The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 191–204

DOI: https://doi.org/10.5351/KJAS.2021.34.2.191
Print ISSN1225-066X / Online ISSN 2383-5818

Using similarity based image caption to aid visual question answering

Joonseo Kang[a], Changwon Lim[1,a]

[a]Department of Applied Statistics, Chung-Ang University

Abstract

Visual Question Answering (VQA) and image captioning are tasks that require understanding of the features of images and linguistic features of text. Therefore, co-attention may be the key to both tasks, which can connect image and text. In this paper, we propose a model to achieve high performance for VQA by image caption generated using a pretrained standard transformer model based on MSCOCO dataset. Captions unrelated to the question can rather interfere with answering, so some captions similar to the question were selected to use based on a similarity to the question. In addition, stopwords in the caption could not affect or interfere with answering, so the experiment was conducted after removing stopwords. Experiments were conducted on VQA-v2 data to compare the proposed model with the deep modular co-attention network (MCAN) model, which showed good performance by using co-attention between images and text. As a result, the proposed model outperformed the MCAN model.

Keywords: visual question answering, multimodal data, co-attention, image captioning, text similarity

1. 서론

시각질의응답(visual question answering; VQA)은 이미지와 질문이 주어 졌을 때 두 가지를 이용하여 답을 맞히는 것으로 이미지의 시각적 요소와 질문의 언어적 요소에 대한 세밀한 이해가 동시에 이루어져야한다 (Antol 등, 2015). 이미지와 텍스트를 통합하여 처리하는 이종 데이터 분석으로는 이미지 텍스트 매칭, 이미지 캡셔닝, 시각질의응답 등이 있다. 다른 이종 데이터 분석과 비교해서 시각질의응답은 정확한 답을 맞히기 위해서 이미지와 텍스트에 대한 더 자세한 분석을 요구한다.

어텐션 메커니즘은 단일 데이터 종류의 분석에서 뛰어난 발전을 가능하게 했다 (Vaswani 등, 2017). 이미지, 텍스트, 음성인식 등 많은 딥러닝 분야에서 활용되어 그성능을 입증하였다. 또한 이종 데이터 분석까지 발전하여 공동-어텐션(co-attention) 네트워크는 많은 이미지-텍스트 분석에 활용되고 있다. 시각적 어텐션과 유사하게 텍스트 어텐션은 텍스트에서 핵심 단어를 학습하게 만든다. 시각질의응답에서도 어텐션 네트워크가 적용되면서 답을 맞히기 위해서 이미지에서 중요한 부분을 찾고 질문에서 핵심 단어를 찾을 수 있게 되어 성능 향상에 중요한 역할을 했다. 최근 시각질의응답 연구에서는 이미지와 텍스트를 동시에 학습할 수 있는 공동 어텐션이 어떻게 구성되고 작동하는지가 중요하게 대두되었다. 하지만 시각질의응답에서 공동 어텐션

This research was supported by the Chung-Ang University Research Scholarship Grants in 2019.
[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: clim@cau.ac.kr

Published 30 April 2021 / journal homepage: http://kjas.or.kr

모델들은 이미지의 특정 부분과 질문의 특정 단어의 상관관계를 잘 학습하지 못하는 문제가 있었다 (Lu 등, 2017). 이러한 문제를 해결하기 위해 여러 층의 공동 어텐션을 사용하는 Yu 등 (2017)이 제안한 Multimodal factorized high-order (MFH), Kim 등 (2018)이 제안한 Bilinear attention networks (BANs), Yu 등 (2019)이 제안한 Deep Modular Co-Attention Networks (MCANs) 같은 여러 모델들이 제안되었고, 이미지와 텍스트 간의 상관관계를 나타낼 수 있게 되어 성능향상에 큰 영향을 끼쳤다.

우리는 Herdade 등 (2019)이 제안한 transformer 모델을 이용하여 이미지 캡션을 생성하고 그것들을 시각질의응답에서 답을 맞히기 위해 추가적인 정보로 사용하고자한다. 하지만 일부 캡션은 답을 맞히는데 오히려 방해가 될 수 있기 때문에 질문과 캡션의 유사도를 계산한 후 유사한 캡션만을 활용하는 방법을 제안한다. 제안하는 모델은 기존에 시각질의응답에서 좋은 성능을 보인 MCAN 모델에 캡션을 생성하고 유사도를 계산하여 생성된 캡션을 사용할지에 대해 판단하는 부분이 들어가는 방식으로 구성되어 있다. 기존의 MCAN의 입력 값은 이미지와 질문이었다면 제안하는 모델은 입력 값의 이미지는 동일하지만 텍스트 부분은 질문에 일정한 유사도를 넘은 캡션을 붙인 절(paragraph)로 변경하여 사용한다. 유사도가 일정한 기준을 넘지 못한다면 질문만 사용하게 된다. 이렇게 해서 이미지, 질문, 캡션 간의 공동-어텐션이 이뤄지는 효과를 바탕으로 성능향상을 기대할 수 있다. 또한 캡션을 활용하여 답만 맞추는 것이 아닌 어떻게 답을 예측했는지 보다 더 논리적으로 보여줄 수 있음을 기대할 수 있다. 본 논문에서 우리는 VQA-v2 데이터에 대하여 기존에 좋은 성능을 보인 BAN, Wu 등 (2019)이 제안한 Generating Question Relevant Captions (GQCR), MCAN 모델과 그 성능을 비교하고자 한다.

본 논문의 총 5장으로 구성되어 있다. 2장에서는 선행 연구로 기존의 모델들에 관하여 설명하고, 3장에서는 제안하는 모델의 세부적인 부분 및 구성을 자세하게 설명한다. 4장은 실험에 사용한 데이터에 관한 설명과 실험에 적용한 파라미터 처리 및 결과를 제시한다. 5장은 이 논문의 결론을 서술한다.

