유해화학물질 분류를 위한 AI 학습 모델에 관한 연구

심보리, 김연진, 전윤주, 유성민, 조숙경*, 김경배

서원대학교, *(주)네스토즈

{bori0725, rladuswls13, jo394}@seowon.ac.kr, smin9830@gmail.com, skyindb@naver.com, gbkim@seowon.ac.kr

A Study on AI Learning Modes for the Classification of Hazardous Chemical Substances

Sim Bo Ri, Kim Yeon Jin, Jeon Yoon Ju, Cho Sook Kyoung*, Kim Gyoung Bae

Seowon Univ., *Nestors Co. Ltd.

요 약

유해화학물질의 유통량 증가로 사고 발생이 해마다 증가함에 따라 사고 수습과 대응을 위해 신속한 유해화학물질 판독이 중요해지고 있다. 본 논문에서는 유해화학물질 사고 영상과 이미지를 이용하여 유해화학물질 분류에 적합한 인공지능 학습 모델을 제시한다. 이를 위해 공개 소프트웨어인 오렌지를 사용하여 'Logistic Regression', 'Random Forest', 'Neural Network', 'SVM' 등의 머신러닝 알고리즘을 적용해 7가지 유해화학물질 이미지를 분류하고, 모델별 성능 지표 및 분류 정확도를 평가하였다. 그 결과로 'Logistic Regression'의 분류 정확도가 가장 높았으며, 학습 데이터의 특성에 따라 인터넷에서 수집된 물질 이미지와 발현 특성이 유사한 물질의 경우 분류 정확도가 저하되는 것으로 나타났다.

Ⅰ. 서 론

유해화학물질 관련 사고는 유해화학물질의 유통량이 늘어나면서 따라서 관련 사고가 해마다 증가하고 있다[1,2]. 유해화학물질 사고의 특성상 유해화학물질의 종류가 사고 수습 및 대응에 있어서 가장 중요한 정보가 되기에 물질에 대한 신속하고 정확한 판독이 요구된다[2]. 이러한 문제를 해결하기 위해 현장에 출동하는 소방관들에게 신속하고 올바른 초기대응을 지원하기 위해 소방청은 과학기술정보통신부와 정보통신산업진흥원 주관으로 "AI융합 유해 화학물질 판독시스템" 과제를 수행 중에 있다[3]. 해당 시스템은 AI를 활용하여 화학물질의 종류와 사고 유형을 판독하는데, 이를 위해 이미지 데이터가 활용된다. 이 데이터는 물질별로 직접 수집된 이미지, 인터넷 상에서 수집된 이미지, 그리고 소방청에서 제공한 이미지로 구성되어 있다.

인공지능을 기반으로 한 유해화학물질 판독을 위해서는 유해화학물질 사고 영상과 이미지를 수집하고, 데이터에 대한 전처리를 수행하여 학습에 적합한 형태로 학습용 데이터를 구축해야 한다. 특히, 유해화학물질 사고에 신속하고 정확한 판독을 위해 인공지능 시스템을 개발하는 과정에서 가장 중요한 요소 중 하나는 적절한 학습 데이터를 기반으로 한 유해화학물질 분류 모델이다. 이 모델은 정확한 화학물질의 식별과 분류를 담당하며, 이를 위해 구축된 학습 데이터를 효과적으로 활용해야 한다.

본 논문에서는 AI를 활용한 유해화학물질 판독 시스템을 개발하기 위해 유해화학물질 분류에 적합한 인공지능 학습 모델을 연구하였다. 이를 위해 파이썬 등의 코드 및 AI 알고리즘에 대한 전문 지식이 부족한 비전공자 및 일반인들도 쉽게 활용할 수 있는 공개 소프트웨어인 오렌지(Orange)를 활용하였다[4]. 유해화학물질 이미지 데이터를 오렌지에서 'Logistic Regression', 'Random Forest', 'Neural Network', 'SVM' 머신러닝 알고리즘에 학습시키고 유해화학물질 이미지 분류 모델링을 실시하였다. 학습 데이터셋과 테스트 데이터셋을 이용하여 각 알고리즘의 특성 및 성능 비교와 유해화학물질 분류의 정확도를 평가하였다.

