JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 1, pp. 57-67, January 2021
https://doi.org/10.9708/jksci.2021.26.01.057

Text Augmentation Using Hierarchy-based Word Replacement

Museong Kim*, Namgyu Kim*

*Graduate Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
*Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

Recently, multi-modal deep learning techniques that combine heterogeneous data for deep learning analysis have been utilized a lot. In particular, studies on the synthesis of Text to Image that automatically generate images from text are being actively conducted. Deep learning for image synthesis requires a vast amount of data consisting of pairs of images and text describing the image. Therefore, various data augmentation techniques have been devised to generate a large amount of data from small data. A number of text augmentation techniques based on synonym replacement have been proposed so far. However, these techniques have a common limitation in that there is a possibility of generating a incorrect text from the content of an image when replacing the synonym for a noun word. In this study, we propose a text augmentation method to replace words using word hierarchy information for noun words. Additionally, we performed experiments using MSCOCO data in order to evaluate the performance of the proposed methodology.

▸ **Key words**: Deep Learning, Generative Adversarial Network, Text to Image Synthesis, Data Augmentation, WordNet

[요 약]

최근 딥 러닝(Deep Learning) 분석에 이질적인 데이터를 함께 사용하는 멀티모달(Multi-modal) 딥 러닝 기술이 많이 활용되고 있으며, 특히 텍스트로부터 자동으로 이미지를 생성해내는 Text to Image 합성에 관한 연구가 활발하게 수행되고 있다. 이미지 합성을 위한 딥러닝 학습은 방대한 양의 이미지와 이미지를 설명하는 텍스트의 쌍으로 구성된 데이터를 필요로 하므로, 소량의 데이터로부터 다량의 데이터를 생성하기 위한 데이터 증강 기법이 고안되어 왔다. 텍스트 데이터 증강의 경우 유의어 대체에 기반을 둔 기법들이 다수 사용되고 있지만, 이들 기법은 명사 단어의 유의어 대체 시 이미지의 내용과 상이한 텍스트를 생성할 가능성이 있다는 한계를 갖는다. 따라서 본 연구에서는 단어가 갖는 품사별 특징을 활용하는 텍스트 데이터 증강 방안, 즉 일부 품사에 대해 단어 계층 정보를 활용하여 단어를 대체하는 방안을 제시하였다. 또한 제안 방법론의 성능을 평가하기 위해 MSCOCO 데이터를 사용하여 실험을 수행하여 결과를 제시하였다.

▸ **주제어**: 딥 러닝, 생성적 적대 신경망, 이미지 합성, 데이터 증강, 워드넷

• First Author: Museong Kim, Corresponding Author: Namgyu Kim
*Museong Kim (kim27416@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
*Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2020. 11. 30, Revised: 2021. 01. 12, Accepted: 2021. 01. 12.

Copyright © 2021 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인공지능 시대에 접어들면서 비구조적 데이터인 비정형 데이터의 수요와 공급이 기하급수적으로 증가하게 되었으며, 이에 따라 다양한 유형의 비정형 데이터를 분석하기 위해 딥 러닝(Deep Learning) 알고리즘을 활용하는 연구가 활발하게 수행되고 있다. 딥 러닝 알고리즘은 여러 층(Layer)을 쌓아 만든 신경망 모델을 사용하며, 층을 지날 때마다 데이터의 특징을 발견하고 유의미한 표현을 학습하는 과정으로 이루어진다. 이러한 딥 러닝 기술은 자연어 처리(Natural Language Processing), 음성 인식(Speech Recognition), 이미지 분류(Image Classification), 객체 감지(Object Detection) 등에 널리 활용되고 있다.

최근의 딥 러닝 기술은 서로 다른 특징 차원을 가진 데이터를 동시에 학습하는 멀티모달 딥 러닝(Multimodal Deep Learning)에 관한 연구로 확장되는 경향을 보인다. 멀티모달 딥 러닝은 하나의 특징 차원을 가진 데이터를 학습하는 싱글모달 학습(Single Modal Learning)과 달리 텍스트, 이미지, 오디오, 비디오 등의 다양한 데이터를 상호보완적으로 사용하여 학습 성능을 향상시킬 수 있다. 특히 텍스트와 이미지 데이터를 함께 다루는 멀티모달 딥 러닝에 대한 연구가 활발하게 수행되고 있으며, 대표적 응용으로 “Text to Image” 합성이 있다. Text to Image 합성은 입력 텍스트에 대응하는 적절한 이미지를 출력하는 기술로, 생성적 적대 신경망(GAN: Generative Adversarial Network)[1]을 바탕으로 다양한 연구가 이루어지고 있다.

이러한 Text to Image 합성 기술의 잘 알려진 응용 사례로 ReStGAN[2]을 들 수 있다. ReStGAN은 아마존이 개발하여 의류 검색 시스템에 적용한 알고리즘으로, 고객이 입력한 제품 설명과 일치하는 의류를 생성한다. <Fig. 1>은 ReStGAN을 활용한 의류 검색의 예시로, 처음에 “Women’s black pants”을 입력하고 다음으로 “Petite”, “Capri”를 차례로 입력했을 때, 기존 알고리즘인 StackGAN[10]에 비해 훨씬 더 입력 텍스트에 부합하는 의류 이미지를 생성해내는 것을 보인다.

이처럼 Text to Image 합성은 다양한 분야에서 활용 가능성이 높은 기술로 많은 관심을 받고 있지만, 텍스트의 의미를 제대로 반영하는 이미지를 생성하는 것은 상당히 어려운 일이다. 이는 동일한 이미지를 텍스트로 설명할 때 다양한 단어들이 사용될 수 있으며, 동일한 단어라도 문맥에 따라 다른 의미로 해석될 수 있기 때문이다. 즉, 텍스트 데이터의 특징과 이미지 데이터의 특징을 잘 매핑(Mapping)하는 것이 가장 중요한 관건이다.

