Journal of The Institute of Electronics and Information Engineers Vol.58, NO.9, September 2021 https://doi.org/10.5573/ieie.2021.58.9.34
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2021-58-9-4

특허문서 자동분류를 위한 딥러닝 개별 모델 분류기와 앙상블 분류기의 성능비교

(Performance Comparison of Deep Learning Individual Model Classifier and Ensemble Classifier for Automatic Classification of Patent Documents)

김 성 훈*, 김 승 천**

(Sunghoon Kim and Seungcheon Kim©)

요 약

기술 혁신의 급속한 증가와 이에 따른 출원의 증가로 인해 특허문서 자동분류기는 특허를 분류할 때 개별 발명가와 변리사 모두에게 매우 유용하다. 본 연구에서는 특허전문가의 관점과 유사한 관점을 통해 특허문서 분류를 수행하는 모델을 선택하였다. 주요 키워드의 존재 여부 및 존재 빈도를 나타낼 수 있는 모델인 MLP, 근접거리에 에 존재하는 키워드들의 존재 및 존재 빈도를 나타낼 수 있는 모델인 CNN 그리고 문장의 구성, 키워드의 순서를 나타낼 수 있는 모델인 LSTM, Attention, Transformer을 선택하여 최대한 특허전문가의 분류관점과 유사한 관점으로 분류할 수 있는 모델을 선택하였다. 본 연구에서는 3가지 앙상블 방법을 사용한다. 앙상블 방법은 각 분류기의 결과를 독립적으로 처리하는 배깅[8]을 사용하여 각각 voting, summation 그리고 weighting 의 3가지 방법을 사용하여 각각의 성능을 측정하였다. 본 연구에서 사용된 3개의 데이터셋중에 2개의 데이터셋(데이터셋 #2, 데이터셋 #3)은 앙상블의 정확도가 높았고 1개의 데이터셋 (데이터셋 #1)은 개별모델인 Attention 모델의 정확도가 높았다. 이러한 결과는 각 데이터셋의 분류 관점(키워드에 의한 분류, 문장 구성에 의한 분류)에 기인한 것으로 판단할 수 있다. 사용자 특허분류에 따른 모델의 생성이 아니라 기존 모델을 활용한 앙상블 모델을 사용한다면 데이터셋별 최적의 정확도를 내거나 최적의 정확도를 내는 단일모델에 근접한 정확도를 낼 수 있을 것으로 예상되며 이러한 앙상블 모델 방식이 특허문서 분류기를 상품화 하는데 적합할 것이다.

Abstract

Due to the rapid increase in technological innovation and the corresponding increase in applications, the automatic patent document classifier is very useful for both individual inventors and patent attorneys when classifying patents .In this study, a model for classifying patent documents was selected from a viewpoint similar to that of a patent expert.MLP, which is a model that can indicate the existence and frequency of existence of major keywords, CNN, which is a model that can indicate the existence and frequency of existence of keywords in close proximity, and LSTM, which is a model that can indicate the structure of sentences and order of keyword , Attention, and Transformer were selected to select a model that can be classified as similar to that of patent experts as much as possible. In this study, three ensemble methods are used. The ensemble method used bagging [8], which independently processes the results of each classifier, and measured each performance using three methods: voting, summation, and weighting. Among the three datasets used in this study, two datasets (Dataset #2, Dataset #3) had high ensemble accuracy, and one dataset (Dataset #1) had high accuracy of the attention model, an individual model. was high. This result can be judged to be due to the classification viewpoint (classification by keyword, classification by sentence structure) of each dataset If an ensemble model using an existing model is used instead of the generation of a model according to the user patent classification, it is expected that the optimal accuracy for each data set or accuracy close to that of a single model with optimal accuracy can be achieved. It would be suitable for commercializing a document classifier.

Keywords : Patent classification, Ensemble, Deep learning

*학생회원, 한성대학교 스마트융합컨설팅학과(Department of Smart Convergence Consulting Hansung University)
**평생회원, 한성대학교 스마트융합컨설팅학과(Department of Smart Convergence Consulting Hansung University)
© Corresponding Author(E-mail: kimsc@hansung.ac.kr)
※ 본 연구는 한성대학교 교내학술연구비 지원과제임.
Received ; July 4, 2021 Revised ; July 21, 2021 Accepted ; July 26, 2021

Ⅰ. 서 론

기술 혁신에 대한 권리를 보호하는 특성을 가진 특허는 대부분의 기업에서 중요한 자산으로 간주됩니다. 특허는 또한 기술 발전과 다양화를 대표할 수 있는 충분한 소스를 제공하기 때문에 기술 및 혁신 확산에 중요한 역할을 한다[1].

특허문서 분류는 특허문서를 효율적으로 검색, 분석, 구조화하기 위해 특허문서를 미리 정의된 클래스로 할당하는 작업이다. 특허문서 클래스는 IPC(International Patent Classification), CPC(Cooperative Patent Classification) 등의 공식적인 클래스와 각 조직 및 개인의 필요에 의해 생성되는 사용자 정의 클래스로 나뉠 수 있다.

