KSME

http://doi.org/10.21797/ksme.2022.24.4.153

ISSN 1229-4225
eISSN 2671-9134

한국체육측정평가학회지. 제24권 4호, 2022, 153-165

야구 경기 승패 예측을 위한 합성곱 신경망(CNN) 최적화 연구*

김주학, 조선미**, 강지연 명지대학교

국문초록

최근 야구 종목의 인공지능 기반의 승패 예측 연구는 점진적인 발전을 보이며, 머신러닝에서 딥러닝까지 다각도의 연구가 진행되고 있다. 인공지능의 학습 성능은 설계된 모델의 하이퍼 파라미터에 값에 영향을 받으며, 최적의 하이퍼 파라미터 값을 찾는 것은 인공지능 모델 설계에서 필수적인 과정이다. 이 연구는 야구 경기의 승패 예측을 위한 합성곱 신경망(CNN)의 최적화 모델을 개발하는 연구로, 성능 최적화를 위해 하이퍼 파라미터 튜닝 방법을 적용하였다. 연구의 목적 달성을 위해 이 연구는 크게 세 단계로 구분된다. 첫 번째 단계는, Sequential 기반의 합성곱 신경망 모델을 1차 개발하는 단계이다. 두 번째 단계는 첫 번째 단계에서 개발한 모델의 하이퍼 파라미터 항목을 조절하여 성능을 비교하는 실험을 10회 진행하여, 최적의 하이퍼 파라미터 값을 찾는 단계이다. 실험결과 최적 성능의 하이퍼 파라미터는 필터(커널) 크기 '3*3', 학습비 '8:2', 배치 사이즈 '32', 에포크 '10'으로 결정하였다. 마지막 단계는, 결정한 하이퍼 파라미터를 적용하여 최적의 야구 승패 예측을 위한 합성곱 신경망 모델을 개발하는 단계로, 최종 모델의 성능은 정확도 '84.79', 정밀도 '84.84', 재현율 '84.58', F1 score '84.78'로 확인되었다.

주요어: 인공지능, 딥러닝, 빅데이터, 인공신경망, 경기분석

Ⅰ. 서론

빅데이터 시대의 도래와 더불어 데이터의 융합적 관점에서의 인공지능(AI: Artificial Intelligence)은 4차 산업 시대의 핵심 사안으로 대두되었다. 빅데이터와 인공지능을 핵심으로 하는 지능정보기술이 우리 삶의 다양한 분야에 보편적으로 활용됨으로써 인공지능과 관련된 기술은 지속적으로 진보하고 있다.

인공지능을 직관적으로 사변하면, '자동화', '예측', '인식', '판단', '특성추출' 등의 사고로 이어진다(조선미, 2022). 스포츠 영역에서 '예측'은 가장 중요한 키워드 중 하나이며 다양한 상황적 변수가 존재하는 스포츠 경기에서 데이터 분석이 경기력 향상과 직결된다는 것은, 주지의 사실이다. 스포츠 빅데이터와 인공지능은 경기기록으로 제한적이었던 스포츠 과학에 혁신을 가져오며 새로운 정보원으로서 시각과 통찰, 예측에 대한 가능성을 제시한다.

야구는 공격과 수비가 투구에 따라 비교적 독립적으로 이루어지기 때문에, 축구와 같이 연속적으로 이루어지는 경기에 비해, 기록의 생산과 관리에 효과적

* 이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019S1A5A2A03055515).

** 교신저자 조선미(liff99@gmail.com)
경기도 용인시 처인구 명지로 116, 명지대학교

인 스포츠이다. 이러한 이유로 야구 데이터를 대상으로 하는 분석은 타 종목에 비해 개괄적인 경기력 분석(이영훈, 2007; 이상인, 2015)에서 경기력과 연봉과의 관계와 같은 스포츠 산업 영역 연구(이만규, 2006; 장재열, 2016)까지 다각적인 연구가 진행되었다.

특히, 최근에는 빅데이터와 인공지능 기법을 활용한 야구의 승패 예측 연구가 점진적으로 수행되고 있다. 김원종, 최연식, 유동희(2018)는 데이터 마이닝 기법을 활용하여 한국프로야구의 승패 예측 모형을 구축하는 실험을 진행하였다. 이 연구에서는 승패 예측 모형 개발을 위해 여러 가지 인공지능 기법을 비교하였으며, 연구결과 인공신경망을 적용한 예측 모형이 가장 높은 성능을 보였다. 김태훈, 임성원, 고진광, 이재학(2020) 역시 머신러닝과 딥러닝 기법의 적용과 비교를 통해 한국프로야구의 승패 예측 모델을 개발하는 연구를 진행하였다. 이 연구에서는 앞서 기술한 김원종 외(2018)의 연구와 달리 머신러닝을 활용한 연구가 높은 성능을 보였다.

이러한 두 연구결과의 차이는, 인공지능을 적용한 야구 경기의 승패 예측이라는 주제는 동일하나, 적용된 알고리즘과 연구 대상 자료가 상이한 즉, 전혀 다른 연구 설계로부터 기인한 결과라고 볼 수 있다. 이는 인공지능 연구에서, 연구 자료의 수준과 범위, 종류, 그리고 사용 알고리즘과 하이퍼 파라미터(hyper parameter)에 따라 다른 결과가 도출될 수 있음을 시사하며, 인공지능의 예측 성능 향상을 위한 실험 연구의 필요성을 제시한다.

그 밖에 서영진, 문형우, 우용태(2019)는 스포츠 영역의 기존 인공지능 연구 자료에 대하여, 단편적인 데이터 사용의 문제점을 개진하였다. 이 연구에서는 정확한 승패 예측을 위해, 타자와 투수 간의 특성 및 팀의 컨디션 등 승패와 상관이 높은 기타 정보가 분석의 대상에 포함되어야 함이 논의되었다.

보편적으로 '예측'은 단편 정보가 아닌 종합적이고 통찰적 시각의 사고로 이루어진다. 인간의 뇌를 모방한 기술인 인공지능 역시 다양한 자료의 종합적 분석이 가능하도록 분석 자료를 수집하고, 적용해야 할 것이다. 그러나 선행연구를 종합하였을 때 대부분의 연구는 단편적인 경기 데이터를 사용하거나, 원데이터(raw data)가 아닌 정리된 통계데이터를 사용하는 특징을 보였다. 조선미(2022)는 스포츠 영역의 인공지능 연구에서 여러 가지 상황적 변수를 적용한 예측을 위해 원데이터 수집이 필수적이라고 개진한 바 있다.

