LSTM 과 GRU 를 활용한 프로그래밍 언어 예측

이만규, 한대진, 배주영, 서희택, 나웅수*
컴퓨터공학부 소프트웨어학과, 공주대학교
aksrb973@gmail.com, djhan8733@gmail.com, bjs338570@gmail.com, suhy1245@naver.com,
*wsna@kongju.ac.kr

Programming Language Prediction Using LSTM and GRU

Mangyu Lee, Daejin Han, Juyoung Bae, Huitaek Seo, Woongsoo Na*
Kongju University

요 약

본 논문은 딥러닝을 이용하여 사용자들이 미래 프로그래밍 언어의 선택 비중을 예측하는 연구를 진행하였다. 빠르게 발전하는 IT 분야의 핵심인 프로그래밍 언어의 사용 비중을 예측한다는 것은 앞으로의 연구주제와 분야 선택에서의 트랜드 파악에 도움이 될 수 있고 또한 언어 사용량 분석을 통해 같은 강점을 가진 언어 주 더 나은 선택지 제공에도 도움이 된다. 우리가 제안하는 예측 방법은 GitHub 에서 월별 사용자들이 프로젝트에 사용한 프로그래밍 언어의 시계열 데이터로 전처리하여 딥러닝 알고리즘 중 하나인 LSTM 과 GRU 의 비교 분석을 하였다. 분석 결과 두 모델 모두 실제 데이터와 유사하게 보이지만 LSTM 같은 경우에는 급격한 변화에 대해 예측을 하지 못하는 반면 GRU 는 급격한 변화에 반응하는 모습을 보이고 있다.

Ⅰ. 서 론

IT 분야는 특히나 빠르게 발전하는 분야로 IT 의 핵심인 프로그래밍 언어의 인기 상승은 곧 다른 언어의 퇴색을 의미한다. 언어들의 사용량이 상승하거나 감소한 언어를 예측이 가능하다면 그 해 또는 앞으로의 시대적 흐름을 알 수 있는 중요한 요인으로써 활용이 가능해 진다.

본 연구에서는 GitHub 에서 사용자들이 프로젝트를 진행할 때 가장 많이 선택한 언어를 월별로 수집하여 진행하고자 한다. 이에 대해 딥 러닝을 이용하여 프로그래밍 언어의 사용량 궤적 분석과 각 언어의 상호관계를 통해 미래의 언어에 대한 사용량을 예측하여 앞으로의 연구의 방향과 기술의 발전 방향성을 제시하고자 한다.

Ⅱ. 연구 방법

본 연구에서는 Github API 를 이용해 Github 의 User 가 생성한 Repository 에 구성된 언어 종류를 시계열 데이터로 수집하여 사용하였다. 이 데이터 셋에 LSTM 모델과 GRU 모델을 사용하였다.

본 연구에서 제안한 모델은 두가지로 LSTM(Long Short Term Memory)을 활용한 모델과 GRU(Gated Recurrent Unit)를 활용한 모델이 있다. LSTM 은 기존의 RNN 과 같은 체인구조로 되어 있지만 RNN 의 장기 의존성 문제(long-term dependencies)를 해결하기 위해서 나온 모델로 좀 더 거시적으로 과거 데이터를 고려하여 미래 데이터를 예측하기 위해 나온 모델이다. LSTM 의 특징은 단순한 한 개의 layer 가 아닌 4 개의 layer 가 서로 정보를 주고 받는 구조로 되어 있다는 것이다. GRU 의 경우 복잡했던 LSTM 의 구조를 간단화 시킨 모델이다. GRU 의 특징은 LSTM 에서 출력, 입력, 삭제라는 3 개의 게이트를 사용하였지만 GRU 에서는 업데이트 게이트와 리셋 게이트 두가지 게이트만 존재 하기 때문에 학습 속도가 빠르다는 장점이 있다. 두 모델의 손실 함수로 Hobber loss 를 손실 함수로 사용하였다. Hobber 는 L1 과 L2 의 장점을 취하면서 단점을 보완하기 위해 제안된 손실 함수로 모든 지점에서 미분이 가능하면서 이상치에 강건한(robust) 성격이 보이는 손실 함수이다.

Ⅲ. 결론

본 연구는 Ruby 와 Javascript 언어에 대해 LSTM 모델과 GRU 모델을 적용하여 약 2020 년 5 월부터 2021 년도 12 월까지의 사용량에 대한 예측을 진행하였다.

그림 1,2 는 Ruby 언어를 20 개월에 LSTM 모델과 GRU 모델을 사용하여 예측을 진행한 그림이다.

