딥 러닝을 활용한 한국형 흑색종의 분류

백경대, 유OO, *양OO
연세대학교 보건과학대학 의공학부

K. D. Baek, Yoo, *Yang

Yonsei University Biomedical Engineering

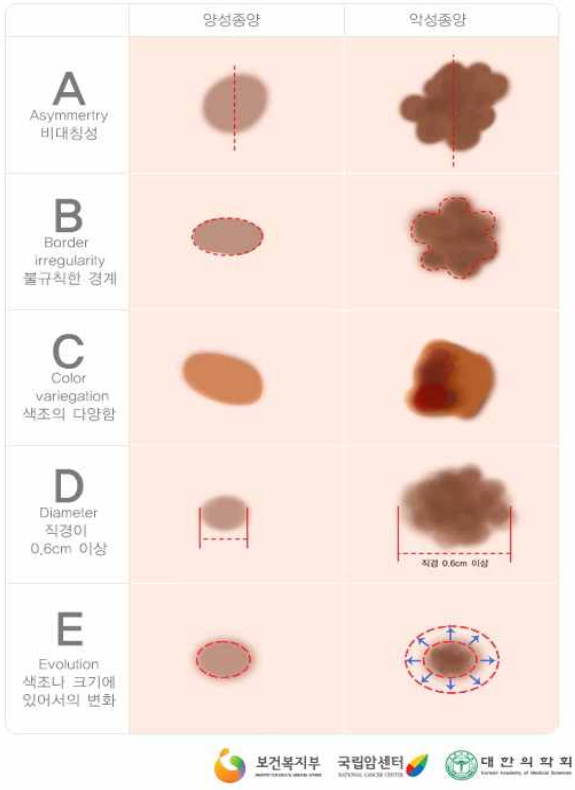
1. 서론

피부에 발생하는 악성 흑색종은 주로 기저층에 산재하여 있는 멜라닌세포에서 발생한다. 멜라닌세포 유래의 악성종양으로서 멜라닌 세포가 존재하는 곳에는 어느 부위에서난 발생할 수 있으나 피부에서 가장 많이 발생한다. 악성 흑색종은 발생 초기부터 다른 장기로의 전이가 가능할 뿐만 아니라, 조직 침범 깊이가 예후와 밀접한 관계가 있는 악성도가 높은 종양이므로 초기의 진단과 치료가 가장 중요하다.

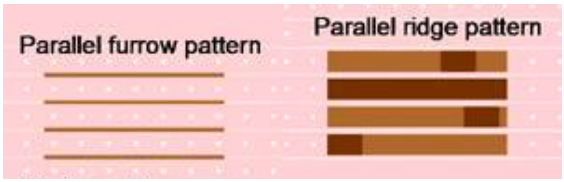
하지만 악성 흑색종은 숙련된 전문의조차 조직검사가 아닌 임상적인 소견만으로는 80-90%만이 진단이 가능한 진단이 상당히 어려운 질병이다. 흑색종의 진단에는 ABCD관찰법이 도움이 되는데 최근에는 진행도(Evolution)를 추가하여 ABCDE원칙이라고도 한다. (Figure 1) ABCD관찰법은 다음과 같이 비대칭성(Asymmetry), 불규칙한 경계(Border irregularity), 다양한 색상(Color variegation), 직경이 0.6cm이상(Diameter)으로 구성되어있다. (질병관리본부, 2011)

'ABCDE원칙'은 주로 몸에 발생한 흑생종을 구분하는데 사용된다. 하지만 국내에서는 외국의 경우와 다르게 주로 몸이 아닌 손과 발에서 흑색종이 발병한다. 손과 발에 발생하는 악성 흑색종은 몇 가지 패턴으로 분류되는데 그중에서 가장 대표적인 패턴은 'furrow'와 'ridge'이다. (Dermoscopy basics and melanocytic lesions, Hong Kong J. Dematol. Venereol. Figure 2) 이처럼 흑색종 병변의 생김새가 다르기 때문에 'ABCDE원칙'을 손, 발의 악성 흑색종에 적용하기엔 어려운 점이 있다. 발병원인 또한 다르고 치료반응도 다르게 나타난다. 이를 한국형 흑색종이라고 부른다. 하지만 서양에서는 흔히 나타나지 않기 때문에 다른 흑색종들에 비해 한국형 흑색종의 연구는 활발히 진행되지 않고 있다

