http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.26, No.2, May 2021, pp.19-43
https://doi.org/10.7838/jsebs.2021.26.2.019

공연예술에서 광고포스터의 이미지 특성을 활용한 딥러닝 기반 관객예측

Deep Learning-Based Box Office Prediction Using the Image Characteristics of Advertising Posters in Performing Arts

조유정(Yujung Cho)*, 강경표(Kyungpyo Kang)**, 권오병(Ohbyung Kwon)***

초 록

공연예술 기관에서의 공연에 대한 흥행 예측은 공연예술 산업 및 기관에서 매우 흥미롭고도 중요한 문제이다. 이를 위해 출연진, 공연장소, 가격 등 정형화된 데이터를 활용한 전통적인 예측방법론, 데이터마이닝 방법론이 제시되어 왔다. 그런데 관객들은 공연안내 포스터에 의하여 관람 의도가 소구되는 경향이 있음에도 불구하고, 포스터 이미지 분석을 통한 흥행 예측은 거의 시도되지 않았다. 그러나 최근 이미지를 통해 판별하는 CNN 계열의 딥러닝 방법이 개발되면서 포스터 분석의 가능성이 열렸다. 이에 본 연구의 목적은 공연 관련 포스터 이미지를 통해 흥행을 예측할 수 있는 딥러닝 방법을 제안하는 것이다. 이를 위해 KOPIS 공연예술 통합전산망에 공개된 포스터 이미지를 학습데이터로 하여 Pure CNN, VGG-16, Inception-v3, ResNet50 등 딥러닝 알고리즘을 통해 예측을 수행하였다. 또한 공연 관련 정형데이터를 활용한 전통적 회귀분석 방법론과의 앙상블을 시도하였다. 그 결과 흥행 예측 정확도 85%를 상회하는 높은 판별 성과를 보였다. 본 연구는 공연예술 분야에서 이미지 정보를 활용하여 흥행을 예측하는 첫 시도이며 본 연구에서 제안한 방법은 연극 외에 영화, 기관 홍보, 기업 제품 광고 등 포스터 기반의 광고를 하는 영역으로도 적용이 가능할 것이다.

ABSTRACT

The prediction of box office performance in performing arts institutions is an important issue in the performing arts industry and institutions. For this, traditional prediction methodology and data mining methodology using standardized data such as cast members, performance venues, and ticket prices have been proposed. However, although it is evident that audiences tend to seek out their intentions by the performance guide poster, few attempts were made to predict box office performance by analyzing poster images. Hence, the purpose of this study is to propose a deep learning application method that can predict box office success through performance-related poster images. Prediction was performed using deep learning algorithms such as Pure CNN, VGG-16, Inception-v3, and ResNet50 using poster images published on the KOPIS as learning data set. In addition, an ensemble

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2020S1A3A2A02093277).
* First Author, Master Student, School of Management, Kyung Hee University(yujung251@khu.ac.kr)
** Co-Author, Bachelor Student, School of Management, Kyung Hee University(kpkang0646@naver.com)
*** Corresponding Author, Professor, School of Management, Kyung Hee University(obkwon@khu.ac.kr)
Received: 2021-01-14, Review completed: 2021-04-07, Accepted: 2021-04-19

with traditional regression analysis methodology was also attempted. As a result, it showed high discrimination performance exceeding 85% of box office prediction accuracy. This study is the first attempt to predict box office success using image data in the performing arts field, and the method proposed in this study can be applied to the areas of poster-based advertisements such as institutional promotions and corporate product advertisements.

키워드 : 공연예술, 흥행 예측, CNN, VGG-16, Inception-v3, ResNet50
Performing Arts, Box Office Prediction, CNN, VGG-16, Inception-v3, ResNet50

1. 서 론

공연예술 기관에서 콘텐츠에 대한 흥행 예측은 공연예술 산업 활성화에 중요한 이슈이다. 과거 흥행 예측 방법으로는 전통적인 예측 기법인 다중회귀분석[14, 23]과 Bass 모형[14] 외에 Random Forest, KNN, 인공신경망[12] 등 데이터마이닝 알고리즘 등이 제안되어 왔다. 그 결과 Kim and Hong[22]의 연구에서는 판별 알고리즘으로 인공신경망, 다항로짓모형, 판별분석을 채택하였으며, 그 결과 인공신경망모형, 판별분석과 비교하여 다항로짓모형의 흥행 예측력이 더 우수하게 나타났다. 또한 Jeong and Min[12]의 전통적 데이터마이닝 알고리즘 비교 연구에서는 Random Forest가 우수한 것으로 나타났다. 그리고 최근에는 빅데이터 준비 및 활용 가능성이 커지면서 리뷰나 온라인 사이트상에서의 언급 특성을 흥행 예측에 활용하려는 연구들도 많아지고 있다.

그러나 최근의 연구는 영화산업, 공연산업, TV드라마, 미디어 등 특정 영역 또는 한국이나 미국 등 특정 지역에서의 흥행 예측이어서 모형화 결과를 일반화하여 적용하는 데에는 주의를 요하며, 동일한 영역과 지역이라고 할지라도 시간이 흐르면서 예측의 정확도가 하락하는 현상이 있다.

더욱이 지금까지의 흥행 예측 연구는 대부분 출연진[22]이나, 공연장소[25, 42], 일반인 평가[22], 전문가 평가[4] 등 정형적인 공연 특성에 의한 것이었다. 그러나 최근 이미지 기반의 판별 문제가 CNN 등 딥러닝 알고리즘에 의해 해결 가능해짐으로써 공연 포스터와 같은 이미지로도 흥행을 예측하는 시도가 가능해졌다. 예를 들어 포스터의 색채는 장르마다 특성이 있으며[17], 흥행을 예감할 수 있는 포스터의 특징들이 있는 것으로 알려져 있다[10]. 그러나 이러한 포스터 정보의 흥행 예측 유용성에도 불구하고 데이터 분석 기법을 통해 포스터의 특성을 분석하고 예측하는 연구는 거의 존재하지 않는다.

이에 본 연구의 목적은 공연관련 포스터 이미지를 통해 흥행을 예측할 수 있는 딥러닝 방법을 제안하는 것이다. 이를 위해 KOPIS에 공개된 포스터 이미지를 학습데이터로 하여 Pure CNN, VGG-16, Inception-v3, ResNet50 알고리즘들을 통해 예측을 수행하였다. 이에 다음과 같은 두 가지 연구 문제를 제기한다.

첫째, 비정형 데이터인 연극공연 안내 포스터로 연극의 흥행 여부를 판단할 수 있을까?

둘째, 포스터의 이미지 특성과 공연물에 대한 정형화된 데이터를 복합적으로 활용하여 흥행을 예측하는 것은 포스터 이미지 특성만이나 공연물에 대한 정형화된 데이터만으로 예측하는 것보다 더 성능이 우수할 것인가?

2. 선행연구

공연예술 분야에서 흥행성과에 대한 대리 지표(proxy)로는 총 관객수[18, 21]나 영화진흥위원회 Box Office 기준 누적 관객수[22], 개봉 기간[18], 상영 횟수[12] 등이 제시되고 있다. 마케팅에서는 구전효과가 흥행의 안정적으로 영향을 미치는 결정변수이므로[25, 27, 28], 구글이나 네이버 등의 언급 수 등은 유의한 설명 변수이다[42]. 구전효과는 평점과도 관련되는 것으로 국내외 영화 평점 5개 사이트들(네이버, 다음, 왓챠, IMDB, 로튼 토마토)을 분석한 것으로 보기도 한다[4]. 한편 장르, 배우, 감독, 시즌 등의 영향력은 스크린 수에 종합적으로 반영된 것으로 보고 스크린 수로 갈음하기도 한다[25]. 또한 장르와 배급사의 특성에 따라 개봉 방식과 상영 기간을 달리하는 차별화가 어느 정도 나타나고 있었으며, 유형별로 주말 관객 수의 변화 형태에서 의미 있는 차이를 보인다[15]. 이상으로 영화 관련 흥행 예측 결정요인들을 <Table 1>에 정리하였다.

그러나 공연예술 분야의 흥행 관련 연구는 성과 예측 모형을 직접 제시하는 것보다는 연극의 장르(유희적, 교육적, 교훈적 연극), 원작의 유무, 연극의 속성[24] 등 연극의 특성이 흥행에 미치는 연구가 대부분이다[21].