2. 선행 연구

2.1. Visual question answering

시각질의응답연구는 컴퓨터 비전과 자연어 처리를 결합하여 특정한 시각적 환경에 대한 이해를 위해서 제안되었다. 가장 먼저 시각질의응답에서 사용된 모델은 이미지 특징 추출을 위해서 Convolutional neural network (CNN)을 사용하고, 텍스트 특징 추출을 위해서 Long short term memory (LSTM)을 사용했다. 이 특징들은 답변을 분류하는 네트워크에서 요소별 곱셈(element-wise)을 사용하여 결합되었다. 이미지와 질문이 각각 특징으로 표현된 다음 이종 혼합 모델로 융합되어 답을 예측하게 된다 (Antol 등, 2015). 어텐션 네트워크로 인해서 컴퓨터 비전과 자연어 처리 분야에서 더 좋은 성능을 낼 수 있게 되었다 (Mnih 등, 2014; Vaswani 등, 2017). 그 다음으로 컴퓨터 비전과 자연어 처리가 합쳐진 이종 데이터 분야에서도 어텐션 네트워크가 사용되면서 획기적인 발전이 있었다 (Xu 등, 2015; Chorowski 등, 2015). 시각질의응답 분야에서도 어텐션 네트워크를 적용하여 이미지의 특징을 더 잘 추출할 수 있게 되었고 질문에서도 중요한 부분에 집중할 수 있게 되었다 (Lu 등, 2017; Anderson 등, 2018).

시각질의응답에서는 단순히 이미지, 텍스트에 대해서 어텐션 네트워크를 적용하여 성능을 향상시키는데 한계가 있었다. 한계를 넘어서기 위해서 이미지와 질문에 대한 어텐션 네트워크가 동시에 이루어져야 했다. 그래서 공동-어텐션 네트워크라는 것을 제안하여 성능을 향상시킬 수 있었고 시각질의응답에서는 다양한 공동-어텐션을 사용하는 모델들이 제안되었다 (Yu 등, 2017; Kim 등, 2018). 그 중 한 가지가 MCAN 모델이다. MCAN 모델에서는 모듈러 공동-어텐션(Modular Co-Attention; MCA)층을 제안한다. 모듈러 공동-어텐션층은 transformer 모델 (Vaswani 등, 2017)에서 제안한 것에 영감을 받은 자기-어텐션(Self-Attention; SA)과 가

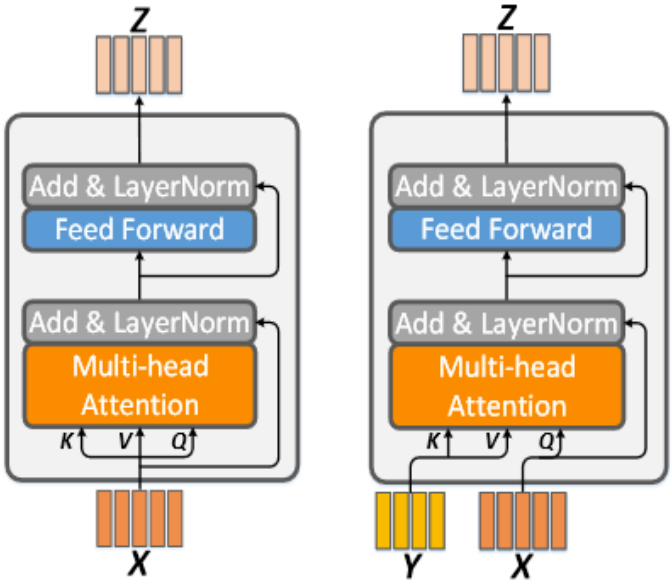

Figure 1: *Self-Attention (SA) and Guided-Attention (GA) architecture (Yu 등, 2019).*

이드 어텐션(Guided attention; GA) 두 가지 유닛을 가진다. Figure 1은 자기 어텐션과 가이드 어텐션 유닛의 구조이다. 각 자기-어텐션의 헤드는 쿼리(Q), 키(K), 밸류(V)값을 다음과 같이 계산한다.

$$Q = XW_Q, \quad K = XW_k, \quad V = XW_V, \tag{2.1}$$

여기서 X는 행렬에 쌓인 모든 입력벡터 $x_1, x_2, \ldots, x_n$을 포함하고 W_Q, W_K, W_V은 투영행렬을 학습한다. x_n은 N 토큰 세트의 n번째 토큰을 의미한다. 시각적 특징의 어텐션 가중치와 각 헤드의 출력 값은 다음과 같이 계산된다.

$$\text{head}(X) = \text{self-attention}(Q, K, V) = \text{softmax}(\Omega_A)V, \tag{2.2}$$

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}}, \tag{2.3}$$

여기서 ω_A은 $N \times N$ 어텐션 가중치 행렬이고 행렬의 요소 w_A^{mn} 은 m번째와 n번째 토큰 사이의 어텐션 가중치이다. 기존의 transformer 모델 (Vaswani 등, 2017)과 마찬가지로 $d_k = 64$의 상수 스케일링 요소를 사용한다. 식 (2.1)–식 (2.3)은 모든 헤드에 대해 독립적으로 계산된다. 8개 헤드의 출력값은 계산 후에 하나의 출력 벡터로 연결되고 학습된 투영 행렬 W_O로 곱해진다.

$$\text{MA}(Q, K, V) = \text{concat}(\text{head}_1, \ldots, \text{head}_h)W_O, \tag{2.4}$$

여기서 head_h는 번째 헤드를 의미하고 MA는 멀티헤드 어텐션을 의미한다. 인코더 층의 다음 구성요소는 각 어텐션 층의 출력 값에 적용되는 point-wise feed-forward 네트워크(FFN)이다.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{2.5}$$

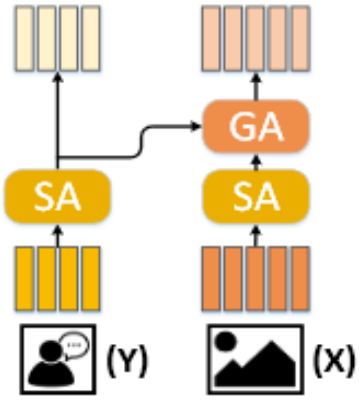

Figure 2: *SA(Y)-SGA(X,Y) architecture (Yu 등, 2019).*

여기서 W_1, b_1, W_2, b_2은 두 개의 fully connected 층의 가중치와 편향이다. 그리고 skip-connection과 층 정규화는 자기 어텐션과 피드포워드 층의 출력 값에 적용된다.

