



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

석사학위 청구논문

지도교수 김원준

GAN을 이용한 텍스트로부터 이미지 변환 방법에 관한 연구

CycleGAN과 BERT 임베딩을 활용한
텍스트로부터 이미지 변환 방법 개선

2021년 8월

건국대학교 정보통신대학원

융합정보기술학과

이정욱

GAN을 이용한 텍스트로부터 이미지 변환 방법에 관한 연구

CycleGAN과 BERT 임베딩을 활용한
텍스트로부터 이미지 변환 방법 개선

Text to Image synthesis using GAN model
Text to Image synthesis with CycleGAN and
BERT embeddings

이 논문을 공학 석사학위 청구논문으로 제출합니다

2021년 5월

건국대학교 정보통신대학원

융합정보기술학과

이정욱

이정욱의 공학 석사학위 청구논문을 인준함

심사위원장 _____ (인)

심사위원 _____ (인)

심사위원 _____ (인)

2021년 5월

건국대학교 정보통신대학원

목 차

표목차	ii
그림목차	iii
ABSTRACT	iv
제1장 서론.....	1
제2장 관련연구.....	3
제3장 제안하는 방법	10
1. BERT 임베딩 모델 생성.....	10
2. 이미지-텍스트 유사도 검증	14
3. 이미지 생성 품질 검증	16
제4장 실험 및 평가	18
1. 이미지 생성 비교	18
2. DAMSM Loss 측정	22
3. Inception Score 측정	24
제5장 결론.....	26
참고문헌.....	27
국문초록.....	31

표 목 차

<표 3-1> BERT 임베딩 학습 파라미터	13
<표 4-1> 모델별 이미지 생성 비교	18
<표 4-2> 모델별 DAMSM 손실 측정	22
<표 4-3> Inception score 측정	24
<표 4-4> FID 측정	24

그림 목 차

<그림 2-1> StackGAN 도식화	4
<그림 2-2> Global Attention 도식화	5
<그림 2-3> AttnGAN 모델 도식화	5
<그림 2-4> Cycle Text-to-Image GAN 도식화	6
<그림 3-1> BERT 임베딩 입력 예시	12
<그림 3-2> BERT 인코더 모델 구조 도식화	12
<그림 3-3> DAMSM 모듈 도식화	15
<그림 4-1> 학습주기별 DAMSM 손실값 변화율	23
<그림 4-2> CycleGAN w/ BERT 이미지 생성 예시	25
<그림 4-3> CycleGAN w/ local BERT 이미지 생성 예시	25

ABSTRACT

Text to image synthesis using GAN model Text to Image synthesis with CycleGAN and BERT embeddings

Lee Jeong wook

Department of Convergence Information Technology
Graduate School of Information and Telecommunications
Konkuk University

Joint embeddings that leverage the relevance between text and image is used as input for conditions in generating image from text using GAN models.

In AttnGAN model, they applied attention-based embeddings which pay attention to relevant words.

In Cycle Text-to-image GAN model, they applied BERT Embeddings which consist of self-attention transformer architecture, and furthermore, they showed advanced quality through cyclic training method of CycleGAN.

In this paper, we propose that we apply domain-specific BERT embeddings in using Cycle Text-to-image GAN model, which is composed of specific words instead of general vocabulary to learn similarities more intensively between text and image.

We showed that our proposed model demonstrates more improved quality in generating image samples from text descriptions.

Keyword : GAN, BERT, image synthesis, image generation

제1장 서론

이미지 생성모델은 GAN Text-to-Image¹⁾ 이후로, StackGAN²⁾, AttnGAN³⁾, Cycle Text-to-Image GAN⁴⁾ 으로 발전되고 있다.

GAN Text-to-Image 에서의 한계로 지적된 해상도 문제를 해결하기 위해 StackGAN은 두 단계로 분리하여 진행하였고, 첫 번째 단계에서는 64x64의 낮은 해상도의 이미지를 생성하고, 이것을 입력으로 해서 두 번째 단계에서는 256x256의 높은 해상도를 생성하는 형태로 진전을 보였다.

StackGAN에서는 제약조건으로서 텍스트-이미지 결합 임베딩 방식을 사용하고, 텍스트 인코딩 방식은 RNN, 이미지 인코딩 방식은 CNN을 사용하였다.

이후로 텍스트-이미지 결합 임베딩 방식에서 밀접한 연관성의 향상을 위해, AttnGAN에서는 텍스트 인코딩시 서로 연관성이 있는 단어에 집중하는 Attention 방식을 활용했다.

자연어처리 분야에서도 Attention을 방식을 활용한 모델들이 연구되었고, Transformer⁵⁾, BERT⁶⁾ 로의 발전을 보였다.

Cycle Text-to-Image GAN에서는 이러한 Attention 방식을 활용한 텍스트 인코딩에 있어서, Self-Attention 방식인 Transformer를 이용한 BERT 임베딩 모델을 사용하여, AttnGAN 보다 더 향상된 품질을 보여줬다.

BERT 임베딩모델은 자연어처리 분야에서 기계번역, 독해, 개체명 인식 등에 뛰어난 성능을 보이므로, 국내에서도 도입하여 많이 활용하고 있다.

또한, Google에서 제공하는 사전훈련된 다국어 모델도 있지만, 국내에서는 한국어 위키백과나 뉴스 등을 학습하여 한국어로 모델을 만들어서 사용한다.

특히, BERT 임베딩 모델은 챗봇상담에도 활용되고 있는데, 특정 분야의 전문성을 강화하기 위해 해당 분야에 특화된 지식과 용어를 학습하여 임베딩모델을 만들어 활용하기도 한다.

1) S. Reed, "Generative adversarial text to image synthesis", arXiv/1605.05396., 2016.

2) H. Zhang, "Stackgan", arXiv:1612.03242., 2017.

3) Tao Xu, "AttnGAN", arXiv:1711.10485, 2017.

4) Trevor Tsue, "Cycle Text-To-Image GAN with BERT", arXiv:2003.12137, 2020.

5) Ashish Vaswani, "Attention is all you need", arXiv:1706.03762, 2017.

6) Jacob Devlin, "BERT", arXiv:1810.04805, 2018.

예를 들면, 금융분야 전문적 상담을 위해 금융 특화된 지식이나 용어 등을 학습하여 금융상담 챗봇에 이용한다.

이와 같이, BERT 임베딩을 특정 분야에 활용하기 위해, Google에서 제공하는 일반적인 임베딩을 사용하기 보다 특화된 분야에 해당하는 전문적인 임베딩 모델을 만들고자 하는 시도가 이루어지고 있다.

Cycle Text-to-Image GAN에서는 BERT 임베딩을 사용하였지만, Google에서 제공하는 사전훈련된 모델을 사용함으로 인해 특정 분야의 텍스트에 이미지를 생성하는 부분에 있어서 섬세하고 구체적인 묘사가 부족할 수 있다.

따라서 본 연구에서는 특정분야에 특화된 텍스트 임베딩을 사용함으로서 보다 더 전문적이고 자세한 이미지 생성이 가능하도록 모델을 만드는데 초점을 두었다.

즉, 본 연구에서는 Cycle Text-to-Image GAN 에서 BERT 임베딩 방식을 사용함에 있어서, 일반 어휘사전을 사용하지 않고, 특정 분야의 어휘를 사용하여 BERT 임베딩을 구성함으로써, 텍스트-이미지 결합간에 연관성과 어휘 집중도를 강화하여 학습함으로써 이미지 합성 품질이 향상되는 것을 실험을 통해 확인해 보고자 한다.

제2장 관련연구

GAN 모델⁷⁾은 적대적 신경망으로서 생성자와 판별자가 서로 보완하면서 적대적으로 모델을 개선하는 방식이다.

생성자는 이미지를 더 진짜로 만들기 위해 노력하고, 판별자는 생성자가 만든 이미지를 가짜로 판단하려고 노력한다.

