

딥러닝을 이용한 한국 주요 매개모기 종 분류

박준영¹, 김동인², 노광래³, 권형욱⁴, 강우철⁵
 인천대학교 임베디드시스템공학과
 인천대학교 생명과학부 & 매개곤충자원융복합연구센터

e-mail:{juny1905¹, kdin34², hello159357³}@naver.com
 {hwkwon⁴, wchkang⁵}@inu.ac.kr

Classification of Korean Vector Mosquito Species using Deep Neural Networks

Jun-young Park, Dong-in Kim, Kwang-rae Roh, Hyeong-wook Kwon Woo-chul Kang
 Dept. of Embedded Systems Engineering, Incheon National University

요 약

기후변화에 따라 매개 질병의 발병 빈도가 증가하고 있으며 모기와 같은 매개체에 의해 전염되는 매개 질병은 인구집단에 대한 중요한 위협 요인이다. 이러한 질병 관리를 위해 지역별 모기 서식 현황을 모니터링 하는 시스템의 필요성이 강조되고 있다. 하지만 현재의 모기 모니터링은 개체 파악을 위한 분류와 동정을 사람이 직접 수행하기에 오랜 시간이 소요된다. 이 연구는 그러한 문제점을 해결하고 미래 매개곤충 서식 현황 파악 시스템의 기반을 마련하기 위해 심층 신경망(Deep Neural Networks)을 활용하여 한국 주요 매개모기 종 분류를 수행하고 결과를 분석하였다. 종 분류를 위한 모델은 잘 알려진 신경망 모델인 DenseNet(Dense Connected Networks)을 사용하였고 이를 직접 촬영한 모기 데이터와 약간의 변형을 가한 모기 데이터를 사용하여 학습시켰다. 학습 데이터를 각각 5배, 20배, 100배로 증강하여 실제 데이터의 부족을 보완하였으며, 이를 통해 최대 99.48%의 정확도를 달성하였다.

1. 서론

최근 하드웨어의 성능이 뒷받침되며 기계학습에 대한 연구가 활발히 진행되어 현실 문제를 해결하는 데에 널리 응용되고 있다. 특히 뇌의 구조를 응용하여 설계된 인공 신경망(Artificial Neural Network)은 인공지능 분야에 많은 영향을 끼쳤다.[1] 이 연구에서는 잘 알려진 합성곱 신경망[2] 모델인 DenseNet[3]을 사용하여 한국의 주요 매개모기 종 분류에 활용하고자 한다. 이를 통해 미래의 매개곤충 모니터링 시스템의 초석이 될 모기 종 분류의 기초적인 방법을 제시하여 미래의 질병관리 시스템 구축에 기여하고자 한다.

2. 기존 연구

기존에 시각 집중(Saliency Map)과 DCNN을 이용한 논병해충의 현지화(Localization) 및 분류(Classification)에 관한 연구가 있었다 [4]. 단 해당 연구는 눈에 주로 출몰하는 여러 곤충의 데이터를 웹에서 수집하여 기존 합성곱 신경망을 개량한 모델을 학습하였지만, 본 연구처럼 특정 지역에 출몰하는 한 과의 주요 개체를 직접 촬영하여 학습시킨 국내외 연구는 찾을 수 없다.

3. 연구 과정

이 연구에서는 질병을 매개할 수 있는 한국 주요 매개모기 5종에 대한 분류를 수행한다. 분류에는 DenseNet 모델(그림 1)이 사용되며, 모델 학습에 있어 시간 비용을 절

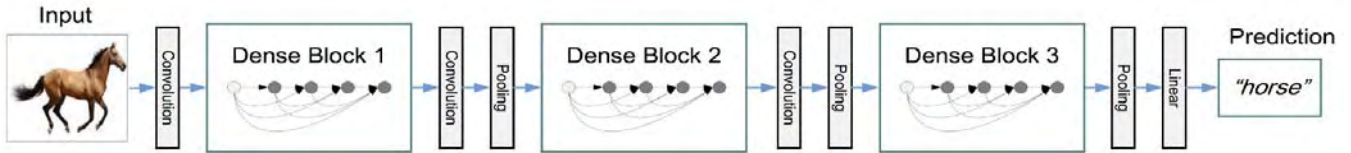
약하기 위해 전체 레이어에 대한 전이학습(Transfer Learning)을 수행하고, 그 과정에서 발생하는 문제점을 보완하기 위해 데이터를 증강하는 등의 실험을 수행한다.

