인공지능(AI) 시스템 규모 산정을 위한 참조법 연구

서병준, 황세진, 한주연

TTA(한국정보통신기술협회)

sbj8388@tta.or.kr, hsejin314@tta.or.kr, hanjy@tta.or.kr

A Study on the reference method for sizing of AI systems

Seo Byoungjun, Hwang Sejin, Han Juyeun

Telecommunications Technology Association

요 약

본 논문은 인공지능(AI) 시스템의 정량적인 규모를 산정하기 위한 참조법을 연구하고, 산정 규모의 일부 파라미터를 변경한 산정값과 TPCx-AI 벤치마크를 실행하여 얻은 실험값과의 비교 연구를 수행한다. 인공지능(AI) 시스템은 우리 시대의 핵심 기술로 자리 잡고, 공공 분야뿐만 아니라 항공 우주, 국방 및 학계 등 다양한 분야에 적용된다.[1] 그러나 인공지능 시스템을 도입하기 위해서는 고액의 비용이 필요하고, 효율적인 시스템 도입을 위해서는 벤치마크를 통한 성능 측정이 필수적이다. 성능 측정을 통한 시뮬레이션법은 시스템 규모의 정확한 값을 얻을 수 있지만, 시험에 필요한 시간과 비용이 많이 소요된다. TPCx-AI 벤치마크는 기업과 공공 부문의 실제 사례를 모델링한 다양한 Use Case를 통해 인공지능 시스템의 데이터 생성, 전처리, 로딩, 학습, 추론, 검증 등의 모든 단계들의 성능을 종합적으로 측정하여 단일 성능 수치로 나타내주는 도구다. 본 논문을 통해 인공지능(AI) 시스템 구축에 필요한 규모를 예측할 수 있는 시스템 규모 산정 방식을 제안하고, TPCx-AI 성능 점수와 비교하는 실험을 통해 두 수치 간의 유사성을 평가한다. 이를 통해 TPC(www.tpc.org) 홈페이지 게재된 성능 점수 리스트 참조를 통해 AI 시스템에 사용되는 하드웨어 및 소프트웨어를 참고하도록 가이드하여 성능 측정에 소요되는 시간과 비용을 절약하고자 한다.

1. 서 론

현재의 행정안전부에서는 국내 공공기관이 컴퓨팅 관련 장비를 도입할 때, 정보 시스템의 정확한 규모를 산정하기 위해 관련 고시에 따라 "정보시스템 하드웨어 규모 산정지침" 표준을 사용하고 있다.[2] 그러나 이 표준은 기존의 OLTP 서버, WEB/WAS 서버, 스토리지 서버에 국한되며, AI 인공지능 시스템을 위한 규모 산정에는 적합하지 않다. 따라서 본 논문에서는 인공지능(AI) 시스템의 정량적인 규모를 산정하기 위한 참조법을 연구하고, TPCx-AI 벤치마크를 사용하여 산정 규모의 파라미터를 변경하면서 인공지능 시스템의 성능을 비교 연구한다. TPCx-AI 벤치마크는 실제 10종의 Use Case를 모델링하여 인공지능 시스템의 모든 단계(생성, 전처리, 로딩, 학습, 추론, 검증 등)의 성능을 측정하여 단일 성능 수치로 나타내는 도구이다. 이러한 연구를 통해 국내 AI 시스템의 규모 산정 방식을 제시하고, 실험을 통해 산정값과 성능 점수 간의 유사성을 평가한다.

2. 소개

2.1. TPC 단체와 TPCx-AI 벤치마크 설명

TPC(Transaction Performance Council)는 서버, 스토리지, DBMS(데이터베이스) 등 컴퓨팅 장비 성능 및 신뢰성에 대한 국제 표준을 제정하고 관련 벤치마크 도구를 개발하는 비영리 단체이다. TPC에서 개발 중인 표준과 도구는 OLTP(TPC-C, TPC-E), OLAP(TPC-H), IoT(TPCx-IoT), BigData(TPCx-BB) 등 전형적인 벤치마크 모델부터 차세대 벤치마크 모델까지 범위를 확대하여 표준과 벤치마크 도구를 개발하고 있다. 그중 TPCx-AI 벤치마크 도구는 얼굴인식, 추천시스템 등 총 10 종의 AI Use Case를 AI 파이프라인(데이터 로딩, 전처리, 학습, 추론 등)을 통해 AI 시스템의 성능을 전반적으로 측정하는 벤치마크 도구이다. 본 벤치마크에서 산출되는 성능 수치는 기본적으로 Use Case마다 AI 추론 요구 정확도를 통과해야 얻어지는 성능 값이다.