이질적인 데이터의 특징을 매핑하기 위해, Text to Image 합성은 기본적으로 방대한 양의 이미지와 텍스트 데이터가 학습에 필요하며, 이때 각 이미지와 이미지를 설명하는 복수의 텍스트가 하나의 쌍(Pair)으로 구성되어야 한다. 하지만 이와 같이 이미지와 텍스트의 쌍으로 구

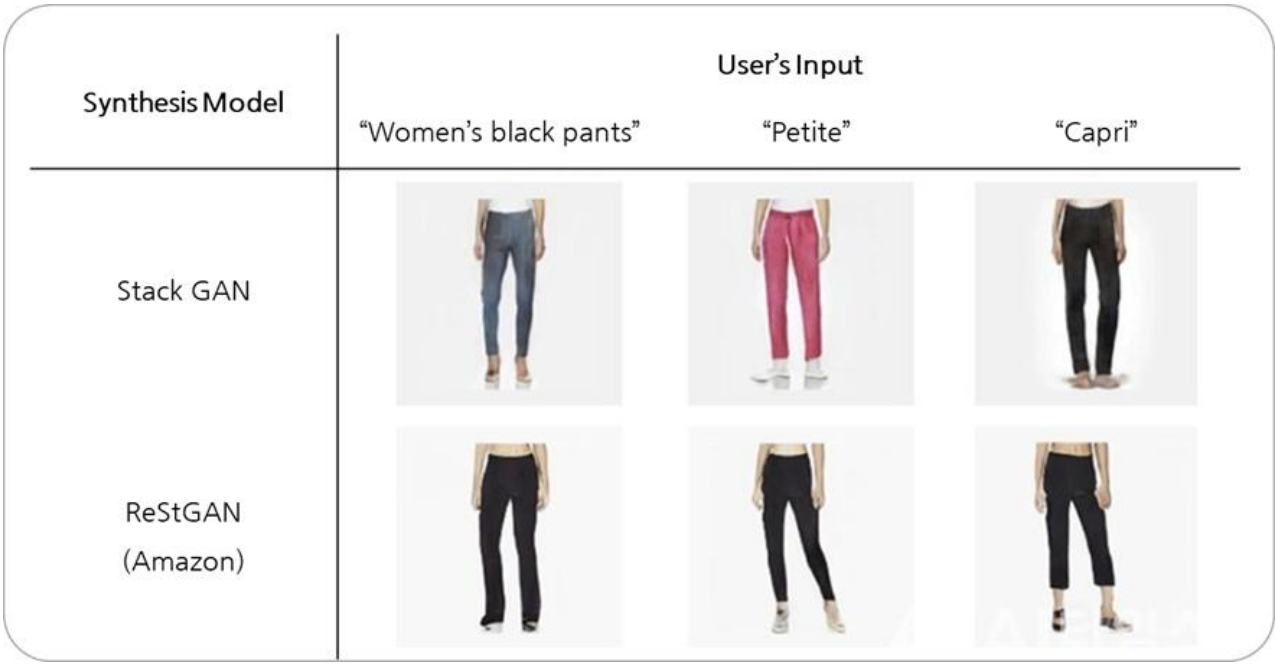

Fig. 1. Examples of Text to Image Synthesis[2]

성된 데이터는 제한적으로 공개되어 있으므로, Text to Image 합성을 위한 충분한 양의 학습 데이터를 확보하는 것은 매우 어려운 일이다.

이와 같은 학습 데이터 부족 문제를 해결하기 위해 자동으로 데이터를 증강하는 방법들이 많이 연구되었으며, 주로 Flipping, Cropping, Rotation 등을 통해 이미지 데이터의 수를 늘리는 증강 방법들이 활용된다[3]. 최근에는 텍스트 데이터를 증강하기 위한 연구도 주목받고 있으며, 대표적인 연구로 시소러스 혹은 임베딩 모델을 사용하여 유의어로 대체하는 어휘 대체 기반의 증강 기법이 있다[4-5]. 하지만 이러한 텍스트 데이터 증강 방법들은 텍스트 대체 과정에서 품사에 따라 각 단어의 계층적(Hierarchical) 관계를 고려하지 못한다는 한계를 갖는다. <Fig. 2>는 이미지와 텍스트가 쌍으로 구성된 데이터에서 명사를 각각 유의어와 상위어로 대체한 가상의 예시이다. 상단의 예는 원본 텍스트의 'car'가 워드넷(WordNet) 상의 유의어인 'cable car'로 대체되어 증강된 텍스트가 원본 이미지의 의미를 왜곡할 수 있음을 나타낸다. 반면 하단의 예는 원본 텍스트의 'car'가 워드넷 상의 상위어인 'motor vehicle'로 대체된 경우로, 원본 텍스트의 의미 왜곡 없이 텍스트가 증강될 수 있음을 나타낸다. 이는 품사에 따라 단어를 계층적으로 대체하는 증강 기법을 통해 증강된 데이터의 품질을 향상시킬 수 있음을 의미한다.

이에 본 연구에서는 기존 텍스트 증강 기법들의 한계를 극복하여 생성 이미지의 품질을 향상시키기 위해 단어의 의미 계층 기반 텍스트 증강 기법을 제안한다. 구체적으로 제안 방법론은 (i) 문장에서 n 개의 단어들을 임의로 선택하고, (ii) 선택된 단어들의 품사를 파악한 후, (iii) 워드넷을 활용하여 품사에 따라 단어를 상이한 방법으로 대체하는 방식으로 텍스트 증강을 수행한다. 또한 제안 기법이 생성 이미지들과 기존의 텍스트 데이터 증강 방법으로 생성된 이미지들에 대한 인셉션 스코어(Inception Score)[6] 비교를 통해 제안 방법의 우수성을 평가한다.

II. Related Research

1. Text to Image Synthesis

딥 러닝은 인공신경망에서 은닉층을 깊게 쌓은 신경망 구조를 활용하여 학습하는 알고리즘으로 합성곱 신경망(CNN: Convolutional Neural Network)[7], 순환 신경망(RNN: Recurrent Neural Network)[8] 그리고 생성적 적대 신경망(GAN: Generative Adversarial Network)[1] 등이 대표적이다. 최근에는 이질적인 데이터 특징들의 표현을 학습하는 멀티모달 학습에 딥 러닝 알고리즘이 많이 활용되고 있으며, 대표적으로 입력 텍스트에 대응하는 적절한 이미지를 생성하는 기술인 Text to Image 합성이 있다.

Text to Image 합성은 생성 알고리즘인 GAN을 바탕으로 다양한 연구가 수행되고 있다. GAN은 생성기 네트워크와 판별기 네트워크가 적대적으로 경쟁하면서 학습을 진행하는 신경망으로, 생성기는 판별기를 속이기 위해 실제와 유사한 가짜 데이터를 생성하고, 판별기는 실제 데이터와 생성된 데이터를 판단하기 위한 학습을 수행한다.

