Research Report

연구보고서 2017-007

인공지능 컴퓨팅 인프라의 중요성과 실태조사 현황 및 시사점

A Report of Importance of AI Computing Infrastructure and Current Status / Implications on Survey

추형석

2018.04.

SPRi 소프트웨어정책연구소

집필기관 및 참여인원 : 소프트웨어정책연구소
추형석

이 보고서는 2017년도 과학기술정보통신부 정보통신진흥기금을 지원받아 수행한 연구결과로 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

목 차

Contents

표 목 차

그 림 목 차

요 약 문

1. 제 목

인공지능 컴퓨팅 인프라의 중요성과 실태조사 현황 및 시사점

2. 보고서의 목적

인공지능과 컴퓨팅 인프라의 관계를 기술적으로 분석하여 그 중요성에 대해 분석한다. 최근 인공지능 기술로 가장 널리 활용되는 심층학습은 경험적으로 결과를 도출하는 것이 일반적이다. 또한 심층학습은 빅데이터에서 패턴을 예측한다는 점에서 필연적으로 많은 계산을 요구한다. 이러한 인공지능의 특징으로 인하여 인공지능 컴퓨팅 인프라의 수요는 전 세계적으로 급증하고 있다. 글로벌 IT 기업은 인공지능 전용 컴퓨팅 환경을 구축하여 클라우드 서비스를 제공하고 있으며, 일본은 195억 엔 규모의 인공지능 클라우드 인프라를 구축할 예정이다. 우리나라 역시 이러한 중요성을 인지하고 인공지능 컴퓨팅 인프라 지원 사업을 추진하고 있으나 양적인 측면이 미흡하다. 이번 보고서에서는 인공지능 관련 중소기업·스타트업의 인공지능 컴퓨팅 인프라의 현황과 향후 수요에 대한 실태조사 결과를 분석한다. 이 결과를 활용하여 현재 국내 중소기업과 스타트업의 컴퓨팅 인프라 수요를 가늠하고, 정책 설계 자료로 활용하는데 목적을 둔다.

3. 보고서의 구성 및 내용

보고서는 크게 두 가지 주제로 구성된다. 먼저 인공지능의 성공요인을 바탕으로 인공지능 연구에 있어 컴퓨팅 인프라의 중요성에 대해 분석했다. 현대 연산처리장치의 발전사와 특성을 바탕으로 ‘왜 인공지능 연구에 컴퓨팅 인프라가 필수적인가?’에 대해 기술적으로 분석하여 논리적인 근거를 제시한다. 특히 인공지능 연구에 가장 활발히 활용되고 있는 그래픽연산처리장치(GPU)의 하드웨어적인 특성을 소개한다.

다음은 인공지능 관련 중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 향후 수요를 파악하기 위한 실태조사 결과를 기술했다. 실태조사 대상을 중소기업과 스타트업으로 한정한 이유는 인공지능 인프라 지원 사업을 설계할 때 가장 먼저 지원해야할 부분으로 고려했기 때문이다. 중소기업이나

스타트업은 대기업과 다르게 인프라 구축 및 운영에 대한 비용을 감당하기에 상대적으로 어렵기 때문이다. 또한 현재 인공지능의 기술을 견인하는 주체는 전 세계적으로 스타트업에 집중돼 있다. 실태조사에서는 현재 보유하고 있는 인공지능 컴퓨팅 장비와 향후 인공지능 연구에 활용할 컴퓨팅 장비의 수요에 대해 조사했다. 조사결과의 의미와 한계를 제시함으로써 지표의 객관성을 확보했다. 또한 정부의 인공지능 컴퓨팅 인프라 지원 방법에 대한 선호도 조사를 통해 현재 산업에서 요구하는 사항이 무엇인지 파악했다.

4. 활 용

인공지능 컴퓨팅 인프라 지원 정책의 설계 자료로 활용

Summary

1. Title

A Report of Importance of AI Computing Infrastructure and Current Status / Implications on Survey

2. Purpose

Analyze the importance of artificial intelligence and computing infrastructure technically. Deep learning, which is the most widely used artificial intelligence technology in recent years, is generally based on empirical results. Also, deep learning requires a lot of computation in that it predicts patterns in big data. Due to the characteristics of this artificial intelligence, the demand for artificial intelligent computing infrastructure is rapidly increasing worldwide. Global IT companies are providing cloud services by establishing a dedicated computing environment for artificial intelligence. On the other hands, Japan will build an artificial intelligence cloud infrastructure of 19.5 billion yen. Korea is aware of this importance and is supporting the artificial intelligence computing infrastructure project, but it is not enough quantitative aspect. In this report, I analyze the current state of artificial intelligence computing infrastructure of artificial intelligence-related SMEs and start-up, Using these results, I aim to measure the demand of computing infrastructure of domestic SMEs and start-ups and utilize them as policy design data.

3. Contents and Details

The report consists of two main topics. First, based on the success factors of artificial intelligence, I analyzed the importance of computing infrastructure in artificial intelligence research. With the evolution and characteristics of modern computation processing devices, it provides a logical basis for the technical analysis of why computing infrastructure is essential for artificial intelligence research. Especially, I introduce hardware characteristics of GPU which is most actively used in artificial intelligence research.

The following describes the status of the artificial intelligence computing infrastructure of small-enterprises and start-ups, and the results of actual surveys to identify future demand. The reasons for limiting the survey to SMEs and start-ups are considered to be the first step in designing an artificial intelligence infrastructure support project. Unlike large companies, SMEs and start-ups are relatively difficult to cope with infrastructure construction and operation costs. In addition, the subject that is currently driving the artificial intelligence technology is focused on start-up around the world. In this research, I investigated the demand for artificial intelligent computing equipment and computing equipment to be used in future artificial intelligence research. By presenting the meaning and limitations of the survey results, the objectivity of the indicators was secured. I also examined the preferences of the government policy on how to support the artificial intelligence computing infrastructure to understand what the industry needs.

4. Application

Utilize policy design of artificial intelligence computing infrastructure

제1장 서 론

제1절 연구배경 및 필요성

지난 2016년 3월 '알파고 쇼크'로 기억될 세기의 대국에서 인공지능 바둑 프로그램 알파고가 세계 최정상 바둑기사 이세돌 9단을 4:1로 꺾었다. 이 기세를 이어 지난 2017년 10월, 인간의 기보를 전혀 학습하지 않은 '알파고 제로'가 바둑을 정복하기에 이르자 인공지능의 가능성에 대한 일말의 의심을 불식시켰다. 이처럼 인공지능의 성능이 급성장하게 된 계기는 무엇일까? 먼저 인공지능의 역사에서 그 단서를 얻을 수 있다.

1. 인공지능의 역사

인공지능은 1956년 다트머스 회의에서 그 개념이 최초로 정립되면서 인간의 지능을 대체할 학문분야로 인정받았다. 1970년대에는 특정 분야에 전문적인 지식을 탑재한 전문가 시스템(Expert System)에 막대한 연구자금이 투입되어 인공지능의 첫 번째 황금기를 맞이했다. 하지만 부풀었던 기대와는 다르게 전문가 시스템이 목표문제를 해결하지 못함에 따라 1980년대 중반까지 내리막길을 걷게 된다.

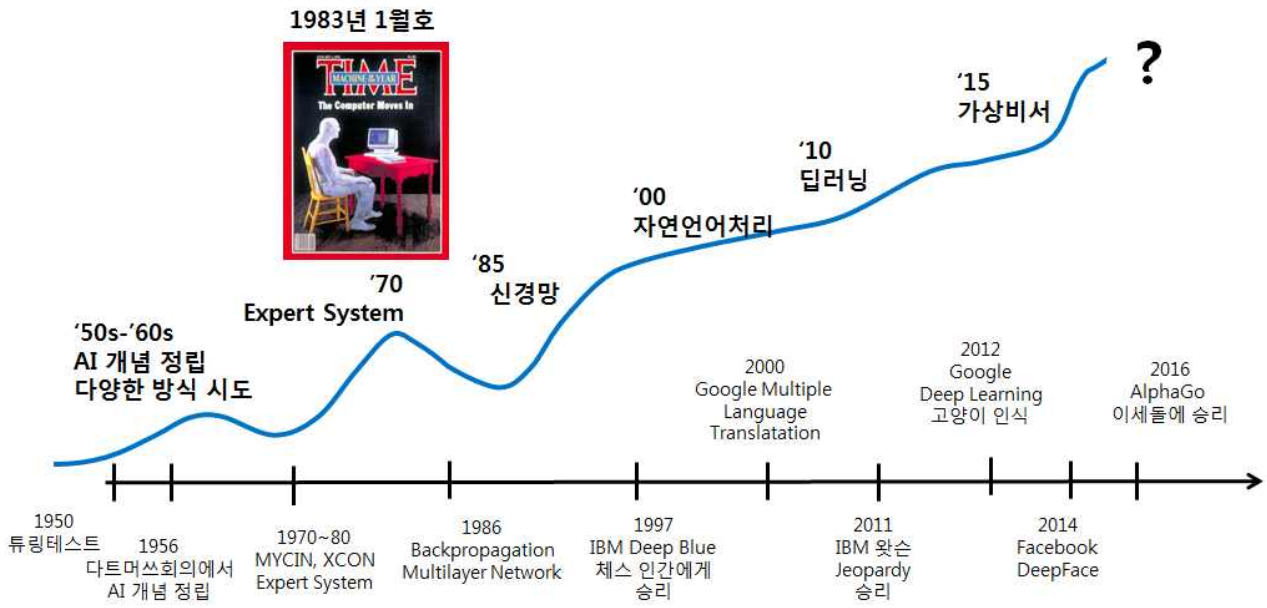
이어 1985년 신경망 학습에 대한 획기적인 연구 성과로 반전을 겪는다. 인공신경망(Artificial Neural Network)으로 잘 알려진 다층 퍼셉트론(Multi-Layer Perceptron)의 학습 방법인 오류역전파법(Error back-propagation method)이 지능적 문제를 해결할 수 있는 가능성을 보여줬기 때문에 인공지능 분야는 다시 한 번 탄력을 받기 시작했다. 그러나 오류역전파법 역시 제한적인 데이터, 컴퓨팅 파워 등 물리적인 한계로 인해 빛을 발하지 못했고 인공지능은 다시 암흑기에 진입했다.

한편, 2006년 캐나다 토론토 대학의 제프리 힌튼 교수는 인공지능 부활의 신호탄을 알리는 논문을 발표했다. 이것은 바로 급부상하고 있는 키워드인 심층학습(Deep Learning)이다. 심층학습은 기존의 인공신경망을 더욱 깊게 구성하는 기술로 그동안 지적되어 왔던 오류역전파법의 한계를 극복했다. 인공신경망을 학습한다는 개념은 곧 비용함수의 전역 최소값(global minimum)을

찾는 문제로 귀결하는데, 초기 값에 대한 의존성이 커서 신경망을 새로 학습할 때 마다 일관성이 없는 결과를 도출했기 때문이다. 이에 힌튼 교수 연구진은 자율학습(unsupervised learning)이라는 개념을 도입하여 문제를 해결하는 방안을 제시했다.

이 자율학습이 적용된 결과가 대중에도 잘 알려진 구글 브레인 프로젝트이다. 구글 브레인은 유튜브 영상에서 고양이를 인식하는 실험을 했는데 학습 과정 중에는 고양이의 정보를 알려주지 않았다. 연구결과 높은 확률로 고양이를 인식하는데 성공했고, 이 결과를 바탕으로 많은 연구진들이 차차 심층학습을 도입하기 시작했다.

현재 인공지능 분야에서는 이미지 인식, 필기 인식, 음성 인식, 영어의 자연어 처리가 가장 성공적인 분야로 꼽힌다. 이미지에서의 객체 인식률은 이미 사람에 버금가는 수준에 도달했고, 중국어와 영어 음성을 동시에 인식하는 기술이 소개됐으며, 필기체 인식은 99%에 육박하여 우편물처리 자동화에 이미 적용됐다.



[그림 1-1] 인공지능의 70년 역사

자료 : 알파고의 능력은 어디에서 오는가?, 소프트웨어정책연구소(2016)

2. 현대 인공지능의 성공요인

두 번의 황금기와 암흑기를 지나 세 번째 황금기를 맞이하고 있는 인공지능은 정보통신기술 혁명의 자연스러운 결과이다. 그 혁명의 저변에는 크게 세 가지 요소가 있다.

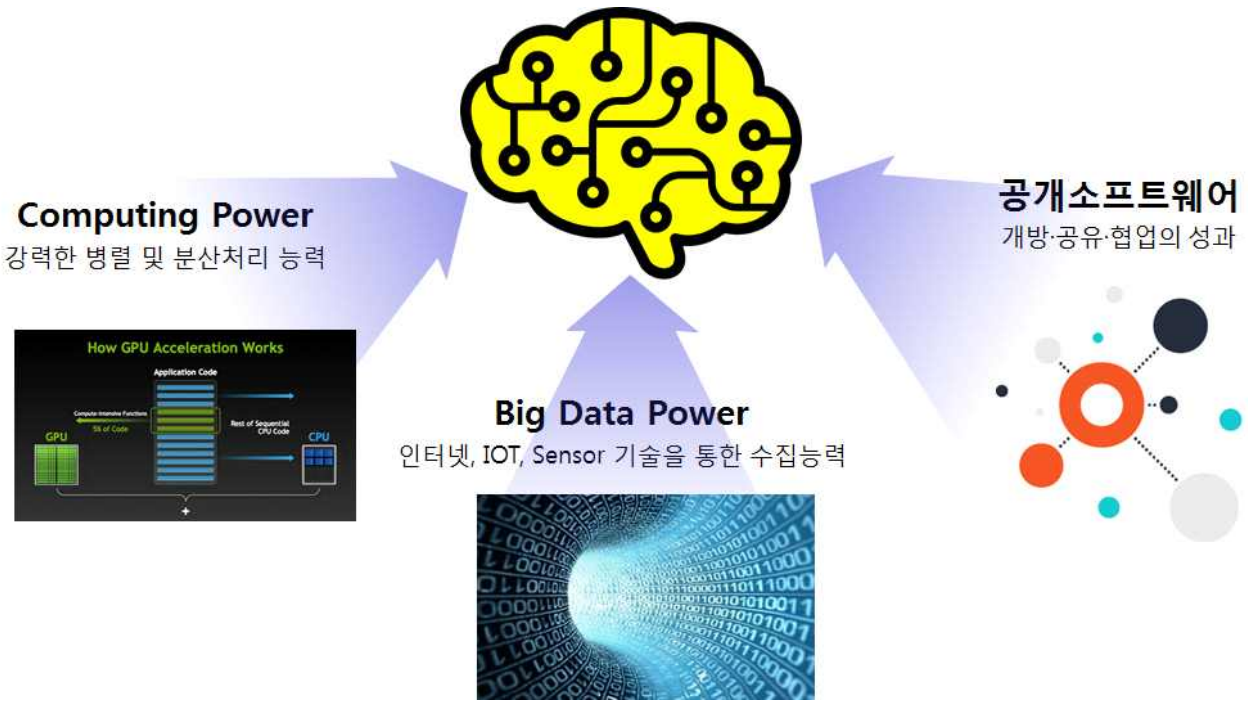
컴퓨팅 환경의 고도화가 그 첫 번째 요소이다. 컴퓨터 연산처리장치는 지수적인 성능 향상을 달성하며 발전했으며, 실리콘칩의 대량 생산에 따라 저렴한 가격으로 보급이 가능해졌다. 발전 속도를 정량적으로 비교해보자면 현재 스마트폰에 탑재된 AP(Application Processor)의 GPU(그래픽연산처리장치, Graphical Processing Unit)는 약 265 GFLOPs[1]의 이론 성능을 보유하고 있다. 1996년 6월 발표된 세계 1위 슈퍼컴퓨터인 일본의 SR2201/1024의 실측성능은 약 220 GFLOPs로 스마트폰의 GPU의 이론 성능에 육박한다. 20년 전 약 500억 원에 달하는 슈퍼컴퓨터가 손안에 들어온 시대가 된 것이다. 이러한 환경변화에도 불구하고 학습기반의 인공지능 연구에는 여전히 막대한 계산량을 필요로 한다. 예를 들어 심층학습 기술은 신경망 구성에 대한 이론이나 원칙이 경험적으로 짐작(Rule of Thumb)되기 때문에 많이 시도해보는 것이 중요하다. 바둑 인공지능 프로그램 알파고의 기보학습에는 약 3주간의 시간동안 50장의 그래픽카드를 활용하여 계산했는데, 이 결과를 얻기까지 수 천, 수 만 번의 반복 계산이 필요했을 것이다. 따라서 인공지능 연구에는 고성능 연산처리장치의 보급이 가장 핵심적인 역할을 했다고 볼 수 있다.

두 번째 요소로는 빅데이터의 대중화이다. 제4차산업혁명에서 데이터의 중요성은 언급을 하지 않아도 될 정도로 공감대가 형성됐다. 인터넷을 통한 데이터의 보급은 새로운 먹거리를 창조할 뿐만 아니라 지능적 의사결정을 위한 인공지능 연구에도 활용하고 있다. 빅데이터 여론 분석을 통해 기업의 이미지와 상품의 피드백을 추측할 수 있으며, 타겟형 소비자 매칭으로 영업 이익을 극대화하는 것이 현실적으로 가능해졌다. 학습기반의 인공지능 분야에서는 자율학습(unsupervised learning)이라는 개념이 등장함에 따라 데이터를 재가공하기 위한 비용이 절감됐다. 기존의 감독학습(supervised learning)은 데이터와 대응 값이 1:1로 쌍을 이뤄야 했기 때문에 학습에 필요한 데이터 셋을

1) FLOP(부동소수점연산수, Floating point Operation)은 연산수를 나타내는 단위로 연산처리장치의 성능을 측정하는 지표로 사용됨. FLOP/s(초당 부동소수점연산수)는 1초에 얼마나 많은 연산을 처리 할 수 있는지를 나타내는 것으로 슈퍼컴퓨터의 성능비교 등에 적용

구성하는 것이 무엇보다 중요했다. 하지만 자율학습에서는 대응 값이 필요하지 않다. 따라서 인공지능 분야에서 빅데이터의 대중화는 로켓엔진의 연료 역할을 하여 기술적 진화를 가속화 한 것이다.

마지막으로 공개소프트웨어가 인공지능 성공의 원동력이다. 현재 인공지능 및 기계학습 관련 공개소프트웨어는 글로벌 IT 기업이 주도적으로 개발하고 있다. 구글의 텐서플로우(TensorFlow), 마이크로소프트의 CNTK(Computational Network ToolKit), 페이스북의 토치(Torch) 등 현재까지 공개된 인공지능 관련 소프트웨어는 50가지가 넘는다. 이러한 공개소프트웨어의 힘으로 인공지능 연구의 진입장벽이 현저히 낮아졌다. 연구자들은 데이터를 확보하여 모델링하는 업무에 집중할 수 있는 환경이 조성된 것이다. 또한 공개소프트웨어는 집단지성을 통해 발전하여 최신 연구까지 반영된 신기술이 적용됐다. GPU와 같은 고성능 계산자원 역시 병렬화 과정이 모두 구현되어 컴퓨팅 인프라를 십분 활용할 수 있다.



[그림 1-2] 인공지능 성공의 원동력

자료 : 알파고의 능력은 어디에서 오는가?, 소프트웨어정책연구소(2016)

3. 인공지능 컴퓨팅 인프라

이 보고서에서는 현대 인공지능의 3가지 성공요인 중 컴퓨팅 환경의 고도화를 집중적으로 다루고자 한다. 앞서 기술했듯이 현대 인공지능의 핵심인 심층학습은 경험적인 결과를 바탕으로 한다. 따라서 많은 경우를 고려하여 다양한 접근으로 시도하는 것이 무엇보다 중요한데, 이것은 필연적으로 막대한 계산량을 요구한다.

