LG 경제연구원

2018. 11. 21

인공지능(AI) 프로세서, 새로운 혁신의 원동력 될까

전승우

오늘날 인공지능은 글로벌 IT 산업의 핵심 화두로 부상하였다. 단순 작업을 넘어 회계나 법률, 진료 등 전문 영역까지 인공지능을 적용하는 사례도 늘고 있다. 미래에는 일상 생활의 거의 모든 분야에 걸쳐 인공지능이 직간접적으로 활용될 것으로 보인다.

그러나 기존 IT 시스템으로는 지속적인 인공지능 발전이 어렵다는 인식도 커졌다. 이런 한계를 넘어서기 위하여 인공지능을 집중적으로 지원할 수 있는 프로세서가 필요하다는 주장이 힘을 얻게 되었다. 글로벌 IT 기업을 중심으로 CPU를 사용하는 대신 인공지능 알고리즘을 전담 처리하는 프로세서를 사용하여 각종 제품 및 서비스를 위한 고성능 인공지능을 구현하려는 움직임이 두드러지고 있다.

현재 인공지능 프로세서 개발 및 활용을 위한 다양한 접근이 이루어지고 있다. 멀티미디어 컨텐츠를 지원하기 위해 등장한 GPU는 현재 가장 주목 받는 인공지능 프로세서다. 딥 러닝 등 인공지능 알고리즘을 효과적으로 처리할 수 있다는 사실이 알려지면서 폭발적 인기를 얻게 되었다. 또한 ASIC 기술을 활용하거나 용도에 맞게 하드웨어 특성을 변경할 수 있는 FPGA을 기반으로 각종 애플리케이션에 특화된 맞춤형 인공지능 프로세서를 만들려는 움직임도 늘고 있다. 나아가 인간 뇌의 신경망 구조와 작동 원리를 모방하여 만든 뉴로모픽 프로세서 역시 차세대 인공지능 프로세서로 각광받고 있다.

소프트웨어 중심의 인공지능 개발로는 지속적인 성능 고도화가 어렵다는 인식이 한층 커질 것으로 보인다. 혁신적인 인공지능을 만들기 위해서는 기존 IT 시스템에 대한 근본적 재검토가 필요하다는 의견이 많다. 소프트웨어는 물론 하드웨어, 특히 모든 IT 기기와 서비스의 중추를 이루는 반도체를 인공지능의 관점에서 접근하려는 노력이 미래 혁신의 원동력으로 부상할 전망이다.

인공지능 프로세서의 부상으로 IT 기업들의 반도체 진출 움직임이 심화될 가능성도 있다. 아직까지는 주력 사업의 인공지능 경쟁력 강화가 주된 목적이지만, 한편으로 인공지능 프로세서 개발이 새로운 사업 진출의 포석이 될 수 있다는 추측도 있다. 향후 많은 기업들이 자사의 인공지능 프로세서 역량을 어떻게 활용할지도 미래 IT 산업의 주요 관심사로 떠오를 것이다.

인공지능 프로세서 전략은 각 기업 차원의 다각적 관점에서 수립되어야 한다. 인공지능의 활용 목적, 필요한 인공지능 구현 방법에 따라 인공지능 프로세서에 대한 접근 방식은 상이하다. 인공지능 프로세서가 미래 인공지능 트렌드에 어떤 변화를 가져올 수 있는지, 그리고 이런 변화가 자사의 사업 영역에 미칠 파급 효과를 면밀히 검토해야 한다. 자사에 적합한 인공지능 전략의 청사진을 만들고 이를 토대로 인공지능 프로세서 역량 확보를 위한 여러 옵션을 마련하는 것이 자사의 인공지능 활용 성과를 극대화하는 방안이 될 수 있을 것이다.

1. 인공지능 프로세서의 부상

오늘날 인공지능은 글로벌 IT 산업의 핵심 화두로 부상하였다. 기계가 사람의 생각과 판단 능력을 가지는 시대가 현실화되고 있다는 전망이 줄을 잇고 있다. 이제는 단순 작업을 넘어 회계나 법률, 진료 등 전문 영역까지 인공지능을 활용하는 사례도 늘고 있다.

최근 들어 인공지능 수준은 더욱 빠르게 발전하고 있다. 음성이나 이미지를 인식하고 분류하는 것은 물론 언어 번역, 자율주행, 기사 작성 등도 수행할 수 있다. 한발 더 나아가 인간의 감정까지 모방하는 컴퓨터 등 인공지능 확산이 거침없이 이루어지고 있다. 이런 추세라면 향후에는 일상 생활의 거의 모든 분야에 걸쳐 인공지능이 직간접적으로 활용될 전망이다. IT 업계뿐만이 아니라 금융, 농업, 자동차, 물류 등 IT 융복합이 활발히 이루어지는 분야의 기업들도 인공지능 기술 역량 축적에 집중하고 있다〈차트 1, 2〉.

특히 인공지능의 부상으로 반도체가 또 다시 큰 주목을 받고 있다. 반도체 산업은 PC와 스마트폰 등 새로운 기기가 등장할 때마다 성장을 거듭하였다. 이들 기기들은 시간이 지날수록 보다 많은 정보 저장 용량과 빠른 연산 실행 속도를 요구하기 때문에, 고성능 반도체를 찾는 수요가 꾸준히 증가하였다. 게다가 스마트폰 시장이 폭발적으로 성장하면서 제품을 작고 가볍게 만들기 위하여 작은 크기에 기능을 집약한 반도체를 개발하려는 기업들의 경쟁도 치열하게 전개되고 있다.

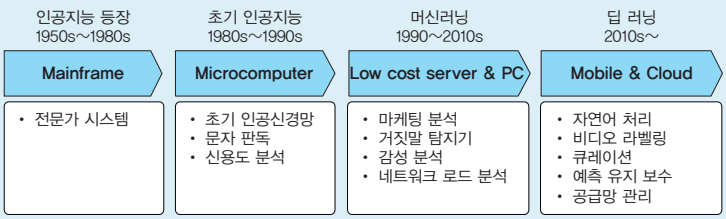
인공지능에 대한 높은 관심은 반도체 산업의 성장을 견인하고 있다. 인공지능의 발전은 무엇보다도 엄청난 정보 수집 및 처리 능력을 전제로 하기 때문이다. 정보를 저장하는 DRAM과 낸드플래시(NAND Flash) 등 메모리 반도체 수요가 폭증하는 가운데, 인공지능 알고리즘을 실행하는 시스템 반도체의 중요성도 커지고 있다. 스마트폰과 TV, 스마트 스피커 등 소비자 기기는 물론 클라우드 데이터센터에서도 인공지능 지원을 위하여 뛰어난 성능의 반도체가 많이 필요할 것