<Figure 1. 흑색종 분류를 위한 ABCDE원칙>



<Figure 2. Furrow, Ridge 패턴>



최근 의료계에서는 이미지로 표현되는 데이터를 분류하는 인공지능이 개발되고 있다. 피부암, 특히 악성 흑색종을 분류하기 위한 이미지 인식기술은 예전부터 연구되어 오던 분야이다. 악성 흑색종을 진단해준다는 스마트폰 어플리케이션을 비롯해 2015년도에는 IBM Research에서 약 95%의 정확도로 악성 흑생종을 구분하는 기술을 개발했다고 발표했다. 다만 추가적인 데이터의 공개가 없어 어떤 기술을 사용했는지, 구체적인 성능 파악은 어려운 부분이 있다.

이러한 프로그램개발의 대부분은 딥 러닝(Deep learning) 기술을 활용하고 있다. 최근에는 딥 러닝과 컴퓨터 기술의 발달로 인해 더욱 높은 정확성을 가진 이미지 인식 프로그램이 나오고 있다. 이와 같은 발전은 피부질환 분류 프로그램의 발전으로 이어졌는데 2017년 2월 Nature에서는 스탠포드 연구진이 개발한 딥 러닝 기반의 피부질환분류 인공지능이 숙련된 피부과 의사수준의 높은 정확도를 보이는 논문이 발표되기도 했다.

이처럼 딥 러닝을 활용하여 피부질환을 분류하는 연구는 현재 활발히 진행되고 있으며 실효성을 보이고 있다. 하지만 전부 해외의 흑색종을 바탕으로 진행되고 있어 한국형 흑색종의 연구는 부족한 실정이다. 따라서 본 연구에서는 손 과발에서 주로 발생하는 악성 흑색종의 대표적인 패턴을 바탕으로 악성 흑색종을 딥 러닝을 이용하여 분류하려고 한다. 딥 러닝 분야에서 가장 많이 사용되고 있는 텐서플로와 기존의 딥 러닝 알고리즘을 활용하여 연구를 진행하고 선행 연구에서 활용된 데이터를 이용하여 보다 높은 정확도를 얻는 것을 목표로 한다. 이로 인해 간단한 과정을 거쳐 높은 정확성을 가지고 진단이 어려운 흑색종을 빠르게 진단할 수 있을 것으로 기대된다.

본 논문에서는 먼저 실험에 사용된 데이터와 텐서플로를 사용한 연구 방법에 대해 기술했다. 딥 러닝의 설정을 변경하여 실험을 진행하였기에 실험 결과별로 딥 러닝의 설정에 대해 설명했다. 결과에 따라 오차가 발생되는 원인을 분석하여 토의 및 결론을 작성했다.

2. 연구 방법

본 연구는 선행 연구(Ridge and furrow pattern classification for acral lentiginous melanoma using dermoscopic images, Sejung Yang)에서 사용된 데이터를 활용하였다. 이 데이터는 2013-2014년도의 연세세브란스 병원의 흑생종 이미지이다. 한국형 흑색종의 연구를 위해 전부 손과 발의 이미지를 이용했으며 대표적인 패턴인 'Furrow'와 'Ridge' 패턴을 가진 이미지를 이용했다. 데이터는 악성 흑생종과 양성의 이미지를 각각 준비했다. 악성 흑생종의 이미지는 350개이고, 양성의 이미지는 374개이다. 약 700개 이상의 데이터 중 80~90%를 학습데이터로, 10~20%를 확인데이터로, 10%를 최종시험데이터로 사용하였다. 실험도중 발견된 손, 발의 데이터가 아니거나, 'Furrow', 'Ridge' 패턴에 맞지 않는 이미지들을 제거하여 학습하였다.

딥 러닝 프로그램으로는 구글에서 오픈 소스로 제공하는 텐서플로기반의 'Inception-v3'를 활용하였다. 2015년, 2016년도 구글의 딥 러닝 프로그램의 비교를 보면 'Inception-v3'는 상위버전과 에러 값의 차이가 0.5%로 크지 않고 데이터를 활용하는데 있어 이득을 가지기 때문에 연구에 사용하기에 적합하다고 판단했다. (Table 1, 2)

학습방법은 전이학습을 이용해 'Inception-v3'에 학습된 가중치를 초깃값으로 설정 후 새로운 데이터셋을 대상으로 재학습을 수행하였다. 학습을 진행할 때는 hyperparameter를 이용해 최적의 학습률을 찾아 보다 빠르게 학습이 100%에 도달하도록 하였다.

