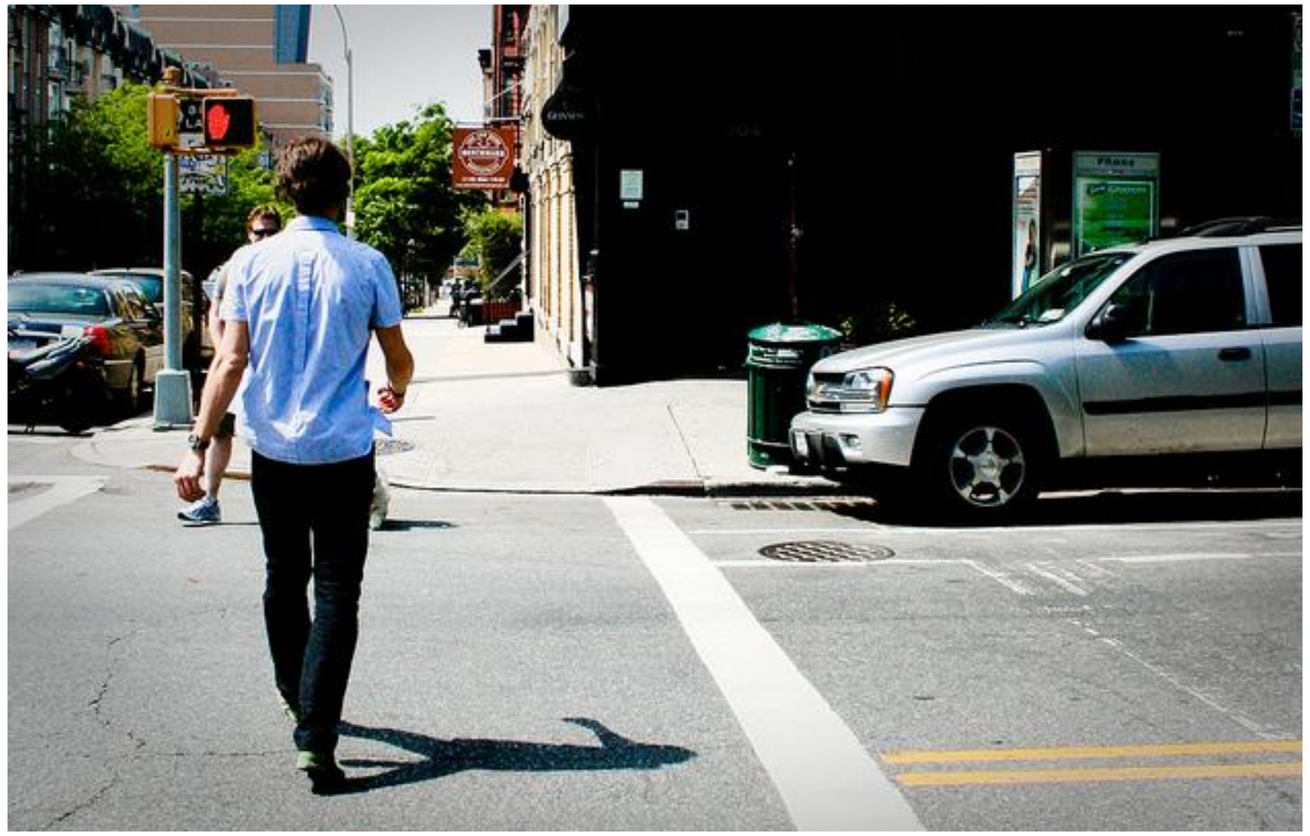
04. Experiment Result (VQA)

Vizwiz			Texevqa		
Shots	Baseline	Accuracy	Shots	Baseline	Accuracy
0	15.4	18.51	0	15.4	18.51
2	-	18.49	2	-	18.49
4	23.2	23.76	4	23.2	23.76

- 논문의 baseline과 거의 일치하는 결과 재현
- 일부 오차들은 seed, random initialization 등에서 비롯된 현상

04. Experiment Result (Image Captioning)

- A group of skiers in the mountains → Correct !



- A man walking down a street → incorrect !

04. Experiment Result (Image Captioning)

- A man sitting in front of a store → Correct !

- An elephant in Temple → Ambiguous

05. 결론 및 한계

05. Conclusion

- Multimodal open-source인 Open-Flamingo 구현
  - Zero-shot & Few-shot 성능 재현
  - Multimodal에 대한 이해도 향상
- Prompt Engineering을 이용하여 classification 성능 향상
  - PromptCap + PromptHate

05. Limitation & Future work

- Colab의 Resource 한계로 다양한 실험 진행의 어려움
  - Image captioning
  - 8-shot 이상 inference 불가

- 논문에 소개된 task 외에 추가 task 진행 계획
  - Video Question Answering
  - Model Architecture 수정 + hyper-parameter 튜닝

Thank You