알파고 개발진인 딥마인드는 구글에 인수된 뒤 풍부한 클라우드 인프라를 활용하고 있다. 알파고 역시 구글 클라우드 계산 자원을 활용한 결과라고 볼 수 있다. 알파고의 초기버전은 최대 1,920개의 CPU와 280장의 GPU를 활용한 결과였다. 이를 규모측면에서 환산해 보면 당시 약 380위 정도에 해당하는 슈퍼컴퓨터에 달한다. 실제로 알파고는 30초의 초읽기 시간 동안 약 10만 수를 예측하고 착수를 결정한다. 이것을 일반 PC에서 처리하기 위해서는 약 80분이 필요한 것으로 볼 때 컴퓨팅 인프라는 인공지능의 필수적인 요소임을 확인할 수 있다.

이러한 중요성으로 인해 일본은 195억 엔(약 1,840억 원) 규모의 인공지능 전용 클라우드를 2018년 1분기까지 구축한다고 밝혔다. 미국이나 중국과 같이 글로벌 IT기업(구글, 바이두)이 상대적으로 약세인 일본은 인공지능 분야에서 선도적인 입지를 다지기 위해 컴퓨팅 인프라를 우선적으로 지원하겠다는 것이다.

현재 우리나라 역시 인공지능과 컴퓨팅 인프라의 중요성을 인지하고 관련된 정부차원의 지원정책을 펼치고 있으나, 그 규모가 미흡한 것이 현실이다. 또한 인공지능을 활용해 제품이나 서비스를 개발하는 기업이 얼마나 많은 수요를 가지고 있는지가 불분명하기 때문에 적정한 지원 규모를 산정하기 어렵다는 점도 있다.

이번 보고서에서는 먼저 다양한 관점에서 인공지능과 컴퓨팅 인프라의 중요성을 설명하고자 한다. 이어 인공지능 관련 중소기업과 스타트업의 인공지능 컴퓨팅 인프라 현황과 향후 수요를 파악하는 실태조사 결과를 분석한다. 총 215개의 설문 대상에서 유효 응답 기업은 72개이며, 이를 토대로 인공지능 컴퓨팅 인프라 현황과 향후 수요를 분석하겠다.

제2절 연구 방법 및 범위

인공지능 컴퓨팅 인프라의 중요성은 고성능컴퓨팅(High Performance Computing, HPC)분야를 위주로 분석했다. HPC는 대규모 과학계산을 수행하기 위한 컴퓨팅 기술로, 일반적으로 슈퍼컴퓨터를 활용하는 컴퓨팅 기술을 일컫는다. 현대 인공지능 컴퓨팅 인프라에서 가장 주목받는 연산처리장치는 GPU다. GPU는 수천개의 사칙논리연산유닛(Arithmetic Logical Units, ALU)를 활용하여 계산하는 매니코어 시스템이다. 따라서 고성능 병렬처리가 필수이며 이는 필연적으로 HPC분야라고 볼 수 있다. 또한 게임 인공지능 분야의 사례를 심층 분석하여 인공지능 컴퓨팅 인프라의 중요성을 강조했다.

이어 인공지능 관련 중소기업·스타트업의 인공지능 컴퓨팅 인프라 실태조사 결과를 분석했다. 실태조사의 주된 목적은 중소기업·스타트업의 인공지능 컴퓨팅 인프라의 현황과 향후수요가 얼마나 되는지에 대한 것을 파악하기 위함에 있다. 이를 토대로 인공지능 인프라 지원정책의 현실적인 규모를 산정할 것이다. 또한 인공지능 컴퓨팅 인프라와 직·간접적으로 연관된 설문과 인공지능 관련 일반 설문도 진행했다. 이번 보고서에서는 주요 설문문항에 대해 문항별 결과를 분석할 것이다.

제2장 인공지능 컴퓨팅 인프라의 중요성

제1절 개 요

인공지능의 부침의 역사를 살펴보면, 컴퓨팅 파워가 부족함으로 인해 암흑기를 맞이했던 과거가 있다. 1980년대 인공신경망 기술은 인간의 뇌 신경망을 모사하여 지능적 행동을 구현한다는 점에 있어서 인공지능의 두 번째 황금기를 이끄는 원동력이 됐다. 인공신경망 기술은 비선형적인 현상을 분류할 수 있는 기능이 이론적으로 입증됐으나, 이를 실제로 구현할 수 있는 계산능력이 매우 부족했다. 또한 데이터의 부족과 학습 방법의 일관성 결여로 인해 인공지능의 암흑기(AI winter)를 초래하기에 이른다.

그러나 컴퓨팅 파워는 무어의 법칙[2]에 의해 지수적인 성장을 하게 된다. 이러한 폭발적인 성장에 힘입어 인공신경망 기술 역시 점차 구현 가능한 영역으로 조명 받게 되기에 이른다. 무어의 법칙에 따르면 지난 20년 동안 5,000배에 가까운 성능이 향상됐다. 무어의 법칙이 법칙으로서 정립되기 까지는 사실 순탄치 않은 여정이었다. 초기의 연산처리장치의 발전은 클럭스피드를 높이는 것으로 진행됐다. 클럭스피드가 3GHz를 넘어서자 발열이 지나치게 증가하여 냉각을 위해 많은 전력을 요구했다. 전력대비 성능의 효율이 급격하게 악화되자, 연산처리장치 생산 기업들은 연산처리장치를 병렬로 탑재하기 시작한다. 현재의 쿼드코어(4개의 CPU)나 옥타코어(8개의 CPU)는 연산처리장치를 병렬화한 것을 의미한다. 이로써 무어의 법칙은 명맥을 이어 갔으나 연산처리장치를 활용하는 측면에서 어려움이 발생하게 된다. 바로 알고리즘의 병렬화에 대한 필요성이다.

알고리즘의 병렬화는 HPC 분야의 전통적인 연구분야다. 특히 슈퍼컴퓨터는 여러 대의 고성능 컴퓨터를 고속 통신망으로 연결하여 활용하기 때문에, 알고리즘의 병렬화가 필수적이다. 기본적으로 슈퍼컴퓨터의 활용에 의해서 발전한 HPC 분야는 HW의 성장과 함께 발전했다. 연산처리장치의 지수적인 성장은 과거에 해결할 수 없었던 문제에 도전할 수 있는 원동력이 됐다. 큰 맥락에서 인공지능 역시 마찬가지라고 볼 수 있다.

---

2) 매 18개월 마다 연산처리장치의 성능이 2배 향상된다는 법칙으로 연산처리장치를 구성하는 트랜지스터의 공정이 18개월 마다 축소되는 것을 의미 → 동일한 면적의 집적도가 향상

그렇다면 알고리즘의 병렬화에 대해 조금 더 살펴볼 필요가 있다. 통상적으로 알고리즘이라고 하면 명령을 순차적으로 처리하는 순차 알고리즘을 의미한다. 순차 알고리즘의 전부 또는 일부를 병렬화 하는 것을 병렬 알고리즘이라고 하는데, 모든 순차 알고리즘의 병렬화 가능한 것은 아니다. 예를 들면, 수학 점화식이 있다. 점화식의 n차 항을 구하기 위해서는 초기 항부터 n-1항의 정보가 필요하다. 이 경우 병렬화가 불가능하다. 현재의 정보가 이전 정보에 의존하여 변경되는 순차 알고리즘의 경우 병렬화가 거의 불가능하다고 볼 수 있다. 병렬화가 불가능하다는 요인은 현재의 멀티코어, 매니코어 시스템을 십분 활용하는데 주요한 병목이 발생한다. 결국 지수적으로 발전한 연산처리장치를 활용하는데 전제돼야 하는 조건은 병렬화 가능성이다.

결론적으로 심층학습 인공지능 알고리즘은 병렬화 가능성이 매우 높다. 먼저 경험적으로 많은 경우를 수행하고 최적의 결과를 도출한다는 점에서 태스크 수준의 병렬화(task-level parallelism)가 용이하다. 또한 심층학습의 세부 알고리즘은 병렬화가 효율이 가장 좋은 것으로 구성된 점 역시 현대 HW의 성능을 십분 활용 가능하다. 비단 CPU 뿐만 아니라 GPU를 비롯해 거의 모든 현대 연산처리장치에서 최대의 성능을 도출할 수 있다는 것이다.

이번 장에서는 인공지능, 특히 심층학습 영역에서 왜 컴퓨팅 인프라가 핵심적인 역할을 하는지에 대해 기술할 것이다. 이어지는 2절에서는 인공지능과 컴퓨팅 환경에 대해 소개하고, 3절에서는 인공지능과 GPU 컴퓨팅, 4절에서는 게임 인공지능 사례를 중심으로 인프라의 중요성을 설명할 것이다.

제2절 인공지능과 컴퓨팅 환경

1. 현대 연산처리장치 개요

연산처리장치의 성능은 초당 부동소수점연산수(FLoating Point Operation per Second, FLOPS)로 측정한다. 부동소수점은 일반적으로 IEEE-754 표준에 의해 단정밀도(32bit, single precision), 배정밀도(64bit, double precision)로 구분된다. 보편적으로 슈퍼컴퓨터의 연산능력을 표현하는 지표로는 배정밀도를 사용한다. 현재 세계에서 가장 빠른 슈퍼컴퓨터인 중국의 선웨이 타이후라이트는 93.0 PFLOPS(페타플롭스, 초당 1천 조 연산)의 성능을 보유하고 있다. 통상적인 CPU나 GPU의 경우에는 GFLOPS(기가플롭스, 초당 1조)나 TFLOPS(테라플롭스, 초당 10억)를 활용한다. 부동소수점연산수는 프로그램의 덧셈과 곱셉 연산의 수로 측정된다. 뺄셈은 음수의 덧셈으로 대체가 가능하다. 나눗셈의 경우는 여러 번의 덧셈과 곱셈으로 구성되기 때문에, 나눗셈을 하나의 연산으로 보기는 어렵다. 예를 들어 100차원의 두 벡터를 내적하는 연산은 덧셈과 곱셈이 각각 100번 수행되는 것으로 총 연산량은 200 FLOP이라고 볼 수 있다. 수치적인 연산의 특징을 살펴보면, 덧셈과 곱셈이 동시에 이루어지는 경우가 많으므로 특정 연산 처리장치는 덧셈과 곱셈을 한꺼번에 처리하는 기능을 탑재하고 있다.

현대 연산처리장치의 눈부신 발전은 손안의 작은 컴퓨터로 인식되는 스마트폰과 과거의 슈퍼컴퓨터의 비교에서도 찾아볼 수 있다. 스마트폰의 두뇌는 어플리케이션 연산처리장치(Application Processor, AP)로 저전력 고성능의 특징을 가지고 있다. 스마트폰은 고해상도 사진 촬영 및 영상 재생, 물체 인식 등 그래픽 처리를 위한 GPU도 탑재하고 있다. 2016년 출시된 갤럭시 S7의 GPU Mali-T880 MP12는 265.2 GFLOPS의 이론 성능을 보유하고 있다. 이 수치는 지난 1996년 세계에서 가장 빠른 슈퍼컴퓨터로 등극한 동경대의 Hitachi SR2201/1024의 실측 성능(220.4 GFLOPS)에 육박한다 [그림 2-1].

앞서 기술했듯이 일반적인 수치적인 계산이 가능한 대표적인 연산처리장치에는 CPU와 GPU가 있다. 고성능 CPU의 경우는 22개의 코어와 1.5 TFLOPS의 이론 성능을 보유하고 있다. CPU는 비단 계산뿐만 아니라 다양한 작업을 처리해야하기 때문에 기본적으로 높은 범용성을 갖는다. 이로 인해 순수한 계산 성능은 동일가격 GPU에 대비하여 약 10배 정도 낮다 [그림 2-2].

[그림 2-1] 삼성 갤럭시 S7에 탑재된 GPU의 이론 성능과 슈퍼컴퓨터의 비교

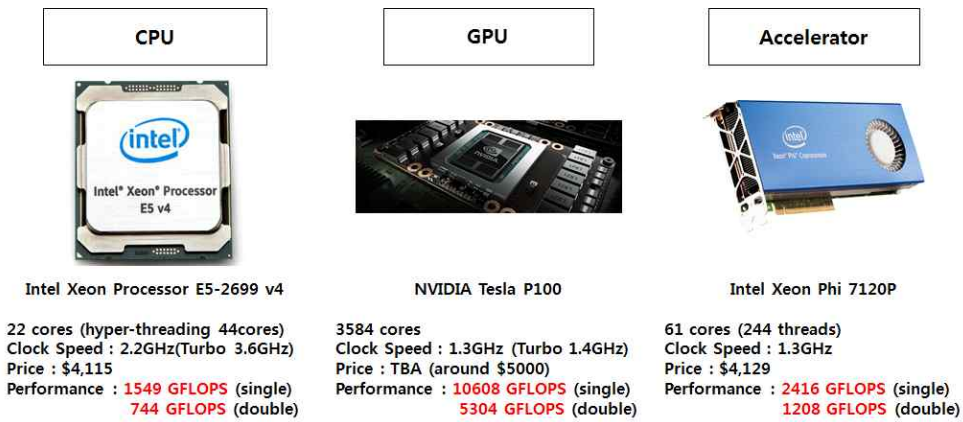
자료 : top500.org, Performance Development https://www.top500.org/statistics/perfdevel/
Samsung Exynos, Wikipedia https://en.wikipedia.org/wiki/Exynos

GPU는 본질적으로 모니터에 그래픽을 출력하는 역할을 한다. 개인용 컴퓨터가 본격적으로 보급된 이후 PC 게임시장이 급격하게 커지면서 화려한 3D 그래픽 처리를 위해 고성능 GPU를 개발하기에 이른다. GPU는 CPU와 다르게 그래픽 용도의 목적이 확고했기 때문에, CPU와 같이 높은 범용성은 필요 없다. 따라서 GPU는 고속 처리를 위한 연산 기능을 극대화하는 방향으로 발전했다. 2006년 슈퍼컴퓨팅 학회(Supercomputing Conference)에서는 GPU를 활용해 과학계산을 시도한 사례가 처음 소개됐다. 2008년에는 세계 최대의 GPU 벤더인 NVIDIA가 GPU에서 프로그래밍 가능한 CUDA(Compute Unified Device Architecture) 툴킷을 공개한다. 이로써 GPU를 과학계산에 접목하는 연구가 활발히 이루어졌으며, 슈퍼컴퓨터에 탑재되기에 이르러 현대 연산처리장치의 핵심적인 역할을 수행하고 있다. GPU가 괄목할 만한 성공에 이르자 가속기 형태의 계산자원을 활용하기 위한 프레임워크인 OpenCL(Open Computing Language)도 공개됐다.

GPU는 구조상 SIMD(Single Instruction Multiple Data) 형식의 병렬화가 가장 일반적이고 효율이 높다. SIMD는 하나의 연산을 수행하는 다수의 데이터를 동시에 처리하는 데이터수준 병렬화(Data-level parallelism)를 뜻한다. 예를 들면, 100차원의 두 벡터의 덧셈은 100개의 서로 다른 데이터가 덧셈이라는 하나의 연산으로 구성되기 때문에 SIMD 방식으로 구현이 가능하다.

CPU와 GPU이외의 연산처리장치는 크게 가속기(Accelerator)의 영역에 속한다. GPU 역시 일반적인 컴퓨터 시스템에서 계산을 가속한다는 관점에서 가속기로 취급할 수 있다. 가속기는 크게 두 가지 형태가 있다. 첫 번째는 Intel에서 개발한 Xeon Phi다. Xeon Phi는 그간 Intel이 개발했던 Atom이나 Pentium 프로세서를 병렬로 집적해 제작한 가속기다. Xeon Phi의 가장 큰 장점은 멀티코어 프로세서에서 작성한 병렬 프로그램을 손쉽게 이식할 수 있다는 것이다. Xeon Phi는 Intel 연산처리장치와 동일한 x86계열의 명령어 집합을 활용하기 때문이다. 이러한 장점으로 인해 전통적인 CPU기반의 슈퍼컴퓨터를 활용하는 영역에서 많은 수요가 있다.

두 번째는 FPGA(Field Programmable Gate Array)다. FPGA를 쉽게 표현하자면 프로그래밍 가능한 연산처리장치라고 볼 수 있다. 따라서 특정 연산에 최적화 가능한 장점이 있다. 더불어 이 장점은 저전력의 특징과도 맞물린다. 그러나 도입에 필요한 높은 비용, 어셈블리어 수준의 복잡한 프로그래밍 등으로 인해 진입장벽이 매우 높다. 최근 OpenCL을 활용해 프로그래밍 장벽을 낮추려는 시도가 있었으나, 최적화에는 여전히 많은 시간을 필요로 한다.



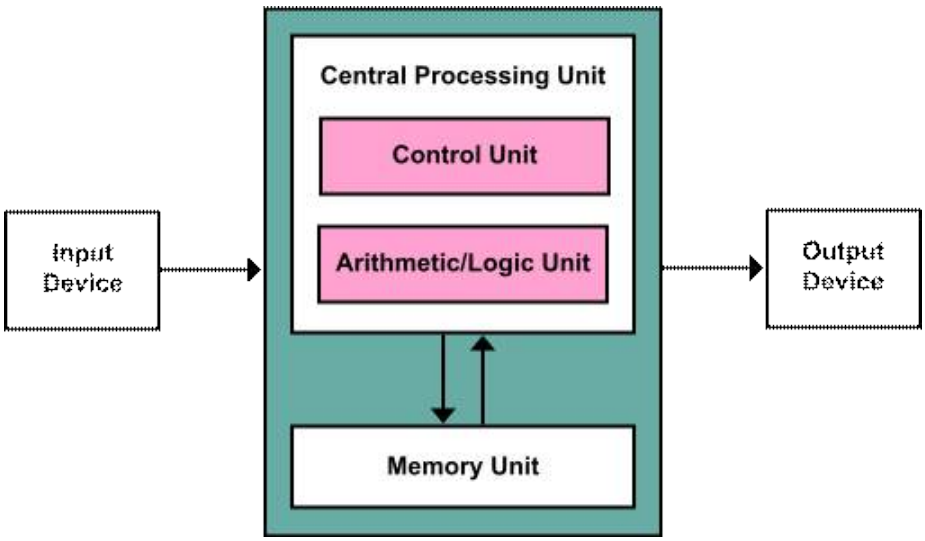
[그림 2-2] 현대 연산처리장치의 성능

자료 : Intel Xeon Processor E5-2699 v4, http://ark.intel.com/products/91317
Nvidia Tesla P100, http://www.nvidia.com/object/tesla-p100.html Intel Xeon Phi 7120P, http://ark.intel.com/products/75799

2. 현대 연산처리장치의 활용 측면에서의 이슈

현대 연산처리장치의 성능을 판단하는 기준은 이론 성능과 실측 성능이 있다. 이론 성능은 말 그대로 연산처리장치의 HW 구성에 따라서 달성할 수 있는 최대 성능을 나타낸다. 실측 성능은 실제로 수치적인 연산이나 응용프로그램을 수행했을 때 나타내는 수치를 말한다. 슈퍼컴퓨터의 성능 역시 이론 성능과 실측 성능으로 측정된다. top500.org에서는 매년 2회(6월과 11월) 1위부터 500위까지의 슈퍼컴퓨터 순위를 공개하는데 그 기준은 실측성능이다. 보편적으로 슈퍼컴퓨터급 시스템에서는 실측 성능으로 이론 성능의 70%이상을 달성한다. 이것을 달성하기 위해서는 고속 통신망 구축, 소프트웨어 최적화 등의 HPC 기술이 필요하다. 슈퍼컴퓨터의 실측 성능은 전통적으로 대규모 선형대수 프로그램(High Performance Linpack[3])으로 측정한다. 선형대수 프로그램은 과학 문제를 해결하기 위해 가장 보편적이고 필수적으로 활용된다.

