이미지 질의응답을 위한 사람과 기계의 시각적 주의 집중 분석

이혁기°[1], 허유정[2], 장병탁[1,2,3]
서울대학교 인지과학 협동과정[1], 서울대학교 컴퓨터공학부[2], 서울대학교 AI 연구원[3]
{hklee, yjheo, btzhang}@bi.snu.ac.kr

A study on analysis of human and machine visual attention map for Visual Question Answering

Hyuk-Gi Lee°[1], Yu-Jung Heo[2], Byoung-Tak Zhang[1,2,3]
Interdisciplinary Program in Cognitive Science[1], Department of Computer Science and Engineering[2]
AI Institute (AIIS), Seoul National University[3]

요 약

이미지에 대한 질의응답(Visual Question Answering, VQA)은 인공지능의 시각, 언어적 처리를 동시에 수행해야하는 멀티모달 학습 문제의 한 예시 이다. 최근 이미지 질의응답 문제에 사람의 선택적 주의 집중 기제를 모방한 주의 기제(Attention mechanism)가 도입되면서 주목할만한 성능의 향상이 보고되었으며, 이후 주의 기제를 보다 정확하게 적용할 수 있는 다양한 방법론이 연구되고 있다. 본 논문에서는 주의 기제의 관점에서 사람의 시각 집중 지도와 기계가 생성한 시각 집중 지도의 상관관계를 비교한 기존 연구를 바탕으로 사람의 시각 집중 지도를 기존 주의 기제 모델에 지도 학습시키는 방법론을 제안한다. 또한 이를 활용하여 기존 모델과의 시각 집중 지도의 상관관계 및 시각 질의응답에 대한 정확도를 비교 분석한다. 제안하는 방법론을 적용한 결과, 추론된 모델의 시각 집중 지도는 사람의 시각 집중 지도와의 정량적 상관관계가 향상되었을 뿐 아니라 정성적으로도 유사한 양상을 보였으며 시각 질의응답에 대한 정확도는 유사한 결과를 보였다.

1 서 론

이미지에 대한 질의응답(Visual Question Answering, VQA)은 주어진 이미지에 대해서 자연어로 묻고 답하는 문제로, 인공지능의 시각, 언어적 처리를 동시에 요구하는 멀티모달 학습 문제의 한 예시이다. 최근 사람이 복잡한 실세계 데이터를 처리 할 때 주어진 정보에 선택적으로 집중하며 효율적으로 처리하는 주의 집중 기제를 모방하는 기계학습 모델이 제안되었고, 주목할만한 성능의 향상을 보였다. 이후, 주의 집중 기제와 관련하여 시각 이미지에 대한 질의응답을 수행할 때 사람과 기계가 주의를 집중하는 지역의 유사성을 비교 분석하는 연구도 이루어져왔다. 하지만, 아직까지는 사람과 기계가 주의를 집중하는 이미지 구역은 다소 차이를 보이며 낮은 상관관계를 보였다 [1].

본 논문에서는 이미지 질의응답을 수행할 때 사람과 기계의 시각적 주의 집중의 상관관계를 개선하고, 이미지 질의응답의 정확도를 확인한다. A. Das et al. [1]이 공개한 이미지 질의응답을 수행하는 사람의 시각 집중 지도의 주석을 바탕으로 사람의 주의 기제를 직접적으로 지도 학습시키는 목적함수를 제안하고, 이를 통해서 사람과 기계가 생성한 주의 집중 지도의 상관 계수 및 이미지 질의응답에서의 성능을 비교 분석한다.

2 관련 연구

2.1 시각 질의응답 (Visual Question Answering)

최근 이미지에 내포된 복잡한 의미를 이해하기 위한 시각 인공지능에 대한 연구가 활발히 진행되고 있으며, 이미지 질의응답(Visual Question Answering)이 이미지 이해의 수준을 측정하기 위해 제안되었다. 초기 이미지 질의응답 연구는 소수의 제한된 물체를 다루거나, 미리 정해진 형식의 질의만을 다루는 제한된 형태로 이루어졌으나 [2], 이후 [3]에서 open-ended, free-form 형태의 이미지 질의응답 데이터를 제안하며 실세계에 가까운 질의응답이 가능해졌으며, 구체적인 성능 평가 지표를 제시함으로써 수치적으로 각 모델에 대한 성능 비교가 가능해졌다.

