설명 문장 생성을 통한 해석 가능한 시각적 질의응답 모델 분석

김단일[O], 한보형[*]
[O]포항공과대학교 컴퓨터공학과,
[*]서울대학교 전기·정보공학부
e-mail: daniel.kim@postech.ac.kr[O], bhhan@snu.ac.kr[*]

Interpretable Visual Question Answering via Explain Sentence Generation

Danil Kim[O], Bohyung Han[*]
[O]Dept. of Computer Science and Engineering, Pohang University of Science and Technology
[*]Dept. of Electrical and Computer Engineering, Seoul National University

● 요 약 ●

본 연구에서는 설명 문장 생성을 통한 해석 가능한 시각적 질의응답 모델을 설계하고 학습 방법을 제시한다. 설명 문장은 시각적 질의응답 모델이 응답을 예측하는 데에 필요한 이미지 및 질문 정보와 적절한 논리적인 정보의 조합 및 정답 추론 과정이 함의되어 있을 것으로 기대한다. 설명 문장 생성 과정이 포함된 시각적 질의응답의 기본적인 모델을 기반으로 여러 가지 학습방법을 통해 설명 문장 생성 과정과 응답 예측 과정 간의 상호관계를 분석한다. 이러한 상호작용을 적극적으로 활용할 수 있는 보다 개선 시각적 질의응답 모델을 제안한다. 또한 학습한 결과를 바탕으로 설명 문장의 특성을 활용하여 시각적 질의응답 추론 과정을 개선함으로써 시각적 질의응답 모델의 발전 방향을 논의한다. 본 실험을 통해서 응답 예측에 적절한 설명 문장을 제시하는 해석 가능한 시각적 질의응답 모델을 제공한다.

키워드: 시각적 질의응답(visual question answering), 해석 가능한 딥러닝(interpretable deep learning)

I. Introduction

시각적 질의응답(visual question answering, VQA) [1, 2, 3]은, 이미지와 이미지 내용과 관련한 질문이 주어졌을 때, 적절한 정답을 예측하는 모델을 설계하는 분야이다. 최근 VQA는 딥러닝(deep-learning) 기반의 모델들이 발전하였다 [1-4]. 하지만, 딥러닝 기반의 모델들은 중간 결정 과정을 사람이 이해할 수 없는 black-box의 특성이 있어, VQA 모델이 정답을 유추해 내기까지 모델이 인간과 같이 질문과 이미지를 제대로 이해하고, 필요한 유추 과정을 제대로 거쳤는지 알기 어렵다. 따라서, 정답 예측의 중간 과정들을 가시 화하고 기존의 문제들을 해결하여 VQA 모델을 더욱 발전시키려는 노력이 이어져 왔다. 본 연구에서는 설명 문장 생성 VQA 모델을 이용하여 자연어 형태의 설명 문장의 생성을 위한 과정과 정답 추론을 위한 과정이 어떻게 상호작용될 수 있는지, 각각의 학습 가중치를 변화시키면서 그 결과를 분석하고 검증하였다.

II. Related works

딥러닝이 발전하면서 시각적 질의응답 분야도 빠르게 발전 할 수 있었다. CNN 기반의 이미지 정보 임배딩과 RNN 기반의 자연어 질문 정보 임배딩을 기본으로, 두 개의 의미를 결합한 멀티모달(multi-modal) 표현을 만들고 응답을 예측하는 것이 딥러닝 기반 VQA 모델의 골자이다 [6]. 이러한 멀티모달 표현은 주어진 이미지 정보를 질문 정보 응답 예측에 잘 기여하도록 수합한 것이라고 볼 수 있다. 이미지와 질문 특징 벡터를 같은 공간에 사영시켜 간단한 덧셈, 곱셈, 이어 붙이기 등의 간단한 연산을 하거나, bilinear pooling 기법 [7], transformer 구조를 따온 co-attention 기법 [8] 등 다양한 멀티모달 표현에 대한 연구가 진행되어 왔다. 어텐션(attention) 기반의 VQA 모델은 질문과 이미지 정보로 부터 질의응답을 위해 이미지의 단위 영역마다 어느정도 정보를 고려할 것인지 가중치(weight)을 결정하여 멀티모달 표 현을 만듬으로써 최근에 좋은 성능을 보이고 있다 [1,5,9] 이후 멀티모달 표현은 정답 예측을 위한 분류기(classification) 다층 퍼셉트론을 통과하여 각 정답 후보에 대한 점수를 예측하게 된다. 학습은 입력부터 점수 출력까지 end-to-end로