야구는 순차적으로 투수가 공을 투구하고, 타자가 공을 타격하는 여러 동작으로 구성된다. 상황적 요인을 고려한 종합적인 예측을 위해, 연속적인 투구 및 투구에 따른 상황과 결과 자료의 수집이 필요하다. 이 연구는 합성곱 신경망(CNN: Convolutional Neural Network) 기법을 적용하여 승패를 예측하는 최적화된 인공지능 모형을 개발하는 데 연구목적이 있다. 이를 위해서는 야구 경기에서 발생하는 투수의 모든 투구와 투구시 발생하는 다양한 상황 정보가 필수적으로 수집되어야 한다. 합성곱 신경망 알고리즘은 주로 이미지, 영상, 자연어처리 분야에서 뛰어난 성능을 보이는 딥러닝 기법이다. 이에 이 연구에서는 야구의 경기단위별로 데이터 행렬을 구성하여 이미지로 변환한 뒤, 합성곱 신경망을 적용하여, 경기 전체의 상황을 고려한 승패 예측을 수행하였다.

합성곱 신경망을 포함한 인공지능 기법은 높은 정확도를 얻기 위하여 하이퍼 파라미터 값을 사용해야 한다. 하이퍼 파라미터는 사람이 직접 입력을 해야 하는 값인데 자료마다 최적값이 다르고, 아직까지 어떤 하이퍼 파라미터를 사용해야 하는지 이론적으로 밝혀진 바가 없다(이재은, 김영봉, 김종남, 2020). 딥러닝 모델의 예측 정확도는 하이퍼 파라미터라고 불리는 변수들의 초기 설정값에 크게 영향받는다. 그래서 사용자들은 딥러닝 모델에 다양한 하이퍼 파라미터 조합을 적용해서, 모델의 정답 예측도를 최대화 해주는 최적의 파라미터 조합을 찾는 절차를 수행한다(손재원, 2020). 딥러닝 모델들은 하이퍼 파라미터의 각각의 값에 따라서 상당히 민감하고, 모든 조합의 테스트는 불가하기 때문에, 대부분 연구자의 경험에 기반하

여 값을 선정하고 테스트한다(조억, 김성범, 2020).

즉, 인공지능 모델의 설정값인 하이퍼 파라미터의 최적화 값을 찾기 위해 다양한 실험을 반복수행한 뒤 개발 목적에 맞고, 예측 성능이 뛰어난 모델을 최종적으로 개발하는 것이 하이퍼 파라미터 최적화 방법의 보편적인 담론이다.

이렇듯 하이퍼 파라미터의 최적화에 대한 표준이 없는 이유는 딥러닝 모델의 목적이나 투입 자료의 수준, 범위에 따라 설정값이 모두 다르기 때문이다. 선행연구에서도 이와 같은 점에 대한 부분을 언급하며, 다양한 각도에서의 모델 개발의 최적화 방법에 대한 비교 연구가 필요함을 개진하였다.

이 연구는 야구 경기의 승패 예측을 위해, 하이퍼 파라미터를 조절하며 최적의 합성곱 신경망 예측 모형을 개발하는 연구이다. 연구를 통해 순차적인 야구 자료의 인공지능 분석 시 요구되는 적합한 하이퍼 파라미터 조합의 가이드를 제시한다. 아울러, 리소스와 시간이 제한된 환경의 실험에서 인공지능 연구의 초기값을 제공하는 데 이바지하고자 한다.

연구의 구성은 크게 이론 검토와 연구방법, 연구결과 및 결론으로 구성하였다. 이론 검토 부분에서는 이 연구의 핵심 이론인 합성곱 신경망과 최적화 방법에 대한 이론적 이해를 다루고, 연구방법 부분에서는 야구 경기의 승패 예측을 위한 합성곱 신경망 개발을 위한 과정을 포함하였다.

Ⅱ. 합성곱 신경망의 이론적 배경

1. 합성곱 신경망

합성곱 신경망(CNN: Convolutional Neural Network)은 딥러닝의 한 종류로 영상 및 이미지 인식에 주로 사용되는 알고리즘이다. 이미지의 특징점을 사전에 추출하지 않고, 입력층에 이미지 데이터를 직접 입력하여 자동적인 특징 추출이 가능한 기법이다(서기성, 2018). 이 과정에서 추출된 특징을 기반으로 영상이나 이미지를 예측, 분류한다.

보편적으로 사용되는 이미지 인식 및 분류 인공신경망은 어떠한 이미지를 1차원으로 변환하여 입력값으로 적용하기 때문에 그 과정에서 공간 정보의 유실이 필연적으로 발생한다. 이는, 2차원 이미지의 주변값의 소실을 의미하며, 복잡한 이미지의 예측이나, 공간 정보가 중요한 이미지의 경우 예측 성능의 저하를 야기한다.

반면, 합성곱 신경망은 일정 크기의 2차원 필터(가로*세로)가 일정 길이만큼 이동하면서 차원을 축소하는 합성곱 연산이 이루어지기 때문에, 비교적 이미지의 공간적인 구조 정보를 보존하면서 학습(조선미, 2022)하는 인공지능 기법이다. 〈그림 1〉과 같이 (가로*세로)로 구성된 필터가 일정 스트라이드 크기만큼 이동하면서 데이터를 추출하기 때문에 공간 정보의 손실을 최소화하며, 특정값의 주변값을 고려한 예측이 가능하다.

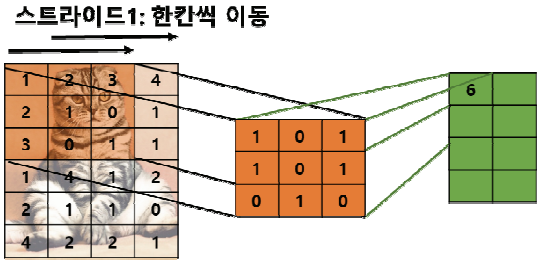


그림 1. (3*3)필터 합성곱 연산

합성곱 연산은 내적(inner product) 연산을 기본으로 수행하며 내적 연산의 수식은 다음과 같다〈식(1)〉.