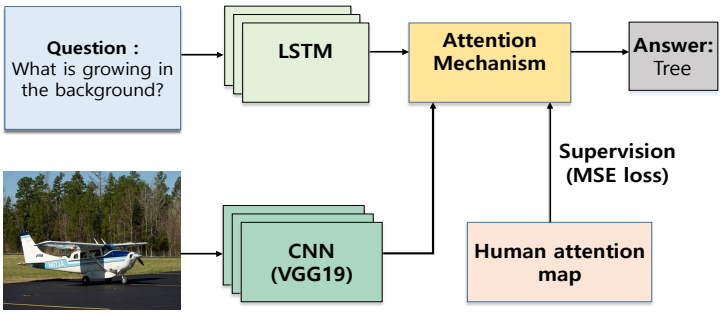

그림 1: 기본 베이스 모델인 SAN에 사람의 주의 지도(Human attention map)를 타겟으로 한 손실함수를 설정한 구조

2.2 주의 기제 (Attention mechanism)

주의 기제는 방대한 데이터 가운데 사람이 중요한 정보에 집중해서 효율적으로 처리하는 것을 모방한 것으로 자연어 처리 분야인 기계 번역에서 처음 제시되었다 [4]. 이후 컴퓨터 비전 분야에서도 시각적 집중이 활발히 적용되기 시작했으며 [5], 이미지에 대한 질의응답을 해결하기 위한 모델에도 적용되고있다 [6, 7]. [6]은 질의 중 정답과 관련성이 높은 단어와 이미지의 지역을 결합하는 방식으로 주의 기제를 적용하였으며, [7]는 질의를 단어, 어구, 문장 단위로 나누어 각각에 대해서 관련성이 높은 이미지 지역을 결합하는 방식으로 주의 기제를 적용하였다.

2.3 사람의 주의 지도 (Human attention map)

이미지 질의응답 문제에 주의 기제가 적용되는 한편, [1]에서는 이미지 질의응답 문제에 대해서 기계와 사람이 주의를 기울이는 지역의 유사도를 비교하는 연구가 이루어졌다. A. Das et al. [1]은 VQA 데이터에 대해서 사람의 시각 집중이 이루어진 지역을 지도로 표시한 새로운 데이터인 Visual Question Answering-Human

Attention Map(VQA-HAT)를 제안하였다. 이미지의 시각 집중 지역에 대한 주석은 사람이 직접 생성해야 하므로 데이터를 만드는데 시간이 오래 걸리고 비용이 비싸기 때문에 데이터의 개수가 많지 않은 단점이 있다. T. Qiao et al. [8]은 데이터가 크지 않다는 한계점을 극복하기 위해 사람과 유사한 시각 집중 지도를 생성하는 Human Attention Network(HAN)을 제안하였다.

3 모델

3.1 Stacked Attention Network (SAN)

본 논문에서는 Stacked Attention Network(SAN) [6]를 시각 질의응답의 기준 모델(baseline model)로 활용한다. SAN은 이미지 임베딩 모듈, 질의 임베딩 모듈, 이미지와 질의 임베딩 모듈을 통해 추출된 벡터에 주의 집중 기제를 적용하는 모듈, 총 세 가지 모듈로 구성되어 있다. 이미지 임베딩 모듈에서는 미리 학습된 VGG19 모델을 사용하여 이미지에서 512x14x14 차원의 이미지 특징 벡터를 추출하였다. 질의 임베딩 모듈에서는 자연어와 같은 순차 데이터를 처리하기 용이한 LSTM 모델을 적용하였고 마지막 은닉상태를 질의 임베딩 벡터로 추출하였다. 이 때, 은닉상태의 크기는 1024, 층은 1개를 사용하였다. 주의 집중 기제 적용 모듈에서는 질의를 쿼리로 삼아 질의에 연관된 이미지 특징 벡터에 가중치를 주는 방식으로 주의 집중 기제를 적용하였다. 주의 집중 기제에 대한 수식은 다음과 같다.

$$h_A = \tanh\left(W_{I,A} v_I \oplus \left(W_{Q,A} v_Q + b_A\right)\right)$$
$$p_I = \mathrm{softmax}\left(W_P h_A + b_P\right)$$

위 식에서 v_I와 v_Q는 각각 이미지, 질의 임베딩 모듈에서 추출된 이미지 임베딩 벡터, 질의 임베딩 벡터이며, $W_{[\cdot]}$와 $b_{[\cdot]}$은 학습되는 가중치와 바이어스(bias) 값이고, p_I는 이미지에 대한 시각 집중 분포를 나타내는 벡터이다.

$$\tilde{v}_I^k = \textstyle\sum_i p_i^k v_i$$
$$u = \tilde{v}_I + v_Q (k = 1)$$
$$u^k = \tilde{v}_I^k + u^{k-1} (k >= 2)$$

위 식에서 $\tilde{v}_I$는 시각 집중 분포를 나타내는 p_I를 기반으로 하는 이미지 v_I에 대한 가중합 벡터이며, u는 최종적으로 이미지 가중합 벡터 $\tilde{v}_I$와 질의 임베딩 벡터 v_Q가 결합된 쿼리 벡터이다. 위 첨자 k는 주의 집중 층수를 나타내며, 본 실험에서는 가장 좋은 성능을 냈던 2개의 주의 집중 층을 사용한다.

