

2023 분석 미니프로젝트 1

족제비즈

21기 분석 미니프로젝트 1
김범준 김도하 서재은 이건하 임수현



CONTENTS*

01 Intro

- 멀티 모달, VQA
- VQA Model

02 학습 과정

- 데이터셋
- 모델 학습 전략

03 실험 결과

- 평가 방법-유형 별 질문
- 모델 결과 비교


Intro

멀티모달, VQA

> VQA: Visual Question Answering

- 이미지 기반 질의응답 AI 모델 개발
- 이미지와 그 이미지에 대한 질문이 주어졌을 때 올바른 답변을 만들어 내는 task

1 이미지 입력



2 질문 입력

What's the animal doing?

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

| | |
|-------------|-------|
| laying down | 0.779 |
| sitting | 0.718 |
| resting | 0.635 |
| laying | 0.580 |
| relaxing | 0.292 |

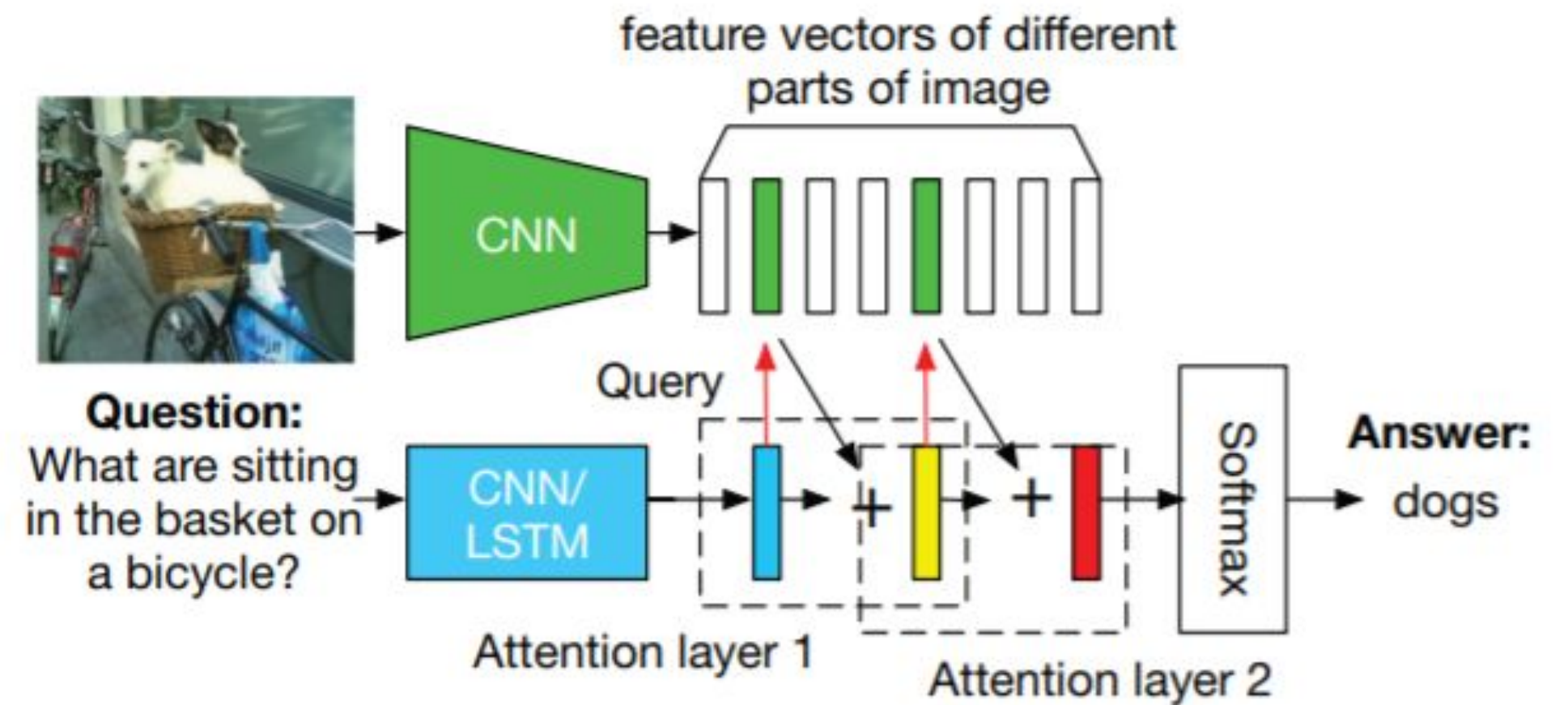
3 정답 출력

Intro

멀티모달, VQA

> VQA Model

- vision model을 이용하여 이미지를 처리하고 language model을 이용하여 질문을 해석하여 이미지와 질문에 맞는 정답을 도출한다



학습과정

데이터셋 전처리



> 데이터셋

01

- 이미지의 id
- 해당 이미지와 관련된 질문이 담긴 csv 파일이 실제 이미지와 함께 데이터셋
- Train 데이터: 107,231이미지, 359521질문
- Test 데이터: 11,915이미지, 40479질문



TRAIN_000029,train_000029,What game is the man playing?,tennis

> 데이터 전처리

02

- Closed Ended Questions 행 제거
- Closed Ended Question: yes/no 답변을 필요로 하는 질문

학습과정

모델 학습 전략



> 모델

- 데이터에서 제공된 Baseline을 기반으로 다양한 Vision Model, Language Model을 조합

Vision Model: ResNet, ViT, EfficientNet, CLIP