GAN의 손실율의 최적화에 대한 계산은 아래 식으로 표현된다.

$$\min \max V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

생성자 입장에서는 $D(G(z))=1$ 에 가깝게 하여 $\log(1 - D(G(z)))$ 를 최소화하려고 하고, 판별자 입장에서는 $D(G(z))=0$ 에 가깝게 하여 $\log(1 - D(G(z)))$ 를 최대화하려고 한다.

따라서, 서로 상반되지만 상호 보완하는 방식으로 파라미터를 최적화 한다.

Conditional GAN⁸⁾은 생성자 방식에 있어서, 임의의 잠재공간(z)을 입력으로 시작하는 것이 아니라, 어떤 제약조건(y)을 통해 조건부로 생성하는 방식을 말한다.

라벨 등 어떤 제약조건을 통해 생성하므로 더 학습이 잘되며, 특히 MNIST 등 숫자이미지 분류모델에 있어서도 빠른 정확도를 보인다. Conditional GAN의 수식은 아래와 같다.

$$\min \max V(D, G) = E_{x \sim p_{data}(x)} [\log D(x|y)] + E_{z \sim p_z(z)} [\log(1 - D(G(z|y)))]$$

GAN Text-to-Image는 텍스트로부터 이미지를 생성하는 방식에 있어서, 텍스트와 이미지를 결합하는 방식을 사용하였다.

이미지와 텍스트를 동일한 차원의 공간에 놓고 일대일 매칭하는 방식으로 접근하였다. 하지만, 이미지의 정확한 정보를 표현하는데 한계가 있어서 64x64의 낮은 해상도 수준에서만 가능했다.

StackGAN에서는 이를 극복하기 위해 두 단계로 나누어서 이미지를 생성하는 방식을 고안하였다.

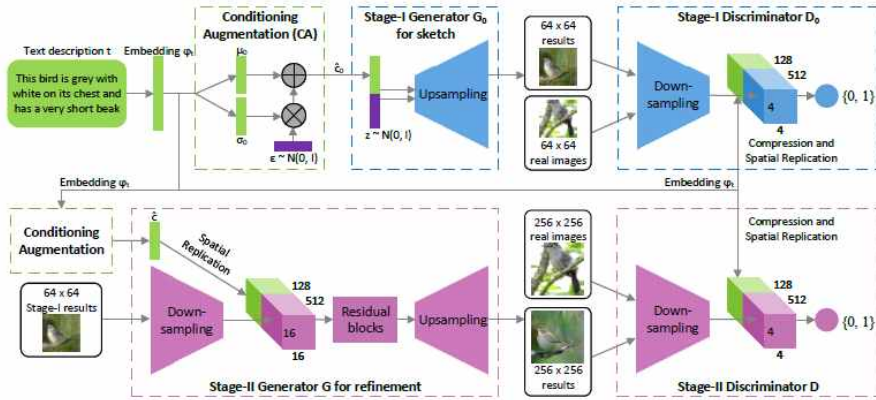
첫 번째 단계에서는 64x64의 낮은 해상도의 이미지를 생성하고, 이것을 입력으로 해서 두 번째 단계에서는 256x256의 높은 해상도의 이미지를 생성하였다.

7) I. Goodfellow, "Generative adversarial nets", NIPS, 2014.

8) M. Mirza, "Conditional generative adversarial nets", arXiv:1411.1784, 2014.

첫 번째 단계에서는 낮은 해상도의 다소 부정확한 대략적인 이미지가 생성되었지만, 두 번째 단계에서는 고화질의 좀 더 정확하고 구체적인 이미지가 생성되었다.

StackGAN의 도식화된 예시는 [그림2-1]과 같다.



[그림 2-1] StackGAN 도식화

이와 같이 StackGAN의 두 단계를 이용한 접근 방식은 Conditional GAN의 제약조건의 원리를 활용한 방식이라고 할 수 있다.

StackGAN++⁹⁾에서는 StackGAN의 두 단계 모델에 국한하지 않고 생성자를 여러 단계로 나누어서 단계별로 해상도를 확장해가는 형태로 개선하였다.

StackGAN++ 이후로 텍스트와 이미지의 밀접한 결합을 위해 텍스트 임베딩을 잘 활용하려는 연구가 진행되었는데, AttnGAN은 텍스트 임베딩에 Attention 방식을 적용하였다.

Attention 방식 인코딩이란, 문장을 단어 집합에서 단어 기준으로 인덱싱하여 단순한 배열로 인코딩하는 것이 아니라, 문장 내의 단어들간의 의미적 관계를 표현하는 방식이다.

Attention 방식 중 Global Attention 방식¹⁰⁾은 문장 내에서 단어들간의 가중치(a_t)를 구하고, 입력 단어에 각각 가중치를 가중합하여 문맥벡터(c_t)를 구하는 방식이다.

가중치(a_t)를 구하는 방식을 수식으로 나타내면 다음과 같다.

9) H. Zhang, "Stackgan++", arXiv:1710.10916, 2017.

10) Thang Luong, "Effective Approaches to Attention-based Neural Machine Translation", D15-1166, 2015.

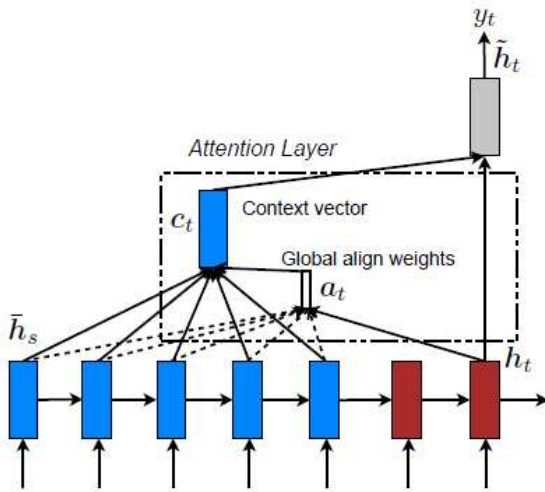
$$a_t(s) = \text{align}(h_t, \bar{h}_s) = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_s \exp(\text{score}(h_t, \bar{h}_{s'}))}, \text{ where } \text{score}(h_t, \bar{h}_s) = h_t^T \bar{h}_s$$

여기에서, h_t 는 목적 은닉벡터이고, h_s 는 소스 은닉벡터이다.

즉, 소스벡터 전체에 대해 가중치를 구하며, 최종 Attention 벡터는 아래와 같이 표현된다.

$$\tilde{h}_t = \tanh(W_c[c_t; h_t])$$

이러한 Global attention model 에 대한 도식화 예시는 [그림2-2]와 같다.

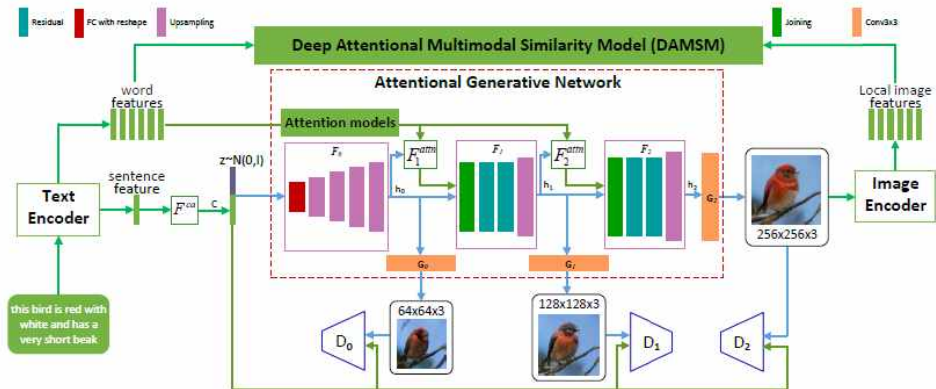


[그림 2-2] Global Attention 도식화

AttnGAN은 문장에 Attention 방식의 인코딩을 적용하여, 텍스트와 이미지간의 밀접한 연관성을 만들어 냈으며, 이러한 핵심모듈을 DAMSM(Deep Attentional Multimodal Similarity Model) 이라 하고, 이 DAMSM 엔코더를 이용하여 이미지 생성모델을 저해상도에서 고해상도로 3단계에 걸쳐 단계적으로 수행하였다.

