황우성

(Woosung Hwang)
ML Engineer

카카오에서 ML 엔지니어로 근무하고 있습니다. 콘텐츠, 장소, 광고 추천부터 요약, QA 등 LLM 기반 서비스까지 다양한 도메인에서의 경험을 쌓았습니다. 최근에는 LLM 성능 최적화와 서비스 적용을 통한 가치 창출에 집중하고 있습니다. 이를 위해 데이터 중심의 최적화 방안을 탐구하고, 최신 기술 동향을 빠르게 적용하여 모델 성능을 지속적으로 개선해가고 있습니다.

이전에는 다음 검색의 키워드 광고 추천 시스템에서 CTR 예측 모델과 키워드 확장 매칭 로직을 적용하여 매출 증대에 직접적으로 기여하고, 광고 효율을 개선한 경험이 있습니다. 또한 실시간 데이터 기반 콘텐츠 추천에 Bandit 알고리즘을 적용하여 CTR 등 주요 개인화 추천 지표를 향상시킨 경험이 있습니다.

저는 새로운 기술에 도전하는 것과, 이론을 넘어 실제 서비스에 적용하여 측정 가능한 성과를 내는 것을 선호합니다. 특히 빠르게 변화하는 최신 연구 동향과 알고리즘을 신속하게 학습하여 실제 문제 해결에 적용합니다. 이 과정에서 얻은 데이터 기반 인사이트로 모델을 꾸준히 개선하며 실질적인 가치를 만들어내는 것을 중요하게 생각합니다.

Contact: zzong2006@gmail.com
HomePage: https://zzong2006.github.io/

Career

Kakao / Machine Learning Engineer

21.04 - 25.05 (4y 1m)

LLM Alignment Team (2023. 10 ~ 2025. 05)

- QA, 요약, 장소 추천 서비스용 언어 모델 개발

추천팀 (Recommendation Team) (2021. 04 ~ 2023. 10)

- Bandit 기반 개인화 추천 및 키워드 광고 추천 시스템 개발

Kakao / Intern (추천팀)

2020.12 - 2021.02 (3m)

- 피코마(piccoma) 서비스 웹툰 추천 알고리즘 고도화 및 성능 개선

Skill

Programming: Python

LLM & Generative AI:
- Model Development & Fine-Tuning: SFT, DPO, GRPO, RLHF, Multi-node Distributed Training
- Application & Prompt Engineering: Prompt Engineering, Function Calling, LLM Workflow Design, RAG
- Data & Evaluation: Synthetic Data Generation, LLM-based Evaluation

Machine Learning & Recommendation:
- Algorithms: Bandit (Contextual UCB, Thompson Sampling), CTR Prediction (FM, DNN, LR)
- Search & Retrieval: ANN, Hybrid Search (Vector Search + BM25), Embedding Model Fine-Tuning

Backend & MLOps:
- Data Systems: Kafka, PySpark, SQLite, MongoDB, RocksDB
- Infrastructure & Deployment: Kubernetes, Airflow, vLLM

Projects

Kakao / Machine Learning Engineer

21.04 - 25.06 (4y 2m)

1. 장소 추천 서비스 모델 개발 (2025. 01. ~ 진행중)

배경

카카오맵 사용자에게 개인화된 장소 추천 경험을 제공하고자 LLM 기반의 대화형 장소 추천 모델 개발을 진행했습니다. 자체 LLM 개발을 통해 관련 서비스에 최적화된 성능 및 빠른 응답 속도 확보를 목표했습니다. 개발된 LLM은 사용자의 검색 의도를 이해하고 내부 시스템과 연동하여 최적의 장소를 추천하는 핵심 역할을 수행합니다.

역할: LLM 모델 개발

자체 LLM 기반 추천 로직 설계 및 구현:

- 사용자의 질의를 분석하여 최적의 API (tool) 를 선택하고 호출하는 핵심 로직(Function Call) 성능 최적화를 위해 파인튜닝 수행

경량 Re-Rank 모델 개발:

- 검색된 장소 목록을 사용자의 선호도 및 대화 맥락을 고려하여 가장 만족도가 높을 순서로 Rerank 하는 모델을 GRPO 학습 전략을 적용하여 자체 개발하고 지속적으로 성능을 개선

대형 모델 대상 GRPO 학습 전략 탐색 및 학습 효율 비교 분석:

- 기존 Re-Rank 모델의 학습 경험을 바탕으로, Agent 기반의 향후 타 서비스들의 통합 시나리오를 대비하여 수백 B (Billion) parameter 규모의 대형 모델에 GRPO 학습 적용 가능성을 탐색.
- Verl, Axolotl 프레임워크를 활용해 GRPO 학습을 시도하고, Multi-node 환경에서 학습 속도 비교 분석 수행.