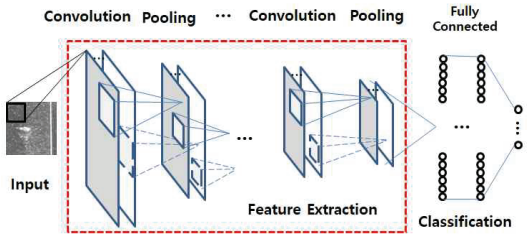
$$(a,b)=\sum_{i=1}^{n}x_iy_i \tag{1}$$
$$=(x_1\ x_2\ \dots\ x_n)\cdot\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}$$
$$=x_1y_1+x_2y_2+\dots+x_ny_n$$

이러한 합성곱 신경망의 특성은, 여러 상황을 종합적으로 고려해야 하는 스포츠 영역의 의사결정 및 예측에 효율적으로 응용된다. 야구 또는 스포츠에서 연속적으로 발생하는 어떤 이벤트는 해당 이벤트와 관련한 다른 속성, 또는 앞뒤의 상황과 상호 유기적인 관계를 맺고 있기 때문이다.

2. 합성곱 신경망의 구조

합성곱 신경망은 일반적인 인공신경망 앞에 하나 이상의 합성곱 계층을 추가한 구조이다〈그림 2〉.

층 사이의 노드 쌍 중 일부만 연결하는 컨볼루션(convolution) 층과 풀링(pooling) 층이 교대로 반복되며 특징 추출과 강인화에 관여하고, 후반부에 분류 목적에 사용되는 층 사이의 노드를 모두 연결하는 완전연결(fully connected)이 위치한다(서기성, 2018). 합성곱 신경망의 컨볼루션 층은 여러 개로 가중될 수 있으며, 각 단계를 구성하는 세부 속성 및 하이퍼 파라미터 등은 분석의 목표나 자료의 특성에 따라 최적화된 구조로 설계된다.



출처:CNN 구조의 진화 최적화 방식 분석

그림 2. 합성곱 신경망 구조

3. 합성곱 신경망 하이퍼 파라미터 최적화

합성곱 신경망 모형에서 높은 정확도를 얻기 위해서는 최적의 하이퍼 파라미터를 설정하는 작업이 필요하다. 하지만 높은 성능을 낼 수 있는 하이퍼 파라미터 값이 정확히 알려진 바가 없으며, 자료마다 최적의 하이퍼 파라미터 값이 달라질 수 있기 때문에 매번 실험을 통해서 찾아야만 한다(이재은 외, 2020).

표 1. 하이퍼 파라미터 최적화 방법

방법	내용
매뉴얼 서치	경험에 의거하여 직접 서치
그리드 서치	탐색 범위 지정 후 일정 간격으로 모든 조합을 조절하며 서치
랜덤 서치	탐색 범위 지정 후 무작위 서치
베이지안 최적화	베이즈 정리를 기반으로 목적함수를 최대화 또는 최소화하며 서치

대표적인 하이퍼 파라미터 최적화 방법은 〈표 1〉과 같다.

조억과 김성범(2020), 이재은 외(2020)는 매뉴얼 서치와 그리드 서치는 모든 조합의 경우의 수를 탐색하기 때문에 많은 시간과 계산량이 요구되어 비효율적이고, 랜덤 서치는 높은 정확도를 보장하지 못하는 연구의 제한이 있음을 논의하였다. 베이지안 최적화는 목적함수와 하이퍼 파라미터의 조합을 반복하면서 순차적으로 업데이트하는 방법으로 비교적 효율적인 최적화를 수행할 수 있다. 또한, 선험적 결과를 바탕으로 통계적인 모델을 구성한 뒤 목적에 맞는 하이퍼 파라미터 튜닝을 통해 고성능의 모델에 해당하는 하이퍼 파라미터의 최적값을 도출한다.

Aszemi & Dominic(2019)은 합성곱 신경망에서 실험 가능한 하이퍼 파라미터 튜닝 항목을 합성곱 신경망 구조영역과 학습영역 두 부분으로 구분하였다〈표 2〉〈표 3〉.

표 2. 하이퍼 파라미터 튜닝 항목-구조영역

항목	내용
Number of Filters	필터의 개수
Kernel Size	필터의 크기: (가로*세로) 커널
Hidden Layer	은닉층 개수
Activation	활성함수

표 3. 하이퍼 파라미터 튜닝 항목-학습영역

항목	내용
Learning rate	학습률: 학습-테스트 비율
Batch Size	배치 사이즈
Momentum	운동량: 학습의 안정성 및 속도
Optimizer	Adam, RMSProp, Nesterov

4. 모델 성능평가 지표

이 연구의 성능평가 지표는 정확도(accuracy), 정밀도(precision), 재현율(recall), F1 score 지표를 사용하였다〈식(2)〉, 〈식(3)〉, 〈식(4)〉, 〈식(5)〉.

$$정확도 = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$정밀도 = \frac{TP}{TP + FP} \quad (3)$$

$$재현율 = \frac{TP}{TP + FN} \quad (4)$$

$$F1\,score = 2 \times \frac{정밀도 \times 재현율}{정밀도 + 재현율} \quad (5)$$

* TP(True Positive) : 실제값 True & 예측값 True(정답)
* FP(False Positive : 실제값 False & 예측값 True(오답)
* FN(False Negative) : 실제값 True & 예측값 False(오답)
* TN(True Negative) : 실제값 False & 예측값 False(정답)
* F1 score : 정밀도와 재현율의 조화평균

Ⅲ. 연구방법

1. 자료수집

야구 경기의 승패 예측 합성곱 신경망 모델 개발을 위하여 미국 메이저리그(MLB)의 스탯캐스트(Statcast)에서 제공하는 2021시즌 전체 투구 자료와 경기별 승패 자료를 수집하였다. 투구 자료는 통계 자료가 아닌 원데이터(raw data)로 투수가 던진 각각의 모든 투구 정보를 포함한다. 경기별 승패 자료는 라벨링을 위한 목적으로 수집하였으며, 메이저리그 30개 팀의 2021시즌 모든 경기의 승패 정보이다.

미국 메이저리그 2021시즌 30개 팀의 정규시즌 경기는 총 2,430경기(한 경기에 홈/어웨이 두팀의 기록 수록)이며, 30개 팀의 모든 투수가 투구한 총 투구수는 719,684건이다.

2. 측정변수

승패 예측을 위한 합성곱 신경망 모델 개발에 사용할 측정변수는 총 25개로 〈표 4〉와 같다.