가이드 어텐션은 두 가지의 입력 특징 $X \in R^{m \times d_x}$, $Y = [y_1; \cdots; y_n] \in R^{m \times d_y}$을 가진다. 여기서 Y는 X에 대한 어텐션 학습을 안내해준다 그리고 X, Y의 모양은 가변적이어서 다른 종류의 특징을 나타내는데 사용될 수 있다(이미지-텍스트). 가이드 어텐션 유닛은 X와 Y의 각각의 쌍의 샘플인 $< x_i, y_j >$을 모델링한다. 식 (2.4)의 멀티 헤드 어텐션은 두 가지 어텐션 유닛에서 중요한 역할을 하고 입력 값의 유형에 따라 다르게 적용된다. 입력 특징 X를 갖는 SA 유닛의 경우 입력 값 $x_i \in X$를 식 (2.4)를 이용하여 MA(x_i, X, X)라는 어텐션 특징으로 계산되고 이는 다른 X의 샘플들과 정규화된 유사도를 계산하여 x_i를 재구성한다고 이해할 수 있다. 이와 유사하게 X와 Y를 입력 특징으로 갖는 GA 유닛의 경우 입력 값 $x_i \in X$가 식 (2.4)를 이용하여 MA(x_i, Y, Y)로 계산되고 Y들의 모든 샘플들과 정규화된 크로스 모달 유사도를 이용하여 재구성된다.

두 가지 유닛의 어텐션을 사용하여 시각질의응답에서 이미지와 텍스트에 대한 크로스 모달 특징을 처리하기 위해서 Figure 2와 같은 방식으로 MCA층을 구성하고 이를 SA(Y)-SGA(X, Y)라고 표기한다. SA(Y)를 통해서 질문 쌍 $\{y_i, y_j\} \in Y$에서 세밀하게 인트라 모달 상호작용을 모델링할 수 있다. 또한 GA(X, Y)를 통해서 이미지의 부분 $x_i \in X$과 질문의 단어 $y_j \in Y$의 인터 모달 상호작용을 모델링 할 수 있다.

2.2. Image captioning

이미지 캡셔닝은 이미지를 잘 설명할 수 있는 문장을 생성하기 위한 작업이다. 이미지 캡셔닝은 시각적 정보와 자연어 처리 정보를 함께 처리해야하는 이종데이터 분석 작업 중 하나이다. 시각적 정보 측면에서 객체 탐지 연구의 발전은 이미지 캡셔닝 발전에 큰 기여를 했다. 자연어 처리 정보 측면에서 어텐션 기반의 순환신경망 네트워크들의 발전 또한 이미지 캡셔닝의 발전에 큰 기여를 했다 (Xu 등, 2015).

먼저 이미지로부터 객체 탐지 방법인 Faster R-CNN과 ResNet-101을 이용하여 특징을 추출해낸다 (Anderson 등, 2018). ResNet-101의 중간 특징 지도(feture map)를 입력 값으로 사용하여 Region Proposal Network (RPN)은 객체 제안에 대한 바운딩 박스를 생성한다. Maximum suppression을 사용하지 않고 임계값인 Intersection over Union (IOU) 0.7을 초과하는 겹쳐진 바운딩 박스를 제거한다. 그 후에 Region of Interest (ROI) 풀링 층은 나머지 바운딩 박스를 동일한 크기로 변환하는데 사용한다 (e.g., $14 \times 14 \times 2048$). 추가 CNN 층은 각 제안된 박스에 클래스 라벨과 바운딩 박스를 조절하는데 적용된다. 그리고 클래스 예측 확률이 임계값 0.2

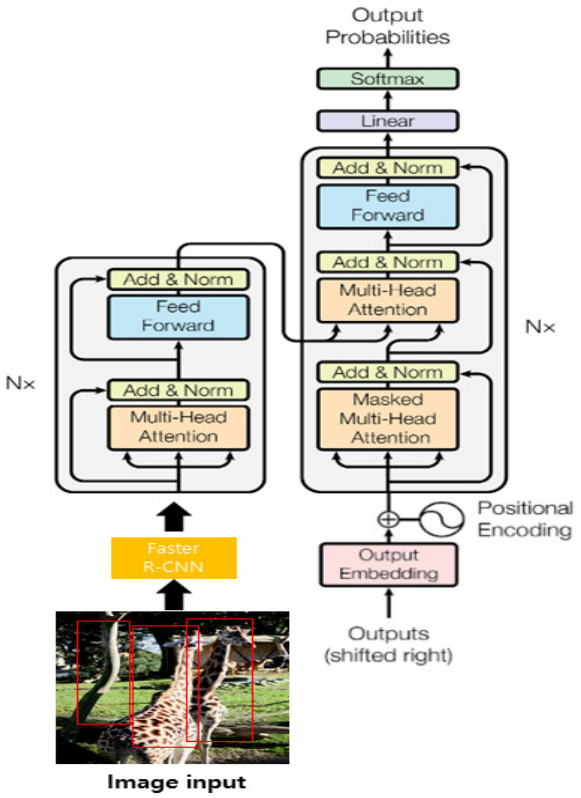

Figure 3: *Standard transformer architecture.*

를 넘지 못한 모든 바운딩 박스들은 사용하지 않는다. 마지막으로 공간 차원에 평균 풀링을 적용하여 각 객체 바운딩 박스에 대해 2048차원의 특징 벡터를 생성한다. 이렇게 생성된 특징 벡터는 transformer 모델의 입력 값으로 사용된다.

