소이넷_2020311595_박제현_1차_과제

0. 기업정보

- **2018년 09월 27일** 설립 (설립 6년차)
- 비상장 기업
- 연매출 2023년 3월 기준 2억 9,706만원
- 중소기업 형태 15명 사원
- 투자 유치 (누적 투자 금액 : **15억원**이상)
  - **엔솔파트너스 2019-06** seed 단계 투자 유치 **3억원**
  - **경기창조경제혁신센터 2021-01** seed 단계 투자 유치
  - **TIPS 2021-04 5억원**
  - **IBK기업은행, 플랜에이치벤처스, 서일이앤엠 2021-07 10억원**

    . . .

1. 기업 대표 기술

- 소프트웨어 기반 AI 실행 **가속 솔루션**
- AI 추론 최적화 엔진

소이네이처

- **AI 머신러닝옵스(MLOps) : 초고속 초경량 추론엔진**
  - **데이터 라벨링**과 **학습자동화 기능** 제공
  - **추론 최적화 엔진**
  - 인공지능 **정확도 증진**의 목적
- **중앙집중적인 서버환경**과 **클라우드** 활용
  - 학습 서버 / 추론 서버 / 엣지 디바이스 환경
- **배포 & 관리 & 모니터링**기능

1. 서버형 시스템

   API를 활용한 추론엔진
2. 서버형 커스터마이즈 시스템

   시스템 구축 여건이 없거나 API 제약이 있는 경우 별도 업무 시스템을 통한 커스터마이징
3. 엣지형 시스템

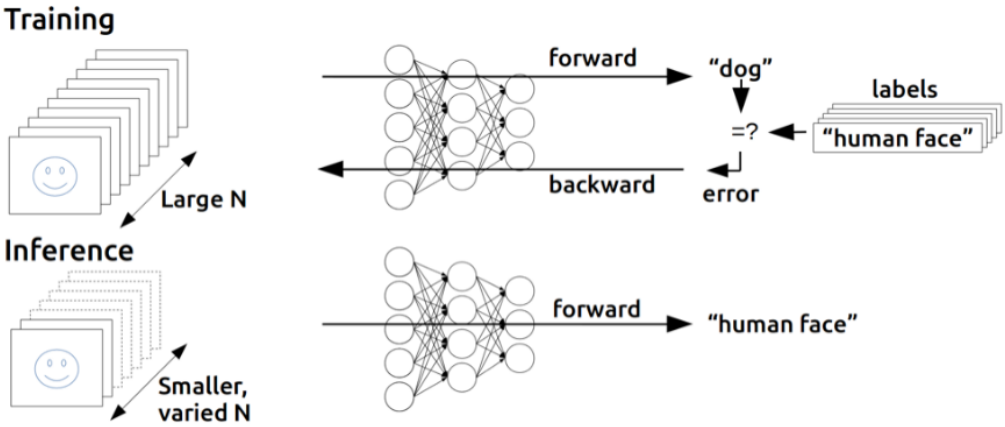
   엣지단에 설치한 리시버와 소이넷의 네트워크 연결 후 작업 진행

- 보유특허
  - **인공지능 실행가속을 위한 인공지능 실행모델 설정방법 및 인공지능 실행가속시스템**

    B-단계

    . . .

2. 기술 동향 및 기본개념



- 머신러닝에서 **추론**은 **학습**과 달리 입력과 피드백으로 구성되지 않고 **입력** 과정에서 weight의 수정등 다양한 작업이 즉각적으로 이루어져 훨씬 속도가 빠르다.
  - 입력 사이즈 : **배치 사이즈**
- SLO (**Service Level Objectives**) : 서비스 단에서 충족해야할 목적
  - 위 단위를 활용해 처리시간 및 정확도 조정
- **배치사이즈**가 큰 추론 작업의 경우 **GPU**를 통한 작업이 훨씬 효율적이다
- 현재 AWS에서 ~~학습~~이 아닌 **추론**을 통한 작업이 비용 절감에 두드러져 추론분야가 대두되는 중이다.

3. 질의사항

- 서버형 시스템 / 서버형 커스터마이징 시스템 / 엣지형 시스템의 개념이 잘 이해가 되지 않습니다.
- Python으로 구현된 로직을 C/C++로 변환해야하는 이유가 궁금합니다.
- 머신러닝에서 멀티스레드가 활용되는 방식이 궁금합니다.