[표1] TPCx-AI Use Case 10종 수행

UseCase	비즈니스 모델	품질측정 지표	요구 정확도 Threshold
1	고객 분류	K-means Cluster	N/A
2	음성 변환	Work Error Rate	50% 이상
3	판매 예측	Forecase Accuracy	94.6% 이상
4	스팸 탐지	Matthews Correlation Coefficient	65% 이상
5	가격 예측	Prediction Quality(RMSLE)	50% 이상
6	하드웨어 고장 탐지	F-score	19% 이상
7	제품 추천	Mean Absolute Error	98.2% 이상
8	쇼핑 유형 분류	Classification Accuracy	65% 이상
9	얼굴 인식	Face Recognition Accuracy	90% 이상
10	허위 금융 거래 탐지	Classification Accuracy	70% 이상

TPCx-AI는 10종의 Use Case에 대한 비즈니스 모델, 품질측정 지표, 요구 정확도를 [표1]에서 제시하고 있다. 10종의 Use Case를 AI 파이프라인(데이터 생성 → 로딩 → 제거 & 변환 → 전처리 → 학습 → 추론 → 평가 → 처리량 → 유효성)을 통하여 AI 시스템의 종합적인 성능 수치를 [표2]에서 제시하는 산정 방식으로 산출한다.

[표2] TPCx-AI 성능 수치 산정 방식

성능 보조 지표	메트릭 내용
SF(Scale Factor)	학습 데이터 용량(용량 구성: 1GB, 3GB, 10GB, 100GB, 300GB, 1000GB, 3000GB, 10000GB)
N	UseCase 수(단위: 개수)
T_{LD}(Load Metric)	데이터 로드 소요시간(단위: 초)
T_{PTT}(Power Training Metric)	10개 Use Case에 대한 모델 학습 소요 시간의 기하평균(단위: 초)
T_{PST}(Power Serving Metric)	10개 Use Case에 대한 모델 추론 소요 시간의 기하평균(단위: 초) ※2회 수행 중 가장 긴 소요 시간 선택
T_{TT}(Throughput Test Metric)	10개 Use Case에 대한 처리량 테스트 평균 소요 시간의 기하평균(단위: 초)
최종 성능 수치 산출식	
AIUCpm@SF = $\frac{SF \times N \times 60}{\sqrt[4]{T_{LD} \times T_{PTT} \times T_{PST} \times T_{TT}}}$	

TPCx-AI 벤치마크의 최종 성능은 분당 Use Case 처리 성능으로 단일 수치로 산출되는데, 학습 데이터 용량(SF), Use Case 수(N), 데이터 로딩 소요 시간(T_{LD}), 10개 Use Case 학습 소요 시간 기하평균(T_{PTT}), 10개 Use Case 추론 소요 시간 기하평균(T_{PST}), 10개 Use Case 처리량 테스트 소요 시간 기하평균(T_{TT})들이 단일 성능 수치 산출에 필수 항목으로 사용된다. 단, 해당 수치는 데이터 용량(SF) 규격 기준에 따른 성능 수치를 나타내기 때문에 같은 데이터 용량(SF) 내에서만 비교할 수 있다.

2.2 국내 AI 시스템 규모 산정을 위한 산정 방식 제시

TPCx-AI 성능 점수는 데이터 용량(SF)에 따라 구분하고, 데이터 용량(SF) 구분은 [표3]에서 1GB부터 10,000GB까지 나눈다.

[표3] TPCx-AI Scale Factor에 따른 데이터 구성과 개수

Scale Factor	비정형 데이터		정형 데이터(14개의 테이블)			데이터 총용량 (GB)
	이미지 개수 (파일포맷:png)	오디오 개수 (파일포맷:wav)	고객정보 개수 (파일포맷:csv)	…	리뷰정보 개수 (파일포맷:csv)	
1	70	387	70,710		134,349	1
3	218	798	145,773		306,117	3
10	1,291	1,965	358,817		825,286	10
30	7,084	4,619	843,356		2,024,064	30
100	46,268	11,787	2,152,033	…	5,595,278	100
300	241,102	26,895	4,910,448		13,749,260	300
1,000	1,303,938	62,539	11,418,023		33,112,258	1,000
3,000	5,368,444	126,906	23,169,807		71,826,411	3,000
10,000	22,531,953	259,980	47,465,671		151,890,144	10,000

위 데이터는 TPC에서 제공하는 소프트웨어인 PDGF(병렬 데이터 생성 프레임워크)를 이용하여 생성되는 데이터이고, 서로 다른 데이터 용량(SF) 결과 간에 비교는 불가능하다.

