BI-MDRG: Bridging Image History in Multimodal Dialogue Response Generation

Hee Suk Yoon[1*], Eunseop Yoon[1*], Joshua Tian Jin Tee[1*], Kang Zhang[1], Yu-Jung Heo[2], Du-Seong Chang[2], and Chang D. Yoo[1**]

[1] Korea Advanced Institute of Science and Technology (KAIST)
[2] KT Corporation
{hskyoon,esyoon97,joshuateetj,zhangkang,cd_yoo}@kaist.ac.kr
{yj.heo,dschang}@kt.com

- Problem / objective
  - 멀티모달 대화 응답 생성 (MDRG)
- Contribution / Key idea
  -

- **Multimodal Dialogue Response Generation (MDRG) 란?**

  Task where the model needs to generate responses in texts, images, or a blend of both based on the dialogue context.

Yoon, Hee Suk, et al. "Bi-mdrg: Bridging image history in multimodal dialogue response generation." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

- **Motivation**

- 기존 연구: 텍스트 모달리티를 중간 단계로 사용함
    1. 이미지를 직접 다루기보다는, 이미지를 설명하는 텍스트 표현으로 변환한 후 처리.
    2. 또는 이미지 응답을 생성할 때도 우선 텍스트를 생성한 다음, 이를 다시 이미지로 변환하거나 설명으로 대신함.

- 기존 연구의 방식 이유:
    1. MDRG task를 위한 데이터셋 부족
        i) 멀티모달 대화 응답을 학습하려면, 다양한 "대화 맥락 + 이미지 + 응답(텍스트 또는 이미지)"의 조합이 필요
        ii) 현재 공개된 데이터셋 중 이 요구를 충족시키는 대규모 고품질 데이터셋이 존재하지 않음.
    2. Powerful pre-trained model 사용
        i) GPT, BERT, T5, BLIP, Flamingo, GPT-4V 등과 같은 사전 학습된 대형 모델들은 대부분 텍스트 중심으로 학습됨.
        ii) 따라서 기존 연구들은 이러한 모델의 성능을 최대한 활용하기 위해 텍스트 기반 처리 방식을 선호함.

- 기존 연구의 문제점:
    1. 이미지 기반 텍스트 응답의 품질 저하
        i) 이미지에 있는 특정 객체나 장면에 근거한 텍스트 응답을 생성할 때, 이미지의 세부 정보를 잃으면 대답이 부정확하거나 관련성이 떨어짐.
    2. 연속적인 이미지 응답 내 객체 일관성이 떨어짐
        i) 이는 이미지 생성 시, 이전 이미지의 객체 상태나 특성을 유지하는 메커니즘이 부족하기 때문임.
        ii) 특히 텍스트만을 기반으로 생성할 경우, 시각적 연속성을 보장할 수 없음.

전유진

Yoon, Hee Suk, et al. "Bi-mdrg: Bridging image history in multimodal dialogue response generation." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

- **Motivation**

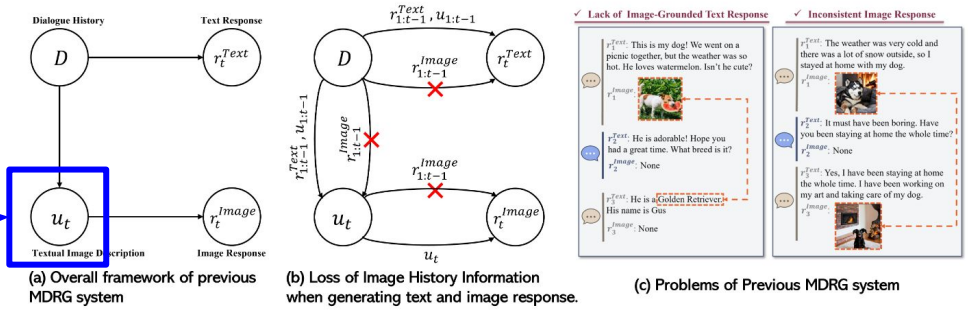

Fig. 1: (a) Outlines the framework of previous Multimodal Dialogue Response Generation (MDRG) systems, which uses the textual descriptions of images (u_t) as an intermediary step toward generating image responses (r_t^{Image}). (b) Highlights the limitations of these systems, particularly their failure to fully leverage image history ($r_{1:t-1}^{\text{Image}}$) in crafting both the textual response (r_t^{Text}) and the image response (r_t^{Image}). (c) Illustrates the consequences of this oversight, including responses that lack grounding in image context and consistency in image-based replies.

얘가 문제

(a) 저 이미지에 대한 텍스트 표현 (중간단) 이 문제
(b) 실제 이미지 히스토리 $r_{1:t-1}^{\text{Image}}$ 를 textual descriptions $u_{1:t-1}$ 로 변환하니까 모델이 시각 정보 제대로 활용 못함.
(c) 문제:
좌) 문제1: 이미지 기반 응답 제대로 못하고 있음
->원인: 이는 모델이 '수박을 먹고 있는 개'라는 텍스트를 통해서만 개를 인식하기 때문
우) 문제2: 이미지 히스토리 속 '개'에 대한 일관성이 유지되지 않는 모델의 불일치 보여줌

Yoon, Hee Suk, et al. "Bi-mdrg: Bridging image history in multimodal dialogue response generation." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

- **Contribution**

1. Bridging Image History in Text Responses (Section 3.1)
2. Bridging Image History in Image Responses (Section 3.2, 3.3)

구체적으로,
3.1 Bridging the Image History for Image-Grounded Text Response
3.2 Citation Module: Bridging the Image History to the Textual Image Description
3.3 Inference Procedure: Bridging the Image History for Consistent Image Response

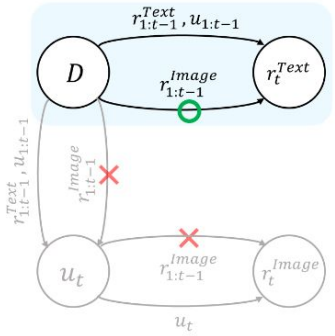

Fig. 2: Bridging Image History to the Text Response.

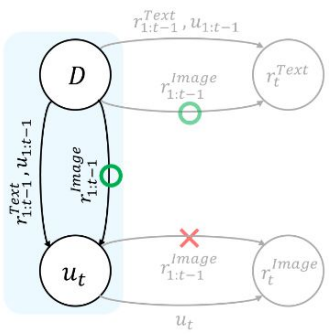

Fig. 5: Bridging Image History to the Textual Image Description.

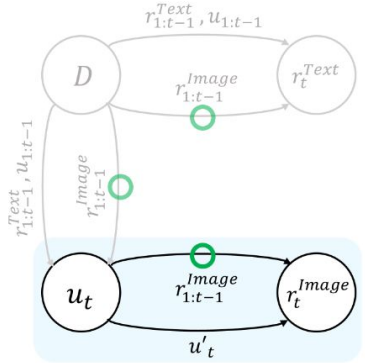

Fig. 6: Bridging the Image History to the Image Response.