<Table 1. 2015 이미지인식 프로그램 성능비교>

Netwrk	Model Evaluted	Crops Evaluted	Top-1 Error	Top-5 Error
VGGNet	2	-	23.7%	6.8%
GoogLeNet	7	144	-	6.67%
PReLU	-	-	-	4.94%
BN-Inception	6	144	20.1%	4.9%
Inception-v3	4	144	17.2%	3.58%

<Table 2. 2016 이미지인식 프로그램 성능비교>

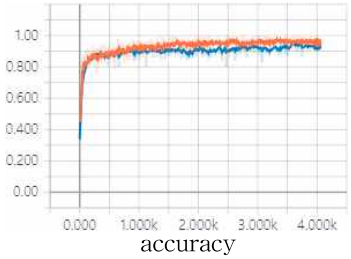
Network	Crops	Top-1 Error	Top-5 Error
ResNet-151	10	21.4%	5.7%
Inception-v3	12	19.8%	4.6%
Inception-ResNet-v1	12	19.8%	4.6%
Inception-v4	12	18.7%	4.2%
Inception-ResNet-v2	12	18.7%	4.1%

'test 3'까지의 학습을 바탕으로 데이터 이미지를 변경시켜 두 가지의 'case'로 나눠 다시 딥 러닝을 수행하였다. 두 가지 'case'로 나눠서 실험을 하게 됨으로 신뢰도, 정확도의 증가와 그에 따라 데이터의 변화가 정확도에 어느 정도의 영향력이 있는지 확인할 수 있다. 'case 1'은 각각의 모든 이미지들을 90°flip, 270°flip, left_right crop, top_bottom crop 중 하나를 이용해 임의로 변형시켰다. 따라서 생성된 값은 기존의 데이터와 동일한 개수의 데이터를 가진다. 'case 2'는 데이터 이미지중에서 임의로 하나를 선택해 90°flip, 270°flip, left_right crop, top_bottom crop들 중 임의로 적용시켰다. 이미지와 1대1 대응이 되는 것이 아니기 때문에 이미지의 개수는 악성흑색종과 양성, 각각 400개로 조정하였다. 그리고 가장 정확도가 높았던 'case 2'를 10번 반복 실험하여 최종정확도의 평균을 구했다. 반복실험을 할 때는 데이터를 랜덤하게 섞어 같은 결과가 나오지 않도록 했다.

3. 연구 결과

'test 1'은 구글에서 오픈소스로 제공하는 프로그램인 'Inception-v3'를 사용해서 손과 발의 악성 흑색종 이미지로 딥 러닝을 수행하였다. 딥 러닝을 수행할 때 learning late를 0.001, batch size를 100으로 설정하는 것이 가장 기본적인 방법이기 때문에 같은 방법으로 learning late와 batch size를 설정하고 진행했다. (figure 3)

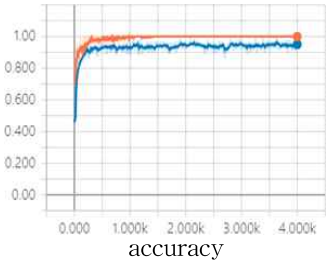
<figure 3. Test 1, 초기 결과값>



-Total Images : 724
-Validation : 20%, Testing : 20%
-Learning rate : 0.001
-Test batch size : 100
-Train accuracy 100% : 도달하지 못함
-Final test accuracy : 88.3%

'test 2'는 hyperparameter를 이용하여 학습이 잘 진행될 때 최적의 learning rate를 찾아 적용하였다. 이렇게 설정한 학습률은 0.038이고 데이터의 크기를 고려해 batch size를 50으로 변경하였다. 그리고 validation과 testing을 10%로 조정했다. 학습률에 변화로 인해 1460steps에서 학습 정확도가 100%로 유지되었고 최종 정확도 또한 93.2%로 상승했다. (figure 4)

