학습데이터 확보를 위한 이종정보 추출기반 데이터 확장 방법 연구

정종진, 박종빈, 이한덕

전자부품연구원

mozzalt@keti.re.kr, jpark@keti.re.kr, mdbow81@gmail.com

A StudyResearch on data extension method based on extracting heterogeneous information to enhance learning data

Jung Jong Jin, Park Jong Bin, Lee Han Duck

Korea Electronics Technology Institute

요 약

본 논문은 인공지능 분석에 필요한 양질의 학습데이터를 확보하고 다양한 각도에서 데이터 분석을 할 수 있도록 임의의 원본(Seed) 데이터가 함의하고 있는 의미를 찾고, 원본(Seed) 데이터 내 포함된 다양한 형상(이미지, 텍스트, 오디오, 비디오 등)이 다른 이종의 데이터를 추출하고 이들의 메타 정보를 파악하여 연관성이 높은 이종의 데이터를 확장해 가는 데이터 확장 기술을 소개한다. 확보된 데이터로부터 이종 정보를 추출하고 데이터가 갖는 의미(메타) 정보를 증식하여 연관 분석에 활용하거나 기존 데이터의 확장 메타 데이터(레이블)로 활용하는 이종 정보 활용기반 데이터 증식에 관한 연구를 포함한다.

Ⅰ. 서 론

4차 산업혁명 시대 데이터가 모든 산업의 발전과 새로운 가치 창출의 촉매 역할을 하는 '데이터 경제(Data Economy)'패러다임의 전환 과정에서 발생하는 대량의 데이터가 데이터 기반 산업·경제 활성화를 견인하는 원동력으로 작용할 전망이다. 세계 주요 선진국 및 기업들은 데이터 경제 선도를 위한 범부처 차원의 주요 데이터의 확보, 활용 확대, 분석 인재양성 등 빅데이터 산업 활성화 대책 마련·추진 중이다. 이런 가운데, 분석 대상 데이터를 스스로 이해하고 데이터 간 유기적 관계 및 상호 연관성을 기계적으로 파악하여 데이터를 확보할 수 있는 연구가 많은 관심을 받기 시작하였다. 따라서 본 논문에서는 빅데이터 또는 인공지능 분석에 필요한 학습데이터를 확장하기 위해, 데이터 내에 포함된 이종의 정보를 추출하고, 이들 정보 간의 상호 연관성을 파악하여 다양한 분야의 데이터 융합을 통해 양질의 데이터를 생성하는 연구에 대해 소개한다.

Ⅱ. 이종정보 기반 데이터 증식 개요

1. 개념

앞서 서론에서도 설명한 바와 같이 이종정보 기반 데이터 증식은 그림 1과 같이 임의의 Seed 데이터 내에 포함된 이종(異種)의 정보 또는 데이터들 간의 상호연관성을 파악하여 이종 데이터까지도 융합하여 양질의 학습 데이터를 확장해 가는 개념이다. 이종 증식이란 개념은
확보한 데이터로부터 정보를 파악하여 이와 유사한 의미를 지닌 다른 형태의 데이터의 양을 확장해 가는 개념이다. 예를 들어 이미지 데이터로부터 텍스트 정보를 추출하거나 비디오 데이터에서 음성 정보를 추출하여, 본래의 이미지나 비디오 뿐만 아니라 다른 종류의 정보를 활용해가며 확장해 가는 개념이라 할 수 있다.