Language Model: GPT-2, BERT

ResNet+ GPT2

ResNet+ BERT

ViT + GPT2

ViT + BERT

EfficientNet+ GPT2

EfficientNet+ BERT

CLIP + GPT2

- VQA Task에서 SOTA성능을 보여주는 BLIP을 fine-tune

Visual Question Answering (VQA) on VQA v2 val

SOTA 성능을 보이고 있고 VQA Task에서도 뛰어난 성능을 보여주고있다

(출처 : <https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-val>)

실험결과

학습된 모델 평가



> 질문 유형

Activity Recognition

(e.g., “Is this man crying?”)

Object Detection

(e.g., “How many bikes are there?”)

Fine-Grained Recognition

(e.g., “What kind of cheese is on the pizza?”)

Commonsense Reasoning

(e.g., “Does this person have 20/20 vision?”,
“Is this person expecting company?”)

Knowledge Base Reasoning

(e.g., “Is this a vegetarian pizza?”)

실험결과

> 전체 맞춘 개수

01

선정된 20개의 질문 중 답을 맞춘 개수 비교

- BLIP v1이 18 / 20개로 가장 많이 맞춤
- ViT+GPT2, ResNet+GPT2가 3/20로 가장 적게 맞춤

| 모델 | ViT+BERT | ViT+GPT2 | ResNet+BERT | ResNet+GPT2 |
|-------|-----------------------------|----------------------------|----------------------------|----------------------------|
| 맞춘 개수 | 1 / 3 / 3 / 2 / 1 → 10 / 20 | 0 / 0 / 1 / 0 / 2 → 3 / 20 | 0 / 3 / 0 / 1 / 1 → 5 / 20 | 0 / 2 / 1 / 0 / 0 → 3 / 20 |

