Taesun **Whang**
NATURAL LANGUAGE PROCESSING & ARTIFICIAL INTELLIGENCE
(+82) 10-4905-1211 | taesunwhang@gmail.com | taesunwhang

Visual Dialog

Multi-view Attention Network for Visual Dialog

- Park, S.*, **Whang, T.***, Yoon, Y., & Lim, H. (2021). Multi-view attention network for visual dialog. Applied Sciences, 11(7), 3009.
- Gihtub : https://github.com/taesunwhang/MVAN-VisDial

01. 개요

- 인간은 어떤 목적을 효율적으로 달성하기 위해 대화를 하는가 하면 단순히 재미를 위해 대화를 나누기도 합니다. 이에 관련 있는 연구분야로는 Task-oriented dialog system, Open-dmain dialog system(chatbot)가 존재하며 NLG, NLU 분야에서 매우 가치 있는 연구 주제입니다. 한편, 실제 인간이 대화를 할 때 시각적 요소의 비중은 언어적 정보만큼 중요합니다.
- Visual Dialog[1]는 시각적 정보를 인지하는 dialog system 연구의 시발점이라 할 수 있습니다. Visual Dialog agent는 이미지, 해당 이미지에 대한 대화 히스토리를 바탕으로 현재 turn에 들어온 질문에 대해 retrieval 또는 generative 방식으로 질문에 대한 가장 적절한 응답을 사용자에게 제공해 주는 것을 목표로 합니다.
- 현재 turn의 질문에 coreference가 존재할 뿐만 아니라 image를 이해하여 답변을 해야 하기 때문에 single-round 형태인 Visual Question Answering (VQA) 보다 더욱 까다롭습니다. Figure 1은 VQA와 Visual Dialog의 차이점을 나타냅니다.

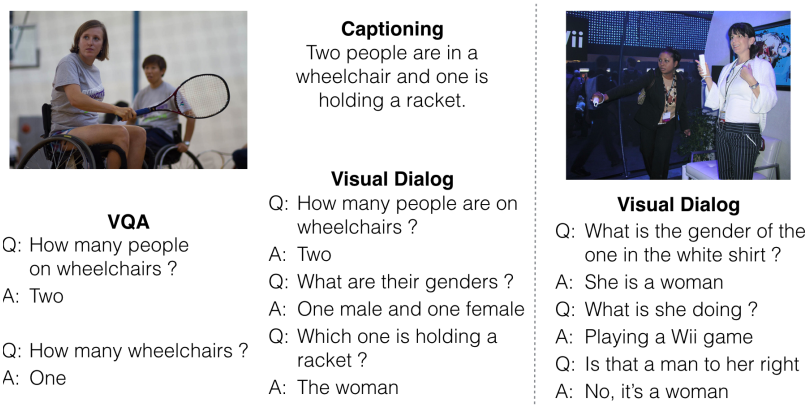

Figure 1: VQA와 Visual Dialog

02. 연구내용

- Figure 2은 Visual Dialog의 예시입니다. 이와 같이 Visual Dialog를 위한 AI agent는 대화가 진행됨에 따라 변화하는 질문의 주제를 명확히 파악해야 하며, Q5, Q6의 "they"와 같은 지시대명사, 대명사 표현들이 야기하는 참조 모호성을 해결하고 더 나아가 이미지까지의 연결하는 능력이 있어야 합니다.
- 이러한 문제점을 풀기 위해, visual co-reference resolution의 관점에서의 방법론들이 제안되었습니다. 기존의 연구들로는 RNN 기반의 모델에서부터 Attention 기반의 모델, Transformer Encoder를 기반으로 하는 모델 등이 있으며, 최근에는 Vision & Language 사전학습 모델 기반으로 하는 연구가 좋은 성능을 보여주고 있습니다.
- 하지만, Visual co-reference를 해결한다고 해서 항상 질문의 의미론적 의도와 주제를 완전히 이해하는 것은 아닙니다. 예를 들어, Figure 2의 Q6 "do they have any other snack?"의 의도는 "snack"의 존재 여부이지, "they"가 무엇을 지칭하는 지에 대해서 묻는 것은 아닙니다.

[1]Das et al. Visual Dialog. CVPR 2017.

