홍길동

Date of birth
1998.01.01

Phone
+82 10-1234-5678

Email
demo@email.com

Address
43, Digital-ro 34-gil,
Guro-gu, Seoul,
Republic of Korea

Language
Korean

Link
linkedin.com

Experience

넥스트 플랫폼 (Next Platform) | AI Labs
머신러닝 엔지니어 (2022년 3월 – 현재)
LLM 기반 차세대 검색 서비스 모델 개발 (2023.01 ~ 현재)
MLOps 플랫폼 고도화 및 운영 (2022.03 ~ 2022.12)

주식회사 데이터리움 (Datarium Inc.)
소프트웨어 엔지니어, 머신러닝 (2020년 2월 – 2022년 2월)
콘텐츠 개인화 추천 시스템 구축 (2020.08 ~ 2022.02)
금융 이상 거래 탐지 시스템(FDS) 개발 (2020.02 ~ 2020.07)

Education

A 대학원 공학 석사 (졸업)
2018년 3월 – 2020년 2월

- **전공:** 인공지능 (Artificial Intelligence)
- **석사 논문:** 효율적인 자연어 이해를 위한 경량화 BERT 모델 연구 (A Study on Lightweight BERT Models for Efficient Natural Language Understanding)

B 대학교 컴퓨터공학부 공학 학사 (졸업)
2014년 3월 – 2018년 2월

- **전공:** 컴퓨터공학 (Computer Science)
- **졸업 프로젝트:** CNN 기반 의류 이미지 분류 및 검색 시스템 개발

ML Engineer

홍길동

Seoul, Korea
+82 10-1234.5678
demo@email.com

넥스트 플랫폼 (Next Platform) | AI Labs

(2022년 3월 – 현재)

1. LLM 기반 차세대 검색 서비스 모델 개발 및 서빙 최적화

(기간: 2023년 1월 ~ 현재)

프로젝트 개요:

기존 키워드 기반 검색 시스템의 한계를 극복하고, 사용자의 의도를 깊이 이해하는 시맨틱 검색(Semantic Search) 서비스를 구축하는 것을 목표로 했습니다. 특히 전문 분야의 문서에 대한 검색 정확도를 높이고, 대화형 검색 경험을 제공하는 것이 핵심 과제였습니다.

담당 역할 및 주요 활동:

도메인 특화 LLM 파인튜닝: 공개된 한국어 LLM(Ko-LLM)을 기반으로, 사내에 축적된 방대한 양의 전문 문서를 활용하여 Fine-tuning을 주도했습니다. 도메인에 맞는 고품질 학습 데이터셋을 구축하고, 모델이 특정 용어와 맥락을 정확하게 학습하도록 유도했습니다.

RAG(Retrieval-Augmented Generation) 파이프라인 설계: LLM의 고질적인 문제인 할루시네이션(Hallucination)을 최소화하고 정보의 신뢰도를 높이기 위해 RAG 파이프라인을 설계 및 구축했습니다. FAISS를 이용한 Vector DB에 문서를 임베딩하여 저장하고, 사용자 질문 시 가장 관련성 높은 문서를 먼저 검색한 후, 이를 컨텍스트로 LLM에 전달하여 답변을 생성하도록 했습니다.

Inference 파이프라인 최적화: 대규모 트래픽 환경에서도 안정적인 서빙이 가능하도록 모델 서빙 파이프라인을 최적화했습니다. PyTorch 모델을 ONNX로 변환하고, Triton Inference Server를 도입하여 GPU 활용률을 극대화하고 동적 배칭(Dynamic Batching)을 적용했습니다. 이를 통해 응답 속도와 처리량을 동시에 개선했습니다.

모니터링 및 성능 분석: Grafana 대시보드를 구축하여 모델의 응답 시간(Latency), 처리량(Throughput), 에러율뿐만 아니라, 답변의 품질(예: Relevance Score)을 지속적으로 트래킹하고 분석하여 개선 포인트를 도출했습니다.

