#1

**안녕하세요 SKT AI Fellowship 4기 설비의 잔여 수명 예측 프로젝트에 참가한 다쿠야키팀의 한양대학교 이서윤, 서강대학교 김서연 학생입니다.**

#2

**오늘 발표는 다음과 같은 순서로 진행하겠습니다. 먼저 연구 과제 아이디어와 데이터셋을 소개 드리고, 어떻게 수행해 나갈 지에 대한 계획을 설명드리겠습니다.**

#3

**RUL이란 Remaining Useful Life의 약자입니다. 저희의 프로젝트는 사전에 RUL을 예측하여 Threshold 이하로 RUL이 떨어지면 장비가 고장나기 전에 사전에 알람을 발생시키는 것을 목표로 하고 있습니다.**

#4

**RUL을 예측하기 위한 방법으로는 모델 기반 접근 방식과 데이터 기반 접근방식이 있는데, 모델 기반 접근 방식은 복잡한 다차원 시스템이 필요하므로 저희는 데이터 기반 접근 방식, 즉 머신러닝 및 딥러닝을 모델을 구현하고, 적절한 Threshold 값을 찾고자 합니다.**

**최적의 모델을 구현하기 위해서 세 가지 사항을 평가하고자 합니다. 먼저 다양한 모델에 대해 예측 오차를 측정하여 이를 줄일 수 있도록 연구 방향을 설정할 것입니다. 둘째는 Overfitting을 확인하는 것입니다. 모델의 Training Error와 Test Error를 비교하고, Metric을 분석하여 Overfitting이 발생했을 때 Early Stop 기준 변경, Dropout 적용 등과 같은 적절한 조치를 취하고자 합니다. 마지막으로 모델의 예측 성능과 더불어 Capacity나 GPU 사용량 등도 추가로 측정하여 더 효율적인 모델을 개발합니다.**

#5

**지금부터는 데이터셋에 대해 소개드리겠습니다.**

**CMAPSS는 W, Xs, Xv, Theta 네 분야의 변수로 이루어진 방정식을 사용해 터보 엔진의 열화에 따라 여러 측정값들과 센서값들을 관측한, RUL 예측 임무에 자주 사용되는 역학 모델입니다.**

#6

**CMAPSS에서 X는 시나리오 설명 작동 조건, Xs는 물리적 특성을 나타내는 측정값, Xv는 가상 센서값, 세타는 모델 건강 파라미터입니다.**

#7

**앞서 설명드린 역학모델을 사용한 CMAPSS 데이터셋은 NASA에서 제공한 터보펜 엔진 성능 저하 시뮬레이션 데이터셋입니다. 이 데이터셋은 작동조건과 고장조건에 따라 FD1,FD2,FD3,FD4의 4개의 서로 다른 세트가 존재합니다. 작동조건과 고장조건이 많을수록 더 RUL 예측이 어려울 것으로 보입니다.**

#8

**저희는 본격적으로 연구 방향성을 설정하기 전에 4개의 각 하위 데이터셋 FD1 #8 ,FD2 #9 ,FD3 #10 ,FD4 #11에 대한 시각화를 진행해보았습니다.**

#10

**(FD3)왼쪽 플롯은 random하게 선정한 10개의 unit에 대한 21개의 센서값들의 이동평균을 나타내고, 오른쪽 플롯은 데이터셋에 존재하는 RUL 값들을 차원 축소했을 때의 산포도입니다.**

#12

**고장 조건이 2개였던 FD2와 FD4의 경우 시간의 흐름에 따라 21개의 모든 센서가 변화했으나, 고장 조건이 1개였던 FD1과 FD3는 변화하지 않는 센서도 존재했고, outlier로 작용할만한 일부 unit의 센서도 발견되었습니다. 작동조건, 고장조건이 많을수록 데이터셋의 난이도가 상승하는 것을 볼 수 있습니다.**

