ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 29, No. 11, pp. 501-508, 2023. 11
https://doi.org/10.5626/KTCP.2023.29.11.501

Vision-Language 모델에서의 Variational 오토인코더와 대조 학습을 이용한 멀티모달 정렬

(Vision-Language Model with Multimodal Alignment using Variational AutoEncoder and Contrastive Learning)

정 은 서 † 김 균 엽 † 강 상 우 ††
(Eunseo Jeong) (Gyunyeop Kim) (Sangwoo Kang)

요 약 기존 Vision-Language 모델 연구에서는 이미지-텍스트 데이터를 각각의 인코더를 두어 독립적으로 학습하는 연구들이 많이 제안되었다. 이러한 모델들은 각 modality를 서로 다른 인코더에서 개별적으로 처리하기 때문에, 이미지 임베딩과 텍스트 임베딩 사이의 연관성이 없고 서로 다른 임베딩 벡터 공간을 가진다는 heterogeneity gap의 문제점이 있다. 이러한 문제점을 해결하기 위한 방법으로 본 논문에서는 기존의 Transformer 기반의 사전학습 모델에 Variational 오토인코더와 대조 학습을 추가로 구조한 모델을 제안한다. 제안 방법론들을 통해 이미지 임베딩과 텍스트 임베딩이 유사한 벡터를 가지도록 모델을 학습하였다. 모델의 학습과 평가는 MSCOCO 캡셔닝 데이터셋을 사용하였고 BLEU, ROUGE-1, ROUGE-L, METEOR를 이용하여 성능을 평가하였다. 본 논문에서 제안하는 Variational 오토인코더와 대조 학습을 차용한 Transformer-VAE-CL 모델이 기존의 Transformer 기반 모델보다 BLEU가 12.1, ROUGE-1이 46.1, ROUGE-L이 44.7, METEOR가 39.9로 더 높은 성능을 보임을 확인하였다.

키워드: 이미지 캡셔닝, 멀티모달, 임베딩 정렬, variational 오토인코더, 대조 학습

Abstract Many previous vision-language model studies have been proposed to train image-text data with each encoder independently. Since these models process each modality independently on different encoders, a heterogeneity gap occurs because there is no association between image embedding and text embedding, and each has different embedding vector spaces. Thus, we propose a Transformer-VAE-CL model by adding variational autoencoder and contrastive learning structures to the existing Transformer-based pretrained model to solve this problem. We train the model using the proposed methodologies so that image embedding and text embedding have similar vector representations. The model training and evaluation use the MSCOCO captioning dataset, and performance is evaluated by BLEU, ROUGE-1, ROUGE-L, and METEOR. The Transformer-VAE-CL model using a variational autoencoder and contrastive learning proposed in this paper obtains 12.1 for BLEU, 46.1 for ROUGE-1, 44.7 for ROUGE-L and 39.9 for METEOR, higher performance than the existing Transformer-based model.

Keywords: image captioning, multimodal, alignment, variational autoencoder, contrastive learning

· 이 성과는 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2022R1A2C1005316).

† 학생회원 : 가천대학교 AI·소프트웨어학과 학생
esjeong153@naver.com
gyop0817@gachon.ac.kr
†† 정 회 원 : 가천대학교 AI·소프트웨어학과 교수(Gachon Univ.)
swkang@gachon.ac.kr
(Corresponding author임)

논문접수 : 2023년 4월 7일
(Received 7 April 2023)
논문수정 : 2023년 8월 11일
(Revised 11 August 2023)
심사완료 : 2023년 10월 5일
(Accepted 5 October 2023)

1. 서 론

최근 많은 연구에서 이미지와 텍스트 데이터를 함께 처리하는 task들이 제안되었다. 예를 들어, 이미지에 대한 자연어 설명인 캡션을 생성하는 이미지 캡셔닝(Image Captioning)이나 주어진 이미지에 대한 질문에 답변을 생성하는 시각 질의 응답(Visual Question Answering) 등과 같은 multimodal task들이 제안되었다. 이러한 multimodal task를 해결하기 위하여, 초기에는 두 modality를 각각의 인코더를 이용하여 독립적으로 처리하였다. 이후, 성능 향상을 위해 두 modality를 하나의 모델로 통합하여 처리하는 모델들이 제안되었다. 이러한 두 modality를 하나의 모델에서 처리하는 Vision-Language 모델은 시각적 추론, 시각 질의 응답, 이미지 캡셔닝 등 다양한 Vision-Language task에서 독립적으로 인코더를 구조한 모델보다 더 높은 성능을 보였다.

하지만 Vision-Language task를 해결하기 위한 모델들은 서로 다른 modality를 처리하기 때문에 다양한 문제점이 발생한다. 그 중 heterogeneity gap이란, 서로 다른 modality들이 서로 다른 임베딩 벡터 공간을 가진다는 문제점을 의미한다. 즉, 이미지-텍스트 쌍이 존재할 때, 이미지에 대한 임베딩과 텍스트에 대한 임베딩이 상이하여 두 modality의 합성이 어렵다는 문제점이다. 해당 문제점이 발생하는 이유는 이미지 모델과 언어 모델이 서로 다른 데이터를 각각의 인코더를 이용하여 독립적으로 학습하기 때문이다. 독립적으로 학습한 두 모델의 결과는 연관성이 없기 때문에 서로 다른 임베딩을 가진다. 이러한 문제점을 해결하기 위해 서로 다른 임베딩을 맞추어주기 위한 alignment 연구가 진행되었다. Alignment란, 같은 의미를 가진 서로 다른 두 modality의 임베딩이 유사한 값을 가지도록 유도하는 과정을 말한다.