표 4. 측정변수

	항목	내용	
상황 변수 (n=10)	inning	이닝	이산
	home_score	투구 전 홈팀 점수	이산
	away_score	투구 전 어웨이팀 점수	이산
	stand	좌타/우타	명목
	P_throws	좌완/우완	명목
	at_bat_number	타석순석	이산
	pitch_number	투구번호	이산
	outs_when_up	아웃카운트	이산
	balls	볼카운트	이산
	strikes	스트라이크 카운트	이산
투구 변수 (n=9)	zone	투구존	이산
	release_pos_y	포수 관점 투구의 수평 움직임	연속
	release_pos_z	릴리즈 z좌표	연속
	release_pos_x	릴리즈 x좌표	연속
	release_speed	릴리즈 속도	연속
	effective_speed	타자 체감 속도	연속
	ax	x 투구 가속	연속
	ay	y 투구 가속	연속
	az	z 투구 가속	연속
결과 변수 (n=6)	pitch_name	구종	명목
	description	투구결과	명목
	post_fld_score	투구 후 수비팀 점수	이산
	post_bat_score	투구 후 홈팀 점수	이산
	delta_home_win_exp	투구 후 기대승리 변화	연속
	delta_run_exp	투구 후 기대득점 변화	연속

3. 자료정리

이 연구에서의 합성곱 신경망은 이미지를 분류하는 모델이다. 따라서 수집된 투구 자료의 이미지 변환을 위해, 팀별 경기의 투구 자료를 데이터 세트(행렬)로 정리하였다.

데이터 세트의 행(row)은 각각의 투구를 의미하고 열(column)은 측정변수 25종을 의미한다. 예컨대 미국 메이저리그 LA다저스팀의 한 경기에서 모든 투수가 총 150개의 투구를 했다면, 행(row)은 첫 번째 투구부터 150번째 투구까지, 열(column)은 각 투구와 관련된 변수 25종인 (150*25) 행렬로 구성하였다.

결과적으로, 30개 팀의 정규시즌 경기 162경기에 대하여 총 4,860개의 데이터 세트가 정리되었다. 정리된 데이터 세트는 모델 개발 과정에서 총 4,860장의 이미지로 변환하였고, 각 이미지는 승리(1), 패배

(0)으로 라벨링 하였다.

4. 합성곱 신경망 모델 개발 및 최적화

야구 경기의 승패 예측을 위한 합성곱 신경망 최적화를 위해 기본 모델을 설계한 뒤 반복적인 하이퍼파라미터 튜닝을 통한 최적화를 진행하였다. 〈그림 3〉은 개발 및 최적화 과정을 도식화한 것이다.

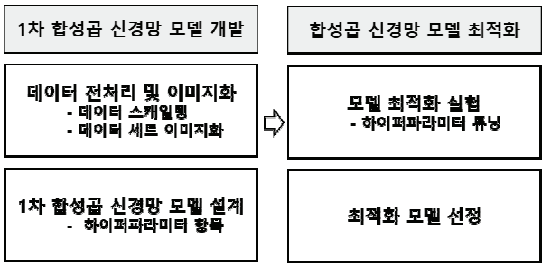


그림 3. 모델 개발 및 최적화 과정

합성곱 신경망 모델 구조는 합성곱 계층, 풀링계층, 활성화 계층의 구현과 컨볼루션 필터의 개수, 필터 사이즈, 패딩 여부, 스트라이드, 풀링레이어, 모델 최적화, 손실함수, 과적합방지, 배치, 에포크, 성능평가 등의 항목으로 구성된다.

5. 자료처리 및 모델 개발 환경

수집된 자료의 데이터 세트는 1차적으로 엑셀로 csv 파일로 정리하였고, 엑셀 파일을 이미지하여 모델을 개발하는 과정에서는 Python 언어를 사용하였다.

구체적인 연구의 물리적 환경은 다음 〈표 5〉와 같다.

표 5. 모델 개발 환경

항목	내용
운영체제	Window11
사용언어	Python 3.8.13
학습 및 검증	CPU i7-10870H (2.2GHz)
	GPU RTX3070
설계 및 개발 프레임워크	karas
	tensorflow 2.8.0
	pandas 1.4.2
	numpy 1.22.3
	opencv-python 4.5.5.64
	matplotlib 3.5.1
	scikit-learn 1.0.2
	seaborn 0.11.2

Ⅳ. 연구결과

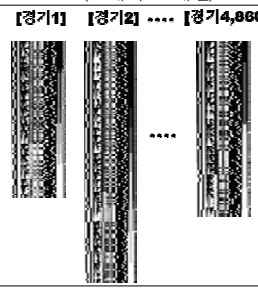
1. 데이터 전처리 및 이미지화

수집 정리된 데이터 세트의 변수 중 인공지능 분석 환경에 맞도록 데이터 전처리 과정을 수행하였다. 측정변수의 범위나 단위가 모두 다르기 때문에 정규화 표준화, 더미 변수화 등의 과정으로 정제하는 과정이다. 데이터 전처리는 연속 및 이산 변수는 '최대최소 정규화(MinMaxScaler)' 방법을 적용하였고〈식(6)〉, 명목 변수의 경우 더미 변수로 변환하는 '원핫인코딩' 방법을 적용하였다.

$$MinMaxScaler = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})} \quad (6)$$

데이터 전처리까지 완료된 데이터 세트는 null 값이 존재하지 않는 완전행렬로 구성하였다. 총 4,860개의 이미지가 생성되었으며 상세사항은 다음 〈표 6〉

표 6. 이미지 변환

항목	내용
사이즈	512*512(pixel)
채널	1(그레이 스케일)
이미지 변환	

가로길이: 측정변수(column)-길이가 같음
세로길이: 투구수(row)-경기당 투구수가 다르기 때문에 길이가 다름
≫ 이미지 상하좌우로 '0'값을 배치하는 패딩을 수행하여 같은 사이즈로 조절함

과 같다. 변환된 이미지는 실제 승패에 따라 [1]과 [0]으로 라벨링 하였다.

2. 1차 합성곱 신경망 모델 설계

수집된 자료를 기반으로 신경망 모델을 개발하였다. 개발 사용한 합성곱 모델은 조선미(2022)가 제안한 keras의 Sequential 합성곱 신경망 모델을 기본으로 하였다. Sequential 모델은 계층을 선형으로 쌓은 모델로 순차적인 데이터 분석에 유용하다고 평가되는 연구방법이다. 이는 순차적인 투구로 구성된 이 연구의 수집 자료의 특성을 반영한 결과이다.

〈그림 4〉는 이 연구에서 1차 설계된 합성곱 신경망 모델을 도식화한 것이다.

모델 구성의 최적화 함수로는 Aszemi & Dominic (2019)이 논의하였던, Optimizer 중 아담(Adam) 최적화 알고리즘을 적용하였다. 딥러닝 기반의 예측 모델은 예측값과 실제값의 차이를 비교하기 위한 손실함수를 모델에 적용한다. 손실함수는 목적함수 또는 오차함수라고 불리기도 하며 손실함수를 통해 '학습 과정에서 인공지능의 오측률이 어느 정도인가'에 대한 확인이 가능하다. 이 연구의 합성곱 신경망에서는 범주형 분류 예측 모델에 적합한 크로스-엔트로피(Cross-Entropy) 함수(조선미, 2022)를 사용하였다. 아울러, 과적합 방지를 위해 드롭아웃(Dropout)을 적용하였다.