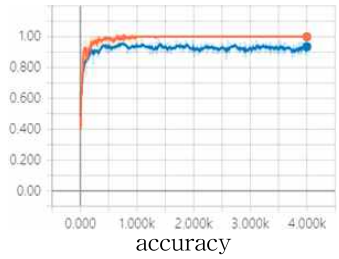
<figure 4. Test 2, 학습률 변경 후 결과값>



-Total Images : 724
-Validation : 10%, Testing : 10%
-Learning rate : 0.038
-Test batch size : 50
-Train accuracy 100% : 1460
-Final test accuracy : 93.2%

'test 2'의 결과값을 바탕으로 잘못 분류된 이미지들 중에 손, 발의 흑생종이 아닌 이미지가 있는 것을 확인했다. 손, 발의 흑색종과 몸의 흑색종은 패턴이 다르기 때문에 같은 방법으로 구분할 수 없다. 따라서 'test 3'에서는 해당 이미지를 삭제하고 손, 발의 흑색종 이미지만을 이용해서 딥 러닝을 다시 수행하였다. 따라서 딥 러닝에 사용된 최종 데이터의 개수는 2개 감소하였고 이외에는 'test 2'와 모두 동일하게 진행하였다. 그럼에도 불구하고 최종 정확도는 94.4%로 1.2%증가했고 학습속도 또한 빨라졌다. (figure 5)

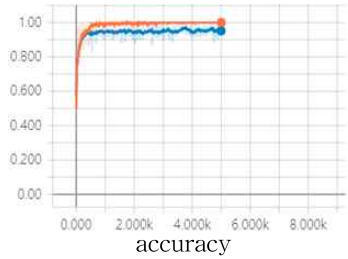
<figure 5. Test 3, 데이터 정리 후 결과값>



-Total Images : 724
-Validation : 10%, Testing : 10%
-Learning rate : 0.038
-Test batch size : 50
-Train accuracy 100% : 1400
-Final test accuracy : 94.4%

'case 1'에서는 악성 흑색종과 양성의 이미지를 각각 모두 변형시켰다. 'test 3'와 비교했을 때 학습속도에서는 크게 차이를 보이지 않았지만 최종 정확도는 96.5%로 2.1%증가했다. 그리고 잘못 분류되는 이미지의 유형은 오히려 'test 3'보다 증가했다. (Figure 6)

<figure 6. Case 1 데이터 결과값>

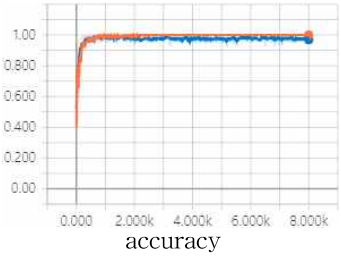


-Total Images : 1444
-Validation : 10%, Testing : 10%
-Learning rate : 0.038
-Test batch size : 50
-Train accuracy 100% : 1450
-Final test accuracy : 96.5%

'case 2'는 임의의 이미지를 뽑아 임의로 변형해서 딥 러닝의 데이터로 사용하였다. 이때 학습 정확도가 100%에 유지되는 지점이 3600steps로 2배 이상 증가했다. 하지만 최종 정확도가 98.1%로 'figure 3'의 결과보다는 3.7%, case 1보다는 1.6% 증가한 결과를 보였다. 또한 분포 사이의 유사성을 나타내는 cross entropy가 거의 일치하는 것을 볼 수 있다. (Figure 7)

이 중 가장 정확도가 높았던 'case 2'를 10회 반복 실험하였다. 10회 최종 정확도의 평균은 95.52%로 약 95%이상의 결과를 정확도를 가진다는 것을 알 수 있었다.

<figure 7. Case 2 데이터 결과값>

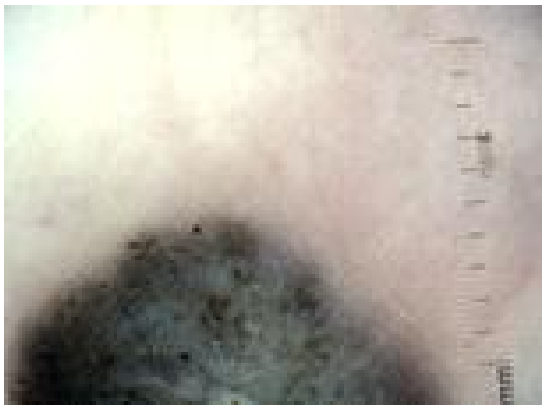


-Total Images : 1522
-Validation : 10%, Testing : 10%
-Learning rate : 0.038
-Test batch size : 50
-Train accuracy 100% : 3600
-Final test accuracy : 98.1%