3.1. DenseNet

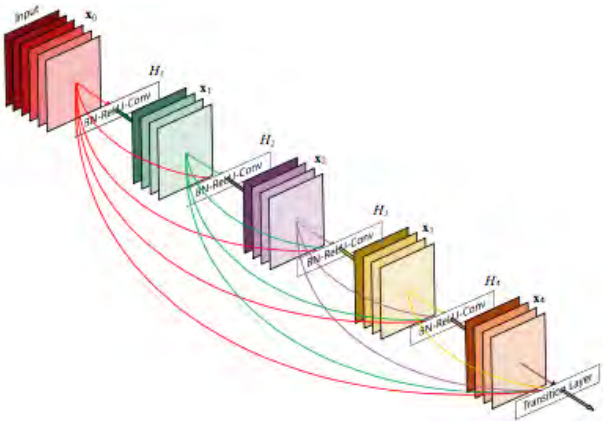
CVPR (Computer Vision and Pattern Recognition) 2017에서 발표된 DenseNet (Densely Connected Networks)은 Convolution 과 Pooling 레이어를 사이에 두고 Dense Block이 존재하는 구조를 가지고 있다. 이와 같이 특징이 증첩된 레이어 구성을 통해 가장 복잡한 특징뿐 아니라 상위 레이어의 덜 복잡한 특징까지 모두 사용하여 전체 이미지 대비 작은 크기를 갖는 특징까지 추출할 수 있다 [3]. 위와 같은 장점을 통해 본 연구에서 수집한 것과 같이 비교적 적은 양의 데이터에서도 고밀도 연결을 통해 일반화 효과를 가져 과적합을 줄여줄 수 있는 DenseNet이 모기종 분류에 적합하다고 판단하였다.

3.2. 전이학습 및 병렬화

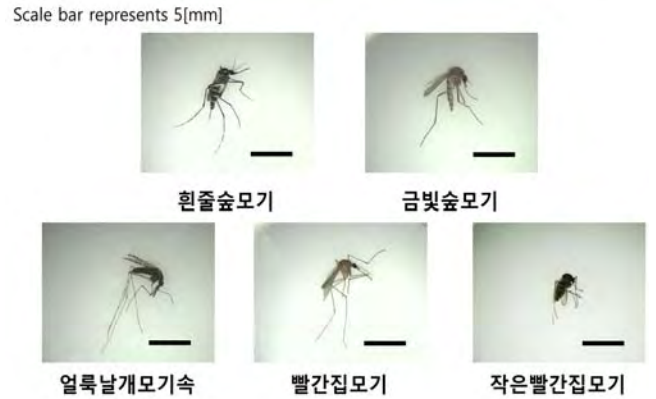
DenseNet과 같은 복잡한 구조를 갖는 신경망 모델을 학습 시키기 위해서는 최소 수일의 시간과 고성능 컴퓨터가 요구된다. 따라서 소규모 연구실 환경에서의 효율적인 모델 학습을 위해 모기의 날개 무늬, 다리의 인편 등과 같이 전체 이미지 크기 대비 작은 영역의 특징을 추출할 필요가 있기에 사전에 ImageNet 데이터 셋 [5]을 통해 사



(그림 1) DenseNet의 구조[3]



(그림 2) Dense Layer의 구성[3]



(그림 3) 모기 데이터 예제

전 학습된(Pre-trained) 가중치들을 받아 미조정(Fine-tuning)을 수행하였다. 미조정은 전이학습 방법 중 하나로, 모델의 전체 레이어를 초기 상태부터 학습해나가는 대신 이미 훈련되어 가중치를 가지고 있는 레이어를 새로운 데이터 셋에 대하여 추가적으로 훈련하는 방법이다. 미조정을 수행하기 위해 마지막 레이어의 Feature 수를 모기 종의 개수인 5개로 수정되었다.

3.3. 데이터 수집

DenseNet 모델을 학습하기 위해 충분한 양의 데이터가 필요하나 웹상에 공개되어 있는 이미지의 경우 자연을 배경으로 찍은 것이 대부분으로 연구실 환경에서의 모기 분류에 적합하지 않은 사진이 다수이고 그 수 또한 적어 연구에 적합하지 않다고 판단하였다. 따라서 본 연구의 목적에 맞는 모기 이미지를 취득하기 위해 인천대 매개곤충자원융복합연구센터(CRCIV)로부터 모기 표본을 얻어 보조도구를 사용해 특정 환경에서의 이미지를 취득하였다. (그림 3)은 학습에 사용된 5종의 모기 데이터의 예제이다.

3.4. 표본

한국에서 주요 질병을 옮기는 모기 5종에 대하여 표본을 수집하였다. 흰줄숲모기(*Aedes albopictus*)와 빨간집모기(*Culex pipiens*)는 사육실에서 채집한 개체를 사용하였고, 얼룩날개모기속(*Anopheles pp.*), 금빛숲모기(*Aedes vexans*), 작은빨간집모기(*Culex tritaeniorhynchus*)는 야생에서 포집한 개체를 냉동보관장치에 보관하여 사용하였다. 이 연구에선 모델 학습을 위해 각 종에 대하여 600장의 사진을 촬영하였다.