Figure 3은 transformer 모델의 구조를 보여준다. Transformer 모델은 인코더-디코더 구조로 구성되어 있으며 인코더, 디코더 모두 6개의 층을 쌓아서 구성된다. 이미지 캡셔닝의 경우 이미지로부터 앞에서 말한 방법을 이용하여 특징벡터를 추출한 것을 입력 값으로 이용하여 출력 값으로 문장(이미지 캡션)을 생성 한다. 모든 이미지 특징 벡터는 입력 임베딩 층을 통해 먼저 처리되며 이는 ReLU 및 드롭 아웃 층과 fully-connected 층을 통과 한다. 이렇게 여러 층을 통과하게 되면 2048차원에서 512차원으로 입력 차원을 줄일 수 있게 된다. 그 후에 임베딩된 특징 벡터는 transformer 모델의 첫 번째 인코더에 대한 입력 토큰으로 사용한다. 그리고 2번째 인코더부터 6번째 인코더까지는 이전 인코더의 출력 토큰을 현재 층에 대한 입력 값으로 사용한다. 각 인코더의 층은 멀티헤드 어텐션 층과 피드포워드 뉴럴 네트워크로 구성되어 있다. 자기 어텐션 층은 8개의 동일한 헤드로 구성되어있다. 각 어텐션의 헤드는 2.1장에서 기술한 식 (2.1)–(2.5)를 이용하여 계산한다.

그 다음 디코더는 인코더의 마지막 층으로부터 생성된 토큰을 입력 값으로 사용하여 캡션 텍스트를 생성한다. Transformer 인코더의 출력 토큰의 차원은 Vaswani 등 (2017)이 제안한 기존의 transformer 모델을 변형 없이 그대로 사용한다. Figure 4는 standard transformer를 이용하여 생성된 캡션들의 예시이다.

Figure 4: *Example of generated captions using the standard transformer model.*

3. 모델 제안

3.1. Representation

본 장에서는 2장에서 기술한 transformer와 MCAN 모델들을 바탕으로 하여 제안하는 새로운 모델인 Deep Modular Co-Attention Networks with similar caption (MCAN-Cap) 모델에 대하여 기술한다. Figure 5는 MCAN-Cap 모델의 구조이다. 입력이미지는 bottom-up 방식으로 지역적 시각 특징들의 집합으로 표현된다 (Anderson 등, 2018). 이 특징들은 Krishna 등 (2016)이 제공한 Visual Genome 데이터 셋에 대하여 사전 훈련된 Faster R-CNN과 ResNet-101을 통해 추출된 중간 단계의 특징들이다. 우리는 탐지된 객체의 확률에 대한 신뢰 임계값을 설정하고 동적 객체 $m \in [10, 100]$를 얻는다. i번째 객체의 경우 탐지된 영역에서 컨볼루션 특징을 평균 풀링하여 특징 $x_i \in R^{d_x}$로 표현된다. 이미지는 특징 행렬 $X \in R^{m \times d_x}$로 표현된다.

입력텍스트를 먼저 단어로 토큰화하고 최대 18단어까지 표시한다 (Teney 등, 2017). 기존 시각질의응답 연구들은 최대 14단어까지 표시했지만 우리는 질문과 캡션을 같이 사용하는 경우도 있기 때문에 최대 18단어까지 사용한다. 질문과 캡션의 각 단어들은 대규모 말뭉치에서 사전 훈련된 300차원의 GloVe 단어 임베딩을 사용하여 벡터로 변환된다 (Pennington 등, 2014) 그 다음으로 생성된 캡션과 질문 간의 유사도를 계산한다. 추가적으로 답변과의 유사도도 같이 계산한다. 유사도는 다음과 같이 구한다 (Li 등, 2018).

$$s(w_i, w_j) = \frac{1}{2}\left(1 + \frac{w_i^T w_j}{\|w_i\| \cdot \|w_j\|}\right), \tag{3.1}$$

$$S(Q, C) = \frac{1}{T_q} \sum_{w_i \in W_q} \max_{w_j \in W_C} s(w_i, w_j), \tag{3.2}$$

$$S(A, C) = \frac{1}{T_a} \sum_{w_i \in W_a} \max_{w_j \in W_C} s(w_i, w_j), \tag{3.3}$$

