ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 23-04-27
194
j.inst.Korean.electr.electron.eng.Vol.27,No.4,561~566,December 2023
https : //doi.org/10.7471/ikeee.2023.27.4.561

이미지-텍스트 쌍을 활용한 이미지 분류 정확도 향상에 관한 연구

A Study on Improvement of Image Classification Accuracy Using Image-Text Pairs

김 미 희*★, 이 주 혁*

Mi-Hui Kim*★, Ju-Hyeok Lee*

Abstract

With the development of deep learning, it is possible to solve various computer non-specialized problems such as image processing. However, most image processing methods use only the visual information of the image to process the image. Text data such as descriptions and annotations related to images may provide additional tactile and visual information that is difficult to obtain from the image itself. In this paper, we intend to improve image classification accuracy through a deep learning model that analyzes images and texts using image-text pairs. The proposed model showed an approximately 11% classification accuracy improvement over the deep learning model using only image information.

요 약

딥러닝의 발전으로 다양한 컴퓨터 비전 연구를 수행할 수 있게 됐다. 딥러닝은 컴퓨터 비전 연구 중 이미지 처리에서 높은 정확도와 성능을 보여줬다. 하지만 대부분의 이미지 처리 방식은 이미지의 시각 정보만을 이용해 이미지를 처리하는 경우가 대부분이다. 이미지-텍스트 쌍을 활용할 경우 이미지와 관련된 설명, 주석 등의 텍스트 데이터가 이미지 자체에서는 얻기 힘든 추가적인 맥락과 시각 정보를 제공할 수 있다. 본 논문에서는 이미지-텍스트 쌍을 활용하여 이미지와 텍스트를 분석하는 딥러닝 모델 제안한다. 제안 모델은 이미지 정보만을 사용한 딥러닝 모델보다 약 11% 향상된 분류 정확도 결과를 보였다.

Key words : deep learning, image classification, data preprocessing, image preprocessing, text preprocessing

Ⅰ. 서론

기존의 딥러닝 기반 이미지 분류 연구는 대부분 이미지 데이터만을 사용하여 이미지의 특징을 추출하고 분류하는 데 초점을 두었다. CNN과 같은 기법을 사용하여 이미지의 픽셀값에서 패턴을 찾아내는 것이 주된 방식이었다[1].

하지만 이런 방식은 이미지 자체의 픽셀값만을 고려하기 때문에, 이미지에 연결된 중요한 문맥 정보나 메타데이터를 무시하게 된다. 예를 들어, 이미지에 포함된 사람

* School. of Computer Engineering & Applied Mathematics, Computer System Institute, Hankyong National University
★ Corresponding author
E-mail : mhkim@hknu.ac.kr, Tel : +82 31-670-5167
※ Acknowledgment
Manuscript received Nov. 21, 2023; revised Dec. 13 2023; accepted Dec. 26, 2023.

이나 물건의 이름, 이미지가 어디에서 촬영되었는지 등의 정보는 이미지 분류에 매우 중요할 수 있지만, 이런 정보는 이미지의 픽셀값에서는 나타나지 않는다.

이러한 한계점을 해결하기 위해 본 논문에서는 사용되지 않는 텍스트 정보를 함께 고려해 이미지와 텍스트를 결합하는 이미지-텍스트 쌍 방식의 딥러닝 분류 모델을 선행연구[2]에서 제안했다. 이미지-텍스트 쌍을 전처리 과정 없이 동시 입력 후 모델 학습을 진행해 이미지만을 사용한 분류 모델보다 높은 정확도를 보였다. 하지만 반복적인 텍스트인 경우에만 분류 정확도가 향상되는 단점이 있었다. 본 논문에서는 전처리 과정을 통해 이미지와 관련된 설명, 주석 등의 텍스트 데이터가 이미지 자체에서는 얻기 힘든 추가적인 맥락과 시맨틱 정보를 제공하여 분류 정확도를 높이고자 한다.또한, 이미지와 텍스트 모델을 강건화하고 다양한 기법을 적용해 정확도를 향상하고자 한다. 자세한 모델 설명은 3장 제안 모델에서 소개한다.

분류 정확도 향상이 가능한 이유는 텍스트 데이터가 이미지에서는 얻기 힘든 추가적인 맥락과 시맨틱 정보를 제공할 수 있기 때문이다[3].

제안 모델을 검증하기 위해 이미지의 시각 정보만을 활용한 모델과, 제안 모델의 분류 정확도 비교 실험을 진행하여 분류 모델의 유효성과 성능을 입증한다.

