한양대학교

SW프로젝트 요약서

프로젝트 기간	2022.02.01. - 2022.11.20. (총10개월)
프로젝트 팀원	정상윤(컴퓨터소프트웨어학부, 4학년) 최가온(컴퓨터소프트웨어학부, 4학년)
지도교수	김은솔 교수님
프로젝트 멘토	(없음)
프로젝트 명	어텐션 기반 광학 문자 인식(OCR) 기술 적용을 통한 기존 BAN-VQA의 성능 개선

한양대학교

프로젝트 내용	본 프로젝트의 이름은 "어텐션 기반 광학 문자 인식(OCR) 기술 적용을 통한 기존 BAN-VQA 의 성능 개선"이다. 시각 질의 응답(VQA; Visual Question Answering) 문제는 특정 이미지가 주어지고 해당 이미지에 대한 질문이 주어졌을 때, 이미지와 질문 텍스트를 이해하여 이를 기반으로 적절한 답을 도출해내는 기술이다. 기존 연구에서 주로 사용한 상호 어텐션(co-attention)은 주어진 이미지와 텍스트 사이의 상호작용은 고려하지 않았던 반면, BAN-VQA 는 저수준 쌍선형 풀링(low-rank bilinear pooling)이 이미지와 텍스트 사이의 접합 어텐션(joint attention)을 계산함으로써 두 입력 채널 간의 상호작용을 고려하는 모델이다. 그러나, BAN-VQA 모델은 특정 질의 유형(question type)에 대해 정확도가 대폭 낮은 것을 확인할 수 있었는데, 대부분 이미지 내의 텍스트와 관련한 질의에서 성능 저하가 나타났다. 이번 프로젝트의 목적은 주어진 이미지에 광학 문자 인식 기술을 적용하고, 주어진 질문의 질의 유형을 모델 내에서 예측하는 방식으로 기존 BAN-VQA 의 성능을 개선하는 것이다. 어텐션 모델(attention model)을 기반으로 하여 어텐션 값이 높게 나타나는 이미지의 특정 부분에만 OCR 를 사용하고 단어 구(Phrase)로 묶어, 단어 정답률을 향상시키고, question type 에 다른 OCR 적용 한계치의 최적값을 찾았다. 이 방식을 통해 'what number is' 질의 유형의 경우 약 13%p 가량 성능이 향상된 결과를 보였다.
기대효과 및 개선방향	본 프로젝트는 기존 BAN-VQA 모델 상에서 확신도(confidence) 값이 낮게 나오는 질의에 대해, 광학 문자 인식(OCR) 기술을 적용하여 이미지 자체에서 인식될 수 있는 일차적인 정보를 이용하는 방식으로 기존 모델에서 성능이 정확도 약 20%로 나왔던 질의 유형의 정확도를 개선했다는 점에서 의의가 있다. 인식된 단어들을 구(phrase) 단위로 묶는 알고리즘과 어텐션 기반의 재구성을 통해 이미지 내의 텍스트 정보를 최대한 활용하였다. 본 프로젝트의 방향성과 모델 성능 개선 결과는 향후 시각 질의 응답 관련 문제에서 주어진 이미지에 포함된 텍스트 정보를 활용하여 해결할 수 있는 일차적인 질의 유형에 대한 성능을 높였다는 점에 기여하였다고 판단된다.

한양대학교

SW프로젝트 결과보고서

프로젝트명	어텐션 기반 광학 문자 인식(OCR) 기술 적용을 통한 기존 BAN-VQA의 성능 개선
프로젝트 요약	시각 질의 응답(VQA; Visual Question Answering) 문제는 특정 이미지가 주어지고 해당 이미지에 대한 질문이 주어졌을 때, 이미지와 질문 텍스트를 이해하여 이를 기반으로 적절한 답을 도출해내는 기술이다. 기존 연구에서 주로 사용한 상호 어텐션(co-attention)은 주어진 이미지와 텍스트 사이의 상호작용은 고려하지 않았던 반면, BAN-VQA 는 저수준 쌍선형 풀링(low-rank bilinear pooling)이 이미지와 텍스트 사이의 접합 어텐션(joint attention)을 계산함으로써 두 입력 채널 간의 상호작용을 고려하는 모델이다. 그러나, BAN-VQA 모델은 특정 질의 유형(question type)에 대해 정확도가 대폭 낮은 것을 확인할 수 있었는데, 대부분 이미지 내의 텍스트와 관련한 질의에서 성능 저하가 나타났다. 이번 프로젝트에서는 주어진 이미지에 광학 문자 인식 기술을 적용하고, 주어진 질문의 질의 유형을 모델 내에서 예측하는 방식으로 기존 BAN-VQA 의 성능을 개선하였다. 마지막으로 주어진 질문에 대해 이미지의 내의 어텐션 값이 높게 나오는 단어를 선택함으로써, 기존 광학 문자 인식만 사용하는 방식에 비해 개선된 성능을 보였다.
프로젝트 기간	2022.02.01. - 2022.11.20. (총10개월)

한양대학교

산출물	졸업 작품 (○), 졸업 논문 ( )
성과물	특허 ( ), 논문( ), SW(○)

학과	학번	학년	이름	연락처
컴퓨터소프트웨어	2018009061	4	정상윤	zetro@hanyang.ac.kr 010-9554-6018
컴퓨터소프트웨어	2019009261	4	최가온	choigaon1028 @hanyang.ac.kr 010-3696-9592