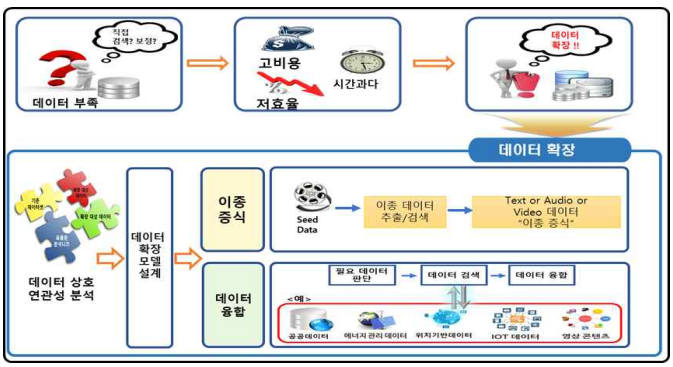

그림 1. 데이터 확장(이종증식 또는 융합검색 개념도

2. 기존연구

학습데이터의 양을 확대를 위한 기존연구중 하나인 GAN(Generative Adversarial Networks)은 2014년 Ian Goodfellow가 발표하였고, 현재는 초기 GAN 보다 더욱 많은 모델들이 개발되고 있다[1]. 이중 2017년 네이버 Clova는 StarGAN이라는 모델을 개발, 인물 사진 한 장으로 인물의 머리색, 표정, 성별 등을 변경하는 이미지를 생성 가능하다. 네이버 웹툰의 '마주쳤다'는 gan기술을 이용, 독자가 자신을 촬영하면 주인공의 얼굴이 독자의 얼굴이 웹툰화 되어 적용되는 기술을 사용 하지만 이는 동종 데이터 증식에만 한정되어 있어 다양한 분석 니즈에 필요한 빅데이터 복합 분석에는 다소 부족한 측면이 있다.

Ⅲ. 이종정보 기반 데이터 시스템 설계

1. 시스템 설계

기존연구에서 설명한 GAN기반 데이터 증식 과정은 일종의 임의의 데이터를 대상으로 임의의 랜덤한 변화를 학습에 사용할 데이터양을 늘리는 과정으로서 동종증식이라 칭할 수 있다. 주로 이미지 인식용으로 주로 활용되지만, 다양한 종류의 데이터를 활용한 융합 분석에는 효율적이지 못한다. 이에 반하여 이종 증식과정은 임의의 데이터, 즉 Seed 데이터에서

다른 형태의 정보를 추출하고, 추출된 정보들을 바탕으로 다양한 방식으로 증식을 함으로써 기존 데이터셋 내에 특정 데이터와 동일한 의미를 갖되 서로 다른 종류의 데이터들이 확장되어 가는, 즉 데이터량의 깊이를 보강하는 일종의 복합적인 데이터의 확장방식이라 할 수 있다. 그림 2는 이종 증식 과정을 설명하고 있다.

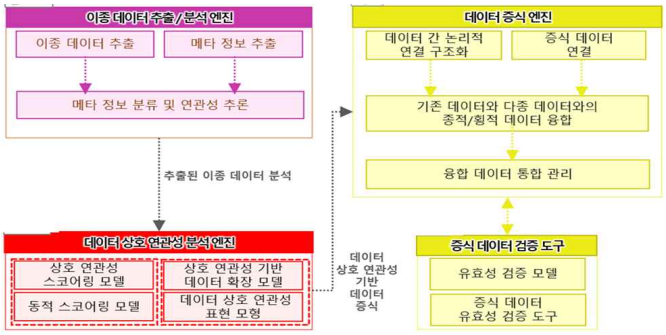

그림 2. 데이터 이종 증식 과정

그림 2에서 볼 수 있듯이, Seed데이터로부터 형태가 다른 이종데이터를 추출하고 연관성을 부여하기 위해서는 Seed에 포함된 이종 데이터와 이에 대한 특징을 기술하는 메타정보를 추출 한 뒤, 이들의 연관성을 추론한 뒤 Seed와 추출된 정보간 관계를 파악하고 관리한다. 그 뒤 자신이 확보중인 데이터 저장소 또는 접근 가능한 다른 곳의 저장소를 탐색하여 추출된 이종정보들과 연계가 있는지에 대한 분석 과정을 거친다. 연관성을 분석하기 위해서 분석 목적에 맞는 상호연관성 스코어링 모델을 적용하고, 이를 기반한 동적 스코어링 모델의 결과에 따라서 연관성이 부여되고 나면 이종정보 유형별로 적합한 증식 모형을 거쳐 실질적인 데이터의 양적 확대가 이루어진다. 데이터 증식엔진을 거치고 난 뒤 연관성 스코어에 따라 다량의 증식된 데이터들이 모여지는데 이 결과가 모두 품질이 높다고 할 수 없다. 따라서 마지막 단계에서는 증식된 데이터들이 분석 목적에 적합한 데이터셋을 이룰 수 있는지 등에 대한 유효성 검증을 거쳐 최종적으로 데이터들이 확장 가능하다.