IPC, CPC 등의 공식적인 클래스는 일반적으로 특허문서가 공개될 때 각국 특허청에 의해서 부여된다. 이러한 공식적인 클래스는 특허의 검색, 분석 등의 작업에 사용된다. 현재 대부분의 공식적 클래스는 특허 전문가에 의해 수작업으로 분류된다.

사용자 정의 클래스는 공식적인 클래스와는 다른 관점에서의 클래스가 필요할 경우 사용목적에 따라 도메인 전문가들에 의해서 클래스가 설계되고 관심 특허가 분류되어 진다. 이러한 사용자 정의 클래스와 클래스에 할당된 특허문서를 바탕으로 경쟁사 특허분석, 기술 동향 분석 등을 수행한다. 대부분의 분류 작업은 특허분류 전문가에 의해 수작업으로 진행되고 있으므로 많은 시간과 비용이 소요된다. 이러한 비용과 시간을 절약하기 위해, 전문가에 의해 분류된 특허 데이터를 학습 데이터로 사용하여 생성된 지도학습 특허문서 자동분류기는 좋은 해결책이 될 수 있다.

기술 혁신의 급속한 증가와 이에 따른 출원의 증가로 인해 특허문서 자동분류기는 특허를 분류할 때 개별 발명가와 변리사 모두에게 매우 유용하다[2].

지도학습 기반의 문서분류기는 스팸메일 분류기, SNS 평판 분류기 등이 사용되고 있으며 이러한 문서분류기는 유의미한 결과를 내고 있다. 그러나 사용자 정의 클래스의 특허문서에 일반 문서분류기를 이용해서 분류할 경우에는 분류 클래스에 따라 정확도가 낮으며 클래스별 편차가 심하여 실제 업무에 적용되고 있지 못한 실정이다. 특허문서를 일반적인 문서분류기로 분류할 때 정확도가 높지 않은 이유는 다음과 같다.

첫째, 훈련에 필요한 충분한 양의 데이터를 확보하기 힘들다. 사용자 정의 클래스를 정의하고 클래스에 따라 특허를 분류하는 작업은 많은 비용과 시간이 드는 작업이다. 그러므로 대규모의 훈련데이터를 확보하는 것은 많은 비용이 드는 작업이므로 훈련에 필요한 충분한 양의 데이터를 확보하는 것이 힘들다.

둘째, 기계학습 모델 선정의 어려움이 있다. 사용자 정의 클래스는 다양한 관점의 사용자 클래스가 존재할 수 있다. 예를 들어 특허에 포함된 기술에 대한 분류, 특허가 적용되는 제품에 따른 분류 또는 특허가 해결하고자 하는 과제에 따른 분류 등의 사용자 클래스가 존재할 수 있다. 이러한 다양한 관점의 분류에 모두 적합한 기계학습 모델을 선정하는 것에 어려움이 따른다.

셋째, 사용자 클래스에 적합한 특허 구성필드를 선정하기에 어려움이 따른다. 특허문서는 제목, 요약, 대표청구항, 전체청구항 그리고 상세설명 등의 텍스트로 이루어진 필드와 출원일, 등록일 등의 날짜 필드등의 다양한 필드로 구성되어 있다. 사용자 클래스 분류에서 이러한 필드 중 어떠한 특허 필드를 사용하는 것이 적합한지에 대해 선정하는 것에 대한 어려움이 따른다.

본 연구에서는 위에 언급한 3가지 문제를 다음과 같은 방법으로 실험하여 그 해결가능성을 확인하고자 한다.

다수의 특허 클래스를 구성하는 상대적으로 적은 수의 훈련데이터를 사용하여 다양한 딥러닝 모델과, 다양한 특허 필드의 조합으로 개별 분류기를 생성하여 성능을 측정한다. 또한 생성된 개별분류기를 선택하여 다양한 방식으로 조합된 앙상블 분류기를 생성하여 성능을 측정한다.

본 연구에서는 클래스 개수가 다른 다수의 특허분류 데이터 셋을 구성하고 이를 이용하여 각 개별분류기와 앙상블 분류기의 성능을 비교하고자 한다. 이를 통해 클래스의 개수와 분류 관점이 다양한 특허분류에 있어서 개별 딥러닝 모델과 앙상블 기법의 성능을 측정하여 다양한 사용자 정의 클래스를 가지는 특허문서를 자동분류 하는 최적의 모델의 찾고자 한다.

Ⅱ. 본 론

1.다양한 사용자 정의 클래스에 적합한 분류기

특허문서 분류를 위해 사용할 수 있는 딥러닝 모델은 다양하다. 본 연구에서는 MLP[3], CNN[4], LSTM[5], Attention[6], Transformer[7] 딥러닝 모델을 선택하였다.