한양대학교

목 차

1. 프로젝트 개요

1.1 프로젝트 목적 및 배경

시각 질의 응답(VQA; Visual Question Answering) 문제는 특정 이미지가 주어지고 해당 이미지에 대한 질문이 주어졌을 때, 이미지와 질문 텍스트를 이해하여 이를 기반으로 적절한 답을 도출해내는 기술이다. 이러한 기술은 자동 응답 시스템, 시각 정보 제공 시스템 등으로 활용될 수 있다. 기존 BAN-VQA 모델은 이미지 내의 텍스트와 관련한 질의에서 정확도가 낮은 것을 확인할 수 있었는데, 이번 프로젝트의 목적은 각 질의 유형별로 기존 모델의 성능을 파악하고 특정 질의 유형에서의 성능을 개선하는 데에 있다.

1.2 프로젝트 최종 목표

본 프로젝트에서는 광학 문자 인식(OCR) 기술을 통해 위에서 정의된 시각 질의 응답 문제에 대한 기존 BAN-VQA의 성능을 개선하고자 한다. 해당 최종 목표를 달성하기 위해 아래의 세부 목표를 설정하였다.

- COCO 데이터셋에 포함된 모든 이미지에 대해 광학 문자 인식 기술 적용
- 인식된 단어를 공간적 문맥에서 병합하여 구(phrase) 단위로 생성하는 기술 적용
- 인식된 단어에 대해 어텐션(attention) 값을 이용한 OCR 적용
- 주어진 질의에서 질문 텍스트를 통해 질의 유형을 예측하는 모델 개발
- 기존 BAN-VQA 모델에서 각 질의별 확신도 값 측면에서 광학 인식 기술을 적용하기 위한 최적의 역치 계산

2. 프로젝트 내용

2-1. COCO 데이터셋에 대한 기존 BAN-VQA의 성능 분석

COCO 데이터셋 validation 2014를 기준으로 하여 기존 BAN-VQA 모델의 성능을 분석하였다. 성능 지표는 스코어(score)를 사용하였으며, 스코어의 계산 과정은 아래와 같다.

2-1-1. 스코어 계산 과정

본 프로젝트에서는 모델의 성능을 측정함에 있어 스코어(score)를 단일 성능 지표로 사용하였다. 특정 질의에 대해 복수의 정답(open-ended)이 존재할 수 있으며, 각 답에 대해 매핑되는 스코어는 다르게 나타난다. 아래의 예시에서 "down", "at table"이라는 단어는 모두 정답으로 인정되나, 각각 1.0, 0.3이라는 다른 스코어가 매핑되었음을 알 수 있다.

```
"262148000": {
    "image_id": 262148,
    "answers": [ "down", "at table", "skateboard", "table" ],
    "scores": [ 1.0, 0.3, 0.3, 0.3 ]
}
```

표 1. 복수 정답과 각 스코어 매핑

Validation Set 2014에는 총 214,354개의 질의 데이터가 포함되어 있으며, 각 질의에는 그것이 속한 질의 유형 정보가 함께 저장되어 있다. 기존 BAN-VQA 모델에서는 이 질의 유형을 정답을 도출하는 과정에서 사용하지 않았으며, 질의 유형에 대한 정보는 해당 데이터셋의 정답지(annotation)에만 포함되어 있다. 기존 BAN-VQA에 대해 각 질의 유형별 평균 스코어를 산출하였으며, 결과는 아래와 같다.