모델 학습용 고품질 데이터셋 구축

- 실제 사용자의 다양한 검색 패턴과 의도를 반영한 합성 데이터를 생성하고, 기존 데이터를 정제하여 자체 LLM의 장소 추천 정확도 및 API 호출(Tool Call) 성능을 향상시키는 데 기여

성과

자체 개발 LLM의 경쟁력 입증

- 32B 파라미터 규모의 자체 개발된 LLM이, API 호출(Tool Call) 정확도 측면에서 GPT-4o 대비 우수한 성능을 달성하여, 자체 모델 확보의 기술적 타당성과 잠재력을 확인

효율적인 경량 모델의 성능 및 속도 확보

- 12B 파라미터의 상대적으로 가벼운 자체 모델로도 GPT-4.1 API 수준의 우수한 Rerank 성능(F1-score 기준)을 달성
- 추론 속도 면에서 32B 모델 대비 약 3배, GPT-4.1 API 대비 약 2.5배 빠른 성능을 보여, 대규모 트래픽 환경에서의 효율적인 서비스 운영 가능성을 입증

대형 모델 학습 효율성 입증

- Multi-node 환경에서 Slurm을 통해 Megatron-LM으로 학습 시, DeepSpeed (ZeRO-3) 대비 일반 모델에서 최소 2배, MoE(Mixture of Experts) 환경에서는 최대 20배의 학습 속도 향상을 확인

2. LLM Router 모델 개발 (2024. 09. ~ 2024. 12.)

배경

대규모 LLM 기반 챗봇 서비스의 운영 비용 절감을 위해, 사용자 쿼리를 최적 모델로 라우팅하는 시스템 개발을 주도했습니다. GPT-4o 와 경량 모델들을 효과적으로 조합하여 서비스 품질 유지와 비용 최적화를 달성했습니다.

역할: 개발 리드

테크니컬 리드로서 시스템 설계부터 성능 최적화까지 전체 개발 과정을 주도했습니다.

설계 및 방향성 설정

- AI Mate, 카나나 등 주요 서비스 적용 시나리오 정의
- 라우터 시스템 아키텍처 설계 및 개발 로드맵 수립
- 4인 개발팀 리드 및 스프린트 운영

데이터 및 모델 개발

- GPT-4o-mini 기반 프로토타입 개발 및 Prompt Engineering
- 분류 모델 기반 라우터 개발 및 성능 최적화
- Instruction Evolution & Magpie 전략 활용한 라우터 튜닝 데이터 생성
- 라우터 성능 평가 지표 설계 및 파이프라인 구축
- Prompt Compression 기술 도입으로 라우팅 효율 개선

성과

비용 효율화 및 서비스 품질 유지

- LLM 운영 비용 60% 절감 (GPT-4 호출 비율 67%로 감소)
- 기존 GPT-4o 대비 99% 성능 유지

빠른 개발 속도

- 2개월 동안 라우터 모델의 설계, 구현, 최적화를 진행하며 실질적인 비용 절감 가능성 입증

2024 If Kakao CEO Keynote 세션에서 '모델 오케스트레이션' 사례로 소개되며 기술적 성과와 잠재력을 인정받음

3. 프로야구봇 요약 서비스 개발 (2024. 03. ~ 2024. 09.)

배경

Daum에서 운영하는 카카오톡 프로야구봇 채널(카카오톡채널 - 프로야구봇)은 KBO 경기 요약 및 실시간 요약 서비스를 제공하기 위해 고도화된 LLM 기반 요약 모델이 필요했습니다.

이 프로젝트는 단순 요약을 넘어서, 200자 이내의 bullet point 형식 등 특정 형식을 준수하면서도, 특정 팀 관점이나 인터넷 커뮤니티 스타일 등 사용자 맞춤형 톤을 구현해야 하는 고도화된 요구사항이 있었습니다.

기존의 LLM 기반 버티컬 모델 개발 경험을 토대로, 경기 요약, 실시간 환호 분석, 한줄 요약 등 세 가지 요약 서비스를 성공적으로 개발 및 런칭하여 채널 활성화와 사용자 경험 개선에 기여했습니다.