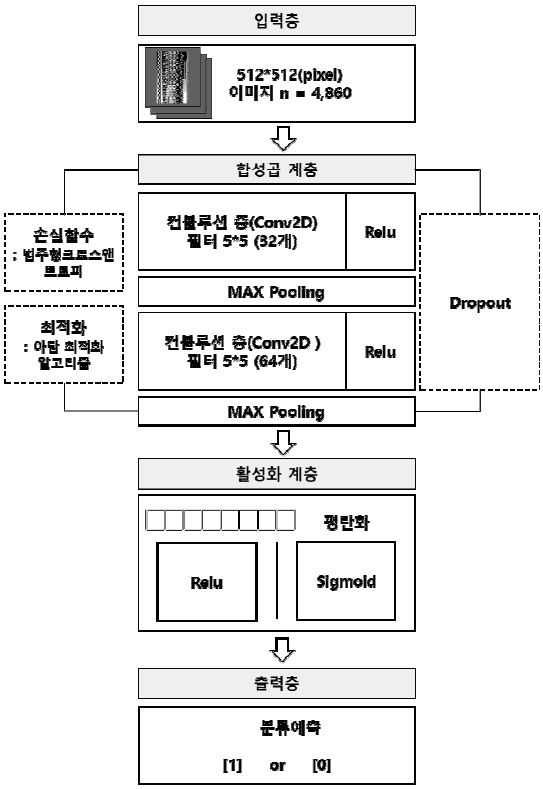


그림 4. 1차 개발 합성곱 신경망 모델 구조

1차 기본 모델의 하이퍼 파라미터 항목은 다음 〈표 7〉과 〈그림 5〉와 같다.

표 7. 1차 개발 합성곱 신경망 모델 하이퍼 파라미터

항목		내용
구조영역	컨볼루션 필터 개수(1층)	32
	컨볼루션 필터 개수(2층)	62
	커널사이즈	5*5
	활성계층	ReLU & Sigmoid
학습영역	학습비	8:2
	투입방식	랜덤
	배치 사이즈	20
	에포크	10
	Optimizer	아담(Adam)

전체 파라미터 : 128,052,482

```
Layer (type)                 Output Shape              Param #
=================================================================
conv_32 (Conv2D)             (None, 508, 508, 32)      832
maxpool_1 (MaxPooling2D)     (None, 254, 254, 32)      0
conv_64 (Conv2D)             (None, 250, 250, 64)      51264
maxpool_2 (MaxPooling2D)     (None, 125, 125, 64)      0
flatten (Flatten)            (None, 1000000)           0
dropout (Dropout)            (None, 1000000)           0
dense (Dense)                (None, 128)               128000128
output_layer (Dense)         (None, 2)                 258
=================================================================
Total params: 128,052,482
Trainable params: 128,052,482
Non-trainable params: 0
_________________________________________________________________
```

그림 5. 1차 개발 합성곱 신경망 모델 요약

1차 개발된 야구 경기 승패 예측 합성곱 신경망 모델에 대하여 수집 자료 중 120경기를 무선 표집하여

표 8. 1차 개발 합성곱 신경망 모델 학습 결과

항목	내용	
전체 데이터	120경기: 팀별 승패 이미지 240	
학습 : 테스트	192 이미지 : 48 이미지	
성능평가	정확도	53.12
	정밀도	51.08
	재현율	52.24
	F1 score	51.65

학습과 테스트(8:2)를 진행하였다〈표 8〉.

1차 개발된 기본 모델 학습 결과, 정확도 53.12, 정밀도 51.08, 재현율 52.24, F1 score는 51.65로 기록되었다. 전체적으로 50%의 확률의 예측률을 보였는데 이 모델이 패배 [0]과 승리 [1]의 분류 예측이라는 점을 고려했을 때, 성능 향상을 위한 하이퍼 파라미터 조절이 필수적으로 요구됨을 확인하였다.

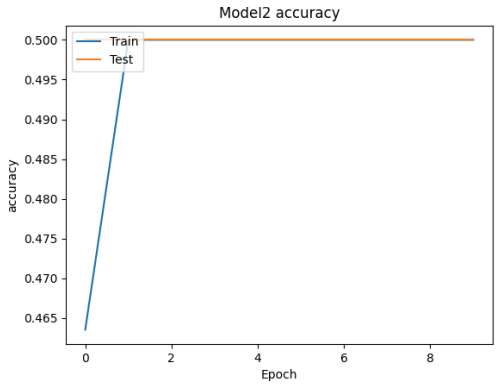


그림 6. 1차 개발 합성곱 신경망 모델 에포크와 정확도

〈그림 6〉은 1차 개발 합성곱 신경망 모델의 에포크와 정확도를 도식화한 것이다. 배치 사이즈는 전체 학습 데이터를 여러 작은 그룹으로 나눈 소그룹에 속하는 데이터의 수를 의미한다. 배치 사이즈에 따라 나눠진 소그룹 별로 학습 모형에 투입되기 때문에 배치 사이즈를 너무 크게 하면 속도 저하가 유발되고, 너무 작게 하면, 가중치 업데이트가 가중되므로 적절한 사이즈를 찾아야 한다. 에포크는 배치 사이즈와 관련이 있는 지표로 전체 학습 데이터가 신경망을 통과한 횟수를 의미한다. 1차 설계된 합성곱 신경망 모델의 에포크와 배치 사이즈에 따른 가중치는 다음과 같다〈표 9〉.

표 9. 배치 사이즈와 에포크에 따른 가중치

항목	내용
학습데이터	192
배치 사이즈	20
에포크	10
가중치	학습데이터 192÷배치 20 = 10번의 가중치
	가중치 10×에포크 10 = 총 100번의 가중치 업데이트

3. 합성곱 신경망 모델 최적화 실험

1차 개발된 기본 모델 학습과 동일한 120개의 경기를 무선 표집하여 최적화를 위한 실험을 진행하였다. 최적화를 위한 실험은 선행연구에서 조사된 하이퍼 파라미터 튜닝 항목을 조절하는 방법으로 학습과 테스트를 반복 진행하면서 가장 좋은 성능을 보이는 합성곱 신경망 모델을 도출하였다.

