CLIP 모델과 Visual7W 데이터셋을 활용한 Visual Question Answering

Visual Question Answering with CLIP Model and Visual7W Dataset

김태훈

개요

- CLIP과 Visual7W 데이터셋을 활용하여 멀티모달 Visual Question Answering 모델 구축
- LoRA를 사용하여 효율적으로 훈련
- Public score 0.7532, Private score 0.74519

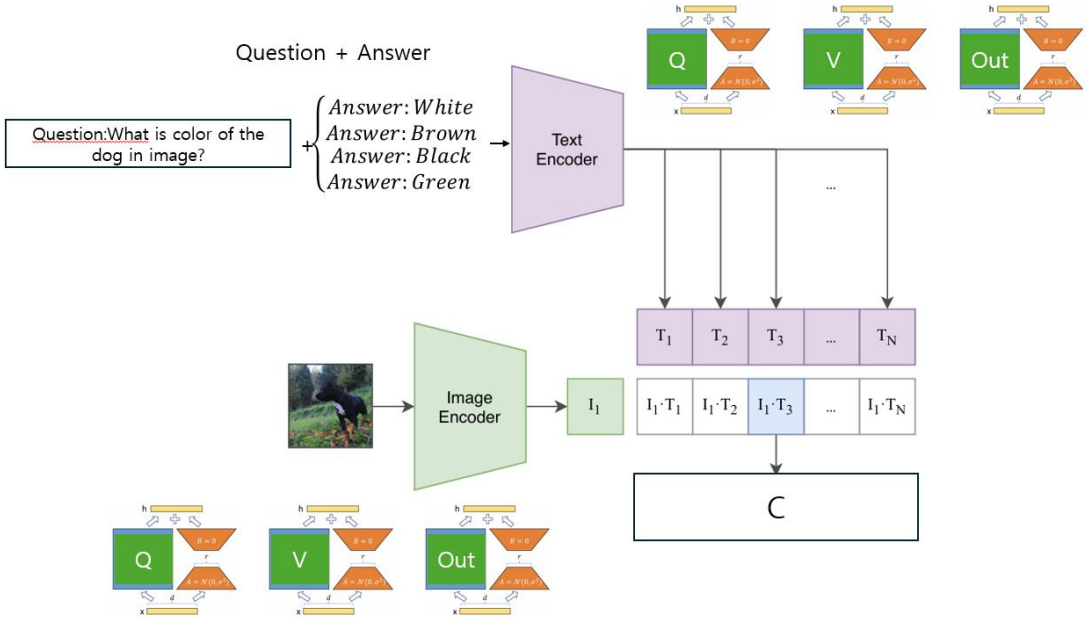


그림1. 모델의 전체적인 구조

Contrastive Language-Image Pre-training(CLIP)

- 이미지와 텍스트를 동시에 학습하여 둘 간의 의미 연결
  - ViT 등의 Image Encoder를 통해 이미지를 인코딩
  - Transformer 기반의 Text Encoder를 통해 텍스트를 인코딩
  - 이미지와 정답 텍스트 간의 코사인 유사도는 가깝게 학습
  - 이미지와 오답 텍스트 간의 코사인 유사도는 멀게 학습