$$S(< Q, A >, C) = \frac{1}{2}(S(Q, C) + S(A, C)), \tag{3.4}$$

여기서 w_i, w_j은 i, j번째 단어, 캡션은 C, 질문은 Q, 답변은 A이다. 캡션과 질문, 답변을 토큰화한 것은 각각

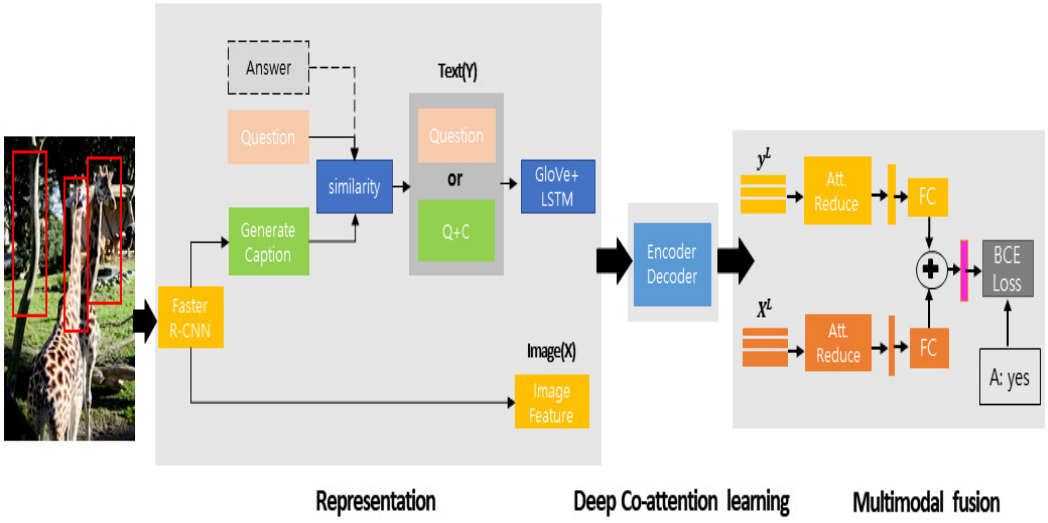

Figure 5: *MCAN-Cap model architecture.*

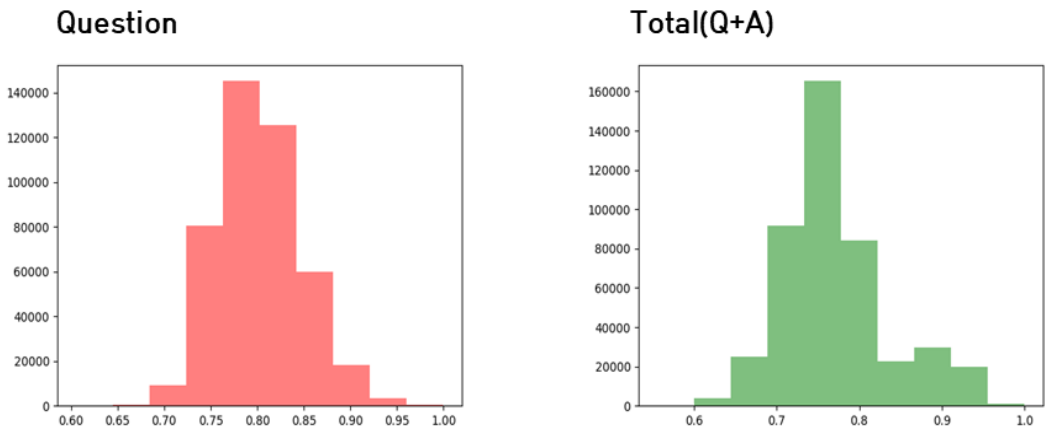

Figure 6: *Question and caption similarity distribution and question and answer with caption similarity distribution.*

$W_c = \{w_1, \ldots, w_{T_c}\}$, $W_q = \{w_1, \ldots, w_{T_q}\}$, $W_a = \{w_1, \ldots, w_{T_a}\}$로 표시하고, 여기서 T_c, T_q, T_a들은 각각 캡션, 질문, 답변의 단어 수이다. 이렇게 계산한 질문과 캡션의 유사도 분포와 질문과 답 그리고 캡션의 유사도 분포는 Figure 6과 같다.

생성된 캡션과 질문과의 유사도를 계산한 후 특정 임계값을 설정하고 유사도가 임계값 이상인 경우에는 질문 뒤에 캡션을 붙인 절을 텍스트 입력 값으로 사용한다. 임계값 미만의 경우에는 캡션은 사용하지 않고 질문만 텍스트 입력 값으로 사용한다.

텍스트 입력 값은 $n \times 300$, $n \in [1, 18]$ 크기의 단어 시퀀스로 임베딩 된다. 그 후에 단어 임베딩은 1층의

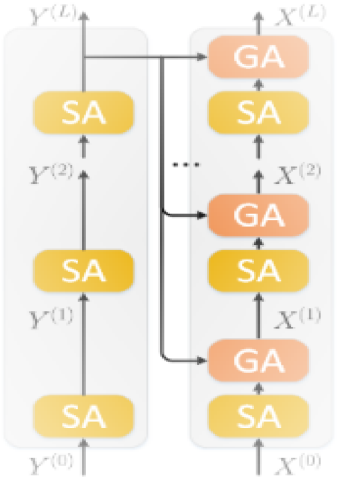

Figure 7: *Encoder-decoder architecture of SA(Y)-SGA(X,Y) (Yu et al., 2019).*

LSTM 네트워크를 통과한다. 그렇게 해서 마지막 단어의 출력 특징만 사용하는 다른 모델들보다 비교적 모든 단어들의 특징을 잘 추출한 텍스트 특징 행렬 $Y \in R^{n \times d_y}$을 출력한다.

가변적인 객체들의 수 m과 텍스트의 길이 n을 처리하기 위해서 제로 패딩을 사용하여 X와 Y을 최대크기로 채운다 ($m = 100$, $n = 18$). 훈련 중에 언더플로우 문제를 방지하기 위해서 모든 소프트맥스 층보다 먼저 확률을 $-\infty$부터 0까지의 패딩 로짓으로 채운다.

3.2. Deep co-attention learning

3.1절에서 구한 이미지 특징 X와 텍스트 특징 Y을 입력 값으로 사용하여 우리는 입력 특징을 깊게 연결된 L개의 MCA 층($\text{MCA}^{(1)}, \text{MCA}^{(2)}, \ldots, \text{MCA}^{(L)}$)을 통과시켜 깊은 공동-어텐션 학습을 할 수 있게 한다. $\text{MCA}^{(l)}$의 입력 특징은 $X^{(l-1)}$와 $Y^{(l-1)}$로, 출력 특징은 $X^{(l)}$와 $Y^{(l)}$로 표기한다. 이러한 방식으로 $\text{MCA}^{(l+1)}$의 입력 값은 $\text{MCA}^{(l)}$의 출력 값이고 재귀적인 방식으로 층을 쌓는다. 이를 표현하면 다음과 같다.

$$\left[X^{(l)}, Y^{(l)}\right] = \text{MCA}^{(l)}\left(\left[X^{(l-1)}, Y^{(l-1)}\right]\right), \tag{3.5}$$

여기서 $\text{MCA}^{(1)}$의 경우 입력 특징은 $X^{(0)} = X$, $Y^{(0)} = Y$이다.

