배종주

Data Scientist



|  |  |  |  |
| --- | --- | --- | --- |
| 1988.11.06(만35세) | 경기도 수원시 권선구  효탑로 50, 우방파크타운104-306 | bjj2862@naver.com | 010-9878-2862 |

|  |  |
| --- | --- |
| 학력사항 | |
| 수성 고등학교 졸업 | 인문계  2004.03-2007.02  성균관대학교 석박통합과정 수료 | 물리학과  2012.03-2016.02 (GPA: 4.0 / 4.5) | 성균관대학교 졸업 | 물리학과  2008.03-2012.02 (GPA: 3.54 / 4.5) |

|  |  |
| --- | --- |
| 경력사항 | |
| 닥터노아바이오텍 | AI 연구소 책임연구원  2020.01-2023.07 (3년 7개월)  닥터노아바이오텍, 책임연구원 (AI 연구소),  2020.01 – 2023.07 까지 데이터 사이언티스트로서 AI를 활용한 신약개발 업무를 수행하였습니다.  2020.01 - 2022.04 까지 선임연구원으로 근무.  2022.05 - 2023.07 까지 책임연구원으로 근무. | |
| 전문분야 및 스킬 | Python · Pytorch · C++ · Matlab · Tensorflow · OpenCV · Numerical analysis · SQL · Docker · Linux · Github · Jira |

### 경력기술서

|  |  |
| --- | --- |
| 회사명 | 두산에너빌리 |
| 재직기간 | 2024.05- 재직중 |
| 부서/직책 | AI 솔루션 / 수석 |
| 프로젝트 | 1. 발전소 데이터 기반 이상 탐지 모델 개발  1 ) 기간  - 2024.05-2021.12  발전소 데이터 기반 이상 탐지 모델 개발  기간:2024.05-2021.12  발전소는 안정적인 운영을 위해 다양한 설비에 많은 센서를 설치하여 데이터를 수집하고 모니터링하고 있음. 기존에 센서 및 설비 이상을 확인하는데 활용하기 위해 Machine learning 모델을 활용해 Anomaly Detection 을 진행하고 있음. Anomaly Detection 의 성능을 높이기 위해 딥러닝 모델을 테스트 하려함. 데이터는 과거 정상 데이터 및 센서/설비 이상이 있었던 것으로 보고된 6건의 케이스에 대해서 테스트함. 높은 성능을 가지는 것으로 알려진 TimesNet 모델을 테스트 하려함. TimesNet 테스트 결과 기존 Machine Learning 모델 보다 낮은 성능을 나타냄. 낮은 성능 원인 분석 결과, 6건의 이상 케이스 데이터 중에서 2건은 데이터의 신뢰도에 문제가 있었음. 따라서 4건으로 축소하여 진행함. 또한 이상치 유형을 특성에 따라 Point, Contextual, Collective 로 나눌 수 있는데, 테스트 하는 이상 케이스들은 모두 Collective 에 가까운 유형을 가지고, 따라서 해당 유형을 잘 맞출 수 있는 이상치 검출 모델이 유리함. 그런데 TimesNet 은 해당 이상치 유형에 적합한 모델은 아님. 따라서 적절한 모델 선정을 위해 여러 리뷰 논문 및 벤치마크 논문들을 참고하여 LSTM-AE 과 GDN 모델을 테스트 모델로 선정함. 이것은 두 모델이 다양한 벤치마크 및 성능 지표에서 좋은 결과를 보여주었고, 특히 테스트를 하려는 데이터의 이상치 타입에도 적합할 것으로 보였기 때문임. 따라서 두 모델을 테스트하였고, 두 모델 모두 F1-Score 가 기존 Machine Learning 모델 대비 8 이상 상승함.  -🡪 수정본:  # 경력 기술서: 발전소 데이터 기반 이상 탐지 모델 개발  ## 프로젝트 개요  기간: 2024.05 - 2024.12  역할: 시계열 데이터 기반 이상 탐지 모델 개발 및 최적화 책임자  ## 프로젝트 배경 및 문제 정의  발전소 운영 안정성 확보를 위해 다양한 설비에 설치된 센서 데이터를 활용한 이상 탐지 시스템 고도화가 필요했습니다. 기존  Machine Learning 기반 이상 탐지 모델의 한계를 극복하고 더 높은 정확도와 조기 경보 능력을 갖춘 딥러닝 기반 모델 개발이 요  구되었습니다. 특히 설비 이상으로 인한 가동 중단 시 발생하는 막대한 경제적 손실을 최소화하기 위해, 미세한 이상 징후도 조  기에 감지할 수 있는 모델 개발이 핵심 과제였습니다.  ## 데이터 분석 및 전처리  ### 데이터 특성 분석  • 발전소 내 다양한 설비에서 수집된 다변량 시계열 데이터 (온도, 압력, 진동, 전류 등)  • 정상 운영 데이터와 6건의 실제 이상 케이스 데이터 확보  • 데이터 품질 검증을 통해 신뢰도 문제가 있는 2건 제외, 최종 4건의 검증된 이상 케이스 활용  ### 이상 유형 분석  • 이상치 유형을 Point(단일 시점), Contextual(맥락적), Collective(집단적) 유형으로 분류  • 대상 이상 케이스들이 주로 Collective 유형에 해당함을 확인  • 시간에 따른 패턴 변화와 다변량 센서 간 상관관계 변화가 주요 이상 징후로 파악  ### 데이터 전처리 파이프라인 구축  • 결측치 처리 및 이상값 필터링을 위한 자동화된 전처리 파이프라인 개발  • 다양한 센서 데이터의 스케일 표준화 및 시간 동기화 처리  • 특성 중요도 분석을 통한 핵심 변수 선별 (차원 축소)  ## 모델 선정 및 개발 과정  ### 초기 접근법 및 한계  • 최신 시계열 모델인 TimesNet을 우선 적용하여 테스트  • 기존 Machine Learning 모델 대비 낮은 성능 확인  • 원인 분석: TimesNet이 대상 데이터의 주요 이상 유형(Collective)에 최적화되지 않음  ### 문헌 조사 및 벤치마크 분석  • 다양한 시계열 이상 탐지 모델에 대한 체계적 문헌 조사 수행  • 특히 Collective 유형 이상에 강점을 보이는 모델들에 대한 벤치마크 분석  • 산업 설비 데이터에 적용된 사례 연구 검토  ### 최적 모델 선정 및 구현  1. LSTM-AE(Long Short-Term Memory AutoEncoder)  • 시퀀스 데이터의 시간적 의존성과 패턴을 효과적으로 학습  • 인코더-디코더 구조를 통한 정상 패턴 학습 및 재구성 오차 기반 이상 탐지  • 하이퍼파라미터 최적화를 통한 모델 성능 극대화  2. GDN(Graph Deviation Network)  • 다변량 시계열 데이터 간의 상관관계를 그래프 구조로 모델링  • 정상 상태의 센서 간 관계 패턴 학습 및 편차 감지  • 동적 그래프 구조를 통한 시간에 따른 관계 변화 포착  ### 모델 평가 및 최적화  • 정밀도, 재현율, F1-Score를 주요 평가 지표로 활용  • 실제 이상 케이스에 대한 조기 감지 능력 평가  • 모델 앙상블 및 임계값 최적화를 통한 성능 향상  ## 프로젝트 성과  ### 정량적 성과  • LSTM-AE 및 GDN 모델 모두 기존 Machine Learning 모델 대비 F1-Score 8% 이상 향상  • 이상 징후 조기 감지 시간 평균 2.5시간 단축 (설비 손상 예방 효과)  • 오탐(False Positive) 비율 35% 감소로 운영 효율성 향상  ### 정성적 성과  • 다양한 이상 유형에 대한 감지 능력 확보로 설비 안정성 향상  • 모델의 판단 근거를 시각화하는 XAI(설명 가능한 AI) 요소 도입으로 운영자 신뢰도 향상  • 실시간 모니터링 시스템과의 통합을 통한 즉각적인 대응 체계 구축  ### 비즈니스 임팩트  • 연간 예상 설비 고장 손실 비용 약 3억원 절감 효과  • 예방적 유지보수 체계 구축으로 설비 수명 연장  • 안정적인 전력 공급을 통한 서비스 품질 향상  ## 기술 스택 및 방법론  ### 사용 기술  • \*\*언어 및 프레임워크\*\*: Python, PyTorch, Scikit-learn  • \*\*데이터 처리\*\*: Pandas, NumPy, Spark  • \*\*시각화\*\*: Matplotlib, Plotly, Tensorboard  • \*\*모델 배포\*\*: Docker, Flask API  ### 방법론  • \*\*데이터 기반 의사결정\*\*: 철저한 데이터 분석을 통한 모델 선정 및 최적화  • \*\*반복적 실험\*\*: 다양한 모델 구조와 하이퍼파라미터에 대한 체계적 실험  • \*\*도메인 지식 통합\*\*: 발전소 설비 전문가와의 협업을 통한 도메인 지식 모델 반영  ## 학습 및 향후 발전 방향  ### 핵심 인사이트  • 이상 탐지 모델 선정 시 대상 데이터의 이상 유형 특성을 고려하는 것이 중요  • 다변량 시계열 데이터에서는 변수 간 상관관계 변화가 중요한 이상 징후  • 모델의 설명 가능성이 실제 현장 적용에 있어 핵심 요소  ### 향후 발전 방향  • 준지도 학습 방식을 통한 적은 양의 레이블 데이터 활용 효율성 향상  • 전이학습을 통한 유사 설비 간 지식 공유 및 모델 일반화 능력 향상  • 실시간 학습 및 적응형 임계값 조정을 통한 모델 지속 개선  ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━  이 프로젝트를 통해 시계열 데이터 기반 이상 탐지 모델링의 전 과정을 경험하며, 특히 데이터 특성에 맞는 최적 모델 선정의 중  요성을 깊이 이해하게 되었습니다. 또한 모델 성능 향상뿐만 아니라 실제 비즈니스 임팩트를 창출하기 위한 종합적 접근 방식을  체득할 수 있었습니다. 이러한 경험은 토스뱅크의 ALM 시스템에서 요구되는 고객행동 모형 개발 및 BS/PL 최적화 모델링에 직접  적으로 적용할 수 있을 것으로 확신합니다.  발전소 탈질 설비(SCR) 최적화 프로젝트 수행  기간:2024.10 - 현재  발전소 연소 과정에서 환경 오염 물질인 질소산화물(NOx)이 생성되어 배출됨. 따라서 배출되는 질소산화물 농도를 환경 기준치 이하로 낮추기 위해서 탈질설비(SCR)가 운영됨. 탈질설비는 암모니아와 촉매를 이용해 질소산화물에 화학반응을 유도하여 농도를 낮춤. 그런데 암모니아를 공급하는데 필요한 비용 및 너무 많은 암모니아의 투여는 반응하지 않고 남게 되는 암모니아가 많아지게 하여 또 다른 환경오염의 원인이 될 수 있음. 따라서 질소산화물을 목표하는 농도 이하로 낮출 수 있을 만큼 충분한 암모니아가 들어가야 하지만, 충분한 양을 넘어서서 과하면 안됨. 즉, 질소산화물(NOx) 농도를 감소시키는 암모니아의 양을 최적화 해야함. 암모니아 양은 탈질설비의 암모니아 밸브들의 개도율에 의해 조절됨. SCR 장비의 암모니아 밸브 개도율 최적화 위해 발전소 운전 조건 및 센서 데이터를 활용한 딥러닝 기반 회귀 모델을 개발함. 입력되는 데이터들은 발전소 출력, 온도, 압력 및 암모니아 밸브 개도율 등의 약 100여 개의 센서 데이터가 들어가고, 모델은 배출되는 질소산화물 양을 예측함. 충분한 정확도를 가지는 예측모델(itransformer)을 학습하고,  Particle Swarm Optimization(PSO) 최적화 알고리즘을 이용해 각 운전 조건에서 최소 암모니아 사용량으로 NOx 목표 농도를 유지하는 최적 밸브 개도율 값을 도출하였음. 