Vision-Language 모델에서 alignment를 위한 방법으로는 대표적으로 대조 학습(contrastive learning)이 있다. 대조 학습이란, 두 임베딩 간의 거리를 조절하는 손실 함수를 추가하여 모델을 학습하는 방법이다. 예를 들어, 자전거를 타고 있는 소녀의 이미지와 어떤 소녀가 자전거를 타고 있다라는 캡션이 있을 때, 두 데이터의 임베딩 벡터 공간을 가깝게 학습하는 것이다. 이와 같이 대조 학습을 통해 학습할 경우, 동일한 정보를 포함한 서로 다른 modality의 임베딩은 유사한 값을 가지도록 학습한다. 해당 방법론을 사용한 대표적인 Vision-Language 모델 연구로는 CLIP[1]과 ALIGN[2] 등이 있다. [1]에서는 대조학습을 이용하여 이미지-텍스트 쌍에 대한 임베딩을 align하는 연구들이 진행되었다. 또한 [2]에서는 대조학습을 이용해 이미지-텍스트 쌍의 통합된 임베딩을 얻기 위한 연구들이 진행되기도 하였다.

본 논문에서는 서로 다른 modality의 임베딩을 alignment하여 heterogeniety gap을 해결하기 위한 방법을 제안한다. 또한 제안 방법을 통하여 이미지에 대한 캡션 생성의 성능향상을 목적으로 한다. 본 논문에서 제안하는 방법은 기존의 Encoder-Decoder 모델 구조에 추가적으로 대조 학습과 Variational 오토인코더(Variational AutoEncoder)[3]를 적용한 방법이다. Encoder-Decoder 모델은 [4]과 같이 이미지 인코더를 통해 이미지 특징 벡터를 추출하여 이미지를 이해하고, 이해한 내용을 기반으로 텍스트 디코더를 통해 이미지에 대한 캡션을 생성한다. 제안하는 방법은 이 과정에서 대조 학습을 통해 이미지와 텍스트 임베딩 사이의 alignment를 수행하는 것이다. 또한 기존의 모델에 Variational 오토인코더를 구조하여 이미지에 대한 임베딩을 생성하는 방법을 제안한다. Variational오토인코더는 이미지 정보를 정규분포를 따르는 특정 크기의 잠재 벡터로 압축하고, 이를 기반으로 캡션을 생성하는 과정에서 임베딩을 학습한다.

2. 관련 연구

2.1 Vision-Language Models

서로 다른 modality를 모두 처리하는 Vision-Language 모델에서 각 modality를 처리하는 방법에 따라 서로 다른 임베딩 벡터 공간을 가진다는 문제점이 발생한다. 이에 따라 최근까지 해당 문제점을 해결하기 위한 연구들이 활발히 진행되었다.

Vision-Language 모델은 서로 다른 임베딩 벡터를 align해주는 방법을 기준으로 모델을 다음 세 가지 카테고리로 분류할 수 있다. 세 가지 카테고리는 Dual-Encoder, Fusion-Encoder, Encoder-Decoder이다. Dual-Encoder는 각 modality에 대한 인코더를 개별적으로 두어, 서로 다른 인코더가 이미지와 텍스트를 독립적으로 학습하는 모델 구조이다. 각 modality를 독립적으로 학습한 두 인코더는 dot product와 같은 간단한 함수를 통해 이미지와 텍스트 사이의 cross-modal attention으로 임베딩을 align하여 유사한 값을 가지도록 유도한다. 대표적인 Dual-Encoder 모델은 CLIP[1]과 ALIGN[2]이 있다. Dual-Encoder 모델의 장점은 두 modality를 독립적인 인코더로 학습하기 때문에, 계산 효율이 높다는 것이다. 하지만 시각 질의 응답과 같이 이미지와 텍스트의 심층적인 정보와 복잡한 추론이 필요한 task의 경우에는, 두 modality의 합성이 어려워 성능이 하락할 수 있다는 단점이 있다. Fusion-Encoder는 각 modality에 대한 인코더를 개별적으로 두는 것이 아니라, 모든 데이터를 하