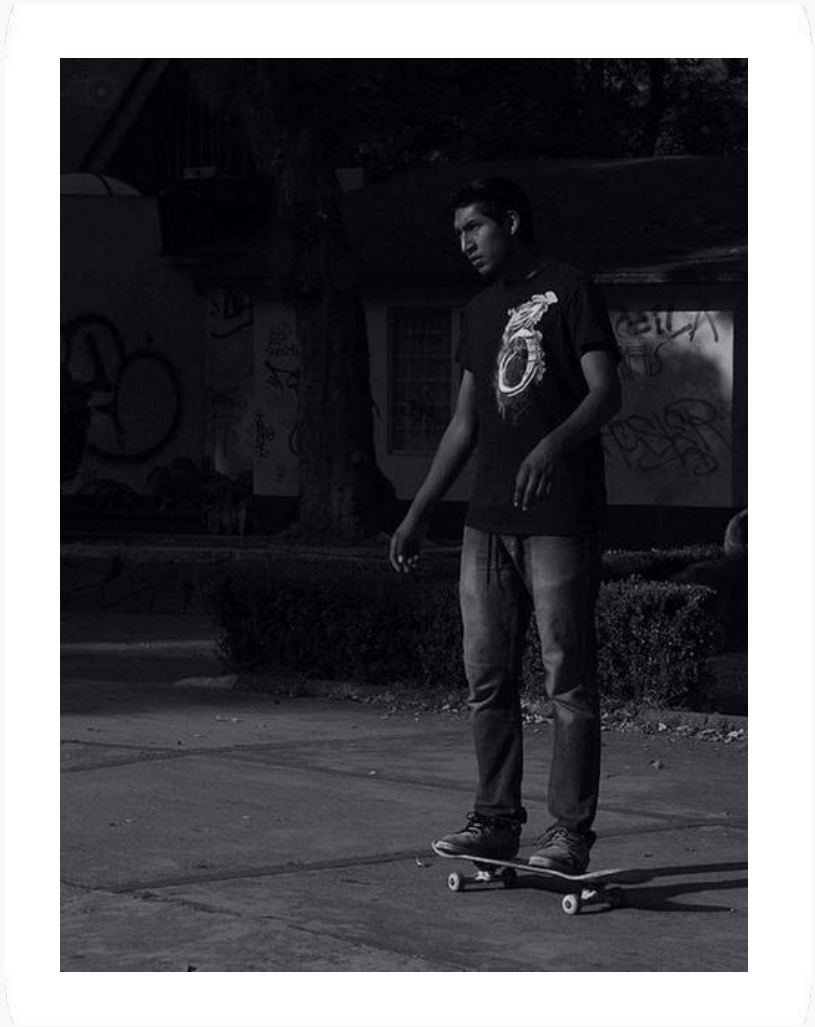
| 모델 | EfficientNet+BERT | EfficientNet+GPT2 | CLIP(ViT)+GPT2 | BLIP v1 |
|-------|-----------------------------|----------------------------|----------------------------|-----------------------------|
| 맞춘 개수 | 0 / 1 / 4 / 3 / 2 → 10 / 20 | 0 / 2 / 3 / 1 / 0 → 6 / 20 | 0 / 1 / 1 / 3 / 1 → 6 / 20 | 4 / 3 / 4 / 3 / 4 → 18 / 20 |

실험결과

질문 유형 별 결과 비교



> Activity Recognition



What activity is the person taking part in?

| 모델 | BLIP v1 | EfficientNet + BERT | ViT + BERT |
|----|---------------|---------------------|---------------|
| 답변 | skateboarding | skateboarding | skateboarding |

실험결과

질문 유형 별 결과 비교



> Object Detection



How many cookies can be seen?

| 모델 | BLIP v1 | EfficientNet + BERT | ViT + BERT |
|----|---------|---------------------|------------|
| 답변 | 0 | 3 | 0 |

실험결과

질문 유형 별 결과 비교



> Fine-Grained recognition



What is in the left corner?

| 모델 | BLIP v1 | EfficientNet + BERT | ViT + BERT |
|----|---------------|---------------------|------------|
| 답변 | surfing board | lamp | lamp |

실험결과

질문 유형 별 결과 비교



> Commonsense Reasoning



What alphabet letter is formed where the two mountain look like they touch each other?