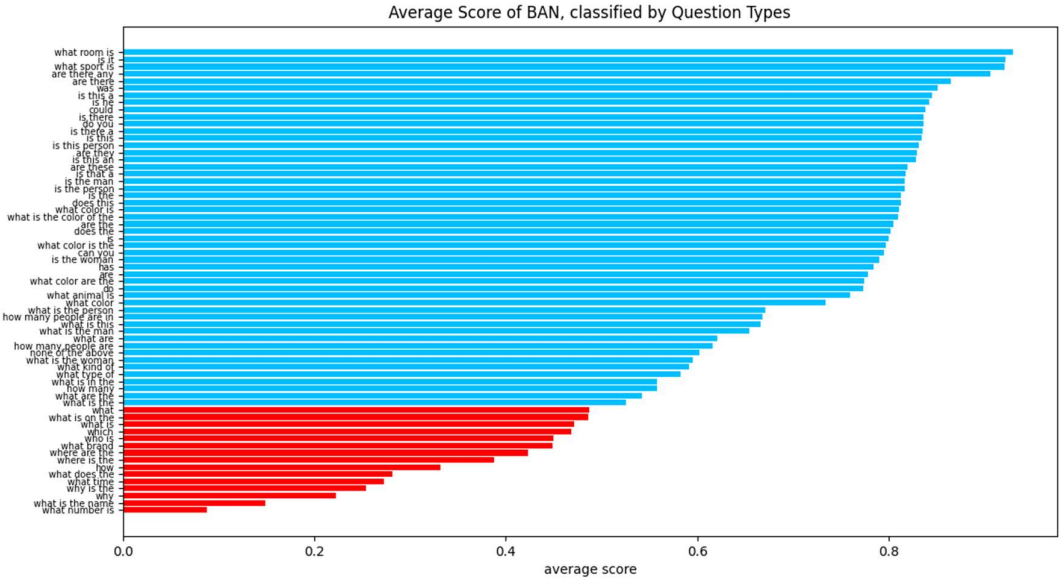

그림 1. BAN-VQA의 각 질의 유형별 평균 스코어

질의 유형 측면에서 관찰한 결과, “what number is”, “what is the name”, “what does the”, “what brand” 등의 질의 유형에서 기존 BAN-VQA의 성능이 낮게 나타나는 것을 확인하였다. 이러한 질의 유형은 대부분 주어진 질의 이미지에 포함된 텍스트 정보를 통해 해결할 수 있는 질의 유형에 해당한다. 해당 질의 유형에서 기존 모델의 성능이 낮게 나오는 원인을 분석하였다. BAN-VQA 모델에서는 결과로 출력하는 데에 있어서 단어 집합(Answer Candidate List)을 정의하고 해당 집합 내의 단어를 결과로 출력한다. 이 단어 집합에 주어진 질의에 대응되는 정답 단어가 존재하지 않아 스코어가 낮게 계산되는 것이 원인인 것으로 파악되었다. 우리는 후자 측면에서 기존 BAN-VQA 모델이 정답을 출력하는 데에 사용하는 단어를 전체적으로 관찰하였다.

본 프로젝트에서는 Keras-OCR를 이용하여 COCO 데이터셋을 구성하는 모든 이미지에 대해 텍스트를 추출하였다. 이를 통해 주어진 질의 이미지에 포함된 텍스트를 구성하는 단어가 정답이 될 수 있음에도 불구하고, 모델에서 사용하는 단어 집합에 포함되지 않아 성능이 낮아지는 경우를 보완하고자 하였다.

2-2. BANS 모델 및 확신도를 기준으로 한 최적의 역치 찾기

기존 BAN 모델 내의 feature을 이용하여, BAN의 답과 실제 답을 매칭시켜 이진 교차 엔트로피(BCE; Binary Cross Entropy)를 통해 확신도(confidence)를 측정할 수 있는 모델(BANS)을 구성하였다. 확신도를 계산한 방법은 아래와 아래의 표는 기존 BAN-VQA 모델과 우리 팀에서 새롭게 제시한 BANS 모델의 훈련 세트와 테스트 세트에 대한 성능 결과이다.

모델명	훈련 세트 스코어	테스트 세트 스코어
BAN-12epoch	80.92	66.43
BANS-12epoch	80.54	66.38

표 2. BANS의 훈련 세트와 테스트 세트에 대한 평균 스코어

BANS의 확신도(confidence)를 OCR를 통해 인식한 단어의 사용 여부에 대한 역치로 사용한다. BANS의 정답 도출 프로세스는 아래와 같다.

2-1-1. BANS의 정답 도출 프로세스

1) 주어진 질의에 대한 모델의 확신도(confidence)를 확인한다.
2) 만약 해당 질의의 확신도가 역치보다 높다면, 기존 모델의 출력을 사용한다.
3) 만약 해당 질의의 확신도가 역치보다 낮다면, 질의 이미지에 대한 OCR 단어의 유무를 확인한다.
    3-1) 3에서 OCR 단어가 존재한다면, 해당 단어들 중 하나를 무작위로 선택하여 결과로 출력한다.
    3-2) 3에서 OCR 단어가 존재하지 않는다면, 기존 모델의 출력을 사용한다.

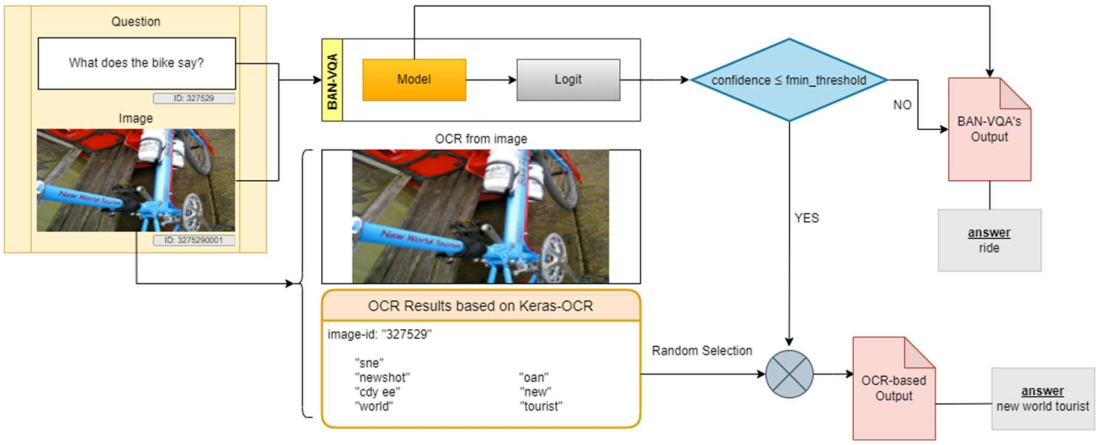

그림 2. BANS + OCR 모델의 스코어 계산 프로세스

위의 정답 도출 프로세스에서 사용하는 역치 중 훈련 세트 데이터에서 가장 스코어가 높게 산출되는 최적의 역치를 계산하였다.