인공지능 적용 분야 1

	단말 기반	하이브리드	클라우드 기반
Consumer /Entertainment/Retail	개인용 VR/게임	개인 비서	광고 맞춤형 커머스
Transprotation /Infrastructure	자율주행차	수송, 원격 제어	교통, 네트워크 분석
Enterprise Operations	배달 드론, 창고 로봇	사이버 보안	판매, 마케팅, 고객 서비스
Oil&Gas/Agriculture	필드 드론, 로봇	기후, 수질, 에너지 제어	센싱 데이터 분석
Industrial/Military	로봇, 코봇, UAV	공정 제어/감시	공정 운영/분석
Medical/Healthcare	의료 이미지, 수질 로봇	의료 분석	건강 분석, 상담

자료: Moor Insights & Strategies

인공지능 발전 트렌드 2



으로 예상된다.

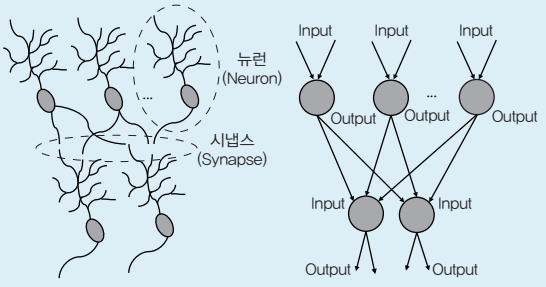
이런 가운데 최근에는 인공지능을 집중적으로 지원할 수 있는 반도체를 개발 및 탑재하려는 움직임이 두드러지고 있다. 인공지능 알고리즘을 효율적으로 구동할 수 있는 전용 반도체 프로세서(Processor)[1]를 사용하여 인공지능 성능을 대폭 끌어 올리는 것이다. 인공지능 저변이 폭넓게 확대되면서 인공지능 프로세서는 글로벌 IT 산업의 핵심 기술로 부상하고 있다.

인공지능 이론에 대한 연구가 활발히 이루어지면서 이를 구현하기 위한 소프트웨어가 관심을 받은 반면, 이를 뒷받침하는 하드웨어는 상대적으로 크게 부각되지 못했다. 그러나 인공지능에 대한 기대 수준이 나날이 커지면서 지금의 IT 시스템으로는 이를 충족하기 어렵다는 주장이 제기되었다. 간단한 인공지능 알고리즘의 탑재는 큰 문제가 없지만, 인공지능의 요구 성능 및 활용 범위가 커지면서 기존 하드웨어 기반으로는 지속적인 발전이 어렵다는 것이다.

특히 이런 문제는 오늘날 인공지능 발전을 주도하는 딥 러닝(Deep Learning) 구현에서 중요 이슈로 부상하였다. 딥 러닝은 인공지능의 주요 연구 분야인 머신러닝(Machine Learning)[2]을 구현할 수 있는 알고리즘으로, 인간 뇌의 신경망(Neural network) 구조를 모방한 것이 특징이다.

뇌신경은 뉴런(Neuron)이라 불리는 작은 세포 단위가 연속적으로 연결되어 있는 구조다. 뉴런은 시냅스(Synapse)라는 연결 세포를 통하여 앞에 연결된 뉴런으로부터 다수의 전기 자극을 입력 받아 이를 저장하거나 혹은 새로운 자극을 만들어 다음 뉴런으로 넘겨주게 된다. 1,000억 개 이상의 뉴런과 100조 개 이상의 시냅스는 이런 과정을 연속 반복하여 정보를 기억하거나 이를 토대로 다음에 해야 할 일을 판단할 수 있다〈차트 3〉.

신경망과 딥 러닝 비교 3



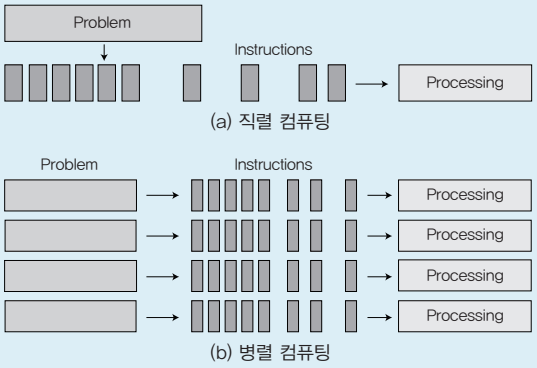
딥 러닝 알고리즘에서 뉴런의 역할을 담당하는 정보 입출력 연산은 그리 어려운 작업이 아니다. 그러나 가장 큰 문제는 수 천에서 수 만 개 이상의 이런 연산을 동시에 처리하는 과정을 반복해야 한다는 것이다. 이런 까닭에 딥 러닝 담당 프로세서의 병렬 컴퓨팅(Parallel computing)[3] 능력이 인공지능 성능을 결정하는 핵심 요건으로 부상하였다.

그러나 이런 특성은 현재 컴퓨터의 핵심 반도체인 중앙프로세서(CPU: Central

1 컴퓨터 동작을 위한 소프트웨어 프로그램을 처리하는 반도체
2 인공지능 연구 분야 중 하나로 컴퓨터가 방대한 데이터를 학습하고 이를 기반으로 새로운 상황을 판단, 예측하는 기술
3 수많은 정보 연산을 동시에 처리하는 컴퓨팅 기술

Processing Unit)의 구조에 적합하지 않다. CPU는 복잡하고 어려운 연산을 입력 순서에 따라 처리하는 직렬 컴퓨팅(Sequential computing) 구조다. 이런 까닭에 CPU가 딥 러닝을 구동하면 처리 속도가 느리고 필요 이상의 에너지를 과다 소모할 수 있다는 문제점이 지적되었다. 게다가 CPU가 본연의 역할 외 인공지능 업무까지 담당하게 되면서 과중한 부담으로 전체 시스템 성능이 저하될 수 있다는 우려도 커졌다〈차트 4〉.

직렬 컴퓨팅과 병렬 컴퓨팅 4



자료: BoA Merrill Lynch

이를 보완하고자 글로벌 IT 업계는 인공지능 구현에 적합한 프로세서를 전담 사용하는 인공지능 가속(AI Acceleration) 방법을 적용하고 있다. CPU 대신 인공지능 알고리즘을 처리할 수 있는 프로세서를 탑재하면 제품 및 서비스에 필요한 인공지능을 효율적으로 구현할 수 있다는 주장이 큰 호응을 얻었다. 따라서 최근까지 개념조차 뚜렷하지 않았던 인공지능 프로세서가 반도체 업계의 유망 테마로 떠올랐다.