Scott Reed는 2016년 GAN을 활용한 간단한 모델 구조를 통해 텍스트로부터 이미지를 생성해내는 방법론을 제안하였다[9]. 하지만 초창기 GAN의 간단한 모델 구조로는 고해상도 이미지를 생성할 수 없다는 한계가 존재한다. 이러한 한계를 해결하기 위해 두 개의 GAN을 쌓아

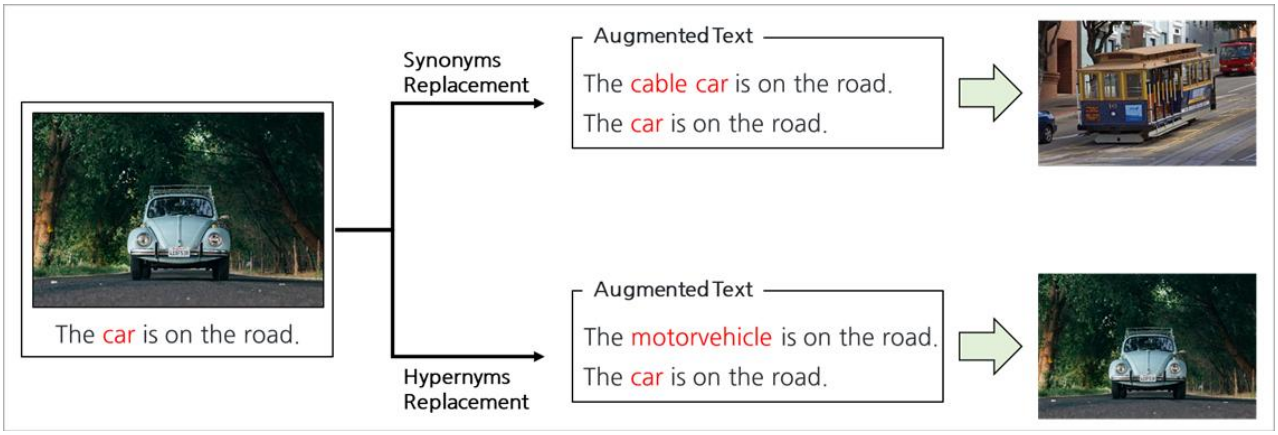

Fig. 2. Examples of Text Augmentation

두 개의 스테이지(Stages)를 구성한 StackGAN[10]이 고안되었다. StackGAN의 첫 스테이지는 입력 텍스트에 해당하는 객체를 스케치하고, 두 번째 스테이지는 첫 번째 스테이지에서 생성된 객체의 잘못된 부분을 수정하고 세부 정보를 추가함으로써 더 나은 고해상도 이미지를 생성한다. 한편, 텍스트를 글로벌 문장 벡터(Global Sentence Vector)로만 인코딩하는 경우 단어 수준에서의 세부 정보를 잘 활용하지 못한다는 한계를 해결하기 위해, 여러 개의 GAN을 쌓은 구조에 어텐션 메커니즘을 적용한 AttnGAN[11]이 고안되었다. AttnGAN은 입력 텍스트로부터 하위 영역의 이미지를 생성할 때, 해당 이미지와 관련된 단어에 더욱 주목하여 높은 가중치를 부여한다. 또한 생성된 이미지와 입력 텍스트 사이의 매칭 손실(Matching Loss)을 계산하여 더 나은 생성기 학습을 유도함으로써, 텍스트의 의미를 더욱 정확하게 반영하는 고해상도 이미지를 생성할 수 있다.

2. Data Augmentation

데이터 증강이란 인위적인 변화를 통해 데이터의 수를 증가시켜 학습에 필요한 충분한 수의 데이터를 확보하는 기법이다. 데이터 증강은 특히 이미지 데이터의 수를 늘리기 위해 널리 사용되었다. 구체적으로 이미지 데이터에 대한 Flipping, Color Space, Cropping, Rotation, Translation 등의 간단한 변형을 통해 이미지 데이터의 수를 늘리는 전통적인 방법뿐 아니라, 학습된 Feature Space에서의 변환, Neural Transfer 혹은 GAN을 활용한 새로운 데이터 생성 등 딥 러닝 기술을 적용한 데이터 증강 알고리즘들도 새롭게 제안되고 있다[12-14].

최근에는 이미지 데이터뿐 아니라 자연어 처리 분야에서도 데이터 증강 기법을 활용하려는 시도가 증가하고 있으며, 대표적인 연구로 어휘 대체 기반의 텍스트 데이터 증강이 있다. 이는 문장 내에 있는 단어를 유의어로 대체하는 기법으로 시소러스나 임베딩 모델 등을 사용한다. 시소러스 기반의 데이터 증강은 주로 워드넷에서 트리 구조로 정의된 유의어 사이의 관계를 사용하며, 이를 통해 문장에 포함된 일부 단어를 유의어로 대체함으로써 유사한 내용의 여러 문장을 생성한다. 하지만 워드넷과 같은 시소러스 기반 데이터 증강은 시소러스 구축에 상당한 비용과 시간이 소요될 뿐 아니라, 시소러스에 포함되지 않은 어휘를 처리할 수 없다는 한계를 갖는다. 한편 임베딩 모델 기반의 데이터 증강은 말뭉치에 대한 학습을 통해 문장에 포함된 단어의 벡터와 유사한 벡터를 갖는 단어를 찾는 방식으로, Word2Vec[15], Fasttext[16], 그리고 Glove[17] 등의 단어 임베딩 알고리즘을 통해 구현된다.