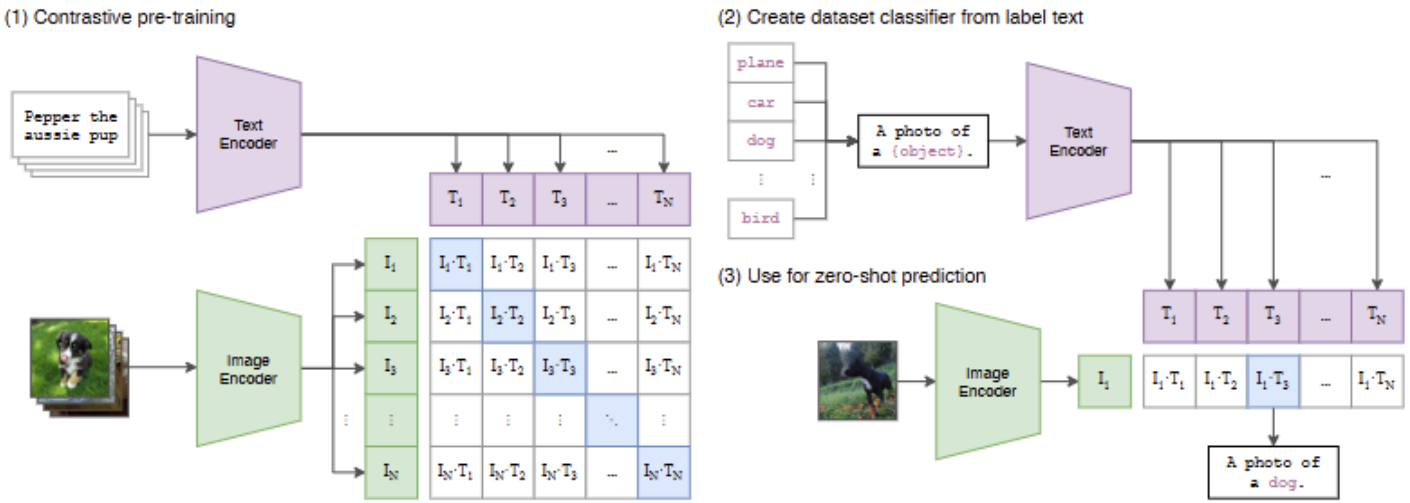


그림2. CLIP 모델의 전체적인 구조

Visual7W dataset

- COCO 이미지 기반의 4지선다형 VQA 데이터셋
- 47300개의 이미지, 139868 개의 질문 – 4지선다 pairs
- VisualGenome 산하 프로젝트
  - CC BY 4.0 으로 배포
- https://ai.stanford.edu/~yukez/visual7w/

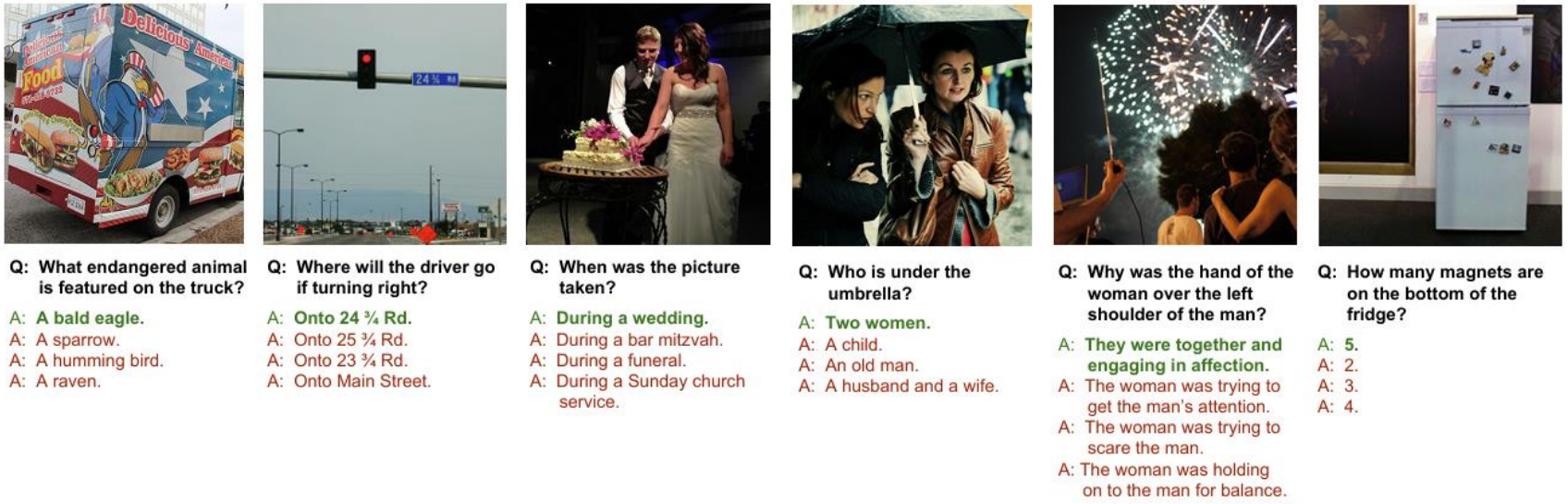


그림3. Visual7W 데이터셋 샘플

LoRA: Low-Rank Adaptation

- CLIP 전체를 finetuning 하기에는 데이터셋이 부족하다고 판단
- 효율적으로 finetuning 하기 위해 LoRA 사용
  - 사전학습된 가중치 W 는 고정
  - 저랭크 행렬 A, B만 학습
  - $W \leftarrow W_0 + \frac{\alpha}{r} BA \ (Rank\ r \ll d)$
    - r, α는 하이퍼파라미터