#13

**두번째로 참고한 데이터셋은 NCMAPSS입니다. 이는 NASA에서 CMAPSS의 한계점을 보완하여 2021년에 출시한 데이터셋입니다. CMAPSS는 이미 성능이 저하된 시스템의 샘플만을 추출하였으나, NCMAPSS는 건강한 상태로 시작하는 궤적의 전체 히스토리를 제공하여 시스템에 존재하는 복잡성의 중요 요인을 파악할 수 있을 것으로 보입니다.**

**NCAMPSS는 기존에 비해 용량이 27GB로 훨씬 커졌고, DS1부터 DS8c 등 총 10개의 하위 데이터셋이 존재합니다. 고장원인과 unit은 오른쪽과 같았습니다.**

#14

**NCMAPSS 데이터셋에 대한 시각화 결과입니다. NCMAPSS는 기존 데이터셋과 유사하나 각 샘플에 대한 unit정보, cycle, Flight Class, health state에 대한 보조정보인 A 변수가 추가된 것을 볼 수 있습니다.**

#15

**unit별로 시간의 흐름에 따라 변화하는 주요 변수 4가지를 확인하였는데, 특히 시간에 따른 Theta의 변화를 봄으로서 무엇이 고장 원인인지 파악할 수 있었고, unit 별 health state를 플로팅하여 몇 사이클일 때 어떤 유닛이 고장나는지도 확인해보았습니다.**

#16

**또한 health state를 이용하여 건강할 때와 건강하지 않을 때 기계의 주요 feature들의 경향성도 확인해보았습니다**.

#17

**이렇게 CMAPSS와 NCMAPSS를 메인으로 RUL 예측 모델을 구현하고자 합니다. 또한 추가적으로 다음같은 여러 기기 노화 데이터셋을 사용해보고자 합니다. 터보팬 뿐 아니라 다른 기계들에 대한 수명 예측도 진행하여 모델의 범용성을 높이고 성능을 높이고자 하기 때문입니다.**

#18

**지금부터는 사용할 모델에 대해 소개드리겠습니다.**

**첫번째 모델은 RVE입니다. RVE는 Recurrent Variational Encoder의 약자로, 모델 구조는 크게 인코더, 잠재공간, regressor 세 부분으로 나눠져 있습니다. 인코더는 LSTM으로 구성되어 있으며, raw input data를 저차원 공간에 매핑하여 잠재 표현을 생성하는 역할을 합니다. 잠재 표현이란 고차원 feature를 가진 input 데이터의 차원을 축소시킨 것으로, input 데이터를 잘 설명할 수 있는 특징을 찾을 수 있습니다. regressor는 활성화 함수와 FC layer로 구성되었으며, 잠재표현을 이용하여 최종적으로 RUL을 예측합니다.**

#19

**타 모델들은 정량적인 RUL 값만 반환하는데, RVE는 RUL 예측 결과 뿐 아니라 가운데 그림처럼 샘플에 대한 잠재공간 속 근접성을 기반으로 특정 순간의 항공기 상태를 시각적으로 진단해줍니다. 이는 다른 모델에서는 부족한 설명력을 올린다는 장점이 있습니다. 또한 타 모델에 비해 CMAPSS 내에서 난이도가 높은 데이터셋인 FD2와 FD4에서 높은 RUL 예측 성능을 보였는데, 이는 잠재 표현으로 인풋 데이터의 feature를 잘 추출했기 때문일 것입니다.**

(왜 FD2,FD4에서 성능이 좋았느냐? = 다른 모델들에 비해 왜 얘가 더 좋냐? -> FD1,FD3보다 FD2,FD4가 고장조건이 많아 더 feature가 많다. 고차원 feature를 가졌을수록 차원을 축소하는 잠재 표현이 효과적이기 때문에)

(VAE도 있는데 걔 쓰지 왜 RVE 쓰냐? -> RVE는 VAE와 다르게 단순 디코더, 즉 분류기 역할 뺴고 regressor를 넣었기 때문에 비슷한 RUL 값을 가진 샘플들도 잘 구분하여 매핑하므로 중복이 적고, 잠재 표현을 잘 표현해준다.)