많은 IT 기업들이 인공지능 프로세서에 큰 관심을 가지고 기술 확보에 뛰어들고 있다. 고성능 인공지능 구현이 학계 및 산업계의 주요 이슈로 부상하면서 인공지능 프로세서에 대한 연구와 투자가 집중되고 있다. 글로벌 반도체 전문 기업은 물론 세레브라스 시스템즈(Cerebras Systems)나 그래프코어(Graphcore) 등 관련 스타트업도 등장하여 인공지능 프로세서 개발 열기를 더하고 있다[4].

2. 인공지능 프로세서 개발 동향

(1) 그래픽처리장치(GPU)의 재조명

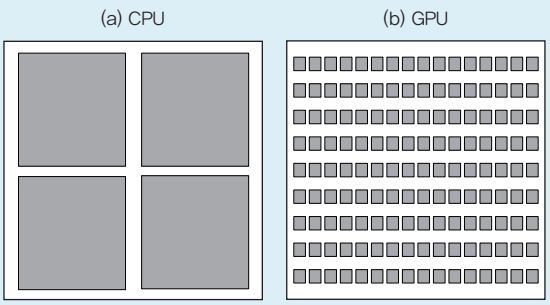
원래 그래픽처리장치(GPU: Graphic Processing Unit)는 CPU가 처리하는 작업을 분담하기 위한 목적으로 개발되었다. 반도체의 발전에 비해 소프트웨어 성능이 낮았던 과거에는 CPU가 전체 소프트웨어의 구동을 담당하였다. 그러나 컴퓨터에 탑재되는 소프트웨어의 크기가 커지고 유형도 다양해지면서 CPU만으로 모든 작업을 담당하기가 매우 어려워졌다.

특히 IT 산업의 성장에 따라 소프트웨어에서 고해상도의 화려한 그래픽이 적용되는 게임, 영화 등 멀티미디어 컨텐츠의 비중이 빠르게 늘었다. 멀티미디어 컨텐츠를 재

4 Aaron Tilley, "AI Chip Boom: This stealthy AI hardware startup is worth almost a billion", Forbes, 2017.08

생하기 위해서는 영상을 구성하는 개개의 픽셀(Pixel)을 표현하는 연산을 동시에 처리해야 한다. 이는 기존 CPU로 원활하게 수행하기 어렵기 때문에, 그래픽 처리를 전담하는 프로세서가 필요하다는 주장이 대두되었다. 신생 반도체 설계(Fabless)기업이었던 엔비디아(Nvidia)는 멀티미디어 컨텐츠 지원을 위하여 수 천 개 이상의 코어(Core)[5]을 연결하는 병렬 컴퓨팅 기술을 선보여 GPU라는 신규 영역을 개척하였다〈차트 5〉.

CPU와 GPU 구조 비교 5



초고화질과 3차원 영상 등 그래픽 기술이 나날이 발전하면서 GPU의 인기도 급상승했다. 컨텐츠 재생에 주로 사용되었던 GPU는 이후 각종 소프트웨어에도 적용되었다. 엔비디아는 GPU의 활용 가치가 무궁무진하다는 점에 착안하여 다목적 GPU 솔루션을 출시하여 큰 인기를 얻었다. 엔비디아가 2006년 선보인 GPU 솔루션 쿠다(CUDA)는 헤지펀드와 과학 연구소 등 여러 기관을 중심으로 복잡한 금융투자 및 기후 모델 시뮬레이션 등 다양한 분야에 활용되고 있다.[6]

무엇보다도 GPU는 강력한 병렬 컴퓨팅의 강점 때문에 인공지능, 특히 딥 러닝에 적합하다는 사실이 알려지면서 폭발적 인기를 얻게 되었다. 기존 CPU를 활용한 딥 러닝 구현에 어려움을 겪던 많은 기업들은 GPU의 잠재력에 주목하게 되었다. 구글과 페이스북 등 많은 IT 기업들이 자사의 비즈니스에 딥 러닝을 적용하기 위하여 GPU 사용을 크게 늘렸다.[7]

현재 GPU 시장의 과반 이상을 점유하고 있는 엔비디아는 데이터센터, 가전, 자동차 등 많은 기기의 인공지능 적용 열풍에 힘입어 매출이 급증하였다. 엔비디아는 자율주행 시스템 드라이브 PX(Drive PX) 등 인공지능 처리에 특화된 각종 GPU 솔루션을 출시하여 인공지능 프로세서 시장의 주도권을 강화하고 있다. 한편으로 AMD 등 다른 반도체 기업들 역시 인공지능 시대를 겨냥한 GPU 개발에 박차를 가하고 있다.

(2) 맞춤형 인공지능 프로세서의 부상

인공지능의 성장은 현재진행형이다. 인공지능의 주류 기술로 부상한 딥 러닝 역시 학계와 기업을 중심으로 변형 및 개선을 거듭하고 있다. 게다가 인공지능의 궁극적 목적인 뇌의 신비를 풀기 위하여 인간의 사고 체계, 뇌의 생물학 특성을 연구하는 움직

5 프로세서 내부에 포함되는 프로그램 처리 장치
6 Aaron Tilley, "The New Intel: How Nvidia went from powering video games to revolutionizing artificial intelligence", Fortune, 2016.11
7 Cade Mets, "The Race To Build An AI Chip For Everything Just Got Real", Wired, 2017.04

임도 활발하다. 이러한 지식의 축적은 새로운 인공지능 이론 및 상용화의 근간이 되기 때문에 인공지능 기술의 발전은 한층 지속될 전망이다.

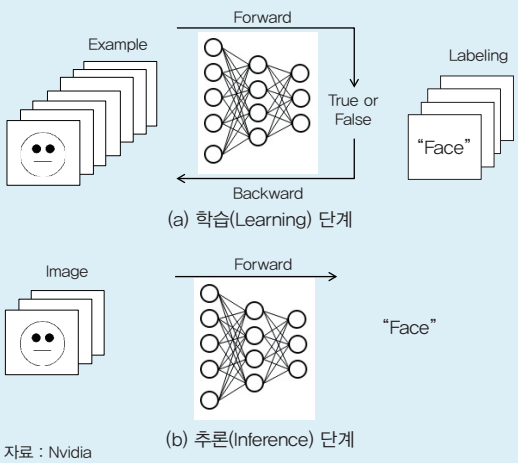
인공지능을 활용하는 어플리케이션도 꾸준히 늘고 있다. 간단한 음성이나 문자 인식은 물론 수많은 정보를 분석하여 시사점을 발굴하는 등 여러 적용 사례도 소개되고 있다. 그러나 동일한 원리의 인공지능이 적용되더라도 집중적으로 처리해야 할 정보 유형 및 처리 방법 등 구현 방식은 각 어플리케이션마다 상당히 다를 가능성이 높다. 만일 이런 특성을 고려하지 않는다면 인공지능 시스템은 간단한 작업을 처리하는 데에도 상당히 많은 시간과 에너지 자원을 낭비할 수 있다.

