장승호

대한민국 | (+82) 10-3550-5156 | shjang1992@hotmail.com | https://github.com/skyser2003

보유 기술

프로그래밍 언어: Python, TypeScript, Go, Rust, C++
LLM 추론 최적화: vLLM, TensorRT-LLM
오케스트레이션 / 워크플로우: Kubernetes, Docker, GCP

경력

소프트웨어 엔지니어, MLOps / AI 엔지니어 엔씨소프트 – 대한민국 2017 년 9 월 – 2024 년 5 월

- **LLM 최적화 및 배포** (2023 년 6 월 – 2024 년 5 월)
  - TensorRT-LLM/vLLM 을 활용하여 Transformer 모델 추론 최적화하여 서빙 성능 극대화
  - 모델 양자화 및 프리픽스 캐싱과 같은 최적화 기술을 적용
  - 스트레스 테스트 세부 조정 및 자동화를 통해 처리량과 시스템 안정성을 강화
- **마이크로서비스 마이그레이션 및 Helm 통합** (2022 년 6 월 – 2022 년 12 월)
  - 기존 Kustomization 기반 Kubernetes 배포를 Helm 으로 마이그레이션하여 다중 환경 구성을 모듈식 템플릿으로 통합
  - ArgoCD 를 사용하여 CI/CD 워크플로우를 자동화하고, 스테이징 및 프로덕션 클러스터 간의 배포 복잡성을 간소화
- **머신러닝 모델 (CPU) 서빙 및 최적화** (2021 년 6 월 – 2024 년 5 월)
  - Tritonserver 를 사용하여 배치 및 스트리밍 추론 요청을 위한 CPU 기반 머신러닝 모델 서빙 인프라를 구축
  - Tritonserver 를 대체하기 위해 Rust/Go 로 batch 추론 서버를 개발함으로써 확장 가능한 서비스 인프라를 설계하고, 지연 시간 및 안정성 문제를 해결하여 상당한 성능 향상을 달성
  - 서버 튜닝과 프로파일링을 수행하여 ML 서빙 파이프라인의 효율성을 극대화
- **머신러닝 플랫폼 개발** (2019 년 6 월 – 2021 년 5 월)
  - Jupyter Notebook 통합과 GPU 자원 할당을 지원하는 Kaggle 유사 ML 학습 및 평가 플랫폼을 설계하고 구현
  - Kubernetes 및 Elasticsearch 를 활용하여 실시간 모니터링, 로그 집계, 하이퍼파라미터 튜닝 기능을 갖춘 포괄적인 ML 작업 관리 플랫폼을 구축
- **GraphDB 애플리케이션 개발** (2019 년 3 월 – 2019 년 5 월)
  - Neo4j 를 활용해 실시간 야구 경기 통계 데이터를 동적으로 업데이트하는 GraphDB 애플리케이션을 개발
  - 노드 및 관계 쿼리 구조를 재설계하여 데이터베이스 성능을 개선하고, 업데이트 및 검색 속도를 향상
- **웹 서버 개발** (2017 년 9 월 – 2018 년 8 월)
  - Node.js 및 TypeScript 를 사용하여 AI 기반 야구 애플리케이션 Paige 의 백엔드를 개발
  - FCM 기반 알림 서비스를 구현
  - Jenkins 와 Docker 를 활용한 CI/CD 파이프라인을 자동화

게임 개발자 IMC 게임즈 – 대한민국 2012 년 9 월 – 2015 년 8 월

- **스피릿위시 (모바일 MMORPG)** (2015 년 3 월 – 2015 년 8 월)
  - C# 및 Protobuf 를 활용하여 월드 및 채널 시스템을 포함한 서버 아키텍처를 개발
  - 커스텀 더미 클라이언트를 이용한 스트레스 테스트 설계 및 실행을 통해 서버 안정성과 확장성을 확보
- **그라나도 에스파다 (PC MMORPG)** (2012 년 10 월 – 2015 년 2 월)
  - 멀티스레드 프로그래밍을 통해 XML 파싱 효율성을 개선하고 로딩 시간을 단축
  - C++ 템플릿을 사용하여 실시간 XML 리로드 시스템 및 자동 패킷 관리 프레임워크를 구현
  - ELK 를 이용하여 모니터링을 지원하는 중앙 집중 로그 시스템을 구축

학력

연세대학교 – 컴퓨터과학 학사, 심리학 부전공 2010 년 3 월 – 2017 년 2 월

언어

- **한국어:** 모국어
- **영어:** 유창 (TOEFL iBT 111, 2009 년 3 월 취득)
- **프랑스어:** 중급 (DELF B2, 2007 년 6 월 취득)

기타 개인 프로젝트

- **C++ 템플릿 라이브러리** (GitHub: FTL): C# 스타일의 속성, Unity 스타일의 벡터 클래스 및 Enum 열거자를 C++ 템플릿 라이브러리로 구현.
- **개인 딥러닝 프로젝트** (GitHub: ML): InfoGAN, DCGAN 및 개선된 WGAN 을 사용하여 픽셀 스타일의 게임 캐릭터 이미지를 생성.
- **넥슨 개발자 컨퍼런스 2014 발표** (NDC 2014): 초급 프로그래머도 유지 관리하기 쉬운 프로젝트 생성에 대한 발표 진행.
- **홀로렌즈 데모 프로젝트** (GitHub: Artan): 피아노 레슨 프로그램 및 3D 포트리스 게임을 구현.
- **개인 슬랙 봇** (GitHub: Ditto Bot Rust): 웹 페이지 파싱 및 미리보기 생성, 사용자별 채팅 카운트 추적, ChatGPT API 통합을 Rust 로 구현.
- **몬스터헌터 스킬 계산기** (GitHub: MHR Skill Calculator): Vue3 및 Rust 로 개발된 몬스터헌터: Sunbreak 게임의 최적 방어구 및 스킬 조합 계산 도구.