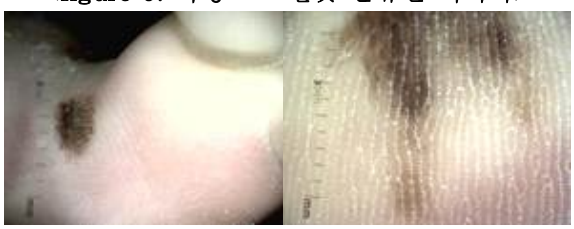
'case 2'를 반복 실험하였을 때, 3회까지의 최종 정확도평균은 96.93%, 5회는 96.2%, 7회는 95.45, 10회는 95.52%로 약 95%이상의 결과를 보여줬다.

4. 결론 및 토의

<figure 8. 양성으로 잘못 분류된 이미지>



<figure 9. 악성으로 잘못 분류된 이미지>



'figure 8'과 'figure 9'은 'case 2'에서 최종적으로 잘못 분류된 이미지들이다. 'figure 8'은 악성 흑색종이지만 양성으로 잘못 분류된 것이고, 'figure 9'은 양성이지만 악성 흑색종으로 잘못 분류된 이미지이다. 'figure 8'의 흑색종에서 보이는 패턴이 본 연구에서 중점적으로 연구하는 'Furrow'와 'Ridge' 패턴과 유사하지 않은 모습을 보이고 있다. 또한 'figure 9'의 오른쪽 이미지를 보면 하얀색으로 각질이 많이 보이는 것을 확인할 수 있다. 이처럼 각질, 이미지의 화질, 흑색종의 패턴차이와 이미지의 배경등과 같은 특이점들이 딥 러닝의 과정에서 노이즈로 발생할 수 있다.

딥 러닝 학습을 진행할 때 최종정확도의 기여도는 최적의 학습률을 찾는 것이 가장 많은 정확도의 상승을 가져왔다. 두 번째로 많은 영향을 끼친 것은 데이터 변형을 통한 학습량 증가였다. 약 700개 가량의 제한적인 데이터를 이용하더라도 데이터변형을 통해 추가적인 정확도의 상승을 기대할 수 있다.

딥 러닝의 정확도 상승을 위해선 최적의 학습률을 찾아 최적화 시키는 것이 중요하고 적절한 데이터를 활용해 학습해야했다. 에러를 유발할 수 있는 데이터는 삭제하고 데이터 변형을 통해 적은 데이터로도 높은 정확도를 얻을 수 있었다. 다만 데이터의 전처리가 이뤄지지 않은 점, 딥러닝을 수행하기에 약 700개의 데이터가 충분하지 않았다는 점을 보완하기 위해 추후연구가 필요하다. 또한 선행연구에서 케라스를 사용해 딥 러닝을 진행한 것과 달리 본 연구는 텐서플로를 이용했기 때문에 케라스를 사용한 연구와의 비교가 가능하다.

본 연구를 통해 딥 러닝을 이용한 한국형 흑색종의 분류정확도를 상승시킬 수 있었다. 또한 비침습적인 흑색종의 진단 및 분류방법을 개발했다. 나아가 원격의료에서 흑색종의 진단 가능성을 제시했다.

5. 참고 문헌

Ridge and furrow pattern classification for acral lentiginous melanoma using dermoscopic images, Sejung Yang, 2017

Dermatologist-level classification of skin cancer with deep neural networks, Andre Esteva, 2017

Rethinking the Inception Architecture for Computer Vision, Christian Szegedy, 2015

Inception-v4, Inception-ResNet and
the Impact of Residual Connections on Learning, Christian Szegedy, 2016

Dermoscopy basics and melanocytic lesions, Hong Kong Journal of Dermatology & Venereology, 2013

딥 러닝을 활용한 이미지 빅 데이터 분석 연구, 김윤진, 중앙대학교, 2017

러닝 텐서플로 (딥러닝 영상처리와 NLP부터 텐서보드 시각화, 멀티스레딩, 분산처리까지), Hope Tom 외 2명, 한빛미디어, 2018

딥러닝의 정석 (텐서플로와 최신기법으로 배우는 알고리즘 설계), Nikhil Buduma, 한빛미디어, 2018