Dual Supervision 을 이용한 이미지 객체 간 관계 추출

김민규 [1], 장민수 [1], 전희국 [2], 임동혁 [3]
[1] 광운대학교 인공지능응용학과 석사과정
[2](주)핀다
[3] 광운대학교 정보융합학부 교수

mkkim1678@kw.ac.kr, jjkj1026@kw.ac.kr, heegook@finda.co.kr, dhim@kw.ac.kr

Relation Extraction between Image Objects using Dual Supervision

Min-Kyu Kim[1], Min-Soo Jang, Hee-Gook Jun[2], Dong-Hyuk Im[3]
[1]Dept. of Artificial Intelligence Applications, Kwangwoon University
[2]Finda, Seoul, Korea
[3]School of Information Convergence, Kwangwoon University

요 약

비디오, 오디오, 이미지, 텍스트 등의 비정형 데이터는 데이터 구조가 없어 데이터 자체만으로는 내용에 대한 질의 처리가 힘들어 정형 데이터로 변환하는 과정이 필요하다. 관계 추출 작업은 문장 내 단어 간 속성 또는 관계를 예측하여, 문장을 구조적으로 표현한다. 자연어처리 기법인 Dual Supervision 모델은 인간이 레이블한 데이터와 기계가 레이블한 데이터를 기반으로 기존 모델보다 적은 리소스로 관계를 예측한다. 해당 자연어 처리 모델을 이미지 처리에도 적용하여 기존 방법보다 적은 리소스를 이용하여 이미지에 대한 내용을 구조적으로 나타내는 모델을 제안하였으며, 실험을 통해 효율적인 이미지 객체 관계 추출이 가능함을 확인하였다.

1. 서론

비정형 데이터는 다양한 산업 분야에서 가치 있는 정보를 제공하며, 비정형 데이터의 중요성은 계속해서 증가하고 있다[1]. 비디오, 오디오, 이미지, 텍스트 등 다양한 형태의 비정형 데이터는 자체만으로 내용에 대한 질의 처리가 어렵기 때문에, 이를 정형 데이터로 변환하는 과정이 필요하다. 이러한 변환 과정은 머신러닝과 인공지능 기술을 활용하여 진행되며, 이를 통해 비정형 데이터에서 유용한 정보를 추출하고 분석할 수 있다[2, 3].

특히 관계 추출 작업은 문장 내 단어 간 속성 또는 관계를 예측하여 문장을 구조적으로 표현하는데 중요한 역할을 한다. 하지만 대부분의 기존 모델들은 대량의 학습 데이터와 상당량의 컴퓨팅 리소스가 필요하다는 한계점이 있었다. 이에 본 연구에서는 Dual Supervision[4]을 도입하여 인간이 레이블한 데이터와 기계가 레이블한 데이터를 함께 사용함으로써 기존 방법보다 적은 리소스로 정확한 관계 추출을 가능하게 하는 방법론을 제안한다.

또한 본 연구에서 제안하는 방법론은 이미지 처리에도 활용될 수 있다. 이미지 내 객체 간의 상호작용과 관계 등 구조적인 정보를 파악하는 것은 컴퓨터 비전 분야에서 중요한 주제로 여겨져 왔으나, 많은 양의 학습 데이터와 컴퓨팅 리소스가 요구되었다[5]. 그러나 Dual Supervision 모델을 사용함으로써 이러한 요구사항을 크게 준수할 수 있으며, 결과적으로 이미지에 대한 내용도 구조적으로 나타낼 수 있게 되었다.

본 논문에서는 자연어 데이터의 관계 추출에서 사용되는 Dual Supervision 모델을 이미지 데이터 처리에도 적용하여, 비정형 이미지 데이터의 구조화 과정을 개선하였다.

2. 데이터 수집 및 전처리

본 연구에서는 Visual Genome 데이터를 활용한 실험을 진행한다. 이 데이터셋은 Knowledge Base 의 이미지 데이터셋으로, 객체, 속성, 관계 및 Scene 그래프 등으로 구성되어 있으며 이미지의 내용을 언어로 연결하는 기능이 있다[6]. 이 중에서 이미지 객체 탐지 작업에 사용될 단어를 제외한 나머지 객체와 해당 객체의 관계를 제거하여 실험을 수행한다.