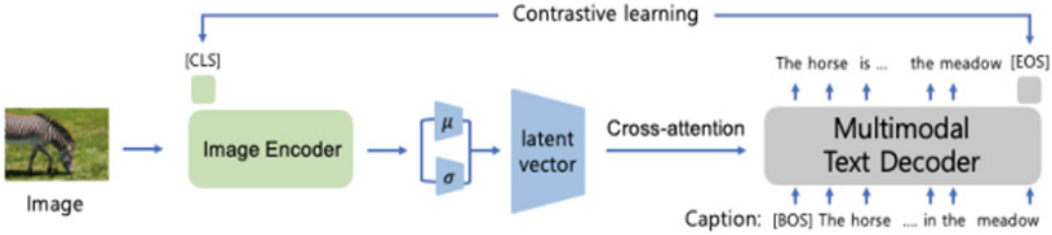

그림 1 제안 모델의 전체 구조도

Fig. 1 Overall architecture of the proposed model

나의 인코더로 한 번에 학습하는 모델 구조이다. 이미지-텍스트 쌍의 데이터를 하나의 인코더로 임베딩하여, multi-layer self-attention과 cross-attention을 수행함으로 임베딩을 align한다. Fusion-Encoder의 대표적인 모델로는 UNITER[5] OSCAR[6] 등이 있다. [5, 6]과 같이 모든 데이터를 하나의 모델로 처리하는 Fusion-Encoder 모델의 장점은 이미지-텍스트 쌍에 대한 통합적인 임베딩 정보를 학습할 수 있다는 것이다. 이와 같이 통합적인 임베딩 벡터를 학습한다면 Vision-Language task를 보다 더 높은 성능으로 수행할 수 있다. Encoder-Decoder 모델의 경우는 이미지 인코더에서 이미지를 이해하고, 텍스트 디코더에서 이미지에 대한 텍스트를 생성하는 모델의 구조이다. Encoder-Decoder 모델은 cross-attention을 통해 서로 다른 modality에 대한 임베딩 벡터를 align한다. 대표적으로 CoCa[7] 모델이 있다.

최근 Vision-Language 모델의 연구에서는 모델의 구조를 개선하는 것뿐만 아니라, 대조 학습을 통해 이미지-텍스트 쌍에 대한 임베딩을 학습한다. [1]에 의하면 대조 학습은 이미지-텍스트 쌍에서 positive 쌍에 대해서는 임베딩 벡터 공간을 가깝게 학습하고, negative 쌍에 대해서는 임베딩 벡터 공간을 멀게 학습하는 방법론이다. Dual-Encoder의 경우, 일반적으로 이미지 데이터의 [CLS] 임베딩과 텍스트 데이터의 [CLS] 임베딩 사이의 대조 학습을 통해 서로 다른 임베딩 벡터 공간을 유사하게 만들어준다. 하지만 Encoder-Decoder의 경우에는 이미지의 [CLS] 임베딩과 텍스트의 [EOS] 임베딩 사이의 대조 학습을 통해 서로 다른 임베딩 벡터를 유사하게 만들어준다. 이와 같이 Vision-Language 모델은 downstream task를 해결하기 위한 모델 구조에 맞는 대조 학습 방법론을 통해 서로 다른 이미지와 텍스트의 정보를 학습할 수 있다.

2.2 AutoEncoder & Variational AutoEncoder

생성기반의 모델은 새로운 데이터를 생성할 때, 해당 학습 데이터가 가지는 실제 분포와 같은 분포에서 샘플링 된 값으로 새로운 데이터를 생성한다. 즉, 원본 데이터와 근사한 확률분포를 가진 새로운 데이터를 생성한다. 생성 모델인 오토인코더[3]는 원본 데이터를 의미있는 더 작은 차원으로 압축하는 것을 목표로 하는 manifold learning[3]을 기반으로 한다. Manifold learning이란, 고차원의 데이터를 모두 포괄하는 저차원의 subspace가 있다고 가정하고, 고차원 데이터를 저차원의 벡터 공간으로 mapping하여 학습하는 방법이다. 이와 같이 오토인코더는 주어진 고차원의 데이터 차원을 저차원의 특정 크기 잠재 벡터로 압축하여 학습하는 방법론이다.

Variational 오토인코더[3]는 오토인코더와 동일하게, 원본 데이터와 유사한 데이터를 생성하는 것을 목표로 한다. 하지만 오토인코더와의 차별점이 존재한다. 오토인코더와 다르게 Variational 오토인코더는 주어진 데이터에서 표준 정규분포를 따르는 평균과 표준편차를 구하고, 사전확률과 유사하도록 새로운 데이터를 생성한다. 따라서 Variational 오토인코더는 정규분포를 따르는 평균과 표준편차를 이용하여 고차원의 데이터를 저차원의 특정 크기의 잠재 벡터로 압축하기 때문에, 오토인코더보다 원본 데이터를 더 잘 정돈된 잠재 벡터로 압축할 수 있다.

2.3 Transformer with Variational AutoEncoder

Transformer[8] 모델의 등장 이후, 다양한 downstream task에서 Transformer 기반 모델이 높은 성능을 보였다. 이에 따라 이미지와 텍스트 데이터 모두 Transformer 기반의 모델로 처리하는 연구들이 제안되었다. 생성 task 또한 Transformer 기반의 모델들을 통해 해결하는 연구들이 등장하였다. Transformer VAE[9], T-CVAE[10] 등과 같이 Variational 오토인코더를 Transformer 모델에 적용하는 연구들이 제안되었다.

3. Transformer-VAE-CL

이번 절에서는, 본 논문에서 제안하는 방법론을 설명한다. 3.1은 이미지 캡셔닝을 수행하기 위한 본 논문의

전체적인 프로세스를 서술한다. 3.2는 Variational 오토인코더를 Transformer에 적용한 제안 방법론에 대한 설명이다. 3.3은 이미지 캡셔닝에서 대조학습을 적용한 방법론을 서술한다.