역할: 개발 리드

경기 요약 서비스

- 2-Stage 경량 오픈소스 모델 구조 설계
- 합성 데이터 생성 및 레이블링 팀과 협업해 고품질 학습 데이터 확보 (모델 학습 (SFT / DPO) 목적)
- 평가 자동화를 위해 airflow 기반 파이프라인 제안 및 구축
- vLLM 서빙을 통한 안정적 응답 제공 및 중국어 응답 방지 패치 제안
- Self-rewarding modeling 및 Mergekit 을 활용한 성능 고도화 시도

환호 분석 서비스 (실시간 요약)

- 프롬프트 엔지니어링(PE)을 통한 GPT 모델 기반의 요약 모델 개발
- 반복적인 프롬프트 개선을 위한 평가 파이프라인 구축
- 서비스 기획팀과 협업해 환호 포인트 기준 및 요약 품질 조율

한줄 요약 서비스

- 인터넷 커뮤니티 스타일 구현을 위해 크롤링 및 합성 데이터 수집

성과

모델 개발 성과

- KBO 경기 요약: 휴먼 피드백 기준 GPT-4-Turbo 대비 61% 선호율 기록

채널 성과 (2023년 대비)

- 채널 친구 수 40% 증가 (+10만 이상)
- 월 평균 활성 유저 수 52% 증가 (+7만 이상)

카카오 본사 내 오픈소스 LLM을 바탕으로 진행한 첫번째 실서비스 런칭 사례

4. QA LLM 모델 개발 (2023. 06. ~ 2024. 01.)

배경

카카오 전사 단위 과제로 진행된 QA(Question Answering) 챗봇 개발 프로젝트는 오픈소스 LLM을 활용해 대규모 사용자 트래픽을 처리할 수 있는

모델을 설계하고 구현하는 데 중점을 두었습니다. 이 챗봇은 오픈채팅 서비스에 적용되어 제공될 예정이였기 때문에, 낮은 운영 비용을 유지하면서도 정확하고 유용한 응답을 제공하는 고품질 AI 모델 개발이 요구되었습니다.

역할

QA 모델 개발 및 고도화

- 오픈소스 LLM(Llama 2)을 활용한 QA 모델 설계 및 구현
- RAG(Retrieval-Augmented Generation) 시스템을 적용하여 성능 향상 및 할루시네이션 문제 해결

RAG 시스템 고도화

- BM25와 Vector Search를 융합한 Hybrid Search 설계
- 도메인 데이터를 활용한 임베딩 모델 파인튜닝
- RAG 평가를 위한 RAGAS 지표 활용
- 데이터 청킹 전략 비교(Paragraph vs. Sliding Window)

데이터 수집

- 고품질 데이터 확보를 위해 휴먼 레이블링 프로젝트 리드
- 레이블링 가이드라인 설계 및 데이터 수집 프로세스(툴) 구축
- 의료 및 축구 등 특정 도메인 데이터 기반 QA 모델 구축

평가 및 개선

- QA 성능 평가를 위한 Partial Matching, Rouge 등 지표 구현
- TruthfulQA 지표를 응용해 할루시네이션 개선 효과 측정

성과

LLM QA 품질 입증

- 의료 도메인에서 Llama2-7b로 개발한 경량 QA 모델이 GPT-3.5-turbo와 동급 이상의 품질(65% 비율)을 달성
- 휴먼 레이블링 데이터로 학습된 모델이 합성 데이터 기반 모델 대비 QA 성능(Rouge)에서 +30% 이상 개선
- DPO 학습을 통해 모델 응답의 할루시네이션 감소 효과를 입증

RAG 성능 개선 입증

- 소량의 도메인 데이터를 활용한 파인튜닝으로 Retrieval 성능(RAGAS) 최대 +30% 향상
- Hybrid Search는 구성 요소 간의 단점을 보완해 안정적인 성능을 제공한다는 점을 확인
- 데이터 청킹에서는 Window 방식이 일반적으로 더 일관된 성능을 제공함을 확인

5. 키워드 광고 추천 모델 개발 (2022. 07. ~ 2023. 03.)

배경

다음 검색에 노출되는 키워드 광고의 매출 향상을 목표로, 기존 광고 랭킹 로직과 매칭 알고리즘을 개선하는 프로젝트를 진행했습니다.

기존 시스템은 입찰가 중심으로 광고 순위를 결정하고, 일부 검색 쿼리에서만 광고가 노출되는 한계가 있었습니다.

이에 따라 광고 기대 수익 예측 모델과 매칭 로직을 개선하고, 데이터 기반 의사결정 프로세스를 통해 매출 성장을 도모했습니다.