〈Table 1〉 Determinants of Box Office Prediction

Variable	Description	Citation
Nationality	Dummy(KR, US, etc.)	Kim and Hong[22]
Genre	Dummy(Comic, SF or Action, Horror or Thriller, Romance, Drama, etc.)	Kim and Hong[22], Kim[15]
Rating	Dummy(G, PG-12, PG-15, R,)	Kim and Hong[22]
Director	The average number of audience of the director's movie before 3 years	Kim and Hong[22]
Actor	The average number of audience of the actor's movie before 3 years	Kim and Hong[22]
Fame Of Main Actor	Appearance frequency of the actor corresponding to role in other movies(Directors with proven box office success prefer famous actors)	Chon et al.[5]
Distributer	The average number of audience of the distributor's movie before 3 years	Kim and Hong[22]
Screen Count	The number of screens released nationwide	Yu and Lee[42] Kim and Hong[22] Kwon[25] Park et al.[31]
Portal Ratings	Public's movie expectations rating in portal site(out of 10)	Kim and Hong[22]
Social Media	The number of search term documents among blog documents	Kim and Hong[22]
Google Index	The number of movie mentions in Google Trends	Yu and Lee[42]
Naver Index	The number of movie mentions in Naver Trends	Yu and Lee[42]
Auds Before Release	The number of audiences booked before the movie release	Kwon[25]
News Count	The number of articles mentioning the movie over a certain period on a specific site such as Naver news, etc.	Kwon[25]
NetizenRt BeforeRelease	Netizens' rating before the movie release (Movie ratings provided by Naver API)	Kwon[25] Cho et al.[3] Chon et al.[5]
Production Cost	Expenses incurred in making the movie	Park et al.[31]
Expert Ratings	Expert rating for the movie	Park et al.[31] Chon et al.[4]
Synopsis	Feature of synopsis	Lee and Kim[27]
Length of Review	The size of the reviews mentioned for the movie	Cho et al.[3]

Rhine and Murnin[32]은 공연예술의 참석률을 높이기 위한 목적으로 공연 시작시간, 행사기간, 발표일 등에 대한 선호를 조사하고, 결합분석을 통해 소비자들이 짧은 공연시간과 토요일 공연에 대한 선호가 높고 성별, 연령 등에 따라 선호하는 공연의 속성이 다름을 파악하였다. de Rooij and Bastiaansen[6]는 인터뷰를 통해 문화적, 사회적인 측면으로 소비동기를 심층적으로 분석하였다. 국내에서는 영화 시장보다 침체되어 있는 공연예술 분야의 문제점과 해결방안을 모색하기 위한 연구들이 주로 수행되었다[26]. 또한 Kwon et al.[24]는 관객의 유형에 따라 다르게 나타나는 공연 관람 구매 의도를 파악하여, 관람동기와 공연의 품질이 구매 의사에 영향을 미치고 관객 유형이 조절 효과를 보임을 검증하였다. Song[35]은 국내 공연사업에서 공연장에 대한 서비스 품질을 높이고 마케팅 전략으로써 활용하기 위해 시설, 인적서비스, 주차 시설 등 공연장 만족도에 유의한 영향을 미치는 요인들을 파악하였다. Kim et al.[20]은 선행연구를 통해 고객 만족 요인과 재관람 의도 요인을 파악하고, 관람 만족과 재관람 의도 간의 인과관계를 밝혔다. Yoo and Kim[40]은 연극의 품질을 독립변수로, 만족도를 종속변수로 설정한 연구모형을 바탕으로 타인이 공연 만족도에 조절변수 역할을 하는지 검증하였다. 이처럼 선행연구들은 주로 설문이나 면접 자료를 기반으로, 관객들이 공연에 대해 느끼는 만족감에 영향을 미치는 요인을 도출하기 위한 분석으로 이루어졌다고 할 수 있다. 즉, 흥행에 영향을 미치는 공연의 특성을 파악하는 연구는 지속적으로 이루어지고 있지만, 관련 특성을 이용한 흥행 예측 모형 연구는 거의 이루어지지 않았다. 예측 모형 연구가 활발하지 않은 이유는 영화만큼 예측 관련 데이터가 풍부하지 않고[35], 설문조사[24]나 면접 방법[6] 등을 통해 자료를 확보하기 때문에 성능이 좋은 예측 모형을 만들기가 용이하지 않기

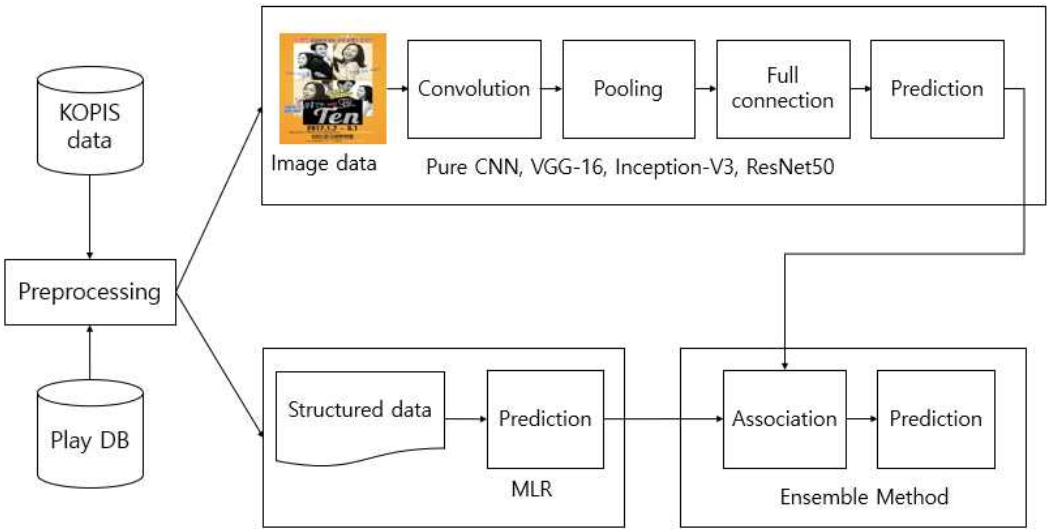
〈Table 2〉 Related Studies

Citation	Description	Methodology
Rhine and Murnin [32]	Preference for the start time, event period and presentation date is investigated in order to increase the attendance rate of the performance	question investigation, conjoint analysis, regression analysis
Lee and Chung [26]	Identifying the problem factors of the performing arts market and presenting solutions	question investigation, discriminant analysis, regression analysis
Kwon et al. [24]	Understanding the intention of purchasing a performance according to the type of audience	question investigation
de Rooij and Bastiaansen[6]	Analysis the motivation for consumption in performing arts	interview
Song [35]	A study on the factors affecting the satisfaction of the performance hall	question investigation
Kim et al. [20]	Identifying the relationship between the factors of watching the play and the client's intention to watch the show	question investigation, discriminant analysis,
Yoo and Kim [40]	An analysis of the influence of others on the satisfaction of play	question investigation, factor analysis

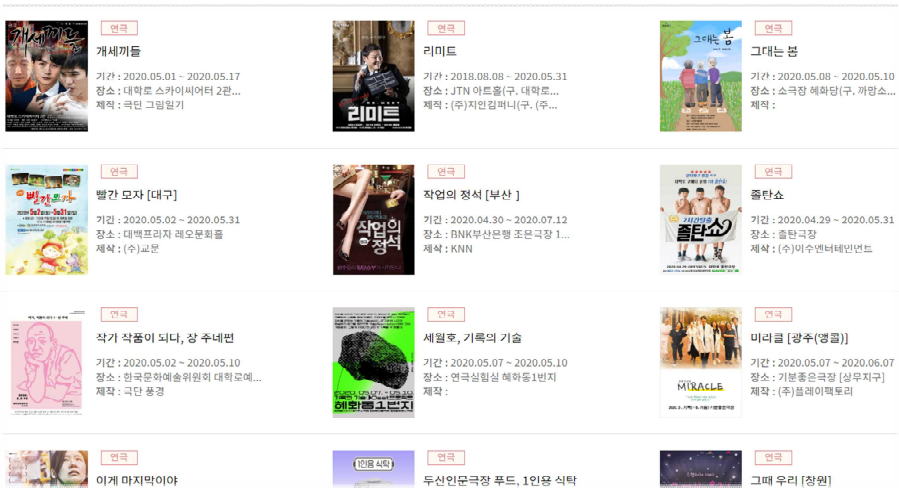
때문이다. 설문 기반의 자료는 공연에 대한 부정적 의견을 드러내기 쉽지 않고[19], 관객들의 성향과 사전지식 등에 따라 만족감을 느끼는 상황이 주관적일 수 있기 때문에 객관적인 데이터를 기반으로 한 흥행 예측 연구가 필요하다. 더욱이 기존의 흥행예측 연구에서 공연을 홍보하는 포스터의 이미지나 담겨진 내용이 흥행에 영향을 주는지에 대한 이미지 기반의 연구는 거의 존재하지 않는다. 최근 이미지 기반 추론을 강력하게 지원하는 CNN 계열의 딥러닝 기법이 소개되어 다양한 응용 방법들이 소개되고 있으나[8] 포스터 이미지에 근거한 추론 연구는 그 가능성에도 불구하고 거의 다루어지지 않았다.

3. 연구 방법

본 연구에서는 공연 포스터 이미지를 기반으로 흥행을 예측하기 위해 <Figure 1>과 같이 예측을 수행하였다. 먼저 연극 속성에 관한 오픈 데이터를 수집하고 전처리를 한 후에 포스터 이미지 데이터에 대해서 복수의 CNN 계열 알고리즘들로 흥행을 예측하고, 정형화된 데이터는 회귀분석을 통해 흥행을 예측하였다. 그리고 두 가지의 흥행 예측 결과를 바탕으로 앙상블기법을 사용하여 예측을 수행하였다. 앙상블기법은 최근 SNS나 신문기사, 영화 리뷰와 같은 비정형 데이터가 흥행 예측에 영향을 주는 요인[16]으로 사용되고 있는 상황에서, 정형 데이터와 비정형 데이터를 함께 예측모형에 사용할 수 있도록 하는 방법론이 될 수 있다. 또한 설문을 기반으로 한 선행 연구와 달리 연극공연에 대한 정량적 데이터를 사용하여 자료수집 비용과 시간을 단축시킬 수 있고, 공연 포스터 이미지 기반 흥행 예측 결과를 함께 반영하여 관객의 흥미를 유발시키는 비정형 데이터가 예측 모형에 포함되도록 하였다.