본 논문은 2장에서 제안 모델에서 사용된 기술들을 설명하고, 3장에서는 제안하는 분류 모델을 소개한다. 4장에서는 제안 모델의 실험 결과를 분석하고, 5장에서 결론을 맺는다.

II. 배경 지식

1. CNN(Convolutional Neural Networks)

CNN은 딥러닝 모델의 한 종류로, 주로 이미지 인식과 처리에 사용된다. CNN은 입력된 이미지에서 다양한 특징을 자동으로 학습하고 분류하는 능력을 갖추고 있다. 이런 특징은 계층적으로 구성되어 있어, 낮은 계층에서는 간단한 패턴을, 높은 계층에서는 복잡한 패턴을 학습하게 된다. CNN은 이미지를 그대로 사용하기 때문에, 이미지의 공간적 구조 정보를 유지하면서 학습할 수 있다[4]. 본 논문에서는 이러한 CNN의 특징을 이용해 이미지 데이터에 포함된 특징을 학습하고 학습된 정보들을 텍스트 데이터와 결합하여 이미지를 분류 정확도를 향상하고자 한다.

2. LSTM(Long Short-Term Memory)

LSTM은 순환신경망(RNN: Recurrent Neural Network)의 한 종류로, 시퀀스 데이터를 처리하는 데에 매우 효과적이다. LSTM은 기존 RNN의 단점인 장기 의존성 문제를 해결하기 위해 고안되었다[5]. 본 논문에서는 이미지 자체에서 얻기 힘든 추가적인 맥락과 시각 정보 등을 활용하기 위해 LSTM을 사용한다.

이 두 모델(CNN과 LSTM)은 각각 이미지와 텍스트를 처리하는 데 특화되어 있어, 이미지-텍스트 쌍을 처리하는 데 유용하게 사용할 수 있다. CNN은 이미지의 특징을 추출하고, LSTM은 텍스트의 시퀀스 정보를 처리하여 더욱 풍부한 정보를 제공함으로써 이미지 분류의 정확도를 향상하는데 기여할 수 있다.

III. 제안 모델

그림 1은 본 논문에서 제안하는 모델의 상위단계 설계이다. 모델 입력값은 이미지와 텍스트 데이터를 사용하며, 이미지와 텍스트 데이터는 각각 전처리 과정을 거친 후, a) image model과 b) text model로 입력된다. a)와 b)에서 각각 이미지와 텍스트 데이터의 특징을 학습한 후, 가중치를 결합하여 완전 연결 층으로 넘어간다. 마지막으로 학습된 가중치들을 바탕으로 이미지를 분류한다.

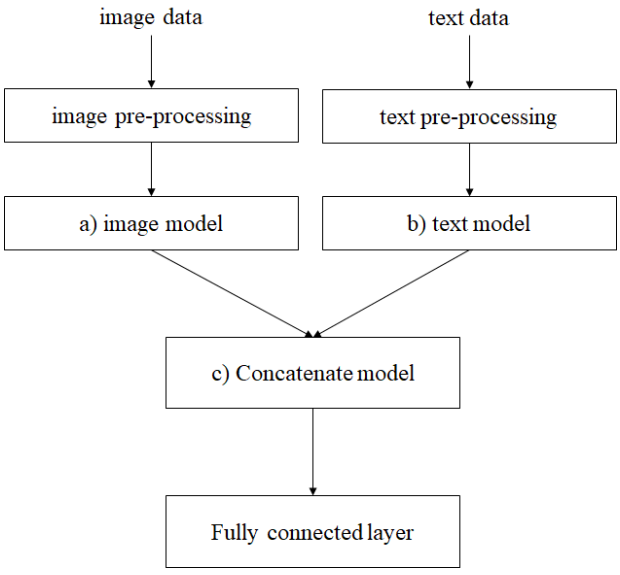

Fig. 1. High level design of proposed model.
그림 1. 제안 모델의 상위 단계 설계

1. 데이터 전처리

그림 1의 이미지 모델과 텍스트 모델에 입력하기 위해

서는 데이터 전처리 과정이 필요하다. 이미지 모델의 입력값은 CNN의 구조상 모두 같은 크기의 이미지를 사용해야 하기에 이미지 크기 조절 과정을 거쳐(64×64)의 크기의 이미지로 변환된다.