이외에도 기계 번역을 활용하여 원래 문장의 의미를 보존하면서 의역을 통해 다르게 표현된 문장을 추가하는

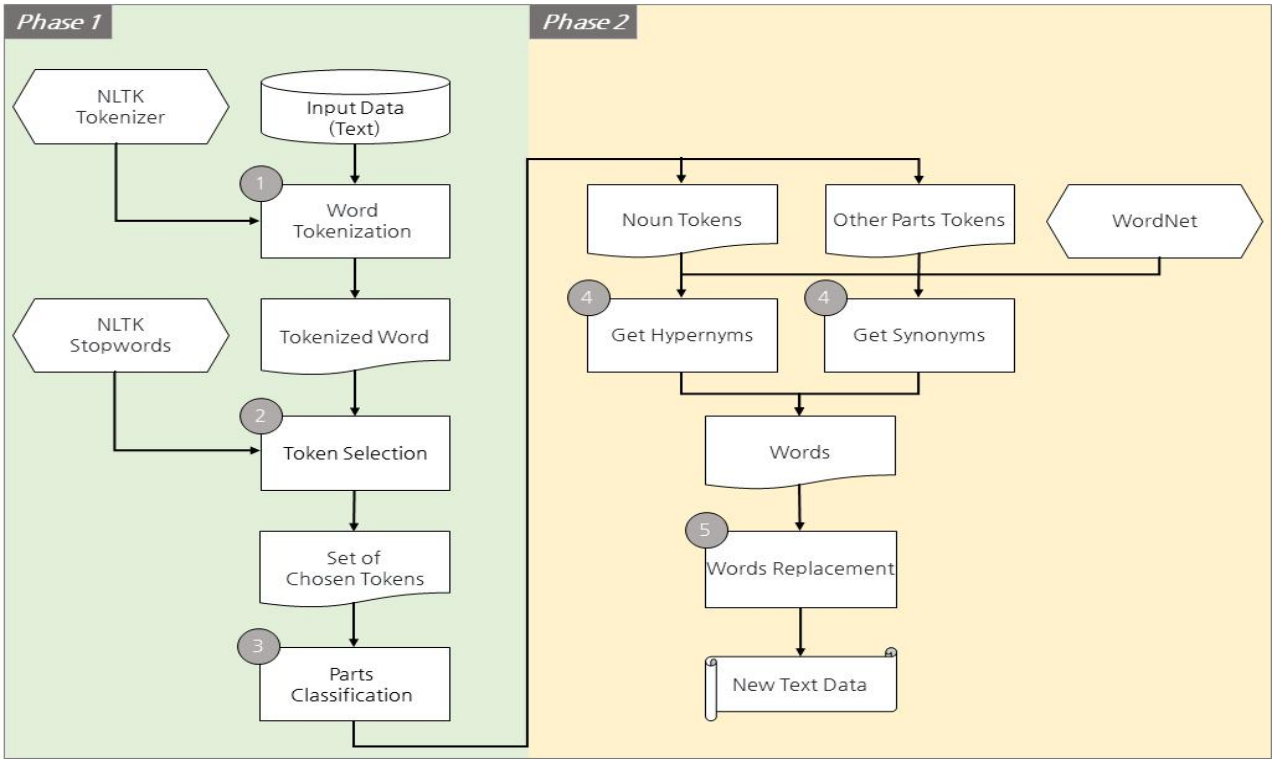

Fig. 3. Overall Research Process

역 번역 기반 증강, 그리고 BERT[18], GPT2[19] 등 대규모 사전 학습이 이루어진 언어 모델을 미세 조정하는 언어 모델 기반 증강 등 딥 러닝 기술을 활용한 텍스트 데이터 증강 연구가 활발히 수행되고 있다[20-22].

하지만 이러한 텍스트 데이터 증강 방법들은 단어들 간의 유의한 관계를 기반으로만 텍스트 증강이 이루어지고, 단어의 계층적 관계를 고려하지 못한다는 한계를 갖는다.

III. Proposed Method

1. Research Process

본 장에서는 본 논문에서 제안하는 단어의 의미 계층 기반 텍스트 증강 기법의 원리를 간단한 가상 예와 함께 설명한다. 제안 방법론의 전체적인 과정은 <Fig. 3>와 <Fig. 4>와 같다.

<Fig. 3>는 텍스트 증강 과정에서 토큰의 품사 정보와 계층적 관계를 활용하는 제안 방법론의 개요를 나타내며, <Fig. 4>은 제안 방법론을 알고리즘으로 표현한 것이다. 제안 방법론은 입력받은 텍스트 데이터를 토큰화(Tokenize)한 뒤 n 개의 토큰(Token)을 선택하여 품사를 정의하고 분류하는 Phase 1, 그리고 워드넷을 활용하여 품사에 따라 계층적으로 단어를 대체하는 Phase 2의 두 단계로 구성된다. 구체적으로 Phase 1은 입력받은 원본 텍스트 데이터를 (1) NLTK[23]의 토크나이저(Tokenizer)를 사용하여 토큰으로 분리하고, (2) 분리된 토큰들의 집합에서 NLTK의 불용어에 해당되지 않는 n 개의 토큰들을 선택한 뒤, (3) 선택된 단어들에 대해 품사를 정의하는 품사 태깅을 통해 명사와 다른 품사를 구분하는 작업을 수행한다. 이후 Phase 2에서는 (4) 선택된 토큰들의 품사가 명사이면 상위어를, 명사 외의 품사이면 유의어를 워드넷에서 추출하고, (5) 추출된 상위어 혹은 유의어를 (2)에서 선택된 토큰들과 대체하는 과정을 통해 최종적으로 새로운 텍스트 데이터를 생성하게 된다.

각 과정에 대한 구체적인 작동 원리는 본 장의 이후 절부터 가상의 예시와 함께 설명하며, 실제 데이터를 적용한 제안 방법론의 성능 평가 결과는 4장에서 소개한다.

```
def Preprocessing(Input_Text, Selection_N):

    Tokens = word_tokenizer(Input_Text)

    Tokens = [Token for Token in Tokens if Token not in Stopwords]
    Chosen_Tokens = random.sample(Tokens, Selection_N)

    for token in Chosen_Tokens:
        if pos_tag(token) in Noun_list:
            Noun_Tokens.append(token)
        else:
            Other_Tokens.append(token)

    return Noun_Tokens, Other_Tokens


def Word_Replace(Noun_Tokens, Other_Tokens):

    for noun in wordnet.synsets(Noun_Tokens):
        hyper_word = noun.hypernyms()
        replace_word = Input_Text.replace(Noun_Tokens, hyper_word)
        New_Text.append(replace_word)

    for other in wordnet.synsets(Other_Tokens):
        syn_word = other.hypernyms()
        replace_word = Input_Text.replace(Other_Tokens, syn_word)
        New_Text.append(replace_word)

    return New_Text


def main():

    Selection_N = 1
    Input_Text = "./MSCOCO.txt"
    Preprocessing(Input_Text, Selection_N)
    Word_Replace(Noun_Tokens, Other_Tokens)

main()
```