3.1 Image Captioning

이미지 캡셔닝은 이미지가 주어졌을 때, 이미지에 대한 설명인 캡션을 자연어로 생성하는 task이다. 이미지 캡셔닝은 일반적으로 Encoder-Decoder 모델 구조로, 전체적인 프로세스는 다음과 같다. 먼저 이미지 인코더가 주어진 이미지를 이해하여 이미지 특징 정보들을 추출한다. 이를 텍스트 디코더에 전달하여, 전달받은 이미지 정보를 기반으로 이미지를 설명하는 자연어 캡션을 생성한다. 모델을 학습할 때에는, 이미지와 정답 캡션 쌍의 임베딩을 학습한다. 모델을 평가할 때에는, 이미지와 캡션 텍스트의 시작 토큰인 [EOS]를 기반으로, 이미지에 대한 캡션을 auto-regressive하게 생성하여 평가한다.

본 논문에서 제안하는 모델의 구조는 그림 1과 같다. 이미지 인코더는 주어진 이미지에서 특징 벡터들을 추출하여 이미지를 이해한다. 이미지 인코더는 Transformer 기반의 이미지 인코더 모델이다. 해당 모델에는 patch 단위로 자른 이미지와 [CLS] 토큰이 입력된다. 이후 이미지 인코더는 각 patch와 토큰에 해당하는 임베딩을 반환한다. [CLS] 토큰에 대한 임베딩은 i_{CLS}로 표기하며, x번째 patch에 대한 임베딩은 i_x로 표기한다. 해당 임베딩을 통해 이미지 임베딩을 구축한다. 이후 이미지 임베딩을 feed-forward network에 입력하여 특정 크기의 잠재 벡터로 압축한다. 이렇게 추출한 이미지 잠재 벡터를 이용하여 텍스트 디코더에서 캡션을 생성한다. 텍스트 디코더는 전달받은 이미지 특징 벡터 임베딩과 정답 캡션 사이의 cross-attention을 통해 캡션을 생성한다. 해당 과정에 대한 식은 다음과 같다.

$$I = \begin{cases} W_{im} * i_{CLS} + b \\ W_{im} * [i_{CLS}; i_1; i_2; \cdots; i_{p-1}; i_p] + b \end{cases} \quad (1)$$

$$L_{lm} = -\sum_{k=1}^{N} \log p(C_k | C_1, \cdots, C_{k-1}, I) \quad (2)$$

식 (1)은 이미지 잠재 벡터를 구하는 식이다. 해당 식은 이미지에 대한 [CLS] 혹은 전체 patch 임베딩 벡터를 이용하여 잠재 벡터를 추출한다. 식 (1)에서 p는 patch 개수를 의미한다. W_{im}과 b는 잠재 벡터를 만들기 위한 feed-forward network의 학습 가능한 파라미터이다. 이미지 잠재 벡터는 식 (1)과 같이 두 가지 방법으로 구성하여 각각 실험을 진행하였다. 첫 번째는 [CLS] 임베딩 벡터만을 이미지 임베딩으로 사용한 경우이다. 두 번째는 [CLS] 임베딩을 포함한 모든 patch 임베딩을 이미지 임베딩으로 사용한 경우이다. 식 (2)는 압축한 이미지 잠재 벡터를 이용하여 이미지를 설명하는 올바른 캡션을 auto-regressive하게 생성하기 위한 언어 모델 손실 함수이다. 식 (2)에서 C_x는 캡션의 x번째 토큰을 의미한다. n은 캡션에서의 단어의 갯수이다. 식 (2)는 현재 시점까지의 캡션 토큰과 이미지 잠재 벡터가 함께 입력으로 들어왔을 때, 그 다음 토큰을 생성하기 위한 손실 함수로 구성하였다.

3.2 Transformer with Variational AutoEncoder

본 논문에서 제안하는 방법론은 인코더의 아웃풋을 그대로 디코더에서 사용하는 기존 모델과 다르게 Transformer 기반의 사전학습 모델에 Variational 오토인코더를 적용한 방법이다. 제안하는 방법론을 통하여 인코더에서 얻은 정보들을 Variational 오토인코더로 효과적으로 압축하고, 이를 디코더에 전달하는 방법을 제안한다. 본 논문에서는 Encoder-Decoder 모델에서 주어진 고차원의 이미지 데이터를 특정 크기의 잠재 벡터로 압축할 때, Variational 오토인코더를 통해 이미지의 정보를 학습한다. 압축한 이미지 임베딩을 이용하여 텍스트 디코더가 이미지에 대한 올바른 캡션을 생성하기 위해서는, 이미지 인코더가 이를 효과적으로 압축하여 텍스트 디코더에 전달하는 것이 중요하다. Variational 오토인코더를 통해 잠재 벡터를 압축한다면 학습 이미지와 유사한 확률분포를 포착할 수 있다. 즉, 압축한 잠재 의미 벡터를 통해 생성모델은 학습 데이터의 확률분포과 근사한 특성을 가진 데이터를 생성할 수 있다.

