논문 2023-1-5 http://dx.doi.org/10.29056/jsav.2023.3.05

인공지능 기반 SW 유사성 탐지 모델 설계에 관한 연구

안철범*, 김진홍**†

A Study on AI-based SW Similarity Detection Model Design

Chulbum Ahn*, Jinhong Kim**†

요 약

현재 우리는 4차 산업혁명이라 불리는 시대에 살고 있으며, 이러한 시대를 대변하는 핵심 기술로 인간의 두뇌를 이용하여 수행할 수 있는 지능 활동의 대부분을 컴퓨터가 대신하는 인공지능을 활용하고 있다. 인공지능에 의해 다양한 시스템 및 SW는 점차적으로 진화하고 있으며, SW 사용성은 매우 빠르게 증가하고 있다. 하지만, SW 사용성이 증가함에 따라 이에 대한 복제 및 오용 등의 문제가 발생됨에 따라 연구소 및 기업에서는 SW 유사성 발생되고 있다. 따라서, 본 연구에서는 SW 유사성을 위해 인공지능을 활용한 유사성 탐지 모델을 설계하고자 한다. 본 연구의 유사성 탐지 모델 설계는 유사성 탐지 모델에서의 Few-Shot Learning은 데이터가 충분한 DataSet을 사용하여 메타 러닝을 진행하고, 각 클래스에 포함된 데이터가 적은 데이터를 분류하여, 인간의 뇌를 합리적으로 표현하는 딥러닝의 인지와 학습을 기반으로 한다.

Abstract

Currently, we live in an era called the Fourth Industrial Revolution, and as a key technology representing this era, we use artificial intelligence instead of computers for most of the intelligence activities that can be performed using the human brain. Various systems and SW are gradually evolving due to artificial intelligence, and SW usability is increasing very rapidly. However, as SW usability increases, problems such as reproduction and misuse occur, and research institutes and companies are demanding SW similarity. Accordingly, this study aims to design a similarity detection model using artificial intelligence for SW replication emotion. The similarity detection model design of this study is based on the Few-Shot Learning in a similarity detection model proceeds with meta-learning using DataSet with sufficient data, the data with less data contained in each class, and the recognition and learning of deep learning that rationally expresses the human brain.

한글키워드 : 4차 산업혁명, 인공지능, SW 사용성, SW 유사성, 탐지 모델 설계

keywords : Fourth Industrial Revolution, AI, SW usability, SW similarity, Detection Model design

1. 서 론

다양한 지능형 시스템이 오늘날 4차 산업혁명 중심으로 기업, 연구소, 사용자 기반으로 활용됨에 따라, 이에 대한 서비스는 증가하고 있으며, 그들의 SW 개발 형상 및 사용성은 매우 빠르게 증대하고 있다. 대표적으로, 네이버, 구글, 아마존

* 서일대학교 정보통신공학과
** 배재대학교 AI·SW창의융합대학 소프트웨어학과
† 교신저자: 김진홍(email: jinhkm@pcu.ac.kr)
접수일자: 2023.03.02. 심사완료: 2023.03.15.
게재확정: 2023.03.20.

및 스마트 플랫폼 환경에서의 앱들은 플랫폼을 중심으로 SW 개발자에 의해 생성되고 있고, 이에 파생되는 데이터 서비스는 너무나 방대한 양을 보유함에 사용자 중심의 서비스를 제공하기 위해 인간의 두뇌를 이용하여 수행할 수 있는 지능 활동의 대부분을 컴퓨터가 대신하는 인공지능을 활용하고 있다. 하지만 SW 사용성이 증가함에 따라 이에 대한 복제 및 오용 등의 문제가 발생됨에 따라 연구소 및 기업에서는 SW 감정을 요구하는 사례라 빈번히 발생되고 있다[1-3]. 따라서, 본 연구에서는 인공지능을 활용한 유사성 탐지 모델을 설계함으로서 효율적인 SW감정 유사성 탐지 모델을 인공지능을 기반으로 설계하고자 한다. 또한, 유사성 탐지 모델 설계 측면을 고려하여 인간의 뇌를 합리적으로 표현하는 딥러닝의 인지와 학습을 기반으로 이에 제안하고자 한다.