Fig. 4. Research Process Algorithm

2. Word Tokenization and Token Selection

본 절에서는 <Fig. 3>의 단계 중 입력받은 텍스트 데이터에 대한 토크나이징(단계 1), 그리고 분리된 토큰들 중에 n 개의 토큰들을 선택하는 과정(단계 2)을 소개한다. 단어 대체 기반의 텍스트 증강을 위해서는 문장 형태의 텍스트 데이터를 단어 단위로 분리하는 작업이 필요하다. 보통 텍스트 데이터 세트인 코퍼스에서 의미 있는 단위로 나누는 작업을 토큰화라고 하며, 토큰화를 거쳐서 나오는 산출물을 토큰이라 부른다. 토큰의 단위는 목적에 따라 다르지만, 본 연구에서는 단어와 동일한 의미로 사용된다. 텍스트를 의미 있는 단위로 나누기 위해 NLTK에서 제공하는 토크나이저를 사용하여 문장을 토큰으로 분리한다. 이렇게 분리된 토큰들 중에서 최종적으로 대체하고자 하는 단어의 개수만큼 선택하게 되며, 이때 NLTK의 영어 불용어 사전을 사용하여 불용어를 제외한 n 개의 토큰들을 Fig. 6에서처럼 랜덤하게 선택한다. <Fig. 5>는 이미지와 이미지에 해당되는 텍스트의 쌍으로 구성된 데이터의 예이며, <Fig. 6>는 이 데이터에 대해 토크나이저와 불용어 사전을 사용하여 입력 텍스트로부터 최종 토큰 집합을 구성한 파일럿 실험 결과의 일부이다.

Fig. 5. Example of Data

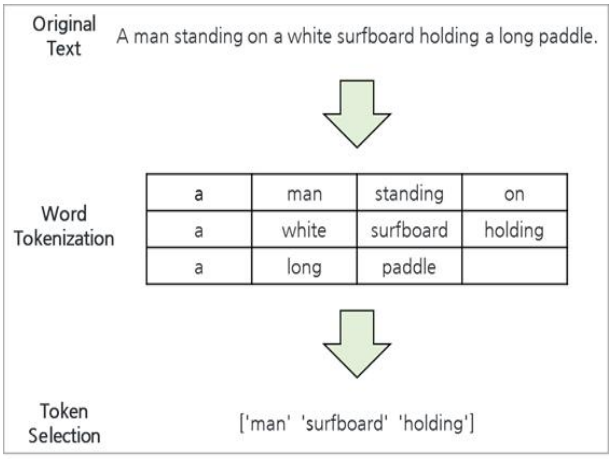

Fig. 6. Example of Word Tokenization and Token Selection

<Fig. 6>는 'a man standing on a white surfboard holding a long paddle'이라는 문장을 토큰으로 분리한 뒤, 최종적으로 'man', 'surfboard', 그리고 'holding'의 3개의 토큰을 선정한 결과를 보인다. 토큰의 선정은 불용어를 제외한 나머지 토큰 중 임의의 n 개 토큰을 선택하는 방식으로 이루어지며, 토큰의 수인 n은 하이퍼파라미터(Hyperparameter)로 직접 설정할 수 있다.

3. Token Classification

본 절에서는 <Fig. 3>에서 선택된 토큰들의 품사를 정의하는 품사 태깅과 명사와 다른 품사를 나누는 품사 구분 (단계 3)을 소개한다. 본 연구는 품사에 따라 토큰의 대체 방식이 달라지기 때문에 품사 태그가 굉장히 중요하다. 품사 태그의 정확도는 데이터의 크기나 단어의 수 등 상황에 따라 상이하게 나타나지만 일반적인 영어 코퍼스에서는 NLTK의 품사 태그 정확도가 약 97%로 나타난다 [24]. 본 연구에서는 높은 수준의 정확도를 보이고 있는 NLTK의 품사 태그를 사용하여 품사 식별을 진행한다. 그 결과 <Table 1>과 같이 'man'은 'NN', 즉 명사로, 'surfboard'는 'NN', 즉 명사로, 그리고 'holding'은 'VBG', 즉 동사로 품사가 식별된다. 이렇게 품사가 식별된 토큰들은 품사를 기준으로 2개의 카테고리로 분류되는데 하나는 명사 카테고리이고 다른 하나는 명사 이외의 품사 카테고리이다. 명사의 범위는 NLTK의 품사 태그 리스트에서 단수 명사인 'NN', 복수 명사인 'NNS', 단수 고유명사인 'NNP', 복수 고유명사인 'NNPS'로 지정한다. <Table 1>에서 세 토큰 중 'man'과 'surfboard'는 명사인 'Noun'으로, 그리고 다른 토큰은 명사 외의 품사를 나타내는 'Others'로 구분된 것을 확인할 수 있다.

Table 1. Example of Parts Classification

Token	man	surfboard	holding
POS	NN	NN	VBG
Class	*Noun*	*Noun*	*Others*

4. Word Replacement

본 절에서는 원본 텍스트의 단어들을 품사에 따라 상이한 방식으로 대체하는 <Fig. 3>의 (단계 4)와 (단계 5)를 소개한다. 명사와 명사 외의 품사로 구분된 토큰들은 워드넷의 단어 연관 정보를 바탕으로 단어 대체가 이루어지는데, 명사로 분류된 토큰은 상위어로, 명사 외의 품사로 분류된 토큰은 유의어로 대체한다. 이때, 유의어는 여러 개가 나올 수 있으며 본 연구에서는 임의추출로 토큰당 하나의 유의어를 선정하였다. <Fig. 7>은 품사에 따라 계층적으로 단어 대체가 이루어진 가상의 결과로, 명사인 'man'과 'surfboard'는 각각 상위어인 'person'과 'board'로, 그리고 명사가 아닌 'holding'은 유의어인 'keeping'으로 대체되었음을 확인할 수 있다.

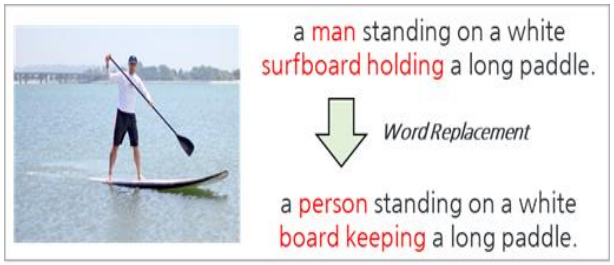

Fig. 7. Example of Word Replacement

IV. Experiment

1. Experiment Overview

본 장에서는 3장에서 소개한 제안 방법론을 실제 데이터에 적용한 결과 및 제안 방법론의 성능 분석 결과를 소개한다. 실험에는 MSCOCO[25] 데이터 세트를 사용하였다. MSCOCO는 객체 인식, 분할 그리고 이미지 캡셔닝 연구에 주로 사용되며 약 33만 건의 데이터가 공개되어 있다. 본 연구에서는 2014년에 공개된 데이터에서 이미지당 5개씩 부여된 텍스트를 사용하였으며, 텍스트 데이터 증강은 훈련용(Training) 이미지 데이터 약 8만 건에 부여된 텍스트에만 적용하였고 검증용(Validation) 이미지 데이터 약 4만 건에 부여된 텍스트에는 적용하지 않았다. 실험 환경은 Python 3.6과 NLTK 패키지를 바탕으로 구축하였으며, Text to Image 합성 모델은 Pytorch 기반으로 구현된 AttnGAN을 사용하였고, AttnGAN 모델에 다양한 텍스트 증강 방법을 적용하면서 실험을 진행하였다.