그렇다면 이론 성능과 실측 성능의 차이에 대해서 더 깊게 살펴볼 필요가 있다. 현대 연산처리장치가 작동하는 방식은 폰노이만 아키텍처로부터 시작한다 [그림 2-3]. 연산처리장치의 성능은 크게 두 가지 요소로 좌우된다. 첫 번째는 연산처리장치 자체의 계산 성능이다. 두 번째는 연산처리장치에서 연산을 수행하기 위해 정보를 전송하기 위한 메모리 관련 장치의 전송 성능이다.

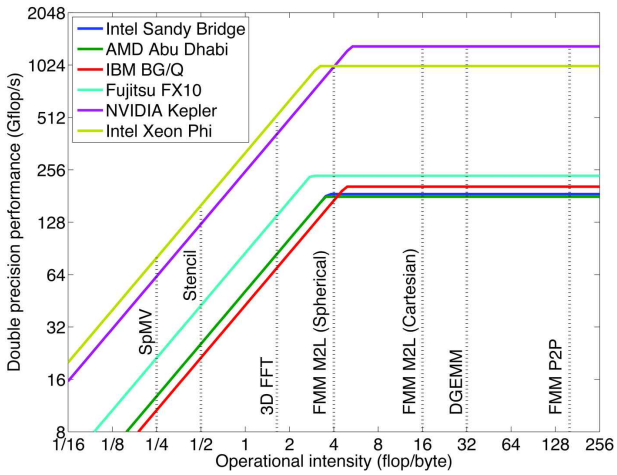


[그림 2-3] 폰노이만 연산처리장치의 개념도
자료: Von Neumann Architecture, Wikipedia. https://en.wikipedia.org/wiki/Von_Neumann_architecture

3) Linear Algebra Package의 약자로 밀집행렬 연산으로 구성

과거 연산처리장치는 메모리 전송 속도는 빠르나 연산처리장치의 계산 능력은 상대적으로 느렸다. 그러나 연산처리장치의 성능은 무어의 법칙에 의해 지수적으로 성장한 반면, 메모리 전송 장치의 성능은 선형적으로 향상됐다. 이로 인해 현대 연산처리장치는 계산 성능은 매우 탁월하나 계산을 위한 메모리 전송 속도는 느려지는 경향성을 갖는다. 이러한 성능의 불균형으로 인해 알고리즘의 계산 집적도에 따라 계산의 효율이 극명하게 나뉜다.

연산강도(Arithmetic Intensity)는 알고리즘의 총 연산수와 메모리 전송량의 비율로 결정된다. 계산강도가 높을수록 전송된 메모리를 재사용하여 계산하는 것으로 분석할 수 있다. 연산강도와 연산처리장치의 성능은 루프라인(Roofline) 모델로 표현할 수 있다. 여기서 루프(Roof)의 의미는 연산처리장치가 HW적으로 달성할 수 있는 최대 이론 성능이다. 앞서 기술했다시피 연산강도가 낮으면 연산처리장치의 최대성능을 사용하지 못하기 때문에, 연산강도가 높아질수록 선형적으로 그 성능이 증가한다. 일정 연산강도 이상에서는 이론 성능을 달성할 수 있기 때문에 상수로 수렴하게 된다. 여기서 선형적 증가가 상수로 변하는 시점은 해당 연산처리장치의 메모리 대역폭과 연산성능의 비율로 결정된다. [그림 2-4]는 다양한 연산처리장치와 알고리즘에 대한 루프라인 모델이다.



[그림 2-4] 다양한 계산자원의 루프라인

자료: How will the fast multipole method fare in the exascale era? (2013.07)

[알고리즘 설명]

SpMV : 희소행렬(Sparse Matrix)과 벡터의 곱

Stencil : 편미분방정식의 수치해법에 적용되는 연산

3D FFT : 3차원 고속 푸리에 변환(Fast Fourier Transformation)

FMM : 고속 멀티폴 방법(Fast Multipole Method) - n-body 연산에 활용

DGEMM : 배정밀도 행렬곱 연산(Double-precision GEneralized Matrix Multiplication)

정리하자면 루프라인 모델에서 선형적으로 증가하는 부분은 메모리 전송에 소요되는 시간 동안 연산처리장치가 유휴상태(idle)에 처한다고 볼 수 있다. 루프라인 모델에서의 가장 큰 가정은 알고리즘 자체가 병렬화 가능한 것을 전제로 한다.

루프라인 모델을 다른 시각에서 한 번 더 살펴보면 우리가 간과하기 쉬운 내용을 짚어볼 수 있다. 시중에 가장 보편적으로 판매되고 있는 쿼드코어(4개의 연산처리장치) CPU를 예로 들어보자. 예를 들어 12차원의 두 벡터의 원소 별 합을 구하는 알고리즘을 생각해보자. 이것은 곧 12번의 덧셈이고, 서로 독립적인 연산이기 때문에 병렬화가 가능한 알고리즘이다. 12번의 덧셈을 4개의 코어에 3번의 덧셈을 각각 할당하여 계산하는 상황을 가정해보면, 1개의 코어에서 12번의 덧셈을 수행하는 시간보다 4배 빠른 결과가 나올 것이라고 판단할 수 있다. 그러나 이것은 사실과 다르다. 현대 연산처리장치에서는 오히려 1개의 코어를 사용하나 4개의 코어를 사용하나 소요되는 시간은 비슷하게 측정된다. 12차원의 벡터가 아닌 12만 차원의 벡터를 적용해도 동일한 결과가 나온다. 그 이유는 연산강도를 계산해보면 명확해진다. 두 벡터의 합을 구하는 알고리즘의 연산강도는 1/16[4]으로 고정되기 때문이다. 이 수치를 [그림 2-4]에 대입해 보면 거의 모든 계산자원에서 상당히 낮은 수준의 성능을 나타낸다. 따라서 성능 향상은 거의 기대할 수 없는 것이다. 위 예제를 표로 정리하면 <표 2-1>과 같다.

<표 2-1> 병렬 계산의 효율성

Q. 12개의 원소를 갖는 두 벡터의 합을 구하는 연산이 있다고 가정하자. 만약 4개의 코어를 갖는 연산처리장치를 활용한다면 단일 코어 대비 4배의 성능향상을 달성할 수 있는가?
A. 반드시 4배의 성능향상을 보장할 수 없다. 위 연산의 연산강도를 계산해보면 약 1/16으로, [그림 2-4]에서 추정할 수 있듯이 메모리전송량에 의해 최대 성능을 달성하기 어렵다.

4) 위 문제를 수식으로 나타내면 $z = x + y$, 여기서 x, y, z는 12개의 원소를 갖는 벡터임. 12개의 원소를 갖는 실수형 벡터의 메모리 크기는 4byte×12 = 48byte. 계산에 필요한 벡터 x, y와 결과 값을 저장하는 벡터 z가 계산을 위해 연산처리장치의 레지스터에 전송되고, 결과 값 z를 다시 내려 받아야 하기 때문에 4번의 12차원 벡터 전송이 발생. 따라서 메모리전송량은 총 192byte인 반면 계산량은 12번이기 때문에 연산강도는 1/16

제3절 인공지능과 GPU 컴퓨팅

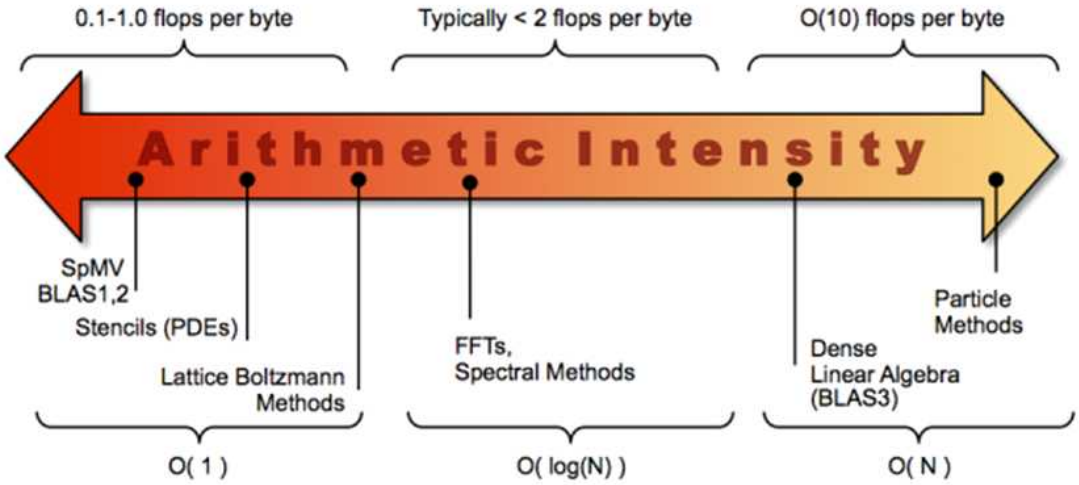
GPU는 본질적으로 3D 게임이나 CAD와 같은 그래픽 관련 작업에 최적화된 연산처리장치이다. 따라서 CPU와는 다르게 범용성이 낮다. 특히 GPU는 수 천 개의 사칙논리연산 유닛을 탑재한 매니코어 시스템으로, 병렬처리가 필수적이다. 앞서 소개했듯이 GPU는 2009년부터 본격적으로 과학계산에 활용됐다. 같은 가격의 CPU대비 10배 이상의 계산 성능을 보유한 GPU는 차세대 연산처리장치로 큰 주목을 받아왔다. 특히 많은 계산을 필요로 하는 컴퓨터 시뮬레이션 분야, 나아가 HPC 시스템을 활용하는 슈퍼컴퓨터의 영역에서 GPU를 활용하는 연구가 활발히 진행됐다.

그러나 GPU는 학계의 기대와는 다르게 매우 제한된 영역에서만 활용됐다. 이것은 GPU의 구조적 특징에서 기인한 것으로 GPU에 최적화된 소수의 알고리즘만이 같은 가격의 CPU대비 10배 이상의 성능을 확보할 수 있었다. 결국 CPU에서도 병렬화가 어렵거나 성능이 보장되지 않는 알고리즘은 GPU에서 역시 성능향상을 기대할 수 없다는 것이다. 물론 GPU의 높은 계산 성능을 적극적으로 활용하고자 하는 노력도 있었다. 하지만 HW 아키텍처에 기반하는 최적화로 인해 SW 개발에 소요되는 비용이 매우 컸다.

GPU의 구조적 특징에 대해 더 자세히 살펴보자. GPU는 일반적으로 컴퓨터의 주변기기에 해당되어 일반적으로 PCIe(Peripheral Component Interconnect Express)에 연결된다. PCIe는 GPU를 활용함에 있어 가장 먼저 직면하는 병목이다. PCIe의 메모리 대역폭은 GPU 메모리의 대역폭보다 수십 배 이상 차이가 나기 때문이다. 또한 GPU 메모리 대역폭에 비해 GPU의 연산능력이 수 배 높기 때문에 상당히 높은 연산강도를 갖는 알고리즘이 아니라면 최대 성능을 달성하기 더욱 어려운 요인을 가지고 있다. 정리하자면 GPU는 연산능력이 매우 뛰어나나 이를 보장해주는 메모리 대역폭이 상대적으로 낮은 이유로 활용에 있어 매우 제한적인 요소로 작동한다.

GPU는 메모리 전송 속도와 계산 성능의 지나친 불균형을 해소하기 위해서 메모리 전송 속도를 향상시키는 전략을 꾀하고 있다. 고대역폭 메모리(High Bandwidth Memory, HBM) 등을 활용하여 PCIe 병목을 줄이는 접근을 적용하고 있다.

[그림 2-5]는 알고리즘별로 다양한 연산강도를 나타낸다. GPU에서 거의 이론치에 가까운 성능을 내는 알고리즘은 밀집행렬 연산이나 n-body 시뮬레이션 등이 있다. 희소행렬-벡터곱은 사칙연산보다 메모리 전송에 소요되는 시간이 높기 때문에 메모리 대역폭에 의한 성능으로 제한된다.



[그림 2-5] 다양한 알고리즘의 연산강도

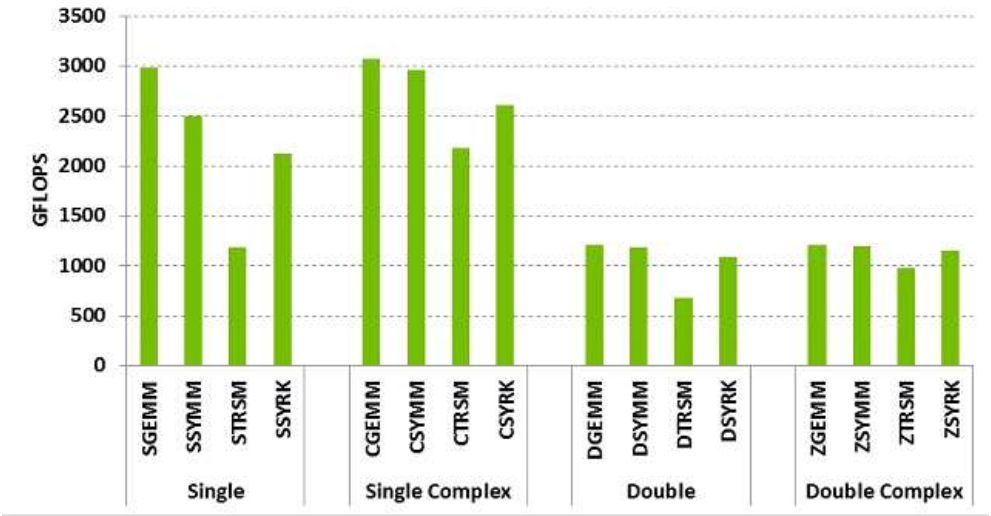
자료: Roofline Performance Model, https://crd.lbl.gov/departments/computer-science/PAR/research/roofline/

정리하자면 현대 연산처리장치를 십분 활용하기 위한 전제조건은 매우 많다. 먼저 현대 연산처리장치는 연산처리장치의 병렬화, 연산성능과 메모리 전송속도의 불균형이라는 큰 특징을 가지고 있다. 따라서 특정한 알고리즘이 연산처리장치의 최대 성능을 활용하기 위해서는 높은 병렬화 가능성과 연산강도를 보유해야한다. 이 두 가지 요소로 인해 연산처리장치의 활용측면이 상당부분 제한되는 것이다. 혹은 기존의 순차 알고리즘을 연산처리장치의 HW 구조에 맞게 재설계하는 것도 방법이다.

심층학습으로 대두되는 현대 인공지능은 결론적으로 연산강도가 높고 병렬화 가능성이 높은 알고리즘으로 구성돼있다. 따라서 GPU의 성능을 십분 활용할 수 있는 응용분야로 각광받음에 따라, HPC분야에서도 인공지능의 가능성에 대해 주목하고 있다. 심층학습의 대표적인 학습 방법인 오류역전파법은 세부적으로 기초선형대수루틴(Basic Linear Algebra Subprogram, BLAS)으로 구성된다. BLAS는 크게 세 가지 수준으로 분류된다. BLAS-1은 벡터-벡터 연산, BLAS-2는 행렬-벡터 연산, BLAS-3은 행렬-행렬 연산으로 구분된다. 연산강도를 살펴보면

[그림 2-5]와 같이 BLAS-1과 BLAS-2는 낮은 연산강도를 보유한 반면 BLAS-3는 매우 높은 연산강도를 갖고 있다. 인공지능 학습 알고리즘의 대부분이 행렬 연산이 BLAS-3에 속하기 때문에 연산처리장치의 성능을 최대한 사용할 수 있다.

사실 BLAS는 병렬처리를 필수로 요구하는 슈퍼컴퓨팅 분야에서 지속적으로 발전해왔다. 수치해석적인 연산을 위해 BLAS는 가장 기저에 있는 역할을 하기 때문이다. NVIDIA는 GPU 전용 BLAS인 cuBLAS[5]를 개발하여 자사 GPU에 최적화된 성능을 보여준다. [그림 2-6]은 NVIDIA K40m 모델을 활용하여 정밀도별 cuBLAS의 연산 성능이다.



[그림 2-6] 유효숫자별 cuBLAS의 성능 실측치

자료 : cuBLAS Performance https://developer.nvidia.com/cublas

NVIDIA는 cuBLAS를 기반으로 cuDNN(cuda Deep Neural Network)[6]이라는 심층학습 라이브러리를 제공한다. 현재 NVIDIA GPU는 인공지능 학계에서 가장 널리 활용되고 있으며, cuDNN은 다양한 인공지능 공개SW와 연계되어 GPU 활용에 대한 진입장벽을 낮췄다.

5) cuBLAS, NVIDIA, http://docs.nvidia.com/cuda/cublas/
6) cuDNN : Efficient Primitives for Deep Learning, https://arxiv.org/pdf/1410.0759.pdf

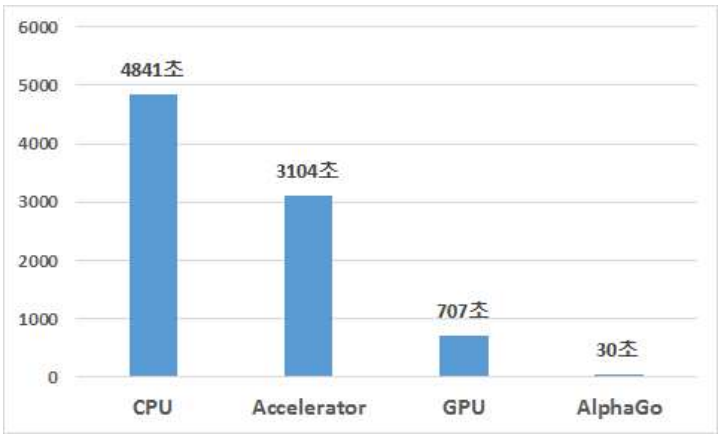
제4절 인공지능 사례 분석 - 게임 인공지능을 중심으로

지금까지 현대 연산처리장치의 지수적인 성장과 그 방향, 그리고 이에 따른 활용측면에서의 제약사항, 인공지능 학습 알고리즘과 GPU를 설명했다. 정리하자면 인공지능 학습 알고리즘은 현대 연산처리장치를 십분 활용할 수 있기 때문에 특히 GPU를 많이 활용하고 있다. 그렇다면 인공지능의 계산량 측면은 어떨까? 인공지능은 빅데이터를 학습한다는 점에서 일견 많은 계산량이 필요할 것으로 추정된다. 또한 그간 역사적인 대결을 수행했던 게임 인공지능 역시 슈퍼컴퓨터급 자원을 활용한 결과였다. 이러한 사실을 바탕으로 할 때 인공지능에 필요한 컴퓨팅 인프라는 상당히 많을 것이라고 추정이 가능하다. 이번 절에서는 구체적인 게임 인공지능 사례를 바탕으로 인공지능의 계산량에 대해 분석할 것이다.