2. 딥러닝 인지 및 학습

2.1 딥러닝 인지

딥러닝(Deep Learning)은 딥러닝 신경망 모델 기반의 기계 학습을 차별화한 것으로 입력 데이터에 대한 특정 추출과 문제 해결을 위한 복잡한 함수를 학습하기 위해 다수의 층을 갖는 신경망 구조를 일컫는다 [4]. 이는 기계 학습 방법에 의해 특정 벡터를 추출하여 입력으로 사용하는 학습된 신경망 성능을 결정할 수 있다. 딥러닝은 다수의 층이 있는 신경망을 사용하여 학습을 통해 입력으로 주어진 원 데이터로부터 적합한 특징을 추출하고 동시에 문제 해결을 위한 모델을 만드는 기계학습 기법이다 [5-9]. 이에 반해, 딥러닝 신경망은 원 데이터를 그대로 입력으로 사용함에 있어, 신경망 안에서의 특징 추출이 이루어지며 입력에 대응하는 목적 추출이 나오도록 하는 학습이 이루어진다. 즉, 딥러닝 신경망에서는 문제 해결에 사용할 특징이 학습을 통해 결정되기에 기존 기계학습 알고리즘에 비해 더 나은 성능을 제공할 뿐 만 아니라, 성능 개선을 위한 원 데이터를 전처리하여 입력으로 사용하는 특징이 있다. 다층 신경망을 학습하고자 할 경우 보통 가중치 초기값으로 0에 가까운 랜덤 값을 사용한다. 특정 구조의 신경망을 동일한 학습 데이터로 학습시키더라도 가중치의 초기값에 따라 학습된 신경망의 성능 차이가 나타난다. 다층 신경망의 가중치 초기값을 결정하는 효과적 방법으로는 첫째로, 제한적 볼츠만 머신(Restricted Boltzmann Machine)이 있다. 학습 데이터의 입력값만을 사용하여 제한적 볼츠만 머신이 입력값을 재현할 수 있도록 학습하며, 이때 결정된 가중치값을 가중치의 초기값을 사용하며, 제한적 볼츠만 머신을 초기값의 사전 학습을 하는데 사용한다. 두 번째로, 입력노드 개수 n_i와 출력노드 n_{i+1}의 정보를 반영하여 아래의 식 1)에서와 같이 무작위로 값을 설정하는 것으로 학습 성능의 좋은 결과를 예측할 수 있다 [10].

$$U\left[-\sqrt{\frac{6}{n_i+n_{i+1}}},\sqrt{\frac{6}{n_i+n_{i+1}}}\right]$$

$$\frac{N(0,1)}{\sqrt{n_i}}\ ,\ \frac{N(0,1)}{\sqrt{n_1}/2}\ \cdots\cdots(1)$$

식 1은 균등분포 U에 따라 무작위로 값을 선택하는 것으로, 예를 들면 노드 개수가 각 각 4, 7인 인접한 층 사이의 가중치를 구간에서 $[-\sqrt{6/11},\sqrt{6/11}]$에서 무작위로 값으로 초기화하는 것이다. N(0, 1)은 평균이 0이고, 분산이 1인 정규분포에서 무작위로 값을 선택하는 것을 의미한다. 이는 제이비어 초기화 방법으로서 N(0,1)/$N(0,1)/\sqrt{n_i/2}$ 허 초기화라 한다. 세 번

째로는 직교 행렬을 만들어 가중치를 초기화하는 방법으로 인접하는 층 사이의 가중치를 나타내는 행렬 W가 n * m의 크기일 때, 직교 행렬을 나타낼 수 있다. 먼저 W의 각 원소를 평균 0, 분산 1인 정규분포에서 무작위로 표본추출해서 채운 후 W를 특이값 분해하여, 직교하는 n차원 벡터를 m개 선택함으로서 직교 행렬을 만드는 방식이다. 이는 컨볼루션 신경망의 컨볼루션 필터를 초기화 할 때 주로 활용 된다 [11-13]. 컨볼루션 신경망의 구조를 요구사항에 의한 다양한 형태로 설계가능하며 이는 아래와 같다.