2. Pilot Experiment

본 절에서는 데이터의 양과 표현의 차이에 따른 성능을 가늠하기 위해 도형을 생성해내는 간단한 파일럿 실험을 진행하였다. <Fig. 8(a)>는 이미지 데이터 180건과 이미지 당 2개의 텍스트가 존재하는 데이터 세트를 구축한 결과의 일부이며, 이러한 데이터 세트로 학습하여 추론한 결과가 <Fig. 8(b)>이다. 결과를 보면 학습에 필요한 데이터가 충분하지 않아서 도형의 형태조차 제대로 생성하지 못한 것을 확인할 수 있다. 한편 <Fig. 9(a)>는 <Fig. 8(a)>에서 구축된 이미지 데이터 180개을 10,040개로 증강시킨 후 이미지당 6개의 텍스트가 존재하는 데이터 세트를 구축한 결과의 일부이며, 이러한 데이터 세트로 학습하여 추론한 결과가 <Fig. 9(b)>이다. <Fig. 8>에 비해 <Fig. 9>은 이미지와 텍스트 데이터의 양과 표현이 풍부해졌으며, 결과적으로 모양과 색상 측면에서 입력으로 주어진 문장에 부합하는 도형 이미지가 생성된 것을 확인할 수 있다.

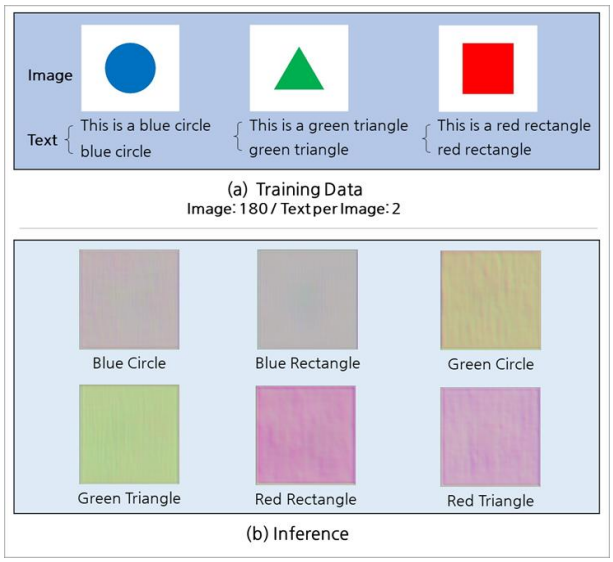

Fig. 8. Image Generation from Original Data

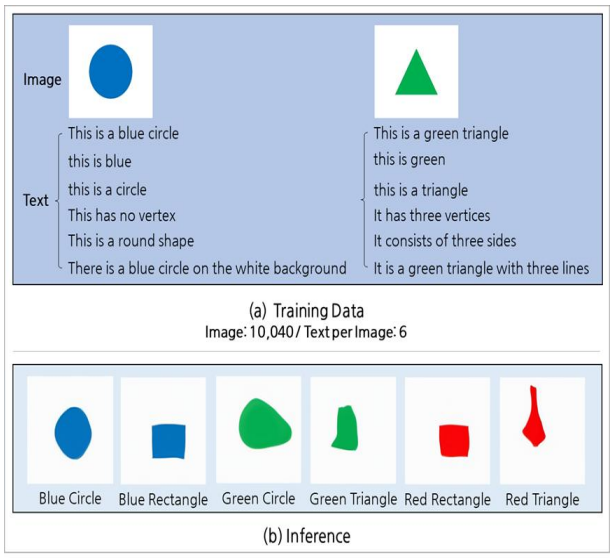

Fig. 9. Image Generation from Augmented Data

3. Preprocessing

본 절에서는 MSCOCO의 2014년 데이터에서 훈련용 이미지 약 8만 건에 부여된 텍스트를 NLTK 토크나이저를 사용하여 토큰 단위로 분할한 뒤, 대체하고자 하는 n개의 토큰들이 선택된 집합의 결과를 제시한다. <Fig. 10>은 훈련용 데이터에서 임의의 이미지와 해당 이미지에 부여된 다섯 개의 텍스트이며, <Fig. 11>은 텍스트를 소문자로 변환한 뒤, NLTK 토크나이저를 사용하여 토큰 단위로 분리한 결과를 보여준다.

① A flat screen and a keyboard and mouse on the desk
② a slim computer and keyboard sitting on a desk
③ A desktop with its monitor turned off on a table.
④ A computer desk with a monitor and keyboard on it
⑤ Flatscreen computer monitor and other electronics on a wooden desk.

Fig. 10. A Sample Image with Five Captions

①	a	flat	screen	and
	a	keyboard	and	mouse
	on	the	desk	
②	a	slim	computer	and
	keyboard	sitting	on	a
	desk			
③	a	desktop	with	its
	monitor	turned	off	on
	a	table		
④	a	computer	desk	with
	a	monitor	and	keyboard
	on	it		
⑤	flatscreen	computer	monitor	and
	other	electronics	on	a
	wooden	desk		

Fig. 11. Results of Word Tokenization

	Selected Token
①	screen
②	slim
③	monitor
④	keyboard
⑤	desk

Fig. 12. Selected Tokens

<Fig. 12>는 분리된 토큰들 중, NLTK에서 제공하는 Stopwords에 해당하지 않는 토큰을 각 문장별로 1개씩 임의추출한 결과이다.

4. Word Replacement

<Table 2>은 <Fig. 3>의 단계 (3)~(4)에 해당하는 과정의 결과, 즉 품사 식별 및 분류를 통해 명사와 다른 품사의 토큰들을 구분하고, 품사에 따라 선택적인 대체 기준을 적용하여 대체 단어를 찾은 결과를 보여준다. 대체하고자 하는 토큰들의 품사 태깅 결과, 'slim'은 'JJ', 즉 형용사로, 'screen', 'monitor', 'keyboard', 'desk'는 'NN', 즉 명사로 정의되었다. 명사로 식별된 단어들은 상위어로 대체하며, 명사를 제외한 품사들은 유의어로 대체한다. 대체 단어는 워드넷에서 탐색하여 추출하며, 유의어 및 상위어가 복수일 경우 임의로 하나만 추출한다. 그 결과 'slim'은 'slender'로 유의어가 추출되었고, 'screen', 'monitor', 'keyboard' 그리고 'desk'는 각각 'display', 'display', 'device', 그리고 'table'로 상위어가 추출된 것을 확인할 수 있다.

