서버리스 앙상블 추론 시스템

Serverless Inference System for Ensemble Models

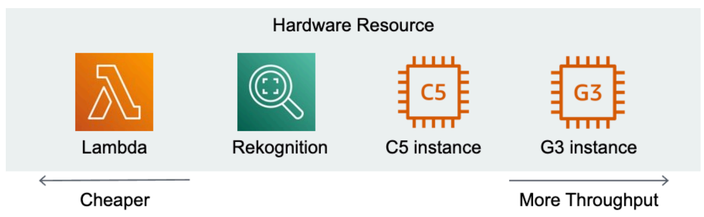
Sungjae Lee*, Jaeghang Choi, Kyungyong Lee
{odobenus, chl8273, leeky}@kookmin.ac.kr

연구배경 및 목표

딥러닝 추론 시스템은 사용자의 요구사항에 따라서 **제한된 지연 시간**안에 **높은 정확도**를 보이는 것이 중요하다. 모델의 정확도를 향상시키기 위해서는, 단일 모델을 사용하는 것 보다 여러 모델을 앙상블 방법으로 합쳐 사용하는 것이 효과적이다. 하지만 딥러닝 앙상블 모델은 일반적으로 직렬화된 시스템으로 앙상블 할 모델이 증가할수록 지연시간이 비례하게 증가한다. 본 연구는 이러한 직렬화된 시스템의 단점을 보완하기 위해 **서버리스 컴퓨팅**의 병렬적인 특성을 활용한 딥러닝 앙상블 추론 시스템을 제안한다.

클라우드 환경의 딥러닝 추론 시스템

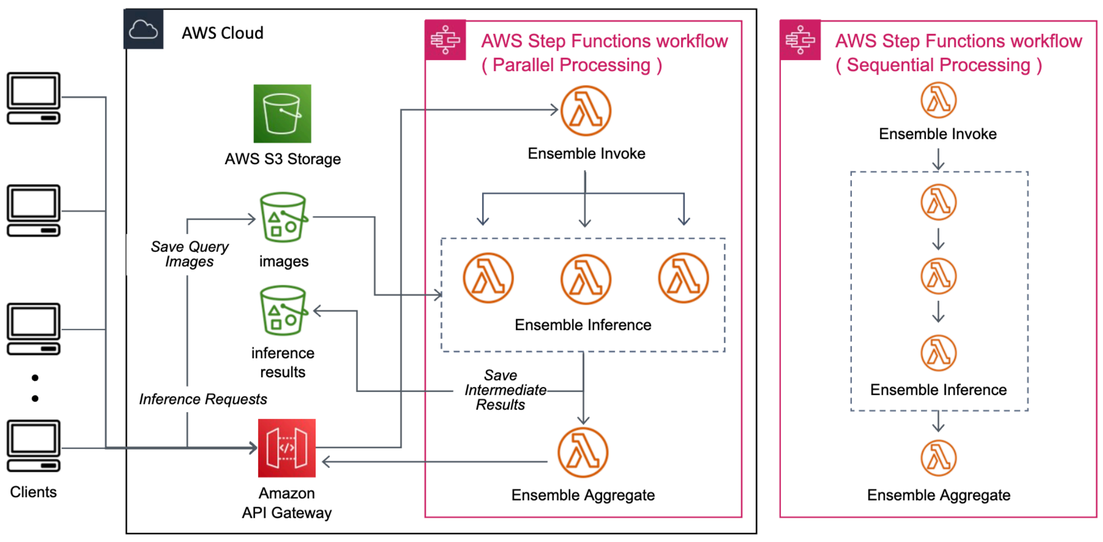
딥러닝 추론 시스템은 사용자의 **SLO(Service Level Objective)** 와 시스템의 **구성 비용**을 고려해야 한다. **클라우드 환경**에서 시스템을 구성할 때 기존 시간당 비용을 지불해야하는 가상머신 대여 방식과 다르게, **서버리스 환경**에서는 요청 당 비용을 지불하기 때문에 적은 횟수의 추론 요청을 유연하게 처리할 수 있다.



[클라우드 환경의 하드웨어 리소스 특성]

서버리스 환경에서는 완전관리형의 **함수** 단위로 시스템을 구성할 수 있어 **병렬화**된 시스템 구성에 매우 유용하다. 배포된 함수는 요청량의 증가에 따라 가상머신(MicroVM)을 실행, 중단 하며 **높은 확장성**을 보장한다. 따라서 이러한 장점은 딥러닝 앙상블 모델을 구성하기에 최적의 조건이 된다.

딥러닝 앙상블 모델의 서버리스 추론 시스템



[딥러닝 추론 앙상블 시스템 구조도]

본 연구에서는 다양한 앙상블 구조 중 배깅 형태의 앙상블 방법으로 시스템을 구성한다. **AWS Step Functions** 서비스로 앙상블 시스템을 구현하였다.
추론 작업은 다음의 과정으로 작동한다.

1. 클라이언트 요청에 따라서 **추론 단계를 Invoke** 하는 람다 함수 실행
2. 추론 단계에서 각각 병렬화, 직렬화 방식으로 작업 진행
3. 추론이 완료된 후 앙상블 람다 함수에서 유사도 **결과를 앙상블**하여 정확도 평가

직렬화된 시스템은 기존의 일반적인 구조로 추론 모델이 순차적으로 작업한다. 하지만 **병렬화된 시스템**은 모델 함수에 동시에 요청을 보낸 후 동시에 독립적으로 작업을 진행한다. 그다음 모든 함수가 작업이 끝난다면 모델들의 추론 결과를 앙상블 함수에 전달한다. 모든 함수가 동시에 작업을 진행하여 순차적인 구성보다 효율적이다.