BANS + OCR	
확신도 최적 역치	최적 역치에서의 평균 스코어
-2.720	83.419

표 3. BANS + OCR의 훈련 세트와 테스트 세트에 대한 평균 스코어

구분	BAN-VQA	BANS + OCR
전체 질의에 대한 평균 스코어	66.433	66.474
OCR 가능한 질의에 대한 평균 스코어	4.950	8.412

표 4. 전체 질의 / OCR 적용 가능한 질의에 대한 평균 스코어 비교

3. 프로젝트의 기술적 내용

3-1. 광학 문자 인식 기술(OCR) 단어 구 생성

사람이 그림에 나타나 있는 문자를 적을 때, 일반적으로 단어 구(句, phrase) 형태로 묶어 적는다. 하지만, 광학 문자 인식 기술(OCR)은 그림에서 문자를 뽑아내는데, 띄어쓰기 단위를 구분하지 않고 한 단어씩 뽑아낸다. 정답 확률을 높이기 위하여, 광학 문자 기술로 뽑아낸 단어들을 단어 구로 생성하는 기법을 고안했다. 우리는 DBSCAN 알고리즘에서 착안하여, 이미지 공간 상에서 이웃하는 다른 단어들을 묶어, 하나의 단어 구로 표현하는 알고리즘을 개발하였다. 단어의 폰트 사이즈와 유사한 크기의 단어가 주변에 있을 경우, 같은 단어 묶음으로 간주하여, 단어 묶음을 군집화함으로써, 글자의 위치 순서에 맞게 배열하여 단어 구를 생성한다.

아래의 그림에서 안장에 적힌 단어를 OCR 기술을 이용하여 "new", "world", "tourist"로 출력한다. OCR-Phrase 기법을 적용하여, 해당 단어를 "new world tourist"로 하나의 단어 구로 합쳐준다.

그림 3. 주어진 질의 이미지 상에 나타낸 어텐션 박스

인식된 단어를 구 형태로 병합하는 OCR-Phrase 알고리즘을 통해, 기본 BAN 모델을 이용한 기존 OCR 보다 확신도 [-5.0, -2.0] 구간에서 최대 약 40p%만큼 성능이 향상되었음을 확인할 수 있었다.

광학 문자 인식 기술을 적용한 질의에 대해 평균적인 스코어를 유한한 범위 내에서 역치를 바꿔가며 계산하였고, 그 결과를 시각화한 것은 아래와 같다.

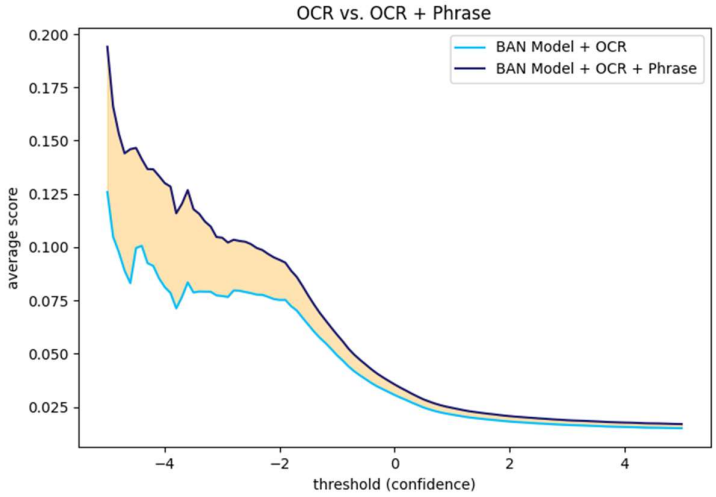

그림 4. 기존 모델과 비교한 OCR-Phrase 기법의 성능

OCR-Phrase 기법을 기반으로 기존 프로세스를 실행할 때 사용할 역치 중 훈련 세트 데이터에서 가장 스코어가 높게 산출되는 최적의 역치를 계산하였다.

BANS + OCR-Phrase	
확신도 최적 역치	최적 역치에서의 평균 스코어
-2.597	83.446

표 5. BANS + OCR-Phrase의 훈련 세트와 테스트 세트에 대한 평균 스코어

구분	BAN-VQA	BANS + OCR-Phrase
전체 질의에 대한 평균 스코어	66.433	66.516
OCR 가능한 질의에 대한 평균 스코어	4.950	10.624

표 6. 전체 질의 / OCR 적용 가능한 질의에 대한 평균 스코어 비교

3-2. 어텐션을 이용한 단어 가중치 부여

BAN 모델에서는 그림에서 특징적인 부분에 어텐션을 부여하고 이들은 각각 다른 값을 가진다. 따라서, 우리는 OCR의 정보를 어텐션 값이 높은 영역을 각 질의별로 탐색하며, 해당 영역의 단어에 상대적으로 가중치를 부여하는 방식으로 기존 모델을 개선하였다. 아래의 그림은 질의의 Question ID를 입력하면 해당 질의 문장, 질의 이미지, 실 정답과 모델이 출력한 정답을 시각화하는 그래픽 유저 인터페이스이다. 질의 이미지 내의 각 박스는 어텐션을 의미하며, 이미지 내에서 어텐션 값이 가장 높은 부분을 빨간색으로 표현하였으며, 파란색 방향으로 갈수록 어텐션 값이 낮아져 답을 도출하는 데에 영향이 상대적으로 적은 부분이라고 할 수 있다.

아래 그림들은 질의의 정답이 이미지 속 텍스트에 있는 경우이다. BANS을 통해 얻은 아래 결과에서 정답에 가까운 텍스트 위치인 그림 5의 기차 앞부분과 그림 6의 표지판의 영역의 어텐션 값이 높게 나왔다. 따라서, 어텐션이 높은 부분의 OCR 단어는 주어진 질문 텍스트와의 연관성이 높은 단어로 기존에 비해 주어진 질의의 정답에 가까운 단어를 모델이 출력할 것이므로 이를 통해 성능이 개선될 것으로 파악하였다.