3.2 Attention supervision loss

그림 1과 같이 SAN을 기준 모델로 활용하여 추론된 모델의 시각 집중 지도(Attention map)가 사람의 시각 집중 지도(Human attention map)와 유사해지게 유도하는 지도 학습 손실함수를 구성하였다. 이 때, 사용되는 목적함수는 MSE(Mean-squared error)이다.

$$L_{mse}(I, q, \alpha_h) \;=\; (\alpha_h^{'}(I, q) \;-\; \alpha_h)^2$$

위 식에서 α_h는 사람 시각 집중 지도를 뜻하며, $\alpha_h^{'}$는 SAN의 마지막 주의 집중 층에서 얻어진 시각 주의 집중 지도이다. I와 q는 $\alpha_h^{'}$과 각각 한 쌍을 이루는 이미지 특징벡터, 질의 특징벡터이다.

$$L \;=\; L_{cls} + \lambda\, L_{mse}$$

최종적으로 모델을 학습시키기 위한 목적함수는 위와 같이 정의되며, L_{cls}는 이미지 질의응답에 대한 정답을 찾기 위한 목적함수이고 λ는 제어가능한 변수이다.

4 실험

4.1 실험 데이터

본 논문에서는 모델의 시각 집중 기제를 학습하기 위해 VQA-HAT [1]를 사용하였다. VQA-HAT은 VQA 1.0 데이터셋 [3]에 제시된 이미지와 질문에 대해 사람이 직접 응답을 수행하면서 나타난 시각 집중 특징을 이미지에 히트맵으로 표시한 데이터셋이다. 이미지와 질문이 한 쌍을 이루며, 각각 58,475쌍의 훈련세트와 1,374쌍의 검증세트로 구성된다. VQA 1.0은 약 80k의 훈련 이미지, 40k의 검증 이미지, 81k의 테스트 이미지로 구성되었으며, 각 이미지에 대해서 3개의 질문이 주어지고 각 질문에 대해 10명의 사람이 답한 10개 답변으로 이루어져있다.

4.2 평가 지표

모델의 성능을 평가하기 위해 시각 집중 지도의 상관성을 나타내는 Mean rank-Correlation과 시각 질의응답의 정확도 지표를 사용한다. 먼저, Mean rank-Correlation은 정답으로 주어진 사람 시각 집중 지도(Human attention map)와 모델의 시각 집중 지도(Attention map)사이의 유사성을 계산하기 위해 사용되며, Mean rank-Correlation 식은 다음과 같다.

$$Correlation = \frac{1}{3}\left(1 \;-\; \frac{6\sum_{i=1}^{3}(\alpha_h^{'} \;-\; \alpha_i)^2}{n^2(n-1)}\right) \qquad (1)$$

위 식에서 n은 rank의 개수로 본 실험에서는 시각 집중 지도(Attention map)를 14x14로 구성하여 196 이다. VQA-HAT의 검증세트는 1쌍의 이미지, 질의응답에 대해서 3명의 시각 집중 지도로 구성되어 있으므로 3명의 평균값을 적용하기 위해 $\frac{1}{3}$을 곱해준다.

시각 질의응답의 정확도는 주어진 질문이 Open-Ended이기 때문에 3명 이상의 사람의 답변과 일관되는 답변을 추론한다면 1의 정확도를 부여한다. 정확도 지표의 수식은 다음과 같다.

$$Accuracy = min\{\frac{\#\ humans\ provided\ answers}{3}, 1\} \qquad (2)$$

4.3 실험 결과

모델 학습을 위해 제안된 목적함수의 제어 가능 변수인 λ 값을 변경하면서 실험을 했을 때, 표 1에서 보듯이 λ 값이 커질수록 모델이 생성하는 시각 집중 지도(Attention map)와 사람 시각 집중 지도(Human attention map) 사이의 상관성인 Mean rank-Correlation 값이 커짐을 알 수 있다.