〈Figure 1〉 Research Process



〈Figure 2〉 Performance Materials Managed by KOPIS

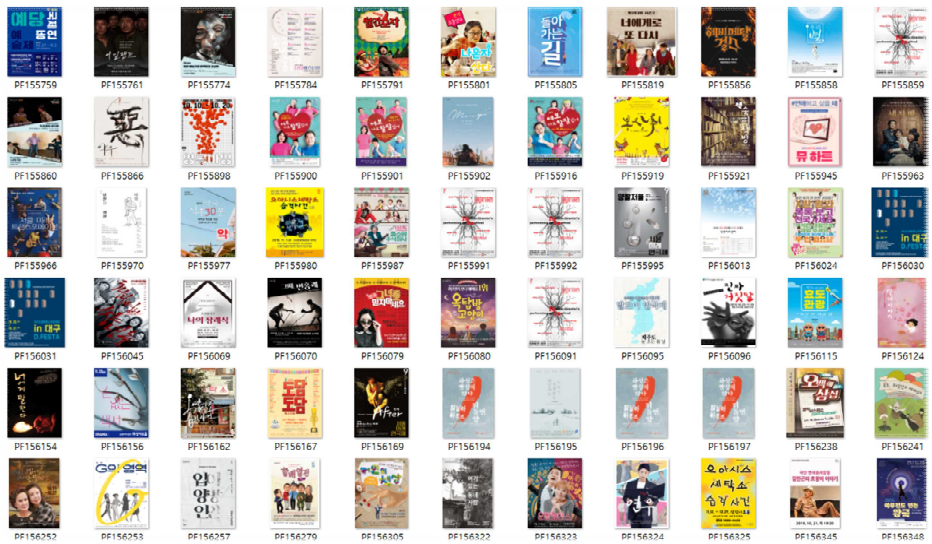
3.1 데이터 수집

연극공연에 대한 정보는 공연예술통합전산망인 KOPIS(http://www.kopis.or.kr)에서 관리하고 있다. 연극 관련 정보는 <Figure 2>와 같다.

KOPIS에서는 API Key를 발급하며 이를 활용하여 다음과 같이 연극정보, 연극상세정보 등을 확보할 수 있다. 크롤링 관련 코드는 <Appendix A>와 같다. 크롤링 결과 2016년부터 2019년 공연된 연극물 1,975건을 수집하였으며(<Figure 3> 참고), 이 중 연극ID가 중복되고 포스터 이미지가 없는 공연을 제외한 1,848건을 분석에 활용하였다. 이때 KOPIS를 통해 수집된 자료 중에서 연극 공연 안내 포스터는 URL 주소만을 제공하고 있기 때문에, 포스터 이미지 파일을 다운로드하여 확보하는 코드는 <Appendix B>와 같으며, 수행한 결과 확보한 포스터 이미지의 예는 <Figure 4>와 같다.

prfpdfrom	fcltynm	mt20id	prfnm	cate	prfdtcnt	entrpsnm	prfpdto
2019.01.10	똑딱의 극장(구. 해무소소극장)	PF162980	히네들	연극	4	연극집단 무궁댄미	2019.01.20
2017.01.07	KBS 수원아트홀	PF162923	탠: 열흘간의 비밀 [수원]	연극	1	극단 몸의 공작소	2017.03.01
2018.11.16	KNN씨어터	PF162849	라이어 1탄 [부산]	연극	4	KNN	2019.02.24
2019.06.15	나루아트센터	PF162833	탕신이 주인공 10주년 특별공연	연극	1	(주)도양앤터테인먼트	2019.06.15
2019.11.17	국립극장	PF162610	국립극장 창극아카데미(심화반) 수료공연	연극	1	국립극장	2019.11.17
2019.11.01	광명시민회관	PF162501	그 술의 심연	연극	2	한국예술종합학교	2019.11.01
2019.10.25	광명시 광명 5동 일대	PF162497	안녕, 광명	연극	7	고래리틀머웃는다	2019.11.03
2019.10.12	김정문화회관	PF160191	뽕 급은 포포마저씨 [김정]	댄극	2	극단 선돌	2019.10.12
2019.09.06	서귀포예술의전당	PF160168	세여자 [서귀포]	연극	3	극단 한강아트컴퍼니	2019.09.07
2019.11.26	서귀포예술의전당	PF160163	세여자 [서귀포]	연극	2	극단 한강아트컴퍼니	2019.11.27
2019.12.20	에저또소극장 [부산]	PF159556	검정고무신 [부산]	연극	1	극단 에저또	2020.01.19
2019.11.27	허수아비소극장	PF159019	알츠, 하이! 뭐?	연극	1	루씨드림 문화예술협동조합	2019.12.01
2019.11.30	한홀림소극장	PF158969	제9회 한홀림 골목연극제, 뭇생긴 남자	연극	1	극단 한홀림	2019.12.04
2019.11.27	소극장 6번출구	PF158821	극한 마미들	댄극	2	교육극단 미야기	2019.11.30
2019.11.28	레몬트리소극장	PF158750	유령 [부산]	연극	2	공연전문예알	2019.12.07
2019.11.27	세종문화회관	PF158728	개벽: 민초의 노래	연극	1	(재)전라남도문화관광재단(도립국악단)	2019.11.27
2019.11.28	명작극장 (구 아소홀레이씨어터)	PF158650	아버지를 찾습니다	연극	4	(사)한국연극배우협회	2019.11.30
2019.11.02	보림인형극장	PF158603	엄마 마중 [파주]	연극	2	인형극단 누렁소	2019.12.01
2019.11.22	Theater 연바람	PF158566	연극 있다-있다 페스티벌, 육주	연극	6	푸른연극마을	2019.11.30
2019.10.18	촉제소극장 [창원](구. 창원마트택들러)	PF158565	셀룸홈즈 [창원]	연극	20	(주)예술공장	2019.12.15
2019.11.27	동래문화회관	PF158515	해븐	댄극	1	동래문화회관	2019.11.27
2019.11.21	연극실험실 혜화동1번지	PF158512	혜화동1번지 기획전 가을페스티벌, 젊은연극: 가제	연극	9	풍악 프로젝트	2019.12.01

〈Figure 3〉 Examples of Collected Structured Data



〈Figure 4〉 Example of Collected Poster Images

한편 공연 배우의 유명도를 객관적으로 측정하기 위해 플레이DB 사이트(http://www.playdb.co.kr/artistdb)를 활용하였다. 이 사이트에는 뮤지컬 배우, 연극배우, 뮤지션, 성악가 등 각 장르별 출연 인물의 지명도를 주간조회순, 누적조회순별로 집계하여 공개하고 있다. 연극배우의 경우 총 10,560명에 대한 조회 결과가 순위별로 나와 있어, 해당 사이트를 근거로 하여 누적조회순으로 순위에 든 연극인이 주연인 경우 2점, 그렇지 않은 경우이거나 DB상에 출연진 정보를 공개하지 않은 경우에는 1점을 부여하였다. 출연진 정보를 제공하지 않았다는 것은 대부분 대학이나 일선학교에서 재학생들이 실험적으로 공연하는 경우가 많다.

3.2 데이터 전처리

CNN을 활용하여 포스터 이미지를 기반으로 흥행을 예측하기에 앞서 수집한 이미지 중에서 중복되는 이미지들은 제거하였다. 같은 연극공연이 여러 지역에서 공연되거나 재공연 되는 연극들은 포스터 이미지의 디자인이 동일한 경우가 있었다. 또한 연극의 제목과 일시, 장소 등의 정보를 담고 있지 않고 공연 포스터라고 보기 어려운 이미지들도 제거하였다.

KOPIS와 플레이DB 사이트로부터 얻은 공연에 대한 상세 정보와 더불어 연극 포스터 이미지의 속성에 따라 공연의 흥행 여부를 예측해 보기 위해 다음과 같은 방법들을 사용하여 포스터 이미지에 대한 정보를 추출하였다. 공연 포스터 이미지에서 추출한 첫 번째 데이터는 포스터에 포함되어있는 글자의 개수이다. 글자 정보를 추출하기 위해 Google Cloud Vision에서 제공하는 Vision API를 활용하여 TEXT_DETECTION 기법을 사용하였다.

Vision API에서는 이미지로부터 자동으로 텍스트를 감지하고 추출하는 OCR(Optical Character Recognition) 기술을 지원한다. 따라서 이 OCR을 지원하는 API를 활용해 포스터에 포함되어 있는 공연일시, 장소, 티켓 가격과 같은

정보를 활용하고자 하였다. 하지만 OCR의 성능은 이미지를 처리하는 시스템에도 영향을 받지만 이미지의 품질에 가장 큰 영향을 받기 때문에 해상도가 좋지 않은 파일들에 대해서는 텍스트의 높은 인식률을 보이지 못하였다. 그러나 저해상도의 포스터일 경우 사람이 보기에도 글씨가 작아서 잘 보이지 않거나 포스터의 내용을 정확히 인지하지 못하는 경우가 있음을 가정하여, 추출되는 텍스트의 내용이 아닌 문자의 개수를 분석에 활용하였다. 이미지로부터 텍스트 정보를 추출한 예는 <Figure 5>와 같다.