AttnGAN 모델을 도식화하면 [그림2-3]과 같다.

즉, AttnGAN은 StackGAN++의 단계적 생성모델을 유지한 채로, 텍스트 임베딩방식에 Attention 방식을 적용한 형태이다.



[그림 2-3] AttnGAN 모델 도식화

MirrorGAN¹¹⁾은 이미지 생성 모델을 이용하여 다시 텍스트를 생성하여 비교함으로써, Text-to-Image 의 손실계산 외에 추가적으로 Image-to-Text 방식의 손실계산을 보정하여, 더 향상된 모델로 발전하였다.

AttnGAN에서의 DAMSM 모듈은 MirrorGAN에서 STREAM (Semantic Text Regeneration and Alignment Module) 모듈로 개선되었다.

한편, Image-to-Image 변환 방식으로서 CycleGAN¹²⁾은 생성된 이미지로 다시 이미지를 합성해서 원본과 비교하여 손실오차를 줄이는 방식으로 이미지 생성 품질을 개선한 모델인데, 이러한 순환방식을 Text-to-Image 에 활용하고자 하는 연구도 진행되었다.

Cycle Text-to-Image GAN은 Text-to-Image 방식이지만, MirrorGAN에서 사용된 STREAM 모듈을 이용해서 이미지로부터 텍스트를 생성하는 Image-to-Text 방식을 통해 다시 원본 텍스트와 비교하는 방식을 추가하여, CycleGAN 의 순환 일관성(Cycle Consistency) 손실을 이용하여 이미지생성 품질을 개선했다.

Cycle Text-to-Image GAN 모델을 도식화하면 [그림2-4]와 같다.

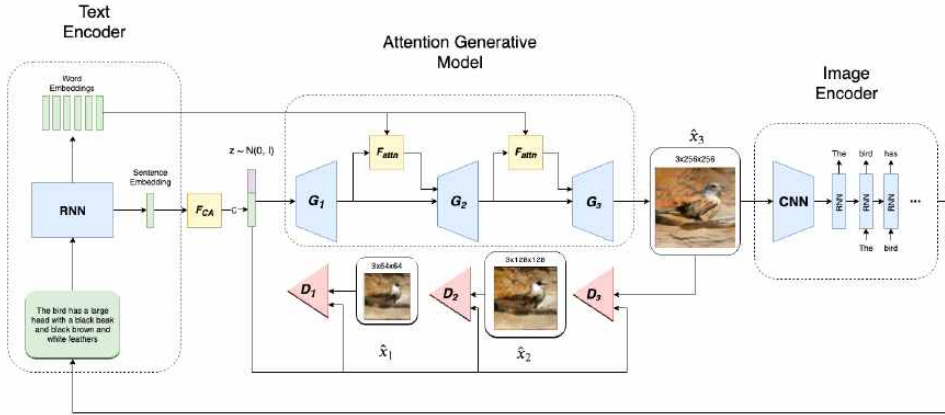
또한, AttnGAN 모델에서는 텍스트를 인코딩할 때, LSTM¹³⁾, GRU¹⁴⁾ 등에 Global Attention을 적용하여 사용했지만, 텍스트 임베딩 방식은 Self-Attention 방식인 Transformer를 활용한 BERT 임베딩을 적용하였다.

11) Tingting Qiao, "Mirrorgan", arXiv:1903.05854, 2019.

12) Jun-Yan Zhu, "CycleGAN", arXiv:1703.10593, 2017.

13) Sepp Hochreiter, "LSTM", Neural computation 9.8, pp.1735-1780, 1997.

14) Kyunghyun Cho, "GRU" arXiv:1406.1078, 2014.



[그림 2-4] Cycle Text-to-Image GAN 도식화

Transformer 방식은 Self-Attention 방식으로 문장 내 단어 간의 가중치를 구하여 가중합하는 방식으로 RNN, LSTM 등 순차적인 모델을 사용하지 않고, 다중 블록으로 분할(Multi-head attention)하여 병렬처리를 가능하게 하여 인코더-디코더 성능을 높였다. 또한, RNN의 장점인 단어위치와 순서를 고려하기 위해 포지셔널 인코딩(Positional encoding)이 추가되었다.

BERT (Bi-directional encoder Representative Transformer)[8]는 Transformer 방식 중 encoder 모듈을 활용한 방식으로, BERT 임베딩을 위한 학습모델은 MLM(Masked Language Model), NSP(Next Sentence Prediction) 두 단계를 거친다.

MLM은 문장 중 일부 단어를 일정 비율로 마스킹을 하여 적당한 단어를 예측하면서 문법적, 의미적 관계를 학습한다. NSP는 다음문장인지 여부를 반복하여 학습함으로써 앞뒤 문장간의 의미 관계를 학습하는 방식이다.

이러한 양방향(Bi-directional) 학습을 통하여, BERT 임베딩 방식은 번역, 독해, 개체명인식 등 다양한 분야에서 우수한 성능을 보여주고 있다.

Cycle Text-to-Image GAN은 텍스트 임베딩 모델을 BERT 임베딩 방식 외에 다른 사전학습된 모델(pretrained model)을 이용하여 전이학습을 통해 다양하게 세부조정(fine tuning)을 할 수 있는 장점이 있다.

예를 들어, 최근에 BERT 이후로 더 발전하고 있는 자연어 처리 모델인 ALBERT¹⁵⁾, RoBERTa¹⁶⁾ 등을 적용해 볼수 있다.

15) Zhenzhong Lan, "ALBERT", arXiv:1909.11942, 2019.

16) Yinhan Liu, "RoBERTa", arXiv:1907.11692, 2019.

국내에서도 이미지 합성에 관한 유사한 연구가 진행되었다.

“텍스트 매핑을 이용한 스케치 기반의 얼굴 이미지 생성”¹⁷⁾ 논문에서는 스케치를 통한 얼굴 이미지를 생성하는데 있어서, 얼굴의 특징이 되는 키워드 텍스트를 입력 제약조건으로 주었으며, Conditional GAN을 이용한 Image-to-Image 변환 방식인 Pix2Pix¹⁸⁾ 모델을 활용하였다.

다만, 입력으로 사용된 텍스트는 자연어가 아닌 이미지의 특징을 나타내는 분류를 위한 키워드라는 한계가 있다.

“텍스트 자질을 활용한 이미지 생성”¹⁹⁾에서는 StackGAN 를 활용하는데 있어서, 사람의 야구 동작의 지역조건, 위치 등 텍스트 설명 데이터를 직접 제작하여 이미지 생성에 활용한 논문이다.

텍스트 임베딩은 Word2Vec²⁰⁾을 이용하였으며, StackGAN의 1단계에서는 64x64의 키포인트 이미지를 스케치하고, 2단계에서는 키포인트와 텍스트 임베딩을 합쳐 256x256의 고화질로 정제된 이미지를 생성하였다.

다만, 텍스트 임베딩이 기본적인 단어임베딩인 Word2Vec를 활용하는데 그쳤으며, 새로운 데이터셋을 GAWWN²¹⁾에서 제시하는 키포인트 제약조건에 맞추어 직접 제작하여 StackGAN 모델을 적용해 본 점에 의미를 둔 연구 사례라고 할수 있다.

“Self-Attention을 적용한 문장 임베딩으로부터 이미지 생성 연구”²²⁾에서는 텍스트 임베딩에 LSTM을 이용해서 텍스트 문장 내에서 직접 Attention을 사용하였고, GAN Text-to-Image 보다는 나은 품질을 보였으나, BERT 임베딩 활용이나 CycleGAN 형태의 시도에는 접근하지 못했다.