Visual Genome 에는 총 7,699 개의 다양한 객체 종류가 포함되어 있으며, 이들 중에서 각각의 등장 횟수를 기준으로 상위 80 개의 객체 종류를 추출하여 실험에 활용한다. 전체 이미지 중 상위 80 개 객체가 포함된 71,204 개의 이미지를 선정하였다. 이들은 훈련, 검증 및 테스트 데이터로 사용하기 위해 7:2:1 비율로

분할하였다.

Visual Genome 이 제공하는 Scene 그래프 데이터는 개별 이미지의 객체 간 관계를 나타내며, 이것은 Knowledge Base 구축에 활용된다. Scene 그래프 데이터에서 불필요한 데이터를 제거하여 트리플 구조(<객체 1, 관계, 객체 2>)로 전처리한다.

각 이미지에서 바운딩된 객체들을 문자열로 나열하고, 이를 토큰 리스트 형태로 반환한다. 반환된 리스트를 바탕으로 POS(Part Of Speech) 태깅과 NER(Named Entity Recognition) 태깅 과정을 거쳐 각각 별도의 리스트로 생성한다[7].

개별 트리플 구조에서 객체 1 은 주어(Subject)로, 객체 2 는 목적어(Object)로 설정하고, 이들에 해당하는 토큰 리스트 내의 인덱스를 저장한다.

토큰 리스트에 대응하는 이미지의 고유 값(ID), 객체 간의 관계, 객체들의 인덱스, 태그들을 활용하여 학습에 사용할 최종 데이터프레임을 생성한다.

딥러닝 모델은 숫자형 데이터만 처리할 수 있기 때문에, 객체와 관계를 나타내는 텍스트 데이터를 변환하는 과정이 필요하다. 이를 위해 정수 인코딩 및 워드 임베딩 작업을 수행한다. 또한 워드 임베딩 작업을 통해 단어들 사이의 의미적 관계를 학습할 수 있으므로, 모델의 성능을 개선할 수 있다.

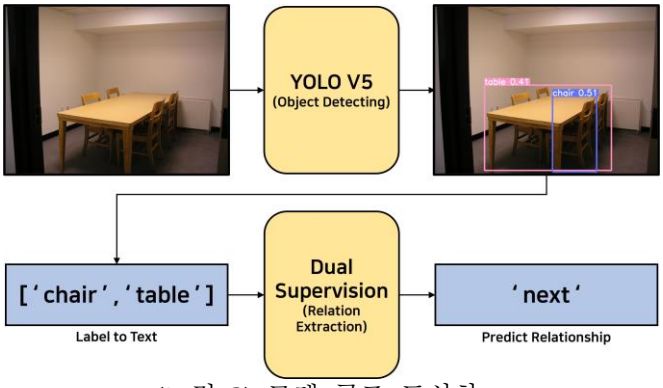
이번 연구에서는 사전 학습된 GloVe[8]를 사용하여 워드 임베딩을 진행한다. 객체와 관계를 나타내는 텍스트 중 GloVe 에 학습되지 않은 텍스트는 이미 학습된 텍스트 중 동의어로 치환하여 임베딩 작업을 진행한다.

3. 모델 설계 및 학습

이미지 객체 탐지 모델로는 YOLO v5[9]를 사용한다. Visual Genome 의 객체 바운딩 박스는 좌상단 위치의 x, y 값과 바운딩 박스의 너비, 높이로 구성되어 있다. 그러나 YOLO 모델 학습에 사용되는 좌표는 바운딩 박스의 중심점 x, y 와 바운딩 박스의 너비, 높이로 구성되므로 기존 데이터셋의 좌표를 변환해야 한다. 기존 x 좌표와 너비를 더한 후 2 로 나누어 바운딩 박스의 가로 중앙 좌표를 계산한다. 계산된 가로 중앙 좌표를 원본 이미지의 너비로 나누면 YOLO x 좌표가 된다. y 좌표도 같은 방식으로 계산한다. 너비와 높이는 기존 너비와 높이를 각각 원본 이미지의 너비와 높이로 나누어 좌표 변환 작업을 완료한다.