역할

광고 기대 수익 예측 모델 개선

- Batch 환경에서 PySpark와 Airflow를 활용하여 광고 CTR을 계산하고, 이를 기반으로 기존 입찰가 중심의 광고 노출 순위를 기대 수익 기준으로 최적화하는 로직을 설계 및 구현
- 오프라인 실험에서는 LR, DNN, FM, GBRT 등 다양한 모델의 성능을 비교 분석

로직 커버리지 확대

- 개선된 로직이 전체 트래픽에 적용될 수 있도록, 로직 적용에 따른 광고 노출 순위 변화 분석 결과를 제공하여 설득.
- 광고주 및 내부 부서(대행사영업팀, 운영파트 등)와 협력하여 기대 수익 기반 로직의 단계적 배포 계획 수립.

확장 매칭 로직 개선

- Word2vec, BERT 임베딩을 활용해 검색 쿼리와 광고 간의 매칭 커버리지 개선.
- 분산 ANN(Approximate Nearest Neighbor) 서버 설계로 1,000만 단위 광고의 실시간 유사도 검색 구현
- 비즈니스 로직 대응: 쿼리별 매칭 제어가 가능한 운영툴을 개발

성과

매출 증대

- 연간 +7억 원의 추가 매출 달성 (개선된 광고 노출 로직 기여분)
- 로직 커버리지 확대로 +28억 원까지 매출 증가
- 온라인 A/B 테스트에서 광고 CTR +4% 상승 확인
- 오프라인 실험에서 FM 모델로 기존 대비 AUC +12% 성능 향상 확인

광고 커버리지 확대

- 광고 노출 커버리지 2배 가까이 확대 (35% → 64%, 오프라인 실험)

운영 프로세스 혁신

- 품질지수 기반 광고 노출 기준을 새롭게 수립하는데 기여

6. Bandit 기반 개인화 추천 시스템 개발 및 고도화 (2021. 05. ~ 2023. 03.)

배경

카카오톡 뷰탭(현 오픈채팅 탭), 피코마(piccoma) 등 다양한 서비스에서 사용자 맞춤형 추천 기술의 중요성이 커짐에 따라, Bandit 모델을 활용하여 개인화 추천 성능을 고도화하는 프로젝트를 진행했습니다.

CTR, 사용자 전환율, 체류시간 등 주요 KPI를 향상시키고 사용자 만족도를 개선하는 것을 목표로 삼았습니다.

역할

Bandit 기반 개인화 추천 시스템 적용

- 시간당 약 백만 단위 트래픽을 안정적으로 처리하는 Kafka 기반 실시간 데이터 파이프라인을 설계 및 구축하여 사용자 행동 데이터를 학습 데이터로 활용
- Topic Modeling 기반 사용자 Context 데이터를 regression 기반 Bandit 모델의 feature로 활용.

프레임워크 최적화

- Sanic 기반 서빙 API 및 학습 모듈 리팩토링을 통해 시스템 자원 효율성 개선
- 학습 데이터 관리 방식을 RocksDB에서 SqliteDB로 변경하여 I/O 부하를 감소시키고 개발 생산성 향상.

알고리즘 고도화

- 사용자 선호도가 변화하기 쉬운 도메인에 모델 파라미터 감쇠(Decay) 옵션 구현
- 추천 가능한 아이템 수가 변동이 심한 도메인에 Empirical Draw 옵션을 구현하여 추천 결과의 다양성과 효율성 조율.
- 신규 사용자 대상 Cold-Start 추천을 위해 성별/연령 기반 앙상블 로직 도입.

운영 및 배포

- Kubernetes 기반 클러스터 환경에서 추천 서비스를 안정적으로 배포 및 운영

성과

사용자 지표 향상

- 카카오톡 뷰탭 기준 클릭수 +16.49% (+500만 이상), 구독수 +30.62% (+8천 회), 사용자 체류시간 +13.71% 증가.
- Cold-Start 추천 로직 도입으로 신규 사용자 전환율 +3% 개선.
- 다양한 도메인에서 CTR 최대 +5.27% 개선 (e.g., 뉴스, 연예 섹션).

시스템 효율화

- 서빙 API 및 학습 모듈 리팩토링으로 I/O 부하 및 메모리 사용량 감소.
- 시스템 효율화가 모델 업데이트 속도 향상으로 이어져 카카오톡 뷰탭 서비스 기준 CTR +2%, 사용자 수 +3.4% (+1만 이상) 증가.

개발 생산성 향상

- 추천 알고리즘 모듈화 및 정책(policy) 독립 설계로 신규 알고리즘 실험 속도 개선에 기여.