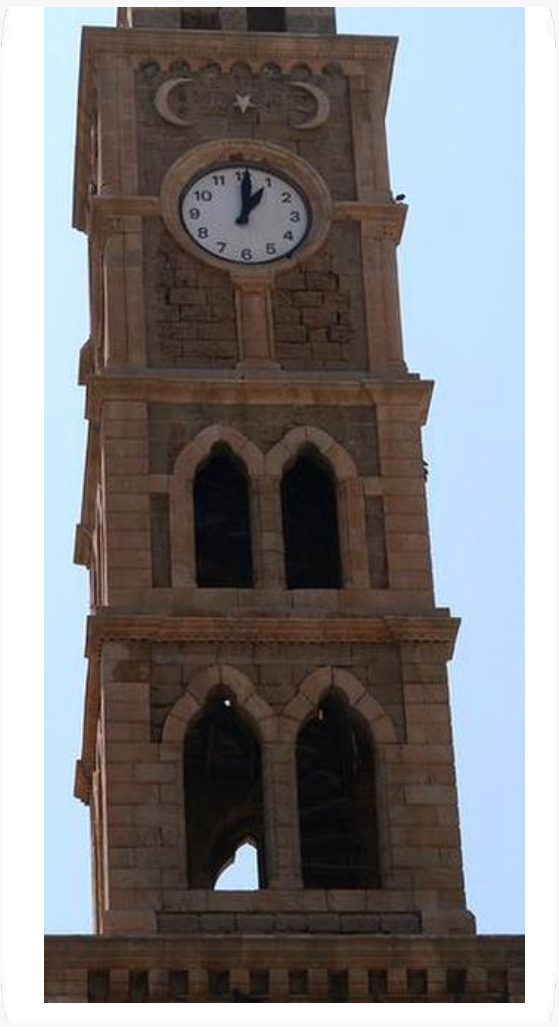
| 모델 | BLIP v1 | EfficientNet + BERT | ViT + BERT |
|----|---------|---------------------|------------|
| 답변 | V | none | T |

실험결과

질문 유형 별 결과 비교



> Knowledge Base Reasoning



What time is it?

| 모델 | BLIP v1 | EfficientNet + BERT | ViT + BERT |
|----|---------|---------------------|------------|
| 답변 | 10:00 | 11:25 | 4:15 |

결론

Vision Model과 Language Model의 조합

01

- GPT보다 BERT를 사용했을때 성능이 좋게 나왔다.
 - bert-base-uncased vs GPT2
 - Loss를 비교하면 같은 조건에서 BERT가 GPT보다 많이 낮음 (0.1 이상 차이)
 - 리소스의 제약으로 학습량이 적었는데 약 10배 이상 큰 모델인 GPT가 충분히 최적화할 시간이 적었다.
- ResNet 보다 EfficientNet, ViT가 성능이 더 우수하게 나왔다
 - 이미지 모델의 크기와 성능이 비례함.

BLIP

02

Vision Model과 Language Model을 2-track으로 적용할 때보다 BLIP모델을 사용했을때 성능이 월등하게 뛰어났다.

- Zero-Shot 성능 우수.
- 학습해도 큰 차이 없음.
 - Task가 복잡하지 않기 때문에 그 이상의 성능 향상 미미.

| Method | Pre-train # Images | Flickr30K (1K test set) | | | | | |
|---------------------------|-----------------------|-------------------------|-------|-------|------|------|------|
| | | TR | | | IR | | |
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| CLIP | 400M | 88.0 | 98.7 | 99.4 | 68.7 | 90.6 | 95.2 |
| ALIGN | 1.8B | 88.6 | 98.7 | 99.7 | 75.7 | 93.8 | 96.8 |
| ALBEF | 14M | 94.1 | 99.5 | 99.7 | 82.8 | 96.3 | 98.1 |
| BLIP | 14M | 94.8 | 99.7 | 100.0 | 84.9 | 96.7 | 98.3 |
| BLIP | 129M | 96.0 | 99.9 | 100.0 | 85.0 | 96.8 | 98.6 |
| BLIP _{CapFilt-L} | 129M | 96.0 | 99.9 | 100.0 | 85.5 | 96.8 | 98.7 |
| BLIP _{ViT-L} | 129M | 96.7 | 100.0 | 100.0 | 86.7 | 97.3 | 98.7 |

2023 분석 미니 프로젝트1 최종발표

감사합니다.

이미지 기반 질의응답

21기 족제비즈