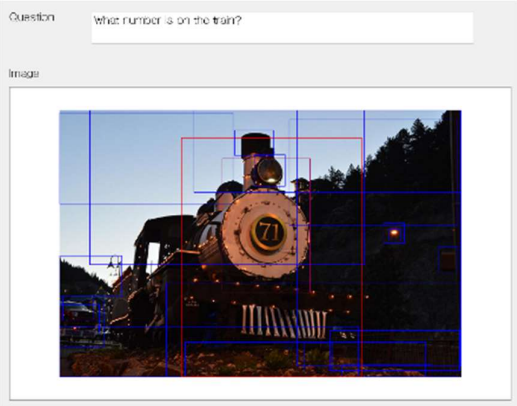

그림 5. 어텐션 시각화-1

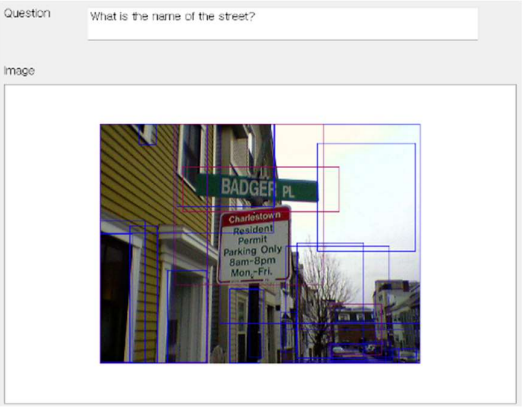

그림 6. 어텐션 시각화-2

OCR-Attention 기법을 기반으로 기존 프로세스를 실행할 때 사용할 역치 중 훈련 세트 데이터의 모든 질의에 대해 가장 스코어가 높게 산출되는 최적의 역치를 계산하였다.

BANS + OCR-Attention	
확신도 최적 역치	최적 역치에서의 평균 스코어
-2.249	83.498

표 7. BANS + OCR-Phrase의 훈련 세트와 테스트 세트에 대한 평균 스코어

구분	BAN-VQA	BANS + OCR-Phrase
전체 질의에 대한 평균 스코어	66.433	66.568
OCR 가능한 질의에 대한 평균 스코어	7.284	16.094

표 8. 전체 질의 / OCR 적용 가능한 질의에 대한 평균 스코어 비교

광학 문자 인식 기술(OCR, OCR-Phrase, OCR-Attention)을 적용한 질의에 대해 각 기법별로 평균적인 스코어를 유한한 범위 내에서 역치를 바꿔가며 계산하였고, 그 결과를 시각화한 것은 아래와 같다.

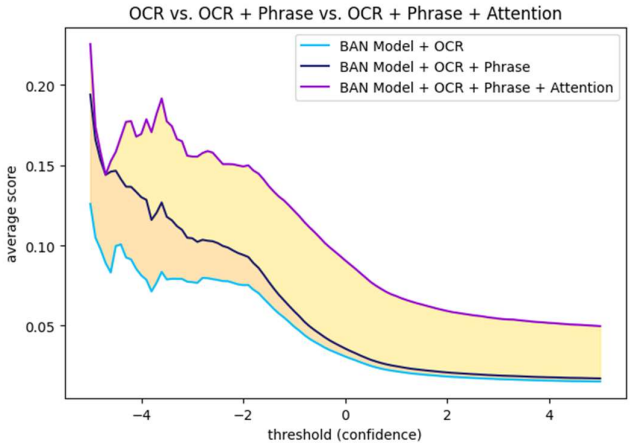

그림 7. 기존 모델과 비교한 OCR-Attention 기법의 성능

3-3. 질의 유형별 최적의 역치 계산

위의 과정에서는 모든 질의에 대해 평균 스코어가 가장 높게 나오는 역치를 최적 역치의 기준으로 하여 계산하였다. 그러나, 확신도 기준의 역치를 유한한 범위 내에서 바꿔가며 각 질의 유형별로 스코어를 계산할 때에는 각 질의 유형별로 분류하였을 때에는 최적 역치가 다르게 나타났다.

따라서, 질의가 주어지면 해당 질의가 소속된 질의 유형을 판단하고 해당 질의 유형에 따라 다른 최적 역치를 적용하는 정책을 고안했다. 아래는 기존 BAN-VQA에서 가장 낮은 성능을 보인 하위 10개의 질의 유형에 대한 각각의 최적 역치와 해당 역치에서 계산된 평균 스코어이다.