Variational 오토인코더에서 잠재 의미 벡터를 만드는 방법은 다음과 같다. 주어진 이미지에 대하여 표준 정규분포를 따르는 평균과 표준편차를 얻는다. 얻은 평균과 표준편차를 정규분포로부터 샘플링한 epsilon 값과 더하고 곱한다. 이를 통해 이미지 확률분포에서 얻은 평균과 표준편차가, 평균이 0이고 표준편차가 1인 정규분포를 따르도록 잠재 벡터를 만든다. 이를 reparameterization trick이라고 한다. Reparameterization trick은 이미지에서 얻은 확률분포로 특정한 크기의 잠재 벡터를 생성할 때, 항상 같은 확률분포를 가지는 것을 방지하기 위한 목적이다. 해당 학습을 위한 손실 함수는 아래의 식과 같다.

$$L_{KL} = D_{KL}[q(I \mid D) \mid\mid p(D)] \quad (3)$$

위와 같이, 이미지에서 표준 정규분포를 따르는 평균과 표준편차를 통해 이미지 임베딩을 잠재 벡터로 압축하면, 보다 더 정돈된 확률분포의 잠재 벡터를 얻을 수 있다. 이는 사전분포와 사후분포의 차이가 가까워지도록, 식 (3)와 같이 두 확률분포 간의 차이를 구하는 KL

Divergence loss를 통해 이미지 인코더를 학습하기 때문이다. 식 (3)에서 D는 입력 데이터의 이미지이며, I는 압축된 잠재 벡터이다.

3.3 Contrastive Learning

본 논문에서는 대조 학습을 이용하여 multimodal 학습의 문제점인 임베딩 벡터의 heterogeneity gap을 해결한다. 대조 학습은 학습된 임베딩 벡터 공간 상에서 positive 쌍은 가깝도록, negative 쌍은 멀도록 학습하는 방법이다. 이러한 학습을 통해 서로 다른 임베딩 벡터 공간을 가지는 두 modality의 임베딩을 유사하도록 align한다. Vision-Language모델에서의 대조 학습은 한 batch 내에서 하나의 정답 이미지-텍스트 쌍을 positive로, 나머지 쌍을 negative로 간주하여 학습한다. 일반적인 Vision-Language 모델에서의 대조 학습은 Dual-Encoder로 수행한다. Dual-Encoder 구조에서는 이미지 인코더의 [CLS]와 텍스트 인코더의 [CLS] 사이의 대조 학습을 통해 모델을 학습한다. 본 논문에서 제안하는 방법은 Encoder-Decoder 모델 구조에서 이미지와 텍스트의 서로 다른 임베딩 벡터를 align해주기 위한 방법이다. 본 논문에서는 한 batch 내의 positive 쌍의 이미지 [CLS]와 텍스트 [EOS]를 가깝도록 유도하고, negative 쌍의 이미지 [CLS]와 텍스트 [EOS]를 멀게 유도하는 손실 함수를 사용한다. 이미지 [CLS] 임베딩은 이미지의 전체적인 정보가 함축되어 있고, 텍스트 [EOS] 임베딩은 텍스트의 전체적인 정보가 함축되어 있기 때문에 해당 방법으로 대조 학습을 진행한다. 해당 손실함수는 아래의 식과 같다.

$$L_{con} = -\frac{1}{N}\Big(\sum_{i}^{N}\log\frac{\exp(I_{CLS}^{i^T}C_{EOS}^{i}/\sigma)}{\sum_{j=1}^{N}\exp(I_{CLS}^{i^T}C_{EOS}^{j}/\sigma)} + \sum_{i}^{N}\log\frac{\exp(C_{EOS}^{i^T}I_{CLS}^{i}/\sigma)}{\sum_{j=1}^{N}\exp(C_{EOS}^{i^T}I_{CLS}^{j}/\sigma)}\Big) \quad (4)$$

식 (4)은 image-to-text, text-to-image의 대조학습 손실함수를 의미한다. N은 batch의 크기를 나타낸다. I_{CLS}^{i}, I_{CLS}^{j}는 i, j번째 이미지의 [CLS] 토큰 임베딩을 의미한다. C_{EOS}^{i}, C_{EOS}^{j}는 i, j번째 캡션의 [EOS] 토큰 임베딩을 의미한다. 따라서 I_{CLS}^{i}와 C_{EOS}^{i}는 positive 쌍을 의미하고, I_{CLS}^{i}와 C_{EOS}^{j}는 negative 쌍을 의미한다. Positive 쌍의 개수는 N개이며, negative 쌍의 개수는 N^2-N개다. X^T는 행렬 X의 transpose를 의미한다. σ는 학습 가능한 파라미터인 temperature를 의미한다. 모델 학습에서의 최종 손실 함수는 식 (5)와 같다.

$$L = \alpha * L_{lm} + \beta * L_{KL} + \gamma * L_{con} \quad (5)$$

L_{lm}은 식 (2)에서의 언어 모델 손실 함수를 의미한다. L_{KL}은 식 (3)에서의 KL Divergence 손실 함수를 의미한다. L_{con}은 식 (4)에서의 대조 학습의 손실 함수를 의미한다. α, β, γ는 각 손실 함수에 가중치를 부여하기 위한 상수이다.