1) Conv-ReLU-Pool-Conv-ReLU-Pool-Conv -ReLU-Pool-FC-SM
2) Conv-Pool-Conv-Pool-Conv-FC-FC-SM
3) Conv-Pool-Conv-Pool-....-FC-FC-SM
4) Conv-ReLU-Pool-....-Conv-...-FC-SM

2.2 딥러닝 학습

딥러닝 기반 컨볼루션 신경망은 다차원의 데이터를 입력받아 이에 대응하는 분류를 출력하는 불류 문제에 주로 활용된다 [14]. 이 경우 컨볼루션 신경망의 마지막층은 소프트맥스층으로 구성되는데, 각 노드는 하나의 부류를 나타내고, 노드의 출력값은 입력이 해당 부류에 속할 확률 값으로 간주할 수 있다. 소프트맥스를 출력에 사용할 경우에는 학습을 위한 목적 함수 E로 학습 데이터 출력 t_{ik}와 컨볼루션 신경망 출력 $y_k(x_i, w_i)$의 교차 엔트로피를 사용 한다 [15]. 교차 엔트로피는 학습에 대한 음의 로그 가능도에 해당하며, 이에 대한 식2)로 정의된다.

$$E(w) = -\log \sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik} \log y_k(x_i, w) \cdots\cdots (2)$$

여기에서 t_{ik}는 i번째 학습 데이터의 k번째 출력 성분을 나타내고, $y_k(x_i, w)$는 i번째 학습 데이터의 x_i에 대한 출력 t_i는 one-hot 벡터로 표현한다. 즉, i번째 학습 데이터가 k번째 부류에 속하는 것이라면, k번째 요소만 1이고 나머지는 0인 벡터로 t_i가 표현된다. 다만, 컨볼루션 신경망이 회귀 문제에 사용되는 경우에는 일반적으로 소프트맥스 층이 사용되지 않으며, 목적 함수는 학습 데이터의 출력 t_{ik}와 컨볼루션 신경망의 출력 $y_k(x_i, w)$의 차이의 제곱으로 식 3과 같이 정의된다.

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} (t_{ik} - y_k(x_i, w))^2 \cdots\cdots (3)$$

이 외 다양한 딥러닝 학습으로 컨볼루션 신경망 모델은 LeNet, ImageNet, AlexNet, VGGNet 등이 있다.

3. 유사성 탐지 모델 설계

유사성 탐지 모델에서의 Few-Shot Learning은 데이터가 충분한 DataSet을 사용하여 메타 러닝을 진행하고, 각 클래스에 포함된 데이터가 적은 데이터를 분류하기 위한 학습 방법이다 [16-18]. 이는 특징 벡터 학습과 벡터 사이의 거리를 비교할 뿐 만 아니라, 유사성 점수를 비교하여 동일여부를 탐지하는 모델이다. 이러한 모델은 DataSet 자체가 방대해야만 가능하며 특히 이미지 기반의 형태로 Siamese Networks의 구조를 갖는 좋은 장점이다. 이에 대한 구조는 그림 1과 같다.

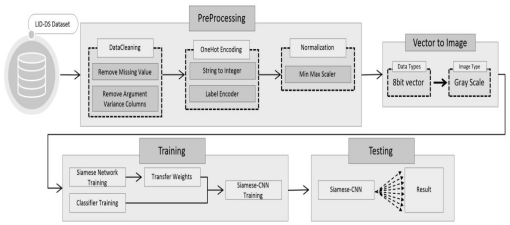

그림 1. 호스트 기반 탐지 모델 구조
Fig. 1. Host-based Detection Model Structure

그림 1처럼 Few-Shot Learning 모델의 구조를 통해 본 연구의 유사성 탐지 모델은 딥러닝 인지 및 학습에서의 컨볼루션 신경망을 기반으로 그림 2와 같이 제안한다.