각 모델 추론 시간의 집합을 $T \ni \{t_1, t_2, \ldots t_n\}$라고 정의하면 각 시스템 별 추론 작업에 소요되는 시간은 다음과 같다.

- 직렬화된 시스템: $\sum_{i=1}^{n} t_i$, 병렬화된 시스템: $max(T)$

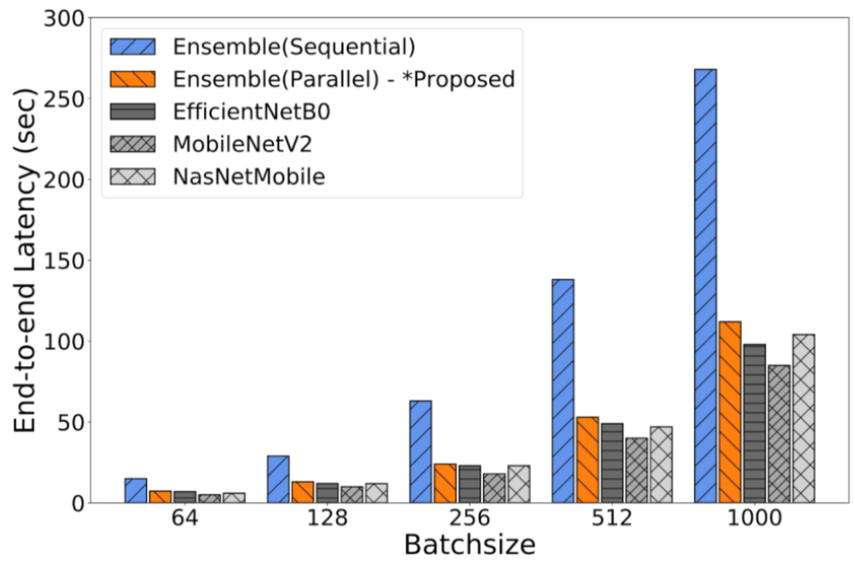
실험 과정 및 결과

본 실험은 딥러닝 앙상블 모델의 정확도와 처리 속도를 측정한다. 사전 학습된 이미지 분류 모델 EfficientNetB0, MobileNetV2, NasNetMobile을 사용하여 50000개의 ImageNet 평가 데이터를 가지고 추론을 수행한다.

	Ensemble	EfficientNetB0	MobileNetV2	NasNetMobile
Accuracy (%)	77.36	74.89	71.49	74.29

[앙상블 모델과 단일 모델의 정확도 비교]

딥러닝 앙상블 모델은 모든 단일 모델보다 **높은 정확도**를 가짐이 확인된다. 단일 모델 중 가장 높은 정확도를 가진 EfficientNetB0 모델 보다 2.47% 높으며 가장 낮은 정확도를 가진 MobileNetV2 모델 보다 5.87% 높은 정확도를 보이며 많은 정확도 향상이 가능함을 보인다.



[배치 크기에 따른 모델 추론 지연시간 비교]

딥러닝 앙상블 모델 중 병렬화된 시스템은 직렬화된 시스템보다 최대 2.6배 빠른 추론 속도가 나타난다. 단일 모델과 비교시 추론 속도가 더 소요되지만 직렬화된 시스템에 비해 확연히 **적은 추론 속도로 높은 정확도의 추론 작업**이 가능하다. 특히 배치 크기가 1000인 경우에서 단일 모델 중 가장 큰 지연시간을 갖는 NasNetMobile에 비하여 7.7%의 적은 추가 시간 소요만으로 3.07% 향상된 정확도를 보인다.

향후 계획

본 연구를 통해 **딥러닝 모델의 앙상블 구성**을 통해 **정확도 향상**이 가능하며 서버리스 환경에서 병렬적인 모델 실행 구조를 구현하여 **단일 모델의 지연시간 수준**에서 처리가 가능함을 확인할 수 있었다.
본 연구에서 더 나아가 다음의 추가적인 발전이 가능하다.

1. 모델의 개수, 사용자의 추론 요청이 증가함에 따라 일정 수준의 지연 시간을 보장할 수 있는지 파악이 필요
2. 이미지 분류 워크로드를 넘어 둘 이상의 데이터 타입으로 고도화된 멀티모달 워크로드도 적용 가능성 고려
3. 서버리스 환경의 제한된 자원에서 성능 최적화 방법 고안 필요

사사

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 ICT 연구과제 (2017-0-00396) 및 한국연구재단 이공분야 기초연구사업(NRF-2020R1A2C1102544)의 지원을 받아 수행됨.

참고문헌

- Zhang, C., et al. "Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving", USENIX ATC '19
- Crankshaw, D., Wang, X., et al "Clipper: A low-latency online prediction serving system", USENIX NSDI '17
- Perez, F., Avila, S., & Valle, E. "Solo or ensemble? choosing a cnn architecture for melanoma classification", CVPR 2019 Workshop
- Ratner, A., Alistarh, D., Alonso, G., Andersen, D. G., Bailis, P., Bird, S., ... & Talwalkar, A. "MLSys: The new frontier of machine learning systems", arXiv preprint 2019