일반적으로 특허전문가가 특허문서를 분류하는 관점은 다양하게 존재 한다. 그중 형식적인 관점으로는 특허문서를 구성하는 주요 키워드에 의한 분류, 주요 키워드와 주변에 있는 키워드간의 관계 그리고 주요키워드의 순서 즉 문장구성에 의한 분류관점이 존재할 수 있다. 본 연구에서는 위와 같은 특허전문가의 관점과

유사한 관점을 통해 특허문서 분류를 수행하는 모델을 선택하였다. 주요 키워드의 존재 여부 및 존재 빈도를 나타낼 수 있는 모델인 MLP, 근접거리에 에 존재하는 키워드들의 존재 및 존재 빈도를 나타낼 수 있는 모델인 CNN 그리고 문장의 구성, 키워드의 순서를 나타낼 수 있는 모델인 LSTM, Attention, Transformer을 선택하여 최대한 특허전문가의 분류관점과 유사한 관점으로 분류할 수 있는 모델을 선택하였다.

앙상블[8] 특허문서 분류기는 여러 단일 모델 특허문서 분류기를 통합하여 분류기를 구축하는 것이다. MLP, CNN, LSTM 등의 개별 모델들로 구성한 특허문서 분류기로 구성 능하다. 그런데 특허문서 분류의 분류기준과 분류 클래스의 개수는 다양하다. 이러한 다양한 특허 데이터셋에 가장 적합한 딥러닝 모델은 무엇인가?를 찾을 수 있다면 단일 딥러닝 모델을 이용하여 특허 문서분류기를 만들면 될 것이다. 그러나 특허분류 데이터 셋에 따라 어떠한 데이터셋은 LSTM에서 가장 높은 성능을 보이고, 어떠한 데이터셋은 Attention에서 가장 높은 성능을 보인다면 특정 단일 모델을 지정해서 생성된 분류기는 특허 데이터셋에 따라 분류 성능이 많은 편차를 보일 수 있다.

사용자 정의 클래스 특허분류의 특성상 클래스의 분류 관점과 클래스의 개수가 다양하다. 이러한 특징으로 인해 단일 모델 분류기가 다양한 데이터셋에서 골고루 좋은 성능을 낼 수 있을지는 의문이다. 이러한 문제에 대한 해결책을 찾기 위해 본 연구에서는 다양한 데이터셋을 만들고 단일 모델을 사용하는 다양한 단일모델 분류기를 만들어 각각에 대해 성능을 측정한다.

또한 각 단일 모델 분류기를 통합한 앙상블 분류기를 생성하여 앙상블분류기의 각 데이터셋에 대해 성능을 측정한다. 이러한 단일 모델 분류기와 앙상블 모델의 결과를 비교하고 이를 바탕으로 다양한 사용자 정의 클래스에 적합한 특허문서분류기의 구조를 제안하고자 한다.

Ⅲ. 실 험

1. 특허 데이터 셋

본 연구에서는 총 3가지 종류의 데이터 셋을 사용하였다. 분류의 신빙성을 확보하기 위해 미국 등록특허의 CPC(Cooperative Patent Classification) 분류 데이터를 사용하였다. CPC는 미국 특허청과 유럽 특허청이 각각 관리하던 분류체계를 통합하여 새롭게 만들어진 분류체계이다.

가. CPC C01 계열 데이터

CPC C01계열 데이터의 클래스는 C01B, C01C, C01D, C01F, C01G 5개의 클래스 데이터를 사용하였다[9]. 각 클래스의 분류기준 및 정의는 미국 특허청과 유럽 특허청의 공동 작업으로 만들어졌으며 새롭게 공개되는 특허들에 대해서도 해당 분류기준에 맞게 정의된다.

CPC C01 계열 데이터를 훈련데이터, 검증데이터 그리고 테스트 데이터로 분리한다. 이를 데이터셋 #1 이라 명명한다. 데이터셋 #1 의 구성은 각 클래스별로 200건을 사용하였고 전체 1,000건의 특허데이터로 구성하였다. 훈련데이터와 테스트 데이터의 비율은 8:2 로 분리하였으며 분리한 기준은 출원일을 기준으로 하였다. 즉 상대적으로 과거의 데이터를 훈련데이터로 사용하였고 최신의 데이터를 테스트데이터로 사용하였다. 이는 실제 특허분류업무에서 이루어지는 과정과 동일하게 이전에 출원된 데이터의 분류를 바탕으로 미래에 출원되는 특허데이터를 분류하는 프로세스와 유사한 환경을 표현하기 위함이다.

나. CPC H01 계열 데이터

CPC H01계열 데이터의 클래스는 H01B, H01C, H01F, H01G, H01H 등 총 14 개의 클래스 데이터를 사용하였다[9].

CPC H01 계열 데이터를 훈련데이터, 검증데이터 그리고 테스트 데이터로 분리한다. 이를 데이터셋 #2라 명명한다. 데이터셋 #2 의 구성은 각 클래스별로 200 건을 사용하였고 전체 2,800건의 특허데이터로 구성하였다. 훈련데이터와 테스트 데이터의 비율은 8:2 로 분리하였으며 분리한 기준은 출원일을 기준으로 하였다.

다. CPC C01 + H01 데이터

CPC C01 계열과 CPC H01 계열 데이터를 결합하여 총 19개의 클래스 데이터를 사용한다. 이를 데이터셋 #3 이라 명명한다. 데이터셋 #3 의 구성은 각 클래스별로 200 건을 사용하였고 전체 3,800건의 특허데이터로 구성하였다. 훈련데이터와 테스트 데이터의 비율은 8:2 로 분리하였으며 분리한 기준은 출원일을 기준으로 하였다.