한편으로 자연어 처리에 있어서, BERT 임베딩을 특정 분야에 특화시켜서 활용하려는 연구도 진행되었다.

“한국어 기술문서 분석을 위한 BERT 기반의 분류모델”²³⁾ 논문에서는 BERT 의 기본 임베딩에 기술분야 국가과제 데이터를 학습하여 기술문서 분류에 특화된 모델로 정확도를 향상시킨 연구사례이다.

17) 김민정, “텍스트 매핑을 이용한 스케치 기반의 얼굴 이미지 생성”, 한국정보과학회, 2017.

18) Isola, Phillip, “Pix2Pix”, arXiv:1611.07004, 2016.

19) 유민환, “텍스트 자질을 활용한 이미지 생성”, 고려대학교 대학원, 2019.

20) T. Mikolov, “Word2Vec” arXiv:1301.3781, 2013.

21) S. E. Reed, “Learning what and where to draw”, arXiv:1610.02454, 2016.

22) 유경호, “Self-Attention을 적용한 문장 임베딩으로부터 이미지 생성 연구”, 한국스마트미디어학회, 2021.

23) 황상흠, “한국어 기술문서 분석을 위한 BERT 기반의 분류모델”, 한국전자거래학회, 2020.

"감정 분석을 위한 BERT 사전학습모델과 추가 자질 모델의 결합"²⁴⁾ 논문에서는 감정 자질 언어를 BERT 기본 임베딩에 결합하여 적용함으로써 영화평 분류나 댓글 감성분석에 더 나은 성능을 내는 것을 확인한 연구사례이다.

"BERT 기반 정유사 뉴스의 감성분석"²⁵⁾에서는 정유사 관련 뉴스 등의 네이버 기사 자료를 BERT 임베딩에 활용하여 정유사 뉴스 관련 감성분석 결과에서 다른 일반적 한국어 임베딩에 비해 더 나은 성능을 내는 것을 확인한 연구사례이다.

24) 이상아, "감정 분석을 위한 BERT 사전학습모델과 추가 자질 모델의 결합", 한국정보과학회, 2020.

25) 백명현, "BERT 기반 정유사 뉴스의 감성분석", 서강대학교 정보통신대학원, 2021.

제3장 제안하는 방법

1. BERT 임베딩 모델 생성

Cycle Text-to-Image GAN 논문에서 사용된 BERT 임베딩 모델은 Google에서 사전훈련(pre-trained)하여 배포한 모델 중 하나인 “bert-base-uncased” 을 사용하였다.

이 모델은 약 25억 단어로 구성된 영문 위키피디아 등을 말뭉치로 학습하여 사전훈련된 모델로서, 어휘사전은 30,522 개의 단어로 구성되어 있다.

그만큼 다양한 분야의 말뭉치로 학습이 되었기 때문에, 특정 분야에서는 단어들 간의 의미적 상관관계가 더 약하게 나타난다고 볼수 있다.

예를 들어, 표준 이미지 데이터셋인 CUB-200(Caltech-UCSD Birds-200-2011) 내의 텍스트는 다양한 새들의 종류와 특징을 묘사한 것인데, 일반적인 어휘사전으로 묘사할때는 의미적 상관관계가 떨어질 수 있다.

따라서, 본 논문에서 제안하는 방식은 텍스트 임베딩을 활용함에 있어서, 사전훈련된 BERT 모델을 사용하지 않고, 새에 특화된 분야에 대해 직접 학습모델을 만들어 임베딩으로 활용함으로서, 문장내 단어들간의 의미적 상관관계를 더 향상시키고자 한다.

1.1 어휘사전 만들기

새에 특화된 분야의 어휘를 준비하기 위해서는 새를 묘사하는 어휘로 구성된 텍스트 말뭉치가 필요하다.

CUB-200 데이터셋 내의 그림설명 텍스트는 그림 카테고리 하나에 40~60여 개의 그림이 들어있고, 각 그림당 서로 다른 표현으로 10문장씩 그림 설명이 포함되어 있다.

전체적으로 약 13만 개의 문장과 약 179만 개의 단어로 구성되어 있으며, 각 그림 설명들이 새의 동작과 특징에 대해 잘 서술하고 있으므로, 전체를 말뭉치로 활용하기로 한다.

이렇게 수집된 말뭉치를 기반으로 어휘사전을 만든다.

어휘사전을 만들때 단순히 띄어쓰기나 어절 단위로 단어를 구분하여 만들면 어휘수가 많아지고 모델 연산을 위한 파라미터가 복잡해지므로, 이러한 문제를 해결하기 위해 Google에서 개발한

SentencePiece를 토큰나이저(Tokenizer) 툴로 사용하기로 한다.

SentencePiece 는 subword units²⁶⁾ 와 unigram language model²⁷⁾ 을 구현한 것으로, 고정된 어휘사전을 사용하지 않고, 말뭉치에서 단어의 발생빈도수에 따라 토큰 어휘사전을 직접 만들고 그 분리된 토큰을 단어의 기본으로 하여 문장을 표시하는 방식이다.

또한, subword units 는 BPE(byte-pair encoding) 방식으로써 단어를 최소 음절 단위를 빈도수에 따라 병합해서 분절하는 형태이며, unigram 모델은 하나의 서브워드 단위를 최소로 분절하는 형태이다.

BPE 나 unigram 방식은 임베딩하기 위한 어휘수를 줄여주며, 이러한 서브워드 방식은 문서에 등장하는 실제 어휘 기준으로 분리하므로 OOV(out-of-vocabulary) 가능성을 낮춘다.

1.2 토큰화

CUB-200 그림설명 텍스트 전체를 말뭉치로 입력했을 때, SentencePiece 에 생성된 어휘사전은 3,766 개의 토큰으로 구성되었다.

이 어휘사전을 기본으로 모든 입력 문장을 토큰단위로 분리하고 숫자 배열화 하여 벡터로 표현하는 작업을 토큰화(Tokenizing)이라고 한다.

문장별 토큰 임베딩을 정해진 길이로 하기 위해, 문장별 토큰수는 최대 토큰 길이로 제한 되고, 최대 토큰 길이보다 작은 토큰 수는 패딩으로 채워진다.

1.3 사전훈련 모델 생성

[그림3-1]은 BERT 임베딩을 위한 사전훈련 모델의 입력에 대한 도식화된 예시이다.

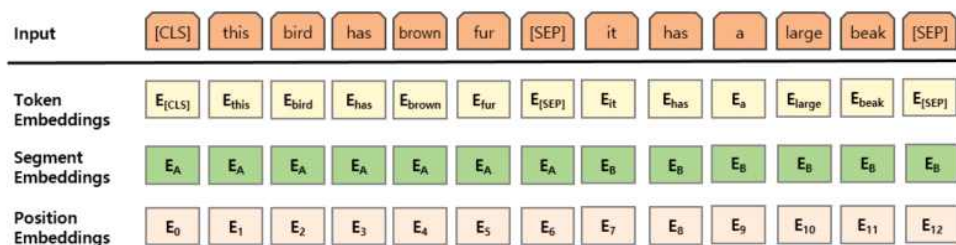
위에서 입력 문장을 CUB-200 데이터셋의 새에 대한 문장 샘플로 대체하여 표현하였으며, SentencePiece 로 분리한 토큰(Token Embeddings)과 문장 구분(Segment Embeddings), 토큰 위치 구분(Position Embeddings)으로 구성된다.

토큰 구분 중 문장시작은 "[CLS]", 문장의 끝은 "[SEP]" 로, 패딩은

26) Rico Sennrich, "Subword Units", arXiv:1508.07909, 2015.

27) Taku Kudo, "Subword Regularization", arXiv:1804.10959, 2018.