#20

**저희는 이러한 RVE 모델의 핵심 구성 요소인 window 길이, smoothing, 내부 뉴런 수, 하이퍼파라미터 등을 변경해가며 여러가지 경우를 실험하고, 평가지표를 확인하여 성능을 측정할 것입니다. 또한 추가적으로 데이터셋을 변경하거나 모델에 새로운 구조나 기법도 적용해 볼 예정입니다.**

(윈도우 길이를 바꾸면 어떻게 되나요? – 과거 시점의 데이터를 얼마나 현재 시점 및 미래시점 측정값 예측에 많이 사용할까에 대한 척도인 것 같습니다. 따라서 윈도우를 길게 하면 과거시점을 많이 보게 되고, 짧게 하면 과거시점을 적게 볼 것이라고 생각합니다.)

(그럼 무조건 긴 윈도우가 좋지 않나요? 적절한 윈도우가 좋을 것 같음. 너무 윈도우가 길면 히스토리에 overfitting될 수 있지 않을까. 또한 샘플마다 살아남아있는 전체 사이클 길이가 다르므로 윈도우를 너무 길게 또는 짧게 하면 전체 사이클이 짧거나 긴 샘플에게는 치명적으로 작용할 수 있을 것 같습니다. 따라서 데이터셋의 일반화된 특성을 추출할 수 있도록 짧지도 길지도 않은 윈도우를 찾는 것이 중요할 듯.)

(새로운 구조나 기법 하신다했는데, RVE에 뭘 하실건가요? -> 이 RVE 저자들이 지정한 하이퍼파라미터 뿐 아니라 GA나 basian과 같은 다양한 하이퍼파라미터 최적화 방법을 적용해서 더 좋은 성능의 하이퍼파라미터를 찾아보고 싶기도 하고, regressor나 encoder의 구조를 변경해보고도 싶습니다.)

#21

**그 다음으로 사용해볼 모델은 stacked DCNN입니다. 이 모델은 크게 두 레벨으로 나뉘어져 있습니다. 레벨 one은 인코더로, feature 추출, 패턴 인식에 특화된 DCNN을 사용하여 항공기 데이터의 고차원 feature를 저차원 feature로 축소하고 노이즈를 제거하여 encoding들을 생성합니다. 이러한 인코딩들을 인풋으로 받는 레벨 two는 DCNN을 사용하며 RUL을 예측합니다.**

(레벨 1 모델은 RUL을 상당히 잘 추정하기 때문에, 레벨 2 모델은 추세를 포착하고 RUL을 개선하기 위해 희소 인코딩 세트만 있으면 된다.)

#22

**stacked DCNN의 output은 RUL 예측입니다. 이 논문은 기존 CMAPSS가 아닌 NCMAPSS를 사용하여 더 많고 정교한 인풋 데이터를 사용했고, DCNN의 장점인 높은 특징 추출력을 발휘하여 6.24의 낮은 RMSE를 달성하였기에 이번 프로젝트에 선정하게 되었습니다.**

#23

**저희는 stacked DCNN의 RUL smoothing, training 데이터 개선 등을 통해 원작 논문에서 아쉽다고 언급한 부분을 개선하고 성능을 측정할 것입니다, 추가적으로 데이터셋을 변경하거나 새로운 모델 구조와 기법도 적용해 볼 예정입니다.**

(RUL smoothing – 노이즈를 줄인다. (칼만 필터 등)

(training 과정에 포함한다 – 훈련때는 어떤 사이클 A보다 낮은 수의 사이클을 갖는 입력은 제외해버려놓고, 검증때는 A 사이클로 채워놓았다. 따라서 채울꺼면 둘다 채우고, 뺄거면 둘다 빼거나, 아니면 더 나은 filling 방법을 사용한다.)