2장에서 설명한 SA(Y)-SGA(X,Y) 구조를 이용하여 transformer 모델에서 영감을 받은 Figure 7과 같은 인코더 디코더 구조의 깊은 공동-어텐션 모델을 구성한다. 각 $\text{MCA}^{(l)}$에서 GA 유닛의 입력 특징 $Y^{(l)}$을 마지막 MCA 층의 텍스트 특징 $Y^{(L)}$로 대체한다. 인코더 디코더 구조는 L개의 SA 유닛을 거쳐 학습된 텍스트 특징 $Y^{(L)}$와 SA유닛을 거쳐 학습된 각 층의 이미지 특징 $X^{(L)}$을 GA 유닛의 입력 값으로 사용하여 층을 쌓아 학습하는 SGA 유닛을 가진 디코더 구조로 이해할 수 있다.

3.3. Multimodal fusion and output classifier

깊은 공동-어텐션 학습단계 이후 출력 이미지 $X^{(L)} = [x_1^{(L)}; \cdots; x_m^{(L)}] \in R^{m \times d}$와 텍스트 특징 $Y^{(L)} = [y_1^{(L)}; \cdots; y_n^{(L)}] \in R^{n \times d}$는 이미 절(텍스트)에 있는 단어와 이미지 영역에 대한 어텐션 가중치로 인해 충분한 정보를 가지게 된다.

그래서 각각의 출력 값 $X^{(L)}$과 $Y^{(L)}$을 2층의 Multi-Layer Perceptron (MLP) 로 구성된 어텐션 감소 모델을 통과시켜서(FC(d)-ReLU-Dropout(0.1)-FC(1)) $\tilde{x}, \tilde{y}$를 얻는다. 예를 들어, $X^{(L)}$의 경우 어텐션 가중치를 얻은 특징은 다음과 같이 구한다.

$$\alpha = \text{softmax}\left(\text{MLP}\left(X^{(L)}\right)\right), \tag{3.6}$$

$$\tilde{x} = \sum_{i=1}^{m} \alpha_i x_i^{(L)}, \tag{3.7}$$

여기서 $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m]$은 학습된 어텐션 가중치이고 같은 방식으로 을 구할 수 있다. 계산된 $\tilde{x}, \tilde{y}$을 이용하여 선형 다종 혼합(linear multimodal fusion) 함수를 다음과 같이 계산 한다.

$$z = \text{LayerNorm}\left(W_x^T \tilde{x} + W_y^T \tilde{y}\right), \tag{3.8}$$

여기서 $W_x, W_y \in R^{d \times d_z}$은 두 개의 선형 투영(linear projection)행렬이고 d_z은 혼합된 특징의 공통 차원이다. Ba 등 (2016)이 제안한 LayerNorm은 훈련을 안정시키기 위한 층 정규화함수이다. 혼합된 특징 z은 시그모이드 함수에 따라서 벡터 $s \in R^N$에 투영된다. 여기서 N은 훈련 세트에서 가장 많이 나오는 대답의 수다. 그 후 다음과 같은 이종 크로스 엔트로피(Binary cross entropy; BCE) 손실함수를 사용하여 혼합된 특징을 이용한 N 가지 방식의 분류기를 훈련시킨다.

$$\text{BCE}(z) = -\frac{1}{N} \sum_{N}^{i=1} y_i \log((h(z_i; \theta)) + (1 - y_i) \log(1 - h(z_i; \theta)), \tag{3.9}$$

여기서, N은 답의 전체 개수이고 $y_i \in \{0, 1\}$는 i번째 정답, h는 시그모이드 함수이다.

4. 실험

우리는 VQA-v2 데이터 셋에 대하여 실험을 진행하였다. 캡션을 이용하였을 때 좋은 성능을 보이는지 확인을 위하여 모든 캡션을 사용하여 실험을 진행하였고, 질문과 캡션과의 유사도를 이용해서 선별한 캡션을 사용한 실험을 진행하였다. 추가적으로 캡션에서 불용어를 제거하고 실험을 진행하였다.

4.1. 데이터 설명

VQA-v2 : 본 논문에서는 이전 버전과 비교하여 시각적 이해를 강조하고 답변의 편향됨을 개선한 VQA-v2 데이터 셋을 사용하여 제안하는 모델을 평가한다. 이러한 개선점은 모델이 질문 및 이미지의 보다 효과적인 공동-어텐션을 사용하도록 유도하고 이는 우리의 모델에서 공동-어텐션을 활용하는 점에서 적합하다고 할 수 있다. 데이터 셋은 훈련데이터 (약 8만개의 이미지, 44만개의 질문과 답), 모델평가에 사용되는 데이터 (약 4만개의 이미지, 21만개의 질문과 답)로 구성된다. 실험 결과는 3가지 종류의 질문(Yes/No, Number, Other)에 대한 정확도와 전체 정확도로 평가한다. Figure 8은 시각질의응답 데이터 셋의 예시이다. 하나의 이미지와 평균 5개의 질문과 답변으로 구성되어 있고 답변을 기준으로 3가지 범주의 질문과 답변이 존재함을 확인할 수 있다.

4.2. 실험 준비

Antol 등 (2015)이 제공한 MSCOCO 데이터 셋에 대해서 사전 훈련된 transformer모델을 이용하여 훈련, 검증 데이터 셋의 약 12만개의 이미지에 대한 캡션을 생성한다. 그 후에 3.1절에서 소개된 유사도를 이용하여