반면, 표 2에서 보듯이 모델의 시각 집중 지도를 훈련시키기 위해 추가적인 목적함수를 구성하더라도 질의응답에 대한 정확도는 크게 증가하거나 감소하지 않고 유사한 수준을 보였다. 다만, λ 값이 특정값을 넘어서면서부터는 기본 모델보다 성능이 일관적으로 감소하는 것으로 나타난다. 이는 목적함수의 비중이 시각 질의응답에서 정답을 찾기보다는 사람을 모방한 시각 집중 지도를 생성하는 방향으로 더 초점이 맞춰졌기 때문인 것으로 보인다.

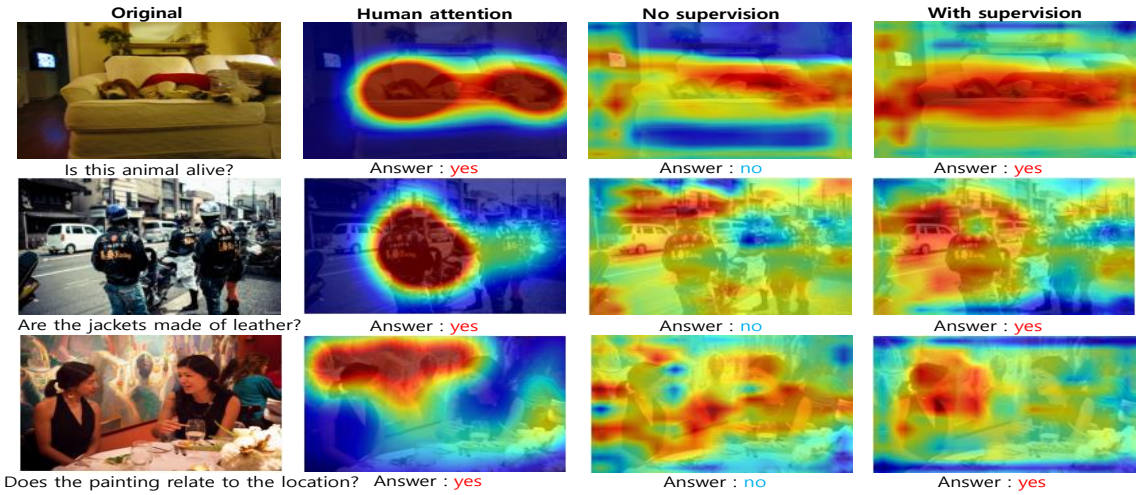

그림 2: 1열은 원본 이미지와 질의, 2열은 사람의 시각 집중 지도, 3,4열은 각각 supervision 유무에 따른 모델의 시각 집중 지도.

Model	Mean rank-Correlation
Random [1]	0.000±0.001
SAN	0.191±0.005
SAN w/ supervision($\lambda = 0.5$)	0.252±0.005
SAN w/ supervision($\lambda = 1.5$)	0.289±0.005
SAN w/ supervision($\lambda = 3.0$)	0.294±0.004
Human [1]	0.623±0.003

표 1: λ 변수의 변화에 따른 Mean rank-Correlation

Model	Accuracy(%)
SAN	49.87
SAN w/ supervision($\lambda = 0.5$)	49.99
SAN w/ supervision($\lambda = 1.5$)	49.40
SAN w/ supervision($\lambda = 3.0$)	49.45

표 2: λ 변수의 변화에 따른 VQA 1.0에 대한 Accuracy

5 결론

본 논문에서는 VQA-HAT 데이터를 바탕으로 주의 기제를 직접적으로 지도 학습시키는 목적함수를 제안하고 이를 통한 사람과 기계의 주의 집중 지도의 상관관계 및 VQA 성능을 비교 분석하였다. VQA-HAT의 데이터가 VQA에 비해서 충분히 크지 않은 관계로 VQA에서의 성능은 최신 모델과 비교하여 다소 낮은 성능을 보였지만, 본 논문에서의 기초 모델과 비교해보았을 때 정량적 성능은 유사한 수준을 유지하면서도 시각 집중 지도 사이의 상관성은 증가하는 것을 확인할 수 있다. 또한, 정성적 결과를 보면 목적함수가 적용되기 전과 비교했을 때 사람의 집중 기제와 유사한 양상으로 변화함을 확인할 수 있다.

6 감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원(2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind), 한국산업기술진흥원(P0006720-GENKO)의 지원을 받았음.

참고 문헌

[1] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?," *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.

[2] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015.

[6] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," *CoRR*, vol. abs/1511.02274, 2015.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," 2016.

[8] T. Qiao, J. Dong, and D. Xu, "Exploring human-like attention supervision in visual question answering," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.