(training fold의 데이터의 균형 개선)

(random gap – 인코딩들 간의 gap까지도 RUL 예측에 반영해서 RUL 예측 성능을 개선해봐라)

(새로운 모델 구조나 기법은? – Grid Search, Random Search, basyen 기법, GA 등 여러 논문에서 여러가지 하이퍼파라미터를 최적화하는 기법을 사용하는 것을 확인하였습니다. 이런 최적화기법도 다양하게 사용해볼 것이고, 인코더와 리그레서 모두에 DCNN을 사용했는데, DCNN 대신 LSTM이나 다른 구조를 끼워넣어 보는 형식도 진행해보고 싶습니다. )

#24

**그 외에도 여러 SOTA 모델 및 기법들을 참고할 예정입니다. 그 중 semi supervised deep architecture는 데이터셋의 labeling 비율을 달리하여 라벨되지 않은 데이터에 비지도 학습을 적용하여 RUL 예측 성능을 높인 구조입니다. 해당 논문에서는 RBM과 LSTM, FNN을 사용하여 Deep Architecture를 구성하였습니다. semi supervised는 데이터 자체의 분포도 잘 모델링하기에 라벨링된 훈련 데이터가 줄어도 높은 RUL 성능을 냅니다. 실제 RUL 예측을 할 때도 고품질의 라벨링 데이터를 얻기 어렵고, 시간이 오래 걸리므로 준지도 학습 또한 연구를 진행해보고자 합니다. 또한 지금까지 살펴본 SOTA 모델들에 사용된 다양한 ML 기법들을 총동원하여 custom한 모델을 제작해보고 싶습니다.**

(또한 성능을 높이기 위한 training 기법으로는 Dropout과 GA를 적용하였습니다.)

(RBM - label이 없는 raw input data에서 추상적 특징을 자동으로 학습하기 위 비지도 사전 훈련 단계로 활용한다. 지도 fine-tuning이 수행되기 전에 good local minimum 근처에서 가중치를 초기화 한다. 그 다음 LSTM – 장기 종속성까지도 학습하여 RUL 예측을 진행한다. FNN – output 출력)

(ML 기법은 뭐뭐가 있느냐? 하이퍼파라미터 최적화 기법들, 가우시안, 볼츠만 등 확률 분포들, CNN, RNN, LSTM, AR 등 예측에 쓰이는 다양한 모델들을 사용해보고 싶습니다. 또한 알고리즘은 제안했는데 코드는 공개하지 않는 논문이 있을 수도 있으므로 코드가 없는 모델을 구현하는 것도 해보고 싶습니다.)

#25

**또한 기계의 고장을 대비하기 위해 알람이 울리는 시점을 설정하는 Threshold 자체도 연구해보고자 합니다. 한 연구결과에 따르면 CMAPSS 데이터셋에서 뉴럴 네트워크로 RUL을 예측하는 경우, RUL 예측값이 평탄하다가 선형적으로 0으로 감소하는 양상을 보였습니다. 성능 저하는 기계의 초기고장이 발생해서야 알아차릴 수 있기에 초기 RUL 값인 Threshold를 상수로 하는 것이 좋습니다. 또한 RUL을 상수로 제한했을 때 오른쪽과 같이 RUL 예측 노이즈가 감소하였으므로 저희도 해당 논문에서 사용한 Threshold 값인 130을 기준으로 하고자 하며, 기계가 고장나기 전에 알람을 울리는 것이 목적이므로 130보다 작은 값들도 Threshold로 설정해보고자 합니다**.