학습 되며 주어진 학습 데이터에 대하여 일반적인 딥러닝의 그라디언트 뒷추적(gradient back-propagation) 방법을 사용하여 학습시킨다. 정답 후보는 학습 데이터 에서 특정 횟수 이상으로 등장한 정답들에 대해서 하나를 예측하는 것으로 설계하고, 일반적인 분류기 학습 방법과 과정을 따른다. 이러한 단순한 골격으로도 50~60%의 정확도를 보여주었다. 본 논문에서는 이러한 기본적인 VQA 모델 골자와 학습 방식을 기반으로 연구를 진행한다.

III. Experiment

VQA v2 데이터셋 샘플들과 정답 레이블, 그리고 VQA-E 데이터셋의 설명 문장 레이블을 혼합하여 학습하고 모델 성능을 검증한다.

멀티모달 표현 이하의 네트워크 부분은 응답 분류 분기와 설명 생성 분기에 대한 정보가 동시에 학습이 이루어지기 때문에 두 분기 각각 독립적으로 학습된 경우에 비해, 서로 간에 긍정 상호작용을 하는 지에 대해 분석하였다. 기존 모델에서 λ = 1 기준으로할 때, 응답 예측 분류 분기를 포함한 모델 (QI-AE)이, 제거한 모델 (QI-E)에 비해 설명 문장 예측 성능이 더 높았다. (QI-E)는 정답 레이블은 무시하고 설명 문장 레이블과 손실함수로만 학습한다. 응답 분류기 예측이 설명 생성을 방해하지 않고 강화한 것은, 정답 유추 과정과 설명 문장 생성 과정이 서로 공유하는 부분이 있음을 알 수 있다. 제안한 손실함수에서 λ 값을 변화시키면서 학습하고 성능을 검증하였는데, 이는 정답 예측과정이 중간과정 설명에 미치는 영향을 보다 정확하게 알 수 있다. λ = 10인 경우 즉, 응답 분류 분기가 설명 생성 분기의 영향에 0.1의 영향을 미치는 경우, 두 분기가 동일한 영향을 미칠때 (λ = 1)에 비해 성능이 상승하는데, 이는 분류기 예측이 생성기 예측에 방해하는 정보도 포함되어 있음을 알 수 있다.

Table 1. Test Result 1

Model	λ	B-1	B-4	M	C
QI-E	-	40.8	9.8	17.0	95.4
QI-AE	0.1	39.6	9.1	15.9	88.1
QI-AE	1	40.6	9.9	16.7	95.8
QI-AE	10	41.1	9.9	17.2	97.2
QI-AE	100	40.8	9.7	17.0	95.9

Table 1은 VQA-E 데이터셋에서 QI-AE 모델을 학습방법을 통하여 학습한 것과 응답 분류기를 떼고 설명 생성 분기만 학습한 QI-E을 VQA-E valiation 셋으로 생성된 설명문장을 검증한 결과이다.

한편 λ = 100 일 때는 λ = 10 에 비해 오히려 성능이 하락하여 QI-E 일때의 성능과 거의 동일한데, 이는 정답유추 과정의 비중이 줄어들면서 그 안에 있는 설명 생성에 방해되는 정보의 제공이 줄기 때문이라고 볼 수 있다. 그래서 λ 값이 커질수록 QI-E 일때의 성능과 흡사해지는 것을 확인할 수 있었다.

