2022년 K-water 대국민 빅데이터 공모전 수행 결과보고서

제 목	심층신경망(DNN)을 활용한 해파리 출현 및 밀도 예측 알고리즘			
부 문	데이터 융합	●	제품 및 서비스 개발	
성 명	팀 장	공민석	010-6436-3711	
		서강대학교	ksjscott@naver.com	
	팀 원	김성현	서강대학교	
		엄다린	서강대학교	
		이정애	서강대학교	
		임영선	서강대학교	

Ⅰ. 과제 목표

2000년 이후 매년 여름, 국내 연안에서는 '해파리 주의 단계 특보'가 발령된다. 최근 연구에 따르면, 전 세계적으로 증가한 해파리 개체수는 상업과 어업, 관광 산업과 같은 다양한 분야에 부정적인 영향을 미친다[1]. 특히 우리나라 연안에서 국지적으로 발생하는 해파리는 해양생태계 교란, 해수욕장 쏘임 사고를 야기할 뿐만 아니라 경제적으로 막대한 피해를 준다. 예를 들어, 해파리의 대량 출현은 원자력발전소의 취수구를 막아 해수의 유입을 방해하여 발전소 가동을 중단시키고, 혼획을 유발해 어획물을 손상시켜 수산업 자원 감소를 야기한다[2]. 해양수산부에 따르면 해파리로 인한 피해 규모가 연간 최대 3천억에 달한다고 한다[1].

이에 대응하여 해양수산부는 해파리로 인한 피해를 최소화하기 위해 해파리 출현 시기와 출현량을 예측하고 이를 기반으로 해파리 피해방지 종합대책을 수립해 이행하고 있다.

그러나, 이러한 해파리 피해방지 대책은 지자체 공무원을 포함한 해파리 모니터링 요원의 신고만을 근거로 시행되므로, 주관적인 요소가

개입할 가능성이 높다. 또한, 해파리가 발견된 이후에 대책이 수립되므로 현재의 피해방지 대책은 예측보다 사후 대응의 성격이 강하며, 정확한 해파리 출현 예측 기간을 알 수 없는 등 해파리 관련 문제를 충분히 예방할 수 없다는 제한점이 있다. 따라서 해파리 피해 예방을 위한 대책에 보완이 필요하다 느껴, 해파리 출현 및 밀도 예측 결과를 제공하고자, 본 과제를 진행하게 되었다.

'심층신경망(DNN)을 활용한 해파리 출현 및 밀도 예측 알고리즘'은 k-water 공공데이터와 공공기관으로부터 수집된 해양 데이터, 그리고 해파리 모니터링 데이터를 가장 대표적인 딥러닝 모델인 DNN에 적용하여 해파리 출현 여부 및 밀도를 예측하는 알고리즘이다. 이를 통해 기존의 해파리 피해방지 대책을 보완함으로써 해파리 쏘임 사고를 예방하고 해파리로 인한 경제적 피해를 최소화하고자 한다.

Ⅱ. 활용 데이터

국내외에서 진행된 선행 연구에 의하면, 표층 수온, 풍속, 우천일수 및 수온과 염분 등 해양환경 요소들을 이용하여 해파리 출현 가능성을 예측할 수 있다[3]. 이에 기반하여 해양환경 요소들을 해파리 출현과 관련된 특징 변수로 설정하고, 해파리 출현 여부와 출현 밀도를 딥러닝을 이용하여 예측하고자 한다. 본 과제에서는 부산 연안에서의 해파리 출현을 예측하며, 사용한 데이터의 수집 기간은 2017년 6월부터 2022년 6월까지이다.

(1) k-water 공공데이터

해양환경 요소 중 "강수량" 데이터를 추출하기 위해 k-water 공공데이터 개방 포털의 "전국 강수 비율분석정보"를 활용했다[4]. 한국수자원공사에서 가뭄 분석 정보 제공을 위해 관측 일시, 금년 평균값 등에

따른 강수 비율 정보 데이터 항목을 제공하며, 본 과제에서는 “관측 일시”와 “관측 지역”, “금년 평균 강수량” 데이터를 사용했다.