터는 놈! 털린 놈!
튀는 놈!
삐뚤어진 도둑들의
경성관 뒤집기 한 판!
201eR
경성 인사이드

〈Figure 5〉 Example of Original Poster Image and Extracted Text Information

두 번째는 포스터 이미지에 대한 색상 정보의 활용이다. 먼저 get_pixel 함수를 사용해 각 픽셀(pixel)마다 가지고 있는 RGB 색상 값을 0~255까지의 숫자로 추출하고 픽셀에서 가장 많이 포함하고 있는 색상이 무엇인지 파악하였다. 이를 기반으로 포스터 전체의 전반적인 색감을 수치로 보기 위해 이미지의 모든 픽셀에 대한 RGB 각 채널별 평균값을 구하여 포스터 이미지에 가장 많이 사용된 색감을 알 수 있었다. 0에 가까울수록 어두운 것을 의미하고 255에 가까울수록 밝은 것을 의미한다. 각각의 픽셀 값으로부터 RGB값을 추출하는 코드는 Appendix C와 같다.

한편, 해당 포스터에 대해 사람이 느끼는 밝기 정도를 계산하기 위해 RGB 값에 가중치를 다르게 주어 휘도(luminance) 값을 계산하였고, 채널의 산술 평균 값도 이미지에 대한 정보로 활용하였다. 결과 예는 <Figure 6>과 같다.

R (red)	G (green)	B (blue)	luminance	((R+G+B)/3)
118.518	54.366	53.353	67.932	80.272

〈Figure 6〉 Example of Original Poster Image and Extracted Color Information

마지막으로 포스터 디자인에 포함된 인물 수를 알아보았다. 포스터에 사람의 사진이 포함되어 있는지 알아보기 위해 사전 훈련된 딥러닝 모델을 사용하여 이미지 내 얼굴 검출을 수행하였다. 얼굴 검출에는 mtcnn(Multi-task Cascaded Convolutional Networks) 딥러닝 모델을 사용하였다[29]. mtcnn은 총 3가지의 CNN 구조를 직렬 프레임 워크에 포함하여 다른 알고리즘들에 비해 더 높은 얼굴 검출 정확도를 가지는 모델이다. mtcnn 모델은 CNN 프레임워크에 데이터를 넣기 전 이미지 사이즈를 재설정하여

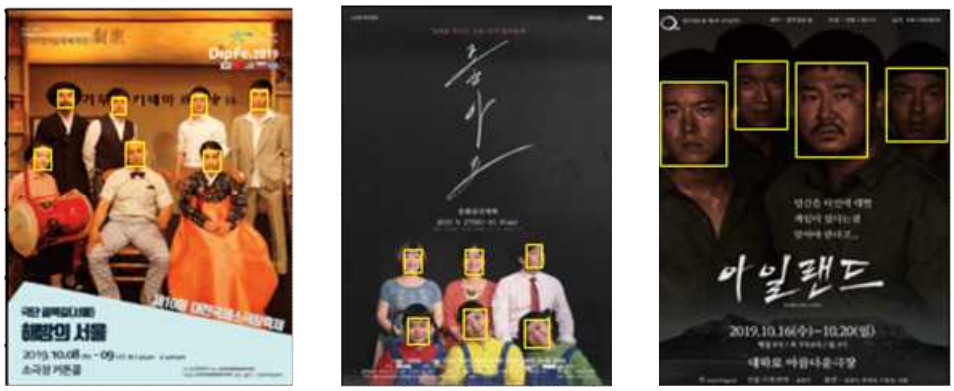
이미지 피라미드(image pyramid)를 만든다. 그리고 Proposal Network(P-Net)에 사이즈가 조정된 이미지들을 넣어 얼굴이 있을 것이라고 예상되는 후보 영역들을 추려낸다. 그리고 이렇게 찾은 여러 개의 Bounding box들에 non-maximum suppression(NMS)을 적용하여 후보군을 줄여나간다. 추가적으로 P-Net과는 다른 CNN 구조인 Refine Network(R-Net)와 Output Network(O-Net)을 차례로 거쳐 더 정교한 Face detection과 Face alignment를 수행한다. 본 연구에서는 파이썬에서 라이브러리로 제공하는 mtcnn을 이용하여 얼굴 검출을 수행하였다. mtcnn 라이브러리로 얼굴 검출기를 생성하면 box, checkpoint, confidence의 키(key)를 가진 json 객체가 출력되는데, 'box'는 이미지 내에 얼굴이 있는 좌표를 나타내고 'checkpoint'는 얼굴의 눈, 코, 입이 있는 위치의 좌표, confidence는 검출의 신뢰도를 나타낸다. 얼굴 이미지의 개수를 분석에 활용할 것이므로 검출된 객체의 개수만을 이용하였다. 검출 결과 예는 <Figure 7>과 같다.

한편, 종속변수는 추정관객수로 보았는데, 추정관객수는 공연횟수와 1회 당 최대 유치 관객수, 즉 공연장 규모의 곱으로 구했다. 연극의 경우 초대권에 의한 관객들의 구성이 매우 높아 매출(수익)액으로 판단을 하기에는 무리가 따르며 해당 정보도 공개되지 않는다는 특징이 있기 때문이다. 그런 후에 각각 Study1과 마찬가지로 관객수 500명 이상으로 추정되면 흥행 성공, 500명 미만은 흥행 실패로 예측하였다. 참고로 500명은 수집한 1,848건의 공연에서 상위 2/3 수준의 실적이다.

4. 연구결과

4.1 Study1: CNN을 활용한 이미지 기반 흥행 예측

Study1은 기존의 연구가 연극 관련 프로필 정보만을 활용하여 흥행 예측을 하려는 것에서 탈피하여 공연 포스터라는 이미지 정보로 흥행 예측을 할 수 있는 딥러닝 모형을 제안하고 그 효과를 측정하는 것이다. 딥러닝 모델은 이미 질병 판별[25], 얼굴 이미지를 활용한 성향 판별[13], 포스터 이미지를 통한 장르 판별[7] 등 다양한 판별 문제에서 활용되고 있다. 본 연구에서 사용하려는 Pure CNN의 기본 파라미터 값은 <Table 3>과 같다.



〈Figure 7〉 Example of Extracted Face Area in Poster Image

〈Table 3〉 Default parameter values of CNN model

Parameter	Value	Information
Class/Labels	2	box-office/box-office failure
Input Image Width	299×299	Fixed size for Inception-3
Training Set	800	
Validation Set	300	
Batch Size		The number of examples in a batch
Epoch	20	A full training pass over all of the samples in the dataset such that each sample has been processed once
Training Iteration	128	Training Set/Batch Size
Activation Function	ReLU, Elu, Selu Linear etc.	
Evaluation Criteria	Accuracy	Correct Predictions/Total Number of Examples

본 연구에서 학습에 사용되는 모델은 CNN을 기반으로 한 모델이다. 수집된 데이터는 추정관객수를 기준으로 관객수가 500명 미만일 경우 흥행실패, 500명 이상일 경우 흥행성공으로 분류하였다. 추정관객수는 공연횟수와 1회당 최대 유치 관객수, 즉 공연장 규모의 곱으로 구했다. 따라서 CNN 모델은 흥행 실패와 흥행 성공 두 개의 클래스로 분류되도록 모델을 설계하였다. CNN 모델을 활용하여 학습을 진행하는 코드는 <Appendix D>와 같다.

한편, 합성곱 신경망(CNN, Convolution Neural Network) 모델 중 이미지 분류를 더 정확히 할 수 있는 모델을 찾아 사용하기 위해 Pure CNN 외에 VGG-16, Inception-v3, Resnet-50을 사용하여 판별을 수행하고 이를 Pure CNN 성능과 비교하였다.

첫째, VGG-16 모델은 모든 convolution layers 층에 3×3 필터를 적용하여 층이 깊어짐에도 불구하고 분류 정확도가 개선될 수 있도록 구성한 모델이다[23]. VGG-16에서 16은 층의 개수를 의미한다. 총 13개의 convolution layers와 3개의 fully-connected layers로 구성되어 있고 stride와 padding 값은 모두 1로 고정되어 있으며 max-pooling 층의 stride는 2이고, 2×2 window를 사용하였다.

둘째, Inception-v3 모델은 구글에서 개발한 GoogleNet에서 조금 더 발전된 형태로 Inception module을 사용하여 학습해야 할 파라미터 수는 줄이고 정확도 성능은 더욱 높인 모델이다[27]. 사이즈(size)가 큰 필터를 사용하는 대신 사이즈가 작은 필터 여러 개를 사용하여 계산 비용을 줄였다. 따라서 42개의 layers를 가지지만 계산비용은 GoogleNet보다 2.5배 정도 높고 더 효율적인 성능을 보인다.

셋째, Resnet50은 residual learning 프레임워크를 사용하여 네트워크 깊이가 증가하여도 계산 복잡도는 많이 증가시키지 않고 더 좋은 정확도 성능을 나타내도록 한 모델이다[7]. 기존 multi-layer perceptrons(MLPs)는 입력값을 정답 레이블에 가깝도록 매핑(mapping) 해주는 함수 H(x)의 최적값을 찾는 방식으로 학습한다면, Residual learning은 잔차 함수 F(x) = H(x)-x를 줄여나가는 방식으로 네트워크를 학습한다.