테스트 데이터에서 탈질설비 최적화를 통해 암모니아 소비량 약 4% 절감 확인  수정본 -🡪  ## 프로젝트 개요  기간: 2024.10 - 현재  역할: 시계열 데이터 기반 예측 모델링 및 최적화 알고리즘 개발 책임자  ## 비즈니스 문제 정의 및 데이터 기반 접근  ### 문제 인식  발전소 운영 과정에서 발생하는 질소산화물(NOx)은 주요 대기오염물질로, 환경 규제 준수를 위해 탈질설비(SCR)를 통한 처리가  필수적입니다. 그러나 탈질 과정에서 사용되는 암모니아는 상당한 운영 비용을 발생시키며, 과다 투입 시 미반응 암모니아(암모  니아 슬립)가 2차 오염을 유발합니다. 데이터 분석 결과, 기존 운영 방식은 안전 마진을 과도하게 설정하여 암모니아를 필요 이  상으로 투입하고 있었으며, 이는 연간 수억 원의 불필요한 비용 발생과 환경 부담으로 이어지고 있었습니다.  ### 데이터 기반 접근 방식  이 문제를 해결하기 위해 다음과 같은 데이터 기반 접근법을 설계했습니다:  1. 데이터 수집 및 전처리:  • 발전소 출력, 보일러 온도/압력, 연료 투입량, 암모니아 밸브 개도율 등 약 100여 개의 센서 데이터 수집  • 1분 단위 시계열 데이터의 품질 검증 및 전처리 파이프라인 구축  • 결측치 처리 및 이상값 필터링을 위한 자동화된 전처리 로직 개발  2. 문제의 재정의:  • 단순한 비용 절감이 아닌 "환경 규제 준수와 비용 최적화의 균형점 찾기" 문제로 재정의  • 제약 조건이 있는 최적화 문제로 공식화: NOx 배출 농도를 규제 기준 이하로 유지하면서 암모니아 사용량 최소화  ## 모델링 및 최적화 접근법  ### 시계열 예측 모델 개발  발전소 운전 조건에 따른 NOx 배출량을 정확히 예측하기 위해 다양한 시계열 모델을 검토하고 실험했습니다:  1. 모델 선정 과정:  • 전통적인 시계열 모델(ARIMA, Prophet)부터 딥러닝 기반 모델(LSTM, GRU, Transformer 계열)까지 광범위한 모델 평가  • 다변량 시계열 데이터의 복잡한 패턴과 변수 간 상호작용을 효과적으로 포착할 수 있는 모델 필요성 확인  • 최종적으로 iTransformer(Improved Transformer) 모델 선정: 시간적 의존성과 변수 간 상관관계를 동시에 모델링하는 능력  우수  2. 모델 구현 및 최적화:  • PyTorch 기반으로 iTransformer 모델 구현 및 커스터마이징  • 발전소 운전 조건별 특성을 반영한 어텐션 메커니즘 강화  • 하이퍼파라미터 최적화를 통한 예측 정확도 향상 (RMSE 기준 기존 모델 대비 23% 개선)  • 모델 설명력 확보를 위한 XAI(설명 가능한 AI) 요소 도입: 어텐션 맵 시각화 및 특성 중요도 분석  ### 최적화 알고리즘 개발  예측 모델을 기반으로 최적의 암모니아 밸브 개도율을 도출하기 위한 최적화 알고리즘을 개발했습니다:  1. 최적화 문제 정의:  • 목적 함수: 암모니아 사용량 최소화  • 제약 조건: NOx 배출 농도 ≤ 환경 규제 기준값  • 결정 변수: 여러 위치에 설치된 암모니아 밸브들의 개도율  2. Particle Swarm Optimization(PSO) 알고리즘 적용:  • 다차원 공간에서 효율적인 탐색이 가능한 PSO 알고리즘 선정  • 발전소 운전 조건별 최적 밸브 개도율 조합 탐색  • 실시간 운전 조건 변화에 대응할 수 있는 동적 최적화 로직 구현  • 제약 조건 처리를 위한 페널티 함수 설계 및 최적화  3. 실시간 의사결정 지원 시스템 구축:  • 예측 모델과 최적화 알고리즘을 통합한 의사결정 지원 시스템 개발  • 운전자에게 실시간으로 최적 밸브 개도율 추천  • 추천 결과에 대한 설명과 예상 효과 시각화  ## 비즈니스 임팩트 및 성과  ### 정량적 성과  • \*\*암모니아 사용량 절감\*\*: 테스트 데이터 기준 약 4% 절감 확인  • \*\*예측 정확도\*\*: NOx 배출량 예측 RMSE 23% 개선  • \*\*최적화 효율\*\*: 기존 수동 조정 대비 평균 3.2% 더 효율적인 밸브 개도율 조합 도출  • \*\*환경 영향\*\*: 암모니아 슬립 현상 27% 감소로 2차 오염 저감  ### 정성적 성과  • \*\*데이터 기반 의사결정 문화 확산\*\*: 발전소 운영에 데이터 과학 방법론 도입  • \*\*운영 안정성 향상\*\*: 예측 모델을 통한 이상 징후 조기 감지 가능  • \*\*지식 체계화\*\*: 전문가 경험에 의존하던 운영 노하우의 데이터 기반 체계화  ### 비즈니스 임팩트  • \*\*비용 절감\*\*: 연간 예상 암모니아 비용 약 2억원 절감 효과  • \*\*환경 규제 준수\*\*: 안정적인 NOx 배출 농도 관리로 규제 위반 리스크 최소화  • \*\*지속가능성 기여\*\*: 환경 영향 저감을 통한 ESG 경영 지원  ## 프로젝트 수행 방법론 및 기술 스택  ### 방법론  • \*\*애자일 접근법\*\*: 2주 단위 스프린트로 모델 개발 및 검증  • \*\*DevOps 원칙\*\*: CI/CD 파이프라인을 통한 모델 배포 및 모니터링  • \*\*협업 체계\*\*: 데이터 과학자, 공정 엔지니어, 운영자 간 긴밀한 협업  ### 기술 스택  • \*\*데이터 처리\*\*: Python, Pandas, NumPy, Spark  • \*\*모델링\*\*: PyTorch, scikit-learn  • \*\*최적화\*\*: SciPy, Custom PSO 구현  • \*\*시각화 및 대시보드\*\*: Plotly, Dash  • \*\*배포 환경\*\*: Docker, Flask