 서울대학교 대학원 의과학과

**학부연구생인턴 보고서**

**연 구 기 간: 28기(2021. 12. 27. ~ 2022. 2. 18.)**

**연 구 제 목: Retinal Fundus image classification**

**성 명: 김수용 (인 또는 서명)**

**지도교수 성 명: 박상민 (인 또는 서명)**

**(※ A4용지 10장 이내, 바탕체, 11포인트, 줄간격 160%. 아래 항목대로 기술하기 힘든 경우에는 해당 항목만을 기술함)**

1. 연구요약 (영문의 경우 : 150단어 이내, 국문의 경우 : 750자 이내)

=> 먼저 RF 데이터셋에 대하여 분석을 합니다. RF 데이터셋은 망막 이미지들을 수만 장 모아놓은 것이며 총 5가지 라벨을 가지고 있고 라벨은 망막의 DR , 즉 당뇨망막증의 심각도에 따라서 5개로 나뉘어져 있습니다. 여기서 crossentrophy 및 bceloss 를 통해서 여러가지 모델의 5개 레이블에 대한 classification 정확도를 높여가는 활동을 해보았습니다. 또한 전이학습을 적용해보기도 하였습니다. 그렇게 선행연구 중 데이터셋에 대한 점과 classification은 마치고 distortion 에 대한 선행연구 즉 perceptual similarity 측정 metric 과 관련한 논문을 간단히 살펴보겠습니다. 우리 팀은 이논문을 참고하여 distortion을 medical image dataset 에 적용할 생각이며 현재 traditional distortion , cnn-based distortion table(논문)에서 절반정도 수행을 하였고 cnn-based distortion 에 의한 alg output task 가 여러가지인데 gan, resnet 등 여러가지모델을 써보고 layer, dropout , learning rate도 조정 해왔습니다.

1. 서론 (연구배경 및 연구목적)

* Distorted dataset 에 대한 2afc, jnd test 를 수행하기 위해 classification 에 모델을 pretrain 한다.
* Distorted dataset 을 만들기 위해 traditional distortion 및 cnn-based 를 수행한다.

1. 선행연구

=> **The Unreasonable Effectiveness of Deep Features as a Perceptual Metric**

**(Richard Zhang1 Phillip Isola12 Alexei A. Efros1 , Eli Shechtman3 Oliver Wang3)**

* 위는 연구를 시행하기 이전에 참고한 논문이며 간단하게 이미지 간의 similarity 측정하는 기존의 metric 을 비판하고 새로운 method 를 제시합니다. 그것은 바로 internal activation of deep feature 입니다. Classification task 에 train 이 된 model 이 이미지 간의 유사성을 측정하는 test 인 2afc , jnd test 에 능하다는 것을 보여줍니다.

2afc, jnd test 를 시행하기 이전에 데이터셋을 기존의 데이터셋과 다르게 구성을 합니다. 데이터셋에 distortion 을 적용하는데 이는 크게 cnn-based, traditional, real algoristhm output 으로 나뉩니다. cnn-based 는 기존의 연구 에서는 별로 없었던 새로운 distortion 입니다. 이것은 artifiact 의 space 를 더 실제와 유사하게 simulate 하기 위한 전략이고 가능성있는 알고리즘 output 도 simulate 합니다. real alg output 은 input 에 corruption 을 적용하면서 나오고 크게 4가지 입니다. 이런식으로 딥러닝 모델의 아웃풋을 plague 하는 artifiact 에 대해서 조사합니다.

이제 본격적으로 2afc, jnd test 에 대해서 살펴보도록 하겠습니다. 먼저 2afc test 로부터 나오는 judgment 는 원본 패치에 두가지 distortion 을 적용하고 어느것이 원본과 더 유사한지를 얼마나 잘 맞추는에 대한 점수입니다. JND 는 2afc 의 주관성을 보완하기 위한 테스트입니다. 이는 한 왜곡이랑 한 원본이 같은지 다른지를 맞추는 시험입니다.

결과를 평가하는데 있어서 먼저 metric 과 network 가 얼마나 잘 작동하는 지 판단해야합니다. 모든 validation test 에는 distortion 2개 원본 1개인 triplet 에 대한 5개의 pairwise judgment 가 포함됩니다. 이것은 모든 judgment 와 함께 알고리즘 일치를 계산합니다.