2016년 1월 바둑 인공지능으로는 최초로 프로 바둑기사와 대결하여 승리한 AlphaGo는 심층학습을 활용했다. AlphaGo는 두 가지 형태의 인공신경망을 활용했다. 정책망(Policy Network)은 전문 바둑기사의 착수 선호도를 학습한 인공신경망이다. 특정 바둑판 상태가 정책 네트워크에 입력으로 들어가게 되면, 출력으로 규칙상 착수 가능한 모든 수에 대해서 착수 선호도(확률 값)가 산출된다. AlphaGo는 이를 위해 16만 개의 바둑 기보에서 2,940만 개의 바둑판 상태를 추출하여 학습에 활용했다. 정책망의 기능은 게임 트리의 폭을 줄이는 용도로, 전문 바둑기사의 관점에서 착수할 지점을 선별하는 역할을 한다. 가치망(Value Network)은 현재 바둑판 상태에서 승리할 근사한 확률을 찾아내는 역할을 한다. 바둑은 거의 무한대 경우의 수를 가지고 있기 때문에, 특정 바둑판 상태의 승률을 정확히 예측하기는 어렵다. 따라서 승률을 예측하기 위한 접근을 취할 수밖에 없다. AlphaGo는 가치망을 학습하기 위해 자체 대국으로 생성된 3,000만 개의 바둑 판 상태를 활용했다. 가치망의 기능은 게임 트리의 깊이를 줄이는 역할을 한다.

AlphaGo의 인공신경망은 한 번 계산하는데 30 GFLOP의 연산량을 필요로 한다. 특히 AlphaGo의 컴퓨팅 환경은 바둑 게임의 30초 초읽기 동안 약 10만 개의 경우의 수를 탐색하기 때문에, 30초 동안 계산해야하는 연산수는 6 PFLOP (6천조 번 연산)[7]이다.

[그림 2-7]은 다양한 연산처리장치에서 6 PFLOP을 처리하는데 소요되는 시간을 표현한다. 일반적인 PC에서 6 PFLOP을 처리하기 위해 필요한 시간은 약 80분으로 AlphaGo 시스템에 비해 약 160배나 많은 시간이 소요된다. [그림 2-7]의 계산자원은 [그림 2-2]에서 소개한 연산처리장치를 대상으로 소요시간을 추정했다. 여기서 계산자원의 성능은 이론성능의 80%[8]로 계산했으며, AlphaGo의 컴퓨팅 환경은 GPU 176개를 활용했다고 가정한다.



[그림 2-7] 계산자원별 소비시간 비교

자료: 인공지능의 핵심 인프라 - 고성능컴퓨팅 환경의 중요성, 이슈리포트 2016-013, 소프트웨어정책연구소(2016)

AlphaGo이전의 게임 인공지능 역시 슈퍼컴퓨터급 컴퓨팅 환경을 활용한 결과다. 슈퍼컴퓨터의 성능 순위는 매년 6월과 11월에 1위부터 500위 까지 top500.org에서 집계한다. 인간과 대결해 승리한 게임 인공지능의 컴퓨팅 환경을 살펴보면 <표 2-2>와 같이 약 2~300위 수준의 슈퍼컴퓨터가 활용됐다.

<표 2-2> 게임 인공지능과 컴퓨팅 환경

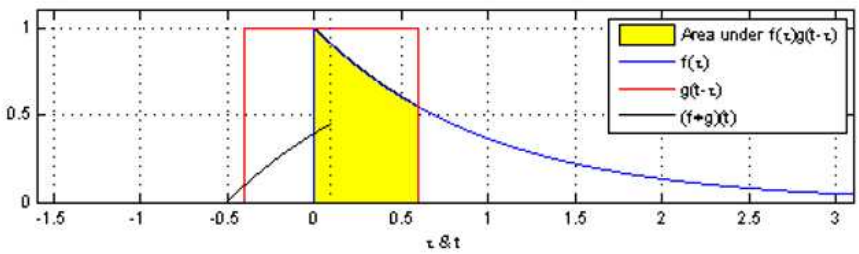
	IBM Deep Blue	IBM Watson	Google AlphaGo[9]
연산처리장치	30 노드 POWER2SC (120Hz) VLSI 체스칩 480개	90 노드 POWER7(3.5GHz, 8코어, 32쓰레드)	약 35~40노드 CPU 1,920 코어 GPU 280 장
연산성능	11.38 GFLOPS	80 TFLOPS	약 300 TFLOPS
top500 순위	259위 (1997 6월)	192위 (2011 6월)	382위 수준 (2015년 11월)

자료: 인공지능의 핵심 인프라 - 고성능컴퓨팅 환경의 중요성, 이슈리포트 2016-013, 소프트웨어정책연구소(2016)

7) 심층학습 알고리즘에 국한된 연산수. 10만 개의 경우의 수를 탐색하기 위해 필요한 심층학습 연산은 20만 번으로 6PFLOP이 필요함.
8) 심층학습은 연산강도가 높은 알고리즘으로 구성되기 때문에 이론성능에 가까운 성능을 보임.
9) AlphaGo의 경우 구글 클라우드를 활용하여 연산을 수행했기 때문에, 정량적으로 정확한 계측이 불가능함

<표 2-2>를 보면 단편적으로 연산처리장치의 경향을 볼 수 있다. CPU의 성능이 제한적이었던 1990년대 후반에는 전용 칩(IBM Deep Blue)을 제작하여 계산에 활용했다. 이후 CPU의 성능이 지수적으로 향상되자 CPU기반의 슈퍼컴퓨터가 등장(IBM Watson)하고, 현대에는 가속기를 탑재한 슈퍼컴퓨터(Google AlphaGo)가 활용됐다. 슈퍼컴퓨터급 컴퓨팅 환경의 수요는 결국 학습에 요구되는 데이터의 양에도 직결된다. IBM Watson은 초당 500 GBYTE[10]에 해당하는 정보를 처리할 수 있으며, AlphaGo는 16만 개의 프로바둑기사의 기보[11]를 한 달 간 학습했다.

이제 구체적으로 AlphaGo의 인공지능 알고리즘에 대해 살펴보자. AlphaGo의 인공신경망은 이미지 인식에 최적화돼 있는 합성곱신경망(Convolutional Neural Network)를 활용했다. Facebook의 얼굴인식 알고리즘(DeepFace) 역시 4백만 장의 얼굴 이미지를 합성곱신경망으로 학습한 결과이다. 합성(Convolution)의 수학적 의미는 두 함수의 합성을 표현한 것으로, 일반적으로는 임의의 두 합수가 겹치는 영역을 적분한 값을 나타낸다 [그림 2-8].

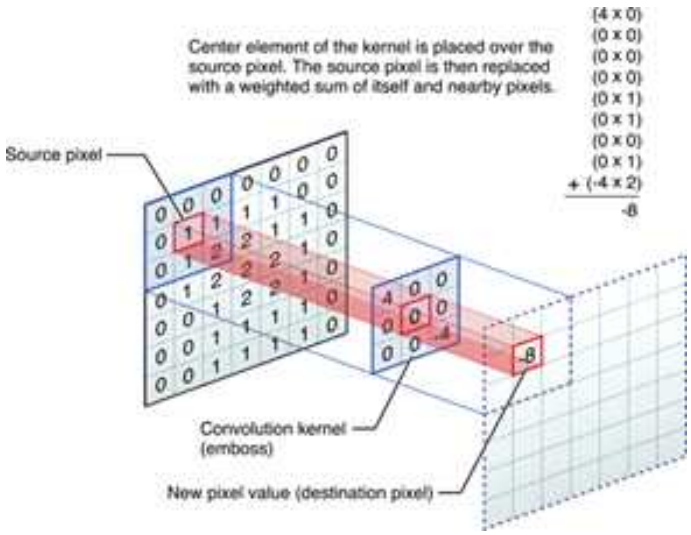


[그림 2-8] 합성(Convolution)의 수학적 모형

자료 : Convolution, https://en.wikipedia.org/wiki/Convolution

이미지에서 합성의 과정은 이미지의 경계를 찾거나(edge detection), 이미지 자체를 흐리게 또는 선명하게 하는데 사용되는 필터를 의미한다. [그림 2-9]는 이미지의 합성 과정을 표현하는데 3x3 행렬로 표현된 합성 필터가 이미지를 이동하며 각 픽셀 값을 곱해서 합하는 연산으로 계산된다. 합성 필터의 역할은 이미지의 특징을 추출하는 것으로, 합성 필터의 값에 따라 부각해서 보고자 하는 관점이 달라진다. 합성곱신경망에서는 바로 합성 필터가 가중치 역할을 하여 학습의 대상이 된다.

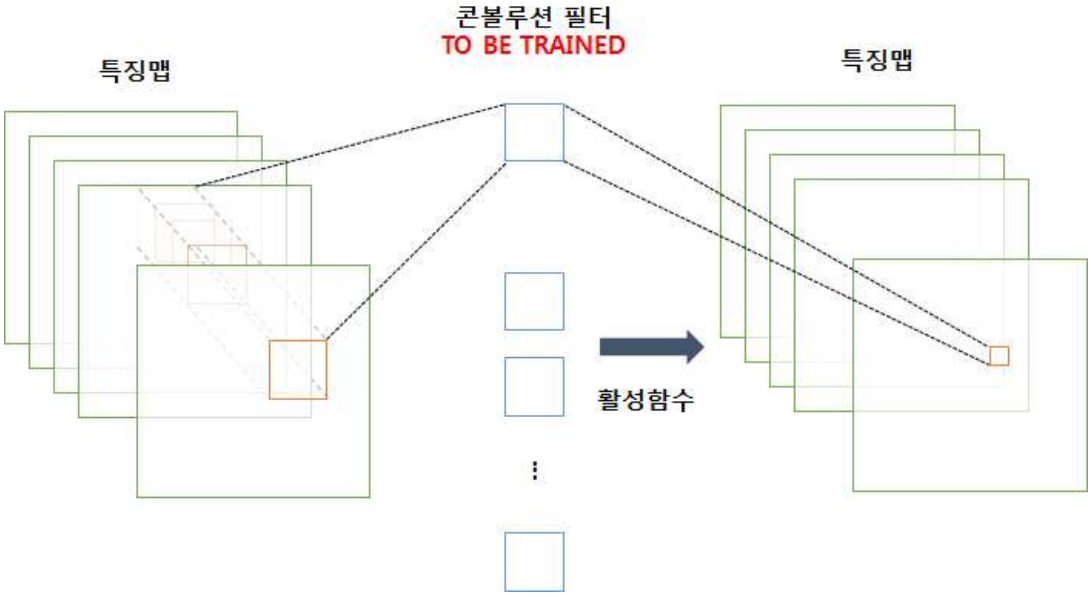
10) 약 백만 권의 책에 해당하는 용량
11) 수치적인 데이터의 용량은 1.85 TBYTE



[그림 2-9] 이미지 합성(Convolution)의 과정

자료 : iOS Developer Library - vImage Programming Guide

[그림 2-10] 합성곱신경망에서 활용되는 합성 필터의 역할을 나타낸다. 여기서 특징 맵 (Feature map)은 이미지의 속성을 나타내는 것으로, 이미지의 가장 대표적인 특징 맵은 RGB(Red, Green, Blue)의 3채널을 활용한다. AlphaGo를 예로 들어보면, AlphaGo에서 활용한 특징 맵은 총 48가지이다. 이것은 백돌, 흑돌, 빈칸, 상수, 활로, 꼬부림, 단수 등의 바둑 정보를 담고 있다.

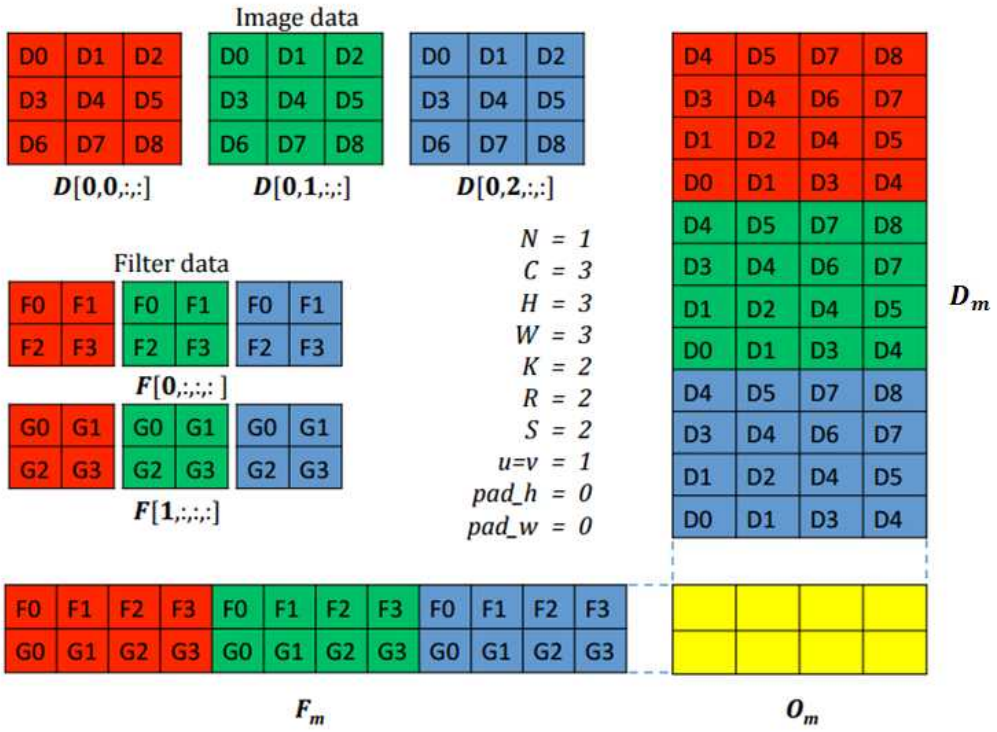


[그림 2-10] 합성곱 층의 개념도

자료: 인공지능의 핵심 인프라 - 고성능컴퓨팅 환경의 중요성, 이슈리포트 2016-013, 소프트웨어정책연구소(2016)

더 세부적으로 합성곱신경망의 과정에 대해 살펴보자. 합성곱신경망에서 이미지에 합성 필터를 적용하는 과정은 먼저 합성 필터와 필터 크기만큼의 이미지의 원소간의 곱을 한 뒤 모두 더하는 것이다 [그림 2-9 참조]. 합성 필터를 행렬이 아닌 벡터 형태로 변환했을 때는 벡터 간의 내적으로 합성 과정을 수행할 수 있다. 합성 필터 벡터와 특정 이미지의 영역 벡터의 내적 결과 값은 곧 다음 특징 맵의 원소가 된다.

이것은 곧 [그림 2-11]과 같이 합성 필터와 이미지를 중복을 허용하여 큰 두 개의 행렬 곱으로 표현할 수 있다. 여기서 이미지는 D, 합성곱 필터는 F라고 볼 때, D_m과 F_m은 행렬을 벡터화 한 것이다. 따라서 이미지 전체에 합성 필터를 적용한다는 것은 D_m과 F_m의 곱으로 표현할 수 있다.



[그림 2-11] 합성곱의 계산 - 행렬 곱

자료 : cuDNN : Efficient Primitives for Deep Learning, https://arxiv.org/pdf/1410.0759.pdf

행렬 곱은 [그림 2-5]의 다양한 연산강도에서 찾아 볼 수 있듯이, 연산강도가 매우 높은 알고리즘에 속한다. 결론적으로 합성곱신경망의 학습(training)이나 추론(inferencing)은 행렬 곱 기반의 연산으로 이루어졌기 때문에, 현대 계산자원의 성능을 최대한 활용할 수 있다.

AlphaGo의 계산량 분석

- AlphaGo의 계산의 대부분은 합성곱신경망의 계산[12)]
  - [그림 9]로 비유해서 표현하자면, 필터 행렬인 F_m의 크기는 192×9, 바둑판기보 행렬에 대응하는 D_m의 크기는 9×69312
  - AlphaGo의 첫 번째 합성곱 층은 19×19 바둑판을 48가지 관점에서 분류한 특징 맵이 입력 값으로 사용[13)]
    * 48가지 관점은 예를 들어 흑 돌의 위치, 백 돌의 위치, 빈 칸의 위치, 활로, 꼬부림 등 이진수로 표현가능
    * 첫 번째 합성 필터의 크기는 5×5이고 총 192개의 서로 다른 필터를 사용하여 다음 합성곱 층에 사용될 192개의 특징 맵을 생성
    * 따라서 48개의 19×19 입력 값에 192개의 5×5 필터와 합성(Convolution) 하는데 필요한 계산량은 약 **4.159 GFLOP**[14)]
  - 두 번째에서 열 세 번째 합성곱 층은 192개의 19×19의 정보가 입력 값이고 필터의 크기와 개수는 각 각 3×3과 192개
    * 첫 번째 합성곱 층과 동일하게 계산할 경우 총 **23.715 GFLOP**[15)]
  - 따라서 특정 바둑판 상태가 입력된 경우 착수 확률분포를 계산하거나 승산을 계산할 때 약 **30 GFLOP의 계산량**이 필요[16)]
    * NVIDIA의 계산전용 그래픽카드인 K40의 cuDNN 실측 성능은 약 1.2 TFLOPS[17)]로 초당 약 40번의 콘볼루션 신경망 계산 가능
- AlphaGo의 합성곱신경망은 정책 네트워크와 가치 네트워크 두 가지로 구성[18)]
  - 정책 네트워크는 프로 바둑기사의 착수 선호도를 예측하는 것으로 특정 바둑판이 입력되면 빈 칸에 대한 착수 확률의 분포를 산출
    * 바둑 게임의 탐색에서 착수 가능한 모든 지점을 고려하지 않고 높은 착수 확률 분포를 갖는 지점을 기준으로 탐색 (게임 트리의 폭을 줄이는 과정)
  - 가치 네트워크는 현재 바둑판 상태의 승산을 근사
    * 예측한 승산이 정확할수록 게임 트리를 깊게 탐색해야하는 소요를 줄일 수 있음 (게임 트리의 깊이를 줄이는 과정)
  - 특정 바둑판 상태에서 성공적인 탐색을 위해서는 정책과 가치네트워크를 모두 계산해야 함
    * 앞서 소개한 합성곱신경망 계산량에 따르면 176개의 GPU[19)]를 활용하면 초당 7,040번의 합성곱신경망의 계산이 가능하나 정책/가치 네트워크를 모두 계산해야 하므로 초당 3,520번의 착수를 계산할 수 있음
  - 따라서 **30초의 초읽기가 주어진다면, 176개의 GPU를 활용하여 약 100,000개의 바둑판 상태에 대하여 착수 확률분포와 승산을 계산**할 수 있음
  - 정책과 가치 네트워크를 학습하는 과정은 위 계산량에 더하여 네트워크의 가중치를 조정하는 오류역전파법에 대한 계산을 고려해야 함

지금까지 AlphaGo에 필요한 연산의 양과 알고리즘적인 특성에 대해 살펴봤다. 정리하자면 이미지 인식에 최적화된 합성곱신경망의 학습 및 추론 과정은 행렬곱 연산으로 대체되어 현대 연산처리장치의 성능을 십분 활용할 수 있다. 또한 합성곱신경망의 구조가 복잡해질수록, 학습에 투입되는 데이터의 양이 많을수록 최종적인 계산량이 증가하기 때문에 결과적으로 굉장히 많은 양의 계산을 필요로 한다.

인공신경망을 활용해 최적의 결과를 도출하는 방법에서 역시 고성능 컴퓨팅 자원을 필요로 한다. 인공신경망의 가장 큰 단점은 입력과 출력의 인과관계를 설명하지 못한다는 점이다. AlphaGo역시 최정상의 바둑실력을 보유하고 있으나 왜 바둑을 잘 두는지에 대해 대답하기 어렵다. 그 이유는 AlphaGo의 성능은 경험적으로 얻어진 결과이기 때문이다. AlphaGo의 합성곱신경망은 13층을 활용했다. 딥마인드 연구진의 이전 논문을 살펴보면 합성곱신경망이 10층인 경우도 테스트를 해봤다. 또한 48개의 특징 맵 중 30개를 사용하는 등 수많은 시도를 했다. 그 결과 최적의 인공신경망의 구조를 도출해낸 것이다. 경험적인 연구결과라는 것은 다른 말로 개선의 여지가 여전히 남아있다고도 볼 수 있다.