모델 학습에는 Keras 내장 사전 훈련 모델을 사용하였다. 사전 훈련된 모델은 'ImageNet Dataset'을 기반으로 훈련되었으며 1000개의 클래스로 분류하도록 설계되어있다. 따라서 본 연구의 목적에 맞게 흥행 성공과 흥행 실패인 두 개의

클래스로 분류하도록 모델을 수정하였고, 모델별로 이미지의 입력 크기(input size)가 다르므로 포스터 이미지 사이즈를 학습 모델의 입력크기에 맞게 설정하였다. 또한 모형을 학습시키는 과정에 있어서 Batch Size는 mini-Batch로 진행하였다. 그 결과 <Table 4>처럼 Pure CNN 모델은 Batch size와 Epochs가 각각 20, 100일 때, VGG-16의 batch size가 30일 때, Inception-v3의 batch size는 40, ResNet50은 batch size가 50일 때 f1-score가 가장 높은 것으로 나타났다.

그러나 <Table 4>에서 ResNet50 모델을 제외한 세 개의 CNN 모델 모두 class '1'의 f1-score보다 class '0'의 f1-score가 현저히 낮은 것을 확인한바, 본 연구를 진행하는 데 이 같은 문제의 가장 큰 원인은 두 class의 데이터 개수가 불균형한 것에 있다고 판단하였다. 이는 흥행 성공에 해당하는 이미지 데이터 개수는 총 1,037개인 반면 흥행 실패에 해당하는 이미지 데이터 개수는 총 495개로 흥행 성공 클래스의 이미지보다 약 2배가량 적었기 때문이다. 따라서 이러한 데이터 불균형 문제를 해결하기 위해 over sampling과 under sampling을 수행하였다. under sampling은 흥행 실패 클래스와 같은 이미지 개수로 흥행 성공 클래스에서 이미지를 랜덤하게 495개 추출하였다. over sampling은 흥행 실패에 해당하는 이미지를 흥행 성공 이미지 개수와 같게 맞추기 위해 중복을 허용하여 542개 추출한 뒤 기존 흥행 실패 이미지 데이터에 추가하였다. 따라서 under sampling의 이미지 개수는 흥행 실패 이미지 데이터 495개, 흥행 성공 이미지 데이터 495개로 총 990개이고, over sampling의 이미지 개수는 흥행 실패 이미지 데이터 1037개, 흥행 성공 이미지 데이터 1037개로 총 2074개의 이미지가 학습에 사용되었다. over sampling 및 under sampling을 통해 도출해낸 결과는 <Table 5>와 같다. 먼저 under sampling 결과 <Table 5>에 나타난 것처럼, Pure CNN 모델의 경우 batch size가 20이고 epochs가 50일 때 f1-score가 가장 높았다. VGG-16은 batch size가 50, Inception-v3는 batch size가 20, ResNet50은 batch size가 40일 때 f1-score가 가장 높은 것으로 나타났다. 또한 over sampling 결과 <Table 6>에 나타난 것처럼, Pure CNN은 batch size가 20, epochs는 100일 때 가장 높은 f1-score를 보였다. VGG-16은 batch size가 30, Inception-v3의 batch size는 40, ResNet50은 batch size는 30일 때 f1-score가 가장 높았다.

〈Table 4〉 Results of CNN Model

CNN Model	Accuracy	f1-score		
		class : 0	class : 1	avg
Pure CNN	0.600	0.350	0.770	0.560
VGG-16	0.658	0.250	0.767	0.508
Inception-v3	0.664	0.278	0.746	0.512
ResNet50	0.322	0.492	0.038	0.265

〈Table 5〉 Results of Under Sampling

CNN Model	Accuracy	f1-score		
		class : 0	class : 1	avg
Pure CNN	0.571	0.520	0.580	0.550
VGG-16	0.571	0.649	0.512	0.581
Inception-v3	0.653	0.625	0.500	0.563
ResNet50	0.571	0.587	0.257	0.422

〈Table 6〉 Results of Over Sampling

CNN Model	Accuracy	f1-score		
		class : 0	class : 1	avg
Pure CNN	0.820	0.530	0.530	0.530
VGG-16	0.825	0.567	0.559	0.563
Inception-v3	0.699	0.609	0.552	0.580
ResNet50	0.501	0.671	0.038	0.356

4.2 Study 2: 회귀분석을 활용한 정형 데이터 기반 흥행 예측

Study 2는 오픈 데이터 중에서 정형화된 데이터로 흥행을 예측한 것이다. 이를 위해 로지스틱 회귀분석 방법을 활용하였다. 종속변수는 Study 1과 동일하게 공연 횟수 정보를 기반으로 추정 관객 수를 계산한 후에 관객 수 500명 이상으로 추정되면 흥행 성공, 500명 미만은 흥행 실패로 보았다. 독립변수는 기존의 문헌에서 흥행 예측과 관련했던 변수 중에서 오픈 데이터로 객관적으로 확보 가능한 변수들로 정했다. 그 결과 공연 시간, 입장 연령, 가격, 오픈런(open run), 요일 수, 공연장 수, 공연장 연혁, 공연장 규모, 공연장위치(수도권 여부), 배우유명도(조연의 배우 DB 내 존재 여부)를 선정하였다. 먼저 변수 간의 상관관계 분석을 수행하였으며, 그 결과는 <Table 7>과 같다.

〈Table 7〉 Result of Correlation Analysis

	Performance time	Admission age	Price	Open run	Number of performance day	Number of theater	History of performance hall	Performance hall size	Located in the capital area	Fame of actor	Number of performance
Performance time	1	.215**	.361**	.074	.340**	-.292**	.002	-.222**	.020	.048	.276**
Admission age	.215**	1	.273**	.049	.278**	-.126*	-.117*	-.188**	.337**	.015	.175**
Price	.361**	.273**	1	.153**	.667**	-.281**	-.039	-.310**	.266**	.026	.467**
Open run	.074	.049	.153**	1	.063	-.036	.001	-.028	.068	-.015	.263**
Number of performance day	.340**	.278**	.667**	.063	1	-.346**	-.063	-.358**	.365**	.033	.487**
Number of Theater	-.292**	-.126*	-.281**	-.036	-.346**	1	-.051	.824**	-.094	.055	-.148**
History of performance hall	.002	-.117*	-.039	.001	-.063	-.051	1	.029	-.103*	.004	-.036
Performance hall size	-.222**	-.188**	-.310**	-.028	-.358**	.824**	.029	1	-.175**	.013	-.162**
Located in the capital area	.020	.337**	.266**	.068	.365**	-.094	-.103*	-.175**	1	.084	.116*
Fame of actor	.048	.015	.026	-.015	.033	.055	.004	.013	.084	1	.067
Number of performances	.276**	.175**	.467**	.263**	.487**	-.148**	-.036	-.162**	.116*	.067	1
Estimated number of audience	.112*	-.013	.162**	.067	.129**	.366**	.050	.506**	.024	.135**	.463**

한편 추정 관객 수에 유의한 영향을 미치는 요인 판별을 위해 로지스틱 회귀분석을 수행한 결과 <Table 8>과 같이 공연 시간과 가격, 1주일 중에 공연하는 요일 수, 해당 공연장이 보유하고 있는 공연장의 수, 그리고 주연 배우의 유명도에 유의하게 영향을 받는 것을 알 수 있었다. 이는 대체로 흥행 가능성이 큰 연극은 오랜 기간 공연을 하고, 또한 그에 걸맞은 가격으로 입장료를 받기 때문으로 보이며, 특히 주연 배우의 유명도가 흥행에 영향을 준다는 부분도 이 분석을 통해 실증적으로 규명할 수 있었다. 한편, 장기 계약 공연을 의미하는 오픈런이 의외로 통계적으로 유의하지 않았는데, 이는 계약 기간이 공연의 흥행에는 크게 영향을 미치지 않고, 오히려 짧은 기간 공연하는 연극에 대한 선호[32]가 높을 수 있음을 추측해 볼 수 있었다.

이상과 같이 도출된 회귀분석 방법에 따른 추정의 정확도는 다음과 같았다. 우선 전체 데이터에서 랜덤하게 75%를 학습데이터로 하여 회귀식을 작성하고, 나머지 25% 데이터로 예측한 결과 Overall Accuracy는 68.4%, F1-score는 0.729의 정확도를 보였다.

〈Table 8〉 Result of Logistic Regression (Dependent variable: Estimated number of audience)

	Coefficient	S.D.	t
(Constant)	-8,052.743	1,662.533	-4.844***
Performance time	39.959	13.397	2.983***
Admission age	-71.902	57.595	-1.248
Price	0.061	0.026	2.373**
Openrun	3,266.908	3,174.272	1.029
Number of performance day	653.945	239.866	2.726***
Number of theater	1,034.618	101.959	10.147***
History of performance hall	38.270	26.428	1.448
Located in the capital area	-98.378	270.097	-0.364
Fame of actor	2,333.560	1,140.297	2.046**
F-value: 14.932***, Adj. R-Square: .251			

4.3 Study 3: 앙상블

Study 3은 네 종류의 CNN 방법과 로지스틱 회귀분석 방법의 앙상블로 예측하여 판정하는 것이다. 앙상블 판정 기준은 다음과 같다. 먼저 회귀분석으로 판정한 결과를 r이라고 하고(0 <= x <= 1), r의 범위를 -5에서 5의 범위로 늘린 것을 x라고 하자. 그리고 다음 식 (1)에 tanh 값을 적용한 z값을 얻었다.

$$z = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad (1)$$

그런 후에 z값이 $-\theta$보다 작으면 미흥행, θ보다 크면 흥행으로 확실히 판정하며, $-\theta \sim \theta$ 사이의 값이 되면 CNN의 판정 결과 및 변형된 x값이 0보다 크면 흥행, 0보다 작으면 미흥행으로 판정한 결과를 모두 합하여 다수결에 의하여 판정하도록 하였다. 그리고 정확도가 최대가 되도록 θ의 값을 최적화하였다. 그 결과 Overall Accuracy는 85.1%, F1-score가 0.782이었다. 이상의 방법들을 비교 요약한 것이 <Table 9>이다. Pure CNN, VGG-16, Inception-v3, ResNet50 등 본 연구에서 고려한 딥러닝 모델은 다양한 sampling 방법 중에서 가장 성능이 좋은 것으로 표기하였다. <Table 9>의 결과 본 연구에서 제시한 대로 이미지 기반 딥러닝과 수치 중심의 회귀분석을 함께 사용한 앙상블 방법이 가장 성능이 좋았다.