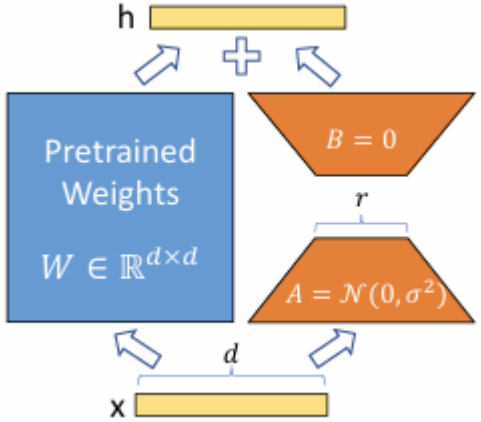


그림4. LoRA 구조

모델 선정

- Laion/CLIP-ViT-L-14-laion2B-s32B-b82k 사용
  - 428M 개의 파라미터
  - 307M 개의 파라미터를 가진 Vision Transformer Large(ViT-L) 기반
  - MIT 라이선스
  - https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K

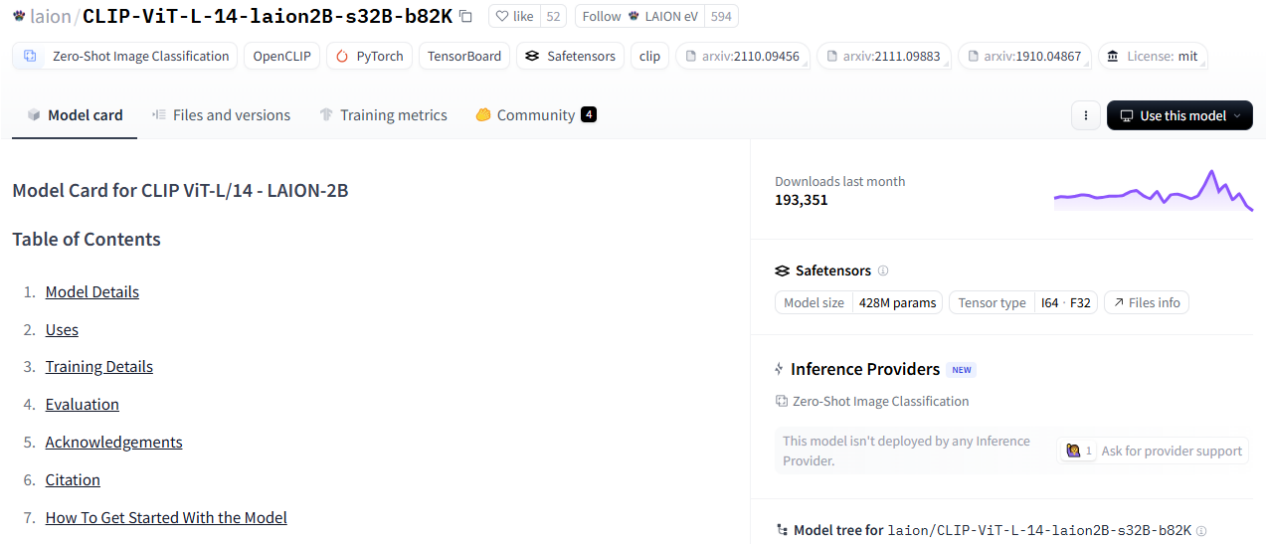


그림5. laion/CLIP-ViT-L-14-laion2B-s23B-b82k huggingface 페이지

데이터 전처리

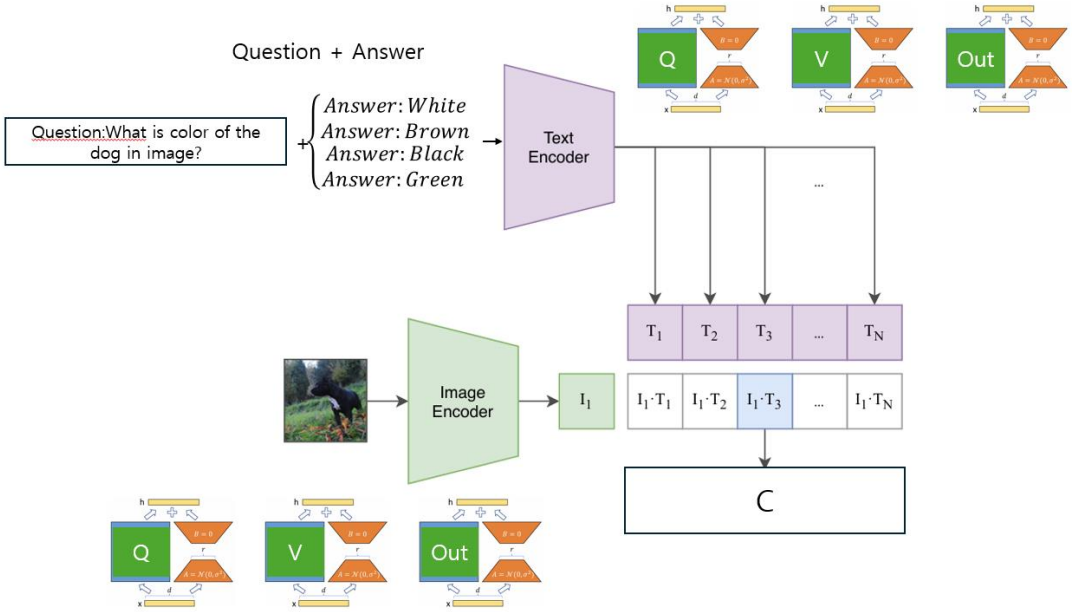
- 질문과 각 선택지를 결합한 텍스트 생성
  - Question: {question} Answer: {choice} 형식의 텍스트 쌍 생성
    - Question: What is the color of the dog? Answer: Purple
    - Question: What is the color of the dog? Answer: White
    - Question: What is the color of the dog? Answer: Black
    - Question: What is the color of the dog? Answer: Brown
  - 정답 레이블 생성하여 정답 텍스트와 이미지간 코사인 유사도가 높아지도록 유도

- Processor를 사용한 데이터 전처리
  - 이미지 리사이즈, 정규화
  - 텍스트 토큰화

- 패딩
  - 배치 데이터 간의 모든 텍스트 입력의 길이를 맞춤

모델 학습

- 모델 학습
  - LoRA를 사용해 Image encoder와 Text encoder의 W_Q, W_V, W_{out}만 학습
  - CrossEntroyLoss 사용하여 정답 텍스트와 코사인 유사도가 가장 높도록 유도
  - Autograd와 GradScaler를 사용하여 Float16 기반의 학습
    - 학습 효율성 증가(학습에 필요한 메모리 및 시간 감소)



모델 학습

- 주요 하이퍼파라미터

표1. 주요 하이퍼 파라미터

하이퍼파라미터	값
Seed	41
Batch size	64
Learning rate	1e-4
Optimizer	AdamW
Weight decay	1.5e-2
Scheduler	ReduceLROnPlateau (mode=min,factor-0.5, patience=2)
Validation Split	0.15
Num_epochs	6
CrossEntropyLoss LabelSmoothing	0.1