기존 모델에서 λ = 1 인 경우, 준지도학습에 의한 효과를 검증한 성능에 대한 표는 Table 2에 나타나 있다. 설명 문장 레이블이 존재하는 VQA-E 샘플에만 학습할 경우 지도학습 셋팅이 되며 (Table 2의 첫 행), 준지도 학습은 설명 문장 레이블이 없는 (VQA-E 데이터셋에 포함되지 않은) 나머지 VQA v2 샘플들로 응답 분류기를 추가 학습하여 (Table 2의 둘째 행), 생성기로 생성된 문장에 대한 성능을 비교한 것이다.

Table 2. Test Result 2

Model	Train Set	B-1	B-4	M	C
QI-AE	VQA-E	40.6	9.9	16.7	95.8
QI-AE	VQA v2	41.2	10.7	17.4	99.9

Table 2는 지도학습을 적용한 VQA-E 데이터셋과 준지도학습을 적용한 VQA v2 모델을 학습하고 VQA-E valiation 셋으로 생성된 설명문장을 검증한 결과이다.

설명 레이블이 없는 VQA v2 샘플 학습 시에 는 설명 생성기부분을 직접적으로 학습하지 않고, 설명 정보 없이 정답 정보만으로 멀티모달 표현을 학습하는 데에도, 생성된 설명 성능이 높아짐을 알 수 있다. 이는 설명을 생성하는데 필요한 이미지, 질문 간의 정보 수합 과정이 응답 예측을 위한 과정과 공통되는 부분이 존재하여, 생성을 방해하기 보다 보완적인 정보를 제공한다고 볼 수 있다.

다음으로 응답 예측 관점에서 응답 분류기 단독학습과 설명 문장 생성기 동시학습, 그리고 준지도학습에 대한 결과 관찰한 응답 예측과 설명 문장 생성에 공통적인 신호를 발견한 것과는 달리, 응답 예측 측면 에서는 설명 문장 생성기 동시 학습 (QI-AE)이 응답 분류 단독학습 (QI-A)에 다소 방해가 된 것을 알 수 있었다. VQA-E 데이터셋에서의 지도학습의 경우 그 경향이 뚜렷하며, 나머지 VQA v2 샘플들을 이용한 준지도학습으로 이러한 간극을 어느정도 극복할 수 있었다. 즉, 응답 분류기 단독학습에 이용되는 추가 샘플이 설명 문장 생성기와 동시 학습되는 샘플 외에 추가로 충분히 존재한다면, 이러한 방해현상을 극복하고 성능이 유지될 수 있는 것이다. 응답 분류기의 성능 저하 현상은, RNN으로 이루어진 생성기의 자연어 문장 생성에 필요한 정보와 과정이 복잡하여 비교적 단순한 분류기의 예측의 흐름이 방해됨을 유추할 수 있다.

Table 3. Test Result 3

Dataset	Model	VQA Acc (%)
VQA v2	QI-A	63.47
	QI-AE	63.45
VQA-E	QI-A	66.31
	QI-AE	66.09

Table 3은 지도학습을 적용한 VQA-E 데이터셋과 준지도학습을 적용한 VQA v2 모델을 학습하고 VQA-E valiation 셋으로 응답 예측 성능을 검증한 결과이다.

Table 4. Test Result 4

Model	VQA ACC	B-1	B-4	M	C
QI-AE	63.45	41.2	10.7	17.4	99.9
QI-E-A	61.35	41.5	10.1	17.0	95.5

Table 4는 VQA-E 데이터셋에서 QI-E-A과 QI-E을 학습방법을 이용하여 학습하고 VQA-E valiation 셋으로 생성된 설명문장을 검증한 결과이다.

기존 모델과 제안된 모델을 각각 VQA v2 데이터셋과 VQA-E 데이터셋을 이용하여 준지도학습을 진행한 후 설명문장을 검증해 보았다. 생성된 문장을 embedding하여 응답 예측에 넣음으로써 응답 예측 전체 과정을 설명 생성에 포함시킨 결과 나타난 성능은 기대에 미치지 못하였다.