〈Table 9〉 Performance Comparison

Method	Accuracy	f1-score
Pure CNN (best)	0.820	0.560
VGG-16 (best)	0.825	0.563
Inception-v3 (best)	0.699	0.580
ResNet50 (best)	0.571	0.422
Regression	0.684	0.729
Ensemble (proposed)	0.851	0.782

5. 결 론

5.1 학술적 의의

본 연구는 공연 관련 빅데이터 중에서 그동안 잘 활용하지 않았던 포스터 이미지를 딥러닝 방법을 통해서 공연 예술물의 성과를 예측해 보았다. 그 결과 over sampling을 적용한 Pure CNN과 VGG-16이 로지스틱 회귀분석에 의한 기존의 판별 방법론보다 더 높은 정확도를 보였으며, 네 종류의 CNN 알고리즘과 로지스틱 회귀분석의 결과를 앙상블한 방법을 사용하여 정확도를 0.82에서 0.85 수준으로 높일 수 있었다. 이에 본 연구는 다음과 같은 몇 가지 학문적 시사점을 가진다.

첫째, 이미지 정보를 활용하여 해당 공연예술의 흥행 여부 예측을 시도한 최초의 연구이다. 이는 이미지 분석 및 판별에 그동안 CNN 계열의 딥러닝 알고리즘들이 기여해 왔는데, CNN 계열 알고리즘의 활용 영역을 문화예술 분야로 확대하였다는 의의가 있을 것이다. 특히 공연예술 분야에서의 딥러닝 활용 성공 사례로 보일 수 있다고 본다. 특히 본 연구에서는 포스터 이미지의 흥행성을 분석하는데, 이는 딥러닝의 이미지 이해(image understanding) 분야와 관련되는바[8], 이미지 이해 연구 분야에도 기여한다고 본다.

둘째, 공연예술에 대한 일반적인 흥행 예측 연구의 존재에도 불구하고[5, 21, 25, 27, 29, 35], 연극 공연에 적합한 별도의 예측 모형 연구가 없는 현 상황에서 연극 공연에 특화된 예측 연구를 수행했다는 점이다. 특히 흥행 예측 연구에 공연 특성들을 활용하거나[27, 35], 공연 포스터 이미지 정보를 활용할 경우 공연의 장르를 예측하는 연구[7]가 있었지만, 본 연구는 이미지 정보를 흥행 예측에 사용했다는 점에서 의의가 있다. 또한 앙상블 기법에 의하여 상호 보완적으로 흥행 예측을 하여 더 우수한 성능을 보일 수 있었다.

셋째, 이미지 판별에 주로 사용되는 CNN 알고리즘과 공연과 관련한 기존의 변수들을 함께 앙상블로 사용할 때 흥행 예측의 정확도가 제고되었다는 것은 향후 예측 연구에서 이미지 기반 및 수치 기반 기계학습 모델의 융복합 활용의 유용성을 보인 것이다. 이는 CNN과 SVM의 복합 활용의 효과를 보이는 최근 연구와도 일관성을 가지는 결과이다[29, 33]. 또한 학습 이미지들의 유사도에 따라 성능이 달라진다는 것을 확인했다는 점에서도 의의가 있다. 포스터 이미지의 경우 연극의 내용에 따라 디자인이 다양하므로 알고리즘을 학습시켜도 정확도가 높게 나오기 어렵다. 반면 포스터 이미지를 중복해서 over sampling을 하였을 때 정확도가 높아진 것으로 보아 데이터 샘플의 개수가 많아지고 포스터 이미지 디자인이 유사한 것들이 있어야 학습 성능이 높아진다는 것을 알 수 있었다.

또한 연극공연과 관련한 빅데이터를 활용하여 포스터 이미지를 딥러닝 방법을 통해서 공연 예술물의 성과를 예측한 본 연구는 연극학이나

예술경영 관점에서 혁신적인 의의를 가진다.

첫째, 연극공연의 포스터 제작은 주로 작품의 흥미와 더불어 관객층에 수요를 발생시키기 위함이기도 하다. 이러한 포스터 제작은 해당 연극공연작품의 질(quality)을 나타내며 이미지나 로고로부터 관람 의도를 이끌어내는 등[2, 41] 공연이 오픈되기 전의 광고효과도 있다. 특히 연극 포스터는 포스터 디자인의 문장과 사진, 삽화(illustration) 등 혼합된 요소를 배열하는 기술, 사진 영상학의 구도 비율과 명암이나 색채의 효과 등을 사용하여 연극작품이 어떠한 스타일로 제작된 공연 작품인지 그 정보를 내포하기도 한다[39].

둘째, 광고학적으로 볼 때 포스터 편집 디자인 요소가 광고학에 기반을 둔 교과서적인 기준이 있으므로, 본 연구에서 제안한 포스터 이미지 평가 딥러닝 모델은 광고디자인 요소와 성과별 상관관계를 학습할 수 있었을 것이다. 더 나아가서 광고디자인 요소별 공연 성과를 예측 가능할 수 있을 것이다. 광고학의 포스터 편집 디자인의 요소로는 면·선·질감 등의 형태, 색채심리, 균형 관계의 비례와 황금비(黃金比), 대립과 착시효과, 균형, 강약과 장단의 동적 변화인 율동, 조화, Unite, 수직축을 중심으로 한 좌우 동일한 형으로 시각적인 안정감을 주며 인간의 시각을 저항 없이 정리하는 대칭, 변형, 수평과 수직과 상승과 하강과 사선 등의 방향, 시각적인 무게중심의 조건 등을 꼽을 수 있다(https://operationnova.tistory.com/53). 만약 학습용 포스터에 대해서 위의 요인들을 레이블링할 수 있다면 딥러닝으로 특정 포스터에 위의 요인들이 의미 있게 존재하는지를 자동 판별함으로써, 이미지와 디자인요소, 그리고 디자인요소와 성과 예측의 인과관계 분석이 가능할 것이다.

5.2 실무적 의의

본 연구는 또한 실무가들에게 다음과 같은 몇 가지 의미 있는 결과를 내었다. 첫째, 공연 홍보 기관은 본 연구의 제안 방법을 활용하여 흥행의 가능성을 높일 방법으로 포스터 제작을 하는 데 도움이 될 것이다. 기관은 포스터 제작을 한 후에 본 연구에서 구축한 추론 모형에 대입하여 흥행 여부를 판단하고, 반복적인 작업에 의하여 가장 효과적인 포스터 디자인을 결정할 수 있을 것으로 기대한다. 예를 들어 본 연구에서 CNN 계열의 알고리즘을 작동시킬 때 이미지 자체의 특성뿐 아니라 이미지의 전체적인 색상 특성, 이미지 내 존재하는 텍스트의 특성, 이미지에 등장하는 인물의 특성 등을 복합적으로 활용하였는바, 포스터 제작 시 공연물의 특성(예: 장르, 연령대, 대표 배우의 지명도 등)과 색상, 인물 배치, 텍스트 노출 사이의 적합성이 중요하다는 것을 파악할 수 있을 것이다.

둘째, 영세한 극단들은 포스터 제작의 효율성을 높이는데 본 연구의 방법이 도움을 줄 수 있다. 대형기획사나 공연단체는 포스터 제작을 전문 광고회사에 부담 없이 맡긴다. 반면, 소형극장에서 공연되는 연극인 경우 포스터제작은 주로 광고학을 전공하지 않은 연극 스텝이나 연출, 기획자가 만드는 일도 있다. 또는 비전공자나 소규모 영세 홍보업체에 맡기며 연극연출가의 의견을 우선시하기 때문에 광고학적인 편집 디자인 요소들을 염두에 두지 않고 포스터를 만드는 경우도 있을 것이다. 따라서 본 연구의 결과는 이러한 영세 극단들에 성공적 홍보를 위한 유용한 단서를 제공할 것으로 본다.

셋째, 연극공연 관련 투자자는 이 방법을 활용하여 공연물의 기본적인 특성(장르, 연령대, 출연진 등) 외에도 포스터 이미지를 통해서 흥행 가능성을 판단하는 데 도움을 받을 것이고, 이에 근거하여 투자 여부나 투자액의 규모를 결정하는 의사결정에 도움이 될 것이다.

다음으로, 본 연구에서 제안한 방법은 빅데이터, AI를 공연예술 분야에 활용한 또 다른 분야를 보여준 것으로, AI 개발자들에게 도움을 줄 것이다. 그동안 AI, 빅데이터 기술은 공연장치 설정이나 관람객에 대한 성향 파악, 공연물을 안내하는 챗봇 개발 등에 활용되어 왔는데, 본 연구로 말미암아 이미지 성격의 공연 관련 데이터로 흥행 예측을 하는 정확도를 높일 수 있는 것이 밝혀진 것이다.