텍스트 데이터의 경우, 텍스트 정제, 토큰화, 어간 추출 및 표제어 추출, 정수 인코딩, 패딩 과정을 거친다. 텍스트 정제 과정은 불필요한 문자, 숫자, 특수 문자 등을 제거한다. 이 과정은 텍스트의 노이즈를 줄이고, 중요한 정보에 집중할 수 있게 한다. 토큰화는 텍스트를 더 작은 단위인 토큰으로 분리하는 과정으로, 토큰은 일반적으로 단어 또는 문장으로 정의된다. 어간 추출 및 표제어 추출은 단어를 기본 형태로 변환하는 과정이다. 표제어를 다양한 언어로 표현할 수 있다. 예를 들어, '수영하다'를 뜻하는 'swim'이 'swim', 'swam', 'swimming' 과같이 다양하게 표현될 수 있다. 다양한 표현을 통일하기 위해 표제어 추출 과정을 거친다. 표제어 추출은 문맥을 고려하여 단어의 표제어(사전 형태)를 찾는 과정이다. 토큰화 과정은 토큰화된 단어를 숫자로 변환하는 과정이다. 모델은 텍스트 데이터를 직접 다루지 못하므로, 각 단어를 고유한 정수로 변환하여 모델이 이해할 수 있도록 하는 과정이다. 패딩 과정이란 LSTM, 즉 순환신경망은 일정한 길이의 시퀀스를 입력으로 받기 때문에 모든 텍스트 데이터가 동일한 길이를 가지게 하는 과정이다. 마지막으로, 단어 임베딩은 정수 인코딩된 단어를 고차원의 벡터로 변환한다. 이 벡터는 단어 간의 의미적 관계를 포착한다.

2. image model

그림 2는 이미지 모델의 상세 설계이다. 이 모델은 CNN[6]의 구조를 가진다. CNN 구조를 통해서 이미지 내의 특징을 파악한다. 본 논문에 사용된 층으로는 convolution layer, maxpooling layer, flatten layer가 있다. convolution layer는 합성곱 연산을 통해, 이미지 분류를 위한 이미지 내의 특징을 파악한다. max-pooling layer는 이미지의 크기를 줄이면서, convolution layer에서 얻은 특징을 더욱 뚜렷하게 해준다. convolution layer와 maxpooling layer를 반복하며 얻은 결과를 텍스트 모델과 결합하기 위해 flatten layer를 적용해 1차원의 데이터로 변환한다. 변환하는 이유는 텍스트 데이터를 같은 차원으로 통일해야 하기 때문이다.

3. text model

그림 3 텍스트 모델은 LSTM[7] 구조를 가진다. 본 논문에서 LSTM은 이미지 분류 모델의 정확도 향상을 위한 텍스트 데이터를 얻기 위해서 사용한다.

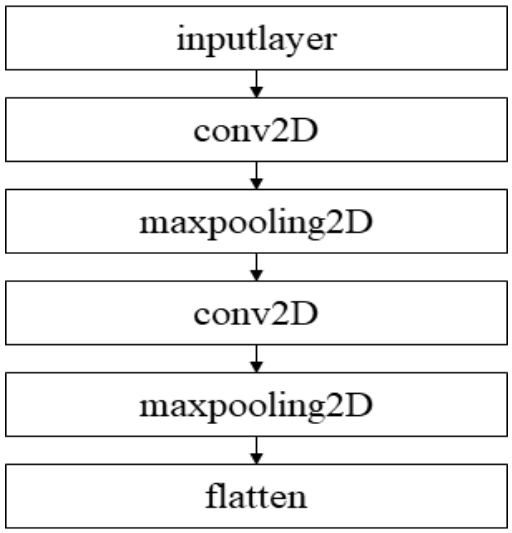

Fig. 2. Detailed design of image model.
그림 2. 이미지 모델의 상세 설계

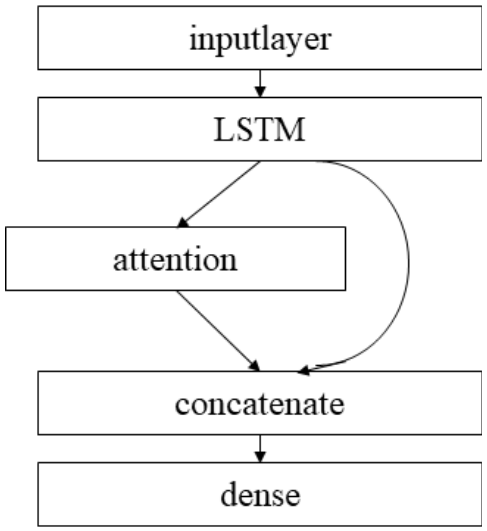

Fig. 3. Detailed design of text model.
그림 3. 텍스트 모델의 상세 설계

또한, LSTM 구조와 함께 attention[8]을 적용해 텍스트 데이터의 가중치를 훈련할 수 있도록 설계했다.