(2) 그 외 해양 수집 데이터

그 외의 해양환경 요소는 국립해양조사원에서 제공하는 ‘바다누리 해양정보 서비스’에서 추출했다[5]. ‘바다누리 해양정보 서비스’는 관측소별 실시간 해양 관측 자료를 제공하는데, 본 과제에서는 데이터를 한 시간 간격으로 수집한 후, 다른 데이터와 기간을 맞추기 위해 이를 일주일 단위로 변환하여 활용하였다. 해양 관측 자료 중 ‘수온’, ‘염분’, ‘기온’, ‘기압’, ‘풍향’, ‘풍속’, ‘최대순간풍속’ 항목을 변수로 사용하였다.

(3) 해파리 데이터

해파리 출현 현황을 파악하기 위해 국립수산과학원에서 제공하는 ‘해파리 출현 및 이동 경로 모니터링 데이터’를 활용했다[6]. 해파리 정보 시스템의 해파리 속보 페이지에서 해파리 모니터링 주간 보고 데이터를 크롤링했으며 본 과제에서는 ‘해파리 종류’, ‘해파리 독성’, ‘출현 해역’, ‘출현 빈도’, ‘모니터링 보고 날짜’, ‘해파리 발견율’ 정보를 사용했다.

Ⅲ. 주요 내용

(1) 융합기법- 모델 소개

최근 인공지능 및 딥러닝 기술이 급격히 발전하면서 이를 다양한 분야에 적용하기 위한 시도가 활발히 진행되고 있으며, 특히 수자원 분야에 적용되는 시계열 데이터를 예측할 때, 물리적 모형에만 의존하지 않고 과거의 데이터를 활용하여 모형을 학습시켜 예측을 진행하는 심층

학습(Deep learning) 연구가 활발하게 진행되고 있다[7].

그중 가장 대표적인 모델인 심층신경망(Deep Neural Network; DNN)은 일반적인 인공신경망을 확장한 딥러닝 기법으로, 다수의 은닉층을 사용하여 복잡한 데이터를 학습하는 것이 특징이며, 데이터의 양이 많을수록 예측 성능이 개선된다는 장점이 있다.

출력층의 활성화 함수로 사용한 소프트맥스 함수는 다중 분류 문제에서 주로 사용되는 함수로, 각 클래스에 속할 확률을 0과 1 사이의 확률로 출력함으로써 해석을 용이하게 한다. 이를 고려하여 소프트맥스를 활용한 DNN 모델을 본 과제에 적합한 모델이라고 판단했다.

(2) 융합기법 - 데이터 추출 과정

- 해파리 데이터 크롤링 과정

국립수산과학원의 해파리 모니터링 주간 보고 크롤링에 필요한 웹페이지의 링크에 주간 보고의 날짜가 포함되어 있었기 때문에, 크롤링하고자 하는 기간의 시작 날짜와 끝 날짜를 입력하여 존재하는 주간 보고 날짜들을 리스트에 저장하였다.

국립수산과학원

해파리 모니터링 주간보고
2018.06.22.~06.28.

문서번호 2018.06.28 * 18 - 08호

담당 부서	기후변화연구과	담 당 자	■과 장: 고우진 ■연구관: 임월애, 연구사: 한창희 ■연구원: 이혜은, 김영진, 황성보	☎ 051) 720-2223 ☎ 051) 720-2236

※ 어업인 해파리모니터링요원 및 지자체의 협조로 취합/분석 자료임

■ 해파리 주간동향 (2018.06.22.~06.28.)