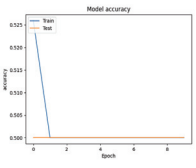
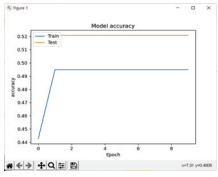
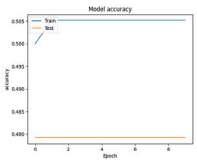
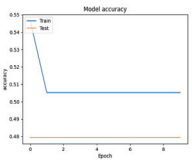
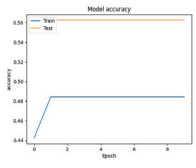
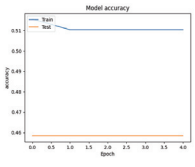
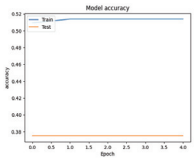
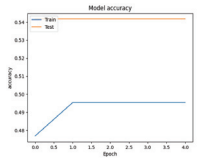
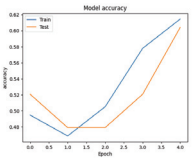
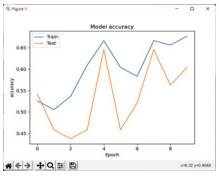
최적화 실험을 위해 조절된 하이퍼 파라미터 변수는 커널 사이즈, 학습비, 배치 사이즈, 에포크 4가지 항목이다.

〈표 10〉은 최적화를 위한 하이퍼 파라미터 튜닝 실험의 결과를 요약한 것이다. 실험결과 [실험9]와 [실험10]에서 최적의 성능을 보였다. 최적 성능의 하이퍼 파라미터는 필터(커널) 크기 '3*3', 학습비 '8:2', 배치 사이즈는 Sequential 모델의 기본값인 '32'로 결정하였다. 〈표 11〉의 에포크와 정확도 그래프를 보았을 때, [실험1]부터 [실험8]까지는 대부분

표 10. 하이퍼 파라미터 최적화 튜닝

실험	에포크	배치	커널	학습비	정확도
실험1	10	10	3*3	7:3	50.00
실험2	10	10	3*3	8:2	52.33
실험3	10	10	3*3	9:1	47.82
실험4	10	20	2*2	7:3	47.95
실험5	10	20	3*3	8:2	56.89
실험6	5	20	5*5	7:3	45.77
실험7	5	20	5*5	9:1	37.33
실험8	5	32	2*2	7:3	54.24
실험9	5	32	3*3	8:2	61.89
실험10	10	32	3*3	8:2	66.20

표 11. 하이퍼 파라미터 최적화 튜닝 실험의 에포크와 정확도

실험1	실험2	실험3	실험4	실험5
				
실험6	실험7	실험8	실험9	실험10
				

'3'이하의 에포크에서 이미 정확도가 결정되는 특징을 보이므로 에포크'5'값과 에포크'10'의 차이가 없다고 판단하였다. 그러나 [실험9]의 에포크와 정확도 그래프에서는 정확도의 값이 에포크가 증가함에 따라 상승곡선을 보이는 특징을 보였다. 따라서, 에포크 값은 '5'와 '10'을 최적화 값의 후보로 선정한 다음 모델 개발 시 분석 대상인 이미지 데이터의 수준과 범위, 양에 따라 조정하는 것이 합리적이라 판단하였다.

4. 최적화 모델 개발

앞서 수행된 모델 최적화 실험을 통해 결정된 하이퍼 파라미터 값을 적용하여, 야구 승패 예측을 위한 합성곱 신경망 모델을 개발하였다.

개발된 모델은 2021시즌 메이저리그의 2,430 경기 데이터를 모두 적용한 승패 예측 합성곱 신경망 모델이며, 최종 에포크 설정을 위해 에포크 값을 '5'와 '10'두 개의 값으로 조정하여 최적의 성능을 재확인하였다. 신경망을 통과한 횟수를 의미하는 에포크는 투입된 데이터의 양에 영향을 받기 때문에 최적화 실험에 적용된 120경기(240이미지)로는 에포크의 최종 결정에 제한이 있다. 따라서 2,430경기(팀당 4,860개)의 이미지를 투입하는 최종 모델에서 두 개의 에포크 값을 재확인 할 필요가 있었다.

〈표 12〉는 에포크 '5'와 '10'으로 조정한 성능평가의 결과이다.

표 12. 최적화 모델 에포크 설정 '5', '10'

항목		내용
전체 데이터		2,430경기: 팀별 승패 이미지 4,860
학습 : 테스트		3,888 이미지 : 972 이미지
성능평가 에포크 '5'	정확도	82.01
	정밀도	82.11
	재현율	81.95
	F1 score	82.03
성능평가 에포크 '10'	정확도	84.79
	정밀도	84.84
	재현율	84.58
	F1 score	84.78

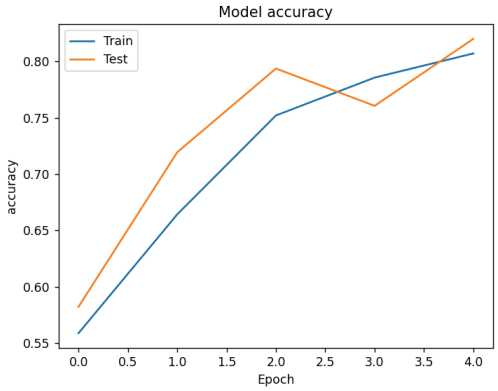


그림 7. 에포크 '5' 설정 정확도

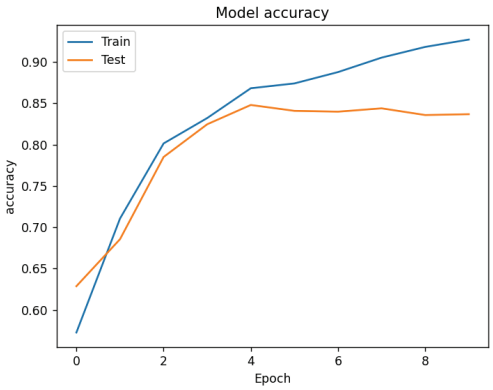


그림 8. 에포크 '10' 설정 정확도

성능평가 결과 에포크 '10'으로 설정한 모델의 예측 성능이 다소 높았음을 확인하였다. 에포크와 정확도 그래프에서도 에포크를 '5'로 설정한 경우 계속해서 상승하는 곡선을 보이는 반면, 에포크 '10'으로 선정한 경우에는 일정 학습 이후에 일관된 테스트 정확도를 확인할 수 있다〈그림 7〉, 〈그림 8〉.