질의 유형	최적 역치에서의 평균 스코어	기존 BAN의 평균 스코어
what number is	17.426	8.751
what is the name	20.675	14.820

what time	27.365	27.285
why	20.730	22.197
why is the	25.117	25.428

표 9. 기존 모델 기준 하위 5개 질의 유형의 성능 변화 (1)

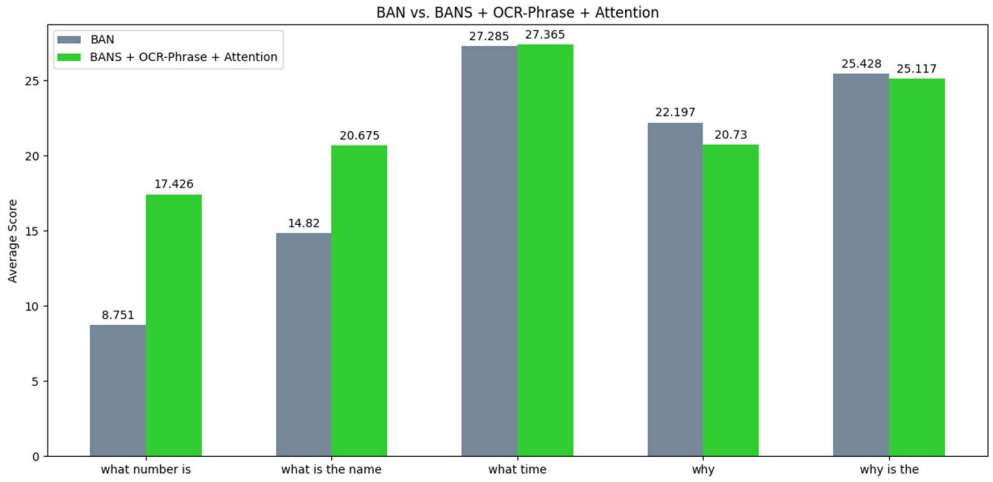

그림 8. 기존 모델 기준 하위 5개 질의 유형의 성능 변화 (2)

3-4 질의 유형 예측 모델 개발 및 적용

기존 모델에서는 주어진 질의에 대해 해당 질의가 소속된 질의 유형 정보는 활용하지 않았다. 본 프로젝트에서 계산한 질의 유형별 최적 역치는 각 질의 유형마다 상이하게 나타났다. 따라서, 질의가 주어지면 입력 받은 텍스트 정보로부터 해당 질의의 질의 유형을 예측할 수 있는 모델을 개발하였다. 질의 유형 예측 모델에서 질의 유형과 해당 질의 유형에 해당하는 최적 역치를 계산하고, 질의 유형마다 다른 역치를 이용함으로써 각 질의 유형에 알맞은 역치를 사용하는 정답 도출 프로세스를 디자인하였다.

질의 유형 예측 모델의 구조는 아래와 같으며, 테스트 데이터셋 기준으로 약 92%의 정확도를 보였다.

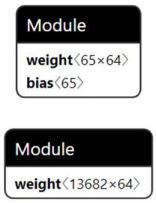

그림 9. 질의 유형 예측 모델의 구조

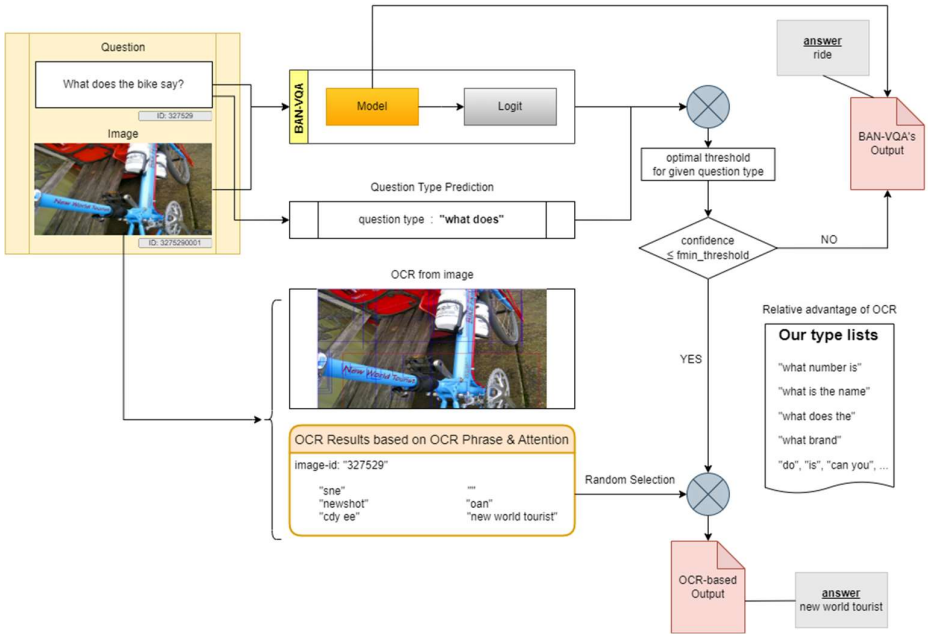

그림 10. BANS + OCR Phrase + Attention과 질의 유형 예측 모델의 정답 도출 프로세스

3-5 데이터셋 전체 측면에서의 광학 문자 인식 기술 적용의 영향력

COCO Validation Set 데이터셋을 구성하는 총 214,354 개의 질의 중 최적 역치 조건을 만족하는 질의를 전체 분모로 하여, 그 중 광학 문자 인식 기술이 적용 가능한 (즉, 해당 질의 이미지 내에 Keras-OCR 를 이용하여 텍스트가 1 개 이상 검출된 경우) 질의의 비율을 계산하였다. 아래의 그림은 각각 BANS 에 OCR, OCR + Phrase, OCR + Phrase + Attention 을 적용한 경우에서의 OCR 적용 가능한 질의의 비율을 나타낸 것이다. OCR 을 적용한 경우와 OCR + Phrase 를 적용한 경우에는 전체 질의 중 약 2.71%이 최적 역치 조건을 만족하였으며, 전체 질의 중 약 2.46%의 질의에서 해당 OCR 기법을 적용 가능한 것으로 나타났다. 반면, 어텐션까지 적용한 BANS 의 경우 해당 최적 역치 조건을 만족하는 질의의 비율은 4.32%로 소폭 증가하였으나, 이 중 OCR 기법을 적용할 수 있는 질의의 비율은 64.77%로 이전에 비해 대폭 낮아졌음을 확인할 수 있었다. 이는 전체 질의 중 2.80%에 해당한다.