변환 완료된 데이터를 입력으로 한 모델 수행 과정에서, 탐지된 객체 레이블(그림 1)을 텍스트로 변환한 후, 이를 Knowledge Base 에서 학습된 Dual Supervision 모델에 입력한다. 입력된 객체 간 적합한 관계를 추출하여 결과값을 도출한다(그림 2).

(그림 1) YOLO 이미지 객체 탐지 결과 예시



(그림 2) 모델 구조 도식화

4. 실험 및 결과

앞서 훈련, 검증, 테스트로 나눠진 이미지를 YOLOv5m 모델을 이용하여 배치 사이즈 64, 총 학습 Epoch 20 으로 학습하여 이미지 객체 탐지를 진행한다.

이번 연구에선 Dual Supervision 의 문장 모델을 이용하여 실험을 진행한다. Feature Encoder 로 BiGRU[10]를 사용하여 입력 텍스트에서 주어(Subject)와 목적어(Object)를 추출한다.

Visual Genome 으로 생성된 최종 데이터프레임은 이미지 고유값(ID)에 따라 훈련, 검증, 테스트 데이터프레임으로 분할하고, 이 분할된 데이터를 이용해 Dual Supervision 모델을 학습한다(Batch size: 64, Epoch: 30, Learning Rate: 0.7, Learning Decay: 0.9).

<표 1> Visual Genome 모델 평가 지표

	Precision	Recall	F1-Score
Visual Genome(Image)	0.4575	0.3809	0.4140
KBP(Sentence)	0.6724	0.4469	0.5288

Dual Supervision 모델을 사용한 이미지 데이터(Visual Genome) 처리 결과는, 자연어 데이터(KBP[11])로 진행한 결과와 유사함을 확인하였다. 이를 통해 이미지에도 해당 모델을 적용할 수 있으며, 효율적인 이미지 객체 관계 추출이 가능함을 확인하였다.

5. 결론

본 연구에서는 컴퓨터 비전 분야의 이미지 관계 추출 모델 학습 과정에서 많은 리소스가 필요한 문제를 해결하기 위해, 기존 텍스트에서 사용되는 관계 추출 모델인 Dual Supervision 을 적용하였다. 이를 통해 적은 리소스로도 이미지 관계 추출이 가능함을 확인하였다. 앞으로의 연구에서는 탐지된 객체의 바운딩 박스 간 거리를 참고하여 토큰 리스트를 생성 후 학습하는 방법에 대해 연구할 계획이다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음 (IITP-2023-2018-0-01417).

참고문헌

[1] 윤희근, 최수정, & 박성배, 의미 유사도를 활용한 Distant Supervision 기반의 트리플 생성 성능 향상, 정보과학회논문지, vol. 43, no. 6, pp. 653-661, 2016.

[2] Zaman, G., Mahdin, H., Hussain, K., and Rahman, A. U., Information extraction from semi and unstructured data sources: a systematic literature review, ICIC Express Letters, vol. 14, no. 6, pp. 593-603, 2020.

[3] Seongyong Kim, Tae Hyeon Jeon, Ilsun Rhiu, Jinhyun Ahn, and Dong-Hyuk Im, Semantic Scene Graph Generation Using RDF Model and Deep Learning, Applied Sciences, vol. 11, no. 2, 826, 2021.

[4] Woohwan Jung, and Kyuseok Shim, Dual Supervision Framework for Relation Extraction with Distant Supervision and Human Annotation, Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 2020, pp. 6411-6423.

[5] Saleh, Alzayat, Marcus Sheaves, and Mostafa Rahimi Azghadi, Computer vision and deep learning for fish classification in underwater habitats: A survey, Fish and Fisheries, vol. 23, no. 4, pp. 977-999, 2022.

[6] Krishna, Ranjay, et al, Visual genome: Connecting language and vision using crowdsourced dense image annotations, International journal of computer vision, vo. 123, pp. 32-73, 2017.

[7] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, 2014, pp. 55–60.

[8] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543

[9] Glenn Jocher, ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Zenodo, Nov. 22, 2022.

[10] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning, Position-aware attention and supervised data improve slot filling, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 35–45.

[11] J. Ellis, J. Getman, J. Mott, X. Li, K. Griffitt, S. M. Strassel, and J. Wright. Linguistic resources for 2013 knowledge base population evaluations. Text Analysis Conference (TAC), Maryland, USA, 2014.