NRF R&D BRIEF

AI 반도체 회로 및 시스템_전기/전자_기초연구본부 선정 R&D 이슈 연구동향(3)

2020-24호 (2020.11.02.) / ICT·융합연구단

1. AI 반도체 회로 및 시스템이란?

○ **인공 지능 소프트웨어**를 보다 효과적으로 구동하는 **설계 기술 및 시스템 구현 기술**

☞ AI 반도체 회로 및 시스템은 학습 및 추론 등 인공지능 알고리즘 구현에 요구 되는 **대규모 데이터 처리를 위한 기존 반도체 회로 및 시스템의 한계점을 극복하기 위해 개발된 새로운 회로 및 시스템 기술**을 의미

☞ AI 반도체 회로 및 시스템의 범위는 인공 신경망 딥러닝 연산의 가속을 위한 디지털 기반 가속기 설계로부터 다양한 형태의 회로와 소자를 이용한 새로운 컴퓨팅에 기반한 시스템의 구현까지 포함

※ 출처 : KISTEP 기술동향브리프 (2019-01호, 인공지능(반도체))

2. 왜 주목받고 있나?

○ 10년 내 모든 기기에 AI 반도체... 한국 '제2의 D램 신화'쓴다. (매일경제 2020.5.5.)

☞ 인공지능 기술의 비약적인 발전으로 **향후 5년 내에 다양한 종류의 응용 프로그램들이 인공지능 기술에 기반하여 개발될 것**으로 예측

■ 이러한 인공지능 응용 프로그램들을 효율적으로 구동하기 위한 **인공지능 반도체 회로 및 시스템 구현 기술의 개발이 절실히 요구**

■ 특히 빅데이터 분석 처리 등에 활용될 대규모 연산 로직을 사용하는 AI 반도체의 경우 전력 사용량이 많아, **저전력, 고효율 회로 설계 기술 개발이 매우 시급**하고, **이를 효율적으로 활용하는 시스템의 개발이 매우 중요**

■ 한국은 반도체 산업 전반에 걸쳐 우수한 인재들이 산학연에 고루 포진되어 있고, 관련 기술 노하우가 많이 축적되어 있어 향후 AI 반도체 회로 설계 및 시스템 개발 산업에서 매우 높은 경쟁력을 갖고 있다고 봄

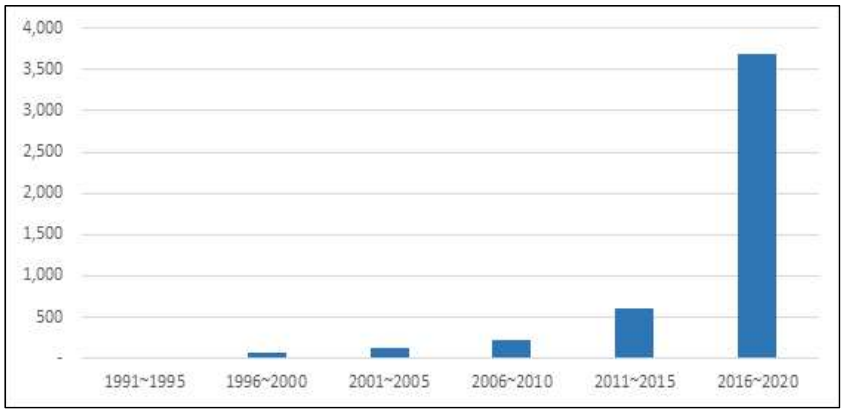
3. 최근 많은 연구가 이루어지고 있나?

○ 국내외를 막론하고 **최근 AI 반도체 회로 및 시스템에 대한 관심이 폭발적으로 늘어남**에 따라 많은 연구가 진행되고 있어 **관련 논문의 수가 지난 수년간 급증**하고 있음.

– AI 반도체 회로 및 시스템 관련 **일정 주기별 논문 출판 수 변화**(IEEExplore*)

* 세계 전기 전자 공학자 협회(IEEE) 논문 데이터베이스(deep learning accelerator, deep learning circuit, deep neural network circuit, neuromorphic circuit) keyword 로 검색

구분	1991~1995	1996~2000	2001~2005	2006~2010	2011~2015	2016~2020
논문 수(건)	20	62	118	226	614	3,690



4. 최근 국내·외 연구 동향은?

○ (미국) 대규모 뉴로모픽 칩 연구(The Robot Report 외 대중매체, 2020.8.19.)