표 1에서 3가지 데이터셋의 구성에 대한 요약정보를 나타내었다.

2. 특허분류를 위한 딥러닝 개별모델

본 연구에서는 5가지의 딥러닝 모델을 사용하여 특허분류에 적용한다. 일반적인 문서분류에 검증된 모델

표 1. 데이터셋 요약
Table 1. Dataset summary.

	class	train	valid	test	total
Dataset #1	5	720	80	200	1,000
Dataset #2	14	2,016	224	560	2,800
Dataset #3	19	2,736	304	760	3,800

이며 실제 특허분류 업무의 관점과 유사한 관점을 모델링 할 수 있는 모델을 이용하였다.

특허문서의 주요 키워드의 존재 및 빈도를 모델링하는 MLP, 근접한 주요 키워드들의 군집을 모델링한 CNN 그리고 문장의 구성을 모델링한 LSTM, Attention, Transformer 모델을 사용하였고 각 모델별로 하이퍼파라메타 조합을 이용하여 모델을 구성하였다. 각 모델들에 영향을 미치는 공통적인 하이퍼파라메타에는 learning rate(0.01 ~ 0.001), batch size(8, 16, 32)가 있으며 각 모델을 구성할 때 순차적으로 사용하였다. 모델별로는 MLP에서는 node의 개수와 층의 깊이를 다양하게 조합하였으며, CNN에서는 단어의 군집 개수, LSTM, Attention, Transformer 에서는 층의 개수, head의 개수, 임베딩 층의 훈련여부 등의 변수를 변화시키면서 다양한 하이퍼파라메타의 조합으로 모델을 만들고 훈련을 수행하였다.

3. 특허분류를 위한 앙상블 분류기

본 연구에서는 3가지 앙상블 방법을 사용한다. 5가지 개별모델에서 생성된 총 15가지의 분류기(모델별 3가지 분류기)의 결과를 조합하는 앙상블 분류기의 성능을 측정한다. 앙상블 방법은 각 분류기의 결과를 독립적으로 처리하는 배깅[8]을 사용하여 각각 voting, summation 그리고 weighting 의 3가지 방법을 사용하여 각각의 성능을 측정한다.

4. 실험방법

3가지 데이터셋, 데이테셋에 각각 3가지 필드조합(제목 + 요약 + IPC, 제목 + 요약 + 대표청구항 + IPC, 제목 + 요약 + 전체청구항 + IPC)을 이용해 데이터를 준비한다.

5가지의 개별모델을 이용해서 몇 가지 하이퍼파라메타의 조합으로 다양한 분류기를 생성한다. 각 모델별로 공통적으로 batch size, learning rate를 변화시켜 적용하였으며 모델별로 층의 개수, 깊이 등의 하이퍼파라메타의 조합으로 MLP(18), CNN(4), LTM(6), ATN(8), TFS(4) 의 총 40개의 분류모델을 생성한다.

각 개별분류기에 하나의 데이터셋당 3가지의 필드조합인 데이터로 실험을 진행한다. 즉 하나의 데이터셋에 대해 총 120개의 분류기가 생성된다. 이를 validation acc와 validation loss를 기준으로 모델별 3개의 분류기를 선택하고 테스트 데이터를 이용해서 각 분류기의 성능(정확도)을 측정한다. 또한 선택된 총 15개의 분류기 결과를 3가지 방법의 앙상블의 방법으로 성능(정확도)을 측정한다.

5. 실험결과

가. 데이터셋 #1의 개별모델과 앙상블 모델의 정확도

CPC C01계열의 데이터셋은 총 2,000개의 특허문서중 720건의 특허데이터로 훈련을 수행하였고, 200개의 테스트 데이터로 성능을 측정하였다.

그림 1은 데이터셋 #1에서의 개별 모델과 앙상블 모델의 테스트 데이터의 분류 정확도를 나타낸다.

MLP(a)의 최대 정확도는 0.795, CNN(b)의 최대 정확도는 0.790, LSTM(c)의 최대 정확도는 0.875, Attention (d)의 최대 정확도는 0.91 그리고 Transformer(f)의 최대 정확도는 0.820 이다.

그림 1의 (f)는 총 15개의 분류모델을 결합한 앙상블 모델의 테스트 데이터의 분류 정확도를 나타낸다. Summation 앙상블 모델이 0.890으로 앙상블 기법 중 최대 정확도를 나타냈다.

데이터셋 #1 에서는 단일 모델들의 테스트 데이터 정확도가 0.79 ~ 0.91까지의 분포를 보였다. 또한 3가지 앙상블 모델에서는 0.875 ~ 0.89까지의 분포를 보였다. 앙상블이 특정 개별모델(Attention) 보다 정확도가 낮은 이유는 키워드의 존재 기반으로 분류하는 MLP, CNN의 모델은 문장에 의해 분류하는 모델인 Attention 등의 분류모델에 비해 데이터셋 #1 에 유효하지 않은 분류기준으로 판단되며 이러한 결과가 단일 모델보다 낮은 앙상블 모델의 정확도의 결과로 나왔다고 판단된다.