3.5. 데이터 수집

모기 데이터의 수집은 평면 촬영이 가능한 전자 모니터

확대경을 사용하였다. 플레이트에 백색 무광 아크릴판을 설치하고 그 위에 모기 개체를 올려 사진을 촬영하였다.

과적합을 해소하기 위하여 개체 사진 1장에 대하여 무작위로 크기, 회전, 밝기를 조정 한 결과를 5장, 20장, 100장으로 확장하여 그 때의 변화를 확인하였다. 사용된 데이터의 구성과 그 수는 (표 2)에서 확인할 수 있다.

<표 2> 증강 후 데이터 셋 구성 (단위 : 장)

	원본	5배	20배	100배
학습	3,000	15,000	60,000	300,000
검증	600	600	600	600

4. 실험

4.1. 실험 환경

DenseNet 모델을 빠르게 학습하기 위해 Nvidia 1080Ti GPU가 4개 장착된 서버를 사용하였다. 구현을 위한 프레임워크는 PyTorch를 사용하였다. 한 Epoch 마다 큰 Accuracy 증가를 얻기 위해 Batch Size를 128로 높이고 8개의 스레드가 병렬로 학습하도록 설정하였다.

<표 3> 학습에 사용된 주요 하이퍼파라미터

매개 변수	값
Batch Size	128
Num. of Workers	8
Optimizer	SGD
Learning Rate	0.001 (x0.7 in 7 epochs)

4.2. 실험 결과

증강을 거치지 않은 데이터 셋에 대하여 첫 7 Epoch 내에 검증 정확도 90%를 달성하였고 20 Epoch 동안 학습

시키는 데에 11분 25초가 소요되었다.

<표 4>는 데이터 원본과 증폭된 데이터를 통한 학습 후의 분류 정확도를 나타낸다. 증강을 거친 후 마지막 20번째 Epoch 에서의 손실률과 학습 중 최대 정확도를 확인한 결과 20배 까지 증강하였을 때 손실이 줄어들고 정확도가 상승한 반면 100배 증강한 데이터 셋은 오히려 손실률이 5배 증강한 데이터 셋 보다 높고 최대 정확도 또한 원본 데이터 셋으로 훈련한 결과와 같게 나타났다. 이 연구에서 사용한 사전 학습 가중치는 ImageNet 데이터 셋으로 클래스 별 약 600~1300장의 학습 이미지에 대하여 사전 학습 되어있다. 따라서 모기 종별 약 1500장의 학습 이미지를 가지는 5배 증강 데이터 셋이 가장 합리적이라고 판단된다. 20배 증강한 데이터 셋의 경우 종별 약 6000장의 학습 이미지를 가져 손실이 줄어들며 정확도가 높아졌지만 기존 대비 많은 양의 이미지에 대해 과적합이 발생하였을 가능성이 있다. 같은 이유로 100배 증강한 데이터 셋의 경우 오히려 정확도가 하락한 것을 확인할 수 있다.

<표 4> 학습 데이터 셋별 정확도

	Top-1 (Acc)	Top-5 (Acc)
원본	97.28%	98.14%
5배	98.21%	97.76%
20배	99.48%	99.34%
100배	97.35%	97.27%

5. 결론

이 연구에선 딥러닝을 통한 모기 개체 분류와 동정을 위해 한국 주요 매개모기 5종에 대한 데이터를 사용하여 DenseNet 모델을 학습하였고 그 결과 최대 99.48%의 정확도를 얻을 수 있었다. 이를 통해 추후 더 많은 종에 대한 데이터를 수집하고 모기 분류에 특화된 딥러닝 네트워크의 구성을 통하여 추론(Inference) 시간을 단축시킨다면 서식 지역의 모기 개체 수를 실시간으로 파악할 수 있는 모기 포집 및 분류기에 응용할 수 있을 것이다.

사사

This research was supported by Incheon National University Institute of Convergence Science & Technology

참고문헌

- [1] Yaan LeCun, Yoshua Bengio and Geoffrey Hinton, "Deep Learning", Nature 521, 436-444, May, 2015.
- [2] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
- [3] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, "Densely Connected Convolutional Networks" Computer Vision and Pattern Recognition; Machine Learning Aug, 2016.
- [4] Ziyi Liu, Junfeng Gao, Guoguo Yang, Huan Zhang and Yong He, "Localization and Classification of Paddy Field Pests using a Saliency Map and Deep Convolutional Neural Network" Scientific Reports volume 6, Article number: 20410 Feb, 2016.
- [5] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255. Jun, 2009.