따라서 풍부한 계산자원을 활용하여 다양하게 시도를 해보는 것이 무엇보다 중요하다. 인공지능 알고리즘 자체 역시 현대의 계산자원에 특화돼 있지만, 경험적인 결과를 도출하기 위해 많은 시도를 해야 한다는 점 역시 컴퓨팅 인프라의 막대한 수요를 간접적으로 나타낸다. 최근 인간의 정보를 전혀 사용하지 않고 학습한 AlphaGo Zero는 데이터를 스스로 생산했다. 그러나 데이터의 학습에는 여전히 고성능컴퓨팅 인프라가 활용됐다. 결론적으로 인공지능 연구에서 컴퓨팅 인프라는 최상의 결과를 도출하기 위해 필수적인 요건이라고 볼 수 있다.

---

12) Mastering the game of Go with Deep neural networks and tree search, Nature (2016)
13) 자세한 신경망 구조는 다음 보고서 참고. AlphaGo의 인공지능 알고리즘 분석, 소프트웨어정책연구소 (2016)
14) 19×19 바둑판 * 48개 특징맵 * 5×5 필터 * 25 덧셈 (원소끼리 곱한 후 더하는 과정, 내적에서의 덧셈) * 192개 필터 * 2개 연산 (활성함수 계산) = 4.159 Gflop
15) 19×19 바둑판 * 192개 특징맵 * 3×3 필터 * 9 덧셈 (원소끼리 곱한 후 더하는 과정, 내적에서의 덧셈) * 192개 필터 * 2개 연산 (활성함수 계산) * 11층 = 23.715 Gflop
16) 콘볼루션 층만을 계산할 때는 이론적으로 27.874 Gflop이 필요하나, 활성 함수의 계산, fully connected layer등을 고려해 볼 때 도합 약 30 Gflop이 필요함
17) cuDNN : Efficient Primitives for Deep Learning, https://arxiv.org/pdf/1410.0759.pdf
18) 자세한 내용은 “AlphaGo의 인공지능 알고리즘 분석” 보고서 참고, https://spri.kr/post/14725
19) 판후이 2단과의 대국 당시 사용된 분산 AlphaGo의 GPU 개수

제3장 중소기업·스타트업의 인공지능 컴퓨팅 인프라의 실태조사[20)]

제1절 개 요

지금까지 인공지능과 컴퓨팅 인프라의 중요성에 대해서 살펴봤다. 현대 인공지능 기술에서 가장 주목받고 있는 심층학습은 필연적으로 대규모 컴퓨팅 인프라를 필요로 한다는 점에서 많은 수요가 있다는 사실은 쉽게 체감할 수 있다. 그러나 그 수요가 구체적으로 얼마인지는 실태조사 없이는 파악하기가 어렵다.

인공지능 컴퓨팅 관련 해외 동향을 살펴보면 우선 상용 클라우드 서비스가 상당 수준 활성화돼있다. 구글, 아마존, 마이크로소프트 등 글로벌 IT기업은 계산 전용 클라우드 인프라를 구축·서비스하고 있다. 특히 심층학습에 성능이 좋은 GPU위주의 환경이 주류를 이루고 있다. 구글은 알파고 대국에서 활용된 심층학습 전용 HW인 TPU(Tensorflow Processing Unit)를 기계학습 클라우드 서비스를 통해 제공하고 있다.

또한 국가차원에서의 인공지능 클라우드 컴퓨팅 인프라 구축도 진행중이다. 일본은 195억 엔(약 1,900억 원) 규모의 인공지능 전용 클라우드 인프라인 ABCI (AI Bridging Cloud Infrastructure)를 2018년 1분기까지 구축한다고 밝혔다. 일본은 자국의 인공지능 연구를 활성화하는 차원에서 신뢰 기반의 클라우드 컴퓨팅 서비스를 지원한다.[21)]

이번 장에서는 인공지능 관련 중소기업과 스타트업의 인공지능 컴퓨팅 인프라 실태조사 결과를 다루고자 한다. 실태조사 대상을 중소기업과 스타트업으로 한정한 이유는 대기업에 비해 비용적으로 인프라 구축이 어렵기 때문이다. 실태조사의 가장 큰 목적은 중소기업·스타트업의 인공지능 연구 개발을 위한 컴퓨팅 인프라의 현황과 향후수요를 파악하는데 있다. 이것을 바탕으로 정부차원의 인공지능 컴퓨팅 인프라 지원 정책을 구체화 할 것이다. 더불어 정부차원의 인공지능 컴퓨팅 인프라 지원 형태에 관한 설문조사도 실시했다.

---

20) 3장의 내용은 '중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 시사점, 소프트웨어정책연구소(2018)'에서 인용
21) 일본 ABCI 도입 배경과 운영 방안, SPRi 해외 전문가 초청 세미나, 소프트웨어정책연구소(2017)

제2절 인공지능 컴퓨팅 인프라 실태조사

1. 실태조사 개요

중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 향후 수요를 파악하기 위해 인공지능 관련 중소기업·스타트업 215개社를 대상으로 정했다. 실태조사 대상을 선정한 모집단은 지능정보산업협회의 협조를 얻었다. 실태조사 대상 215개 기업에 대해서 개별 방문 면접을 통한 설문조사를 수행했으며 유효한 표본은 72개 기업을 확보했다 <표 3-1>.

<표 3-1> 인공지능 컴퓨팅 인프라 실태조사 일반 현황

조사 대상	국내 인공지능 관련 중소기업 · 스타트업 215개社 ※ 조사대상 선정은 지능정보산업협회에서 협조를 얻음
조사 방법	개별 방문 면접을 설문 조사
유효 표본	215개社 중 72개社
조사 기간	2017년 9월 13일 ~ 9월 29일

실태조사 설문 설계에서 중점적으로 고려한 사항은 다음과 같다. 인공지능 컴퓨팅 인프라의 대부분의 수요는 심층학습을 활용하는 여부가 가장 큰 영향을 미친다고 가정했다. 그 이유는 2장에서 설명했듯이 심층학습은 경험적으로 최적의 결과를 도출해내고, 빅데이터를 처리한다는 관점에서 매우 많은 계산을 요구하기 때문이다. 또한 컴퓨팅 인프라가 제 역할을 하기 위해서는 학습을 위한 데이터의 현황도 파악해야하기 때문에 데이터 관련 설문 문항도 설계했다.

인공지능 컴퓨팅 인프라의 현황과 향후 수요를 측정하는 기준은 GPU의 보유량과 향후 수요량으로 파악했다. 심층학습은 CPU를 활용해도 충분히 시험 수준의 연구는 가능하다. 그러나 GPU의 이론 성능은 동일가격대비 CPU보다 10배 이상 높은 것이 일반적이기 때문에, CPU 현황을 조사한다고 해도 결국 GPU의 보유 현황과 향후 수요로 결정되기 때문이다. GPU의 대안으로는 Intel의 XeonPhi, 구글 TPU가 있으나 저변확대가 미흡하기 때문에 GPU만을 조사했다.

심층학습 알고리즘의 특성과 HW

- 심층학습 알고리즘은 병렬처리 효율이 높고 연산강도*가 높은 서브루틴으로 세분화[22)]
  * 연산강도는 해당 알고리즘의 연산수를 메모리 전송량으로 나눈 값으로, 연산 강도가 높다는 것은 메모리 전송량보다 연산이 많다는 것을 의미
  - 현대 연산처리장치는 연산성능이 메모리 전송속도보다 월등히 뛰어나므로, 연산강도가 높은 알고리즘에 적합
  - 특히, 매니코어 시스템인 GPU의 경우 연산처리를 담당하는 산술논리장치(Arithmetic Logical Unit)가 수 천 개 탑재되어 있기 때문에 알고리즘의 병렬화 가능성이 높을수록 기대 성능이 향상
  - 심층학습 알고리즘은 병렬화 효율이 높은 기초선형대수프로그램(Basic Linear Algebra Subprograms)으로 구성되고, 그 중 연산강도가 매우 높은 행렬 곱 연산이 대부분을 차지함

- 따라서, 심층학습은 현대 연산처리장치를 가장 효율적으로 활용할 수 있는 알고리즘이며, 가격대비 연산처리 성능이 높은 HW가 보다 높은 효율을 보유하기 때문에 시판 중인 HW 중에서 GPU가 가장 적합

- (GPU) GPU는 모니터에 그래픽 출력을 담당하는 전용 HW로 3D 작업이나 게임에 주로 활용됐으나, 무어의 법칙에 의해 연산성능이 지수적으로 향상되면서 과학계산에의 활용이 확산

- (가속기) 대표적인 가속기는 Intel의 Xeon Phi와 FPGA(Field Programmable Gate Array) 등이 있으나 비용이 매우 높아 범용성이 낮음

- (TPU) 구글이 개발한 심층학습 전용 HW로 저전력 고성능의 특징을 가지고 있으며, 구글의 기계학습 공개SW인 텐서플로우와 연동이 가능하여 편의성은 높으나 구글의 기계학습 클라우드 플랫폼에서만 활용 가능

- (Neuromorphic Chip) 뇌 신경망을 모사하여 HW로 구현한 뉴로모픽칩은 현재 상용목적으로 출시되지 않았고, 스파이킹 신경망(Spiking Neural Network)를 활용해 학습해야 함

22) 세부 내용은『인공지능의 핵심 인프라 – 고성능컴퓨팅 환경의 중요성』이슈리포트 참고

인공지능 컴퓨팅 인프라의 현황과 향후 수요를 측정하는 방법은 가장 대중화된 GPU 모델 별 개수로 파악했다. 현재 심층학습 분야에서 가장 대중화된 GPU는 NVIDIA의 GPU다. 그 대안으로 AMD가 생산하는 GPU도 있지만 심층학습 관련 SW의 최적화가 미흡하여 저변확대가 진행 중인 상황이다. 대부분의 공개SW 역시 NVIDIA GPU에서의 계산을 주로 지원하기 때문에, 사실 심층학습에 활용되는 HW는 NVIDIA GPU를 활용한다고 여겨도 무방하다. 인공지능 컴퓨팅 인프라의 현황과 향후 수요는 대표적인 세 가지 NVIDIA GPU에 대해서 보유 현황과 향후 수요를 파악하여 이론 성능을 곱한 뒤 모두 합산한 수치로 정량화했다. 설문조사에 고려한 GPU 모델과 이론 성능[23]은 <표 3-2>와 같다.

<표 3-2> 설문조사 GPU 모델

모델명	GTX 1070	GTX 1080	GTX 1080Ti, TITAN X, P100 급
이론성능 (TFLOPS)	5.2	8.2	10.6

자료 : List of NVIDIA graphical processing units, Wikipedia
https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units

또한 정부 차원의 인공지능 컴퓨팅 인프라 지원 정책의 형태에 대한 선호도를 파악하는 설문 문항을 설계했다. 인공지능 컴퓨팅 인프라 지원 정책의 형태는 크게 두 가지로 구분될 수 있다. 먼저 클라우드 형태는 민간 상용 클라우드와 유사한 과금 체계를 부여하여 보다 많은 연구주체가 활용하는 것이다. 중소기업·스타트업의 적극적인 활용을 유도하기 위해서 상용 클라우드 대비 적절한 가격을 설문했다. 두 번째 형태는 제안서 기반의 지원 방식이다. 클라우드 형태는 보다 많은 기업이 인프라의 수혜를 받는 접근이라면, 제안서 기반은 소수의 인공지능 기업이 컴퓨팅 인프라를 점유하는 형태다. 이것은 슈퍼컴퓨터 활용과 유사한 방식으로 컴퓨팅 인프라 활용 제안서를 검토 평가하여 특정 기업에 집중적으로 무상 지원하는 형태이다. 설문조사에서는 두 가지를 병행하여 운영하는 형태에 대한 문항도 설계했으며, 혼합 운영 시 적절한 비율도 조사했다. 클라우드와 제안서 기반 지원 방식에 대한 장·단점은 <표 3-3>과 같다.

23) 이론 성능은 단정도 부동소수점(single precision floating point) 연산을 기준으로 함

<표 3-3> 인공지능 컴퓨팅 인프라 지원 형태의 장·단점

지원 방식	장 점	단 점
클라우드	• 다수가 혜택을 받을 수 있는 환경 • 과금형 방식으로 지속가능한 운영에 용이	• 클라우드 운영에 대한 소요 발생 • 인프라 환경의 규모에 따라 집중 운영에 대한 어려움 발생
제안서	• 특정 기업에 집중적인 지원으로 성과 제고에 용이 • 인프라 환경의 운영이 상대적으로 용이	• 다수가 혜택을 받기 어려움 • 아이디어 구현위주의 단순한 테스트 불가

실태조사의 범위는 크게 두 가지로 나뉜다. 첫 번째는 실태조사의 궁극적인 목적인 컴퓨팅 인프라의 현황과 향후 수요를 파악하는데 있다. 이 부분에서는 심층학습과 GPU 활용여부를 시작으로 <표 3-2>에 해당하는 GPU 현황과 향후 수요를 조사했다. 또한 정부의 인공지능 컴퓨팅 인프라 운영 방식에 대한 설문과 기 추진된 사업에 대한 인지·수혜 여부를 파악했다. 두 번째는 인공지능 R&D 및 기술 수준에 대한 조사도 병행했다. 구체적인 실태조사 범위는 <표 3-4>와 같다.

<표 3-4> 인공지능 컴퓨팅 인프라 실태조사 범위

구 분	조 사 범 위
컴퓨팅 인프라 현황	• 인공지능 연구분야, 데이터의 종류, 데이터의 양 – 현재와 향후수요 • 심층학습, GPU 활용 여부, 인공지능 계산 GPU 규모 – 현재와 향후수요 • 선호하는 정부의 인공지능 컴퓨팅 인프라 지원사업 운영방식 (클라우드, 제안서) • 인공지능 컴퓨팅 인프라 관련 정부지원 사업 인지 여부, 수혜 경험, 애로사항
인공지능 R&D 및 기술 수준	• 인공지능 관련 R&D 예산 비율 – 현재 및 향후 • 인공지능 제품 및 서비스 개발을 위한 SW 사용 현황 • 인공지능 컴퓨팅 인프라 사업에 대한 의견 • 인공지능 기술 분류체계에 대한 해당 영역 설문 • 해당 인공지능 기술 분야에 대한 기업과 우리나라의 기술수준

설문조사 대상 215개 기업에서 유효한 응답을 한 72개 기업의 특성은 다음 <표 3-5>와 같다.

<표 3-5> 72개 실태조사 응답 기업에 대한 특성

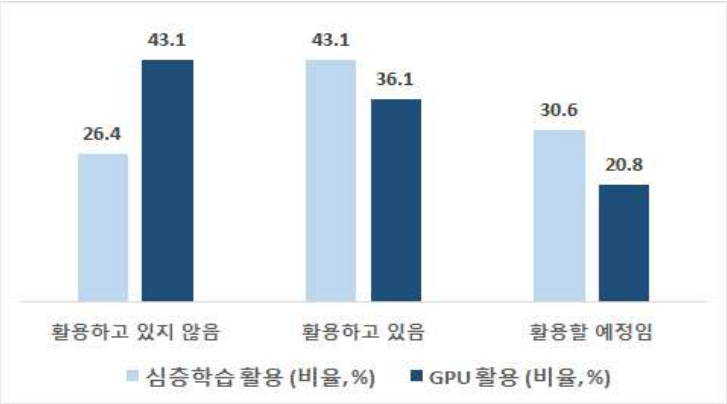
구 분		응답 기업 수	백분율 (단위 : %)
전체		72	100.0
설립년도	2000년 이전	20	27.8
	2000-2010년	25	34.7
	2010년 이후	27	37.5
매출액 기준	10억 미만	17	23.6
	10-30억 미만	17	23.6
	30-50억 미만	6	8.3
	50-80억 미만	4	5.6
	80-120억 미만	3	4.2
	120-200억 미만	9	12.5
	200-500억 미만	8	11.1
	500억 이상	8	11.1
심층학습 활용	비활용	19	26.4
	활용	31	43.1
	활용 예정	22	30.6
GPU 활용	비활용	31	43.1
	활용	26	36.1
	활용 예정	15	20.8
정부지원 인지 여부	인지	26	36.1
	비인지	46	63.9

2. 실태조사 결과

중소기업·스타트업의 인공지능 컴퓨팅 인프라의 조사 결과에 앞서 심층학습과 GPU활용 여부에 대한 결과를 먼저 기술하겠다. 설문조사 설계 상 인공지능 컴퓨팅 인프라의 대부분의 수요는 심층학습을 활용해 제품이나 서비스를 개발하는 기업과 심층학습을 위해 GPU를 활용하는 기업에서 발생할 것이라고 가정했기 때문이다. 이 결과는 간접적으로 국내 인공지능 관련 중소기업·스타트업 중 어느 정도의 비율이 인공지능 기술을 직접적으로 활용하고 있는지에 대해 가늠할 수 있다.

먼저 심층학습을 활용하고 있는 기업은 31개(43.1%) 기업이며, GPU를 사용은 26개(36.1%) 기업으로 조사됐다. 국내 인공지능 연구는 지난 2016년 '알파고 쇼크'가 촉진제 역할을 했기 때문에, 심층학습이나 GPU를 활용하는 기업은 과반에 미치지 못했다. 그러나 활용 예정이라고 답한 수치를 합산해 보면 심층학습을 활용하고 있거나 활용할 예정인 기업은 53개(73.7%) 기업, GPU 활용 및 활용예정 기업은 41개(56.9%) 기업으로 파악됐다. 이 결과는 향후 인공지능과 관련된 인프라의 잠재적인 수요가 있다는 것으로 분석할 수 있다.

심층학습을 활용하고 있는 기업 중 GPU를 활용하는 경우는 74.2%, GPU를 활용하는 기업 중 심층학습을 활용하는 경우는 88.5%로 높은 상관성을 나타냈다. 인공지능 관련 중소기업·스타트업의 심층학습 적용은 상당히 높을 것으로 판단되나, 본격적인 서비스 개발에 필요한 GPU 활용은 다소 소극적으로 분석된다. 그 이유는 심층학습을 시험하는 초기 단계에는 반드시 GPU가 필요한 것은 아니기 때문으로 추정된다.



[그림 3-1] 심층학습 및 GPU 활용 여부 설문 결과

유효한 응답을 한 72개 기업에 대한 컴퓨팅 인프라 현황은 11.37 PFLOPS이며, 향후 수요는 22.04 PFLOPS로 파악됐다. 조사 대상을 215개 기업으로 추정[24]할 경우 인공지능 컴퓨팅 인프라 현황은 25.94 PFLOPS이며, 향후 수요는 47.14 PFLOPS로 조사됐다. <표 3-2>에 해당하는 모델 별 GPU 설문조사 결과는 <표 3-6>과 같다.

<표 3-6> GPU 현황 및 향후 수요 (설문 응답 72개 기업 대상)

모델명	GTX 1070	GTX 1080	GTX 1080Ti, TITAN X, P100 급
현황	194개	117개	887개
향후 수요	553개	230개	1,630개
연산량 환산[25] (현황)	1,008.8 TFLOPS	959.4 TFLOPS	9,402.2 TFLOPS
총계 (현황)	11.37 PFLOPS		
연산량 환산 (수요)	2,875.6 TFLOPS	1,886 TFLOPS	17,278 TFLOPS
총 계 (수요)	22.04 PFLOPS		

<표 3-6> 결과를 토대로 전체 설문조사 대상인 215개 기업으로 추정하기 위해 215개 기업의 2016년 매출액 기준으로 구분한 현황은 <표 3-7>과 같다.