각 어플리케이션의 특성에 적합한 인공지능 시스템을 구현하기 위하여 범용 프로세서를 사용하는 대신, 최근에는 주문형 반도체(ASIC: Application Specific Integrated Circuit)[8] 기술로 프로세서를 개발하려는 움직임이 나타나고 있다. 이런 흐름은 인공지능의 쓰임새가 다양해지고 지금보다 훨씬 뛰어난 성능이 요구되면서 보다 뚜렷해지고 있다.

인간의 뇌는 지식을 배우고 이를 토대로 주어진 상황과 문제를 인식하여 판단할 수 있다. 유사한 원리로 딥 러닝도 크게 두 단계로 구성된다. 첫 번째는 학습(Learning) 단계로 딥 러닝의 신경망이 특정 작업을 수행하기 위해 필요한 기본적인 지식을 배우는 것이다. 두 번째는 추론(Inference) 단계로 학습을 거친 신경망이 외부 명령을 받거나 상황을 인식하면 학습한 내용을 토대로 가장 적합한 결과를 산출하는 것이다. 인공지능은 이 두 과정을 반복 실행하여 더 나은 답을 찾도록 성능을 강화할 수 있다〈차트 6〉.

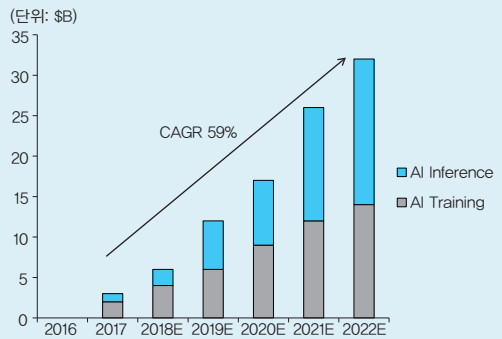
다른 유형의 반도체를 적용하려는 시도도 있지만, 방대한 데이터의 동시 처리가 필요한 학습 단계에서는 병렬 컴퓨팅 성능이 뛰어난 GPU가 가장 많이 사용될 것으로 예상된다. 그러나 추론 단계는 적용 분야에 따라 인공지능 활용 목적 및 기대 효과가 상이하기 때문에 GPU가 반드시 적합한 것은 아니다.[9] 또한 GPU와 동일한 작업을 수행하면서도 훨씬 적은 양의 에너지를 사용하는 프로세서에 대한 관심도 커졌다. 향후 인공지능 적용 확대에 따라 추론 프로세서의 중요성이

딥 러닝의 학습/추론 단계 6



자료 : Nvidia

글로벌 인공지능 프로세서 시장 7



자료: JP Morgan

8 사용자의 특정 용도에 맞게 개별적으로 제작된 반도체. 고속 처리 및 신뢰성 수준 높음.

9 Karl Freund, "A machine learning landscape: Where AMD, Intel, NVIDIA, Qualcomm and Xilinx AI engines live", Forbes, 2017.03

더욱 커질 가능성이 높기에,[10] 다수 기업들은 추론 기능을 중심으로 인공지능 프로세서 개발에 뛰어들고 있다〈차트 7〉.


구글의 텐서플로우 프로세서
자료: Google

구글(Google)은 원래 인공지능 바둑 프로그램 알파고(Alpha Go)를 구동하기 위하여 GPU를 집중적으로 사용하였다. 그러나 이후 개선된 알파고를 위하여 텐서플로우 프로세서(Tensor Flow Processor)라는 독자적 인공지능 프로세서를 개발하였다. 구글은 텐서플로우 프로세서가 기존 CPU와 GPU의 조합보다 15~30배나 빠르게 인공지능의 추론 기능을 수행할 수 있으며, 에너지 소모량도 수십 배 이상 적다고 주장한다.[11] 인텔(Intel) 역시 2016년 인수한 인공지능 프로세서 스타트업 모비디우스(Movidius), 너바나 시스템즈(Nervana Systems)의 기술을 활용하여 ASIC 기반 인공지능 프로세서 시장에 뛰어들었다.

스마트폰 제조 기업들도 독자적으로 인공지능을 위한 맞춤형 프로세서 제조에 뛰어들었다. 2017년 애플(Apple)은 아이폰의 3차원 안면 인식을 위한 인공지능 프로세서를 개발하여 스마트폰 AP(Application Processor)에 탑재하였다. 사람들은 노화나 안경 착용 등으로 타인의 얼굴 특징이 일부 변해도 누구인지를 쉽게 유추하고 인식할 수 있다. 그러나 주어지는 정보로만 판단할 수 있는 기존 컴퓨터에게 이는 매우 어려운 일이다.

이런 문제점을 극복하기 위해서는 사람처럼 얼굴 변화를 추론하여 인식하는 인공지능을 적용해야 한다. 이를 위하여 애플은 스마트폰 AP의 CPU나 GPU를 사용하는 대신 안면 인식을 집중적으로 처리할 수 있는 별도의 인공지능 프로세서 뉴럴 엔진(Neural Engine)을 독자적으로 설계하였다.

한편으로 각 제품이나 서비스에 적합한 인공지능을 구현하기 위하여 FPGA(Field Programmable Gate Array)를 사용하는 사례도 늘고 있다. FPGA는 설계 및 제조 시 특성이 결정되는 일반적 반도체와 달리, 사용자가 용도에 맞게 하드웨어 특성을 유연하게 변경할 수 있는 반도체다. 사실 FPGA는 1980년대부터 등장하였지만 다른 반도체보다 성능이 낮다는 단점 때문에 그간 활용이 제한적이었다.