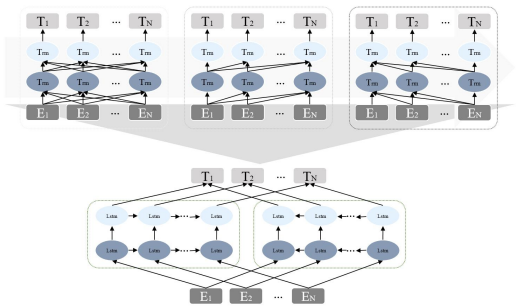

그림 2. SW유사성 탐지 BERT 구조
Fig. 2. SW Similarity Detection BERT Structure

그림 2와 같이 제안한 이유는 SW는 텍스트 기반으로 되어 있으며, Word2Vec등과 같은 임베딩 알고리즘으로 사전에 학습된 임베딩 벡터들을 가져와 사용할 수 있기 때문이다. 만약 태스크에 사용하기 위한 SW가 적다면, 사전 훈련된 임베딩을 사용하면 성능 향상을 기대해 볼 수 있다. 또한, SW 하나의 단어가 하나의 벡터 값으로 매핑 되기에 문맥을 고려할 필요가 없으며, ELMo나 BERT (Semi-Supervised Sequence Learning, Google, 2015)와 LSTM 언어 모델로서 SW 텍스트 분류 추가 학습 방법을 응용함으로서 보다 효과적인 SW 유사성 탐지를 수행할 수 있다 [19-20]. 일례로, 원 SW와 유사성 비교 대상군인 대조군 SW, 우리는 이를 B-SW라 칭한다. SW 유사성 탐지 모델을 통해 Inception-ResNet 모델 구조로 B-SW에 대한 유사성 탐지를 구현하면 그림 3과 같다.

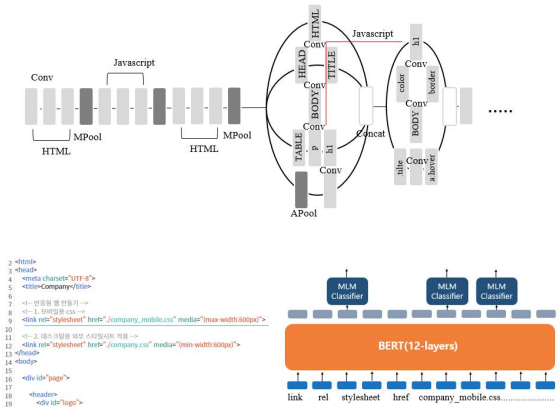

그림 3. Inception-ResNet 모델 기반 B-SW
Fig. 3. Inception-ResNet Model-based B-SW

위와 같이, B-SW는 해당 소스코드(HTML+Javascript)에서 태그 결과를 12개의 층으로 연산할 수 있었다. 이를 통해 마스크드 언어 모델로 사전 훈련(그림 2)을 위해서 인공 신경망의 소스코드 텍스트의 15% 단어를 랜덤으로 마스킹 하였다. 또한, 인공신경망에게 가려진 소스코드를 예측할 수 있었다. 결국 모든 B-SW 소스코드에 대한 세그먼트 임베딩으로 소스코드 상의 Sentence 0 임베딩, 두 번째 소스코드 상의 Sentence 1 임베딩을 더해주는 벡터로 수행한 후 각 태깅 작업을 통해 보다 정확한 SW유사성 탐지를 수행할 수 있었다.

4. 결 론

오늘날의 다양한 SW는 오픈 소스 기반으로 여러 시스템 환경에 활용되고 있다. 하지만, 이러한 SW는 때로는 변형되어 사용되거나 복제되어

사용됨에 따라 SW 감정을 위한 다양한 방법론이 있으나, 본 연구는 인공지능 기반 유사성 탐지 모델에 대해 제안하였다.

본 연구는 텍스트 기반으로 되어 있는 Web 언어를 SW유사성 탐지 BERT 구조로 표현하여 Inception-ResNet 모델 기반 B-SW를 분석하였고, 이에 대한 결과를 그림 4와 같이 유사도 매우 높다는 결과를 시각화 하였다. 향후, SW 유사도 분석을 위해 딥러닝 알고리즘 쉽게 구현하여 대량의 SW 데이터에 대해 학습하고, 적합한 모델을 설계하여 보다 효과적인 유사도 감정을 수행할 수 있을 것으로 기대한다.

참 고 문 헌

[1] Lee Kyu Tae, “A detail item guideline for IT device evaluation”, Journal of Software Assessment and Valuation, Vol.2, No. 1, pp.21-26, June, 2016. http://dx.doi.org/10.29056/jsav.2020.12.04

[2] Lee Kyu Tae,“A Similarity of device driver on Embedded system”, Journal of Software Assessment and Valuation, Vol. 4, No. 1, pp.27-32, June, 2018. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002395870

[3] Kim, Si Yeol, Yoonsoo Kang, “A Study on Application of Summary Procedure in Case of Software Appraisal”, Journal of Software Assessment and Valuation Vol. 15 No. 1, pp.25-34. June, 2019. http://dx.doi.org/10.29056/jsav.2019.12.04

[4] Gu, J.; Lan, C.; Chen, W.; Han, H. Joint Pedestrian and Body Part Detection via Semantic Relationship Learning. Appl. Sci. 2019. https://www.mdpi.com/2076-3417/9/4/752

[5] Gao, H.; Chen, S.; Zhang, Z. Parts Semantic Segmentation Aware Representation Learning for Person Re-Identification. Appl. Sci. 2019, 9, 1239.