<표 3-7> 전체 설문조사 대상 215개 기업의 매출액 분포

구 분	10억 원 미만	10~30 억 원 미만	30~50 억 원 미만	50~80 억 원 미만	80~120 억 원 미만	120~200 억 원 미만	200~500 억 원 미만	500억 원 이상
72개 대상	17	17	6	4	3	9	8	8
215개 대상	54	40	20	15	20	24	23	19

24) 추정 방법은 215개 기업의 2016년 매출액 기준으로 산정
25) GPU 조사 현황에 <표 3>의 이론성능을 곱한 수치

설문조사를 통한 인공지능 컴퓨팅 인프라 현황과 향후 수요 수치의 의미는 해석에 주의가 필요하다. 먼저 국내 인공지능 관련 중소기업·스타트업의 수가 절대적으로 부족하다. 또한 설문에 응답한 72개 기업의 컴퓨팅 인프라 수요를 215개 기업의 수요로 확대한 경우 역시 72개 기업의 대표성에 편향이 있을 수 있다. 일부 편향을 고려하더라도 국내 인공지능 관련 중소기업과 스타트업의 컴퓨팅 인프라 현황과 향후 수요는 215개 기업으로 추정한 것을 활용하는 것이 적절하다.

인공지능 컴퓨팅 인프라의 현황과 향후 수요의 실태조사 결과로 산출된 연산량(FLOPS)은 이론적인 최대치를 의미하는 것이다. 예를 들면 215개 기업의 향후 수요량이라고 측정된 47.14 PFLOPS는 이 수치를 단일 시스템으로 구성해야 하는 것은 아니다. 단정밀도(single precision)으로 47.14 PFLOPS라는 수치는 2018년 도입이 예정된 국가 슈퍼컴퓨터 5호기의 이론 성능[26]에 못미친다. 슈퍼컴퓨터 5호기는 장비 가격만 500억 원이 넘는 고가의 장비이고, 수십 PFLOPS의 연산을 한 번에 처리할 수 있는 성능을 갖는다.

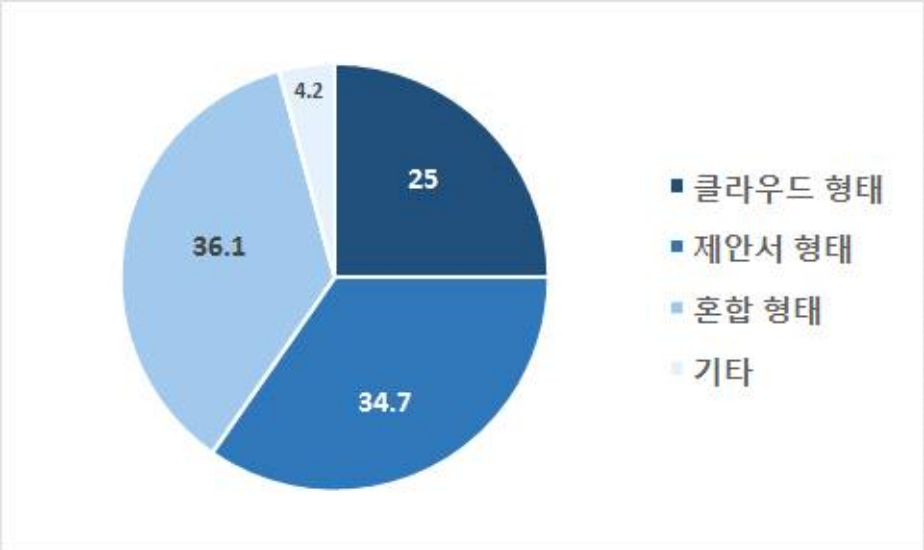
따라서 현황이나 향후 수요는 단일 시스템이라기보다는 분산된 컴퓨터로 봐야한다. 또한 인공지능 컴퓨팅 인프라의 현황과 향후 수요는 단순 합산으로 산출됐기 때문에 활용률 측면은 고려되지 않았다. 활용률은 장비의 가동 현황을 나타내는 것으로 100%일 경우 유휴 상태(idle) 없이 작동하는 것을 의미한다. 이러한 이유로 단순 합산 수치에는 인프라를 구성하는 형태에 대한 정보는 없다. 예를 들면, 컴퓨터 1대에 GPU를 여러 장 활용할 수 있고, 이러한 컴퓨터를 클러스터로 구축할 수도 있다. 이 부분은 설문이 지나치게 기술적인 측면으로 집중되기 때문에, 이번 설문조사에서는 제외했다. 결론적으로 인공지능 컴퓨팅 인프라 현황과 수요로 추정된 수치는 컴퓨팅 인프라 지원 정책을 설계하는데 고려할 자료로 활용될 수 있지만, 반드시 저 수치를 달성해야 한다는 의미는 아니다.

인공지능 컴퓨팅 인프라의 현황과 향후 수요는 매출액 30억 미만의 기업, 2010년 이후에 설립된 기업, 심층학습과 GPU를 활용하는 기업일수록 높게 나타났다. 인공지능 컴퓨팅 인프라 현황의 87.9%(약 10 PFLOPS)가 매출액 30억 미만의 기업이 차지하며, 향후 수요 역시 76.6%(약 16.9 PFLOPS)로 높은 비중을 차지했다.

---

26) 슈퍼컴퓨터 5호기의 이론성능은 25.7 PFLOPS(배정밀도). 단정밀도로 추정할 경우 51.4 PFLOPS

<표 3-3>에서 소개한 두 가지 인공지능 컴퓨팅 인프라 지원 방법에 대한 설문 결과이다 [그림 3-2].



[그림 3-2] 인공지능 컴퓨팅 인프라 지원 방법에 대한 선호도 조사

설문조사 결과 클라우드 형태와 제안서 형태의 혼합 운영을 26개(36.1%) 기업이 선택하여 가장 선호하는 형태로 조사됐다. 혼합 운영시 클라우드와 제안서 형태의 비중은 평균적으로 50:50이었다. 슈퍼컴퓨터 활용과 유사한 제안서 형태는 25개(34.7%) 기업이 선호했다. 혼합 운영이나 제안서 형태를 선택한 응답 기업 51개에 대해서 제안서 형태의 정책에 대한 구체적인 방법을 설문했다.

제안서 형태의 정책에 대한 방법은 크게 두 가지가 있다. ① 컴퓨팅 인프라 활용의 최대치를 제한하되, 보다 많은 기업에 지원한다. ② 지원받는 기업의 수를 제한하되, 컴퓨팅 인프라 활용의 최대치를 제한하지 않는다. 제안서 형태의 구체적인 지원 방법인 ①안과 ②안의 선호도는 50:50으로 중립적인 의견이 대다수였다. 제안서 형태의 컴퓨팅 인프라 적정 활용기간에 대한 조사에서는 '6개월 활용 후 심사를 통해 연장' 하는 방안이 61%(31개 기업 응답)를 차지했다.

일정 비용을 지불하고 사용하는 클라우드 형태는 18개(25%) 기업이 선택했다. 정부가 지원의 클라우드 서비스는 사용 활성화를 위해 상용 서비스보다 저렴한 가격정책이 필요하다. 따라서 혼합 운영이나 클라우드 형태를 선택한 응답 기업 44개에 대해서 적절한 클라우드 사용료를 설문한 결과 상용대비 평균 45%로 분석됐다.

설문 응답 기업이 향후 확보하고자 하는 데이터의 양 9,083 테라바이트(TB)는 현재 구축한 1,837TB 대비 평균적으로 4.8배 증가할 것으로 조사됐다.

데이터는 종류는 크게 정형텍스트, 비정형텍스트, 이미지 및 영상, 신호, 음성 및 청각 데이터로 세분화하여 중복응답을 허용하는 것으로 질문을 설계했다. 자연어처리에 주로 활용되는 정형텍스트와 비정형텍스트는 각각 응답 기업의 50%, 48.5%가 활용하고, 이어 이미지 및 영상 데이터 활용은 응답 기업의 43.1%가 활용했다. 전기 사용, 금융 거래 등 신호 형식의 데이터는 25%가 활용하며, 음성 인식 등 청각에 관련된 데이터는 18.1%의 응답 기업이 활용하는 것으로 분석됐다.

데이터의 양적인 측면에서는 1~2TB[27] 수준으로 데이터를 구축·활용하고 있는 기업(23개, 31.9%)이 가장 많은 것으로 조사됐고, 향후 구축할 데이터 양이 50TB 이상인 경우는 30.6%으로 많은 수요가 발생할 것으로 예상된다. 설문 응답 72개 기업이 현재 활용하고 있는 데이터의 양은 평균적으로 26.76TB이며, 향후 구축 예정인 데이터의 양은 평균 129.76TB이다. 특히 향후 구축 데이터의 양은 심층학습이나 GPU를 활용 예정이라고 응답한 기업에서 높게 나타났다.

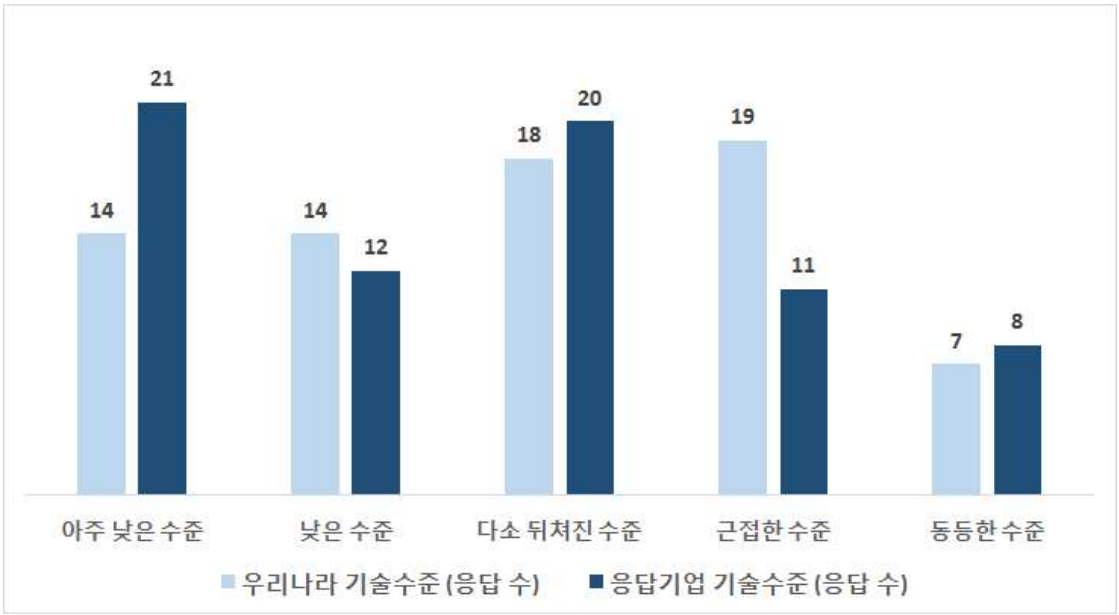
결론적으로 인공지능 컴퓨팅 인프라 구축 시 막대한 데이터를 저장·가공하기 위한 스토리지 서버 역시 구축해야할 필요성 역시 존재하는 것으로 분석된다.

현재 정부가 추진하고 있는 인공지능 컴퓨팅 인프라 지원 사업에 대한 인지 여부도 조사했다. 정부가 추진하고 있는 인공지능 컴퓨팅 인프라 관련 사업은 크게 두 가지로 컴퓨팅 인프라 지원은 아직 본격적으로 시작하지 않은 상황이다. HPC 이노베이션 허브구축 사업의 2017년 예산은 77억 원이며, 20개 스타트업에 컴퓨팅 인프라를 지원할 계획을 갖고 있다. 지능정보산업 인프라 조성 사업의 2017년 예산은 50억 원이며, 지식 베이스, 인공지능 SW 이용환경 구축 등을 수행하고 인공지능 컴퓨팅 인프라는 일부 지원하고 있다.

27) 1TB는 2백만 장의 500킬로바이트(KB) 이미지, 약 백만 권의 책이 저장 가능

정부의 인공지능 컴퓨팅 인프라 사업에 대한 인지여부 조사 결과 63.9%가 인지하고 있지 않은 것으로 조사됐다. 향후 정부의 인공지능 컴퓨팅 인프라 지원 사업에 대한 활용의사를 설문한 결과 63개(87.5%) 기업이 활용의사가 있는 것으로 높게 나타났다.

응답 기업의 인공지능 기술 영역의 최고 기술보유국 대비 국내 수준과 자사의 수준을 설문조사한 결과 모두 '다소 뒤쳐진 수준'으로 조사됐다 [그림 4]. 이 결과는 설문 응답 수의 평균으로 추정했다.



[그림 3-3] 인공지능 기술수준 조사 결과

인공지능 컴퓨팅 인프라 사업관련 의견에서는 연구 인력의 지원 요구가 18.1%로 가장 많았으며, 인프라 운영에 대한 기술적인 완성도에 대한 필요성이 제기됐다 <표 3-8>.

<표 3-8> 인공지능 컴퓨팅 인프라 사업관련 주요 의견

인력 지원	• 연구 개발 인력 지원 필요 – 인공지능 컴퓨팅 인프라 구축에 앞서 인력 지원이 선행돼야 함 – 특히, 빅데이터의 이해를 갖춘 전문 인력 양성을 요구 • 인력양성을 위한 커리큘럼 필요 – 컴퓨팅 인프라를 활용하기 위한 교육 필요
기술 지원	• 사용성에 대한 고려 – 고장이나 오작동에 대한 실시간 대처 필요 – 행정적인 절차에 대한 간소화 – 정부 지원 사업의 홍보 필요 • 기술 지원 및 보안 – 시스템 오류를 해결할 수 있는 기술적 지원 필요 – 데이터의 보안에 대한 신뢰가 있어야 함 • 다양한 플랫폼에의 연동 – 다양한 공개SW 지원, 게임·로봇 등 응용분야와의 연계 필요

제3절 소 결

중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 향후 수요 조사 결과 향후 3년 이내 현재보다 2배 정도의 인프라 수요가 발생할 것으로 예산된다. 인공지능 컴퓨팅의 주요 수요처는 매출액 30억 원 이하의 기업과 2010년 이후에 설립된 기업에에 집중되어 있으며, 심층학습과 GPU를 활용한다고 응답한 기업일수록 컴퓨팅 인프라의 현황과 향후수요가 높게 측정됐다.

특히 자금운용이 상대적으로 어려운 스타트업에 우선적으로 인프라를 공급할 필요성 제기된다. 또한, 심층학습과 GPU를 활용예정으로 응답한 기업에 대한 잠재적인 수요도 고려해야 하므로 신속한 인공지능 컴퓨팅 인프라 구축이 필요할 것이다.

인공지능 컴퓨팅 인프라 관련 설문조사의 결과는 GPU 장비의 단순합산으로 산정된 것이기 때문에, 정책 설계의 유의가 필요하다. 정부의 인공지능 컴퓨팅 인프라 지원의 형태에 관한 설문 결과 클라우드와 제안서 형태를 혼합한 방법을 가장 선호하는 것으로 조사됐다. 클라우드와 제안서 형태 두 가지 방법에서는 제안서 형태를 더 선호하는 것으로 파악된다. 클라우드 형태는 상용 클라우드

서비스의 45% 정도의 요금이 적절한 것으로 분석됐으며, 제안서 형태의 경우 적정 활용 기간은 6개월 활용 후 재심사를 선호했다.

인공지능 연구에 활용하고자 하는 데이터의 양도 현재 대비 향후 4.8배 증가할 것으로 전망되어, 데이터와 관련한 인프라 구축의 병행 필요할 것이다.

제4장 요약 및 시사점

이번 보고서에서는 인공지능과 컴퓨팅 인프라의 중요성에 대해서 심도 있게 분석했다. 현대 인공지능의 핵심인 심층학습은 경험적으로 최상의 결과를 도출해 낸 것이다. 심층학습은 수많은 변수가 존재하기 때문에, 다양한 가능성을 상정하여 많이 시도해보는 것이 가장 일반적인 접근이다. 또한 심층학습의 특성상 대규모 데이터를 필요로 하므로 사실 인공지능에 컴퓨팅 인프라가 중요하다는 사실은 직관적으로 이해할 수 있다.

심층학습의 세부 알고리즘은 현대 연산처리장치를 십분 활용할 수 있는 연산들로 구성된다. 다시 말하자면 컴퓨팅 인프라를 2배 구축하면 기존 보다 2배에 가까운 성능을 기대할 수 있다는 것이다. 이처럼 유연한 확장성(Scalability)을 갖는 인공지능 알고리즘의 특성상 컴퓨팅 인프라가 많이 확보될수록 빠른 연구결과를 도출할 수 있는 밑거름이 될 것이다.

현재 인공지능 분야의 기술 발전 속도는 매우 넓고 빠르게 진행된다. 인공지능은 범용기술(General Purpose Technology)로 발전함에 따라 컴퓨터 공학을 넘어 자연·사회과학까지 영역을 넓히고 있다. 인공지능 연구의 특징은 매우 빠르다는 점과 점진적으로 개선된다는 점에서 SW의 특성과 비슷하다. 인공지능 컴퓨팅 인프라의 확보는 이러한 인공지능의 특성을 반영한 연구 환경에 직결된다고 볼 수 있다.

이번 보고서에서는 상대적으로 자금 운용이 어려운 중소기업·스타트업의 인공지능 컴퓨팅 인프라에 대한 실태조사 결과를 분석했다. 가장 큰 특징은 신생 기업일수록, 매출액 30억 원 미만의 작은 기업일수록 인공지능 컴퓨팅 인프라의 수요가 높게 측정됐다. 설문 대상기업과의 인터뷰에서도 대다수의 스타트업이 컴퓨팅 인프라 지원이 필요하다고 밝힘에 따라, 정부차원의 인공지능 컴퓨팅 인프라 정책은 큰 실효를 거둘 가능성이 높다.이 실태조사 결과는 향후 본격적인 인공지능 컴퓨팅 인프라 지원정책을 설계하는데 기반자료로 활용할 수 있을 것이다.

참 고 문 헌

1. Silver, D. et al., “Mastering the game of Go with Deep neural networks and tree search,”Nature vol 529, pp. 484-489, 28 Jan 2016.
2. cuDNN : Efficient Primitives for Deep Learning, https://arxiv.org/pdf/1410.0759.pdf
3. Barba, Lorena A., and Rio Yokota. "How will the fast multipole method fare in the exascale era." SIAM News 46.6 (2013): 1-3.
4. AlphaGo의 인공지능 알고리즘 분석, 소프트웨어정책연구소 (2016)
5. 국가 슈퍼컴퓨팅 역량 강화 방안, 소프트웨어정책연구소(2017)
6. 인공지능의 핵심 인프라 - 고성능컴퓨팅 환경의 중요성, 소프트웨어정책연구소 (2017)
7. 중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 시사점, 소프트웨어정책연구소(2018)
8. Convolutional Neural Network, http://cs231n.github.io/convolutional-networks/
9. cuBLAS, NVIDIA, http://docs.nvidia.com/cuda/cublas/
10. CUDA Toolkit Documentation, NVIDIA, https://docs.nvidia.com/cuda/cuda-c-programming-guide/
11. Deep Blue https://www-03.ibm.com/ibm/history/exhibits/vintage/vintage_4506VV1001.html
12. Fused Multiply-Add, https://en.wikipedia.org/wiki/Multiply%E2%80%93accumulate_operation
13. Top500, http://www.top500.org
14. Intel Xeon Processor E5-2699 v4, http://ark.intel.com/products/91317
15. Intel Xeon Phi 7120P, http://ark.intel.com/products/75799
16. Nvidia Tesla P100, http://www.nvidia.com/object/tesla-p100.html
17. Roofline Performance Model https://crd.lbl.gov/departments/computer-science/PAR/research/roofline/
18. Samsung Exynos, Wikipedia https://en.wikipedia.org/wiki/Exynos
19. Watson (computer), Wikipedia, https://en.wikipedia.org/wiki/Watson_(computer)

별 첨 1. 실태조사 설문지

통계법 제33조(비밀의 보호 등)
① 통계작성과정에서 알려진 사항으로서 개인 또는 법인이나 단체의 비밀에 속하는 사항은 보호되어야 한다.
② 통계작성을 위하여 수집된 개인 또는 법인이나 단체의 비밀에 속하는 기초자료는 통계작성의 목적 외에 사용하여서는 아니 된다.