그러나 인공지능이 중요한 트렌드로 부상하면서 FPGA를 기반으로 인공지능 프로세서를 만들려는 시도도 늘고 있다. 무엇보다도 FPGA가 하루가 다르게 발전하는 각

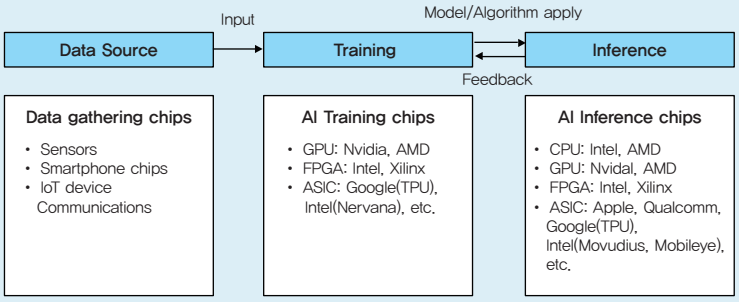
10 Gokul Hariharan, et.al, "Exponential growth from AI adoption in the cloud and at the edge", JP Morgan, 2018.02
11 Norman P. Jouppi, Cliff Young, et.al, "In-Datacenter performance analysis of a tensor processing unit", The 44th International Symposium on Computer Architecture (ISCA), 2017.06

종 인공지능 알고리즘을 적시적으로 지원하기 용이하다는 점이 큰 주목을 받았다. 게다가 기술 수준이 발전하면서 인공지능 구현을 위한 FPGA의 정보 처리 능력도 훨씬 용이해졌다.

이미 많은 기업들이 FPGA에 큰 관심을 보이고 있다. 마이크로소프트(Microsoft)는 애저(Azure) 클라우드 컴퓨팅과 검색서비스 빙(Bing) 등을 지원하는 데이터센터에 FPGA를 탑재하였다. 인텔은 2015년 FPGA 전문 기업 알테라(Altera)를 인수하고 이를 활용한 신제품을 출시하는 등 FPGA를 인공지능 프로세서의 중요 역량으로 간주하고 있다.[12]

ASIC이나 FPGA의 한계를 지적하는 주장도 있다. ASIC으로 만든 프로세서는 특정 업무 수행에 최적화되어 있기 때문에 인공지능이 필요한 다른 분야에 확장 적용하기 쉽지 않다. 따라서 각 제품이나 기능 별로 필요한 ASIC 프로세서를 일일이 새롭게 제작해야 하는 어려움이 따를 수 있다. FPGA 역시 인공지능의 수준이 증가할수록 복잡하고 훨씬 많은 양의 데이터를 처리해야 하기 때문에, 연산 성능을 더욱 끌어 올리는 것이 중요한 과제로 지적된다〈차트 8, 9〉.

딥 러닝 프로세스 및 관련 반도체 8



자료: JP Morgan

주요 프로세서 특성 비교 9

타입	장점	단점	주요 기업
CPU	· 서버와 PC 등 대부분의 범용 컴퓨터에서 사용 가능	· 병렬 컴퓨팅 처리 어려움	인텔, AMD, ARM
GPU	· 고수준 병렬 처리 가능 · 주요 AI 시스템에 적용 중	· 시스템 특성에 맞게 활용하지 못하면 비효율적 · FPGA보다 확장성 낮음	엔비디아, AMD
FPGA	· 하드웨어 특성 재구성 가능 · 지속적으로 성능 요구 수준이 증가하는 작업에 적합	· GPU 대비 낮은 연산 성능 · 주요 AI 시스템에서 활용 빈도 낮음	인텔, 자일링스
ASIC	· 인공지능 처리 성능 및 에너지 효율이 가장 우수	· 비싼 가격 · 활용 분야의 제약 높음	인텔, 구글, 애플, Graphcore

자료: Morningstar

(3) 인간의 뇌를 닮은 프로세서

오늘날 컴퓨터는 폰 노이만 구조(Von Neumann Architecture)라는 컴퓨터 구조 설계 이론에 근간을 두고 있다. 원래 폰 노이만 구조는 정확한 정보와 논리 체계로 구성된 프로그램을 효율적으로 처리하기 위하여 설계되었다. IT 산업의 초창기 이래 폰 노이만 구조는 방대한 정보를 더 빠르게 분석하고 결과를 산출할 수 있도록 발전을 거듭하면서 컴퓨터 시스템 설계의 표준으로 자리잡았다.

스마트폰과 데이터센터 등 인공지능을 탑재하는 IT 시스템 대부분은 폰 노이만 구조를 유지하면서 GPU 등 부가 프로세서를 활용하여 딥 러닝을 수행하고 있다. 즉

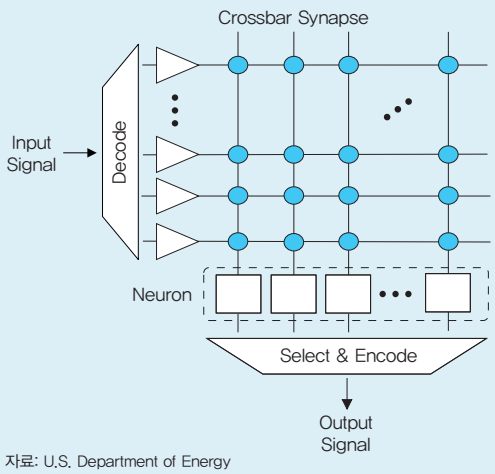
12 "The rise of artificial intelligence is creating new variety in the chip market, and trouble for Intel", The Economist, 2017.02

CPU가 전체 시스템의 동작과 제어를 담당하고, 인공지능 프로세서는 딥 러닝 기능을 집중적으로 처리하도록 구성된 이기종 시스템(HSA: Heterogeneous System Architecture) 방법을 채택하고 있다.[13]

이런 방법은 기존 시스템 구조의 근간을 유지하면서 인공지능을 구현하기 용이한 반면, 인공지능 성능을 강화하기 위하여 비례적으로 더욱 많은 프로세서가 필요하다는 점이 단점으로 지적된다. 2012년 구글의 인공지능 소프트웨어가 아무 정보도 주어지지 않은 상태에서 고양이를 인식하는 법을 배우기 위해서는 무려 16,000개의 프로세서가 필요했다고 한다. 수많은 프로세서를 집적하는 것은 시스템 규모의 증가, 설계 및 개발의 어려움 및 높은 에너지 소모 등 여러 문제를 초래할 수 있다.[14]

이런 이유로 학계를 중심으로 인간의 뇌신경 구조 자체를 그대로 모방한 프로세서를 만들어야 한다는 주장이 등장하였다. 뉴런과 뉴런이 시냅스로 촘촘하게 엮이듯이, 인공지능 프로세서도 정보 처리를 담당하는 개개의 개별 소자가 네트워크로 연결되어 학습 및 의사 결정을 담당하는 구조로 설계되어야 한다는 것이다. 즉 지금까지 소프트웨어로 만들었던 딥 러닝을 반도체 집적회로 기술로 구현하려는 접근법이다〈차트 10〉.