2. 연관성 분석 모델

증식 대상이 되는 데이터들은 다양한 연관성 관계를 가지고 존재 가능하다. 즉 현존하는 데이터 연결 모델 즉, LOD 기반, 데이터 맵 기반, DCAT 기반, Schema.org 기반 등 다양한 형식에 따라 데이터간 연결성을 갖는데 이러한 모델들을 호한 가능토록 데이터 연관성 분석이 가능해야 한다[2]. 그림 3은 이러한 연결관계를 고려한 데이터 확장 기법 예시를 설명한다.

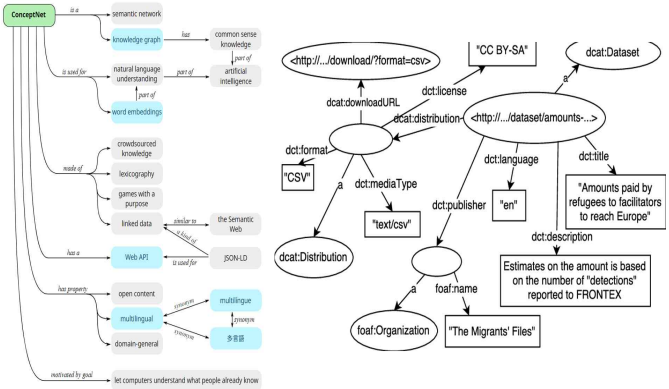

그림 3 데이터 확장 기법(LOD, Data Map, DCAT) 예시

3. 이종 추출 데이터 및 메타 정보 구조

하나의 Seed 데이터들로부터 추출된 다양한 형식의 이종데이터들은 서로 연결성을 가지고 있어야, 하나의 semantic에서 파생된 이종추출 데이터들임을 알 수 있다. 본 논문에서는 하나의 이미지 Seed 데이들에서 추출된 이종 정보를 대상으로 메타 정보 클래스를 정의하는 과정을 다음 그림 4와 같이 실험을 하였다. 그림 4에서 살펴보면 이미지 분류 모델의 성능 평가를 위한 표준 평가데이터인 ImageNet-1k (1000개 이미지 클래스) 데이터 셋을 포함한 다음의 다양한 범용 이미지 데이터 셋을 참고하여 이미지 메타 정보 클래스 정의가 가능하다[3].

그림 4 EIC 데이터 셋에 정의되어 있는 이미지 메타 정보 예
* EIC Dataset (ImageNet의 확장버전으로 약 3,000개의 메타 정보 클래스를 가짐)

Ⅲ. 결론

본 논문에서는 4차 산업혁명 시대 새로운 제품 서비스 개발에 필수인 자율차·스마트 시티 등 영역별 실제 데이터(Real Data)*와 AI 학습용 데이터는 양적으로 부족하여 양질의 데이터를 쉽고 빠르게 확보 가능한 이종데이터 증식에 대한 개념과 이를 실현하기 위한 핵심 기술에 대해 살펴보았다. 현재는 주로 활용되는 학습데이터 양적 확대 방법은 주로 동종 데이터만을 대상으로 한 데이터 증식이 주를 이루고 있어 복잡한 데이터 분석 수요 요구를 충족하고 있지 못하고 있어 이종데이터를 대상, 타 플랫폼을 대상으로 한 데이터 확장이 가능하리라 판단된다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2020-0-0062, 이종 정보 활용 및 데이터 융합을 통한 데이터 증식 기술 개발)

참 고 문 헌

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, Ozair, Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672- 2680).

[2] 박정현, 류승택, 박진완, “LOD(Level-of-Detail)를 통한 카툰렌더링 효과”, 2017, pp. 26-32

[3] Wavelet 기반 LOD 가상객체 표현 시스템. 김기호, 유황빈. 한국정보처리학회 논문지 제7권 제3호(2000.3), pp.766-775