16 생성형 모델을 활용한 AI 헤어스타일러

소속 정보컴퓨터공학부 분과 A 팀명 AICasso 참여학생 박시형, 한지훈, 홍진욱 지도교수 전상률

연구 소개

연구 개요

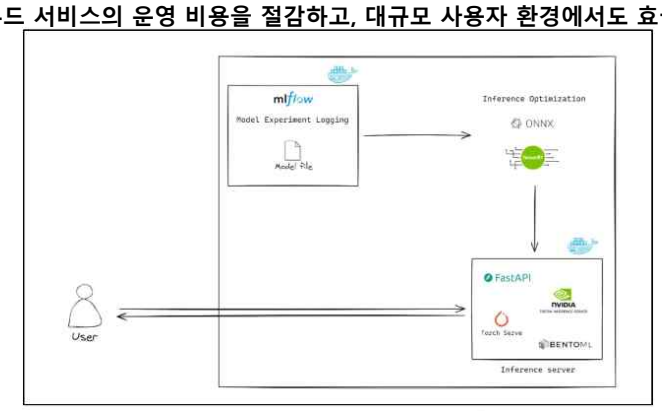
- 연구 목표: 생성형 AI 모델을 활용하여 실시간 헤어스타일 변환 서비스를 제공하는 것을 목표로 했으나, 기술적 한계로 인해 이미지 기반 비실시간 서비스를 개발한다.

- 주요 연구 내용: FastAPI, Streamlit을 활용하여 실시간으로 변환된 이미지를 사용자에게 제공하고, AI 모델의 경량화 및 비용 효율화에 대한 연구를 진행한다.

연구 방법

- 모델 서빙 최적화: FastAPI 및 Streamlit을 사용하여 사용자 요청에 빠르게 응답하는 서비스 구조를 설계하고, GPU 메모리에 모델을 사전 로딩하여 서비스 지연을 최소화한다.

- AI 서비스 비용 효율화: 경량화된 모델을 통해 클라우드 서비스의 운영 비용을 절감하고, 대규모 사용자 환경에서도 효율적인 서비스 운영을 가능하게 한다.



모델 서빙 파이프라인

서비스 개발

서비스 개요

- 생성형 모델인 GAN 기반의 헤어스타일 모델 (HairFastGAN)을 이용하여 웹 서비스 개발.

- 추론 시간 및 서비스 서빙 시간 단축을 위해 FastAPI, Streamlit 등의 프레임워크를 활용한다.

- Face Detection, Crop, Rotation, Alignment 등의 전처리 기술을 적용하여, 추론 및 생성 성능을 개선한다.

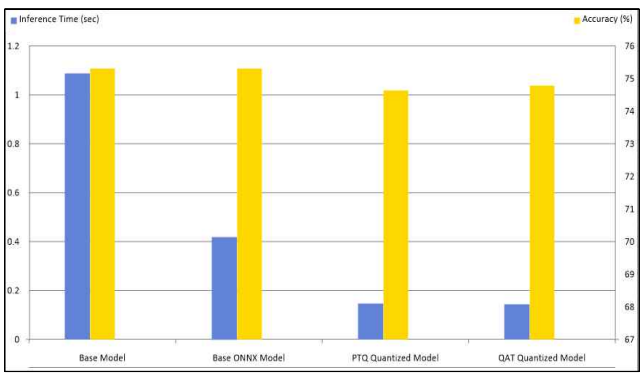
전처리 이전 결과 전처리 이후 결과

경량화 방안

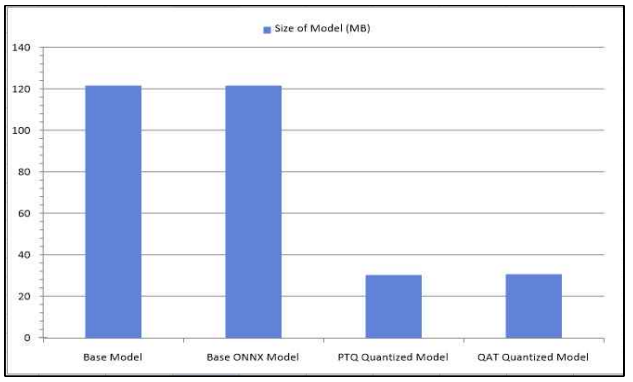
성능 평가 및 비교

- 성능 평가: ONNX 모델이 PyTorch 모델보다 더 높은 처리 성능(RPS)을 보였으며, 추론 속도(응답 시간) 또한 크게 개선되었다.

- 경량화 기법 성능 비교: PTQ는 간단한 구현으로 빠른 경량화가 가능하지만, QAT는 더 높은 정확도와 성능을 제공하며, 실시간 서비스에 적합한 결과를 도출하였다.



추론 시간 및 성능 비교



모델 크기 비교