아울러 학습에 대한 정확도는 에포크 10이 월등히 높으므로 최종 최적화 모델의 에포크는 '10'으로 결정하였다. 최적화된 하이퍼 파라미터 항목을 적용한 모델 구조를 요약하면 다음 〈표 13〉과 〈그림 9〉와 같다.

표 13. 최종 최적화 모델 구조

항목	내용
컨볼루션 필터 개수(1층)	32
컨볼루션 필터 개수(2층)	62
커널사이즈	3*3
활성계층	ReLU & Sigmoid
풀링계층	맥스풀링
학습비	8:2
투입방식	랜덤
배치 사이즈	32
에포크	10
스트라이드	1
손실함수	범주형크로스엔트로피
과적합방지	드롭아웃
이미지 채널	1
Optimizer	아담(Adam)

전체 파라미터 : 130,075,394

```
Layer (type)                 Output Shape              Param #
=================================================================
conv_32 (Conv2D)             (None, 510, 510, 32)      320
maxpool_1 (MaxPooling2D)     (None, 255, 255, 32)      0
conv_64 (Conv2D)             (None, 253, 253, 64)      18496
maxpool_2 (MaxPooling2D)     (None, 126, 126, 64)      0
flatten (Flatten)            (None, 1016064)           0
dropout (Dropout)            (None, 1016064)           0
dense (Dense)                (None, 128)               130056320
output_layer (Dense)         (None, 2)                 258
=================================================================
Total params: 130,075,394
Trainable params: 130,075,394
Non-trainable params: 0
```

그림 9. 최적화된 합성곱 신경망 모델 요약

V. 논의 및 결론

스포츠 영역에서 합성곱 신경망 방법을 활용한 연구는 이미지나 영상처리 분야에 비해 비교적 제한적으로 이루어졌다. 합성곱 신경망은 이미지 인식 또는 분류 분야에서 이미 뛰어난 성능을 증명하고 있다. 합성곱 신경망의 방법론을 스포츠에 적용하기 위하여, 야구 경기별 데이터를 이미지로 변환한 뒤 승리한 경기 이미지와 패배한 경기 이미지로 예측 분류하는 것이 이 연구에서 제안하는 합성곱 신경망 모델의 기본 개념이다.

이에 이 연구는 야구 경기의 승패 예측을 위한 합성곱 신경망의 최적화를 진행하였고, 모델 개발 과정에서 하이퍼 파라미터 최적화 튜닝 방법을 적용하였다. 연구의 목적 달성을 위해, Sequential 합성곱 신경망 모델을 기본으로 개발한 뒤, Aszemi & Dominic(2019)이 제안한 하이퍼 파라미터 항목을 조절하는 실험을 10회 진행하였으며, 그 결과를 바탕으로 최적의 합성곱신경망 모델을 개발하였다. 설계한 모델의 경우 최적 성능의 하이퍼 파라미터는 필터(커널)크기 '3*3', 학습비 '8:2', 배치 사이즈 '32', 에포크 '10'으로 결정하였다.

최종적으로 개발한 모델에 실제 2021시즌 메이저리그의 모든 팀의 투구 데이터를 적용한 결과 정확도 '84.79', 정밀도 '84.84', 재현율 '84.58', F1 score '84.78'의 성능을 확인하였다.

인공지능 연구에서 고성능을 보장하는 하이퍼파라이미터의 최적값을 찾는 것은 정해진 바가 없는 매우 어려운 일이며, 이재은 외(2020), 조억, 김성범, (2020), Aszemi & Dominic(2019) 등의 연구에서도 논의된 바 있다. 이는 개발하고자 하는 인공지능의 목표에 따라 대상 데이터, 알고리즘, 인공지능의 설계 구조가 모두 상이하기 때문이다. 특히, 같은 알고리즘을 사용하더라도, 어떤 데이터가 투입되느냐에 따라서 최적의 하이퍼 파라미터는 차이가 있다. 이에 대부분의 하이퍼 파라미터 최적화는 비슷한 주제의 선험적 연구결과를 초기 모델로 설정하고 값을 조금씩 변형해 가며 최적값을 찾아가는 실험을 진행한다. 따라서 이 연구는 스포츠 데이터를 활용한 합성곱 신경망 연구의 초기 연구로 하이퍼 파라미터 기본 설정값을 제시하는 데 의미가 있다.

종합적이고, 복합적인 판단이 필요한 스포츠 영역의 예측 인공지능은 때론 전체를 바라보는 거시적 시각이 필요하기도 하고, 또 다른 한편으로는 부분을 바라보는 미시적 시각이 필요하기도하다. 따라서, 모델에 투입되는 데이터의 특징에 따른 하이퍼 파라미터 연구는, 데이터 과학 영역의 발전 및 현장 적용의 효율적인 측면에서 파생가치가 높다. 이에 대해 한진, 이건명(2020)은 데이터 집합의 특징을 이용한 예측 모델 개발이 실용적 측면에서 필요함을 개진하였다. 이 연구에서는 야구 경기의 투구 데이터를 원데이터 형태로 순차 수집하였다. 또한, 공간정보의 손실을 최소화하는 합성곱 신경망 모델은 데이터를 분석하는데 주변값을 고려한 계산을 할 수 있는 기법이기 때문에, 이 연구의 데이터 집합이 가지고 있는 투구의 연속성이라는 특징과 종합적 분석을 동시에 반영한 최적화 방안을 제안하였다고 할 수 있다.

예원진, 이성노(2022)는 머신러닝 분류 모형의 예측 성능을 비교한 연구에서 이후 딥러닝 분야에서의 성능을 비교한 연구가 필요함을 제안하였는데, 이 연구의 합성곱 신경망은 대표적인 딥러닝 기반의 인공신경망 알고리즘으로 예원진, 이성노(2022)의 연구와 내용적 함의를 이룬다.

Ji, Zhanga, Shangb & Liu(2021)은 합성곱 신경망 기반의 인코더와 디코더 네트워크의 연구에서, 연구 설계의 확인을 위해 투입 변수의 조정을 통한 비교도 필요함을 주장하였다. 이러한 Ji et al.(2021)의 연구에서 논의된 바와 같이 측정변수를 조정한 예측 모델을 개발하는 후속연구도 진행되어야 할 것이다. 야구 경기의 승패 예측에 영향을 미치는 요인은 이 연구에서 설정한 25종의 측정변수 이외에 다른 상황이나 조건이 있을 수 있기 때문이다.