나. 데이터셋 #2의 개별모델과 앙상블 모델의 정확도

CPC H01계열의 데이터셋은 총 2,800개의 특허문서 중 2,016건의 특허데이터로 훈련을 수행하였고, 760건의 테스트 데이터로 성능을 측정하였다.

그림 2는 각각 단일 모델들을 사용한 분류기의 테스트 데이터의 추론 정확도를 나타낸다.

MLP(a)의 최대 정확도는 0.838, CNN(b)의 최대 정확도는 0.886, LSTM(c)의 최대 정확도는 0.954,

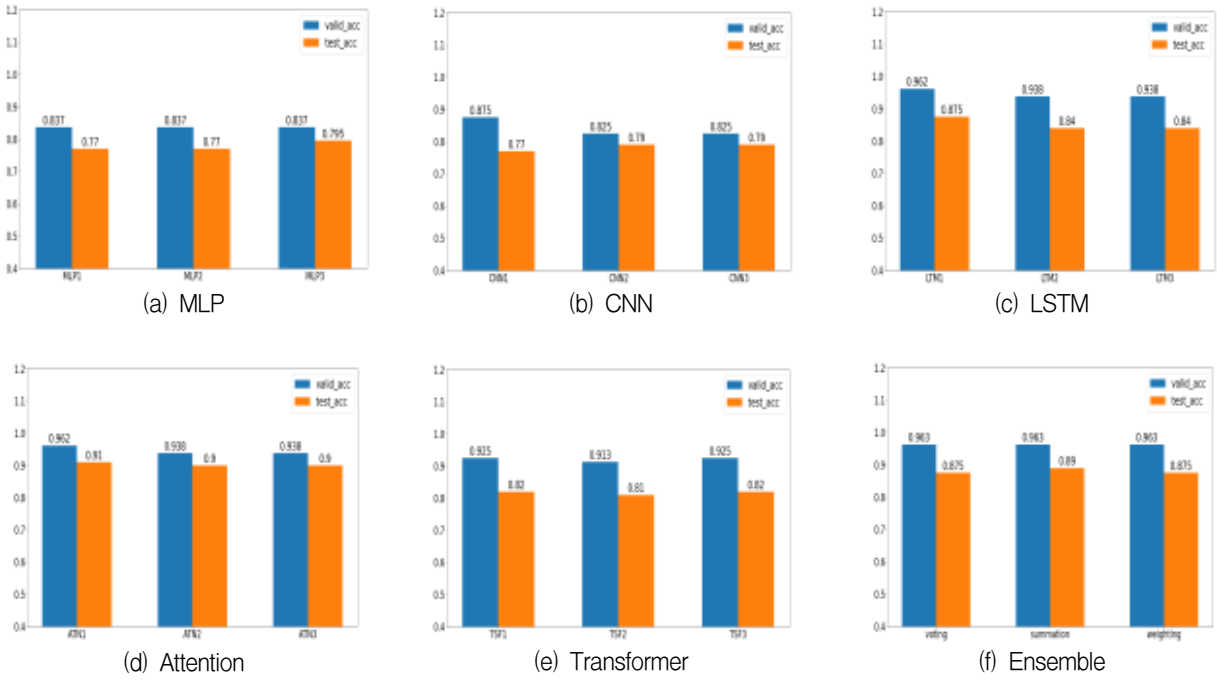


그림 1. 데이터셋 #1의 개별 모델과 앙상블 모델의 정확도
Fig. 1. Accuracy of individual and ensemble models of dataset #1.

Attention(d)의 최대 정확도는 0.957 그리고 Transformer(e)의 최대 정확도는 0.928 이다.

그림 2의 (f)는 총 15개의 분류모델을 결합한 앙상블 모델의 테스트 데이터의 분류 정확도를 나타낸다. Summation 앙상블 모델이 0.961으로 앙상블 기법 중 최대 정확도를 나타냈다.

데이터셋 #2 에서는 단일 모델들의 테스트 데이터 정확도가 0.838 ~ 0.957까지의 분포를 보였다. 또한 3가지 앙상블 모델에서는 0.936 ~ 0.961까지의 분포를 보였다. 데이터셋 #2 에서는 데이터셋 #1 과는 달리 앙상블의 정확도가 모든 개별모델 보다 높게 나왔다. 이러한 결과로 보아 데이터셋 #2 의 분류관점은 특허문서 내의 주요한 키워드의 존재여부를 이용한 분류(MLP, CNN)와 문장의 구성을 통한 분류(LSTM, Attention, Transformer)가 데이터셋 #2의 분류에 모두 유효한 분류 관점이라고 할 수 있다. 이러한 유효한 분류관점에 의해 분류된 결과를 통한 앙상블 모델이 각 개별모델에 비해 높은 정확도를 나타낸다고 할 수 있다. 즉 개별 모델들의 결과가 앙상블 모델의 결과에 긍정적인 영향을 끼쳤으리라 판단된다.