이미지에 대한 설명인 텍스트 데이터는 일련의 단어나 문장으로 구성된다. LSTM의 장기적인 의존성을 활용해 텍스트 데이터 중 이미지 설명에 효과적인 텍스트를 장기적으로 유지한다. 유지한 데이터 중 가중치를 부여하기 위해 attention을 통해 모델이 어느 부분에 집중해야 할지 결정을 돕는다. 또한, attention 기술을 사용함으로써, 모델이 어떤 텍스트에 주목하고 있는지와 모델의 텍스트 해

석 성능을 개선할 수 있다. 그 결과 LSTM과 attention 기술을 결합하여 긴 문장이나 문서에서도 문맥을 잘 이해할 수 있으며, 입력한 텍스트 데이터에서 주요한 부분을 모델이 파악해 더욱 정확한 예측을 할 수 있다.

4. 최종 분류 레이어

제안 모델은 이미지 데이터와 이미지 데이터를 설명하는 텍스트 데이터를 동시 입력하는 이미지 분류 모델이다. 이미지 모델과 텍스트 모델의 학습을 통해 얻은 데이터는 그림 4의 concatenate layer를 통해서 결합한다. 결합한 데이터들은 fully connected Layer로 구성된, 즉 concatenate layer 이후 층들을 통해서 최종 출력을 수행한다. 제안 모델은 마지막 층에서는 분류할 클래스 개수를 정할 수 있다. 그림 5처럼 이미지와 텍스트가 동시에 입력되며, 각각 그림 2와 그림 3을 거친다.

이미지와 텍스트 가중치를 이용해, 그림 4 모델의 결합을 통해서 이미지 분류를 통해 다중 분류를 실행한다. 더 나아가 문장으로 텍스트 전처리를 할 경우, 문장으로 분류도 가능하다.

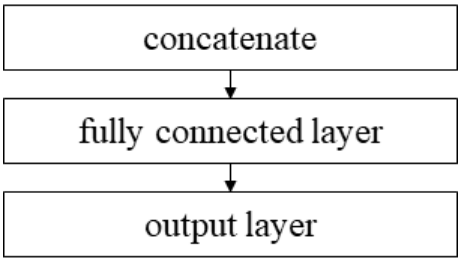

Fig. 4. Final classification layer.
그림 4. 최종 분류 레이어

그림 5는 이미지 분류 결과 예시이다.

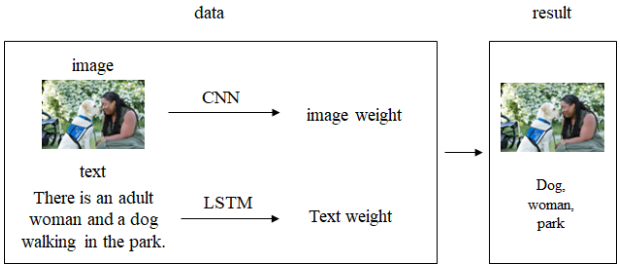

Fig. 5. Image classification results example.
그림 5. 이미지 분류 결과 예시

Ⅳ. 실험 결과 및 분석

실험 데이터는 COYO dataset[9]중 랜덤으로 선택된 이미지-텍스트 쌍을 5,000개를 사용했으며, 학습 데이터와 테스트 데이터는 8:2의 비율로 구성했다.

그림 6 훈련 모델들은 같은 학습 데이터와 테스트 데이터를 사용했으며, 학습률 0.001, 배치 크기 32, 에포크 100을 기준으로 학습했다.

모델 훈련 결과 이미지 데이터만 활용한 이미지 분류 모델인 CNN은 최고 정확도 약 0.8267을 보였다. [2]는 이미지-텍스트 쌍을 활용한 선행 연구이자 본 논문의 선행연구로, 정확도 약 0.8356을 보였다. 제안 모델은 최고 정확도 약 0.9315의 정확도를 보였다. 이미지 데이터만을 사용한 경우보다 약 11% 상승했다.

