2차 프로젝트 멘토 피드백 보고서

■ 프로젝트 주제 : 객체탐지를 활용한 반도체 불량 자동 분류 모델 구현(DATA FLOW)
■ 프로젝트 조원 : 최정인, 김나영, 김예슬, 손아, 한나영

참여기업명	인빅 주식회사	멘 토	추동원, 지효철

Q. 질문
여러 모델들을 활용해보려고 하는데 현업에서는 어떤 기준으로 사용할 모델을 선정하는지 궁금합니다.

A. 답변
하드웨어에서 처리 가능한 속도와 메모리를 만족하는 수준에서 가장 높은 정확도의 모델이 선정 기준이 됩니다. 그리고 발표된 Pre-trained model을 튜닝해서 많이 사용하며 이때 라이센스 문제가 없어야 합니다.

Q. 질문
저희 조가 전처리 했었던 부분들이 1.클래스 재정의, 2.바운딩 박스 재조정, 3.데이터 증식, 세가지 였는데요. 피드백 주신 걸 듣고 이해한 것은 1.클래스 재정의는 의미가 없다고 해석을 하였는데 그럼 전처리는 2.바운딩 박스 재조정, 3.데이터 증식 이 두가지로 충분할지, 이 두 가지 방식 외에 조언해주실 만한 또다른 전처리 방향도 궁금합니다.

A. 답변
1. 클래스 재정의가 필요하다면 해야 합니다. 이건 의미론적으로 접근하면 좋을 것 같습니다. 재정의해서 고객이 납득이 가는지 논의해 보시면 좋을 거 같습니다.
2. 바운딩 박스는 사람이 정의하기 때문에 100% 정확하다고 볼 수 없습니다. 그렇다고 이걸 조정한다고 성능이 좋아질 것이라는 것도 장담하기 어렵고 데이터 자체의 오류나 오염을 증가시킬 가능성이 있습니다. 크게 바운딩 박스가 잘못된게 아니라면 굳이 신경 써서 해야 하는지 의문이 듭니다.
3. 데이터 증식은 필요해 보입니다. 데이터가 너무 저적 특정 불량 클래스가 많다면 학습에 바이어스(편향)이 생길 수 있습니다. 증식 방법은 크롭도 유의미해 보입니다.

Q. 질문

원본 데이터의 양은 늘릴 수 없는 상황입니다. train(학습)데이터는 증식을 통해 늘리지만, vaildation(1차검증)과 test(2차검증)셋은 증식이 불가한 걸로 알고 있습니다. 그런데 이때 검증용셋 안에서 데이터를 증식시켜도 되는지 알고 싶습니다.

ex) test셋 안에 9장의 정상 이미지와 50장의 결함 이미지가 있습니다. 이때 정상 반도체 이미지를 증식하여(회전, 색상변환 등) 50장 정도로 만들고, 결함 이미지를 증식하여(회전, 색상변환, 이미지합성 등) 200장 정도로 만들어, 총 59장->150장의 test셋을 완성해도 괜찮은지 궁금합니다.

검증용 셋은 증식하면 안된다고 알고 있지만, 학습용 셋과 섞이지 않은 이미지로, 검증용 셋 내에서만 증식을 시행하면 검증이 여러차례 이루어질 뿐 모델 성능에는 크게 차이가 나지 않을거라는 생각에 질문 드렸습니다.

A. 답변

[아래 질문들과 공통된 답변이 될 수 있어 여기에 길게 서술하겠습니다.]

현재 가장 큰 문제는 데이터 분포 문제입니다. 예시로 병원에서 가진 10만명의 데이터를 가지고 암을 예측한다고 가정해 봅시다. 10만명의 사람 중에서 100명이 암환자고 9만9천9백명이 암이 없는 정상인이라고 가정해 봅시다. 이 경우 암환자 비율은 0.1%입니다. 이 경우 AI모델이 모든 사람이 암이 없다고 예측하여도 99.9%의 정확도를 가지고 있게 됩니다. 정확도는 높지만 다른 지표(precision, recall, f1-score)는 매우 낮게 나와 쓸모없는 모델이 됩니다. 이처럼 가지고 계신 데이터도 unbalanced dataset problem이 있습니다.(이 주제로 검색해 보시면 많은 자료가 있을 것입니다.) 자주 보이는 불량 패턴은 잘 판단하는데 가끔 보이는 불량 패턴은 잘 판단하지 못할 가능성이 있습니다. 따라서 테스트 데이터셋은 모든 불량 패턴이 고르게 분포해야 이 AI모델이 어느 정도의 성능이 나는지 명확하게 알 수 있습니다. 따라서 2가지 방법이 가능해 보입니다.

◆ Test dataset, val dataset 모두 불량 패턴 클래스를 고르게 샘플링 할 수 있는 경우

가장 좋은 방법입니다. Validation dataset은 학습에 도움을 주는 데이터셋이므로 이 데이터셋이 test dataset보다 적어도 됩니다.(테스트 100개, val 20개) 성능 평가는 무조건 Test dataset으로 하므로 validation dataset은 학습에 도움을 줄 뿐이지 개수가 많든 적든 모델 검증에는 중요하지 않습니다.