종류	출현해역	발견률	독성	비고
노무라입깃해파리	○서해 남해 동해 출현 •저밀도 출현 해역 - 인천 덕적도 근해 - 경남 고성 효암산 연안 - 경남 창원 난포리 연안 경남 소지도, 자란도 부근 - 경북 울릉도 부근	4.11% (6/28) ↑ 2.06% (6/21) ↑ 3.17% (6/14)	강독성	
	○서해 남해 동해 출현 •고밀도 출현 해역 - 충남 홍성 죽도리 부근 - 전북 군산 개야도, 연도, 신시도, 비응도 부근 전북 군산 새만금 연안 전북 부안 위도 부근 전북 고창 광승리 연안 - 전남 보성 득량만, 율포 연안 전남 강진 대구면 연안			

<그림 1. 국립수산과학원의 해파리 모니터링 주간 보고 페이지 모습>

크롤링 결과, 주간 보고 날짜와 해파리 출현 해역, 해파리 출현 밀도 등을 추출해 데이터프레임에 저장했다. 해파리 출현 해역의 경우, 이름을 표현하는 방법이 다양했기 때문에 지역명이 짧을 경우 해당 행정 구역을 앞에 추가하는 등 각 지역의 이름을 통일하였다. 또한 '연안', '부근' 등 지역을 명확히 구분하지 못하는 단어는 해역 이름에서 삭제하였다. 따라서 특정 지역을 입력했을 때 해당 지역이 포함된 해역을 추출하도록 알고리즘을 제작하였다.

초기 크롤링 알고리즘은 종속변수인 해파리 출현 밀도에 따라 '출현', '저밀도 출현', '고밀도 출현'으로 구분하고, 각각 [0, 1, 2]로 변환하여 데이터를 저장했다. 이 결과, 데이터의 크기가 819개였기에 딥러닝 모델에 사용하기에 부족했다. 이에 데이터의 양을 늘릴 필요성을 느껴 두 가지 방식으로 데이터 증강을 진행하였다.

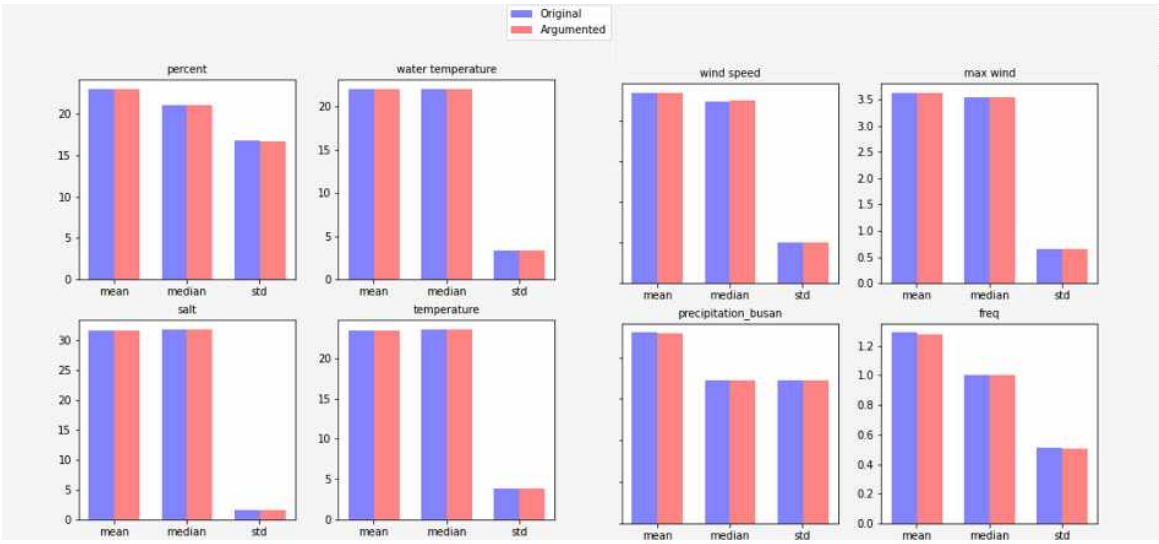
- 데이터 증강 및 통합

첫 번째는 새로운 변수를 추가하는 방법이다. 해당 기간 동안 주간 보고에 명시된 모든 해역에 관하여 해파리가 등장하지 않은 '미출현' 경우를 추가할 경우 데이터의 크기가 91,848개로, DNN모델에 사용하기에 데이터양이 충분하였다. 따라서 해당 변수를 추가하여 '미출현', '출현', '저밀도 출현', '고밀도 출현' 총 4개의 클래스를 각각 [0, 1, 2, 3]으로 변환하여 저장했다.