■ IBM TrueNorth 칩의 뒤를 이어 Intel에서 **Loihi 라는 이름의 대규모 뉴로모픽 칩 발표** (2017.9.25.)
: 인텔사는 딥러닝에 특화되어 있는 대부분의 AI chip 과는 달리 인간의 뇌가 동작하는 방식을 보다 충실하게 구현하는 뉴로모픽 칩을 연구 개발
– 한 이스라엘 연구진은 인텔 Loihi 칩을 사용하여 환자를 돕는 로봇 팔 제작에 성공 (2020.8.19.)

○ (미국) Tensor Streaming Processor 기반 고성능 AI Chip 개발(EETimes 외 대중 매체, 2020.8.14.)

■ 미국 AI chip 스타트업 회사인 Groq은 2020년 International Symposium on Computer Architecture (ISCA) 학회에서 병렬연산을 위해 많은 수의 작은 코어를 사용하던 기존 신경망처리장치(NPU) 들과는 달리 **많은 수의 기능단위(functional unit)을 갖는 하나의 프로세서에 기반한 Tensor Streaming Processor (TSP)를 발표**
■ **기존 그래픽처리장치(GPU)나 신경망처리장치 들에 비해 ResNet50 처리 기준 4배 이상의 성능 향상 보고**
■ INT8 성능의 경우 1 PetaOp/second@1.25GHz 을 가짐을 보고
■ 기존 투자 금액 $67M 에 더해 2020년 8월, 알려지지 않은 금액의 추가 투자 유치 성공

○ (미국) 다양한 종류의 상용 AI 가속기 회로 및 시스템 개발 현황 (VentureBeat 외 대중매체, 2020.7.29.)

■ **구글, 애플, 테슬라와 같은 대기업의 경우 자체 AI Chip을 만들어 자사제품에 사용하는 경향**이 두드러짐
- 구글 : 자사 서버에서 수행하는 AI 컴퓨팅의 성능 향상을 위해 **전용 가속기 칩인 Tensor Processing Unit(TPU)를 개발**하여 실제 사업에 적용
- 애플 : 8-코어 뉴럴 엔진을 탑재한 Bionic A13 **칩을 개발**하여 아이폰 라인에 적용
- 테슬라 : **자사의 자율주행 자동차를 위한 전용칩 개발** 및 적용

■ 다양한 응용분야를 위한 AI **가속기 칩 스타트업 회사 창업** 열풍
- Cerebras Systems: 12인치 웨이퍼 전체를 하나의 칩을 만드는데 사용하여 고성능 AI 슈퍼컴퓨팅 시스템 개발
- Syntiant: 아마존 에코 스피커와 같은 on-device AI device 응용에 사용될 수 있는 저전력 AI Chip을 개발하여 상용화

○ (중국) 범용 인공지능 위한 하이브리드 AI 회로 및 시스템 설계, 자율주행 자전거 구현 (The Science Monitor, 2019.8.1.)

■ 중국 과학자들이 **자체 개발한 인공지능(AI) 칩을 적용한 자율주행자전거를 공개**함. 자율주행 자전거에는 중국 칭화대(Tsinghua University)의 루핑시(Luping Shi) 교수와 연구진들이 개발한 **뉴로모픽 칩(neuromorphic chip) 텐직(Tianjic)이 적용**

■ 범용 인공지능(artificial general intelligence, AGI)을 목표로 개발한 텐직칩은 **기존의 컴퓨터 공학 기반 폰 노이만 디자인과 신경과학 기반 두뇌에서 영감을 얻은 뉴로모픽 아키텍처 두 접근 방식을 하나로 결합한 하이브리드 디자인을 특징**으로 함. 해당 연구 결과는 네이처지에 2019년 7월 31일에 발표

○ (중국) 대학 연구팀 AI 회로 및 시스템 연구 결과 상용화(Techcrunch.com 외 대중매체, 2020.6.25.)

■ Chinese Academy of Sciences(CAS) 연구실의 **AI Chip 연구내용을 바탕으로 창업된 Cambricon 사는 화웨이 사의 휴대폰에 들어가는 상용 AI Chip 라이센싱에 성공** 하는 등의 상업적 가치를 인정받아 2020년 7월 중국 상하이 증권 거래소 STAR Market에서 $369M 가치의 기업공개(IPO)에 성공

■ 중국 칭화대학교 연구팀의 AI Chip 연구내용을 바탕으로 창업된 Deephi Tech는 2018년 7월 FPGA 업계의 최대회사인 미국 Xilinx 사에 밝혀지지 않은 가격에 매각되며 성공적인 투자금 회수(exit)

○ (한국) 저전력 생성적 적대 신경망 (GAN) 전용 칩 연구(The Science Monitor 외 대중매체, 2020.4.6.)