◆ Test dataset만 불량 패턴 클래스를 고르게 샘플링 할 수 있는 경우

데이터가 너무 적어서 Test dataset만 불량 패턴을 고르게 샘플링할 경우에는 어쩔 수 없습니다. 하지만 Test dataset은 클래스가 고르게 분포해야 **어떤 패턴이든지 잘 찾아준다고 검증**할 수 있습니다. validation dataset의 분포가 고르지 않아서 학습 중간 중간 확인하는 모델의 성능에 바이어스가 있어 학습 중에 평가가 어려울 수 있지만 모델의 평가는 Test dataset으로만 하는 것입니다. 그러므로 학습에 어려움이 있을 것을 감안하고 학습할 수 밖에 없습니다.

그리고 Test dataset에는 train dataset에서 augmentation한 데이터가 들어가면 안됩니다. 아예 다른 데이터만 있어야 합니다. 그리고 test dataset은 사용 안한 crop된 데이터는 존재해도 될거 같습니다.

(예, 테스트 데이터셋 구성)

1개의 데이터를 4등분해서 4개의 데이터로 사용 -> 가능

1개의 데이터를 4등분해서 원본 1개 + 4개 데이터 -> 불가능

Q. 질문

train, validation, test 로 나눈 데이터셋중 test 의 데이터 비율이 너무 적다라고 판단되어서 test 데이터 양을 증식해도 된다면 증식할 예정이에요. test 검증 때 모델 결과를 실제와 똑같이 뽑고싶은데 불량품 이미지와 양품 이미지의 비율을 얼마나 둬야할지 감이 잡히지 않아 비율에 대해 질문 드립니

다.

A. 답변

Q. 질문

반도체 결함 탐지 모델을 만들 때 필요한 적정 데이터 수가 10~20만장 정도라고 하셨는데, 저희는 데이터가 300여장 밖에 없는 상황입니다. 그래서 가능한 최대한 데이터를 증식하고자 하는데, 이때 데이터를 10~20만장까지 늘리는 게 모델 성능에 큰 영향이 없는지 궁금합니다.
300여장의 데이터를 십만장 단위로 늘리는 것이 불가능한 일은 아니지만, 원본이 지나치게 적은 상황에서 너무 큰 규모로 늘리는 게 아닐지 하는 생각이 들어 여쭙니다.

A. 답변

너무 많이 증식시키면 안됩니다. 어차피 오버피팅이 생깁니다. 이러한 문제 때문에 프로젝트의 컨셉을 조금 바꿔보면 어떨지 제시해 드린 겁니다. 현재 가진 데이터가 적으므로 기존 모델에 적은 데이터를 가지고 학습하는 transfer learning, few-shot learning등을 강점으로 내세우면 “적은 데이터로도 객체 인식을 잘하는 모델을 만들 수 있습니다!”고 강조할 수 있을 거 같습니다.
이런 것도 있습니다.
https://mldlcvmjw.tistory.com/265

Q. 질문

약간 번외 질문이지만 회사에서 사용하는 GPU의 성능이 궁금합니다. 일반 개인 컴퓨터와 같다고 생각해야 할지, 아니면 모델용으로 훨씬 좋은 성능의 GPU를 사용하는지 알고 싶습니다.
이번에 딥러닝 모델을 돌릴 때 GPU메모리 부족으로 런타임 에러가 발생하는 경우가 많았는데, 그럼

이 모델들은 실제 현장에서 사용이 어려운건지(실 활용보단 학술적인 의미로 존재하는 모델들인지), 혹은 회사에서 사용하는 다른 방법이 있는지 궁금합니다.

A. 답변

메모리 부족으로 런타임 에러가 나면 mini-batch size(배치 사이즈)를 줄어보시기 바랍니다. GPU 메모리 문제로 동작하지 않는 것 같습니다. 너무 최신 모델만 찾고 따라 가시다 보면 메모리 문제로 구동 조차 못하실거 같습니다. 큰 모델의 경우 40GB 메모리를 가진 GPU에 batch를 8개 올려서 구동해야 할 수 있습니다. batch size가 줄어들면 학습 시간이 오래 걸립니다.

회사에서도 때에 따라서 RTX 4090(24GB)같은 컨슈머 GPU도 사용하고 Nvidia Quadro 제품도 사용합니다. diffusion model 같은 거대한 모델을 학습(구동x)하기 위해서는 큰 메모리의 GPU를 사용하지 않으면 학습하지 못하거나 배치를 매우 작게 줘야 동작합니다.

AI모델을 서비스(구동)하는 경우에는 1~2개의 데이터를 빨리 넣어서 빨리 출력해야 하므로 RTX 4090 같은 고성능 GPU도 활용됩니다. 하지만 학습은 몇십개 몇백개의 데이터를 넣어서 학습하므로 4090의 24GB 메모리로도 부족할 수 있습니다.