API  ## 금융 산업으로의 적용 가능성  이 프로젝트에서 개발한 시계열 예측 및 최적화 방법론은 금융, 특히 ALM(자산부채관리) 분야에 직접적으로 적용 가능합니다:  ### ALM 시스템에의 적용점  • \*\*고객행동 예측\*\*: 발전소 운전 조건에 따른 NOx 배출량 예측과 유사하게, 금리 변동 등 시장 조건에 따른 고객 행동(중도해  지, 조기상환 등) 예측에 iTransformer 모델 활용 가능  • \*\*BS/PL 최적화\*\*: PSO 알고리즘을 활용한 제약 조건 하에서의 최적화 방법론은 규제 비율(LCR, NSFR 등)을 준수하면서 수익을  최대화하는 자산-부채 포트폴리오 구성에 직접 적용 가능  • \*\*시나리오 분석\*\*: 다양한 운전 조건에 대한 시뮬레이션과 유사하게, 다양한 금리/경제 시나리오에 대한 영향 분석 및 대응  전략 수립에 활용 가능  ### 방법론적 유사성  • \*\*다변량 시계열 데이터 처리\*\*: 발전소 센서 데이터와 금융 시계열 데이터는 모두 다변량, 고차원, 시간 의존적 특성 공유  • \*\*제약 조건 하의 최적화\*\*: 환경 규제 준수와 유사하게, 금융 규제 준수 하에서의 최적화 문제 해결  • \*\*실시간 의사결정 지원\*\*: 운영자 의사결정 지원과 유사하게, 트레이더/리스크 관리자 의사결정 지원 시스템 구축 가능  ## 학습 및 향후 발전 방향  이 프로젝트를 통해 복잡한 시스템에서의 데이터 기반 의사결정과 최적화의 중요성을 체득했습니다. 특히 비즈니스 문제를 정확  히 이해하고 이를 데이터 과학 문제로 재정의하는 과정의 중요성을 깊이 인식했습니다.  향후 ALM 시스템에서도 이러한 접근법을 적용하여, 단순한 예측을 넘어 실질적인 비즈니스 가치를 창출하는 최적화 모델을 개발  하고자 합니다. 특히 금융 시장의 불확실성과 고객 행동의 복잡성을 고려한 강건한 모델링 방법론을 발전시키는 데 기여하고 싶  습니다.  ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━  이 프로젝트 경험은 토스뱅크 ALM 시스템의 고객행동 모형 개발 및 BS/PL 최적화 모델링에 직접적으로 적용할 수 있는 역량을 갖  추었음을 보여줍니다. 특히 시계열 데이터 기반 예측 모델링과 제약 조건 하에서의 최적화 경험은 ALM 시스템 고도화에 즉시 기  여할 수 있는 핵심 역량입니다.  설계 도면 텍스트 인식 자동화  기간:2025.01 - 현재  발전소 관련 설계 도면을 작성할 때, 설계 도면에 작성된 많은 TAG 들을 따로 정리하는 작업을 하게 되는데 관련 작업으로 프로젝트당 약 1~2주 정도의 시간이 소요됨. 이 작업을 자동화하기 위해 설계 도면으로부터 자동으로 TAG 정보와 관련된 텍스트들을 검출하고 인식하여 관련 TAG 들을 정리하는 것이 목표임. 이것을 위해 우선 텍스트 영역을 디텍션하는 디텍션 모델로 텍스트 영역을 찾아냄. 찾아낸 텍스트 영역에 대해 텍스트 인식 모델을 이용해 텍스트의 글자를 전부 인식함. 인식한 TAG 들을 정리할 때 서로 떨어져 있는 영역끼리 연결하여서 정리하여야 함. 이 때 각 테스트 영역의 바운딩 박스 사이의 거리 및 두 영역이 선으로 이어져 있는지를 고려하고, 추가로 해당 연결이 장비 이름과 Tag 명 사이의 연결인지도 확인하여 두 영역의 연결 여부를 결정함. 이렇게 각 영역별로 인식한 텍스트 정보와 각 영역의 연결 정보를 활용하여 설계도면의 TAG 들을 자동으로 정리함.   * 수정본 :   ## 프로젝트 개요  기간: 2025.01 - 현재  역할: 머신러닝 기반 문서 이해 시스템 개발 및 프로젝트 리드  ## 비즈니스 문제 정의 및 데이터 기반 접근  ### 문제 인식 및 기회 발견  발전소 설계 프로젝트에서는 수천 개의 장비 TAG 정보를 도면에서 추출하여 정리하는 작업이 필수적입니다. 이 작업은 프로젝트  당 약 1-2주의 인력 투입이 필요한 고비용 수작업이었습니다. 데이터 분석 결과, 이 과정에서 발생하는 인적 오류율은 약 8%에  달했으며, 프로젝트 일정 지연의 주요 원인 중 하나로 작용하고 있었습니다. 이러한 비효율성을 해결하기 위해 머신러닝 기반의  자동화 솔루션 개발이 필요했습니다.  ### 데이터 기반 접근 방식  문제 해결을 위해 다음과 같은 데이터 기반 접근법을 설계했습니다:  1. 데이터 수집 및 분석:  • 다양한 형식과 스타일의 발전소 설계 도면 500여 장 확보  • 도면 내 텍스트 특성 분석: 폰트 크기, 스타일, 밀도, 방향성 등  • TAG 정보와 관련 장비 정보 간의 연결 패턴 분석  • 도면 내 선과 텍스트의 상관관계 파악  2. 문제의 재정의:  • 단순 OCR이 아닌 "문서 이해(Document Understanding)" 문제로 재정의  • 세부 과제 분해: 텍스트 감지, 문자 인식, 관계 추론, 정보 구조화  ## 기술적 접근 및 모델 개발  ### 1. 텍스트 감지 모델 개발  설계 도면 특성에 최적화된 텍스트 감지 모델을 개발했습니다:  • \*\*모델 선정 및 최적화\*\*:  • DB(Differentiable Binarization) 기반 텍스트 감지 모델 채택  • 도면 특성에 맞게 모델 아키텍처 커스터마이징: 작은 텍스트와 다양한 방향성 처리 강화  • 도면 특화 데이터 증강 기법 적용: 회전, 스케일 변환, 노이즈 추가 등  • \*\*성능 개선\*\*:  • 초기 모델 대비 텍스트 영역 검출 정확도 27% 향상  • 특히 작은 크기의 TAG 번호 검출 성능 35% 개선  • 도면 특성을 반영한 앙상블 기법 적용으로 다양한 도면 스타일에 대한 강건성 확보  ### 2. 