Question: What is this piece of furniture used for?
Answer: Sitting

Question: What room is this?
Answer: Living room

Question: How many reading materials are laying on the coffee table?
Answer: 3

Question: Did the family forget to turn the TV off?
Answer: No

Question: Is the person who lives here a musician?
Answer: No

Figure 8: *Example of VQA data (https://visualqa.org/).*

생성된 캡션과 질문과의 유사도를 계산한다. Figure 6에서 두 분포간의 차이가 있기 때문에 유사도 임계값을 설정할 때 차이가 있을 것으로 판단을 할 수 있다. 불용어를 제거하고 단어별로 품사를 붙이는 tag 작업은 Loper과 Bird (2002)가 제안한 NLTK패키지를 사용하여 진행하였다. 유사도는 최솟값부터 0.01단위로 변경하며 실험을 진행하였다. 실험에 사용될 초매개변수인 input image feature의 차원 d_x, input question features의 차원 d_y, fused multi-modal features의 차원 d_z는 각각 2,048, 512, 1,024이다. 멀티 헤드 어텐션의 잠재적인 차원 d은 512이고 헤드 수는 8로 설정하고 각 헤드마다 잠재적인 차원은 $512/8 = 64$이다. 우리 실험에서는 MCA 층은 중에서 선택하였으며 답변은 개의 집합에서 선택한다. MCAN 모델 훈련을 위해서 $B_1 = 0.9, B_2 = 0.98$의 Adam을 사용한다. 처음의 학습률은 t가 1부터 시작하는 현재 에폭이고 $(2.5te^{-5}, 1e^{-4})$ 중에서 선택한다. 10번의 에폭 이후에 학습률은 2번의 에폭마다 1/5씩 감소한다. 모든 모델은 14번의 에폭까지 같은 배치사이즈인 64로 학습을 진행하였다. 검증 데이터 셋의 실험결과를 위해서는 오직 훈련 데이터 셋만 사용하여 훈련을 진행하였다.

4.3. 결과

유사한 캡션을 사용했을 때의 성능을 비교하기 위해 우리는 BAN과 GQRC 두 모델과 비교실험을 진행하였다. VQA-v2 훈련, 검증 데이터 셋을 이용하여 실험을 진행하였다. 실험 결과는 Yes/no, Num, Other 그리고 전체 정확도(All)를 이용하여 평가한다. Tables 1–4는 실험의 결과이다.

Table 1은 우리 모델을 다른 시각질의응답 모델들과 성능 비교한 결과 표, Table 2는 다양한 tag를 이용하여 성능을 비교한 결과 표, Table 3은 질문과 캡션과의 유사도를 이용한 성능과 답변과 질문 캡션의 유사도를 이용한 성능을 비교하는 결과표이다. Table 4는 두 가지도 유사도에서 0.1 단위 구간별로 가장 성능이 좋게 나타난 것을 표시한 결과표다. Table 1을 보면 비교 모델인 BAN과 GQRC, MCAN에 비해 불용어를 제거하고 질문과 유사한 캡션만 활용한 우리모델이 yes/no 범주를 제외한 Number, Other 그리고 전체 성능이 가장 좋은 것을 확인할 수 있었다. 이때 임계값은 0.68로 유사도가 임계값 이상인 캡션들만 선별하여 활용하였다. 범주별 비율은 yes/no(38.37%), number(12.31%), other(49.32%)이다. 가장 비율이 높은 other 범주에서 성능이 향상되었으며 모든 모델에서 정답률이 낮은 number 범주에서 1.5% 이상의 성능향상을 확인할 수 있다. 이를

Table 1: VQA-v2 data result

Model	yes/no	Num	Other	All
BAN	82.63	47.04	57.76	66.08
GQRC	82.60	43.90	56.40	65.80
MCAN	84.83	48.52	58.69	67.20
Ours(stopword)	84.71	50.04	58.85	67.41

Table 2: Results by tag

Model	yes/no	Num	Other	All
N	84.51	49.70	58.82	67.28
N+J	84.45	49.60	58.73	67.21
N+V	84.56	49.30	58.80	67.24
J+V	84.35	49.20	58.38	66.93
All caption	84.49	49.97	58.81	67.30

Table 3: Two type similarity results

Model	yes/no	Num	Other	All
MCAN	84.83	48.52	58.69	67.20
All caption	84.49	49.97	58.81	67.30
Total similarity	86.28	49.64	59.35	68.19
Question similarity	84.44	50.00	58.84	67.32

Table 4: Results by similarity

Question		Total	
Question Similarity	All	Total Similarity	All
0.67	67.26	0.66	67.34
0.74	67.32	0.76	68.19
0.86	67.15	0.86	67.56

통해 캡션을 선별해서 활용했을 때 비교적 어려운 문제를 더 잘 맞히게 된다는 점을 확인할 수 있다.

또한, Table 2를 보면 tag별 결과 알 수 있는데, N은 명사, J는 형용사, V는 동사이다. 기초 모델인 MCAN에 비해서 형용사, 동사만 사용한 조합을 제외하고 모두 성능향상이 있었다. 하지만 모든 캡션을 사용했을 때보다 일부 tag만 활용했을 때 성능향상이 없다.

Table 3의 두 가지 유사도에 대한 결과를 보면, 비교 모델인 MCAN에 비하여 답변과 질문과의 유사도를 이용한 모델은 약 1%의 성능향상이 있었다. 하지만 답변을 활용했기 때문에 실제로 활용하기 어렵다. 질문과의 유사도를 이용한 모델은 MCAN모델을 이용했을 때보다는 성능향상이 있었지만 모든 캡션을 사용했을 때보다 큰 성능향상을 확인할 수 없었다. 마지막으로 Table 4를 보면 답변까지 이용했을 경우가 질문만 이용했을 때보다 모두 성능이 좋지만 이는 답변을 이용했기 때문에 당연한 결과라 할 수 있다. 두 가지 유사도 모두 평균 근처에서 가장 성능이 좋았는데 이는 모든 캡션을 활용했을 때보다는 유사도를 이용해서 선별된 캡션을 이용했을 때 성능 향상을 할 수 있음을 알 수 있게 해준다.

5. 결론

본 논문에서는 이미지를 통해 생성된 캡션을 질문과의 유사도를 이용하여 선별하여 사용하는 모델을 제안하였다. 선별된 캡션을 사용하였을 때 기존 시각질의응답 모델들에 비해서 어떻게 답변을 추출했는지에 대한

논리를 더 잘 보여줄 수 있다.