- 이러한 관점에서 본 프로젝트에서는 질의 기반 문맥 정보와 이전 대화 맥락에서의 단서를 활용하고, 연속적 정렬 프로세스를 통해 이미지와 텍스트 정보를 효과적으로 연결하는 모델을 제안하였으며, VisDial v1.0 데이터 셋에서 뛰어난 성능을 보였습니다.

Figure 2: Visual Dialog 예시

- Figure 3은 본 프로젝트에서 제안한 Multi-View Attention Network의 구조도입니다. 모델의 입력으로는 이미지, 이전 대화 기록 그리고 질문이 들어가게 되며, Topic Aggregation module과 Context Matching module을 통해 질문의 의도와 그에 상응하는 이전 대화 정보를 융합적으로 추출하고, Modality Alignment module을 통해 이미지와 텍스트 정보 간의 의미론적 연결합니다. 각 모듈의 효과는 뒤의 Ablation 실험에서도 확인할 수 있었습니다.

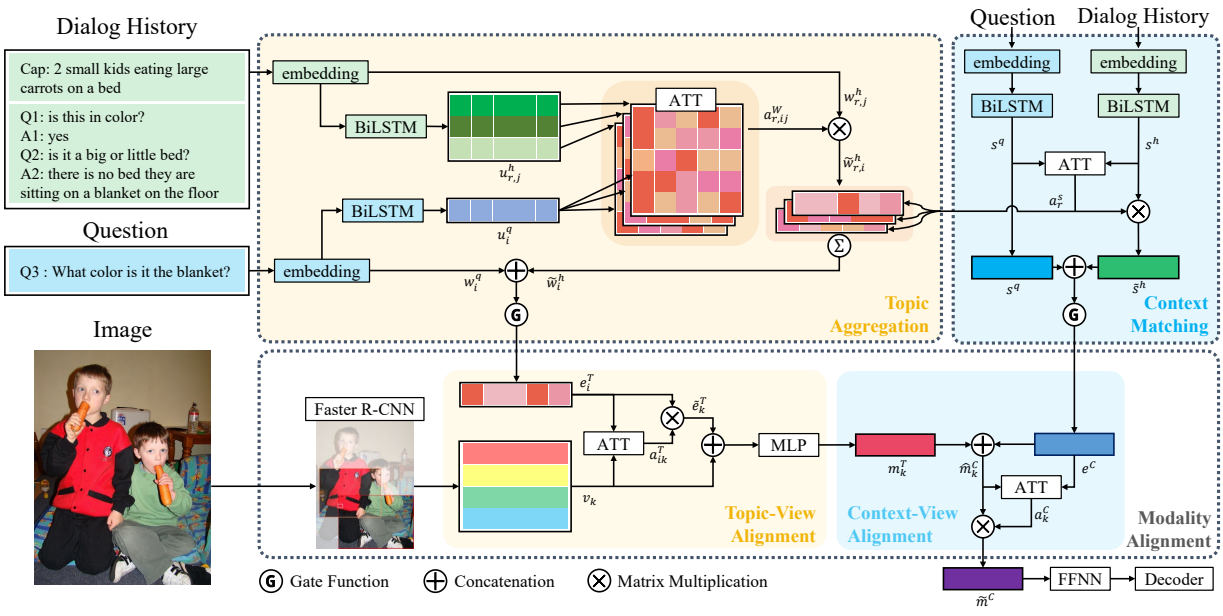

Figure 3: Multi-View Attention Network (MVAN) 구조도

03. 실험 및 결과

- 아래 표와 같이 VisDial v1.0 [2] 데이터에 대해 모델 성능 평가를 진행하였습니다. 기존 RNN 기반의 모델에서부터 attention, graph, Transformer 기반의 모델들까지 baseline으로 설정하고 성능을 비교하였고, 두가지 주요 평가 지표인 MRR과 NDCG에서 각각 좋은 성능을 보였으며 두 평가 지표를 평균 낸 경우 가장 높은 순위를 보여주었습니다.