뉴로모픽 프로세서의 예시적 구조 10



자료: U.S. Department of Energy

뇌신경을 닮은 프로세서라는 의미에서 이런 특징을 갖춘 프로세서를 뉴로모픽 프로세서 (Neuromorphic Processor)라 정의한다. 뉴로모픽 프로세서는 수십 년 전 제안된 아이디어지만 당시에는 인공지능에 대한 낮은 관심 탓에 그리 큰 주목을 받지 못했다. 그러나 딥 러닝의 인기가 폭발적으로 증가하면서 뉴로모픽 프로세서의 잠재력 역시 재조명 받고 있다.

많은 연구에도 불구하고 지금까지 뉴로모픽 프로세서는 본격적인 상용화에 도달하지 못했다. 무엇보다도 신경망처럼 수천만에서 수 억 개가 훨씬 넘는 트랜지스터(Transistor)를 정교하게 연결할 수 있는 기술이 없었기 때문이다. 이런 까닭에 적은 수의 트랜지스터로 제작한 초창기 뉴로모픽 프로세서는 기대와 달리 성능 수준이 매우 낮았다. 그러나 반도체 설계 및 공정 기술이 발전하면서 뉴로모픽 프로세서를 현실로 만들 수 있는 가능성을 발견하게 되었다.

뉴로모픽 프로세서가 각광받는 이유는 바로 언제 어디서나 인공지능을 활용하는 인공지능 확산(Pervasive AI) 시대를 열 수 있는 잠재력이 크기 때문이다. 딥 러닝은 거의 무한정에 가까운 정보를 학습하기 때문에 뛰어난 인식 및 판단 능력을 가질 수

13 최새솔, "인공지능 반도체 산업동향 및 이슈 분석", ETRI, 2017.12
14 Robert D. Hof, "10 breakthrough technologies 2014: Neuromorphic chips", MIT Technology Review, 2014.05

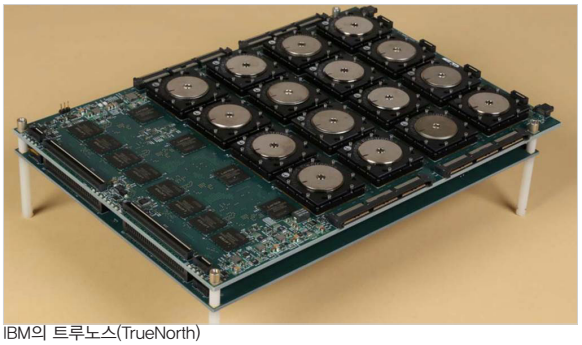
있는 반면, 막대한 에너지 소모를 동반한다. 구글은 알파고를 구동하기 위해서 1,200여 개의 CPU와 176개의 GPU, 920TB(Tera Bytes)의 DRAM 등 엄청난 양의 반도체를 장착한 슈퍼 컴퓨터를 사용했는데, 이 컴퓨터의 에너지 사용량은 무려 12GW에 달했다. 반면 사람이 일상적으로 정보를 기억하고 판단하기 위한 이론적 에너지 소모량은 약 20W에 불과하다고 한다.

미래 인공지능 시대에는 데이터 처리 성능 자체보다 전력 효율성을 더 강조할 수 있다. 모바일이나 가전, 혹은 작은 사물인터넷 기기 하나하나마다 인공지능이 적용될 수 있기 때문이다. 뉴로모픽 프로세서는 신경망 구조를 내재화하고 있기 때문에 이론적으로는 인간처럼 아주 적은 전력만으로도 고성능 인공지능을 수행할 수 있다. 만일 뉴로모픽 프로세서가 상용화된다면 인공지능의 저변이 예상보다 더욱 빠르게 확산될 것이라는 전망도 있다.

현재 기술로는 인간의 신경망을 완벽하게 모사할 수 있는 이상적 뉴로모픽 프로세서를 만들기 어렵다는 것이 대체적 중론이다.[15] 인간 뇌 수준의 뉴로모픽 프로세서를 만들기 위해서는 신경망 구조에 대한 심층적 이해는 물론, 뉴런과 시냅스의 역할을 담당하는 정교한 개별 소자 개발까지 고려해야 하기 때문이다. 기존 반도체 기술로는 뉴런과 시냅스가 실제로 정보를 받아들이고 처리하는 것과 근사한 메커니즘을 구현하기 어렵다. 이런 까닭에 현재 고성능 뉴로모픽 프로세서를 만들기 위하여 멤리스터(Memrister)[16] 등 차세대 소자에 대한 연구도 활발히 이루어지고 있다.

뉴로모픽 프로세서를 활용하여 인공지능 혁신을 가속화할 수 있다는 주장이 재조명되면서 전세계적으로 뉴로모픽 프로세서를 개발하려는 움직임이 등장하고 있다. IBM은 2014년 인간의 뇌를 모방하는 시스템을 연구한 미국 국방부 방위고등연구계획국(DARPA)의 프로젝트 결과를 활용하여 뉴로모픽 프로세서 트루노스(TrueNorth)를 개발하였다. 트루노스는 100만 개의 뉴런과 2억 6천만 개 시냅스로 이루어진 신경망 구조를 가지고 있는 것으로 알려진다.[17]

퀄컴(Qualcomm) 역시 뉴로모픽 프로세서 연구에 뛰어들었다. 퀄컴은 자사의 인공지능 역량을 강화하기 위하여 제로스(Zeroth) 프로젝트를 추진하였다. 이를 통하여 뉴로모픽 프로세서 기술을 연구하였으며, 이중 일부를 자사의 스마트폰 AP 스냅드래곤(Snapdragon)에 적용하였다. 인공지능에 대한 투자를 확대하고 있는 인텔도 캘리포니아 공과대학(Caltech) 등과 협력하여 뉴로모픽 프로



IBM의 트루노스(TrueNorth)
자료: 위키피디아

15 Lee Gomes, "Neuromorphic chips are destined for deep learning–or obscurity", IEEE Spectrum, 2017.05
16 메모리(Memory)와 레지스터(Resistor)의 합성어로 빠른 속도로 정보를 저장할 수 있는 메모리를 만드는 소자
17 Vivek Arya, et.al, "Deep learning and the processor chips fueling the AI revolution", BoA Merrill Lynch, 2016.10

세서 로이히(Loihi)을 개발하며 관련 기술 역량 축적에 나서고 있다.