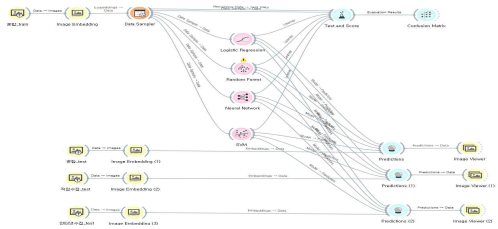
Ⅱ. 본 론

유해화학물질 이미지 데이터는 유해화학물질 직접 실험을 통한 직접수집과 인터넷수집으로 취득할 수 있다. 본 논문에서는 오렌지를 이용하여 유해화학물질 이미지 분류를 위해 염소, 질산, 브로민, 등유, 염화제이구리, 톨루엔, 메탄올 7개의 물질에 대하여 학습 데이터와 테스트 데이터로 나누어 이미지 모델링 작업을 수행하였으며 직접수집 이미지와 인터넷수집 이미지, 그리고 직접수집 이미지와 인터넷수집 이미지를 혼합하여 데이터셋을 구성하였다. 사용한 학습 데이터셋, 테스트 데이터셋의 정보는 [표1]과 같다.

순번	train data (이미지 혼합)		순번	test data 1 (이미지 혼합)	
	물질명	이미지(개)		물질명	이미지(개)
1	염소	1,000	1	염소	200
2	질산	1,000	2	질산	200
3	브로민	1,000	3	브로민	200
4	등유	1,000	4	등유	200
5	염화제이구리	1,000	5	염화제이구리	200
6	톨루엔	1,000	6	톨루엔	200
7	메탄올	1,000	7	메탄올	200
총합		**7,000**	**총합**		**1,400**
순번	test data 2 (직접수집 이미지)		순번	test data 3 (인터넷수집 이미지)	
	물질명	이미지(개)		물질명	이미지(개)
1	염소	200	1	염소	200
2	질산	200	2	질산	200
3	브로민	200	3	브로민	200
4	등유	200	4	등유	200
5	염화제이구리	200	5	염화제이구리	200
6	톨루엔	200	6	톨루엔	200
7	메탄올	200	7	메탄올	200
총합		**1,400**	**총합**		**1,400**

〔표 1〕 데이터셋 정보

오렌지는 데이터 시각화, 머신러닝, 데이터 마이닝 등의 작업을 코드 없이 사용할 수 있는 오픈 소스 프로그램이다. 코드 대신 드래그&드롭 형식으로 위젯들을 연결해 다양한 작업을 수행할 수 있으며 논문에서 사용한 버전은 3.36.2이다. 오렌지에서 이미지 모델링 작업을 수행하기 위한 위젯은 [그림 1]과 같다.



〔그림 1〕 이미지 분류 모델링 위젯

먼저 오렌지를 이용해 이미지 분석을 하기 위해서는 이미지를 임포트한 다음 임베딩을 진행해야 한다. 임베더는 구글에서 개발한 모델인 'SqueezenNet(local)'을 사용하였다. SqueezeNet은 인터넷 연결 없이 사용할 수 있으며 50배 적은 파라미터로 이미지넷에서 AlexNet 수준의 정확도를 달성하는 이미지 인식용 딥 모델이다[4]. 임베딩 이후 'Data Sampler'위젯을 이용해 학습 데이터와 테스트 데이터를 나누고 모델이 훈련 데이터에만 과도하게 적합되지 않도록 'Sampling Type'에서 'Fixed proportion of data'를 80%로 지정해주었다. 데이터 샘플링이 완료되었다면 머신러닝 모델 위젯들과 연결하여 모델 학습을 진행한다. 테스트 데이터들의 분류 결과와 정확도를 살펴보기 위해 테스트 데이터들을 임포트하여 똑같이 SqueezeNet으로 임베딩해준 후 'Predictions'위젯에 학습된 모델들과 연결해 모델별 성능 지표를 확인한다[표2].

test data 1(이미지 혼합) 성능 지표

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.909	0.677	0.670	0.698	0.677	0.628
Random Forest	0.804	0.511	0.502	0.507	0.511	0.431
Neural Network	0.884	0.630	0.618	0.625	0.630	0.571
SVM	0.898	0.605	0.605	0.611	0.605	0.546

test data 2(직접수집 이미지) 성능 지표

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	1.000	0.857	0.806	0.777	0.857	0.852
Random Forest	0.990	0.831	0.805	0.859	0.831	0.812
Neural Network	1.000	0.857	0.806	0.776	0.857	0.852
SVM	1.000	0.866	0.829	0.931	0.866	0.862

test data 3(인터넷수집 이미지) 성능 지표

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.826	0.652	0.564	0.528	0.652	0.624
Random Forest	0.798	0.524	0.461	0.456	0.524	0.464
Neural Network	0.804	0.624	0.531	0.496	0.624	0.589
SVM	0.791	0.499	0.420	0.380	0.499	0.427

〔표 2〕 모델별 성능 지표

이미지 분류 결과 모델별 분류 정확도는 [표3]과 같다.