뿐만 아니라 합성곱 신경망 이외에 다른 딥러닝 알고리즘을 활용한 후속연구가 다양하게 진행되어, 스포츠 영역의 인공지능 예측 연구의 점진적 발전을 기대한다.

참고문헌

김원종, 최연식, 유동희(2018). 데이터 마이닝을 활용한 한국 프로야구 구단의 승패예측과 승률 향상을 위한 전략 도출 연구. **한국스포츠산업경영학회지, 23**(3), 88-104. http://doi.org/10.31308/KSSM.23.3.6

김태훈, 임성원, 고진광, 이재학(2020). 인공지능 모델에 따른 한국 프로야구의 승패 예측 분석에 관한 연구. **한국빅데이터학회지, 5**(2), 77-84. https://doi.org/10.36498/kbigdt.2020.5.2.77

서기성(2018). CNN 구조의 진화 최적화 방식 분석. **전기학회논문지, 67**(6), 767-772. https://www.kci.go.kr/kciportal/po/search/poSereArtiList.kci?sereId=SER000002663&volIsseId=VOL000099242

서영진, 문형우, 우용태(2019). 기계학습 기법을 이용한 한국 프로야구 승패 예측 모델. **한국컴퓨터정보학회논문지, 24**(2), 17-24. http://doi.org/10.9708/jksci.2019.24.02.017

손재원(2020). 클라우드 환경에서 딥러닝 하이퍼 파라미터 최적화 가속을 위한 GPU 스케줄링 프레임워크. 미간행 석사학위논문. 서강대학교 대학원. http://library.sogang.ac.kr/

예원진, 이성노(2022). 2022 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측 성능 비교. **한국체육측정평가학회지, 24**(3), 53-69. http://doi.org/10.21797/ksme.2022.24.3.005

이만규(2006). **세이버 매트릭스를 적용한 프로야구 타자의 경기력과 연봉과의 관계**. 미간행 석사학위논문. 국민대학교 스포츠산업대학원. http://image.kookmin.ac.kr/thesis/2006/361099.pdf

이상인(2015). **한국프로야구 기록 분석을 통한 투수의 경기력 지수 개발**. 미간행 석사학위논문. 명지대학교 기록정보과학전문대학원. http://dcollection.mju.ac.kr/jsp/common/DcLoOrgPer.jsp?sItemId=000000065869

이영훈(2007). 한국프로야구 경기력 결정요인에 관한 실증분석. **한국체육측정평가학회지, 9**(2), 63-77. https://doi.org/10.21797/ksme.2007.9.2.005

이재은, 김영봉, 김종남(2020). 합성곱 신경망에서 이미지 분류를 위한 하이퍼 파라미터 최적화. **융합신호처리학회 논문지, 21**(3), 148-153. http://doi.org/10.23087/jkicsp.2020.21.3.008

장재열(2016). **한국 프로야구 선수 경기력과 연봉에 대한 비용효율성 분석**. 미간행 석사학위논문. 고려대학교 대학원. https://dcollection.korea.ac.kr/public_resource/pdf/000000069522_20221206064818.pdf

조선미(2022). **인공지능 기반 한국프로야구 선발 투수 교체 예측 모형 개발**. 미간행 박사학위논문. 명지대학교 기록정보과학전문대학원. https://dcollection.mju.ac.kr/common/orgView/000000077152

조억, 김성범(2020). 인구 기반 트레이닝 밴딧 알고리즘을 이용한 강건한 하이퍼파라메터 최적화 기법. **대한산업공학회 추계학술대회 논문집**, 2136-2157. https://www-dbpia-co-kr-ssl.openlink.mju.ac.kr/journal/articleDetail?nodeId=NODE10505780

한진, 이건명(2020). 시계열 예측 모델의 특징 벡터 기반의 하이퍼 파라미터 최적화. **한국정보통신학회 종합학술대회 논문집, 24**(1), 78-30. https://www-dbpia-co-kr-ssl.openlink.mju.ac.kr/journal/publicationDetail?publicationId=PLCT00013797

Aszemi, N. M., & Dominic, P. D. D. (2019). Hyperparameter optimization in convolutional neural network using genetic algorithms. *International Journal of Advanced Computer Science and Applications, 10*(6), 269-278. https://pdfs.semanticscholar.org/c02f/877d81f487106cbd437f3f8d46b1496a897f.pdf

Huang, M. L.&Li, Y. Z. (2021). Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches. *Applied Sciences, 11*(10), 1-22. https://doi.org/10.3390/app11104499

Ji, Y., Zhanga, H., Zhangb, Z. & Liu, M. (2021). CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Information Sciences 546*, 835-857. https://doi.org/10.1016/j.ins.2020.09.003

저자정보

김주학(Joo-Hak KIM)
명지대학교 스포츠학부 교수
kjhak@mju.ac.kr

조선미(Sun-Mi CHO)
명지대학교 스포츠기록분석연구센터 분석팀장
liff99@gmail.com

강지연(Ji-Yeon KANG)
명지대학교 스포츠기록분석연구센터 분석부팀장
jyconniek@gmail.com

논문투고일 2022년 12월 06일
심사완료일 2022년 12월 27일
게재확정일 2022년 12월 28일

Abstract

The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science. 2022, 24(4), 153-165

A Study on Optimization of Convolutional Neural Network (CNN) for Win-Loss Prediction of Baseball Game

Joo-Hak KIM, Sun-Mi CHO, Ji-Yeon KANG *Myongji Univ.*

Recently, artificial intelligence-based win-loss prediction study in baseball is gradual development from machine learning to deep learning. The training performance of artificial intelligence is affected by the values of the hyper parameter of the designed model. Therefore, finding optimal hyper parameter values is essential in artificial intelligence model design. This study is to develop an optimization model of a Convolutional Neural Network(CNN) for predicting the win/loss of a baseball game, and the hyper parameter tuning method was applied for performance optimization. This study consists of three steps. The first step is to develop the Sequential-based Convolutional Neural Network model. The second step is to find the optimal hyper parameter value by conducting 10 experiments to compare the performance by adjusting the hyper parameter values of the model developed in the first step. As a result of the experiments to compare, the optimal performance hyper parameter values were determined as filter (kernel) size ‘3*3’, learning ratio ‘8:2’, batch size ‘32’, and epoch ‘10’. The last step is to develop a Convolutional Neural Network model for optimal win/loss prediction by applying the determined hyper parameter values. The performance of the final model was confirmed with accuracy ‘84.79’, precision ‘84.84’, recall ‘84.58', and F1 score ‘84.78’.

Keywords: Artificial Intelligence(AI), Deep Learning(DL), bigdata, artificial neural network, performance analysis