캡션을 활용하는 것은 이미지를 보고 사람이 하는 사고 과정을 좀 더 비슷하게 모델이 따라할 수 있게 해주는 것을 기대하게 한다. 이미지에 있는 잠재적인 정보들을 보다 더 잘 활용할 수 있게 되어 시각질의응답에서 더 좋은 성능을 보일 것을 기대한다. 모든 캡션을 사용했을 때보다 질문과 유사한 캡션을 선별해서 활용했을 때 답변에 더 큰 영향을 미칠 것이라 기대한다.

시각질의응답연구에서 가장 활발히 사용되는 VQA-v2 데이터 셋에 대하여 정확도를 비교하는 실험을 진행하였다. 비교 모델로 BAN과 GQRC, MCAN을 사용하여 제안된 선별된 캡션을 활용하는 모델이 비교모델보다 좋은 성능을 보이는 것을 확인하였다. 또한 다양한 태그를 이용한 실험을 통해 일부 태그만 활용하는 것보다 모든 캡션을 그대로 활용하는 것이 더 좋은 성능을 보이는 것을 확인하였다.

우리는 시각질의응답에서 캡션을 선별하여 활용함으로써 성능을 높일 수 있고, 이미지에서 더 많은 정보를 가질수록 성능향상에 도움이 되는 것을 확인하였다. 비교모델로 최근에 연구가 활발히 진행 중인 Visual Bert기반의 모델을 활용하지 못한 점이 아쉬우며, 더 다양한 유사도를 이용한 실험을 진행하고자 한다.

References

Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, and Zhang L (2018). Bottom-up and top-down attention for image captioning and visual question answering, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086.

Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick L, and Parikh D (2015). VQA: Visual question answering. *In Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Ba JL, Kiros JR, and Hinton GE (2016). Layer Normalization, arXiv preprint arXiv:1607.06450.

Chorowski JK, Bahdanau D, Serdyuk D, Cho K, and Bengio Y (2015). Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems (NIPS)*, 577–585.

Herdade S, Kappeler A, Boakye K, and Soares J (2019). Image Captioning: Transforming Objects into Words. In *Advances in Neural Information Processing Systems*, Mit Press, Cambridge, MA, USA, 11137–11147.

Kim JH, Jun J, and Zhang BT (2018). Bilinear attention networks. In *Advances in Neural Information Processing Systems*, **31**, 1564–1574.

Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS, and Li FF (2016). Visual Genome: connecting language and vision using crowdsourced dense image annotations, arXiv preprint arXiv:1602.07332.

Li Q, Tao Q, Joty S, Cai J, and Luo J (2018). VQA-E: Explaining, elaborating, and enhancing your answers for visual questions, arXiv preprint arXiv:1803.07464.

Loper E and Bird S (2002). NLTK: The natural language toolkit, *ETMTNLP '02: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, **1**, 63-–70.

Lu J, Yang J, Batra D, and Parikh D (2017). Hierarchical question-image co-attention for visual question answering, arXiv preprint arXiv:1606.00061.

Mnih V, Heess N, Graves A, and Kavukcuoglu K (2014). Recurrent models of visual attention. In *Advances in neural information processing systems (NIPS)*, 2204-2212.

Pennington J, Socher R, and Manning CD (2014). GloVe: Global vectors for word representation. *In Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Teney D, Anderson P, He X, and Hengel A (2017). Tips and tricks for visual question answering: Learnings from the 2017 challenge, arXiv preprint arXiv:1708.02711.

Vaswani A, Shazeer M, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017). Attention is all you need. *In Advances in Neural Information Processing Systems*, 6000–6010.

Wu J, Hu Z, and Mooney R (2019). Generating question relevant captions to aid visual question answering, arXiv preprint arXiv:1906.00513.

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015). Show, attend and tell: Neural image caption generation with visual attention, arXiv preprint arXiv:1502.03044.

Yu Z, Yu J, Cui Y, Tao D, and Tian Q (2019). Deep modular co-attention networks for visual question answering, arXiv preprint arXiv:1906.10770.

Yu Z, Yu J, Xiang C, Fan J, and Tao D (2017). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *In Proceedings of the IEEE*, **26**, 2275–2290.

Received December 22, 2020; Revised January 18, 2021; Accepted January 27, 2021

유사도 기반 이미지 캡션을 이용한 시각질의응답 연구

강준서[a], 임창원[1,a]

[a]중앙대학교 응용통통계학과

요 약

시각질의응답과 이미지 캡셔닝은 이미지의 특징과 문장의 언어적인 특징을 이해하는 것을 요구하는 작업이다. 따라서 두 가지 작업 모두 이미지와 텍스트를 연결해 줄 수 있는 공동 어텐션이 핵심이라 고 할 수 있다. 본 논문에서는 MSCOCO 데이터 셋에 대하여 사전 훈련된 transformer 모델을 이용 하여 캡션을 생성한 후 이를 활용해 시각질의응답의 성능을 높이는 모델을 제안하고자 한다. 이때 질 문과 관계없는 캡션은 오히려 시각질의응답에서 답을 맞히는데 방해가 될 수 있기 때문에 질문과의 유사도를 기반으로 질문과 유사한 일부의 캡션을 활용하도록 하였다. 또한 캡션에서 불용어는 답을 맞히는데 영향을 주지 못하거나 방해가 될 수 있기 때문에 제거한 후에 실험을 진행하였다. 기존 시 각질의응답에서 이미지와 텍스트간의 공동 어텐션을 활용하여 좋은 성능을 보였던 deep modular co-attention network (MCAN)과 유사도 기반의 선별된 캡션을 사용하여 VQA–v2 데이터에 대하여 실험을 진행하였다. 그 결과 기존의 MCAN모델과 비교하여 유사도 기반으로 선별된 캡션을 활용했을 때 성능 향상을 확인하였다.

주요용어: 시각질의 응답, 이종 데이터, 공동 어텐션, 이미지 캡셔닝, 텍스트 유사도

이 논문은 2019년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: clim@cau.ac.kr