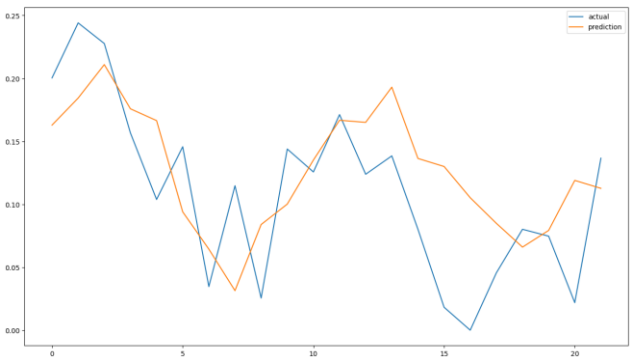

그림 1. LSTM 을 사용한 Ruby 예측 그래프

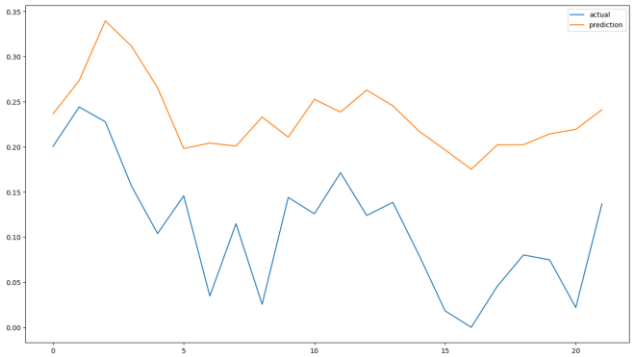

그림 2. GRU 을 사용한 Ruby 예측 그래프

그림 1 과 2 를 확인하였을 때 LSTM 의 경우 전체적으로 본 데이터와 유사하게 나왔지만 급격한 변화가 있는 지점에서는 예측이 되지 않는 문제점이 보였다. 다만 GRU 는 실제 데이터와의 값의 차이는 있지만 경향에 대해서는 모든 부근에 대해 보다 우수한 성능을 보였다. 이는 본 연구에서 사용한 GitHub 의 출시가 2008 년으로 데이터 세트가 10 년간의 월별 언어 선택량으로 데이터의 양이 방대하지 않아 GRU 의 특징인 단기 데이터에는 더 적합하여 LSTM 에 비해 예측 정확도가 높은 것으로 보인다. Loss 값 역시 비교를 진행하여도 LSTM 은 0.0032 가 나왔고 GRU 는 0.0019 로 측정되어 GRU 모델이 더욱 적합한 것으로 보인다.

그림 2 와 3 은 Javascript 언어데 대한 예측을 진행한 그림이다.

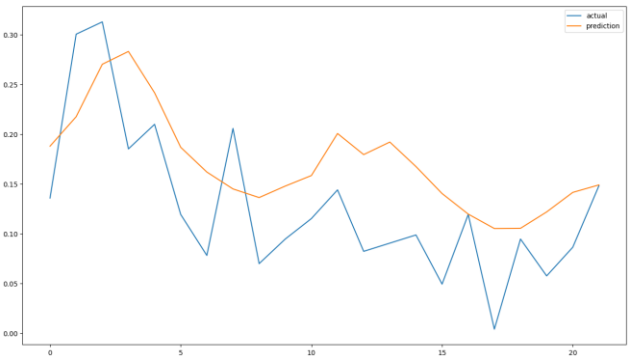

그림 3. LSTM 을 사용한 Javascript 예측 그래프

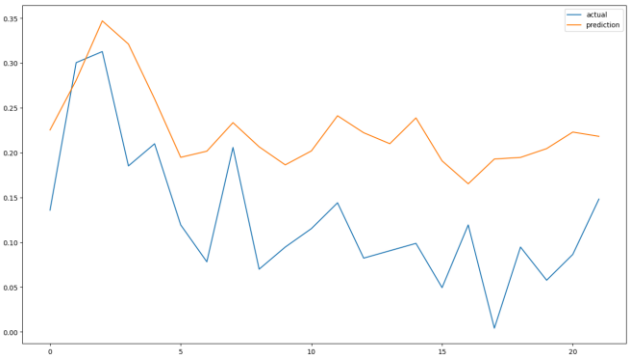

그림 4. GRU 을 사용한 Javascript 예측 그래프

Javascript 에서도 같은 방법으로 기술을 적용하였을 때 GRU 가 LSTM 에 비해 특정 부근에 대한 변화 예측이 더욱 우수한 것으로 보인다. Loss 또한 GRU 가 0.0033 이며 LSTM 이 0.0042 로 나타난다.

IV. 기대효과

본 연구의 기대효과로는 언어 사용의 수요 예측이 가능하다는 점이다. 이를 통해 IT 분야에서의 트렌드 파악과 적절한 언어 제안에 도움을 줄 수 있다. 상승세에 있는 프로그래밍 언어는 곧 해당 언어를 사용하는 분야에서의 상승과 연관되기에 이는 연구 주제와 관련 기술 파악에 도움이 될 수 있다. 또한 유망한 프로그래밍 언어를 선택하고자 하는 입문 프로그래머에게 적절한 언어 선택 제안이 가능하다.

ACKNOWLEDGMENT

이 논문은 2022 년 교육부의 국립대학 육성사업 연구지원에 의해 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2023-RS-2022-00156353).

참 고 문 헌

[1] Byungun Yoon and Yongtae Park, Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information, IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, VOL. 54, NO. 3, AUGUST 2007

[2] Se Ill Cho, A Method to Forecast the Computer Technology Trends based on Computer Languages, Smart Media Journal / Vol.5, No.3 / ISSN:2287-1322

[3] Hochester S. and Schmidhuber J., “Long short-term memory.”, Neural computation 9.8.pp.1735-1780, 1997.

[4] 김건우, 장규삼, 임세민, 안인경, 박주영, 오형철 “GRU 기반 초기 동작 인식”, 대한전자공학회 하계종합학술대회 논문집,2, 016-2,019,2020

[5] 임주완, 김인경, 이명학, 하정민, 이재구, "BERT, LSTM 과 GRU 를 사용한 네트워크 이상 탐지 성능 비교", 한국통신학회 동계종합학술발표회 논문집, 1,268-1,269,2022