ID	-			

중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 수요조사

안녕하십니까? 귀사의 무궁한 발전을 기원합니다.

소프트웨어정책연구소에서는 인공지능과 고성능 컴퓨팅 인프라의 중요성을 강조하며, 중소기업·스타트업을 위한 컴퓨팅 인프라 지원 사업에 대해 정책 연구를 수행하고 있습니다.

이에 본 조사는 중소기업·스타트업의 인공지능 컴퓨팅 인프라 현황과 수요를 파악하고자 합니다. 인공지능은 빅데이터를 학습해 패턴을 도출하고 미래를 예측한다는 점에 있어 막대한 계산량을 필요로 합니다.

따라서 현실적이고 효과적인 인공지능 컴퓨팅 인프라 지원 정책을 위해 귀사의 적극적인 도움을 요청 드립니다. 응답 결과는 정책 개발을 위한 기초자료로만 활용할 예정이며, 그 외의 목적으로는 절대 사용되지 않습니다. 응답하신 내용은 통계법 제33조에 의해 철저히 비밀을 보장 받습니다.

인공지능 컴퓨팅 인프라 정책 결정을 위한 귀중한 자료로 사용할 예정이오니 바쁘시더라도 귀사의 고견을 응답해 주시기 바랍니다.

2017년 9월

주 관 : 소프트웨어정책연구소	조사기관 : ㈜마크로밀엠브레인

사업체 일반 현황

기업체명		**설립년도**	년 월
상장 여부	① 코스피② 코스닥③ 코넥스④ 비상장		

응답자 성명		**응답자 전화번호**	
조사 일시	2017년 월 일 (오전/오후)	_____시 _____분부터	_____시 _____분까지
조사원 성명		**관리자(SV)**	
에디터		**검증원**	

A. 컴퓨팅 인프라 현황 및 전망

인공지능 계산을 위한 컴퓨터

인공지능 연구개발에 활용하고 있는 계산용 서버 및 PC 기준, 클라우드 컴퓨터의 경우도 포함. 단, 웹서버 및 인트라넷용 서버는 제외

문 1. 귀사의 **인공지능 연구를 적용하고 있는 분야를 모두 선택**해 주십시오.

① 의료	② 국방	③ 교육
④ 농업	⑤ 법률	⑥ 웰니스
⑦ 제조	⑧ 홈	⑨ 금융
⑩ 문화관광	⑪ 유통	⑫ 사무관리
⑬ 도시	⑭ 교통	⑮ 에너지
⑯ 안전	⑰ 공공서비스	⑱ 기타 ( )

문 2. 귀사에서 **인공지능 연구에 활용하고 있는 데이터의 종류를 모두 선택**해 주십시오.

① 이미지 및 영상	② 정형 텍스트
③ 비정형 텍스트	④ 음성 및 청각데이터
⑤ 신호	⑥ 기타 ( )

※ 예시
① 이미지 / 영상 : 일반적인 사진, 의료 사진 및 영상, CCTV 영상, 위성 영상 등
② 정형 텍스트 : 일반적인 데이터베이스를 활용 할 수 있는 텍스트 데이터 (구매 이력 등)
③ 비정형 텍스트 : 신문 기사, 문서 등 정형적으로 데이터베이스화 할 수 없는 데이터
④ 음성 : 음성, 음파 등 청각과 관련된 데이터
⑤ 신호 : 시계열적인 특성이 있는 데이터 (시간에 대한 의존성이 있는 경우)
⑥ **기타 : 위 범주에 해당하지 않는 데이터는 자세하게 기술 해주시기 바랍니다.**

문 3. 귀사에서 **인공지능 연구에 활용하고 있는 데이터의 양**을 작성해 주십시오.

( ) 테라바이트* (TeraByte, TB)

* 테라바이트는 1,000기가바이트 (일반 PC에 탑재되는 하드디스크 용량은 1테라바이트)

문 4. 귀사에서 **향후 인공지능 연구에 활용하기 위해 확보하고자 하는 데이터의 양**을 작성해 주십시오.

( ) 테라바이트 (TeraByte, TB)

문 5. 귀사에서는 **인공지능 연구에 딥러닝*을 활용**하고 계십니까?

① 활용하고 있지 않다 ② 활용하고 있다 ③ 활용할 예정이다

* 딥러닝은 기계학습의 한 종류로 빅데이터를 학습하여 패턴을 인식하는 알고리즘. 본 문항에서의 딥러닝은

합성곱신경망(CNN), 순환신경망(RNN), 심층신경망(DNN) 등을 포함하고, 기존 인공신경망(Artificial Neural Network)에서 파생된 신경망 방법론도 모두 포함

문 6. 귀사에서는 **인공지능 연구에 GPU*를 활용**하고 계십니까?

① 활용하고 있지 않다　　② 활용하고 있다　　③ 활용할 예정이다

* GPU는 그래픽연산처리장치(Graphical Processing Units)으로 3D 가속이나 그래픽 처리를 하기 위한 용도로 주로 활용됐으나, 계산용도 특히 딥러닝에 최적화된 연산처리장치로 부상

문 7-8. 귀사에서 **인공지능 계산을 위한 GPU 규모* 현황과 수요**를 아래 표에 작성하여 주십시오.**

* 현 시점에서 가장 널리 활용되는 학습용(training) 연산처리장치는 GPU이기 때문에 GPU의 규모로 현황과 수요를 파악하고자 함

** GPU 규모에 대한 수요 부분은 향후 3년 이내 추진하고자 하는 인공지능 제품 및 서비스 개발에 필요한 규모 작성

※ 성능별 보기 참조

<　응답표　>

<table>
<tr><th>GPU 모델명</th><th>성 능 요 약</th><th>계 산 능 력</th><th>문7. 현재
보유 규모</th><th>문8. 향후
희망 규모</th></tr>
<tr><td>① GTX 1070</td><td>이론성능 : 5.2 TFLOPS
메모리 : 8GB</td><td>이미지 학습 : 초당 100여장
기계번역[1] : 1배 성능 향상</td><td>_____장</td><td>_____장</td></tr>
<tr><td>② GTX 1080</td><td>이론성능 : 8.2 TFLOPS
메모리 : 8GB</td><td>이미지 학습 : 초당 150여장
기계번역 : 1.5배 성능 향상</td><td>_____장</td><td>_____장</td></tr>
<tr><td>③ GTX 1080Ti</td><td>이론성능 : 10.6 TFLOPS
메모리 : 11GB</td><td rowspan="3">이미지 학습 : 초당 200여장[2]
기계번역 : 2배 성능 향상</td><td rowspan="3">_____장</td><td rowspan="3">_____장</td></tr>
<tr><td>④ GTX TITAN X
(Pascal, 10series)</td><td>이론성능 : 10.1 TFLOPS
메모리 : 12GB</td></tr>
<tr><td>⑤ P100 (P40 포함)</td><td>이론성능 : 10.6 TFLOPS
(11.8 TFLOPS)
메모리 : 16GB (24GB)</td></tr>
<tr><td>⑥ 해당사항이
없는 경우</td><td>모델명을 알고 있을
경우 별도 기입</td><td>-</td><td>_____
장</td><td>_____
장</td></tr>
</table>

1) 기계번역 성능 기준 : K80 활용 시 1배 (LSTM + RNN 활용)

2) ③, ④는 데스크탑 보급형 그래픽카드이고, ⑤의 경우는 계산 전용으로 생산하여 워크스테이션 급으로 활용 가능 (⑤는 한 컴퓨터에 최대 8장으로 구성 가능, DGX-1)

문 9. 귀사에서 **인공지능 R&D를 위한 향후 컴퓨팅 인프라 형태**를 어떤 비율로 계획하고 있습니까? 아래의 표에 합이 100%가 되도록 응답하여 주십시오.

클 라 우 드	자 체 구 축	합 계
%	%	100%

문 10. **정부에서 인공지능을 위한 컴퓨팅 인프라를 지원하는 사업을 시행**하는 것을 알고 있습니까?
(예 - 지능정보산업 인프라 조성 사업, 글로벌 HPC 이노베이션 허브 구축 사업)

① 알고 있다 ② 모른다

문 11. [문10의 ①응답자만] **최근 1년간 정부에서 추진하고 있는 인공지능 관련 컴퓨팅 인프라 지원사업의 수혜**를 받은 경험이 있습니까?

① 컴퓨팅 인프라 수혜 경험 있음 ② 컴퓨팅 인프라 수혜 경험 없음

문 12. 귀사에서 인공지능 관련 사업을 수행하시면서 컴퓨팅 인프라와 관련한 애로사항이 있습니까?

전혀 어려움이 없다	그다지 어려움이 없는 편이다	보통이다	어려움이 약간 있는 편이다	어려움이 매우 많다
①	②	③	④	⑤

문 13. 정부에서 인공지능 연구를 위해 클라우드 컴퓨팅 인프라 지원사업을 운영할 시,
다음 중 **귀사에서 선호하는 지원사업 운영 방식**은 무엇입니까?

① 상용보다 저렴한 과금형 클라우드 서비스 ➔ 문 13-1 로 이동
② 제안서 채택으로 일정 규모 컴퓨팅 인프라 무상 지원 ➔ 문 13-2 로 이동
③ ①과 ②의 혼합 운영 ➔ 문 13-1 로 이동
④ 기타의견 ( ) ➔ 문 13-5 로 이동

문 13-1. [문13의 ①, ③ 응답자] 저렴한 과금형 클라우드 서비스를 지원하는 형태로 운영될 시,
상용 클라우드 요금 대비 적절한 과금형 클라우드 요금의 비율은 몇 %라고 생각합니까?

(%)

문 13-2. [문13의 ②, ③ 응답자] **제안서 채택 방식으로 컴퓨팅 인프라를 보급할 경우 적정한 기준**은
어떤 것인지 선택해 주십시오.

① 컴퓨팅 인프라 활용의 최대치를 제한하되, 많은 기업에 지원하는 방식
② 지원받는 기업의 수를 제한하되, 컴퓨팅 인프라 활용의 최대치를 제한하지 않는 방식

문 13-3. [문13의 ②, ③ 응답자] **제안서 채택 방식으로 컴퓨팅 인프라를 보급할 경우 적정한 활용 기간**은 어떤 것인지 선택해 주십시오.

① 1개월 ~ 6개월 ② 6개월 ~ 1년
③ 6개월 활용 후 심사를 통해 연장 ④ 기타의견 ( )

문 13-4. [문13의 ③ 응답자] **과금형 클라우드 방식과 제안서 채택 방식의 혼합 형태로 구성**할 경우
그 비율을 작성해 주십시오.

과금형 클라우드	제안서 채택 방식	합 계
%	%	100%

문 13-5. [문13의 ④ 응답자] 귀사에서 **선호하시는 정부 클라우드 컴퓨팅 인프라 운영 방식**에 대해 자유롭게 제안해 주시기 바랍니다.

문 14. 정부에서 인공지능을 위한 컴퓨팅 인프라를 지원하는 사업을 시행할 경우 **귀사에서는 정부에서 지원하는 컴퓨팅 인프라 활용의사**가 있으십니까?

① 활용의사 있음 ② 활용의사 없음

B. 인공지능 관련 R&D 투자 및 개발 현황

문 15. 귀사에서 **현재 인공지능 관련 R&D에 투입하고 있는 예산 비율**을 선택해 주시기 바랍니다.

① 매출의 5% 이하 ② 매출의 5~10% ③ 매출의 10~20%
④ 매출의 20~30% ⑤ 매출의 30% 이상

문 16. 귀사에서 **향후 인공지능 관련 R&D에 투입할 예정인 예산 규모 비율**을 선택해 주시기 바랍니다.

① 매출의 5% 이하 ② 매출의 5~10% ③ 매출의 10~20%
④ 매출의 20~30% ⑤ 매출의 30% 이상

문 17. 귀사에서 **인공지능 제품 및 서비스 개발을 위해 사용하고 있는 SW를** 모두 선택해주십시오.

① 텐서플로우(Tensorflow) ② 토치(Torch) ③ 띠아노(Theano)
④ 카페(Caffe) ⑤ CNTK/DMTK ⑥ 케라스(Keras)
⑦ 매틀랩(MATLAB) ⑥ R(통계프로그램) ⑦ 기타 ( )

문 18. **최근 1년간 정부에서 추진하고 있는 인공지능 관련 R&D 사업에 선정**되신 적이 있으십니까? 선정된 적이 있으시다면, 지원받으신 연구비 금액을 적어주십시오.

① 정부 R&D 사업 선정 경험 있음 (총 지원받은 연구비 : 원)
② 정부 R&D 사업 선정 경험 없음

문 19. **인공지능 컴퓨팅 인프라 지원 사업에 대한 귀사의 자유로운 의견**을 부탁드립니다. 제시하신 의견은 향후 정책설계의 자료로 활용됩니다.

※ 귀사에서 당면하고 있는 과제를 작성하셔도 좋습니다.

(예시) 컴퓨팅 인프라의 보안이 매우 중요하다. *컴퓨팅 인프라를 직접적으로 활용할 수 있는 HPC(High Performance Computing) 엔지니어가 필요하다.*

C. 인공지능 기술 수준

문 20. 다음 중 **귀사에 해당하는 인공지능 영역을 모두 선택**해 주십시오.

지능정보기술 4대 분야 17대 기술 中			
인공지능 SW(7종)			
① 추론 및 기계학습 ⑤ 복합 지능	② 지식표현 및 언어지능 ⑥ 지능형 에이전트	③ 청각 지능 ⑦ 인간-기계 협업	④ 시각 지능

문 21. 다음 중 **귀사에서 가장 연관도가 높은 데이터·네트워크 기술을 모두 선택**해 주십시오.

지능정보기술 4대 분야 17대 기술 中			
데이터·네트워크(4종)			
① 사물인터넷	② 클라우드 컴퓨팅	③ 빅 데이터	④ 모바일 (5G)

문 22. 귀사의 인공지능 분야에서 **세계 최고수준 대비 우리나라의 기술수준과 귀사의 기술수준**은 각각 어느 위치에 있는지 응답해주십시오.

세계 최고 100점 기준 대비 (%)	우리나라 기술 수준	귀사 기술 수준
① 100점 : 세계 최고 수준	①	①
② 90~99점 : 최고산업기술과 동등한 수준	②	②
③ 80~89점 : 최고산업기술국에 근접한 수준	③	③
④ 70~79점 : 최고산업기술국보다 다소 뒤쳐진 수준	④	④
⑤ 60~69점 : 최고산업기술국보다 낮은 수준	⑤	⑤
⑥ 59점 이하 : 최고산업기술국보다 아주 낮은 수준	⑥	⑥

문 23. 귀사의 **재무현황**에 대해 응답해주십시오.

구분	2016년	2017년 (연말 추정치)	2018년(목표치)
전체 매출액	백만원	백만원	백만원
인공지능 매출액	백만원	백만원	백만원

● 끝까지 설문에 응답해 주셔서 감사합니다 ●

별 첨 2. 실태조사 결과 표

별첨2는 중소기업·스타트업의 컴퓨팅 인프라 현황과 향후수요에 대한 설문결과 데이터로 구성된다.

<표 별첨2-1> 인공지능 연구 분야

[Base: 전체 응답자(n=72), Unit: %]

		(Base)	사무 관리	공공 서비스	금융	교육	제조	에너지	유통	의료	국방	홈	교통	안전	법률	문화 관광	도시	농업	웰니스	기타
전체		(72)	**33.3**	**33.3**	**30.6**	**26.4**	**26.4**	**16.7**	**15.3**	**13.9**	**13.9**	**13.9**	**13.9**	**9.7**	**8.3**	**5.6**	**5.6**	**4.2**	**4.2**	**18.1**
설립년도	2000년 이전	(20)	25.0	15.0	20.0	25.0	15.0	15.0	5.0	10.0	15.0	10.0	20.0	15.0	5.0	0.0	15.0	0.0	5.0	25.0
	2000-2010년	(25)	44.0	52.0	28.0	20.0	24.0	16.0	16.0	8.0	20.0	12.0	12.0	4.0	8.0	8.0	0.0	0.0	4.0	16.0
	2010년 이후	(27)	29.6	29.6	40.7	33.3	37.0	18.5	22.2	22.2	7.4	18.5	11.1	11.1	11.1	7.4	3.7	11.1	3.7	14.8
매출액	10억 미만	(17)	29.4	29.4	35.3	5.9	23.5	5.9	17.6	29.4	11.8	29.4	5.9	23.5	11.8	5.9	5.9	5.9	11.8	17.6
	10-30억 미만	(17)	29.4	41.2	23.5	35.3	35.3	29.4	11.8	11.8	17.6	11.8	5.9	0.0	11.8	11.8	0.0	5.9	0.0	5.9
	30-50억 미만	(6)	33.3	33.3	33.3	50.0	33.3	16.7	16.7	0.0	16.7	33.3	33.3	0.0	0.0	0.0	0.0	0.0	0.0	33.3
	50-80억 미만	(4)	75.0	50.0	50.0	50.0	25.0	25.0	25.0	25.0	25.0	0.0	50.0	25.0	0.0	0.0	0.0	0.0	0.0	25.0
	80-120억 미만	(3)	33.3	100.0	66.7	66.7	0.0	33.3	66.7	0.0	0.0	0.0	33.3	0.0	33.3	0.0	0.0	0.0	0.0	33.3
	120-200억 미만	(9)	44.4	33.3	44.4	11.1	0.0	11.1	11.1	11.1	33.3	0.0	22.2	0.0	11.1	0.0	22.2	0.0	0.0	11.1
	200-500억 미만	(8)	25.0	12.5	12.5	37.5	37.5	12.5	0.0	0.0	0.0	12.5	0.0	0.0	0.0	12.5	0.0	12.5	0.0	25.0
	500억 이상	(8)	25.0	12.5	12.5	12.5	37.5	12.5	12.5	12.5	0.0	0.0	12.5	25.0	0.0	0.0	12.5	0.0	12.5	25.0
딥러닝 활용	비활용	(19)	31.6	26.3	36.8	26.3	36.8	10.5	21.1	10.5	21.1	15.8	10.5	10.5	15.8	10.5	10.5	5.3	5.3	26.3
	활용	(31)	32.3	**41.9**	29.0	**35.5**	22.6	**25.8**	6.5	**19.4**	12.9	9.7	16.1	6.5	9.7	3.2	3.2	0.0	3.2	9.7
	활용 예정	(22)	36.4	27.3	27.3	13.6	22.7	9.1	22.7	9.1	9.1	18.2	13.6	13.6	0.0	4.5	4.5	9.1	4.5	22.7
GPU 활용	비활용	(31)	29.0	29.0	35.5	25.8	29.0	12.9	19.4	12.9	16.1	12.9	9.7	9.7	16.1	9.7	9.7	6.5	9.7	19.4
	활용	(26)	34.6	**42.3**	34.6	**34.6**	23.1	**30.8**	7.7	**19.2**	11.5	11.5	19.2	7.7	3.8	3.8	0.0	0.0	0.0	15.4
	활용 예정	(15)	40.0	26.7	13.3	13.3	26.7	0.0	20.0	6.7	13.3	20.0	13.3	13.3	0.0	0.0	6.7	6.7	0.0	20.0
정부지원 인지 여부	인지	(26)	26.9	34.6	34.6	30.8	26.9	19.2	11.5	19.2	11.5	7.7	23.1	15.4	11.5	7.7	7.7	7.7	0.0	7.7
	비인지	(46)	37.0	32.6	28.3	23.9	26.1	15.2	17.4	10.9	15.2	17.4	8.7	6.5	6.5	4.3	4.3	2.2	6.5	23.9