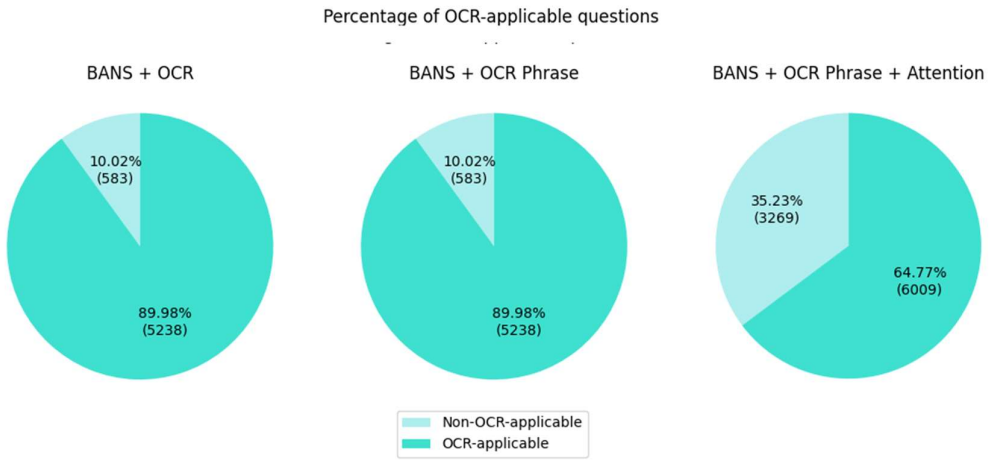

그림 11. 전체 질의 중 OCR 기법 적용이 가능한 질의의 비율

위의 과정은 전체 질의 중 OCR 기법이 적용될 수 있는 질의의 비율을 계산한 것이다. 여기에 OCR 기법을 실제로 적용하여 어느 정도의 성능 개선을 이뤄졌는지를 평균 스코어를 계산함으로써 판단하였다. 구체적인 실험 방법은 아래와 같으며, 하단에 최종 성능 개선을 나타낸 그림을 첨부하였다.

- 대상이 되는 질의는 아래의 두 조건을 만족해야 한다.

  - 최적 역치 조건을 만족해야 한다.

  - Keras-OCR 를 적용하였을 때에 이미지 내에서 검출된 텍스트가 적어도 1 개 이상 존재해야 한다.

- OCR 기법이 적용 가능한 질의에 대해 BANS 와 아래의 3 가지 OCR 관련 기법을 각각 적용한 경우의 평균 스코어를 계산한다.

  - Keras-OCR 만을 적용한 경우 (OCR)

  - OCR Phrase 기법을 적용한 경우 (OCR Phrase)

  - OCR Phrase 와 어텐션을 모두 사용한 경우 (OCR Phrase + Attention)

- OCR 기법이 적용 가능한 질의에 대해 기존의 BAN-VQA 모델이 출력한 정답을 이용하여 평균 스코어를 계산한다.

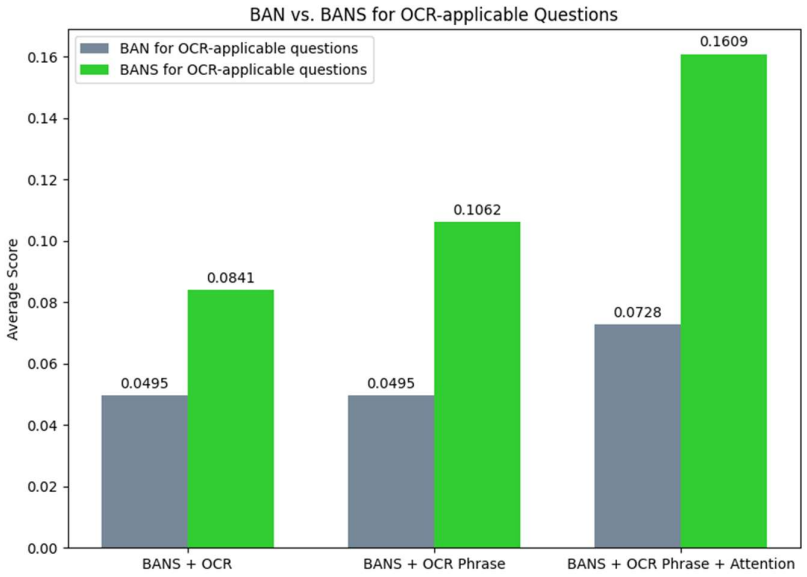

그림 12. OCR 기법 적용 가능한 질의에서의 BANS 성능

4. 프로젝트의 역할 분담

4.1 개별 임무 분담

번호	학과	학번	학년	이름	담당업무
1	컴퓨터소프트웨어	2018009061	4	정상윤	1) 광학 문자 인식 기술 적용 및 구(phase) 단위 병합 관련 알고리즘 개발 2) BAN-VQA 모델 학습 3) 어텐션 기법 디자인 및 적용
2	컴퓨터소프트웨어	2019009261	4	최가온	1) 기존 BAN-VQA 모델 성능 통계 분석 2) 질의 유형 예측 모델 개발 및 적용