증강된 데이터를 통합하는 과정은 다음과 같다. 해파리 데이터는 일주일 단위로, k-water 강수데이터는 한 달 단위로, 해양 수집 데이터는 1시간 단위로 데이터가 존재하였다. 예측값이 '해파리 출현 밀도'이기 때문에, 해파리 데이터를 기준으로 날짜(date)를 통일하였다. 해양 수집 데이터는 일주일의 평균을 구하였고 k-water 강수데이터는 한 달 동안의 feature를 일주일 단위로 확장하여 해파리 데이터와 통합하였다. 이렇게 증강과 통합을 완료한 데이터는 'jellyfish2.csv'로 저장하였다.

두 번째 방식으로는, 부트스트래핑(bootstrapping)을 활용하여 데이터 증강을 진행하였다. 부트스트랩은 무작위 표본 복원 추출 방식을 활용하여, 모집단에서 N 크기의 무작위 데이터를 추출하고 이를 이어 붙이는 방식이다. 우선, 초기 크롤링 결과 데이터에 위의 데이터 통합 과정을 적용하여 jellyfish1.csv에 저장하였다. 이 데이터의 크기를 jellyfish2.csv와 비슷하게 증강하기 위해 N=15, 추출 횟수는 6,000번으로 설정하여 부트스트래핑을 진행하였고, 총 90,000개의 증강된 데이터를 얻어 augmented_jellyfish.csv에 저장했다.

증강된 데이터의 형평성을 확인하기 위해, 원본 데이터인 jellyfish1.csv와 각 feature의 평균, 중간값, 표준편차를 비교하였다. 그 결과, 그림 2와 같이 두 데이터의 통계량 차이가 거의 존재하지 않아, 증강된 데이터인 augmented_jellyfish.csv를 원본 데이터 대신 사용할 수 있다고 판단하였다.



<그림 2. 원본 데이터와 증강된 데이터 비교>

(3) 융합결과

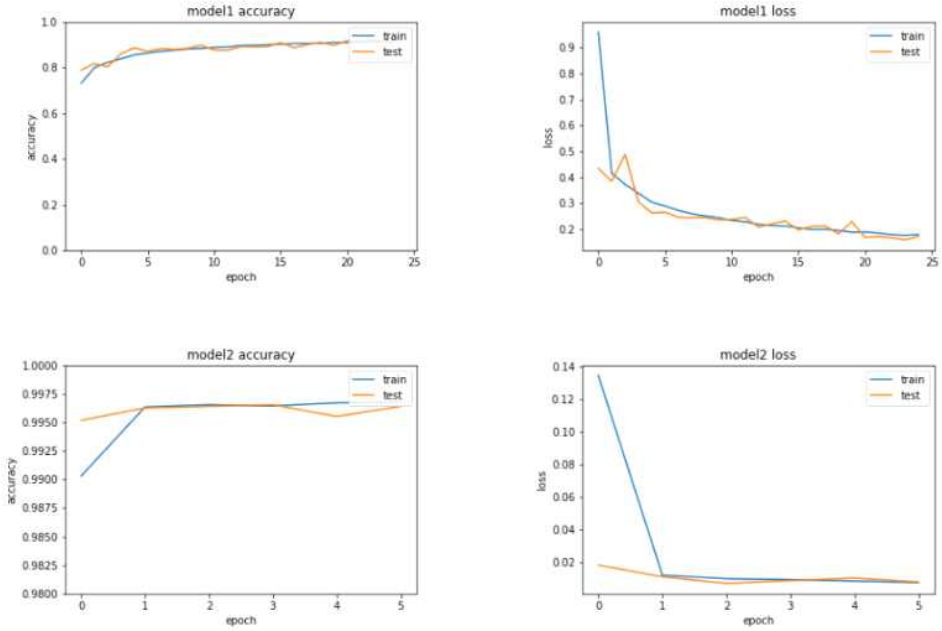
DNN 모델은 딥러닝 프레임워크인 텐서플로(tensorflow)를 활용하여 제작하였다. 각 데이터(augmented_data.csv, jellyfish2.csv)에 맞춘