마지막으로, 이 방법은 공연의 주요 대상(target) 관객층을 고려한 맞춤형 포스터 제작에 활용될 수 있다. 공연 예매 사이트나 애플리케이션에서 관객들의 예매 이력과 같은 과거 데이터를 통해 접속자가 선호하는 공연들의 속성을 파악하고, 해당 공연의 포스터와 비슷한 디자인의 포스터를 노출시킨다면 관객의 표 구매 가능성을 높이는 데 도움이 될 것이다.

5.3 한계점

본 연구는 학술적, 실무적으로 의미 있는 결과를 도출했지만, 다음과 같은 몇 가지 한계점을 지닌다. 첫째, 공연의 흥행 여부에 대한 불확실성이다. 본 연구를 진행함에 있어 공연의 매출 및 투자 규모에 대한 데이터를 얻는 데 한계가 있었다. 따라서 본 연구에서는 공연장 좌석수와 공연 횟수를 곱한 추정 관객수 500명을 공연의 흥행 여부에 대한 기준으로 설정하였다. 그러나 실제 공연의 흥행 여부는 단순히 관객의 수가 아닌, 투자 대비 매출에 근거한다. 공연에 입장한 관객 수와 더불어, 투자액 대비 매출액을 데이터로써 활용할 수 있다면 공연의 흥행 여부를 더욱 정확히 예측할 수 있을 것으로 기대된다. 또한 비록 주관적이기는 하나 흥행에 대한 정의에 따라 500명의 기준도 변경하여 재분석해야 할 수도 있다.

둘째, 데이터의 생성 시기이다. 본 연구에 활용된 공연 관련 데이터는 공연예술통합전산망 KOPIS에서 가져온 2016~2019년까지의 데이터이며 시대적 상황을 반영할 수 있는 데이터는 포함되지 않았다. 이는 본 연구에서 도출한 방법이 COVID-19처럼 공연예술 분야에 타격을 줄 만한 시대적 상황에서 이미지 특성을 통해 공연의 흥행 여부를 예측하는 데 정확도가 다소 떨어질 수 있음을 의미한다.

셋째, 이미지의 특성이다. 2000년대 초반의 공연 포스터와 최근의 공연 포스터를 비교하면 알 수 있듯이, 공연 포스터의 디자인적 특성(폰트, 색상, 배치 등)은 당시의 디자인 트렌드를 반영한다. 따라서 관객 예측 모형을 학습하는 데 사용된 이미지 특성과는 다른, 독창적인 공연 포스터를 사용한 공연의 흥행 여부를 예측할 경우, 정확도가 다소 떨어질 수 있다.

마지막으로, 본 연구는 이미지의 의미를 추출하여 그 내용으로 흥행을 예측한 것이 아니라, 이미지 자체 특성을 가지고 예측한 것이다. 그리고 보조적으로 회귀분석 결과로 앙상블을 함으로써 흥행 예측의 정확도를 높이고, 그 주된 원인을 회귀분석 결과를 통해 찾은 것이다. 이미지의 어떤 부분으로 인해 흥행 가능성이 변동되는지를 보려면 추후 설명 가능한 AI(Explainable AI, XAI) 기술을 고려해야 할 것이다.

공연예술 분야는 인공지능. 특히 딥러닝이 적용될 수 있는 유력한 분야임에도 불구하고,

그동안 많은 연구가 진행되지 못했다. 특히 연극 분야에서는 영화나 드라마 등 인근 분야보다 적용이 활성화되지 않았다. 이에 본 연구는 연극 분야가 보유하고 있는 이미지를 포함한 오픈 빅데이터를 토대로 연극의 활성화를 위한 예측과 의사결정에 도움이 되는 딥러닝 방법을 제안했다는 데 의미가 있을 것이다. 이후에는 더 다양하게 딥러닝 알고리즘이 적용되어 연극 산업의 활성화에 기여할 것으로 본다. 특히 본 연구에서 진행한 관객 예측 모형은 흥행 혹은 실패로 나뉘는 판별형 알고리즘인데, 이후 본 연구에서 활용한 관객 예측 모형을 기반으로 생성형 알고리즘을 구축함으로써 문화예술 분야에 있어서 생성형 적대 신경망(GAN) 연구로 확장할 수 있는 가능성도 있을 것이다.

한국의 연극공연과 홍보 환경은 영세한 편이다. 특히 공연 홍보에 필요한 포스터 제작이 과학적이지 못하고 때로는 연극공연의 수요와 상반되게 만들어지기도 하는 실정이다. 이는 한정된 연극공연 예산 속에 제작비의 많은 부분을 차지하는 홍보 포스터임에도 불구하고, 연극 및 연기 연출의 수월성에 미치지 못하는 포스터 디자인으로 인하여 흥행의 실패를 불러올 수도 있기 때문이다. 그러므로 연극공연의 포스터 이미지를 기반으로 사전에 딥러닝 알고리즘이 포함된 앙상블 방법으로 공연 성과를 예측하는 것은 연극학이나 예술경영 관점에서 혁신적인 의의를 가진다고 하겠다.

References

[1] Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., and Saalbach, A., “Comparison of Deep Learning Approaches for Multi-label Chest X-ray Classification,” Scientific reports, Vol. 9, No. 1, pp. 1-10, 2019.

[2] Chae, S. H., “A Study on the Awareness of and Preference for Make-up Images in Performance Posters and the Intention to Attend a Performance,” Oriental Arts, Vol. 32, pp. 359-389, 2016.

[3] Cho, Y. H., Park, Y. S., and Kim, H. J., “Changes in Review Length based on the Popularity of Movies using Big Data,” Journal of The Korea Contents Association, Vol. 18, No. 5, pp. 367-375, 2018.

[4] Chon, B. S. and Yi, J. Y., “The effects of domestic and global movie rating services on economic performances of movies in korea,” Journal of Communication Science, Vol. 19, No. 4, pp. 227-253, 2019.

[5] Chon, B. S., Park, S. B., and Jo, A. R., “The effects of movie stars on box-office performances,” The Journal of Image and Cultrual Contents, Vol. 18, pp. 363-389, 2019

[6] de Rooij, P. and Bastiaansen, M., “Understanding and measuring consumption motives in the performing arts,” The Journal of Arts Management, Law, and Society, Vol. 47, No. 2, pp. 118-135, 2017.

[7] Gozuacik, N. and Sakar, C. O., “Turkish movie genre classification from poster images using convolutional neural networks,” In 2019 11th International Conference on Electrical and Electronics Engineering(ELECO), pp. 930-934, 2019.

[8] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S., "Deep learning for visual understanding: a review," Neurocomputing, Vol. 187, pp. 27-48, 2016.

[9] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[10] Im, E. K. and Seo, I. S., "Study of pictorial digital image shown on the movie posters -focused on top 10 national films shown in between 2003~2012," Journal of Digital Design, Vol. 12, No. 3, pp. 497-506, 2012.

[11] Jang, S., Wi, J., and Kim, Y., "Korean movie genre classification based on poster images with multi-label training using CNN," Proceedings of HCIK 2019, pp. 746-749, 2019.

[12] Jeong, C. M. and Min, D. K., "A Study on the Performance Evaluation of Machine Learning for Predicting the Number of Movie Audiences," The Jounal of Society for e-Business Studies, Vol. 25, No.2 , pp. 49-63, 2020.

[13] Joo, J., Steen, F. F., and Zhu, S. C., "Automated facial trait judgment and election outcome prediction: social dimensions of face," In Proceedings of the IEEE international conference on computer vision, pp. 3712-3720, 2015.

[14] Kim, B. K. and Lim, C. W., "Prediction of movie audience numbers using hybrid model combining GLS and Bass models," The Korean Journal of Applied Statistics, Vol. 31, No. 4, pp. 447-461, 2018.

[15] Kim, B. S., "Comparison of factors predicting theatrical movie success: focusing on the classification by the release type and the length of run," Korean Journal of Journalism and Communication Studies, Vol. 53, No. 1, pp. 257-287, 2009.

[16] Kim, H. Y. and Seo, D. H., "A topic modeling approach to the analysis of domestic performance," Official Journal of Korean Society of Dance Science, Vol. 36, No. 3, pp. 99-111, 2019.

[17] Kim, S. J., Choi, I. S., and Seok, S. M., "Comparative analysis of movie poster colors according to genre -On thriller/mystery, comedy & melo/romance movies (2010~2015)," Journal of The Korean Society of Illustration Research, Vol. 18, No. 53, pp. 33-42, 2017.

[18] Kim, S. Y., Lim, S. H., and Jung, Y. S., "A comparison study of the determinants of performance of motion pictures: art film vs. commercial film," The Journal of the Korea Contents Association, Vol. 10, No. 2, pp. 381-393, 2010.

[19] Kim, S. Y. and Yi, E. S., "A case study on big data analysis of performing arts consumer for audience development," Journal of the Korea Academia-Industrial cooperation Society, Vol. 18, No. 12, pp. 286-299, 2017.

[20] Kim, S. M., Kim, J. Y., and Choi, M. C., "The study of the spectating factors and

satisfactions effect on the re-spectate Intention," The Journal of Management, Vol. 32, pp. 23-38, 2011.

[21] Kim, T. H. and Shin, H. D., "The impacts of theme, genre, and existence of originals on the number of visitors of children performance: Focusing on the interaction effect of school periods," Journal of the Korea Academia-Industrial cooperation Society, Vol. 14, No. 10, pp. 4762-4768, 2013.