물질	정확도(%)			
	Logistic Regression	Random Forest	Neural Network	SVM
test data 1 (이미지 혼합)				
질산	96	38.5	95.5	53
브로민	50	51	51	50
염소	63.5	54	31	62.5
메탄올	50	23.5	51.5	50
염화제이구리	100	95.5	100	96
톨루엔	52	46.5	50	61.5
등유	62.5	48.5	62	50.5
평균(%)	67.7	51.1	63	60.5
test data 2 (직접수집 이미지)				
질산	100	27	100	100
브로민	100	100	100	100
염소	100	100	100	100
메탄올	0	56	0	6.5
염화제이구리	100	100	100	100
톨루엔	100	98.5	100	100
등유	100	100	100	100
평균(%)	85.7	83.1	85.7	86.6
test data 3 (인터넷수집 이미지)				
질산	98	93	94.5	72
브로민	0	0	0	0
염소	100	64	100	100
메탄올	72	43	65.5	63
염화제이구리	87	87.5	77	15
톨루엔	0	0	0	0
등유	99.5	79.5	100	99.5
평균(%)	65.2	52.4	62.4	49.9

〔표 3〕 모델별 분류 정확도

모델별 분류 정확도를 살펴보면 모든 테스트 데이터에 대해 Logistic Regression(72.9), Neural Network(70.4), SVM(65.7), Random Forest(62.2) 순으로 분류 정확도가 높게 나타났으며[표4], 전반적으로 직접수집 이미지의 경우 분류 정확도가 높으나 인터넷수집 이미지의 경우 분류 정확도가 낮다. 이는 직접수집 이미지의 경우 비슷한 환경에서 촬영된 반면에 인터넷수집 이미지는 다양한 환경에서 촬영되었기 때문으로 예상된다. 더불어 특성(불꽃이나 연기의 색)이 유사한 물질의 경우 분류에 실패하는 경우가 많았다.

data	Logistic Regression	Random Forest	Neural Network	SVM
이미지혼합	67.7	51.1	63	60.5
직접수집	85.7	83.1	85.7	86.6
인터넷수집	65.2	52.4	62.4	49.9
평균(%)	72.9	62.2	70.4	65.7

〔표 4〕 모델별 분류 정확도(2)

Ⅲ. 결 론

본 논문에서는 유해화학물질 사고 영상과 이미지를 이용하여 유해화학물질 분류에 적합한 인공지능 학습 모델에 대한 연구를 수행하였다. 유해화학물질 이미지 데이터를 'Logistic Regression', 'Random Forest', 'Neural Network', 'SVM' 등의 머신러닝 알고리즘에 대하여 학습을 수행하여 유해화학물질 이미지 분류 모델링을 실시하였다. 그 결과, 모델 중에서 'Logistic Regression'모델의 분류 정확도가 가장 높았으며, 학습 데이터의 특성에 따라 인터넷에서 수집된 물질 이미지와 발현 특성이 유사한 물질인 경우에 분류 정확도가 저하되는 것으로 나타났다.

본 연구는 인공지능에 대한 전문 지식이 부족한 비전공자 및 일반인들도 쉽게 활용할 수 있는 공개 소프트웨어인 오렌지를 활용하여 진행되었으며, 향후 유해화학물질 사고 대응의 신속성과 효율성을 향상시켜 화학물질 관련 사고로 인한 피해를 최소화하는 데 기여할 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부와 정보통신산업진흥원 주관으로 소방청 컨소시움에서 수행하는 "AI융합 유해화학물질 판독시스템 사업" (2022~2024)의 지원을 받았음

참 고 문 헌

[1] 화학물질안전원 (https://nics.me.go.kr/)

[2] 김연진, 박봉섭, 김경배, "인공지능기술 기반의 유해화학물질 사고 대응에 관한 연구," 한국통신학회 하계종합학술발표회, pp.359-360, June, 2022.

[3] 김연진, 심보리, 윤아롱, 박봉섭, 김경배, "유해화학물질 판독 시스템을 위한 인공지능학습데이터 구축." 한국통신학회 하계종합학술발표회, June, 2023.

[4] 오렌지 (https://orangedatamining.com/)