먼저 2afc score 에 대해서 살펴보겠습니다. 기존의 metric 에서 나온수치, 인간이 왜곡과 원본을 얼마나 잘 구분하는지에 대한 정도, 모델의 정확도 에 대해서 각각 수치화 하여서 2afc score를 구해냈습니다. 대체로 기존의 metric 보다 새로운 deep feature 를통한 activation 이 더 좋게 인간의 유사성 판단을 따라하기위한 method로 자리매김하고 있고 모델의 용량이 클수록 더 좋은 점수를 보이고 있습니다. 이는 squeezenet 을 다른것과 비교하면 알수있습니다. 그리고 대체로 traditional 보다 input 에 corruption 을 적용하면서 나온 4가지 task 들에 의한 real algoristhm output 이 유사성 판단을 blur 하는데 더 치명적입니다.

다음은 classficaiton 에 대해 train 을 한 것이 perceptual similarity task 에 얼마나 도움이 되는지를 보여주는 또다른 표인데요, 이것은 detection 보다 classification 이 2afc, jnd 테스트와 상관관계가 높은 것을 볼수 있고 0.928 수치 부분에서 jnd test 를 2afc test 와 같이 하면 judgment 수집에 더 효과적인걸 볼수 있습니다.

위 연구는 고차원 이미지 인식에 대해서 해결하기위해 clasffication task 관련 train 이 효과적이라는 것을 보여주고 judgment 하는데 있어서 rerpresentation 을 배우는 것 또한 효과적인 것을 보입니다. 즉 sementic 과 perceptual task의 특성이 서로 연관이 되어 있다고 볼수 있습니다.

1. 실험방법
2. 데이터셋 파악 (Retinal Fundus dataset)

* 저는 retinal fundus dataset 에 대해서 classification 을 해보았습니다. 5개의 class 로 split 되어있고 환자번호 그리고 왼쪽 눈, 오른쪽 눈에 따라서 라벨링 되어 있었습니다. 몇몇 환자는 왼쪽 오른쪽중 하나의 이미지 데이터가 누락이 되어있었습니다. 주로 이 데이터셋이 만들어지고 classification task 에 대해서 train 하는 연구가 많이 이루어진 이유는 인간의사가 한명한명 dr 을 진단을 하면 시간이 많이 소모가 될수가 있습니다. 즉, 많은 환자를 한꺼번에 빨리 처리하는 시스템이 필요해졌고 당뇨병 진단및치료인프라 가 부족한데 그지역이 당뇨병환자가 많은 경우가 많아서 그 필요성이 더 커지고 있는 추세입니다. 데이터셋과 dr 과관련된 용어들을 먼저 살펴보면 npdr, pdr 이 두가지가 진짜 중요한데 p 가 proliferative 즉 증식을 의미하고 severity 레벨을 결정하는 결정적 요소입니다. 증식성이 높을수록 그환자가 위험한상태에 잇다고 보면 됩니다. Severity 레벨은 정상에서부터 증식성 dr 까지 5개로 나뉘고 데이터셋도 이 레벨에 따라서 5개의 class로 split 했습니다. Severity 레벨 5가지 에 대해서 자세히 살펴보겠습니다. 숫자가 높을수록 dr 을 환자가 더 심각하게 앓고 있는 것이고 0는 normal 즉 dr 이 없는 상태이고 mild 부터 npdr 한 특성 즉 비증식성 당뇨망막증을 보이고 있습니다. Moderate level 에서는 1년이내에 레벨 4 즉 pdr 증식성 당뇨망막증으로 될 확률이 16프로 정도이고 severe level 에서는 1년이내에 50프로확률로 pdr 에걸릴수가 있습니다. 단 여기서 중요한점은 레벨 4는 레벨1부터 3까지 , 레벨 3은 레벨 1부터 2까지의 특징마저도 보유하고 있습니다. 즉 0단계(정상) 제외하고 나머지 질병이 있는 단계들은 자신보다 낮은 숫자의 등급의 특징을 가지고 있습니다(단 낮은숫자중 0은제외). 마지막으로 이미지 size 에 대해서 파악을 해보면 500\*500 pixels 로 되어있습니다.

1. Classification Task

* 위 연구는 medical image 가 아닌 일반적인 image dataset 에만 distortion 및 classifciiation 에 pretrain 된 model을 가지고 실험했습니다. 우리는 이연구를 그 연구가 일반적인 image 에 한것에만 국한하지 않고 medical image dataset 으로 확장을 해나갈 계획입니다.