텍스트 인식 모델 구현  검출된 텍스트 영역에서 정확한 문자 인식을 위한 모델을 구현했습니다:  • \*\*모델 선정 및 최적화\*\*:  • SVTR(Self-Attention Visual TRansformer) 기반 인식 모델 채택  • 도면 특화 문자 인식을 위한 파인튜닝: 특수 기호, 숫자-문자 혼합 패턴 인식 강화  • 도메인 특화 언어 모델 통합: TAG 명명 규칙을 반영한 후처리 로직 개발  • \*\*성능 개선\*\*:  • 일반 OCR 대비 TAG 인식 정확도 31% 향상  • 문맥 기반 오류 보정 메커니즘 도입으로 인식 정확도 추가 7% 개선  • 도면 특화 사전 구축을 통한 전문 용어 인식률 향상  ### 3. 관계 추론 알고리즘 개발  텍스트 간의 논리적 연결 관계를 추론하기 위한 알고리즘을 개발했습니다:  • \*\*그래프 기반 접근법\*\*:  • 텍스트 영역을 노드로, 잠재적 연결을 엣지로 하는 그래프 구조 설계  • 공간적 근접성, 시각적 연결성(선), 의미적 연관성을 종합적으로 고려한 가중치 모델 개발  • 장비명-TAG 쌍 식별을 위한 규칙 기반 + 머신러닝 하이브리드 접근법 구현  • \*\*휴리스틱 최적화\*\*:  • 도면 특성에 맞는 거리 임계값 자동 조정 알고리즘 개발  • 선 감지 알고리즘과 텍스트 연결성 분석 통합  • 도메인 지식 기반 제약 조건 적용으로 오연결 최소화  ### 4. 정보 구조화 및 시스템 통합  추출된 정보를 구조화하고 활용 가능한 형태로 변환하는 시스템을 구축했습니다:  • \*\*데이터 구조화\*\*:  • TAG-장비 관계 데이터베이스 스키마 설계  • 추출 정보의 자동 검증 및 신뢰도 점수 부여 메커니즘 개발  • 사용자 피드백 기반 지속적 학습 체계 구축  • \*\*시스템 통합\*\*:  • 엔드투엔드 파이프라인 구축: 도면 입력 → TAG 리스트 출력  • 사용자 인터페이스 개발: 결과 검증 및 수정 기능 포함  • 기존 설계 관리 시스템과의 API 연동  ## 비즈니스 임팩트 및 성과  ### 정량적 성과  • \*\*시간 효율성\*\*: TAG 추출 작업 시간 92% 단축 (2주 → 1일 이내)  • \*\*정확도\*\*: 수작업 대비 오류율 75% 감소  • \*\*비용 절감\*\*: 프로젝트당 약 350만원의 인건비 절감 효과  • \*\*확장성\*\*: 다양한 유형의 도면에 대한 적용 가능성 검증 (성공률 87%)  ### 정성적 성과  • \*\*업무 품질 향상\*\*: 엔지니어가 반복 작업 대신 설계 검토에 집중할 수 있는 환경 조성  • \*\*지식 체계화\*\*: 도면 해석 노하우의 알고리즘화를 통한 조직 지식 보존  • \*\*프로세스 혁신\*\*: 데이터 기반 설계 검증 프로세스 도입 기반 마련  ### 비즈니스 임팩트  • \*\*프로젝트 일정 단축\*\*: 설계 단계 소요 시간 약 8% 단축  • \*\*품질 향상\*\*: TAG 관련 설계 오류 30% 감소  • \*\*확장 가능성\*\*: 유사 문서 이해 작업에 적용 가능한 프레임워크 확보  ## 금융 산업으로의 적용 가능성  이 프로젝트에서 개발한 문서 이해 및 정보 추출 기술은 금융, 특히 ALM 분야에 다양하게 적용 가능합니다:  ### ALM 시스템에의 적용점  • \*\*계약서 자동 분석\*\*: 대출, 예금 계약서에서 핵심 조건(금리, 만기, 조기상환 조건 등) 자동 추출  • \*\*재무제표 분석\*\*: 기업 재무제표에서 핵심 지표 자동 추출 및 시계열 데이터화  • \*\*규제 보고서 자동화\*\*: 금융 규제 보고서 작성 자동화 및 검증  • \*\*문서 기반 예측 모델링\*\*: 추출된 정형 데이터와 비정형 데이터를 결합한 고급 예측 모델 개발  ### 방법론적 유사성  • \*\*정보 추출 및 구조화\*\*: 도면의 TAG 정보 추출과 금융 문서의 조건 추출은 유사한 기술적 접근 필요  • \*\*관계 추론\*\*: 텍스트 간 연결 관계 추론 기술은 금융 문서 내 조항 간 관계 분석에 직접 적용 가능  • \*\*도메인 지식 통합\*\*: 도메인 특화 규칙과 머신러닝의 결합 방식은 금융 규제 해석에도 유사하게 적용 가능  ## 기술 스택 및 방법론  ### 사용 기술  • \*\*프로그래밍 언어 및 라이브러리\*\*: Python, PyTorch, OpenCV  • \*\*텍스트 감지/인식\*\*: PaddleOCR 커스터마이징, CRAFT, SVTR  • \*\*이미지 처리\*\*: OpenCV, scikit-image  • \*\*그래프 분석\*\*: NetworkX  • \*\*데이터 관리\*\*: PostgreSQL, pandas  • \*\*시스템 통합\*\*: Flask API, Docker  ### 방법론  • \*\*애자일 개발\*\*: 2주 단위 스프린트로 기능 개발 및 검증  • \*\*사용자 중심 설계\*\*: 엔지니어와의 지속적 피드백 루프 구축  • \*\*점진적 자동화\*\*: 수작업 → 반자동 → 완전 자동화 단계적 접근  ## 학습 및 향후 발전 방향  이 프로젝트를 통해 비정형 데이터에서 구조화된 정보를 추출하고 이를 비즈니스 가치로 전환하는 과정의 중요성을 체득했습니다  . 특히 도메인 지식과 머신러닝 기술의 효과적인 결합이 성공의 핵심 요소임을 확인했습니다.  향후 ALM 시스템에서도 이러한 접근법을 적용하여, 다양한 금융 문서에서 자동으로 정보를 추출하고 이를 고객 행동 예측 및 자  산부채 최적화에 활용하는 방안을 발전시키고자 합니다. 특히 문서 이해 기술과 시계열 예측 모델을 결합하여 더 정확하고 설명  가능한 예측 모델을 개발하는 데 기여하고 싶습니다.  ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━  이 프로젝트 경험은 토스뱅크 ALM 시스템에서 요구되는 데이터 모델링 및 시스템 최적화 역량을 갖추었음을 보여줍니다. 특히 비  정형 데이터에서 가치 있는 정보를 추출하고 이를 의사결정에 활용하는 능력은 ALM 시스템의 고객행동 모형 개발 및 BS/PL 최적  화에 직접적으로 기여할 수 있는 핵심 역량입니다. |