<표 별첨2-2> 인공지능 연구 활용 데이터 종류

[Base: 전체 응답자(n=72), Unit: %]

		(Base)	정형 텍스트	비정형 텍스트	이미지 및 영상	신호	음성 및 청각데이터	빅데이터
전체		(72)	**50.0**	**45.8**	**43.1**	**25.0**	**18.1**	**1.4**
설립년도	2000년 이전	(20)	40.0	30.0	55.0	25.0	20.0	0.0
	2000-2010년	(25)	60.0	52.0	36.0	28.0	20.0	0.0
	2010년 이후	(27)	48.1	51.9	40.7	22.2	14.8	3.7
매출액	10억 미만	(17)	47.1	52.9	35.3	23.5	5.9	0.0
	10-30억 미만	(17)	58.8	52.9	47.1	41.2	23.5	0.0
	30-50억 미만	(6)	16.7	33.3	66.7	0.0	33.3	16.7
	50-80억 미만	(4)	50.0	50.0	25.0	25.0	25.0	0.0
	80-120억 미만	(3)	66.7	66.7	33.3	33.3	33.3	0.0
	120-200억 미만	(9)	55.6	55.6	66.7	22.2	33.3	0.0
	200-500억 미만	(8)	75.0	37.5	25.0	0.0	12.5	0.0
	500억 이상	(8)	25.0	12.5	37.5	37.5	0.0	0.0
딥러닝 활용	비활용	(19)	52.6	42.1	31.6	21.1	10.5	5.3
	활용	(31)	45.2	**54.8**	**58.1**	22.6	**32.3**	0.0
	활용 예정	(22)	54.5	36.4	31.8	31.8	4.5	0.0
GPU 활용	비활용	(31)	54.8	41.9	25.8	22.6	16.1	3.2
	활용	(26)	50.0	**53.8**	**57.7**	26.9	**26.9**	0.0
	활용 예정	(15)	40.0	40.0	53.3	26.7	6.7	0.0
정부지원 인지 여부	인지	(26)	53.8	57.7	53.8	26.9	23.1	0.0
	비인지	(46)	47.8	39.1	37.0	23.9	15.2	2.2

<표 별첨2-3> 인공지능 연구 활용 데이터 양

[Base: 전체 응답자(n=72), Unit: %, TB]

		(Base)	1TB 미만	1-2TB 미만	2-5TB 미만	5-50TB 미만	50TB 이상	모름	평균	합계
전체		(72)	**9.7**	**31.9**	**12.5**	**25.0**	**18.1**	**2.8**	**26.76**	**1,873**
설립년도	2000년 이전	(20)	10.0	45.0	15.0	10.0	15.0	5.0	**31.79**	604
	2000-2010년	(25)	8.0	28.0	12.0	32.0	20.0	0.0	26.46	**662**
	2010년 이후	(27)	11.1	25.9	11.1	29.6	18.5	3.7	23.36	607
매출액	10억 미만	(17)	17.6	29.4	17.6	17.6	17.6	0.0	26.67	453
	10-30억 미만	(17)	0.0	35.3	11.8	29.4	23.5	0.0	21.06	358
	30-50억 미만	(6)	0.0	50.0	0.0	33.3	16.7	16.7	4.60	23
	50-80억 미만	(4)	0.0	25.0	0.0	50.0	25.0	0.0	22.75	91
	80-120억 미만	(3)	33.3	33.3	0.0	33.3	0.0	0.0	3.67	11
	120-200억 미만	(9)	11.1	44.4	22.2	0.0	22.2	0.0	34.72	313
	200-500억 미만	(8)	0.0	25.0	12.5	25.0	37.5	12.5	25.00	175
	500억 이상	(8)	25.0	12.5	12.5	37.5	12.5	0.0	56.13	449
딥러닝 활용	비활용	(19)	10.5	42.1	10.5	10.5	15.8	10.5	**38.51**	**655**
	활용	(31)	9.7	29.0	12.9	29.0	19.4	0.0	18.18	563
	활용 예정	(22)	9.1	27.3	13.6	31.8	18.2	0.0	29.76	655
GPU 활용	비활용	(31)	16.1	35.5	16.1	12.9	12.9	6.5	25.82	749
	활용	(26)	3.8	26.9	7.7	34.6	26.9	0.0	**29.00**	**754**
	활용 예정	(15)	6.7	33.3	13.3	33.3	13.3	0.0	24.67	370
정부지원 인지 여부	인지	(26)	11.5	26.9	11.5	26.9	23.1	0.0	22.93	596
	비인지	(46)	8.7	34.8	13.0	23.9	15.2	4.3	29.02	1,277

<표 별첨2-4> 향후 인공지능 연구 활용을 위한 확보 데이터 양

[Base: 전체 응답자(n=72), Unit: %, TB]

		(Base)	1TB 미만	1-2TB 미만	2-5TB 미만	5-50TB 미만	50TB 이상	모름	평균	합계
전체		(72)	**2.8**	**8.3**	**13.9**	**41.7**	**30.6**	**2.8**	**129.76**	**9,083**
설립년도	2000년 이전	(20)	0.0	15.0	10.0	55.0	15.0	5.0	64.21	1,220
	2000-2010년	(25)	4.0	12.0	16.0	36.0	32.0	0.0	101.06	2,527
	2010년 이후	(27)	3.7	0.0	14.8	37.0	40.7	3.7	205.26	5,337
매출액	10억 미만	(17)	5.9	0.0	23.5	35.3	35.3	0.0	164.16	2,791
	10-30억 미만	(17)	0.0	5.9	11.8	41.2	41.2	0.0	57.35	975
	30-50억 미만	(6)	0.0	33.3	16.7	0.0	33.3	16.7	420.90	2,105
	50-80억 미만	(4)	0.0	0.0	0.0	75.0	25.0	0.0	36.25	145
	80-120억 미만	(3)	33.3	0.0	0.0	66.7	0.0	0.0	5.00	15
	120-200억 미만	(9)	0.0	22.2	22.2	22.2	33.3	0.0	114.33	1,029
	200-500억 미만	(8)	0.0	0.0	12.5	50.0	25.0	12.5	164.57	1,152
	500억 이상	(8)	0.0	12.5	0.0	75.0	12.5	0.0	109.00	872
딥러닝 활용	비활용	(19)	0.0	15.8	10.5	52.6	10.5	10.5	85.16	1,448
	활용	(31)	3.2	6.5	19.4	32.3	38.7	0.0	116.55	3,613
	활용 예정	(22)	4.5	4.5	9.1	45.5	36.4	0.0	**182.84**	**4,023**
GPU 활용	비활용	(31)	6.5	9.7	12.9	45.2	19.4	6.5	58.27	1,690
	활용	(26)	0.0	7.7	11.5	38.5	42.3	0.0	131.11	3,409
	활용 예정	(15)	0.0	6.7	20.0	40.0	33.3	0.0	**265.63**	**3,985**
정부지원 인지 여부	인지	(26)	3.8	3.8	23.1	23.1	46.2	0.0	164.65	4,281
	비인지	(46)	2.2	10.9	8.7	52.2	21.7	4.3	109.14	4,802

<표 별첨2-5> 딥러닝 활용 여부

[Base: 전체 응답자(n=72), Unit: %]

		(Base)	활용하고 있지 않음	활용하고 있음	활용할 예정임
전체		(72)	**26.4**	**43.1**	**30.6**
설립년도	2000년 이전	(20)	45.0	35.0	20.0
	2000-2010년	(25)	16.0	52.0	32.0
	2010년 이후	(27)	22.2	40.7	37.0
매출액	10억 미만	(17)	17.6	35.3	**47.1**
	10-30억 미만	(17)	17.6	**70.6**	11.8
	30-50억 미만	(6)	50.0	16.7	33.3
	50-80억 미만	(4)	25.0	50.0	25.0
	80-120억 미만	(3)	0.0	66.7	33.3
	120-200억 미만	(9)	33.3	44.4	22.2
	200-500억 미만	(8)	25.0	37.5	37.5
	500억 이상	(8)	50.0	12.5	37.5
딥러닝 활용	비활용	(19)	100.0	0.0	0.0
	활용	(31)	0.0	100.0	0.0
	활용 예정	(22)	0.0	0.0	100.0
GPU 활용	비활용	(31)	54.8	16.1	29.0
	활용	(26)	3.8	**88.5**	7.7
	활용 예정	(15)	6.7	20.0	73.3
정부지원 인지 여부	인지	(26)	3.8	69.2	26.9
	비인지	(46)	39.1	28.3	32.6

<표 별첨2-6> GPU 활용 여부

[Base: 전체 응답자(n=72), Unit: %]

		(Base)	활용하고 있지 않음	활용하고 있음	활용할 예정임
전체		(72)	**43.1**	**36.1**	**20.8**
설립년도	2000년 이전	(20)	65.0	25.0	10.0
	2000-2010년	(25)	32.0	40.0	28.0
	2010년 이후	(27)	37.0	40.7	22.2
매출액	10억 미만	(17)	47.1	35.3	17.6
	10-30억 미만	(17)	17.6	64.7	17.6
	30-50억 미만	(6)	50.0	16.7	33.3
	50-80억 미만	(4)	25.0	50.0	25.0
	80-120억 미만	(3)	33.3	66.7	0.0
	120-200억 미만	(9)	55.6	11.1	33.3
	200-500억 미만	(8)	50.0	25.0	25.0
	500억 이상	(8)	75.0	12.5	12.5
딥러닝 활용	비활용	(19)	89.5	5.3	5.3
	활용	(31)	16.1	**74.2**	9.7
	활용 예정	(22)	40.9	9.1	50.0
GPU 활용	비활용	(31)	100.0	0.0	0.0
	활용	(26)	0.0	100.0	0.0
	활용 예정	(15)	0.0	0.0	100.0
정부지원 인지 여부	인지	(26)	19.2	57.7	23.1
	비인지	(46)	56.5	23.9	19.6

<표 별첨2-7> 현재 보유 인공지능 계산 GPU 규모

[Base: 전체 응답자(n=72), Unit: TFLOPS]

		(Base)	평균				합계			
				GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)		GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)
전체		(72)	**52.6**	**14.2**	**13.3**	**130.6**	**11,370**	**1,009**	**959**	**9,402**
설립년도	2000년 이전	(20)	11.4	19.2	12.3	3.7	684	364	246	74
	2000-2010년	(25)	9.6	12.7	14.4	1.7	720	317	361	42
	2010년 이후	(27)	123.0	12.1	13.1	343.9	9,966	328	353	9,286
매출액	10억 미만	(17)	97.5	21.7	6.3	264.4	4,970	369	107	4,494
	10-30억 미만	(17)	98.6	2.8	11.1	281.8	5,027	47	189	4,791
	30-50억 미만	(6)	4.8	0.9	13.7	0.0	87	5	82	0
	50-80억 미만	(4)	33.5	0.0	82.0	18.6	402	0	328	74
	80-120억 미만	(3)	5.8	17.3	0.0	0.0	52	52	0	0
	120-200억 미만	(9)	8.9	22.0	0.0	4.7	240	198	0	42
	200-500억 미만	(8)	23.8	39.7	31.8	0.0	571	317	254	0
	500억 이상	(8)	0.9	3.0	0.0	0.0	21	21	0	0
딥러닝 활용	비활용	(19)	40.4	3.3	3.0	114.9	2,303	62	57	2,184
	활용	(31)	69.3	18.7	25.4	164.5	6,447	562	787	5,099
	활용 예정	(22)	39.7	17.5	5.2	96.4	2,620	385	115	2,120
GPU 활용	비활용	(31)	26.5	6.0	4.5	69.1	2,468	187	139	2,141
	활용	(26)	82.2	19.8	31.5	196.1	6,413	494	820	5,099
	활용 예정	(15)	55.3	21.8	0.0	144.2	2,490	328	0	2,162
정부지원 인지 여부	인지	(26)	67.5	19.8	7.9	174.9	5,267	515	205	4,547
	비인지	(46)	44.2	11.0	16.4	105.5	6,103	494	754	4,855

<표 별첨2-8> 현재 보유 인공지능 계산 GPU 규모 - 추정

[Base: 전체 응답자(n=215/모집단), Unit: TFLOPS]

		(Base)	평균				합계			
				GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)		GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)
전체		(72)	**49.8**	**14.7**	**14.0**	**120.7**	**32,093**	**3,135**	**3,017**	**25,941**
설립년도	2000년 이전	(20)	11.4	18.2	12.0	4.7	2,021	1,036	707	278
	2000-2010년	(25)	10.8	13.8	17.0	1.5	2,492	1,067	1,311	113
	2010년 이후	(27)	116.8	13.1	12.7	324.7	27,580	1,032	998	25,550
매출액	10억 미만	(17)	97.5	21.7	6.3	264.4	15,788	1,173	339	14,276
	10-30억 미만	(17)	98.6	2.8	11.1	281.8	11,827	110	444	11,273
	30-50억 미만	(6)	4.8	0.9	13.7	0.0	291	17	273	0
	50-80억 미만	(4)	33.5	0.0	82.0	18.6	1,508	0	1,230	278
	80-120억 미만	(3)	5.8	17.3	0.0	0.0	347	347	0	0
	120-200억 미만	(9)	8.9	22.0	0.0	4.7	640	527	0	113
	200-500억 미만	(8)	23.8	39.7	31.8	0.0	1,643	912	731	0
	500억 이상	(8)	0.9	3.0	0.0	0.0	49	49	0	0
딥러닝 활용	비활용	(19)	44.9	3.4	2.6	128.7	7,221	183	142	6,897
	활용	(31)	59.2	18.3	27.7	132.2	16,552	1,664	2,578	12,310
	활용 예정	(22)	40.6	18.9	4.4	98.6	8,320	1,288	297	6,734
GPU 활용	비활용	(31)	27.6	5.5	3.6	73.8	7,628	509	335	6,784
	활용	(26)	69.6	21.9	33.6	154.1	16,686	1,694	2,682	12,310
	활용 예정	(15)	60.1	21.6	0.0	158.8	7,779	932	0	6,847
정부지원 인지 여부	인지	(26)	59.4	20.7	7.6	149.7	15,151	1,759	651	12,741
	비인지	(46)	43.5	10.8	18.2	101.6	16,942	1,376	2,366	13,200

<표 별첨2-9> 향후 희망하는 계산 GPU 규모

[Base: 전체 응답자(n=72), Unit: TFLOPS]

		(Base)	평균				합계			
				GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)		GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)
전체		(72)	**102.0**	**41.1**	**26.6**	**240.0**	**22,040**	**2,876**	**1,886**	**17,278**
설립년도	2000년 이전	(20)	51.2	30.9	41.4	82.7	3,069	588	828	1,654
	2000-2010년	(25)	26.5	55.7	20.8	3.8	1,989	1,394	500	95
	2010년 이후	(27)	209.6	34.4	20.7	575.1	16,981	894	558	15,529
매출액	10억 미만	(17)	152.9	37.6	12.1	409.0	7,798	640	205	6,954
	10-30억 미만	(17)	178.1	8.5	22.6	505.1	9,082	135	361	8,586
	30-50억 미만	(6)	33.7	87.5	13.7	0.0	607	525	82	0
	50-80억 미만	(4)	40.6	0.0	82.0	39.8	487	0	328	159
	80-120억 미만	(3)	9.0	20.8	2.7	3.5	81	62	8	11
	120-200억 미만	(9)	73.5	46.2	0.0	174.3	1,985	416	0	1,569
	200-500억 미만	(8)	61.6	72.2	112.8	0.0	1,479	577	902	0
	500억 이상	(8)	21.7	74.3	0.0	0.0	520	520	0	0
딥러닝 활용	비활용	(19)	79.6	0.8	4.7	233.2	4,537	16	90	4,431
	활용	(31)	143.0	37.3	51.9	344.0	13,303	1,082	1,558	10,664
	활용 예정	(22)	63.6	80.8	10.8	99.3	4,200	1,778	238	2,184
GPU 활용	비활용	(31)	74.2	25.8	8.5	188.4	6,904	801	262	5,841
	활용	(26)	148.8	33.4	64.9	353.1	11,604	801	1,624	9,180
	활용 예정	(15)	78.5	84.9	0.0	150.5	3,532	1,274	0	2,258
정부지원 인지 여부	인지	(26)	124.5	36.2	14.2	324.5	9,711	905	369	8,438
	비인지	(46)	89.3	43.8	33.7	192.2	12,328	1,971	1,517	8,840

<표 별첨2-10> 향후 희망하는 계산 GPU 규모 - 추정

[Base: 전체 응답자(n=215/모집단), Unit: TFLOPS]

		(Base)	평균				합계			
				GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)		GTX 1070	GTX 1080	GTX 1080Ti /GTX TITAN X (Pascal,10series)/ P100 (P40 포함)
전체		(72)	**95.1**	**40.5**	**26.6**	**219.3**	**61,312**	**8,520**	**5,651**	**47,141**
설립년도	2000년 이전	(20)	49.3	30.3	40.8	78.2	8,754	1,718	2,412	4,624
	2000-2010년	(25)	25.2	50.3	22.5	3.4	5,824	3,880	1,685	260
	2010년 이후	(27)	198.0	38.3	19.8	537.1	46,734	2,923	1,554	42,257
매출액	10억 미만	(17)	152.9	37.6	12.1	409.0	24,771	2,032	651	22,088
	10-30억 미만	(17)	178.1	8.5	22.6	505.1	21,369	318	849	20,202
	30-50억 미만	(6)	33.7	87.5	13.7	0.0	2,024	1,751	273	0
	50-80억 미만	(4)	40.6	0.0	82.0	39.8	1,826	0	1,230	596
	80-120억 미만	(3)	9.0	20.8	2.7	3.5	541	416	55	71
	120-200억 미만	(9)	73.5	46.2	0.0	174.3	5,293	1,109	0	4,183
	200-500억 미만	(8)	61.6	72.2	112.8	0.0	4,253	1,659	2,593	0
	500억 이상	(8)	21.7	74.3	0.0	0.0	1,235	1,235	0	0
딥러닝 활용	비활용	(19)	88.4	0.9	4.1	260.2	14,211	48	219	13,944
	활용	(31)	122.6	36.0	53.0	282.0	34,259	3,186	4,812	26,261
	활용 예정	(22)	62.7	77.4	9.1	101.6	12,843	5,286	620	6,936
GPU 활용	비활용	(31)	73.7	21.7	6.9	192.5	20,332	1,993	631	17,709
	활용	(26)	125.0	35.1	64.7	279.1	29,965	2,641	5,020	22,303
	활용 예정	(15)	85.1	90.1	0.0	165.3	11,015	3,887	0	7,129
정부지원 인지 여부	인지	(26)	105.0	35.8	13.5	266.6	26,792	2,964	1,146	22,682
	비인지	(46)	88.6	43.6	35.3	188.3	34,520	5,556	4,505	24,459

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

93000
9 788961 084048
ISBN 978-89-6108-404-8