4.2 개발 일정

순서	업무별 수행 기간		세부 사항
	시작일	종료일	
1	2022.02.01	2022.03.08	"VQA_Visual Question Answering" 논문 리뷰
2	2022.03.09	2022.03.31	"Bilinear Attention Network" 논문 리뷰
3	2022.03.09	2022.04.30	BAN-VQA 코드 분석 https://github.com/jnhwkim/ban-vqa
4	2022.03.10	2022.03.15	각 질문에 대해 매핑된 서로 다른 이미지의 개수 계산 및 분포 파악

5	2022.04.01	2022.04.08	각 질문에 대한 정확도 분석 (accuracy = # of correct / # of total)
6	2022.04.09	2022.04.19	각 이미지에 대한 정확도 분석 (계산 방식은 위와 동일)
7	2022.04.20	2022.04.30	기존 BAN-VQA 모델의 각 질의 유형 별 정확도 계산 및 상대적 약점 파악
8	2022.05.01	2022.5.21	테스트 용 GUI 구현 - question_id를 입력하면 그에 매핑되는 이미지와 모델이 출력한 답변을 출력 - attention을 직사각형의 형태로 표시
9	2022.05.22	2022.06.10	기존 모델의 정확도가 낮은 질의에 대해, 단어 리스트(word-of-bag)에 답에 해당하는 단어가 존재하지 않아 틀린 답을 내놓은 경우에 대해 조사
10	2022.06.11	2022.06.17	COCO 데이터셋의 모든 이미지에 대한 광학 문자 인식(OCR) 처리 적용 및 JSON 파일 변환
11	2022.06.18	2022.07.03	(weight 2배)
12	2022.07.04	2022.08.07	모든 question에 대해 BAN-VQA 모델에 입력으로 넣어 entropy, logit을 구한 후 그 분포를 평면 상에 표시함으로써 분포 파악 및 entropy 기준으로 하여 fmin 계산
13	2022.08.08	2022.08.10	질의 유형 별로 fmin 개별 파악
14	2022.08.11	2022.09.14	공간적 문맥상 동일한 단어를 묶는 OCR Combination 기법 고안 및 구현 및 데이터 셋에 적용
15	2022.09.15	2022.09.23	각각의 질의에 대해 매핑된 이미지에서 OCR Combination 기반으로 인식된 단어들 중 실제 정답에 해당하는 단어가 포함된 빈도 측정
16	2022.09.24	2022.10.02	기존 BAN-VQA 모델, OCR 적용한 경우, OCR Combination 적용한 경우 각각에 대하여 모든 question에 대한 score 계산

17	2022.10.03	2022.10.14	기존 BAN-VQA 모델에 confidence head를 추가하여 OCR 적용 방식 개선 및 confidence 기준 fmin 계산
18	2022.10.15	2022.10.26	BAN-VQA 모델의 어텐션(attention) 값에 대한 Top-1을 이용하여 OCR 적용에서의 성능 개선
19	2022.10.27	2022.10.26	entropy, confidence / OCR, OCR Combination, OCR Attention 방법에 대한 종합적 성능 평가
20	2022.11.10	2022.11.29	최종 보고서 작성 및 소스 코드 정리

5. 결론 및 기대효과

본 프로젝트에서는 주어진 질의 이미지에 광학 문자 인식 기술을 적용함으로써 기존 BAN-VQA의 성능을 개선하고자 하였다. 모델이 주어진 질의에 대해 갖는 불확실성이 상대적으로 작을 때에는 기존 모델이 출력한 정답을 사용하나, 불확실성이 크다고 판단되는 경우 광학 문자 인식 기술을 통해 검출된 단어들을 정답 단어로 사용하고자 하였다. 불확실성의 기준으로 확신도(Confidence)를 사용하였으며, OCR 단어를 사용하기 위한 최적의 확신도 값의 경계에 해당하는 최적 역치(Optimal Threshold)을 계산하였다.

기존의 낮은 OCR 성능을 보완하기 위해 DBSCAN 알고리즘에 착안하여 이미지 공간 상에서 인접한 단어들을 묶어 구(phrase, 句)를 형성하는 OCR-Phrase 기법을 고안하였다. 이후, 모델의 어텐션 값을 최대한 활용하기 위해 주어진 질의에 대해 질의 이미지 상에서 어텐션 값이 높게 나타나는 영역의 텍스트를 모델이 출력하는 정답 후보로 설정함으로써 성능 개선을 이루었다.

마지막으로, 주어진 질의의 질의 유형 정보를 활용하고자 각 질의 유형마다 최적 역치를 계산하였으며, 입력 받은 질의 텍스트를 통해 해당 질의 유형을 예측하는 모델을 추가하였다. 결과적으로 “what number is”, “what is the name” 등과 같이 이미지 내의 텍스트를 일차적으로 활용하여 답을 낼 수 있는 질의 유형에 대해 성능 개선을 확인할 수 있었다. 그러나, OCR를 통해 검출한 텍스트를 무작위로 선택하여 모델이 출력하는 정답으로 결정한다는 점에서 일차원적 질의에서만 효율성을 높였다는 한계점이 존재한다. 이러한 부분은 추후 연구에서 모델이 사용하는 단어 세트를 넓히거나 OCR로 검출된 단어에 가중치를 부여함으로써 완화할 수 있을 것으로 예상한다.