[6] Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234 - 241. https://www.mdpi.com/2076-3417/9/6/1239

[7] Bouindour, S.; Snoussi, H.; Hittawe, M.M.; Tazi, N.; Wang, T. An On-Line and Adaptive Method for Detecting Abnormal Events in Videos Using Spatio-Temporal ConvNet. Appl. Sci. 2019, 9, 757.

[8] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 2015, 28, 91 - 99. https://dl.acm.org/doi/10.5555/2969239.2969250

[9] Radac, M.B.; Precup, R.E. Data-Driven Model-Free Tracking Reinforcement Learning Control with VRFT-based Adaptive Actor-Critic. Appl. Sci. 2019, 9, 1807. https://www.mdpi.com/2076-3417/9/9/1807

[10] Wei, C.; Ni, F.; Chen, X. Obtaining Human Experience for Intelligent Dredger Control: A Reinforcement Learning Approach. Appl. Sci. 2019, 9, 1769. https://www.mdpi.com/2076-3417/9/9/1769

[11] Zheng, H.T.; Chen, J.Y.; Liang, N.; Sangaiah, A.K.; Jiang, Y.; Zhao, C.Z. A Deep Temporal Neural Music Recommendation Model Utilizing Music and User Metadata. Appl. Sci. 2019, 9, 703.https://www.mdpi.com/2076-3417/9/4/703

[12] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", Nature, vol. 521, no. 7553, pp. 436-444, May 2015. https://pubmed.ncbi.nlm.nih.gov/26017442/

[13] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends perspectives and prospects", Science, vol. 349, no. 6245, pp. 255-260, 2015. https://www.science.org/doi/10.1126/science.aaa8415
[14] K. Nagpal et al., "Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer" in CoRR, Nov. 2018. https://www.nature.com/articles/s41746-019-0112-2
[15] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep reinforcement learning: A brief survey", IEEE Signal Process. Mag., vol. 34, no. 6, pp. 26-38, Nov. 2017. https://ieeexplore.ieee.org/document/8103164
[16] M. Gheisari, G. Wang and M. Z. A. Bhuiyan, "A survey on deep learning in big data", Proc. IEEE Int. Conf. Comput. Sci. Eng. (CSE), pp. 173-180, Jul. 2017. https://www.sciencedirect.com/science/article/abs/pii/S1566253517305328
[17] S. Pouyanfar, "A survey on deep learning: Algorithms techniques and applications", ACM Comput. Surv., vol. 51, no. 5, pp. 92, 2018. https://dl.acm.org/doi/10.1145/3234150
[18] R. Vargas, A. Mosavi and R. Ruiz, "Deep learning: A review", Proc. Adv. Intell. Syst. Comput., pp. 1-11, 2017. DOI: 10.20944/preprints201810.0218.v1
[19] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives", IEEE Access, vol. 2, pp. 514-525, 2014. DOI: 10.1109/ACCESS.2014.2325029
[20] Y. LeCun, K. Kavukcuoglu and C. Farabet, "Convolutional networks and applications in vision", Proc. IEEE Int. Symp. Circuits Syst., pp. 253-256, May/Jun. 2010. https://ieeexplore.ieee.org/document/5537907

저 자 소 개

안철범(Chulbum Ahn)

2010.2 단국대학교 전자·컴퓨터공학과 박사
2018.3-현재 : 서일대학교 교수
<주관심분야> 데이터통신응용, 빅데이터, 네트워크 보안, 인공지능

김진홍(Jinhong Kim)

2006.2 성균관대학교 컴퓨터공학과 박사
2022.3-현재 한국SW감정평가학회 부회장
2021.3-현재 한국저작권위원회 감정인
2020.3-현재 : 배재대학교 교수
2017.3-2020.02 : 서일대학교 교수
<주관심분야> 인공지능, 빅데이터, 지능형 소프트웨어