[22] Kim, Y. H. and Hong, J. H., "A study for the development of motion picture box-office prediction model," Communications for statistical applications and methods, Vol. 18, No. 6, pp. 859-869, 2011.

[23] Kim, S. H. and Han, J. M., "An analysis of motion picture box office performance : focusing on korean movies released in 2012," Journal of Social Science, Vol. 53, No. 1, pp. 191-214, 2014.

[24] Kwon, H. I., Jung, S. G., and Choi, Y. S., "Effects of factors of purchase intention to viewing performance by audience type," The Journal of the Korea Contents Association, Vol. 15, No. 2, pp. 139-150, 2015.

[25] Kwon, S. J., "Factors influencing cinema success: using news and online Rates," Review of Cultural Economics, Vol. 17, No. 1, pp. 35-55, 2014.

[26] Lee, E. M. and Chung, Y. K., "The study of the revitalization of performing arts audiences: focused on the obstructive factor and remedy for performing arts," Journal of Arts Management and Policy, Vol. 42, pp. 211-247, 2017.

[27] Lee, J. H. and Kim, Y. J., "A study on the factors in box-office of the audience-driven attributes," Journal of the Korea Entertainment Industry Association, Vol. 7, No. 1, pp. 1-9, 2013.

[28] Lee, J. K., Choi, Y. G., Jung, E. H., Jo, J. S., Ko, H. J., and Jo, H., "The effect of buzz volume on success of movie," The Journal of Internet Electronic Commerce Research, Vol. 19, No. 1, pp. 211-221, 2019.

[29] Panahi, M., Sadhasivam, N., Pourghasemi, H. R., Rezaie, F., and Lee, S., "Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression(SVR)," Journal of Hydrology, Vol. 588, pp. 125033, 2020.

[30] Park, S. and Shim, H., "Movie poster classification into genres via convolutional neural network," Proceedings of the The Korean Institute of Information Scientists and Engineers, pp. 890-892, 2017.

[31] Park, S. H., Song, H. J., and Jung, W. K., "The determinants of motion picture box office performance: evidence from korean movies released in 2009-2010," Journal of Communication Science, Vol. 11, pp. 231-258, 2011.

[32] Rhine, A. S. and Mumin, P. M., "Day, duration, and start time: are the arts providing what their audiences require?," Arts and the Market, Vol. 8, No. 1, pp. 19-29, 2018.

[33] Saif, M. A., Medvedev, A. N., Medvedev,

M. A., and Atanasova, T., “Classification of online toxic comments using the logistic regression and neural networks models,” In AIP conference proceedings, AIP Publishing LLC, Vol. 2048, No. 1, pp. 060011, 2018.

[34] Simonyan, K. and Zisserman, A., “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[35] Song, E. A., “A study of the factors explaining the long-run theatre,” Journal of the Korea Entertainment Industry Association, pp. 132-136, 2013.

[36] Sudharshan, P. J., Petitjean, C., Spanhol, F., Oliveira, L. E., Heutte, L., and Honeine, P., “Multiple instance learning for histopathological breast cancer image classification,” Expert Systems with Applications, Vol. 117, pp. 103-111, 2019.

[37] Szegedy et al., “Going deeper with convolutions,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

[38] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., “Rethinking the inception architecture for computer vision,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826, 2016.

[39] Yi, H. C., “Examining current challenges of Korean movie posters and seeking improvements for changes -Referencing the Hollywood movie poster production system-,” Journal of Communication Design, Vol. 53, pp. 252-263, 2015.

[40] Yoo, M. H. and Kim, J. W., “A Study on the factors influencing satisfaction of the play audience: focusing on the Influence of others,” Journal of Arts and Cultural Management, Vol. 10, No. 2, pp. 29-58, 2017.

[41] Yoon, J. W., Jeon, J. W., and Park, K. S., “Effects of font fit and familiarity on moviegoer’s attitudes and intentions,” Media and Performing Arts, Vol. 9, No. 1, pp. 1-28, 2014.

[42] Yu, J. P. and Lee, E. H., “A model of predictive movie 10 million spectators through big data analysis,” The Korea Journal of BigData, Vol. 3, No. 1, pp. 63-71, 2018.

[43] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y., “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, Vol. 23, No. 10, pp. 1499-1503, 2016.

〈Appendix A〉 Crawling Code

```
library(XML)

#### 연극정보 수집
read_xml("http://www.kopis.or.kr/openApi/restful/prfstsPrfBy?service=d3afbeb305014b438f9798e51a1a3b
6d&cpage=1&rows=399&stdate=20160601&eddate=20191130&shcate=AAAA")
doc<-xmlParse(pg)
xmldf <- xmlToDataFrame(nodes = getNodeSet(doc, "//prfst"))
xmldf
write.csv(xmldf, file = "연극정보.csv", row.names=TRUE)

#### 연극상세정보 수집
for (i in 1:nrow(xmldf)) {
  xmlurl <- paste("http://www.kopis.or.kr/openApi/restful/pblprfr/",
xmldf[i,3],"?service=d3afbeb305014b438f9798e51a1a3b6d", sep="")
  pg1 <- read_xml(xmlurl)
  doc<-xmlParse(pg1)
  xmldf_1 <- rbind(xmldf_1, xmlToDataFrame(nodes = getNodeSet(doc, "//db")))
}
write.csv(xmldf_1, file = "연극상세정보.csv", row.names=TRUE)

#### XML 해석
# LOADING TRANSFORMED XML INTO R DATA FRAME
for (i in 1:nrow(xmldf_1)) {
  xmlurl <- paste("http://www.kopis.or.kr/openApi/restful/prfplc/",
xmldf_1[i,18],"?service=d3afbeb305014b438f9798e51a1a3b6d", sep="")
  pg2 <- read_xml(xmlurl)
  doc<-xmlParse(pg2)
  xmldf_2 <- rbind(xmldf_2, xmlToDataFrame(nodes = getNodeSet(doc, "//db")))
}
write.csv(xmldf_2, file = "공연장정보.csv", row.names=TRUE)
```

〈Appendix B〉 Code for Collecting Poster Images

```
library(RCurl)
library(jpeg)
library(png)
library(imager)
library(stringr)
library(ggplot2)

data <- read.csv("연극상세정보.csv", header=T)
data$mt20id
data$prfnm
data$poster

for (i in 1:length(data$poster))
  {
    x <- data$poster[i]
    sign <- TRUE
    if ((str_sub(x, start= -3) == "png") || (str_sub(x, start= -3)== "PNG"))
    {
      filename = paste(data$mt20id[i],".png", sep="")
    }
    else if ((str_sub(x, start= -3) == "gif") || (str_sub(x, start= -3)== "GIF"))
    {
      filename = paste(data$mt20id[i],".gif", sep="")
    }
    else if ((str_sub(x, start= -3) == "jpg") || (str_sub(x, start= -3)== "JPG") || (str_sub(x, start=
-4)== "jpeg") || (str_sub(x, start= -4)== "JPEG"))
    {
      filename = paste(data$mt20id[i],".jpg", sep="")
    }
    else
    {
        print("No images found")
        sign = FALSE
    }

    if (sign == TRUE)
    {
      download.file(as.character(data$poster[i]),filename, mode = 'wb')
      plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
      rasterImage(jj,0,0,1,1)
    }
}
```

〈Appendix C〉 Extracted Color Information in Poster

```
X,Y= img.size
rimg = []
gimg = []
bimg = []
for x in range(X):
    for y in range(Y):
        pixelRGB = img.getpixel((x,y))
        R,G,B = pixelRGB
        rimg.append(R)
        gimg.append(G)
        bimg.append(B)
print("R : ",sum(rimg)/len(rimg))
print("G : ",sum(gimg)/len(gimg))
print("B : ",sum(bimg)/len(bimg))
```

〈Appendix D〉 CNN Code

```
#층 설정 및 활성함수
mode = keras.Sequential([
        Conv2D(64, kernel_size=(3, 3), padding='same', activation='relu'),
        MaxPooling2D(pool_size=(2, 2)),
        Conv2D(64, kernel_size=(3, 3), padding='same', activation='relu'),
        MaxPooling2D(pool_size=(2, 2)),
        Conv2D(64, kernel_size=(3, 3), padding='same', activation='relu'),
        MaxPooling2D(pool_size=(2, 2)),
        Flatten(),
        Dropout(0.25),
        Dense(128, activation='relu'),

        Dropout(0.5),
        Dense(2, activation='softmax')
        ])
# compile
mode.compile(optimizer='Adam',
                        loss='sparse_categorical_crossentropy',
                        metrics=['accuracy'])

# fit
history = model.fit_generator(
        train_generator,
        steps_per_epoch=35,
        epochs=100,
        validation_data=validation_generator,
        validation_steps=16)
```

저 자 소 개

조유정 (E-mail: yujung251@khu.ac.kr)
2020년 수원대학교 응용통계학과 (학사)
2020년～현재 경희대학교 빅데이터응용학과 석사과정
관심분야 데이터마이닝, 빅데이터분석, 딥러닝

강경표 (E-mail: kpkang0646@naver.com)
2016년～현재 경희대학교 경영학과 학사과정
관심분야 계량경제, 머신러닝, 빅데이터분석

권오병 (E-mail: obkwon@khu.ac.kr)
1988년 서울대학교 경영학과 (학사)
1995년 한국과학기술원 경영과학과 (박사)
1996년～2004년 한동대학교 경영경제학부 부교수
2004년～현재 경희대학교 경영대학 교수
관심분야 AI응용, 빅데이터분석, 소셜네트워크분석