[표4] 국내 AI 시스템 규모 산정 방식

구분	산정 항목	내용	보정값 적용 범위	일반값
A1	데이터 용량	정형, 비정형 학습 데이터 용량(GB)	1 ~ 10,000	-
A2	클러스터 보정	AI 시스템 성능 향상을 위한 클러스터 노드 확장	▪ 클러스터 내 노드 수	산정값
A3	처리자 수	AI 시스템 동시 요청에 따른 처리자 수	▪ 2개~9개: 0.05 ~ 0.2 ▪ 10개~99개: 0.10 ~ 0.3 ▪ 100개~200개: 0.16 ~ 0.37	산정값
A4	ML/DL 업무 보정	머신러닝(ML), 딥러닝(DL) 관련 AI 서비스 용도에 따른 업무 비율	▪ ML 용도: 1 ~ 7 ▪ DL 용도: 1 ~ 3 ▪ ML+DL 용도: 5 ~ 10	10
A5	AI 시스템 요구 정확도	학습된 모델에서 요구되는 추론 정확도	▪ 추론 정확도: 0.5 ~ 1	산정값
A6	학습/추론 업무 보정	AI 시스템에서 요구되는 학습과 추론의 업무 비율	▪ 학습 용도: 20 ~ 80 ▪ 추론 용도: 10 ~ 20 ▪ 학습+추론 용도: 50 ~ 100	산정값
산정식	AIUCpm@A1 = (A2 * A3 * A4 * A5 * A6)			

국내 AI 시스템 규모를 산정하기 위해 크게 6개 산정 항목을 [표4]와 같이 제안한다. 본 산정 가이드를 토대로 국내 AI 시스템의 규모를 구하고, TPC 홈페이지(www.tpc.org)에 게재된 성능 점수 리스트를 참조하여 AI 시스템에 사용될 수 있는 HW와 SW를 참고하도록 가이드한다.

3. 실험 환경 구성

제안한 AI 시스템 규모 산정 값에 따른 성능 수치 비교 분석을 위해 서버 1대에서 실험을 진행하였다. 실험 환경(HW, SW) 정보는 [표5]와 같다.

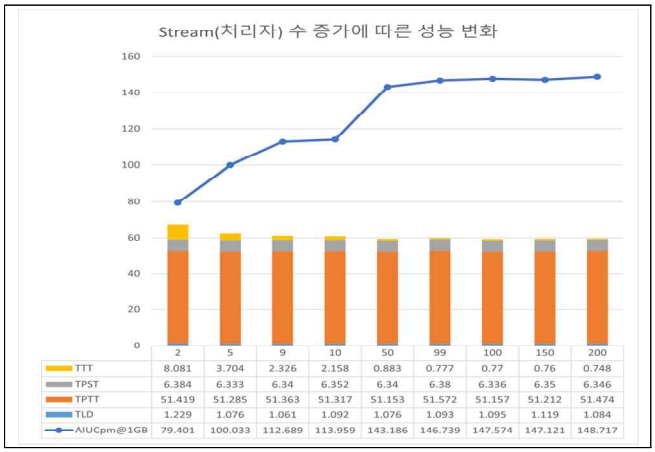
[표5] 실험 환경(HW, SW) 정보

하드웨어(HW)				소프트웨어(SW)		
CPU	Intel(R) Xeon(R) Gold 6140 * 2EA			OS	CentOS Stream 8	
MEM	128 GB	HDD	600 GB	Platform	Python 3.11	Miniconda v24.1.2
					Java 1.8.0_362	SBT v1.9.9