Table 2. Replacement Candidates for Each Word

Token	screen	slim	monitor	keyboard	desk
POS	NN	JJ	NN	NN	NN
Class	*Noun*	*Others*	*Noun*	*Noun*	*Noun*
Word	display	slender	display	device	table

<Table 3>는 선택된 토큰들을 추출된 단어들로 대체하여 각각의 텍스트들을 새로운 텍스트로 증강한 결과이며, 'Origin'으로 표기된 텍스트는 원본 텍스트, 'Augmentation'으로 표기된 텍스트는 제안 방법론을 통해 증강한 텍스트이며, 이와 같은 방식을 통해 이미지의 의미에 대한 왜곡 없이 증강한 텍스트를 이후 프로세스인 훈련에 활용한다.

Table 3. Augmented Text Captions

Num	Type	Text
①	Augmentation	a flat display and a keyboard and mouse on the table
	Origin	a flat screen and a keyboard and mouse on the desk
②	Augmentation	a slender computer and keyboard sitting on a desk
	Origin	a slim computer and keyboard sitting on a desk
③	Augmentation	a desktop with its display turned off on a table
	Origin	a desktop with its monitor turned off on a table
④	Augmentation	a computer desk with a monitor and device on it
	Origin	a computer desk with a monitor and keyboard on it
⑤	Augmentation	flatscreen computer monitor and other physics on a wooden table
	Origin	flatscreen computer monitor and other electronics on a wooden desk

5. Performance Evaluation

본 절에서는 품사에 따라 계층적으로 단어를 대체한 제안 방법론과 모든 품사를 동일 계층의 유의어로 대체하는 기존 텍스트 증강 방법의 성능을 분석한 결과를 소개한다. 전체 실험 프로세스는 <Fig. 13>과 같다.

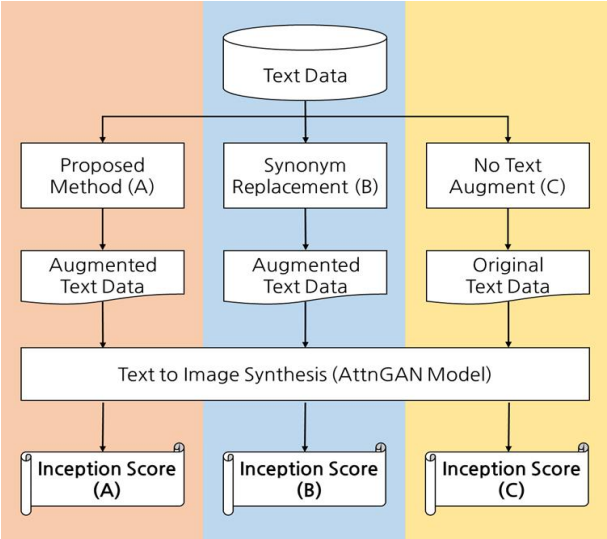

Fig. 13. Overall Process of Performance Evaluation

<Fig. 13>의 (A)는 제안 방법론을 통해 증강된 텍스트 데이터를 사용하여 Text to Image 합성을 진행하는 과정이며, (B)는 유의어로 대체하는 어휘 대체 기반의 증강 방법을 통해 증강된 텍스트 데이터를 사용하여 Text to Image 합성을 진행하는 과정이다. 마지막으로 (C)는 증강을 하지 않은 원본 텍스트 데이터를 사용하여 Text to Image 합성을 진행하는 과정이다. 최종적으로 Text to Image 합성에 사용되는 이미지당 텍스트의 개수는 각각 (A) 10개, (B) 10개, (C) 5개이며, 학습은 AttnGAN[11] 모델을 사용하여 진행하였다. 또한 텍스트로부터 생성된 이미지의 품질 평가에는 인셉션 스코어(Inception Score)[6]를 사용하였다. 인셉션 스코어는 생성 이미지의 품질을 평가하는데 주로 사용되는 평가 척도로 생성된 이미지의 품질과 다양성을 기준으로 측정되며, 인셉션 스코어가 높을수록 좋은 성능을 나타내는 것으로 해석할 수 있다. 세 가지 모델을 적용하여 생성한 결과 이미지에 대한 인셉션 스코어가 <Fig. 14>과 <Fig. 15>에 나타나있다.

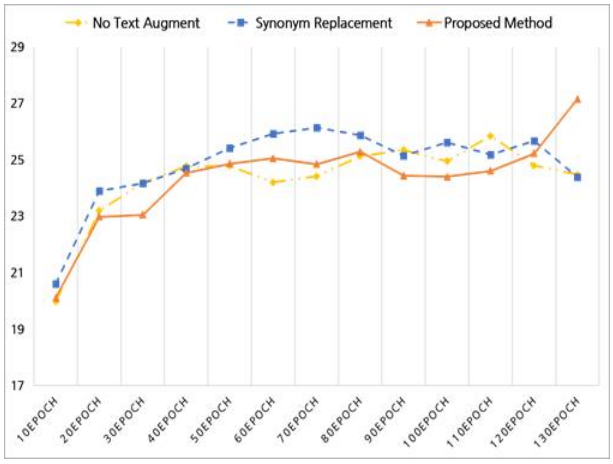

Fig. 14. Comparison of Inception Score

<Fig. 14>은 실험에서 비교한 세 가지 모델을 통해 생성한 이미지에 대한 인셉션 스코어를 10 에폭(Epoch) 단위로 나타낸 비교 그래프이다. 중반 에폭까지는 기존 텍스트 증강 방법을 사용한 모델 (B)와 증강을 하지 않은 모델 (C)의 인셉션 스코어가 제안된 방법론의 인셉션 스코어보다 다소 높게 나타나지만, 학습이 충분히 진행된 130 에폭에서는 제안 모델의 인셉션 스코어가 가장 높게 나타남을 확인할 수 있다. 가장 높은 인셉션 스코어를 나타내는 에폭은 각 모델마다 서로 다르며, 인셉션 스코어의 최댓값을 비교한 결과도 <Fig. 15>과 같이 제안 방법론의 성능이 가장 우수함을 나타내고 있다.