| 프로젝트 | 2. 약물 생성 모델 개발  1 ) 기간  - 2021.07-2022.01: 약물 생성 모델 개발  2) 배경 및 문제인식  - 기존 약물과 효과는 유사하면서, 약물의 흡수, 분배, 대사, 배설에 유리한 약물의 필요성  - 여러 약물로 이루어진 복합제의 어려움 개선(약물간 상호작용, 용량 조절, 특허 문제  3) 연구 목표 및 방법  - 복합제의 특징을 모두 가질 수 있는 약물 생성하여 약물 수 줄이기.  - Variational autoencoder를 활용하여 약물들의 latent vector 표현 및 두 vector 다양하게 interpolation하여 두 약물의 특징을 다양하게 가질 수 있는 약물들 생성.  - 신약 후보와 유사한 특징을 가지는 약물들 생성하여 부작용 예측 후 성공 가능성이 높은 약물 후보들 선별.  4) 연구 근거:  - 여러 약물로 구성된 복합제의 문제 및 특허 문제 해결.  - Variational autoencoder의 효율적인 약물 조합 기능 활용.  - 다양한 약물 생성 및 부작용 예측을 통한 후보군 선별로 비용 효율성 증대  5) 프로젝트 수행  - 혼자서 수행  - 생성 모델(Variational autoencoder)을 활용하여 화학적, 경제적으로 유효한 신약 후보 물질 생성.  - 화합물-독성 데이터를 활용하여 신약 후보 물질의 독성 여부 예측.  6) 프로젝트 결과  - Variational autoencoder를 활용하여 약물들의 latent vector 표현 및 두 약물의 특징을 가질 수 있는 다양한 약물들 생성.  - 신약 후보와 유사한 특징을 가지는 약물들을 생성 후, 약물의 '흡수, 분배, 대사, 배설, 독성'을 예측하고, 성공 가능성이 높은 약물 후보들 선별. |
| 프로젝트 | 3. 약물-약물 상호작용(부작용) 예측 시스템  1 ) 기간  - 2020.05-2020.08: 약물-약물 상호작용 예측 시스템  - 2020.09-2021.01: 외부 프로젝트 - 대원제약과 협력하여 신약과 기존 약물들 사이의 부작용 예측  2) 배경 및 문제인식  - 복합제 개발 시 약물들 사이의 예상치 못한 상호작용(부작용)이 발생할 수 있음.  - 신약 개발 및 투여 시 기존 복용 약물과의 예상치 못한 상호작용(부작용)이 발생할 수 있음.  - 약물-약물 상호작용이 없는 것으로 알려진 데이터 부족  - 약물-약물 상호작용 예측에 적합한 약물 features 탐색 필요.  3) 연구 목표 및 방법  - 부작용 심각도 및 상호작용 종류 예측 모델 개발.  - 기존 처방 약물들의 정보 활용하여 상호작용이 없는 데이터 생성.  - Multi-Task Learning 활용하여 과적합 방지 및 일반화 능력 향상.  - 약물 사이 구조 정보 유사성 비교 및 주요 상호작용 단백질 관련 feature  4) 연구 근거:  - 기존 동시 처방 약물들은 상호작용이 없을 가능성이 높음.  - 약물 상호작용과 관련한 주요 요소는 약물 구조 유사성 및 상호작용 단백질 유사성.  - Multi-Task Learning 을 통해 데이터 부족 문제를 보완하고 일반화 능력 향상.  5) 프로젝트 수행  - 팀원 1명과 함께 프로젝트 진행.  - 개발에 주도적인 역할 수행.  - 딥러닝(Multi-Task Learning)을 활용하여 약물-약물 사이의 부작용 데이터 학습 및 새로운 약과 기존 약물들과의 부작용 예측.  6) 프로젝트 결과  - 약물-약물 상호작용(부작용의 심각도 및 상호작용 종류) 예측 모델 개발.  - 기존 처방 약물들의 정보를 활용하여 상호작용이 없는 데이터 생산.  - Multi-Task Learning 모델을 활용한 알려지지 않은 약물 쌍의 상호작용 예측.  - 약물 사이 구조 정보 유사성 비교 및 주요 상호작용 단백질 관련 feature 활용을 통해 높은 예측 정확도 달성.  - 외부 프로젝트(대원제약과 협력)에서 신약과 기존 약물들 사이의 부작용 예측에 성공적으로 적용 |

|  |  |
| --- | --- |
| 회사명 | 닥터노아바이오텍㈜ |
| 재직기간 | 2020.01-2023.07(4년차) |
| 부서/직책 | AI 연구소 / 책임연구원 |