4. 실 험

이번 절에서는, 본 논문에서는 제안 모델에 대한 실험 결과를 비교 및 분석한다. 본 논문에서 해결하고자 하는 downstream task는 이미지 캡셔닝으로, 이미지에 대한 캡션을 생성하는 task이다. 4.1은 실험에서 사용한 데이터셋과 실험 환경에 대한 설명이다. 4.2는 제안 모델인 Variational 오토인코더를 Tranformer 기반 모델에 적용할 때, 잠재 벡터 크기 설정에 따라 달라지는 성능 결과를 분석한다. 4.3은 제안 모델인 Variational 오토인코더와 대조 학습을 Transformer 기반 모델에 모두 적용했을 때의 성능 결과를 분석한다. 4.4는 기존의 alignment를 수행하는 모델과 제안하는 모델의 성능 결과를 비교한다. 모든 실험은 PyTorch 프레임워크를 사용하여 진행하였다.

4.1 데이터셋 및 실험 환경

본 논문에서 fine-tuning에 사용한 데이터셋은 MSCOCO_captioning 2014[11]으로, 이미지 캡셔닝에서 일반적으로 사용하는 데이터셋이다. 해당 데이터셋은 이미지-캡션 쌍으로 구성된다. 학습용 데이터로 50,000개, 평가용 데이터로 10,000개, 테스트용 데이터로 10,000개의 이미지-캡션 쌍을 사용하여 모델을 fine-tuning하였다.

본 논문에서 제안하는 모델의 구조는 Encoder-Decoder 기반으로, 이미지 인코더는 Vision Transformer, 텍스트 디코더는 GPT2-base[12] 모델을 사용하여 실험하였다. 모델은 batch size를 48로 하여, 30 epoch만큼 학습하였다. Optimizer는 Adam optimization으로, learning rate 값은 1e-4 혹은 1e-5로 실험하여 둘 중 높은 값을 결과로 사용하였다. 각 실험에서 사용한 learning rate는 표에 명시하였다. Cosine Annealing LR scheduler을 통해 learning rate를 조절하였다. 이미지의 크기는 모두 224*224로 transform하여 실험하였다. σ는 이전 연구[2, 7]에 따라 초기값은 0.07으로 설정하였다. 손실 함수의 α, β, γ는 각각 0.7, 0.1, 0.2 혹은 0.6, 0.1, 0.3으로 설정하여 실험을 수행하였다. α, β, γ는 0.7, 0.1, 0.2로 설정하였을 때, 더 높은 성능을 보였다. 따라서 모든 실험의 결과는 더 높은 성능을 보인 hyperparameter 값으로 설정한 실험 결과를 기재하였다. 모델의 성능은

표 1 제안 방법론에 대한 성능 평가표
Table 1 Performance table for the proposed methodology

	learning rate	BLEU	ROUGE-1	ROUGE-L	METEOR
Encoder-Decoder (baseline)	le-5	11.3	45.6	43.8	39.1
Baseline+VAE	le-5	11.0	45.1	43.3	38.6
Baseline+VAE+CL	le-5	**12.1**	**46.4**	**44.7**	**39.9**

표 2 제안 방법론에서 latent vector type과 latent vector size에 따른 성능 평가표
Table 2 Performance table for the proposed methodology according to latent vector type and size

	latent vector type	learning rate	latent vector size	BLEU	ROUGE-1	ROUGE-L	METEOR
Baseline+ VAE	CLS embedding	le-4	64	9.33	42.0	40.2	35.3
			128	8.99	42.1	40.4	35.4
			768	9.05	41.6	39.7	35.0
	patch embedding	le-5	64	10.6	44.4	42.7	37.7
			128	10.5	44.3	42.5	37.4
			768	**11.0**	**45.1**	**43.3**	**38.6**

표 3 제안 방법론에서 대조 학습을 적용함에 따른 성능 평가표
Table 3 Performance table for the proposed methodology with contrastive learning

	latent vector type	learning rate	latent vector size	BLEU	ROUGE-1	ROUGE-L	METEOR
Baseline+ VAE+CL	CLS embedding	le-4	64	8.56	41.0	39.1	34.5
			128	8.94	42.5	40.6	35.6
			768	8.86	42.3	40.5	35.5
	patch embedding	le-5	64	11.1	45.3	43.5	38.7
			128	**12.1**	**46.4**	**44.7**	**39.9**
			768	11.9	46.4	44.5	39.9

BLEU(Bilingual Evaluation Understudy)[13], ROUGE-1(Recall-Oriented Understudy for Gisting Evaluation)[14], ROUGE-L[14], METEOR(Metric for Evaluation of Translation with Explicit Ordering)[15]를 통해 측정하여 평가하였다.

4.2 실험결과

4.2.1 제안방법론에 대한 성능 평가표

Baseline은 Encoder-Decoder 기반의 모델로, Vision Transformer와 GPT2 사전학습 모델 구조이다. Baseline+VAE는 제안 모델으로, 기존 baseline에 Variational AutoEncoder를 추가한 모델이다. 표준 정규분포를 따르는 평균과 표준편차를 이용하여 특정 크기의 잠재 벡터로 압축하여 이미지 캡셔닝 task를 수행하였다. Baseline+VAE+CL은 제안 모델으로, 기존 baseline에 Variational AutoEncoder와 Contrastive Learning을 추가한 모델이다. 해당 실험에 대한 결과는 표 1과 같다.