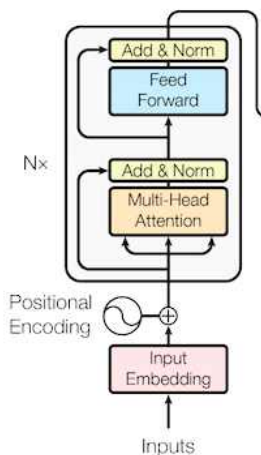
"[PAD]"로 구분한다. 문장구분은 첫 번째 문장인지 두 번째 문장인지를 구분하며, 위치 구분은 문장 내 토큰의 순차적인 위치 순서이다.



[그림 3-1] BERT 임베딩 입력 예시

BERT 모델은 Transformer 의 인코더-디코더 구조에서 인코더만을 활용하여 여러 개의 블록으로 쌓은 구조이다.

[그림3-2]는 BERT 인코더 모델의 구조를 도식화한 예시이다.



[그림 3-2] BERT 인코더 모델 구조 도식화

BERT 인코더는 N 개수만큼 Transformer의 인코더를 쌓은 구조이며, Multi-Head Attention과 Feed Forward 두 개의 서브 층으로 구성된다.

Multi-Head Attention 은 Self-Attention을 여러개를 병렬적으로 연결한 구조이고, Feed Forward 는 완전 연결(Fully-Connected) 신경망이며, Add & Norm은 잔차 연결(Residual connection) 및 정규화(Normalization)를 의미한다.

CUB-200 데이터셋의 텍스트 말뭉치를 이용하여 BERT 임베딩의 사전훈련에 필요한 모델 설정과 파라미터 구성은 [표3-1]과 같다.

[표 3-1] BERT 임베딩 학습 파라미터

구분	파라미터 항목	설명	값
모델 설정	hidden_size	은닉계층 차원 수	768
	num_attention_heads	어텐션 헤드 수	12
	num_hidden_layers	은닉계층 개수	12
	vocab_size	어휘 개수	3,766
훈련용	num_steps	스텝 수	1,000,000
	batch_size	배치 개수	96
	learning_rate	학습률	1e-5
	max_sequence_length	시퀀스당 최대 토큰 수	64
	max_predictions_per_seq	시퀀스당 최대 예측 개수	20
	masked_lm_prob	마스킹 비율	0.15

위와 같은 방식으로 새로운 CUB-200 데이터셋의 그림설명 텍스트 말뭉치를 입력으로 BERT 임베딩 모델을 생성한 후, 이것을 텍스트 임베딩으로 활용하여 각 텍스트에 대한 이미지를 생성하여 개선된 점을 이미지-텍스트 유사도, Inception 점수 등 측정값으로 검증해 본다.

2. 이미지-텍스트 유사도 검증

AttnGAN 논문에서 문장과 이미지간의 유사도를 이용하여 관련성을 표현하기 위해 DAMSM(Deep Attentional Multimodal Similarity Model) 모듈을 도입하였다.

이미지와 텍스트 임베딩 벡터를 동일 차원으로 정의하여, 이미지의 서브 영역과 단어 어휘간의 문맥 벡터를 표현하여 이미지-텍스트간의 유사도 점수를 계산하는 방법을 사용하였다.

먼저, 텍스트 임베딩을 e , 이미지의 특징벡터를 v 라고 할 때, 유사도 계산을 위한 행렬은 $s = e^T v$ 로 표현할수 있고, $s_{i,j}$ 는 i 번째 단어와 j 번째 이미지 서브영역의 내적값으로 정의할 때, 이미지 서브영역에 대한 문맥 벡터 c_i 는 이미지의 서브영역과 문장의 i 번째 단어에 대한 동적 표현으로서 아래와 같이 표현된다.

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \text{ where } \alpha_j = \frac{\exp(\gamma_1 \overline{s_{i,j}})}{\sum_{k=0}^{288} \exp(\gamma_1 \overline{s_{i,k}})}, \text{ where } \overline{s_{i,j}} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})}$$

결국, 문장(D)과 이미지(Q) 간의 관련성을 표현하기 위해, 텍스트와 이미지 서브영역에 대해 유사도 $R(c_i, e_i)$ 는 코사인 유사도를 이용하여 아래와 같이 표현할 수 있다.

$$R(c_i, e_i) = (c_i^T e_i) / (\|c_i\| \|e_i\|)$$

또한 위 식을 이용하여, 전체 이미지-텍스트간 유사도 점수는 아래와 같이 표현할 수 있다.

$$R(Q, D) = \log \left(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i)) \right)^{\frac{1}{\gamma_2}}$$

이 유사도 점수를 이용하여, 이미지에 대한 문장의 사후확률을 구하면, 다음과 같다.

$$P(D_i | Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^M \exp(\gamma_3 R(Q_i, D_j))}$$

이 사후확률에 대해 손실함수로 표현하면, 아래와 같다.

$$L_1^w = - \sum_{i=1}^M \log P(D_i | Q_i)$$

또한, 역으로 문장에 대한 이미지의 사후확률을 이용하면, 손실함수를 아래와 같이 표현할 수 있다.

$$L_2^w = - \sum_{i=1}^M \log P(Q_i | D_i)$$

위와 비슷하게 문장과 이미지에 대한 손실을 표현하기 위해,
임베딩 벡터 e 대신 문장벡터 \bar{e} 로,
서브영역 이미지 벡터 v 대신 전역 이미지벡터 \bar{v} 를 사용하여
유사도 점수를 표현하면, 아래와 같다.

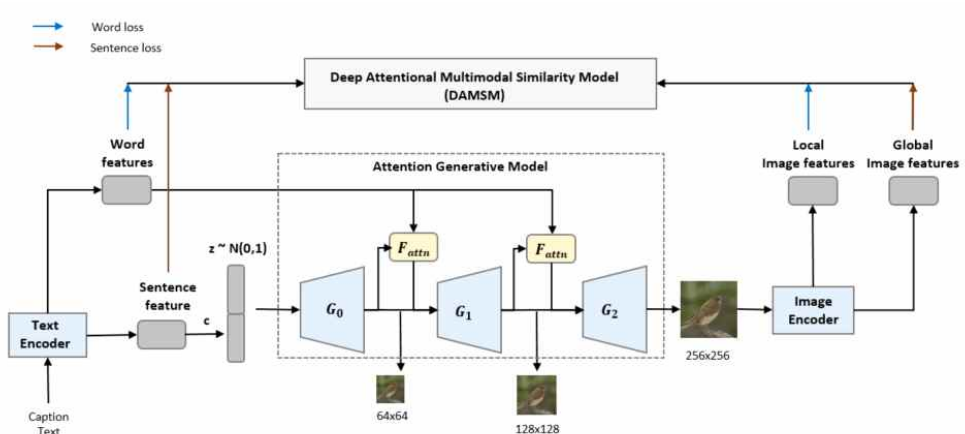
$$R(Q, D) = (\bar{v}^T \bar{e}) / (\|\bar{v}\| \|\bar{e}\|)$$

위의 식을 이용하여, 단어와 이미지 서브영역의 손실값 $L_1(w)$, $L_2(w)$ 의 계산식과 비슷한 방식으로 문장과 이미지 전체영역의 손실값 $L_1(s)$, $L_2(s)$ 도 계산할 수 있다.

결국, DAMSM 의 전체 손실의 계산은 위의 네 가지의 손실의 합
의 형태로 아래와 같이 표현된다.

$$L_{DAMSM} = L_1^w + L_2^w + L_1^s + L_2^s$$

AttnGAN 모델에서 DAMSM 모듈을 도식화 하면 [그림 3-3]와 같다.



[그림 3-3] DAMSM 모듈 도식화

위 그림에서 단어와 이미지 서브영역간(word loss), 문장과 전체 이미지간(sentence loss)의 유사도 손실계산 흐름을 구분하여 표시하였다.

이러한 DAMSM 의 손실값을 측정함으로써, 직접 훈련한 BERT 모델이 Google 의 사전훈련된 BERT 모델에 비해 더 손실값이 적고 이미지-텍스트가 유사도가 높음을 검증한다.