LoRA Rank	32
LoRA Alpha	64
Target Module	Q_proj, v_proj, out_proj

결과

- 모델 크기
  - 총 434,104,577(약 434.1M)개의 가중치
  - LoRA 학습을 통해 총 6,488,064(약 6.5M)개의 가중치만 업데이트

- 학습 결과
  - 2 에포크에서 가장 좋은 검증 정확도 달성

표2. 학습 결과

데이터	정확도
Train set	82.73%
Validation set	78.37%
Public score(SCPC)	75.32%
Private score(SCPC)	74.519%

감사합니다

References

- A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision. 2021. Available: https://arxiv.org/abs/2103.00020
- Y. Zhu et al., Visual7W: Grounded Question Answering in Images. 2016. Available: https://arxiv.org/abs/1511.03416
- E. Hu et al., LoRA: Low-Rank Adaptation of Large Language Models. 2021. Available: https://arxiv.org/abs/2106.09685

Appendix

- License of laion/CLIP-ViT-L-14-laion2B-s32B-b82K
  - MIT License

그림6. MIT License

Appendix

- License of Visual Genome
  - Creative Commons Attribution 4.0 International License
  - http://creativecommons.org/licenses/by/4.0/
    - https://homes.cs.washington.edu/~ranjay/visualgenome/about.html

- License of COCO image
  - Creative Commons Attribution 4.0 License
  - https://creativecommons.org/licenses/by/4.0/legalcode
    - https://cocodataset.org/#termsofuse

Appendix

- License of packages

Packages	License
Pandas	BSD-3-Clause
Pytorch	License
Pillow	MIT-CMU
Transformers	Apache-2.0
Numpy	License
Tdqm	License
Scikit-learn	BSD-3-Clause
Peft	Apache-2.0