이러한 결과가 산출된 데 대한 예상되는 원인으로는 응답 예측 분석기에 임베딩된 문장에 의미상 오류가 있는 경우와 임배딩된 정보가 정답 예측을 방해되는 noise를 많이 포함하고 있는 경우를 예상할 수 있다. 특히 학습 초기에 설명 문장 생성기가 더욱 임의의 문장을 생성하므로, 응답 분류기가 학습 초기에 잘못된 optimum으로 빠질 수 있는 원인으로 작용할 수 있다. 성능 저하의 또 다른 원인으로 생성기의 문장 생성과 임배딩에 사용되는 정보가 응답 예측 관점에서 필요 이상의 방해되는 정보들이 가해질 수 있기 때문이다.

하지만, 설명 문장 생성 관점에서, 우리 모델은 응답 예측으로부터 흐르는 그라디언트 정보를 설명 생성기가 받기 때문에, 설명 문장 생성이 응답 예측의 중간과정으로 온전히 포함되게 되고, 이는 기존 모델에 비해 응답 예측 내용과 설명 문장 내용이 보다 의미상 일관성이 있도록 유도할 수 있었다. 따라서 모델이 틀린 답을 예측 과정에 대한 보다 신뢰할 수 있는 설명 문장을 제시한다.

Table 5. Test Result 5

질문	Is this a motorcycle or bike?	
모델	기존 모델	제안 모델
응답 예측	bike	motorcycle
설명 문장	a motorcycle paralleled on the side of the street	a motorcycle is parked next to a garage with a log of books

Table 5는 기존 모델과 제안된 모델을 각각 VQA v2 데이터셋과 VQA-E 데이터셋을 이용하여 준지도학습을 진행한 후 검증한 결과이다.

IV. Conclusions

본 연구를 통해, 예측 과정을 가시화할 수 있는 설명 문장 생성을 동반한 시각적 질의응답 모델을 기반으로 설명 문장 생성과 응답 예측 상호간의 영향 및 정보교환에 대해 분석해 보았다. 설명 문장 생성과 응답 예측을 동시에 학습하고, 그 상호작용을 분석함으로 두 개의 목표가 서로 상호보완적인 정보와 과정을 교류하여 성능을 강화할 수 있음을 실험을 통해 확인하였다. 또한, 설명 문장의 언어적, 논리적 정보를 이용하여 응답 예측을 강화하고, 응답 예측 정보를 전달받아 더욱 정답과 일관성 있는 설명을 생성할 수 있는 구조의 모델을 제안하여 실험하였다. 제안된 모델은 생성된 문장의 오류 전파와 생성기 학습에 의한 방해 요인 등이 응답 예측의 성능을 저하시킬 수 있을 것으로 예측된다. 따라서 향후의 연구에서는 설명 문장 생성의 오류 전파 현상을 극복하기 위해 설명 생성 분기 이하에서 생성된 설명 문장이 일정 수준 이상의 의미를 가질 때까지 예비학습을 진행할 것을 제안한다. 또한 응답 분류기의 신호를 이용한 응답 예측이 포함된 적대적 생성 학습 방식을 적용하여 분류기의 정보와 생성기가 양성 피드백이 될 수 있도록 모델을 설계해 볼 수 있다. 이러한 방식으로 모델의 한계점을 극복한다면, 설명 정보를 활용한 보다 높은 응답예측 모델이 개발될 수 있을 것이다.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In The IEEE International Conference on Computer Vision (ICCV), December 2015.

[2] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In The IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), June 2016.

[3] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In The IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), June 2016.

[4] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question- guided spatial attention for visual question answering. In European Conference on Computer Vision, October 2016.

[5] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. CoRR, abs/1611.00471, 2016.

[6] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In Bastian

Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII, volume 9912 of Lecture Notes in Computer Science, pages 727{739. Springer, 2016.

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Conference on Empirical Methods in Natural Language Processing, (EMNLP), pages 457{468. The Association for Computational Linguistics, November 2016.

[8] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. CoRR, abs/1906.10770, 2019.

[9] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), June 2018.