3. 시사점

(1) 하드웨어 중심 인공지능 개발 트렌드 확산

지금까지의 인공지능은 주로 소프트웨어와 서비스를 중심으로 발전하였다. 딥 러닝 등 새로운 알고리즘과 인공지능을 활용한 다양한 서비스의 등장은 그간 큰 주목을 받지 못했던 인공지능을 IT 산업의 중심으로 끌어올리는 데 결정적 역할을 수행했다. 그러나 향후에는 시스템의 근간인 하드웨어 차원에서 인공지능에 접근하기 위한 노력도 많은 관심을 받게 될 것이다.

기존의 IT 시스템은 인간의 판단과 명령을 기반으로 빠르고 정확하게 정보를 처리할 수 있는 수동성을 전제로 한다. 반면 인공지능 시스템은 인간의 개입이 최소화되고 모든 정보 처리가 시스템 스스로 이루어지는 자율성을 지향한다. 이를 위해서는 시스템이 풍부한 데이터를 학습 및 활용할 수 있는 능력을 필수적으로 가져야 한다. 특히 자율성 수준이 높아질수록 데이터 처리 능력도 훨씬 향상되어야 한다.

그러나 소프트웨어 중심 인공지능 개발로는 이런 능력을 지속적으로 강화하기 어렵다. 지금보다 뛰어난 인공지능을 만들기 위해서는 기존 IT 시스템에 대한 근본적 재검토가 필요하다는 의견이 많다. 이런 이유로 소프트웨어는 물론 하드웨어, 특히 모든 제품과 서비스의 중추를 이루는 반도체를 인공지능의 관점에서 접근하는 인공지능 프로세서의 중요성이 강조될 수 있다.

역사적으로 반도체 시장은 IT 산업의 변화를 촉발하는 패러다임의 등장으로 큰 변곡점을 겪었다. 인터넷, 이동통신, 빅데이터, 사물인터넷 등 새로운 패러다임이 부상할 때마다 정보 저장 및 처리 역량이 크게 강조되었다. 이에 대응하기 위하여 메모리와 각종 시스템 반도체의 성능을 강화하려는 움직임이 꾸준히 이어져 왔다. 인공지능 역시 다른 어떤 기술보다도 많은 양의 정보 처리 능력을 필요로 하기 때문에, 첨단 인공지능을 만들기 위한 반도체의 중요성은 꾸준히 강조될 것이다. 많은 연구에서도 인공지능의 확산 수준이 미래 반도체의 소재, 설계 및 공정 기술에 대한 관심과 투자에 큰 영향을 미칠 것으로 예상한다.

인공지능의 발전으로 지금까지와 근본적으로 다른 기술적 접근이 각광받을 수 있다. 폰 노이만 구조 등 오랜 시간에 걸쳐 IT 업계를 지배해 온 컴퓨터 기술 메커니즘의 궤도와 다른 변화가 출현하는 것이다. 그래픽 처리 등 일부 분야에 주로 사용되었던

병렬 컴퓨팅 기술이 인공지능 트렌드를 맞이하여 재조명되었듯이, 뉴로모픽 프로세서 등 더욱 우수한 인공지능을 구현하려는 시도 역시 추상적 아이디어를 넘어 상용화 차원에서 큰 관심을 받게 될 전망이다. 이런 흐름에 대응하고자 현재 반도체 시장을 이끄는 기업들은 물론 새로운 비즈니스 기회를 모색하려는 기업들 역시 인공지능 프로세서 기술 개발에 박차를 가할 것으로 보인다.

(2) IT 기업의 반도체 진출 가속화

PC 시대를 지나 모바일 시대로 접어들면서 부품과 소프트웨어, 완제품 간 연관성은 강화되고 있다. 이제는 완제품의 차별화 경쟁력 제고를 위하여 소프트웨어 역량을 축적하는 것에서 나아가 자체 필요에 맞는 반도체까지 설계하려는 움직임이 늘고 있다. 애플은 모바일 AP에 이어 그간 외부 기업에 의존하였던 GPU나 전력관리반도체(PMIC) 등 아이폰에 포함되는 주요 반도체를 직접 설계하면서 반도체 내재화 영역을 넓히고 있다. 다른 기업들 역시 주력 제품의 성능 제고를 위해 반도체 기술 역량 확보에 큰 관심을 보이고 있다.

인공지능 프로세서의 부상으로 IT 기업들의 반도체 진출 움직임이 심화될 수 있다. 인공지능 구현이나 적용을 위한 기술 방식은 각 기업의 비즈니스 및 사용 환경마다 매우 다를 것으로 보인다. 아무리 뛰어난 인공지능 알고리즘을 개발하더라도 이를 뒷받침하는 반도체가 적합하지 않다면 원하는 성능 구현이 쉽지 않다. 게다가 인공지능을 구성하는 기술 역시 개선과 발전을 거듭하고 있기 때문에, 반도체 역량의 외부 의존으로는 변화 트렌드를 빠르게 대응하기 어렵다는 인식도 늘고 있다.

이런 까닭에 반도체가 주력이 아니었던 여러 기업들이 독자적인 인공지능 프로세서 개발을 모색하고 있다. 구글은 보다 빠르고 안정적으로 인공지능 서비스를 구현하기 위하여 자체 설계한 프로세서를 클라우드 데이터센터에 접목하였다. 이를 통하여 전체 시스템에 걸쳐 인공지능 구현 역량을 확보하는 동시에, 인텔과 엔비디아 등 반도체 전문 기업에 대한 의존도를 낮추는 발판을 마련할 것으로 보인다. 마이크로소프트는 증강현실 시스템 홀로렌즈(Hololens)를 위하여 인공지능 프로세서를 개발하고 있다. 자율주행차 시장이 확대될수록 인공지능 프로세서의 역할이 커질 것이라고 판단한 테슬라(Tesla) 역시 엔비디아의 자율주행 솔루션에서 벗어나 인공지능 프로세서를 개발하고 있다고 밝혔다.[18]

아마존(Amazon) 역시 인공지능 프로세서를 개발하고 있다는 추정도 있다.[19] 아마존 웹 서비스(Amazon Web Service)에서 인공지능 서비스 알렉사(Alexa)가 차지하는

18 Tom Simonite, "Musk says Tesla is building its own chip for autopilot", Wired, 2017.12
19 Matthew Lynley, "Amazon may be developing AI chips for Alexa", Techcrunch, 2018.02

비중이 늘고 있으며, 인공지능 스피커 에코(Echo) 등 하드웨어 사업에서도 인공지능의 중요성이 강조되고 있다. 게다가 주력 사업인 물류, 배송 등에서도 인공지능을 활용할 여지는 커질 전망이다. 이런 까닭에 아마존 역시 인공지능 역량의 전면적 강화, 특히 반도체 기반의 하드웨어 기술 확보에 큰 관심을 가진 것으로 보인다. 최근에는 페이스북(Facebook)도 인공지능 서비스 강화를 위하여 데이터센터에 탑재할 자체 프로세서 개발에 착수한 것으로 보인다[20]〈차트 11〉.