(왜 상수로 하냐? 상수로 제한하다가 감소해야 그 지점부터 그래프가 꺾여 언제 고장이 발생한 지점인지를 파악가능함)

#26

**이제 저희의 인력 운영계획에 대해 말씀드리겠습니다. 전자공학부인 저와 컴퓨터공학과인 김서연 학생의 역량을 발휘하여 연구과제를 진행할 것이고, 특히 실제 설비에 대한 전문지식을 가지고 있지는 않기에 예지정비솔루션 및 설비 데이터에 대해 중점적으로 멘토링을 받고 싶습니다.**

#27

**이러한 방향성을 가지고 5월부터 10월까지 모델 구현과 성능 테스트, 멘토링을 거치며 유종의 미를 거두고자 합니다.**

#28

**마지막으로 연구비 사용 계획을 말씀드리겠습니다. 저희의 장비에는 GPU가 없어서 무료 colab으로 train 코드를 실행시켜보았으나 RAM이 부족하여 데이터셋을 로드할 수조차 없었습니다. 따라서 저희는 더 향상된 환경을 가진 colab pro+를 통해 협업 및 코딩을 진행하고자 합니다.**

#29

**그리고 colab pro + 에서 구동할 수 없을 만큼 큰 데이터셋이나 모델의 경우에는 GPU가 탑재된 데스크탑이나**

#30

**AWS cloud를 구매하여 연구를 진행하고자 합니다. 총 예산은 600만원내에서 진행할 수 있도록 구성하였고, 제한된 비용과 환경에서도 최적의 솔루션을 도출하고자 노력하겠습니다.**

#31

**그러면 QnA를 받으며 발표 마치도록 하겠습니다. 긴 발표 들어주셔서 감사합니다.**

**-------------------------------------------------------------------------------------------------------------**

**포괄질문**

Q. 역할 분담은 어떻게 했는지, 1차 준비하면서 힘들었던 점, 배운 점?

- 이서윤 학생은 SOTA 모델 논문 탐색과 이해를 하여 연구방향성을 제시하여 계획을 수립하고 할 일들을 설정하였습니다. 김서연 학생은 SOTA 모델의 코드를 분석하고 개선사항을 파악하였습니다. 또한 데이터셋을 전처리 및 시각화하였습니다. 1차를 준비하면서 우선 논문에 쓰인 기술들을 이해하고 어떻게 개선해야할지를 고민하는 것이 어려웠습니다. 그러나 이 기회를 통해 원하는 문제를 해결하기 위해 연구 방향성을 잡고 더 좋은 기법은 찾아가는 과정을 경험하게 되었습니다.

Q 앞으로 SKT 하게되면 코치들한테 뭘 부탁하고 싶은지.

실제 설비에 모델을 도입하는 데 있어 전문지식을 가지고 있지는 않기에 예지정비솔루션 및 설비 데이터에 대해 중점적으로 멘토링을 받고 싶습니다.

Q 지원 동기는? 꼭 해당 skt ai fellowship이어야만 한 이유는?

저희 팀은 SKT Fellowship이 다른 공모전과는 다르게 실제 현업에서 쓰이는 과제를 연구하고 SKT 회사 서비스에 실제로 적용될 수도 있다는 점이 특별하다고 생각하였습니다. 연구 과제에 대해서도 단순히 이미 있는 모델을 차용하는 것이 아니라 현존하는 모델보다 뛰어난 성능을 내기 위해 저희만의 새로운 시도를 하고, 그 때의 성능을 평가하는 과정을 반복하며 머신러닝 엔지니어로서의 역량을 키우고 싶습니다. 뿐만 아니라 SKT라는 권위있는 회사 소속이신 인공지능, 데이터 전문가 멘토님께 5개월동안 연구비와 멘토링을 받을 수 있다는 점도 매력적이었습니다. 그래서 꼭 skt AI 4기 Fellow가 되고 싶습니다.

Q 해당 주제를 선택한 이유는?