3. 이미지 생성 품질 검증

Cycle Text-to-Image GAN에서는 생성된 이미지 품질을 정량적으로 측정하기 위해 Inception score²⁸⁾ 을 활용하였는데, 계산식은 아래와 같다.

$$IS(G) = \exp(E_x D_{KL}(p(y|x) \| p(y)))$$

여기에서 D_{KL} 은 KL 발산(Kullback-Leibler Divergence)인데, 두 개의 확률분포 $P(x)$ 와 $Q(y)$ 의 차이를 표현하는 것으로 아래처럼 표현될 수 있다.

$$D_{KL}(P||Q) = - \sum_{x \in X} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

위에서 x 는 제안한 모델에 의해 생성된 이미지 샘플이고, y 는 주어진 x 에 대하여 이미지 분류모델인 Inception-v3²⁹⁾에 의해 예측된 라벨값이다.

결국, Inception score 는 주변확률분포 $p(y)$ 와 조건부 확률분포 $p(y|x)$ 의 차이를 의미하는 KL 발산을 측정한 것이며, 숫자가 클수록 다양하고 의미있는 이미지를 만들어 낸다는 것을 뜻한다.

또한, 두 확률분포 사이의 거리를 나타내는 척도로 FID(Frechet Inception Distance)³⁰⁾가 있는데, FID는 두 정규분포 사이의 거리를 의미한다.

먼저, 두 단일변량(univariate)의 정규분포의 거리는 다음과 같다.

$$d(X, Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

여기에서 μ 와 σ 는 정규분포의 평균 및 표준편차를 나타내며, μ_X 와 μ_Y , 그리고 σ_X 와 σ_Y 는 각각 정규분포를 따르는 X , Y 의 평균과 표준편차를 의미한다.

두 다변량(multivariate)의 정규분포의 거리는 다음과 같이 표현되다.

$$FID = \|\mu_X - \mu_Y\|^2 - \text{Tr}\left(\sum_X + \sum_Y - 2\sqrt{\sum_X \sum_Y}\right)$$

여기에서 Tr 은 행렬의 대각합(trace)이고, Σ_X 와 Σ_Y 는 공분산 행렬(covariance matrix)을 의미한다.

GAN에서 FID를 활용할 때 X 는 실제이미지 데이터셋, Y 는 모델을

28) Tim Salimans, "Improved Techniques for Training GANs", arXiv:1606.03498 , 2016.

29) Christian Szegedy, "Inception-v3", arXiv:1512.00567 , 2015.

30) Martin Heusel, "TTUR(two time-scale update rule)", arXiv:1706.08500, 2017.

통해 생성된 이미지 데이터라고 할 수 있다.

IS는 집합 자체의 우수함을 나타는 반면, FID는 생성된 집합과 실제 데이터 집합의 분포와의 차이를 계산한 것으로 두 집합 사이의 거리를 나타내며, FID는 작을수록 좋다고 할 수 있다.

본 논문에서는, 직접 학습한 BERT 모델을 통해 생성된 이미지의 Inception score 와 Google 의 사전학습된 모델을 통해 생성된 이미지의 Inception score를 비교하여 직접 학습한 BERT 모델을 통해서 생성된 이미지가 더 다양하고 의미있는 이미지를 만들어 낸다는 것을 검증해 본다.

또한, 추가적으로 FID를 비교 측정함으로써 생성된 이미지의 분포와 원본이미지의 분포가 유사한지를 확인해 본다.

제4장 실험 및 평가

1. 이미지 생성 비교

Cycle Text-to-Image GAN 논문에서 사용한 모델명은 CycleGAN w/ BERT 이며, 본 논문에서 제안한 모델은 CycleGAN w/ local BERT 라고 표현하기로 한다. 더 정확한 표현은 CycleGAN w/ local-pretrained (or domain-specific) BERT이다.





본 논문의 제안에 따라, GAN 이미지 생성모델 중에 AttnGAN, CycleGAN w/ BERT, CycleGAN w/ local BERT 세가지 모델을 비교한다.






AttnGAN 모델은 LSTM 언어임베딩을 활용하여 Global Attention을 적용하였고, CycleGAN w/ BERT 모델은 이 구조에서 언어임베딩 모델을 Google 의 사전학습된 BERT 언어모델을 사용하였으며, CycleGAN w/ local BERT 모델은 언어임베딩 모델을 특정 분야(새)에 특화하여 임베딩 모델을 직접 제작하여 사용하였다.


이미지 생성을 위한 텍스트 샘플은 CUB-200 데이터셋 중에서, 3가지 카테고리리를 선택하였는데, 001.Black Footed Albatross ~ 003.Sooty Albatross 까지이다.

[표4-1]은 위의 실험 데이터셋 중 3개의 카테고리에 대해, 각 모델을 통해 생성하여 이미지를 비교한 예시이다.

[표4-1] 모델별 이미지 생성 비교

Category	AttnGAN	CycleGAN w/ BERT	CycleGAN w/ local BERT	Ground Truth
001 Black Footed Albatross				

	bird has brown body feathers, white breast feathers and black beak			
				
	this bird has grey neck, head, wings and back, it has white around its bill, and a long tall bill that is curved and black at its tip.			
002 Laysan Albatross				
	this bird has a curved white bill, a white belly, and black primaries.			
				
003 Sooty Albatross	the bird has a long white bill and long black secondaries.			
				
	this bird is white and grey in color with a curved black beak, and white eye rings.			

				
	this small bird has a black flat bill, fuzzy black feathers and small feet.			

첫 번째 카테코리(Black Footed Albatross)의 위쪽 샘플에서 CycleGAN w/ local BERT 모델은 다른 모델에 비해 갈색 몸 깃털(brown body feathers)과 흰 가슴 깃털(white breast feathers), 검은색 부리(black beak)를 잘 묘사하였다.

다만, 텍스트 설명에는 없으나, 실제 사진과 비교하면 긴 부리를 묘사하지는 못했다. 오히려 긴 부리는 CycleGAN w/ BERT 모델에서 잘 묘사하였다.

첫 번째 카테고리의 아래쪽 샘플에서 CycleGAN w/ local BERT 모델은 다른 모델에 비해 회색 목(gey neck), 머리(head), 날개(wings), 등(back)을 잘 묘사하였고, 부리 주위의 흰색(white around its bill)과 휘어진 긴 부리(long tall bill)와 부리 끝에 검정색(tall bill that is curved and black at its tip)을 잘 묘사하였다.

두 번째 카테고리(Laysan Albatross)의 위쪽 샘플에서 CycleGAN w/ local BERT 모델은 휘어진 흰 부리(curved white bill), 하얀 배(white belly), 검정색 주 날개(black primaries)를 특징적인 부분은 어느정도 묘사하였으나, 몸이 너무 통통하게 표현되어 실제 사진에 비해 부자연스럽다.

두 번째 카테고리의 아래쪽 샘플에서 CycleGAN w/ local BERT 모델은 긴 흰 부리(long white bill)와 긴 검정색 부날개(long black secondaries)를 특징적인 부분은 어느 정도 묘사하였으나 머리와 다리 부분에 불필요한 형태가 추가적으로 표현되어 실제 사진에 비해 부자연스럽다.

세 번째 카테고리(Sooty Albatross)의 위쪽 샘플에서 CycleGAN w/ local BERT 모델은 흰색과 회색 바탕에(white and grey in color) 휘어진 검정색 부리(curved black beak)를 잘 묘사하였고, 흰 눈테(white eye rings)도 표현하였으나 검정색에 가깝게 표현되었다.

세 번째 카테고리의 아래쪽 샘플에서 CycleGAN w/ local BERT 모델은 검은색 평평한 부리(black flat bill), 보송보송한 검정색 깃털(fuzzy black feathers)과 작은 발톱(small feet)을 잘 표현하였다.

2. DAMSM Loss 측정

AttnGAN이 사용하고 있는 DAMSM 모듈은 텍스트 인코더와 이미지 인코더를 사용하여 텍스트와 문장간의 유사도를 최대화 하기 위한 모듈이다.