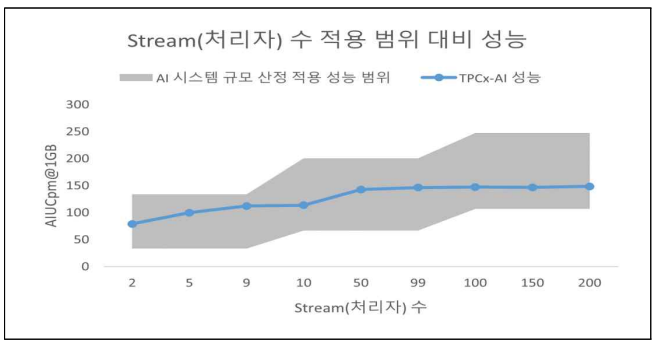
4. 규모 분석 및 실험 결과

본 실험은 Stream(처리자) 수가 증가함에 따라 TPCx-AI 벤치마크의 성능이 어느 정도 영향을 받는지 확인한다. Stream(처리자) 수를 조정하여 성능 변화를 확인하고, 처리자 수에 따른 규모 산정 값 적용 범위를 파악한다. 실험에서는 환경 설정(데이터 용량=1, 클러스터 보정=1)과 벤치마크 특성(ML/DL 업무 보정=10, AI 시스템 요구 정확도=0.67, 학습/추론 업무 보정=100)을 고정하며, 최대 200개의 Stream을 사용하여 결과를 분석한다. 결과는 Stream(처리자) 수가 증가함에 따라 성능이 비례적으로 증가하지 않음을 보여준다. 이는 Stream(처리자) 수가 처리량 단계에서는 영향을 미치지만, 다른 단계(데이터 로딩, 학습, 추론)에서는 영향을 받지 않기 때문이다. 결과는 [그림2]와 같이 Stream(처리자) 수 증가에 따라 TPCx-AI 성능 결괏값이 AI 시스템 규모 A3(처리자 수) 적용 성능 범위 안에 속하는 것을 볼 수 있다. 다만, Stream(처리자) 수를 100개 이상으로 증가시키는 경우, [그림1]과 같이 T_{TT}(처리량)의 성능이 0.7 이상으로 올라가지 않아 종합적인 성능 점수가 일정 이상으로 올라가지 않는 것을 확인할 수 있다.

[그림1] TPCx-AI 실험에 따른 성능 변화



[그림2] AI 시스템 규모 산정 적용 성능 범위 대비 TPCx-AI 성능 수치



이는 본 실험에서 구성한 하드웨어(CPU Core 수: 36개 / Thread 수: 72개) Thread 수에 따른 병목 현상으로 파악된다. 하지만 TPCx-AI 벤치마크는 단일 노드가 아닌 멀티 노드 클러스터 환경에서도 구동 가능하다. Stream(처리자) 수 증가에 따른 병목 현상은 클러스터 구성과 하드웨어 스펙 향상으로 보완이 가능하다는 점을 고려하여 A3(처리자 수) 보정 값을 0.37 범위까지 선택할 수 있게 확장 가능성을 열어두었다.

5. 결론 및 향후 연구

본 논문에서는 TPC 단체와 TPCx-AI 벤치마크를 소개하고, 국내 AI 시스템 규모 산정을 위한 규모 산정 방식을 제안하여 TPCx-AI 성능 수치에 참조할 수 있도록 가이드하였다. 그리고 규모 산정 수치의 적절성 확인을 위해 실험을 통하여 TPCx-AI 성능 수치가 A3(처리자 수) 적용 범위에 따른 규모 산정 적용 성능 범위 안에 속하는지 확인하였다. 정확한 AI 시스템 규모 산정을 위해서는 정밀한 수치 획득을 위한 다수의 실험과 AI 산업계와 학계에서 요구되고 있는 추가적인 산정 항목들, 그리고 실측에 필요한 TPCx-AI 벤치마크 도구 기술의 고도화가 필요하다. TPCx-AI 벤치마크 도구와 표준은 현재까지도 TPC 위원들과 산업계, 학계의 수요를 바탕으로 개발을 진행하고 있다. 본 연구를 바탕으로 국내 AI 시스템 규모의 더 정확한 산정 방식 개발을 위해 추가적인 TPCx-AI 성능 실험(클러스터 멀티 노드 수 변화에 따른 성능 증가 추이)과 국내 AI 산업계와 학계의 전문가 의견을 수렴하여 국제 TPC 단체 활동을 통해 TPCx-AI 벤치마크 개발 연구를 수행할 예정이다.

ACKNOWLEDGMENT

이 논문은 2024년도 "과학기술정보통신부"의 재원으로 "정보통신기획평가원"의 지원을 받아 수행된 연구임[세부사업: 정보통신융합산업-ICT산업기반확충(정진)-글로벌ICT혁신클러스터조성-HPC이노베이션허브]

참 고 문 헌

[1] TPCx-AI Specification v1.0.3.1

[2] 정보시스템 하드웨어 규모산정지침(TTAK.KO-10.0292/R3)