ML의 발전에 따라 앞으로 시계열 데이터가 가지는 잠재력이 무궁무진할 것이라 생각하였고, 공학도의 입장으로서 기기 및 설비의 데이터를 정제 및 분석하여 더 나은 세상을 만든다는 모토와도 일치하였기에 설비 잔여 수명 예측 주제를 선택하였습니다.

Q. 팀명 이유는?

서연 학생이 좋아하는 다쿠아즈와 서윤 학생이 좋아하는 다코야키 음식을 합쳐 만든 이름입니다. 좋아하는 음식처럼 더 나은 성능을 내는 모델을 만들 때에도 열정을 다하자는 의미가 담겨있습니다.

\*\*시계열 데이터란? 이 데이터를 다루는데 가장 중요한 포인트는(기술은?)\*\*

시게열데이터란 시간의 흐름에 따라 순차적으로 발생한 관측치의 집합을 말합니다. 가장 중요한 포인트는 시점에 따라 데이터들이 연관 되어있다는 점입니다. 따라서 데이터간의 의존성의 패턴을 잘 나타내는 데이터 분석 기법이나 Recurrent한 DL 기술을 사용하여 연구가 많이 진행되는 것 같습니다. (RNN, LSTM, AR)

Q. 유사한 프로젝트를 진행한 경험은? 거기서 내 역할은? 배운 점은?

서윤 학생의 경우 칼만 필터를 이용하여 측위 데이터를 추정해본 경험이 있고, 딥러닝 모델의 성능 최적화와 커스텀 데이터셋을 제작하여 전이 학습을 진행하여 서비스를 배포해보았습니다. 이 과정에서 머신러닝 학습 시 데이터셋이 매우 중요하며, 성능을 향상시키기 위해 필요한 요소들에 대해 알게 되었습니다.

서연 학생의 경우 데이터 전처리부터 시작하여 데이터셋 시각화 업무를 맡았었고, 머신러닝 모델의 하이퍼파라미터를 튜닝하며 성능을 평가하여 신용카드 사기 검출, 영화 감성분석 등의 테스크를 진행해보았습니다. 특히 사전 훈련된 모델들의 layer를 직접 뜯어보며 구현해본 경험으로 머신러닝 모델의 성능을 끌어올리는 법을 배우게 되었습니다.

Q. 데이터 분석가 / 머신러닝 엔지니어?(관련 직종)에게 가장 필요한 역량은?

머신러닝 엔지니어 – 최신 논문들에 대한 관심과 연구 동향 파악 능력이 있어야 되고, MLOps와 같이 ML모델 뿐 아니라 데이터 분석, ML 모델 학습 및 배포단계까지 아우를 줄 알아야 한다고 생각합니다.

데이터 분석가 – 효율적인 데이터 적재 능력(빅데이터), 시각화 및 의사소통 능력, 중요한 feature를 잘 관통해서 인사이트 도출

Q. 포부..?ㅎ

저희 팀은 현재 인공지능 엔지니어를 꿈꾸고 있고, 휴학을 하여 인공지능 공부에 전념하고 있습니다. 따라서 SKT Fellowship에 집중하여 개발을 진행할 수 있고, 거주지도 가까워서 원활한 소통이 이뤄진다는 장점이 있습니다. 꼭 이번 SKT Fellowship 4기 fellow으로 활동하며 잔여수명 예측 프로젝트에서 좋은 성과를 내서 그랜드뷰 서비스에 기여를 하며 엔지니어로서의 꿈을 이루고, 인공지능 전문가로 성장하고 싶습니다.

Q. \*\*예상 결과 및 기대 효과는\*\*?

RMSE를 기존 SOTA 논문의 모델의 것보다 낮추는 것이 가장 큰 목표이고, 주어진 연구 환경 내에서 머신러닝 모델의 더 효율적인 관리 및 유지를 하는 것이 두번째 목표입니다. 그 후에 그랜드뷰에 탑재하여 제조기업의 설비 관리에 도움을 주고 싶습니다.