다. 데이터셋 #3의 개별모델과 앙상블 모델의 정확도

CPC H01계열과 C01계열의 데이터를 결합한 데이터가 데이터셋 3이다. 총 3,800개의 특허문서중 2,736건의 특허데이터로 훈련을 수행하였고, 560개의 테스트 데이터로 성능을 측정하였다.

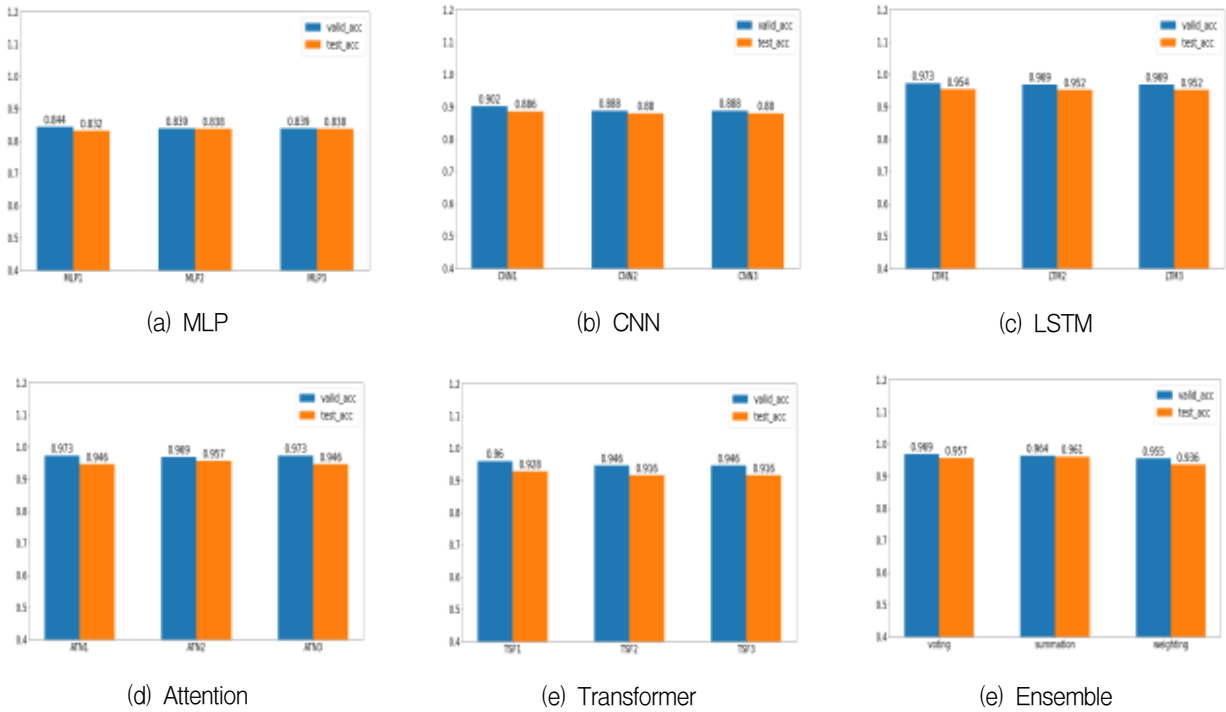
그림 3은 각각 단일 모델들을 사용한 분류기의 테스트 데이터의 추론 정확도를 나타낸다.

MLP(a)의 최대 정확도는 0.812, CNN(b)의 최대 정확도는 0.846, LSTM(c)의 최대 정확도는 0.928, Attention(d)의 최대 정확도는 0.938 그리고 Transformer(e)의 최대 정확도는 0.920 이다.

그림 3의 (f)는 총 15개의 분류모델을 결합한 앙상블 모델의 테스트 데이터의 분류 정확도를 나타낸다. Summation 앙상블 기법이 0.941으로 앙상블 기법 중 최대 정확도를 나타냈다.

데이터셋 #3 에서는 단일 모델들의 테스트 데이터 정확도가 0.812 ~ 0.938까지의 분포를 보였다. 또한 3가지 앙상블 모델에서는 0.929 ~ 0.941까지의 분포를 보였다.