| 프로젝트 | 1. 약물에 대한 세포 반응: 이미지 세그멘테이션을 통한 염증 분석(PM)  1 ) 기간  - 2021.09-2021.12: 신경염증(Neuroinflammation)에 대한 이미지 분류 시스템 개발  - 2022.01-2022.04: 외부 프로젝트(지놈앤컴퍼니, 신경 염증에 대한 약효 평가) 수행  - 2022.11-2023.07: 약물에 대한 세포 반응 - 이미지 분할을 통한  2) 배경 및 문제인식  - 세포 염증에 대해 약물의 영향을 파악하는 것은 염증 완화와 관련된 신약 개발에 큰 역할을 할 수 있음.  - 약물의 염증 효과 확인에 많은 시간 및 비용이 소요  - 세포의 염증 유무에 따른 형태 변화 확인이 필요  - 이미지에서 세포의 밀도 및 다양성이 부족  3) 연구 목표 및 방법  - 약물이 세포 염증에 미치는 영향을 효율적으로 파악  - 이미지 분석을 활용하여 약물의 영향을 분석  - Instance Segmentation을 사용하여 세포의 염증 상태 및 형태적 특징을 분석  - 생성모델을 통해 세포 이미지 데이터 증강  4) 연구 근거:  - 염증(Neuroinflammation)은 신경세포 또는 신경 경로와 관련된 복잡한 원인을 가진 염증  - 세포(Microglia)는 신경 염증 유무에 따라 형태가 변함  - LPS(Lipopolysaccharides)는 세포에 염증을 유발하는 요소  - Diffusion-Model은 세포 이미지의 패턴과 밝기 분포를 잘 포착하고 재현  5) 프로젝트 수행  - PM으로서 전체 과정 기획 및 리드  - AI 팀 인원 2명, 실험팀 인원 3명과 함께 진행.  - 실험팀과 의견을 조율하면서 데이터 생성 퀄리티 높임.  - "이미지 분류 모델 -> 외부 과제 수행 -> 이미지 분할을 통한 이미지 분석"의 단계를 거치면서 더 고도화되고 정확도 높은 시스템 완성.  6) 프로젝트 결과  - 약물 농도에 따라 농도에 따라 세포의 형태적 특징 중요도가 변화.  - 약물 농도 증가에 따라 염증 없는 세포 비율이 증가 확인.  - 이미지 분석과 기술적 접근을 통해 약물의 염증 완화 효과를 성공적으로 평가  - 완성된 시스템은 약물 발견 및 개발에 효과적으로 활용  - 특허: 세포 이미지에 기반하여 약효를 평가하기 위한 방법 및 장치(제 10-2440373호) |
| 프로젝트 | 2. 약물 생성 모델 개발  1 ) 기간  - 2021.07-2022.01: 약물 생성 모델 개발  2) 배경 및 문제인식  - 기존 약물과 효과는 유사하면서, 약물의 흡수, 분배, 대사, 배설에 유리한 약물의 필요성  - 여러 약물로 이루어진 복합제의 어려움 개선(약물간 상호작용, 용량 조절, 특허 문제  3) 연구 목표 및 방법  - 복합제의 특징을 모두 가질 수 있는 약물 생성하여 약물 수 줄이기.  - Variational autoencoder를 활용하여 약물들의 latent vector 표현 및 두 vector 다양하게 interpolation하여 두 약물의 특징을 다양하게 가질 수 있는 약물들 생성.  - 신약 후보와 유사한 특징을 가지는 약물들 생성하여 부작용 예측 후 성공 가능성이 높은 약물 후보들 선별.  4) 연구 근거:  - 여러 약물로 구성된 복합제의 문제 및 특허 문제 해결.  - Variational autoencoder의 효율적인 약물 조합 기능 활용.  - 다양한 약물 생성 및 부작용 예측을 통한 후보군 선별로 비용 효율성 증대  5) 프로젝트 수행  - 혼자서 수행  - 생성 모델(Variational autoencoder)을 활용하여 화학적, 경제적으로 유효한 신약 후보 물질 생성.  - 화합물-독성 데이터를 활용하여 신약 후보 물질의 독성 여부 예측.  6) 프로젝트 결과  - Variational autoencoder를 활용하여 약물들의 latent vector 표현 및 두 약물의 특징을 가질 수 있는 다양한 약물들 생성.  - 신약 후보와 유사한 특징을 가지는 약물들을 생성 후, 약물의 '흡수, 분배, 대사, 배설, 독성'을 예측하고, 성공 가능성이 높은 약물 후보들 선별. |
| 프로젝트 | 3. 약물-약물 상호작용(부작용) 예측 시스템  1 ) 기간  - 2020.05-2020.08: 약물-약물 상호작용 예측 시스템  - 2020.09-2021.01: 외부 프로젝트 - 대원제약과 협력하여 신약과 기존 약물들 사이의 부작용 예측  2) 배경 및 문제인식  - 복합제 개발 시 약물들 사이의 예상치 못한 상호작용(부작용)이 발생할 수 있음.  - 신약 개발 및 투여 시 기존 복용 약물과의 예상치 못한 상호작용(부작용)이 발생할 수 있음.  - 약물-약물 상호작용이 없는 것으로 알려진 데이터 부족  - 약물-약물 상호작용 예측에 적합한 약물 features 탐색 필요.  3) 연구 목표 및 방법  - 부작용 심각도 및 상호작용 종류 예측 모델 개발.  - 기존 처방 약물들의 정보 활용하여 상호작용이 없는 데이터 생성.  - Multi-Task Learning 활용하여 과적합 방지 및 일반화 능력 향상.  - 약물 사이 구조 정보 유사성 비교 및 주요 상호작용 단백질 관련 feature  4) 연구 근거:  - 기존 동시 처방 약물들은 상호작용이 없을 가능성이 높음.  - 약물 상호작용과 관련한 주요 요소는 약물 구조 유사성 및 상호작용 단백질 유사성.  - Multi-Task Learning 을 통해 데이터 부족 문제를 보완하고 일반화 능력 향상.  5) 프로젝트 수행  - 팀원 1명과 함께 프로젝트 진행.  - 개발에 주도적인 역할 수행.  - 딥러닝(Multi-Task Learning)을 활용하여 약물-약물 사이의 부작용 데이터 학습 및 새로운 약과 기존 약물들과의 부작용 예측.  6) 프로젝트 결과  - 약물-약물 상호작용(부작용의 심각도 및 상호작용 종류) 예측 모델 개발.  - 기존 처방 약물들의 정보를 활용하여 상호작용이 없는 데이터 생산.  - Multi-Task Learning 모델을 활용한 알려지지 않은 약물 쌍의 상호작용 예측.  - 약물 사이 구조 정보 유사성 비교 및 주요 상호작용 단백질 관련 feature 활용을 통해 높은 예측 정확도 달성.  - 외부 프로젝트(대원제약과 협력)에서 신약과 기존 약물들 사이의 부작용 예측에 성공적으로 적용 |