주요 기업의 인공지능 프로세서 개발 11

기업	장점
애플	· 2017년 안면인식 기능 구현을 위한 인공지능 프로세서 뉴럴엔진(Neural Engine)이 탑재된 스마트폰 AP A11 Bionic 발표
구글	· 2016년 클라우드 컴퓨팅의 인공지능 서비스를 지원하는 프로세서 TPU(Tensor Processing Unit) 탑재. 이후 지속적으로 TPU 버전 업그레이드 추진
마이크로소프트	· 증강현실 헤드셋 홀로렌즈에 언어와 영상 인식 기술을 구동하는 인공지능 프로세서 HPU(Holographic Processing Unit) 탑재 · FPGA를 자사의 캐터펄트(Catapult) 서버에 적용해 검색 엔진 및 인공지능 서비스 성능 강화 추진
화웨이	· 2017년 스타트업 캠브리콘(Cambricon)과 협력하여 인공지능 프로세서 NPU(Neural Processing Unit)를 탑재한 스마트폰 AP 기린 970 발표
바이두	· 클라우드 컴퓨팅용 인공지능 프로세서 쿤룬(Kunlun) 발표
테슬라	· 자율주행 기능 오토파일럿(Auto Pilot) 지원용 프로세서 개발

이들 기업들은 대체로 자사 사업의 인공지능 경쟁력 강화를 위하여 반도체 프로세서 개발에 나서고 있다. 그러나 한편으로 인공지능 프로세서가 새로운 사업 진출을 위한 포석이 될 수 있다는 예상도 있다. CPU나 스마트폰 AP 등 시스템 구성에 필수적인 반도체가 하나의 플랫폼이 되어 연관 하드웨어 및 소프트웨어의 성장과 발전을 이끌었듯이, 인공지능 프로세서도 인공지능 고도화는 물론 전후방 IT 산업의 판도 변화를 이끌 플랫폼으로 성장할 수 있다는 것이다. 이런 관점에서 많은 기업들이 인공지능 프로세서 역량을 어떻게 활용할지도 미래 IT 산업의 주요 관심사로 부상하게 될 것이다.

(3) 자사에 적합한 인공지능 프로세서 전략 수립 필요

미래에도 새로운 기술과 아이디어로 인공지능을 구현하려는 움직임은 계속될 전망이다. 인간에 근접한 수준의 인공지능을 만들기 위해서는 하드웨어와 소프트웨어 등 시스템 전반에 걸쳐 혁신이 이루어져야 한다는 의견이 많다. 이런 차원에서 인공지능 프로세서 등 새로운 접근법으로 인공지능의 잠재력을 실험하려는 기업들의 움직임도 빨라질 것이다.

인공지능과 연관 깊은 제품 및 서비스가 주력인 기업일수록 인공지능 프로세서 및 이를 통해 출현하게 될 새로운 트렌드와 밀접하게 관련될 것이다. 따라서 인공지능 프로세서의 확산이 자사의 인공지능 전략에 어떤 영향을 미칠지를 주의 깊게 관찰하는 것이 필요하다. 인공지능 프로세서가 미래 인공지능 트렌드에 어떤 변화를 가져올 수 있는지, 그리고 이런 변화가 자사의 사업 영역에 미칠 파급 효과를 면밀히 분석해야 할 것이다.

인공지능 프로세서 대응 전략은 각 기업 차원의 다각적 관점에서 수립되어야 한다.

20 Mark Gurman , "Facebook is forming a team to design its own chips", Bloomberg, 2018.04

인공지능의 적용 목적, 그리고 구현 방법에 따라 인공지능 프로세서에 대한 접근 방식은 다를 수 있다. 실제로 필요한 인공지능의 수준은 산업, 제품별로 상이하며 이를 활용하는 수단 역시 클라우드 컴퓨팅, 엣지 컴퓨팅(Edge computing)[21], 단말 기기 등 한층 다양해질 전망이다. 따라서 자사에 적합한 인공지능 전략의 청사진을 만들고 이를 토대로 인공지능 프로세서 역량 확보 옵션을 마련하는 것이 비즈니스의 인공지능 활용 효과를 극대화하는 방안이 될 수 있다.

예컨대 다목적으로 활용되는 고사양의 인공지능 프로세서보다는 특정 목적의 기기 탑재 등 필요한 요구 조건에 맞는 인공지능 프로세서 개발이 가시적 성과 달성을 위해 더 유리할 수 있다.[22] 또한 독자 프로세서 개발도 필요하지만, 인공지능 기술 유형이 다양화되는 점을 감안하여 인공지능 프로세서를 제조하는 다수 기업과의 유기적 협력을 기반으로 필요한 인공지능을 빠르게 자사 제품에 접목하는 전략도 고려할 수 있다. 나아가 인공지능이 시스템 전반의 주요 기능과 깊숙이 연관될 가능성이 높기 때문에, 자사에 필요한 인공지능 프로세서가 에너지 효율, 네트워크 연결성, 보안 등 여러 측면에 미칠 영향도 살펴 보아야 할 것이다.

미래 인공지능이 담당하게 될 역할은 더욱 늘어날 것이다. 간단한 업무를 넘어 지금까지 다루지 않았던 방대한 정보를 바탕으로 빠르고 정확한 판단과 실행이 필요한 전문 영역까지 적용되는 등 IT 기기 및 서비스의 인공지능 의존도는 심화될 전망이다. 인공지능 프로세서의 부상은 인공지능 고도화 시대의 중추적 기반이 될 수 있기 때문에, IT 산업은 물론 여러 분야의 기업들이 인공지능의 중요성을 각인하고 인공지능 기반 비즈니스 경쟁력 강화를 깊이 고민하는 계기가 될 것이다. www.lgeri.com

21 클라우드 컴퓨팅과 같은 중앙집중식 정보 처리와 달리, 데이터가 생성되는 단말과 근접한 곳에서 정보를 처리하는 기술
22 김용균, "반도체 산업의 차세대 성장엔진 AI 반도체 동향과 시사점", 정보통신기술진흥센터, 2018.01

LG 경제연구원