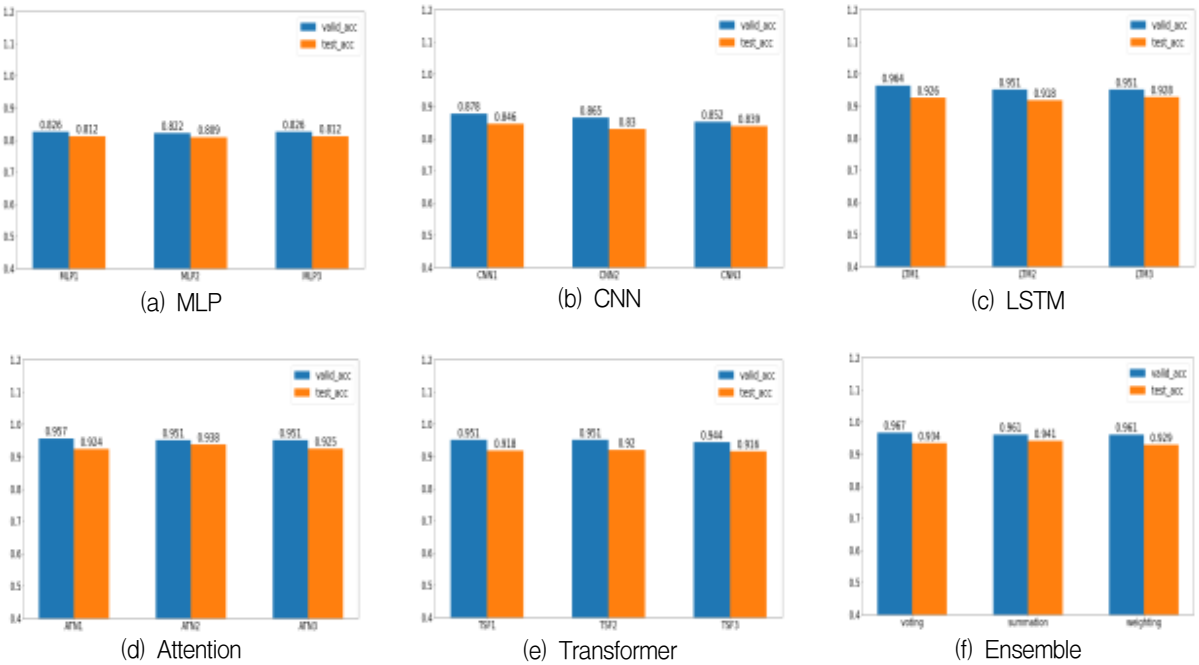
또한 3가지 앙상블 모델에서는 0.929 ~ 0.941까지의 분포를 보였다. 데이터셋 #3 에서는 데이터셋 #2 과 유사하게 앙상블의 정확도가 모든 개별모델 보다 높게 나왔다. 이러한 결과로 보아 데이테셋 #3 의 분류관점은 특허문서 내의 주요한 키워드의 존재여부를 이용한 분류(MLP, CNN)와 문장의 구성을 통한 분류(LSTM, Attention, Transformer)가 데이터셋 #3의 분류에 모두 유효한 분류 관점이라고 할 수 있다. 이러한 유효한 분류관점에 의해 분류된 결과를 통한 앙상블 모델이 각



(a) MLP (b) CNN (c) LSTM

(d) Attention (e) Transformer (e) Ensemble

그림 2. 데이터셋 #2의 개별 모델과 앙상블 모델의 정확도
Fig. 2. Accuracy of individual and ensemble models of dataset #2.



(a) MLP (b) CNN (c) LSTM

(d) Attention (e) Transformer (f) Ensemble

그림 3. 데이터셋 #3의 개별 모델과 앙상블 모델의 정확도
Fig. 3. Accuracy of individual and ensemble models of dataset #3.

개별모델에 비해 높은 정확도를 나타낸다고 할 수 있다. 즉 개별 모델들의 결과가 앙상블 모델의 정확도 증가에 긍정적 영향을 끼쳤다고 판단된다.

라. 전체 데이터셋의 정확도 비교

그림 4는 각 데이터셋 별로 개별모델의 최대, 최소 정확도와 앙상블 정확도를 비교하였다. 데이터셋 #1에서는 개별 모델의 최대 정확도가 앙상블의 정확도 보다 0.02 높다. 반면 데이터셋 #2와 데이터셋 #3에서는 개별

모델의 최대 정확도 보다 앙상블의 정확도가 각각 0.003, 0.002 높게 나타났다. 사용된 3개의 데이터셋 중 두 개의 데이터셋에서는 앙상블의 정확도가 개별모델의 정확도보다 높게 측정되었다. 개별 딥러닝모델이 앙상블 모델보다 높은 경우의 정확도차이는 0.02이다. 앙상블 모델과 개별모델의 최소 성능을 비교하면 앙상블 모델이 0.12 ~ 0.14 정도의 높은 정확도를 보였다. 개별 모델의 평균값과 비교하였을 때 앙상블이 항상 높은 성능을 보였다.

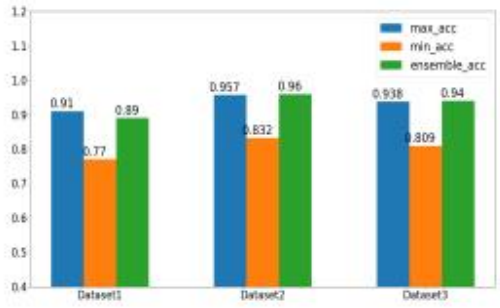


그림 4. 개별 모델의 최대, 최소 정확도와 앙상블 정확도 비교

Fig. 4. Comparison of maximum and minimum accuracy of individual models and ensemble accuracy.

본 연구에서 사용된 3개의 데이터셋중에 2개의 데이터셋(데이터셋 #2, 데이터셋 #3)은 앙상블의 정확도가 높았고 1개의 데이터셋 (데이터셋 #1)은 개별모델인 Attention 모델의 정확도가 높았다. 이러한 결과는 개별 데이터셋의 결과에서 언급한 것처럼 각 데이터셋의 분류 관점(키워드에 의한 분류, 문장 구성에 의한 분류)에 기인한 것으로 판단할 수 있다.