Baseline인 Encoder-Decoder 모델은 BLEU가 11.3, ROUGE-1이 45.6, ROUGE-L이 43.8, METEOR가 39.1의 성능을 보였다. Baseline에 VAE를 추가한 모델은 BLEU가 11.0, ROUGE-1이 45.1, ROUGE-L이 43.3, METEOR가 38.6의 성능을 보였다. Baseline 모델보다 적은 폭의 성능 저하가 나타났다. 본 논문에서 제안하는 모델인 Baseline+VAE+CL은 BLEU가 12.1, ROUGE-1이 46.4, ROUGE-L이 44.7, METEOR가 39.9의 성능을 보였다. 이는 baseline 모델과 baseline 모델에 VAE를 추가한 제안 모델의 성능 결과를 비교했을 때, 적은 폭의 성능 향상을 보인다.

4.2.2 Latent vector type과 size에 따른 성능 평가표

해당 실험 결과는 제안하는 모델에서 VAE를 이용하여 잠재 벡터를 생성할 때, type과 size를 다르게 실험한 결과이다. 잠재 벡터 type은 식 (1)에서의 각 잠재 벡터 생성 방법에 따른 분류이다. 식 (1)과 같이 잠재

표 4 기존의 alignment 수행 모델과 제안 방법론의 성능 평가표
Table 4 Performance table for the Dual-Encoder, Encoder-Decoder and proposed methodology with contrastive learning

	learning rate	BLEU	ROUGE-1	ROUGE-L	METEOR
Dual-Encoder	le-5	**32.7**	46.2	44.1	**63.3**
Encoder-Decoder (baseline)	le-5	11.3	45.6	43.8	39.1
Baseline+VAE+CL	le-5	12.1	**46.4**	**44.7**	39.9

벡터를 이미지 인코더의 [CLS] 임베딩을 사용하여 생성하는지 혹은 전체 임베딩 벡터를 사용하여 생성하는지에 따라, CLS embedding과 patch embedding으로 나뉜다. 잠재 벡터 size는 주어진 고차원의 이미지를 어떤 크기의 잠재 벡터로 압축할지에 따라 64, 128, 768으로 나뉜다.

해당 실험에 대한 결과는 표 2와 같다. CLS 임베딩을 사용한 경우, 64차원의 잠재 벡터를 생성할 때의 성능이 BLEU가 9.33, ROUGE-1이 42.0, ROUGE-L이 40.2, METEOR가 35.3으로 가장 높은 성능을 보였다. Patch 임베딩을 사용한 경우, 768차원의 잠재 벡터를 생성할 때의 성능이 BLEU가 11.0, ROUGE-1이 45.1, ROUGE-L이 43.3, METEOR가 38.6으로 가장 높은 성능을 보였다. 따라서 patch 임베딩을 사용하여 잠재 벡터를 생성하는 경우에는 잠재 벡터의 크기를 768차원으로 설정하고 생성하는 것이 가장 좋은 성능을 보였다. 해당 실험의 결과를 통해 CLS 임베딩으로 잠재 벡터를 생성하는 것보다 patch 임베딩으로 잠재 벡터를 생성하는 것이 더 높은 성능의 결과를 보임을 확인하였다.

4.3 대조학습에 대한 실험 결과

표 3은 제안 모델에 대조 학습을 적용하여 이미지 캡셔닝을 수행한 실험 결과표이다. 해당 실험은 Encoder-Decoder 모델 구조에서, 이미지 인코더의 [CLS]와 텍스트 디코더의 [EOS] 사이의 대조 학습과 VAE를 모두 이용하여 이미지 캡셔닝을 수행하였다. 본 실험에서도 잠재 벡터 type과 size를 다르게 설정하여 결과를 확인하였다.

해당 실험에 대한 결과는 표 3과 같다. CLS 임베딩을 사용한 경우에는 128차원의 잠재 벡터를 생성할 때 BLEU가 8.94, ROUGE-1이 42.5, ROUGE-L이 40.6, METEOR가 35.6으로 가장 높은 결과를 보였다. Patch 임베딩을 사용한 경우에도 128차원의 잠재 벡터를 생성할 때 BLEU가 12.1, ROUGE-1이 46.4, ROUGE-L이 44.7, METEOR가 39.9으로 가장 높은 성능을 보였다. 해당 실험의 결과를 통해 제안 모델에 대조 학습을 적용하여 이미지와 텍스트 임베딩을 유사하게 만드는 것이 대조 학습을 적용하지 않은 모델보다 높은 성능을 보임을 확인하였다. 또한 CLS 임베딩으로 잠재 벡터를 생성하는 것보다 patch 임베딩으로 잠재 벡터를 생성하는 것이 더 높은 성능의 결과를 보임을 확인하였다.