Cycle Text-to-Image GAN 에서는 DAMSM과 동일한 모듈을 사용하고, 이 외에 CycleGAN의 순환 일관성 손실 계산을 위한 Text Regeneration 블록이 추가되어 STREAM 모듈로 개선되었다.

결국, 텍스트 Attention을 활용한 세 가지 모델에 대해, 텍스트와 문장간의 유사도 측정을 위해서 DAMSM 모듈의 손실값을 공통적으로 비교해 볼 수 있다.

모델별로 이미 학습된 DAMSM 인코더를 사용하여, 10개의 데이터셋 샘플에 대해 이미지-텍스트 유사도를 활용한 DAMSM 손실값을 측정해 본다.

본 실험에서는 50~200 주기(epoch) 범위 내에서 학습된 이미지 및 텍스트 인코더를 활용하여 학습 정도에 따라 손실값의 차이를 비교해 본다.

DAMSM 손실값을 얻기 위해서 문장내 각 단어와 이미지 서브영역에 대한 매칭확률에 대한 손실 $L_1(w)$, $L_2(w)$ 과 문장과 이미지 전체에 대한 $L_1(s)$, $L_2(s)$ 를 구하여 모두 더한다.

[표4-2]는 단어와 문장에 대한 DAMSM 손실을 측정한 값이다.

[표 4-2] 모델별 DAMSM 손실 측정

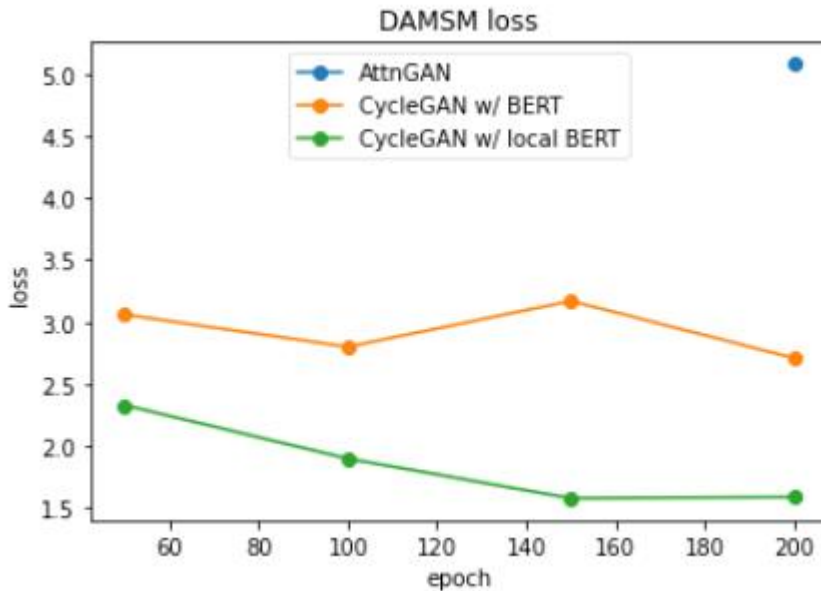
Model		DAMSM loss				
		word loss		sentence loss		total loss
	epoch	L ₁ (w)	L ₂ (w)	L ₁ (s)	L ₂ (s)	
AttnGAN	200	1.32	1.41	1.11	1.23	5.08
CycleGAN w/ BERT	50	0.76	0.52	0.99	0.79	3.06
	100	0.62	0.39	1.01	0.78	2.80
	150	0.76	0.45	1.09	0.87	3.17
	200	0.58	0.41	0.93	0.79	2.71
CycleGAN w/ local BERT	50	0.65	0.62	0.65	0.41	2.33
	100	0.52	0.45	0.57	0.36	1.90
	150	0.41	0.35	0.50	0.32	1.58
	200	0.41	0.30	0.52	0.36	1.59

동일한 학습주기(200 epoch) 대해 DAMSM 손실은 AttnGAN 이 제일

크고, CycleGAN w/ BERT, CycleGAN w/ local BERT 순으로 작아짐을 확인할 수 있다. 특히, AttnGAN 은 200 epoch의 학습주기에도 불구하고 다른 두 모델의 50 epoch 단계에서 보다 손실값이 크게 나타났다.

CycleGAN w/ BERT 와 CycleGAN w/ local BERT는 각 학습주기별로 비교했을 때, local BERT를 사용한 모델의 손실값이 더 작게 나타났으며, 또한 학습 주기(epoch)가 커질수록 손실값이 더 작아짐을 확인할 수 있다. 다만, CycleGAN w/ BERT는 200 epoch에서 최소값을 갖긴 했지만, 100 epoch 단계와 별 차이가 없고, CycleGAN w/ local BERT는 150 epoch 단계 이후로 크게 감소하지 않았다.

[그림4-1]은 위 표의 주기별 손실값 변화율을 그래프로 표시하였다.



[그림 4-1] 학습주기별 DAMSM 손실값 변화율

결국, GAN 훈련에서는 사용된 DAMSM 인코더는 100~200 epoch 범위 내에서 적당한 학습주기를 선택할 수 있으며, 본 논문에서는 200 epoch로 학습된 모델을 사용하였다.

결국, 위의 측정 결과에 의하면, CycleGAN w/ local BERT 모델에서는 텍스트임베딩과 그림의 서브영역에 대한 유사도가 커져서 다른 모델보다 이미지 생성 품질이 향상될 수 있음을 보여준다.

3. Inception score 측정

Inception score를 측정하기 위해, 각 모델별로 3가지 카테고리에 대해서 이미지를 생성한 후, Inception score 측정모듈에 입력하여 계산한다.

[표4-3]과 [표4-4]는 데이터셋 카테고리별 생성한 샘플이미지에 대하여, 600 주기(epoch)로 학습된 각 모델별로 Inception score를 측정하여 비교하였다.

카테고리별 이미지 개수는 약 60개이고, 이미지 한 개당 10개의 캡션이 있으며, 각 캡션별로 이미지가 생성되므로 생성된 이미지 전체 개수는 원본 이미지 개수 x 10 개이다.

[표 4-3] Inception score 측정

Category	원본 Image 수	Inception score			
		생성 Image 수	AttnGAN	CycleGAN w/ BERT	CycleGAN w/ local BERT
001. Black footed Albatross	60	600	3.21	4.57	4.73
002. footed Albatross	60	600	3.06	3.88	4.21
003. Sooty Albatross	58	580	3.44	4.20	4.75

Inception score 는 AttnGAN이 제일 낮고, CycleGAN w/ BERT, CycleGAN w/ local BERT 순으로 증가한다.

즉, CycleGAN w/ local BERT 모델에서 더 다양하고 고른 이미지가 생성되는 것을 정량적으로 확인할 수 있다.

[표 4-4] FID 측정

Category	원본 Image 수	FID			
		생성 Image 수	AttnGAN	CycleGAN w/ BERT	CycleGAN w/ local BERT
001. Black footed Albatross	60	600	226.4	197.5	205.7
002. footed Albatross	60	600	228.0	160.3	184.6
003. Sooty Albatross	58	580	211.6	176.8	185.8

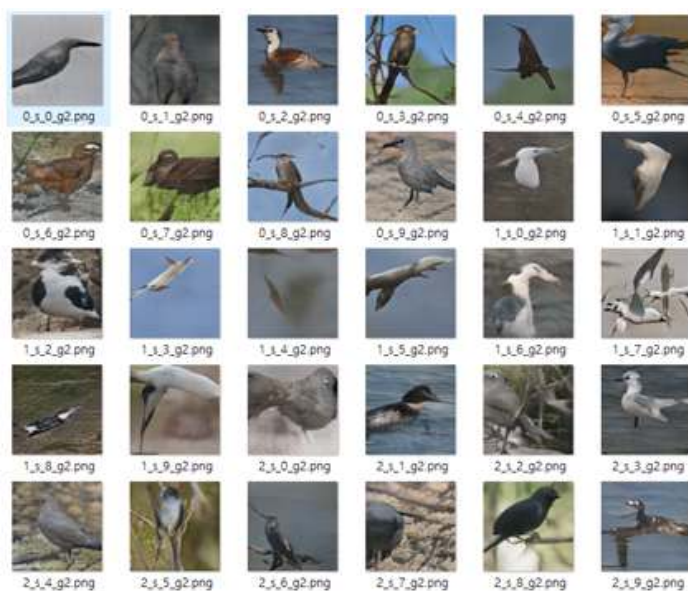
한편, FID 는 AttnGAN이 제일 크게 나왔고, CycleGAN w/ BERT, CycleGAN w/ local BERT 순으로 감소할 것으로 예상하였으나, CycleGAN w/ local BERT 가 FID가 크게 나왔다.