Pretrained model 을 만들어가는 과정을 만들어 나가기 위해서 classification 의 정확도를 높여가는 방법이 무엇인지 이나은 선생님과 고민해나가고 만들어 나갔습니다. 먼저 resnet 18, 34, 50 ,101 을 인수인계 받은 코드 내에서 각각 classification 을 실행해나가 보았습니다. 결과적으로 resnet 50 가 가장 우수한 성과를 보였고 모델을 고정해두고 hyperparamer 를 바꾸어 나가며 실험을 해나갔습니다. 먼저 data augumentation 부분을 살펴보면 이미지데이터셋이 500\*500으로 큰 사이즈인걸 고려하여 256\*256 으로 resize 하였습니다. 또한 randomhorizontalflip 을 적용하였고 이미지 픽셀 mean, std value 를 고려하여 normalize 를 하였습니다. Batch size 는 gpu, accuracy 모두를 고려 보았을 때 256 일 때 가장 model 이 좋은 성과를 보였고 256으로 고정 해두고 주로 몇 epoch 마다 learning rate 를 줄이는 scheduler 를 통해서 accuracy 를 높여나갔습니다. Optimizer 의 경우에는 adam 과 sgd 를 써보았으며 sgd 에서 더 좋은성과를 보였습니다. Criterion 을 처음에는 crossentropy 로 시작을 하였으나 80프로의 정확도 마저 달성하지 못해 bceloss로 전환을 해보고 80프로 이상의 정확도를 달성해보았습니다. 또한, rf datset 이 4단계의 경우 1-3, 3단계는 1-2 단계의 특징 마저도 보유하고 있는점을 고려하여 0과1로 데이터셋을 새롭게 라벨링을 하였습니다. 예를 들면, 4단계의 경우 [01111], 3단계는 [01110] 이런식으로 하였습니다. 그리고 transfer learning 을 통해서 88프로의 정확도까지(vgg-16) 달성을 해보았습니다. Sota 통계에 비해서(92프로) 정확도가 낮았지만 그래도 라벨이 5개, 많다는 점을 고려하면 좋은 성과를 거두었습니다.

1. Making distortions

[1] Traditional distortions

텍스트이(가) 표시된 사진

자동 생성된 설명

* 우리 실험에 적용 해보았던 distortion 은 위 선행연구에 나왔던 distortion입니다. 위 표에서 photometric distortion 과 어떻게 RF dataset 에 적용했는지를 살펴보도록 하겠습니다. 총 4가지인데 이는 모두 pytorch library를 통해서 손쉽게 구현이 가능했습니다. Lightness shift 의 경우에는 밝기를 조정하는 것으로 데이터셋마다 랜덤으로 밝기가 조정되었습니다(어두운것과 밝은것이 섞임). 이외에 채도, 대비, 색상 모두 이미지 마다 랜덤으로 적용했습니다.

Noise의 경우에는 gaussian noise class를 만들고 평균과 표준편차의 수치를 조정해나가면서 mean=0.1, std=0.8로 고정하고 모든 이미지에 noise 를 적용했습니다.

Blur 는 torch transform 의 gaussian blur를 썼고 이미지 사이즈에 (500) 에 맞추어서 500으로 kernel-size 를 맞추었습니다.

Sub type= Spatial 인 distortion 경우에는 affine warp 부터 살펴보면 저는 pytorch library 를 통해서 전체이미지를 30도기울였습니다. 이때 방향은 랜덤입니다.

Compression 은 jpeg로 png 파일을 압축시켰습니다.

[2] auto encoder

=> gaussian noise 로 distorted 된 이미지를 auto encoder 로 복원하고 그 이미지에 대해서 model 이 classficiation 하는 task 를 수행하였으며 auto encoder 의 loss도 같이 측정했습니다. Gaussian noise(0.1,0.4) 에 대해 autoencoder 로 실험을 해보았습니다.

1. 실험결과

|  |  |
| --- | --- |
|  |  |
| Model | Pretrained VGG-16 |
| Batch Size | 128 |
| Input image size | 224\*224 |
| Optimizer | SGD  (Momentum :0.9,  Learning Rate : 0.001) |
| Epoch | 50 |
| Criterion | Binary Cross Entropy Loss |
| Augumentation | Resize(256), RandomHorizontalFlip,  RandomResizedCrop  (size=224, scale=(0.7, 1.0)) |
| Best Train Accuracy | 0.8932 |
| Best Valid Accuracy | 0.8794 |
| Best Train loss | 0.1645 |
| Best Valid loss | 0.1857 |

* 위는 distortion 없는 RF 데이터셋에 대한 classification task



=> 위는 앞페이지 표의 조건 중 data augumentation 쪽에만 변화를 줌-> 데이터셋에 gaussian noise를 적용 🡪 이후 auto encoder로 가우시안 노이즈 제거한 reconstructed image 를 통한 train & validation실시