| 프로젝트 | 4. 단백질-화합물 결합 예측 시스템 개발  1 ) 기간  - 2020.01-2020.09: 단백질-화합물 사이의 결합 예측  - 2021.02-2021.07: 외부 프로젝트(아모레퍼시픽, 특정 단백질에 대한 결합 억제제 예측)  - 2022.06-2022.12: 단백질-화합물 사이의 결합 예측 정확도 향상  2) 배경 및 문제인식  - 특정 질병은 관련 단백질의 활동으로 인해 발생  - 단백질과 결합하는 화합물(신약) 개발 필요성 인식  - 단백질과 화합물 사이의 결합 방식 이해 중요성 인식  3) 연구 목표 및 방법  - 단백질-화합물 결합 예측 시스템 개발  - 학습 데이터를 다양한 결합 포즈와 화합물로 증가(Data augmentation)  4) 연구 근거:  - Graph Attention Model을 활용하여 원자 사이의 관계를 효과적으로 학습  - 일반화 성능, 결과 해석 용이성, 단백질 및 화합물의 특정 원자나 구성 요소가 결합에 미치는 영향 파악 가능  5) 프로젝트 수행  - 팀원 1명과 협력하여 주도적인 역할 수행  - 결합 데이터 및 모델 활용하여 단백질-화합물 결합 친화도 학습 및 예측  - 상호작용 구체적 분석 및 시각화 수행  6) 프로젝트 결과  - 단백질-화합물 결합 데이터 기반 예측 모델 개발  - 상호작용 메커니즘 이해 향상 및 시각화를 통한 분석 가능  - 외부 프로젝트(아모레퍼시픽)에서 특정 단백질 결합 억제제 예측에 적용 |

|  |  |
| --- | --- |
| 대학원 | |
| 물리학과  석박통합수료  2012.3 – 2016.2 | 성균관대학교 물리학과 석박통합과정 수료(2012. 03 - 2016. 02)  - 고체이론연구실, 연구분야: 응집및통계물리-상온초전도체 이론  - 이론적인 모델들을 시뮬레이션하거나 실험 결과에 대한 최적화 수행  - 논문: Quantitative Determination of the Pairing Interactions for High Temperature Superconductivity in Cuprates. 공동저자. Science Advances 04 Mar 2016 |
| 초전도 상태에서는 전자들이 쌍을 이루는 것이 알려져 있다. 이렇게 전자들이 쌍을 이루어 움직이는 현상의 메커니즘이 어떻게 작동하는지를 정확하게 수치로 표현하는 것을 연구.  1. 데이터 수집: 초전도체 내의 전자 쌍의 움직임에 대한 실험 데이터 수집  2. 모델: 전자 쌍을 형성하는 역할을 할 수 있는 전자-원자진동의 상호작용을 표현하는 모델  3. 수학적표현: 모델은 복잡한 편미분 방적식으로 주어지고, 이 방정식을 풀기위해 수치해석적으로 접근  4. 수치해석: 베이지안 접근 방식과 비선형 최적화 방법으로 모델의 변수들을 계산.  5. 모델검증: 다양한 조건에서의 데이터를 사용하여 모델의 정확도를 검증.  6. 결과해석: 모델을 바탕으로 초전도체 내의 전자 쌍의 움직임에 대한 원리 파악  7. 활용기술: Python, C++, Mathematica, Origin, Linux, Quantum Mechanics, Mathematics and Statistics |

|  |  |
| --- | --- |
| 군대 | |
| 국방부 국직부대  연구소  2016.6 – 2019.5 | 공군 중위 전역. 국방부 신호분석 연구소에서 근무 (2016. 06 - 2019. 05)  - 역할: 장교로 복무하면서 데이터 관련 업무를 수행하였습니다.  - 성과: 디지털 신호 처리, 암호 해독, 궤적 예측 등 클러스터링, 최적화, 시뮬레이션 등의 작업을 수행했습니다. |
| 1. 데이터 전처리: 다양한 방법으로 암호화 및 부호화 된 신호 데이터 정리  2. 모델 및 예측: 데이터 클러스터링을 통한 분류 및 예측  3. 해독: 클러스터링 결과를 참고하여 복호화 및 암호 해독  4. 수치해석: 해독된 데이터에 대해 최적화 방법 활용  5. 모델검증: 다양한 데이터를 사용하여 모델 정확성 검증  6. 결과해석: 데이터 분석을 통한 인사이트 확보  7. 활용기술: Python, Matlab, Linux, Mathematics and Statistics |

자기소개서

|  |
| --- |
| 안녕하세요. 데이터 사이언티스트 배종주입니다.  저는 도전적인 과제에 흥미를 느낍니다. 문제에 대해 깊게 생각하고, 문제 해결을 위해 끈질기게 달라붙는 편입니다.  저는 닥터노아바이오텍의 AI 연구소에서 신약 개발을 위해 다양한 프로젝트를 수행했고, PM 으로서 성공적으로 프로젝트를 완료했습니다. 이 과정에서 딥러닝, 머신러닝 및 이미지 처리 등을 활용하여 바이오/제약 관련 데이터를 다루었고, 다른 팀과의 협업 및 소통을 통해 프로젝트에서 실질적인 성과를 이루어 냈습니다.  국방부 국직부대 연구소에서 장교로 복무하며 디지털 신호 분석, 암호 해독, 궤적 예측 등의 업무에서 신호처리, 클러스터링, 암호 알고리즘 분석, 시뮬레이션 및 최적화 등을 수행했습니다.  성균관대학교 물리학과 및 석박통합과정을 수료하였습니다. 학부/대학원 과정에서 배운 물리, 수학 및 수치해석 등의 프로그래밍 지식을 데이터 사이언스 분야에서 활용하고 있습니다. 또한, 다양한 분야의 데이터에 대해 이론적 및 경험적인 방법을 적용하여 문제를 해결하는 데에 능숙합니다.  이러한 제 역량과 경험을 바탕으로, 저는 두산에너빌리티 디지털솔루션팀의 AI 및 데이터 사이언티스트 업무에 있어서 가치를 창출할 수 있는 사람이라고 생각합니다. |

현재 연봉: 7,800 만원 (인센티브 및 성과급 없음), 희망 연봉: 8,000 만원

출근 가능일: 4월