[2] https://visualdialog.org/data

Model	AVG ↓	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
LF [7]	12	45.31 (13)	55.42 (12)	40.95	72.45	82.83	5.95
HRE [7]	12	45.46 (12)	54.16 (13)	39.93	70.45	81.50	6.41
MN [7]	11	47.50 (11)	55.49 (11)	40.98	72.30	83.30	5.92
GNN [12]	10	52.82 (10)	61.37 (10)	47.33	77.98	87.83	4.57
CorefNMN [4]	9	54.70 (9)	61.50 (9)	47.55	78.10	88.80	4.40
RVA [6]	8	55.59 (8)	63.03 (7)	49.03	80.40	89.83	4.18
DualVD [33]	7	56.32 (7)	63.23 (5)	49.25	80.23	89.70	4.11
Synergistic [9]	6	57.32 (3)	62.20 (8)	47.90	80.43	89.95	4.17
CAG [14]	5	56.64 (6)	63.49 (4)	49.85	80.63	90.15	4.11
DAN [5]	4	57.59 (2)	63.20 (6)	49.63	79.75	89.35	4.30
HACAN [34]	3	57.17 (4)	64.22 (3)	50.88	80.63	89.45	4.20
FGA [13]	2	56.90 (5)	**66.20** (1)	**52.75**	**82.92**	**91.07**	**3.80**
MVAN (ours)	1	**59.37** (1)	64.84 (2)	51.45	81.12	90.65	3.97
Synergistic † [9]	5	57.88 (4)	63.42 (5)	49.30	80.77	90.68	3.97
CDF † [11]	2	59.49 (2)	64.40 (4)	50.90	81.18	90.40	3.99
DAN † [5]	2	59.36 (3)	64.92 (3)	51.28	81.60	90.88	3.92
FGA † [13]	2	57.20 (5)	**69.30** (1)	**55.65**	**86.73**	**94.05**	**3.14**
MVAN † (ours)	1	**60.92** (1)	66.38 (2)	53.20	82.45	91.85	3.68

- MVAN은 크게 세 개의 모듈로 이루어져 있으며, Figure 4은 각 모듈의 Attention score에 대한 시각화입니다. 해당 시각화를 통해, 답변 추론 시 AI agent가 이미지의 어느 영역과 대화 히스토리의 어떤 단어들에 집중하는지 간접적으로 해석할 수 있었습니다.

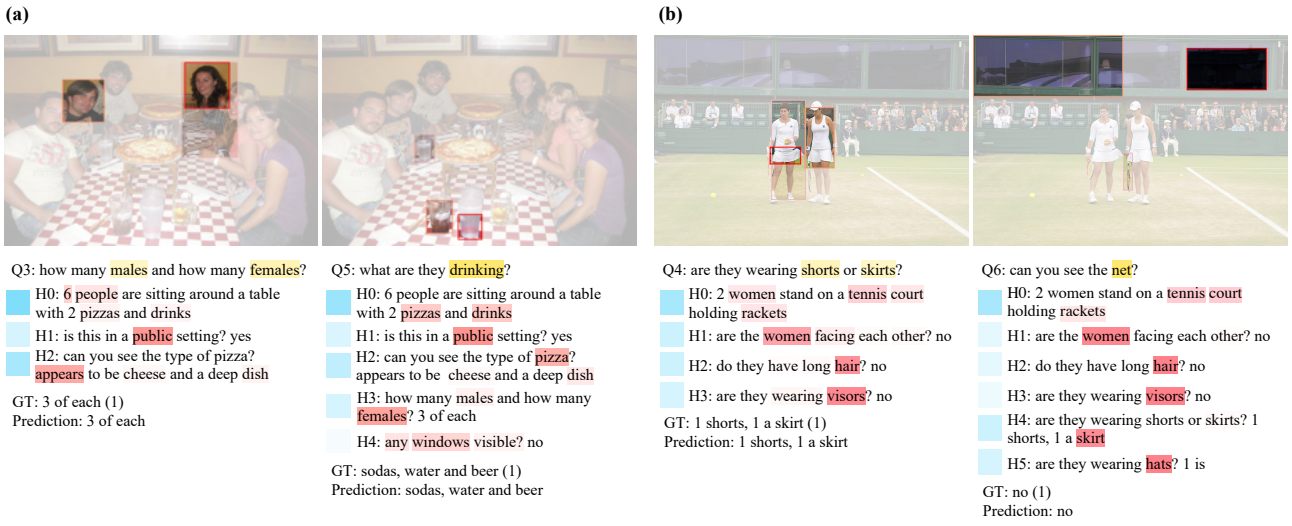

Figure 4: MVAN 모듈별 Attention score 시각화. Attention score는 강조표시로 나타내었으며, 진할수록 높은 Attention score를 의미합니다.