(2) traditional distortions

옥외설치물, 옅은, 무척추동물, 어두운이(가) 표시된 사진

자동 생성된 설명 original무척추동물, 옥외설치물, 어두운, 흐린이(가) 표시된 사진

자동 생성된 설명 Ghost옅은, 어두운, 무척추동물, 켜진이(가) 표시된 사진

자동 생성된 설명light

옥외설치물, 별이(가) 표시된 사진

자동 생성된 설명 텍스트, 별, 옥외설치물, 흐린이(가) 표시된 사진

자동 생성된 설명 Gaussian white noise

Chromatic aberration

무척추동물, 옥외설치물, 어두운, 옅은이(가) 표시된 사진

자동 생성된 설명 옥외설치물, 어두운, 무척추동물, 옅은이(가) 표시된 사진

자동 생성된 설명 Homography

Bilateral filter

어두운, 옥외설치물, 옅은, 켜진이(가) 표시된 사진

자동 생성된 설명 옥외설치물, 어두운, 별이(가) 표시된 사진

자동 생성된 설명saturation

contrast

무척추동물, 옥외설치물이(가) 표시된 사진

자동 생성된 설명 Linear warp 옥외설치물, 어두운, 옅은, 무척추동물이(가) 표시된 사진

자동 생성된 설명 Shifting

지도이(가) 표시된 사진

자동 생성된 설명 compression 텍스트, 별, 옥외설치물, 흐린이(가) 표시된 사진

자동 생성된 설명 uniform white noise

무척추동물이(가) 표시된 사진

자동 생성된 설명 gaussian blur 무척추동물, 옥외설치물이(가) 표시된 사진

자동 생성된 설명 color shift

무척추동물, 옥외설치물, 옅은, 어두운이(가) 표시된 사진

자동 생성된 설명 affine warp 옥외설치물, 별, 어두운, 흐린이(가) 표시된 사진

자동 생성된 설명 gaussian colored noise

5. 고찰

처음 받았던 코드에서 모델만 바꿀때에는 resnet 50 가 가장 우수한 성과를 보였습니다. Batch size 는 gpu, accuracy 모두를 고려 보았을 때 256 일 때 가장 model 이 좋은 성과를 보였습니다. Optimizer 의 경우에는 adam 과 sgd 를 써보았으며 sgd 그리고 learning rate = 0.001 에서 가장 좋은성과를 보였습니다. Scheduler 의 경우에는 전체 epoch 의 30프로 지나는 지점마다 줄이는 것이 가장 효과적이었습니다. Criterion 을 처음에는 crossentropy 로 시작을 하였으나 80프로의 정확도 마저 달성하지 못해 bceloss로 전환을 해보고 80프로 이상의 정확도를 달성해보았습니다. 그리고 라벨 변환, transfer learning 을 통해서 88프로의 정확도까지(vgg-16) 달성을 해보았습니다. Sota 통계에 비해서(92프로) 정확도가 낮았지만 그래도 라벨이 5개, 많다는 점을 고려하면 좋은 성과를 거두었습니다. Cnn based Auto encoder 실험에 대해서는 이미지를 복원하는 과정에서 mseloss 를 썻지만 encoder loss 가 줄어들지 않는 결과가 나왔고 결국 accuracy 가 epoch 를 많이 지나도 별로 개선되지않는 모습을 보였습니다.

6. 결론

레이블이 5개인거 감안하면 distortion 이 안된 데이터셋이었다하더래도 좋은 성과를 보였지만 distortion 이 적용된 이미지를 reconstruct 해서 cnn based distortion 을 만드는데 있어서 나오는 loss 에 대한 문제를 해결하지 못한점은 아쉽습니다. 보고서 낸 이후에는 다른 선생님 분들과 협력하여 이 loss 에 대한 문제를 해결 할 필요가 있고 제가 위에서 시행했던 여러 traditional distortion 에대해서도 cnn-based distortion 과 융합을 해야하는지 할 수 있는지를 파악할 필요가 있을 것입니다. 현재, 이나은 선생님과 gan model 을 통해서 autoencoder 를 만들기로 한 상황이어서 auto encoder loss 문제에 해결할 발판을 마련하고 있습니다. 또한, 현재 위 에서 보여주는 distortion 들은 육안으로 보이게 하기위해 distortion function input 에 큰 수치를 넣었고 (예 : 가우시안 노이즈 표준편차) 이는 우리 메디컬 데이터셋에는 적합하지 않은 수치일 수도 있습니다. 그 수치 value 자체에 대해서도 discussion 이 이루어질 필요가 있을것입니다.