기술 스택: PyTorch, Hugging Face Transformers, FAISS, Kubernetes, Docker, Triton Inference Server, Grafana

ML Engineer

홍길동
Seoul, Korea
+82 10-1234.5678
demo@email.com

성과 및 결과:

주요 도메인 검색 결과의 정확도 20% 향상 및 할루시네이션 현상 40% 감소

서빙 최적화를 통해 모델의 P95 Latency를 15% 단축하고 서빙 비용을 10% 절감

신규 검색 서비스의 성공적인 런칭에 핵심적으로 기여하여 사용자 만족도 조사에서 긍정적 평가 획득

2. MLOps 플랫폼 고도화 및 운영 자동화

(기간: 2022년 3월 ~ 2022년 12월)

프로젝트 개요:

당시 여러 팀에서 개별적으로 머신러닝 모델을 개발하고 배포하여 파이프라인이 파편화되어 있었고, 실험 관리 및 배포에 많은 수작업이 필요했습니다. 이를 해결하기 위해 모델 개발부터 배포, 모니터링까지의 전 과정을 표준화하고 자동화하는 MLOps 플랫폼을 고도화하는 프로젝트를 진행했습니다.

담당 역할 및 주요 활동:

CI/CT/CD 파이프라인 구축: Jenkins(CI), Airflow(CT), ArgoCD(CD)를 연동하여 코드 커밋 시 자동으로 모델 학습, 테스트, 배포가 이루어지는 파이프라인을 구축했습니다.

실험 관리 및 모델 레지스트리 도입: MLflow를 도입하여 모든 실험의 파라미터, 성능 지표, 아티팩트를 체계적으로 기록하고 관리했습니다. 또한, 모델 레지스트리를 통해 검증된 모델만 스테이징/프로덕션 환경으로 승격될 수 있는 워크플로우를 정립했습니다.

모니터링 시스템 강화: Prometheus와 Grafana를 이용해 서빙 중인 모델의 성능 저하(Drift), 데이터 분포 변화 등을 감지하는 모니터링 시스템을 구축하여 모델의 신뢰성을 확보했습니다.

기술 스택: MLflow, Kubernetes, Jenkins, Airflow, Docker, Prometheus, Grafana

성과 및 결과:

아이디어 구상부터 프로덕션 배포까지 걸리는 리드 타임을 평균 2주에서 3일로 80% 이상 단축

모델 배포 과정의 휴먼 에러를 제거하여 배포 안정성 대폭 향상

데이터 사이언티스트들이 인프라 작업 대신 모델링에 집중할 수 있는 환경을 제공하여 팀 전체의 생산성 증대

ML Engineer

홍길동

Seoul, Korea
+82 10-1234.5678
demo@email.com

주식회사 데이터리움 (Datarium Inc.)

(2020년 2월 – 2022년 2월)

1. 콘텐츠 개인화 추천 시스템 구축 및 고도화

(기간: 2020년 8월 ~ 2022년 2월)

프로젝트 개요:

모든 사용자에게 동일한 콘텐츠를 노출하던 기존 방식에서 벗어나, 사용자 개개인의 취향과 행동 패턴을 기반으로 맞춤형 콘텐츠를 추천해주는 시스템을 구축하여 서비스의 인게이지먼트를 높이는 것이 목표였습니다.

담당 역할 및 주요 활동:

데이터 파이프라인 설계: AWS S3에 저장된 대용량 사용자 행동 로그(클릭, 체류 시간 등)를 Spark을 이용해 주기적으로 가공하고, 모델 학습에 필요한 피처를 생성하는 배치 파이프라인을 Airflow로 구축했습니다.

추천 모델 개발 및 서빙: 초기에는 Spark MLlib의 ALS(Alternating Least Squares)를 이용한 협업 필터링 모델을 적용하여 빠르게 시스템을 구축했습니다. 이후, 사용자 및 아이템의 메타데이터를 함께 활용하기 위해 TensorFlow로 Wide & Deep 모델을 구현하여 추천 성능을 고도화했습니다. 최종 추천 결과는 빠른 조회를 위해 Redis에 저장하고 API를 통해 제공했습니다.

A/B 테스트 및 성능 평가: 새로운 추천 로직의 효과를 검증하기 위해 A/B 테스트 프레임워크를 설계하고, CTR, 전환율 등의 핵심 지표를 기반으로 모델 성능을 정량적으로 평가하며 지속적으로 개선했습니다.

기술 스택: TensorFlow, Spark, Scikit-learn, Airflow, Redis, AWS S3, PostgreSQL

성과 및 결과:

추천 시스템이 적용된 영역의 CTR(클릭률) 10% 상승

개인화 추천을 통한 사용자 1인당 콘텐츠 소비량 15% 증가 및 매출 5% 증대에 직접적으로 기여