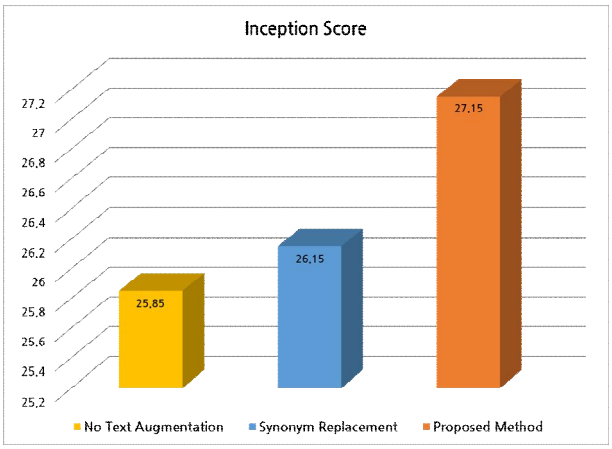

Fig. 15. Best Inception Score of Each Method

실험 결과 소량의 원본 데이터를 사용한 모델에 비해 텍스트 증강을 적용한 모델의 성능이 높게 나타났음을 확인하였다. 또한 텍스트 증강의 경우 본 연구에서 제안한 단어의 의미 계층 기반 텍스트 증강 방법이 기존의 유의어 대체 기반의 텍스트 증강 방법에 비해 우수한 성능을 나타냄을 확인하였다.

V. Conclusions

최근 이질적인 데이터를 딥 러닝 분석에 활용하는 멀티모달 딥 러닝에 대한 연구가 활발히 수행되고 있으며, 대표적으로 Text to Image 합성이 있다. 이미지 합성을 위한 딥 러닝 학습에 필요한 충분한 데이터를 확보하기 위해 다양한 연구들이 수행되고 있으며, 일반적으로 데이터 증강 기법이 많이 사용되고 있다. 이에 본 연구에서는 이미지와 텍스트가 쌍으로 존재하는 데이터에서 텍스트 데이터를 증강할 때, 이미지의 의미를 보존하면서 증강하기 위해 단어의 의미 계층 기반 텍스트 증강 기법을 제안하였다. 또한, 제안 기법을 통해 생성한 이미지가 기존의 유의어 대체 증강 기법을 통해 생성한 이미지에 비해 우수한 품질을 보임을 실험을 통해 확인하였다.

이미지 데이터 증강에 대한 연구가 다수 수행되고 있는 것에 비해 텍스트 데이터 증강에 대한 연구는 상대적으로 부족한 상황이며, 대부분의 텍스트 데이터 증강 연구는 유의어 대체에 기반을 두어 제한적으로 수행되고 있다. 본 연구에서는 단어가 갖는 품사별 특징을 활용하는 텍스트 데이터 증강 방안, 즉 선택적 품사에 대해 단어의 의미 계층 정보를 활용하는 방안을 새롭게 제시하였다. 또한, 제안 기법은 제한된 수의 이미지에 대해 더 많은 텍스트 정보를 생성함으로써, 이미지와 텍스트를 함께 다루는 멀티모달 딥 러닝 분석의 정확도 향상에 기여할 수 있을 것으로 기대한다.

본 연구의 추후 연구에서는 다음의 측면에 대한 고려가 이루어져야 한다. 우선 본 연구는 단어의 품사를 명사와 그 외의 품사로만 구분하여 방법론을 적용하였다. 추후 연구에서는 단어의 다양한 품사가 갖는 특징을 보다 적극적으로 활용할 필요가 있다. 또한 제안 방법론의 성능 평가를 위해 인셉션 스코어를 사용했는데, 추후 연구에서는 보다 다양한 관점에서의 성능 평가를 수행하여 제안 방법론의 성능을 확인할 필요가 있다.

ACKNOWLEDGEMENT

This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF)

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” Advances in Neural Information Processing Systems 27, 2014.

[2] S. Surya, A. Setlur, A. Biswas, and S. Negi, “ReStGAN: A Step towards Visually Guided Shopper Experience via Text to Image Synthesis,” 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar, 2020.

[3] C. Shorten and T. M. Khoshgoftaar, “A Survey on Image Data Augmentation for Deep Learning,” Journal of Big Data, No. 60, Feb, 2019.

[4] X. Zhang, J. Zhao, and Y. LeCun, “Character-level Convolutional Networks for Text Classification,” Advances in Neural Information Processing Systems 28, 2015.

[5] W. Y. Wang and D. Yang, “That’s So Annoying!!!: A Lexical and Frame Semantic Embedding-based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors Using #petpeeve Tweets,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2557-2563, Sep, 2015.

[6] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,”

Advances in Neural Information Processing Systems 29, 2016.
[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324, 1998.
[8] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of Recurrent Neural Network Language Model," Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5528-5531, 2011.
[9] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," arXiv:1605.05396, May, 2016.
[10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to Photo Realistic Image Synthesis with Stacked Generative Adversarial Networks," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5907-5915, 2017.
[11] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine Grained Text to Image Generation with Attentional Generative Adversarial Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1316-1324, 2018.
[12] T. DeVries and G. W. Taylor, "Dataset Augmentation in Feature Space," arXiv:1702.05538, Feb, 2017.
[13] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying Neural Style Transfer," arXiv:1701.01036, Jul, 2017.
[14] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks," arXiv:1810.10863, Oct, 2018.
[15] T. Mikolov, C. Kai, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, Jan, 2013.
[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," arXiv:1607.04606, Jul, 2016.
[17] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532-1543, 2014.
[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, May, 2019.
[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, Feb, 2019.
[20] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," arXiv:1904.12848, Jun, 2020.
[21] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Not Enough Data? Deep Learning to the Rescue!," arXiv:1911.03118, Nov, 2019.
[22] V. Kumar, A. Choudhary, and E. Cho, "Data Augmentation Using Pre-trained Transformer Models," arXiv:2003.02245, Mar, 2020.
[23] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," arXiv:cs/0205028, May, 2002.
[24] Y. Tian and D. Lo, "A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports," 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER), Mar, 2015.
[25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," European Conference on Computer Vision, pp. 740-755, 2014.

Authors

Museong Kim received the B.A. degree in Management Information Systems from Kookmin University in 2020 and currently enrolled in Graduate School of Business IT, Kookmin University.

Museong Kim is interested in natural language processing, image processing, multi-modal learning, and deep learning.

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively.

Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He is currently a dean of the Graduate School of Business IT at Kookmin University. He is interested in text mining, deep learning, and data modeling.