Ⅳ. 결론 및 토론

특허문서 분류의 특징은 다양한 관점의 분류와 클래스의 다양성에 있다. 본 연구에서 이러한 특징을 가진 특허문서 분류의 어려움을 극복하기 위해 다양한 딥러닝 모델을 적용하였고 또한 개별 모델들의 앙상블을 이용하여 3가지 데이터셋에 대해 성능을 비교하였다.

본 연구에서 선택한 3가지 데이터셋에 대해서 각 개별모델과 앙상블의 정확도를 비교하였을 때 2가지 경우에 있어서 앙상블이 가장 높은 정확도를 보였고 1가지 경우에서는 개별모델(Attention)이 앙상블 모델에 비해 높은 정확도를 보였다.

데이터셋 #3는 실제로 데이터셋 #1 과 #2를 결합한 데이터셋이다. 데이터셋 #1은 개별모델, 데이터셋 #2는 앙상블에서 가장 높은 정확도를 보였고 이 두 개의 데이터셋을 결합한 데이터셋 #3 에서는 앙상블 모델이 높은 정확도를 보였다. 이러한 결과는 다양한 관점으로 분류된 특허데이터셋에 항상 최고의 정확도를 내는 단일모델을 찾기는 힘들며, 다양한 단일모델의 앙상블을 통해서 개별단일 모델보다 높은 정확도를 내거나 최고 정확도를 내는 단일모델과 유사한 정확도를 나타낼 수 있을 것으로 보인다.

지도학습 기반의 특허문서 자동분류기가 실제 특허산업 현장에 사용되기 위해서는 사용자 관점의 다양한 분류체계에 대응, 데이터의 입력에서 출력까지 자동으로 이루어질 수 있는 시스템 그리고 분류의 변화에 대한 즉각적인 대응이 이루어 질 수 있어야 한다. 본 연구에서는 이러한 점을 극복하기 위해서 사용자 특허분류에 따른 모델의 생성이 아니라 기존 모델을 활용한 앙상블 모델을 제안하였다. 변화가 다양한 사용자정의 특허분류에 앙상블 모델을 사용한다면 데이터셋별 최적의 정확도를 내거나 최적의 정확도를 내는 단일모델에 근접한 정확도를 낼 수 있을 것으로 예상되며 이러한 앙상블 모델 방식이 특허문서 자동분류기를 상품화 하는데 적합할 것이다.

REFERENCES

[1] A. J. C. Trappey, F. C. Hsu, C. v. Trappey, and C. I. Lin, “Development of a patent document classification and search platform using a back-propagation network,” Expert Systems with Applications, vol. 31, no. 4, pp. 755 - 765, Nov. 2006, doi: 10.1016/j.eswa.2006.01.013.

[2] J. Yun and Y. Geum, “Automated classification of patents: A topic modeling approach,” Computers and Industrial Engineering, vol. 147, p. 106636, Sep. 2020, doi: 10.1016/j.cie.2020.106636.

[3] J. L. McClelland and D. E. Rumelhart, Parallel Distributed Processing : Explorations in the Microstructure of Cognition: Psychological and Biological Models. MIT Press, 1987.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, Proc. IEEE, vol. 86, no. 11, pp. 2278 - 2324, Nov. 1998, doi: 10.1109/5.726791.

[5] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735 - 1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[6] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, no. 3rd International Conference on Learning Representations
[7] A. Vaswani et al., “Transformer: Attention is all you need,” Advances in Neural Information Processing Systems 30, no. Nips, pp. 5998 - 6008, 2017.
[8] L. Bbeiman, “Bagging Predictors,” Machine Learning, 24, 123-140 (1996)
[9] European Patent Office, United States Patent and Trademark office., "Cooperative Patent Classification", [Online]. Available: https://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions/table

저 자 소 개

김 성 훈(학생회원)
1997년 경희대학교 화학공학과 학사 졸업(공학사)
2000년 경희대학교 화학공학과 석사 졸업(공학석사)
2018년～현재 한성대학교 스마트융합컨설팅학과 박사 과정
2016년～현재 (주)윕스 서비스개발부 부서장
<주관심분야 : 인공지능, 딥러닝, NLP, 특허분석>

김 승 천(평생회원)
1994년 2월: 연세대학교 전자공학과 학사 졸업(공학사).
1996년 2월: 연세대학교 전자공학과 석사 졸업(공학석사)
1999년 8월: 연세대학교 전기컴퓨터공학과(공학박사)
2000년 1월～2001년 1월 Univ. of Sydney Research Fellow
2001년 2월～2003년 8월 LG전자 DTV/DA 연구소 선임 연구원
2009년 7월～2010년 7월 Univ of Oregon 방문교수
2003년 3월～현재 한성대학교 IT 융합공학부 교수
<주관심분야 : 네트워크 보안, 블록체인 서비스, 사물인터넷 보안, 5G 이동통신망 서비스>