DNN 모델 model 1, model 2를 생성하여 데이터 학습 및 테스트에 활용하였다. 전처리를 위해 우선 사이킷런의 라벨인코더(LabelEncoder)를 사용해서 각 데이터를 범주형으로 인코딩 후, 케라스의 to_categorical 함수를 통해 배열(array) 형태로 변환했다. 이렇게 변환된 데이터를 model 1과 model 2에 입력값으로 넣고 학습을 진행하였다. 두 모델 모두 4개의 layer를 쌓고, 학습률은 0.001로 설정했으며, 논문에 근거해 활성함수는 relu함수, 옵티마이저는 Adam, 손실함수는 categorical cross-entropy를 사용하였다[8]. Output layer의 최종 활성함수는 softmax 함수를 사용하였다. model 1과 model 2의 layer의 node 수는 다르게 설정하였는데, 이는 augumented_data.csv의 경우 예측 변수가 [0, 1, 2]로 3개인 반면, jellyfish2.csv는 예측 변수가 [0, 1, 2, 3]로 4개이기 때문이다.

model 1과 2 모두 에포크(epoch)를 25로 설정하여 학습을 진행하였다. 다만, model 1과 model 2의 학습 과정 중간에 early stop이 발생하여 각각 5번, 6번의 에포크 학습이 진행되었다.

(4) 결과

model 1의 학습 데이터의 정확도(accuracy)는 92.44%이며, 테스트 데이터의 정확도는 92.87%이다. model 2의 학습 데이터 정확도는 99.69%이고, 테스트 데이터의 정확도는 99.66%이다.

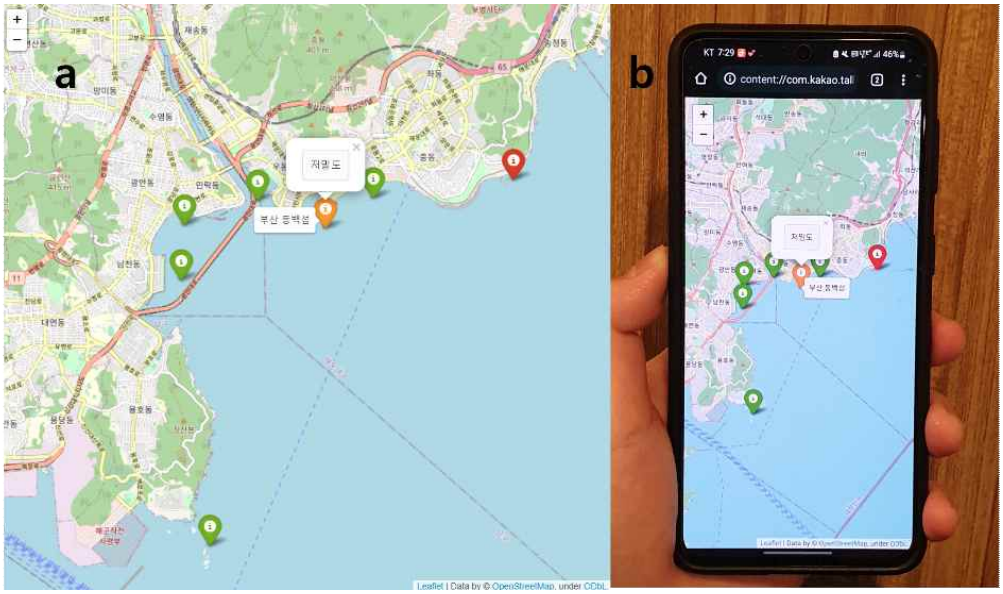


<그림 3. 각 모델 별 정확도와 loss>

두 모델 모두 높은 정확도를 보이며, 딥러닝 모델에서 흔하게 발생하는 과적합 문제도 발생하지 않았다. model 1에서 쓰인 augmented_data.csv는 부트스트래핑을 활용해 원래의 데이터를 증강한 데이터고, model 2에서 쓰인 jellyfish2.csv는 원래의 데이터에 포함되지 않았던 '미출현' 클래스를 학습레이블에 추가함으로써 데이터를 간접적으로 증강한 데이터이다. 서로 다른 방식으로 증강한 데이터에 대해 모두 높은 테스트 정확도를 보이므로, 두 모델 모두 우수한 예측 모델로 판단할 수 있다.