Awards

(2024. 03) DACON 국내 LLM 경진대회 - 3위 수상

- 도배 하자 질의 응답 처리 : 한솔데코 시즌2 AI 경진대회
- 참가자 500명 이상, 비공개 리더보드 1위 달성

(2023. 07) 카카오 사내 AI 해커톤 '2023 24K' - 1위 수상

- 카카오, 사내 해커톤 '2023 24K' 진행
- 팀 리드로 참여, 총 참가자 300명 이상
- Stable Diffusion 을 활용한 카카오톡 이모티콘 생성 서비스 제안

Presentation (사내 세미나 및 기술 공유 활동)

1. Stable Diffusion 기반 이모티콘 생성 모델 개발 사례 발표 (2023)

Diffusion 모델을 활용해 카카오톡 IP 기반 이모티콘 생성 프로세스를 혁신한 사례를 공유하며 창의적인 AI 활용 방안을 제시.

2. 고품질 데이터로 LLM 질의응답(QA) 성능 극대화 (2023)

약 500명의 임직원을 대상으로 진행된 세미나에서 고품질 데이터셋의 중요성과 이를 통한 LLM 의 질의응답 성능 향상 전략을 소개.

3. Stable Diffusion 모델을 활용한 디자인 어시스턴트 개발 사례 발표 (2024)

약 120명의 참석자를 대상으로 디자인 업무를 지원하는 AI 어시스턴트 개발 과정과 결과를 공유하며 실제 업무 활용 가능성을 제안.

4. LLM 파인튜닝 및 플랫폼 활용 강의 (2024)

전사 16개 조직의 대표 인원들을 대상으로 LLM 파인튜닝 기술과 관련 플랫폼 활용 방법을 교육.

대표 인원들이 실무에 적용 가능한 지식을 습득하고 각 조직으로 전파할 수 있도록 맞춤형 강의를 기획 및 진행.

Kakao / Intern

2020.12 - 2021.02 (3m)

피코마 웹툰 추천 알고리즘 고도화

Random Walk with Restart (RWR) 기반의 TPA 로직 개선

- Alias Method, 병렬 처리(numba, cupy 활용), 그래프 가지치기 등 최적화 기법 적용
- 응답 속도 2배 이상 향상 및 메모리 사용량 50% 절감 달성
- (참고) TPA: https://arxiv.org/pdf/1708.02574

개선된 알고리즘 실 서비스 배포 후 CTR(클릭률) 향상 확인

테크 블로그를 통한 인턴 경험 공유

charlie의 추천팀 인턴 생활기 - tech.kakao.com

—

Education

Yonsei University (Wonju Campus) / M.S., Computer Science

2019.03 - 2020.08 (1y 5m)

주요 연구 분야: 데이터베이스

- 시퀀스 데이터 간 유사도 측정 시간을 단축시키기 위해, 특정 도메인의 특성을 활용한 유사 시퀀스 매칭 방법 연구

부가 연구 분야: Telecommunication

- MIMO (Multiple-Input Multiple-Output) 시스템에서의 optimal precoder selection 알고리즘 연구

Toronto University (Ontario) / Exchange Student

2016.08 - 2017.05 (9m)

Yonsei University (Wonju Campus) / B.S., Computer Engineering

2011.03 - 2013.04 / 2015.04 - 2019.02 (6y)

- 최우등졸업 (평량평균 4.0 이상 / 4.3 만점, 단과대 및 학년별 상위 1% 이내)
- 데이터베이스 연구실 인턴 경험: 시계열 데이터를 활용한 차원 축소 기법 효율성에 대한 연구

Papers

- 황우성, 임효상. (2021). 단조 증가 성질을 지닌 데이터 도메인에서의 집합 유사 시퀀스 매칭 방법. 정보과학회.
- 황우성. (2020). "색인 구조를 활용한 효율적인 데이터 시퀀스 매칭 기법 연구." 국내석사학위논문 연세대학교 대학원
- W. Hwang, J. Park, J. Kim and H. -S. Lim, (2020) "Optimal Precoder Selection for Spatially Multiplexed Multiple-Input Multiple-Output Systems With Maximum Likelihood Detection: Exploiting the Concept of Sphere Decoding," IEEE Access.
- 황우성, 임효상. (2020). 단조 증가 성질을 지닌 데이터 도메인에서의 집합 유사 시퀀스 매칭 방법. 한국정보과학회 학술발표논문집.
- Hwang, W. and Lim, H.-S. (2018) “Clustering Performance Analysis for Time Series Data: Wavelet vs. Autoencoder,” 한국정보처리학회