4.4 기존 연구와 제안 방법에 대한 실험 비교

표 4는 기존의 Vision-Language 모델에서 alignment를 수행하는 모델인 Dual-Encoder, Encoder-Decoder의 성능과 제안 모델의 성능을 비교한 실험 결과표이다. Dual-Encoder 모델은 Vision Transformer와 BERT 사전학습 모델을 사용하여 실험하였다. Encoder-Decoder 모델은 Vision Transformer와 GPT2 사전학습 모델을 사용하여 실험하였다. 제안 모델인 Baseline+VAE+CL은 Encoder-Decoder 모델에 Variational AutoEncoder와 Contrastive Learning을 추가한 모델이다. 해당 실험에서 learning rate는 높은 성능을 도출하는 le-5로 지정하여 실험하였다. 해당 실험에 대한 결과는 표 4와 같다. 두 modality를 각 인코더를 두어 실험한 Dual-Encoder에서는 BLEU가 32.7, ROUGE-1이 46.2, ROUGE-L이 44.1, METEOR가 63.3으로, BLEU와 METEOR에서 제안 방법보다 큰 폭으로 높은 성능을 보였다. 하지만 모델의 생성 성능을 평가하는 ROUGE-1, ROUGE-L에서는 제안 방법이 Dual-Encoder 모델보다 높은 성능을 보였다. 해당 실험의 결과를 통해 제안 모델에서의 성능 결과가 서로 다른 modality를 독립된 두 개의 인코더로 처리하는 Dual-Encoder 모델과 비교했을 때 유의미한 결과를 보이는 것을 확인하였다.

5. 결 론

본 논문에서는 주어진 이미지에 대한 자연어 캡션을 생성하는 이미지 캡셔닝 task에서 이미지-텍스트 쌍에 대한 각각의 임베딩이 상이하다는 heterogeneity gap 문제점을 해결하기 위한 방법론을 제안한다. 제안하는 방법론은 Transformer 기반 모델에 modality 간의 임베딩 격차를 줄일 수 있는 방법인 Variational 오토인코더와 대조 학습을 적용한 'Transformer-VAE-CL' 모

델이다. ‘Transformer-VAE-CL은 고차원의 이미지를 특정 크기의 잠재 벡터로 압축하는 Variational 오토인코더와 이미지의 [CLS] 임베딩 벡터와 텍스트의 [EOS] 임베딩 벡터 사이의 대조 학습을 통해 모델을 학습하였다. 모델의 학습과 평가는 MSCOCO 캡셔닝 데이터셋을 사용하여 수행하였다. Variational 오토인코더만을 사용한 경우에서는 baseline보다 적은 폭의 성능 하락이 발생하였지만, 대조 학습을 추가로 사용한 모델에서는 baseline보다 높은 성능을 보였다. 해당 실험을 통해 제안하는 방법론 중 Variational 오토인코더와 대조 학습을 모두 사용하는 것이 성능 개선을 보일 수 있음을 확인하였다.

References

[1] Radford, Alec, et al., "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning*. PMLR, pp. 8748-8763, 2021.

[2] Jia, Chao, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," *International Conference on Machine Learning*. PMLR, pp. 4904-4916, 2021.

[3] Kingma, Diederik P., and Max Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[4] Li, Chenliang, et al., "mplug: Effective and efficient vision-language learning by cross-modal skip-connections," arXiv preprint arXiv:2205.12005, 2022.

[5] Chen, Yen-Chun, et al., "UNITER: UNiversal Image-Text Representation Learning," *European Conference on Computer Vision*, pp. 104-120, 2020.

[6] Li, Xiujun, et al., "Oscar: Object-semantics aligned pre- training for vision-language tasks," *European Conference on Computer Vision,* pp. 121-137, 2020.

[7] Yu, Jiahui, et al., "Coca: Contrastive captioners are image-text foundation models," arXiv preprint arXiv:2205.01917, 2022.

[8] Vaswani, Ashish, et al., "Attention is all you need," *Proc. of the 31th International Conference on Neural Information Processing Systems,* pp. 6000-6010, 2017.

[9] Jiang, Junyan, et al., "Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning," *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 516-520, 2020.

[10] Wang, Tianming, and Xiaojun Wan, "T-CVAE: Transformer-Based Conditioned Variational Autoencoder for Story Completion," *International Joint Conference on Artificial Intelligence,* pp. 5233-5239, 2019.

[11] Lin, Tsung-Yi, et al., "Microsoft coco: Common objects in context," *European Conference on Computer Vision,* pp. 740-755, 2014.

[12] OpenAI [Online] Available: https://github.com/openai/gpt-2.

[13] Papineni, Kishore, et al., "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* pp. 311-318, 2002.

[14] Lin, Chin-Yew, "Rouge: A package for automatic evaluation of summaries," *Association for Computational Linguistics,* pp. 74-81, 2004.

[15] Banerjee, Satanjeev, and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization,* pp. 65-72, 2005.

정 은 서

2022년 가천대학교 소프트웨어학과 졸업(학사). 2022년~현재 가천대학교 AI·소프트웨어학부 석사과정. 관심분야는 자연어처리(Natural Language Processing), 멀티모달(Multimodal), 이미지 캡셔닝(Image Captioning), 시각 질의 응답(Visual Question Answering) 등

김 균 엽

2020년 가천대학교 소프트웨어학과 졸업(학사). 2022년 가천대학교 AI·소프트웨어학부 졸업(석사). 2022년~현재 가천대학교 AI·소프트웨어학부 박사과정. 관심분야는 자연어처리, XAI, 대화시스템 등

강 상 우

2012년 서강대학교 컴퓨터공학과 졸업(박사). 2012년~2016년 서강대학교 연구교수. 2016년~현재 가천대학교 AI·소프트웨어학부 부교수. 관심분야는 자연어처리(Natural Language Processing), 음성 대화 처리(Speech Dialogue Processing), 정보 검색(Information Retrieval), 기계독해(Machine Reading Comprehension), 기계번역(Machine Translation) 등