이것은 CycleGAN w/ local BERT 모델에서 생성된 품질이 상당수가 잘 나오기도 하지만, 일부 이미지는 품질이 안 좋은 것도 많이 나타난다. 결국, 전체적으로는 품질의 좋고 나쁨의 분포가 더 커서 원본이미지의 품질 분포와 차이가 많이 날수 있다.

[그림 4-2]는 CycleGAN w/ BERT 로 생성된 그림이고, [그림 4-3]은 CycleGAN w/ local BERT 로 생성된 그림 샘플이다.

상당수는 좋은 품질의 이미지가 생성되지만, 일부는 낮은 품질의 이미지도 발견된다.

이러한 이유는 텍스트 임베딩이 전반적으로 고르게 학습되어 있지 않기 때문으로 보이며, 좀 더 세부적이고 특화된 분야의 텍스트로 충분히 학습된다면, 다양한 텍스트에 대해 이미지 생성 품질이 고르게 더 향상될 것으로 판단된다.



[그림4-2] CycleGAN w/ BERT 이미지 생성 예시



[그림4-3] CycleGAN w/ local BERT 이미지 생성 예시

제5장 결론

이미지 생성모델에 대한 연구는 GAN을 활용한 Text-to-Image 방식으로 단계를 나눠서 진행한 StackAN 이후, AttnGAN에서 텍스트 임베딩에 Attention을 활용함으로 진전을 보였다.

텍스트 임베딩에 Attention을 활용하는 방법에 있어서 Self-Attention 방식인 Transformer 블록을 활용한 BERT 임베딩은 자연어 처리분야에서도 큰 진전을 보였는데, CycleGAN w/ BERT에도 활용되어 향상된 이미지 합성 품질을 보여줬다.

본 논문에서는 BERT 임베딩을 Google 의 사전훈련된 모델을 그대로 사용하지 않고, 직접 특정 분야의 언어가 담긴 말뭉치를 활용하여 특화된 BERT 임베딩모델을 만들어 적용하였다.

이미지와 문장간의 유사도 손실을 측정결과 손실이 감소하였으며, Inception 측정결과도 높게 나와서 다양하고 좋은 품질의 이미지가 생성됨을 확인하였다.

다만, CUB-200 데이터셋 텍스트 전반에 있어서 좋은 이미지 생성 결과가 나온 것은 아니며, 오히려 일부 이미지들은 이전 모델보다 낮은 품질의 결과가 나오기도 하였다.

이것은 특정 분야의 말뭉치를 CUB-200 데이터셋의 캡션 텍스트에만 의존하였기 때문에, 새의 형태를 묘사하는 일부 정교한 언어에 대한 학습이 덜 되었기 때문으로 판단된다.

따라서, 향후 연구에서는 이러한 문제를 극복하기 위해, 새를 묘사하는 백과사전의 용어나 설명 등 다양한 언어를 말뭉치에 추가하여 BERT 임베딩 모델을 만들어 적용하면 텍스트-이미지간의 좀 더 정교한 표현을 통해 상호 연관성이 잘 표현되어 보다 향상된 품질의 이미지 생성이 가능할 것으로 예상된다.

참 고 문 헌

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets”, NIPS, 2014.

M. Mirza and S. Osindero. “Conditional generative adversarial nets”, arXiv:1411.1784, 2014.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis”, arXiv/1605.05396., 2016.

H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”, arXiv:1612.03242., 2017.

H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks”, arXiv:1710.10916, 2017.

Thang Luong, Hieu Pham, Christopher D. Manning, “Effective Approaches to Attention-based Neural Machine Translation”, D15-1166, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is all you need”, arXiv:1706.03762, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding”, arXiv:1810.04805, 2018.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan,

Xiaolei Huang, and Xiaodong He. "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks", arXiv:1711.10485, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks", arXiv:1703.10593, 2017.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. "Mirrorgan: Learning text-to-image generation by redescription", arXiv:1903.05854, 2019.

Trevor Tsue, Samir Sen, Jason Li, "Cycle Text-To-Image GAN with BERT", arXiv:2003.12137, 2020.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, "Improved Techniques for Training GANs", arXiv:1606.03498 , 2016.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision", arXiv:1512.00567 , 2015.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", arXiv:1609.08144, 2016.

Zhenzhong Lan, Mingda Chen, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", arXiv:1909.11942, 2019.

Yinhan Liu, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692, 2019.

Isola, Phillip, et al. "Image-to-image translation with conditional

adversarial networks.”, arXiv:1611.07004, 2016.

S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, “Learning what and where to draw”, arXiv:1610.02454, 2016.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv:1301.3781, 2013.

Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” Neural computation 9.8, pp.1735–1780, 1997.

Kyunghyun Cho, “Learning phrase representation using RNN encoder-decoder for statistical machine translation,” arXiv:1406.1078, 2014.

Rico Sennrich, “Neural Machine Translation of Rare Words with Subword Units”, arXiv:1508.07909, 2015.

Taku Kudo, “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”, arXiv:1804.10959, 2018.

Martin Heusel, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”, arXiv:1706.08500, 2017.

김민정, 임형석, 박영준, 주예슬, 구명완, “텍스트 매핑을 이용한 스케치 기반의 얼굴 이미지 생성”, 한국정보과학회, 2017.

유민환, 강재우, “텍스트 자질을 활용한 이미지 생성”, 고려대학교 대학원, 2019.

유경호, 노주현, 홍택은, 김형주, 김판구, “Self-Attention을 적용한 문장 임베딩으로부터 이미지 생성 연구”, 한국스마트미디어학회, 2021.

황상흠, 김도현, “한국어 기술문서 분석을 위한 BERT 기반의 분류모델”, 한국전자거래학회, 2020.

이상아, 신호필, “감정 분석을 위한 BERT 사전학습모델과 추가 자질 모델의 결합”, 한국정보과학회, 2020.

백명현, 구명완, "BERT 기반 정유사 뉴스의 감성분석", 서강대학교
정보통신대학원, 2021.

GAN을 이용한 텍스트로부터 이미지 변환방법에 관한 연구 CycleGAN과 BERT 임베딩을 활용한 텍스트로부터 이미지 변환 방법 개선

텍스트로부터 이미지 합성을 위한 GAN 모델에 있어서, 텍스트-이미지간 밀접한 결합도 향상을 위해, 제약 조건으로 텍스트-이미지 결합 임베딩 방식을 사용한다.

AttnGAN에서는 텍스트 인코딩시 서로 연관성이 있는 단어에 집중하는 Attention 방식의 임베딩 모델을 GAN에 적용하였고, Cycle Text-to-Image GAN에서는 Self-Attention 방식인 Transformer 구조를 이용한 BERT 임베딩 모델을 사용하였으며, 또한 CycleGAN의 순환 훈련방식을 이용해 AttnGAN 보다 더 향상된 품질을 보여줬다.

본 논문에서 제안한 방식은, Cycle Text-to-Image GAN에서 BERT 임베딩 모델을 사용함에 있어서 일반적인 어휘사전이 아닌 특화된 분야의 어휘들로 구성된 문장 임베딩을 사용하여, 텍스트-이미지 결합간에 어휘 집중도를 높여 학습함으로써, 보다 더 향상된 품질을 확인하였다.