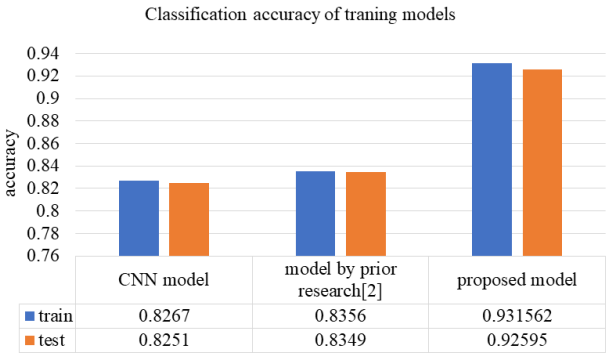

Fig. 6. Classification accuracy of training models.
그림 6. 훈련모델들의 분류 정확도

제안 모델은 다중 입력을 받는 딥러닝 모델로 이미지를 처리하여 클래스를 분류하는 모델이다. 이미지 처리 부분은 conv2d와 maxpooling2d layer를 통해서 이미지 처리한다. conv2d로 이미지의 특징을 추출하며, maxpooling2d로 차원을 줄이고 중요한 특징을 반복한 뒤 flatten layer를 통해서 1차원으로 변환한다. 텍스트 처리 부분은 LSTM과 attention 기술을 사용해 처리한다. 텍스트 데이터를 LSTM을 통해서 텍스트 데이터를 처리하며, attention 기술로 LSTM의 출력에 가중치를 할당하면서 더욱 중요한 텍스트 정보에 가중치를 둘 수 있게 된다. 마지막으로 결합 부분인 concatenate 층에서는 이미지 처리와 텍스트 처리의 결과값을 결합한다.

제안 모델이 정확도가 높아진 이유는 보다 많은 정보를 활용했기 때문이다. 이미지뿐만 아니라 텍스트 정보도 함께 활용하고 있기 때문에 이미지 정보만을 활용한 모델보다 더 많은 정보를 활용하여 예측을 수행한다. 또한 이미지와 텍스트 정보가 상호 보완적인 경우이기에 더욱 높은 성능 보인다.

Ⅴ. 결론

본 논문에서는 이미지-텍스트 쌍을 활용한 이미지 분류 정확도 향상에 관한 연구를 진행했다. 이미지-텍스트 쌍 데이터를 전처리 과정을 통해 이미지 분류 모델과 텍스트 모델에 입력한 다중 입력 모델을 구축해 이미지 데이터의 가중치와 텍스트 데이터의 가중치를 결합해 이미지 분류를 진행했다. 이미지 데이터만의 시각 정보에만 의존해 진행했던 기존 이미지 분류 방식과 비교하면 이미지 자체에서는 얻기 힘든 정보를 텍스트 데이터로부터 추가로 학습해 기존 기법과 비교해 더 높은 분류 정확도를 보였다.

향후 연구에서는 텍스트 전처리에 따른 이미지 분류 모델의 결과를 연구하고자 한다. 또한, 이미지 분류를 목표로 하는 것이 아닌 이미지-텍스트 쌍을 활용하여 객체 탐지, 이미지 세그먼테이션 등의 컴퓨터 비전 문제도 해결하고자 한다.

References

[1] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2*016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016.
DOI: 10.1109/CVPR.2016.90

[2] J. H Lee, M. H Kim, "Image classification model utilizing text to improve image classification accuracy," *Annual Conference of KIPS 2023,* p.4, 2023.

[3] J. Johnson, A. Karpathy and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4565-4574, 2016.
DOI: 10.1109/CVPR.2016.494

[4] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems 25 (NIPS 2012),* p.14, 2012. DOI:10.1145/3065386

[5] Sepp Hochreiter, Jürgen Schmidhuber, "Long short-term memory," *Neural computation, 9(8)*, pp.1735-1780, 1997.
DOI: 10.1162/neco.1997.9.8.1735

[6] Yann LeCun Leon Bottou Yoshua Bengio and Patrick Haner, "GradientBased Learning Applied to Document Recognition," *Proceedings of the IEEE,* 86(11), pp.2278-2324, 1998.
DOI: 10.1109/5.726791

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of theNorth American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp.4171-4186, 2019.
DOI: 10.48550/arXiv.1810.04805

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems 30 (NIPS 2017),* pp.15, 2018.
DOI: 10.48550/arXiv.1706.03762

[9] COYO-700M: Image-Text Pair Dataset "COYO dataset," [Internet], https://github.com/kakaobrain/coyo-dataset

BIOGRAPHY

Mi-Hui Kim (Member)

1997 : BS degree in Computer Science and Engineering, Ewha Womans University.
1999 : MS degree in Computer Science and Engineering, Ewha Womans University.
1999~2003 : Researchers at Switching & Transmission Technology Lab.(ETRI)
2007 : Ph.D. degree in Computer Science and Engineering, Ewha Womans University
2009~2010 : postdoctoral researcher of the department of computer science, North Carolina State University
2011~present : School of Computer Engineering & Applied Mathematics, Computer System Institute, Hankyong National University.

Ju-Hyeok Lee (Member)

2022 : BS degree in Computer Science and Engineering, Hankyong National University
2022~present : MS student in School of Computer Engineering & Applied Mathematics, Hankyong National University