Ⅳ. 결과 및 기대효과

해파리 출현 예측 결과를 이용하여, 해파리 피해 방지 대책을 사후 대응보다는 사전 대응의 방식으로 변화시킬 수 있을 것이라 기대한다.



<그림 4. 지도 시각화 예시>

그림 4 의 a 처럼 해파리 출현 예측 데이터를 지도에 표시하여, 해파리의 출현 및 밀도에 대해 쉽게 알아볼 수 있도록 할 수 있다. 더 나아가, 이를 해파리 예측에 사용되는 변수인 기온, 수온, 기압 등 해양환경 정보와 함께 시각화한다면 해파리 피해방지 종합대책을 즉각적으로 시행할 수 있을뿐더러 해파리로 인한 피해를 최소화할 수 있을 것이다. 예를 들어, 특정 해변의 해파리 출현 밀도가 일정 수준 이상으로 예측될 경우, 사람들의 출입을 통제하여 해파리 쏘임 사고와 같은 인명 사고가 일어나지 않도록 예방할 수 있으며, 해파리 출현 예측 결과를 어업인들이 활용함으로써 어업 활동의 피해를 최소화할 수 있다.

더불어, 현재 진행 중인 '해파리 web 신고 서비스'와 해파리 예측 지도를 연동하거나, 그림 4 의 b 와 같은 스마트폰 애플리케이션을 제작한다면, 해파리 예측뿐만 아니라 해파리 출현 신고도 더욱 편리해질 것이라 기대한다.

본 모델은 부산 지역의 데이터를 사용하여 예측 모델을 개발하였으나, 추후 더 넓은 범위로 확장하여 적용할 수 있다. 데이터가 전국적으로 확대될 경우 국가적 차원에서 해파리 예측을 통해 해파리의 종류

및 개체수 동향을 파악할 수 있다. 이를 이용하여 국내 해파리에 관한 연구 및 고위험 해파리의 개체 수를 줄이기 위한 다양한 대책 설립에 기여할 수 있다.

참고문헌

[1] 김대영, 이정삼, 김도훈, “해파리 피해 실태 및 산업적 이용 방향”, 수산해양교육연구, 제26권 제3호, 통권69호, 2014, pp. 587-590

[2] KBS 뉴스. 해파리 떼에 기름값까지… 엎친 데 덮친 어민들 [웹사이트]. Available: https://news.kbs.co.kr/news/view.do?ncd=5499423

[3] KIM, Bong-Tae, EOM, Ki-Hyuk, HAN, In-Seong, PARK, Hye-Jin, “An Analysis of the Impact of Climatic Elements on the Jellyfish Blooms”, The Korean Society for Fisheries and Marine Sciences Education, Volume 27, Issue 6, 2015, pp. 1755-1763

[4] k-water 공공데이터 개방포털 [웹사이트]. Available: https://opendata.kwater.or.kr/open/data/list/view.do

[5] 바다누리 해양정보 서비스 [웹사이트]. Available: http://www.khoa.go.kr/oceangrid/gis/category/reference/distribution.do#none

[6] 국립수산과학원 해파리 속보 [웹사이트]. Available: https://www.nifs.go.kr/bbs?id=jellynews

[7] 이명진, 김종성, 유영훈, 김형수, 김삼은, 김수전, "DNN 및 LSTM 기반 딥러닝 모형을 활용한 태화강 유역의 수위 예측", Journal of Korea Water Resources Association, Vol. 54, No. S-1, 2021, pp. 1061-1069

[8] 장하영, 유은경, 김혁진, Hyeock-Jin, “딥러닝 학습에서 최적의 알고리즘과 뉴론수 탐색”, 디지털융복합연구, Vol. 20, No. 4, 2022, pp. 389-396