■ KAIST 유회준 교수 연구팀이 **생성적 적대 신경망(GAN: Generative Adversarial Network)을 저전력, 효율적으로 처리하는 인공지능(AI: Artificial Intelligent) 반도체를 개발**
- 연구팀은 이번 반도체 칩 개발을 통해 이미지 합성, 스타일 변환, 손상 이미지 복원 등의 생성형 인공지능 기술을 모바일 기기에서 구현하는 데 성공
- 연구 결과는 2020년 2월 17일 미국 샌프란시스코에서 개최된 국제고체회로설계 학회(ISSCC)에서 발표

○ (한국) 토종 신경망처리장치(NPU) 스타트업 "샘플 성능 벤치마크 기대이상" (ZDNet, 2019.11.7.)

■ 퓨리OOOO는 최근 한국 기업으로는 유일하게 **글로벌 AI칩 벤치마크 테스트 'MLPerf'에 참가해, 경쟁력 있는 성능 지표를 인정받음.**
- MLPerf는 구글, 바이두, 하버드, 스탠포드 등 유수의 기업 및 대학들이 주최하는 글로벌 AI칩 성능 테스트임.
- 2019년 MLPerf에는 전세계 26개 기업이 참가 신청했으나, 퓨리OO를 포함한 13개 기업만이 조건을 충족해 결과를 제출

5. 기초연구사업 지원 현황은?

※ 연관 키워드 : AI 반도체, 뉴로모픽, 가속기, 하드웨어 가속기, 하드웨어 가속기, 딥러닝, 뉴럴네트워크

○ 지원과제 현황

구분	2017	2018	2019	2020
지원과제수(건)	11	16	26	38
지원액(백만 원)	726	1,236	2,089	3,255

※ 연구사업통합지원시스템(e-R&D)에서 '뉴로모픽, 가속기, 하드웨어 가속기, 하드웨어가속기, 딥러닝, 뉴럴네트워크, 뉴럴 네트워크' 키워드 검색 후 전문가 검토를 통해 지원과제 추출

연구책임자	과제명	사업명	총 연구기간
정OO 연세대학교	Domain Wall Motion 시냅스 기반의 On-Chip 지도-자율 통합학습 뉴로모픽 SoC 개발	중견연구	20170301 ~20200229
김OO 한국과학기술원	근사 컴퓨팅과 DRAM 내부 프로세싱을 활용한 초저에너지 심층 신경망 가속 프로세서 개발	중견연구	20170301 ~20200229
송OO 연세대학교	인공 지능 및 빅 데이터 처리를 위한 메모리-가속기 구조 개발	신진연구	20170301 ~20200229
이OO 울산과학기술원	내장형 기계학습을 위한 스파이크 기반 인공신경망 연구	이공학개인기초	20170601 ~20200531
김OO 광운대학교	모바일 인공지능 플랫폼을 위한 에너지 효율적인 근사연산 기반의 뉴로모픽 로직 연구	이공학개인기초	20170601 ~20200531
이OO 광운대학교	에뮬레이션 기반 전력분석을 위한 전력 계산 하드웨어 자동생성	이공학개인기초	20170601 ~20200531
김OO 충북대학교	생체모방 산화물 용액 공정 기반의 고성능 트랜지스터와 저항 변화형 메모리를 이용한 웨어러블 플랫폼 적응형 뉴로모픽 시스템 집적화 연구	이공학개인기초	20170601 ~20200531

연구책임자	과제명	사업명	총 연구기간
이OO 광운대학교	양이온 시냅스 소자 기반 뉴로모픽 패턴인식 시스템의 소자/아키텍처 플랫폼 개발	신진연구	20170901 ~20200831
선우OO 아주대학교	모바일 시각지능 지원을 위한 기계 학습 전용 프로세서 설계 연구	중견연구	20170301 ~20200229
강OO 울산과학기술원	근사연산을 이용한 저전력 딥러닝 하드웨어 설계	이공학개인기초	20170601 ~20200531
박OO 한국과학기술원	확장 가능한 저전력 멀티 모드 인공지능 프로세서 개발	전략공모	20171101 ~20221031
박OO 세종대학교	차세대 응용을 위한 데이터 중심 가속기 기반 컴퓨팅 시스템 구조 설계	중견연구	20180301 ~20210228
정OO 한국뉴욕주립대학교	회선신경망 기반의 영상분석용 저전력 하드웨어 가속기 신구조 개발	이공학개인기초	20180601 ~20200531
공OO 경북대학교	고성능, 저전력 이종컴퓨팅을 위한 시스템 수준 설계 및 기법 연구	이공학개인기초	20180601 ~20210531
구OO 홍익대학교	기계 학습 어플리케이션을 위한 Near Data Processing (NDP) 시스템에 대한 연구	신진연구	20180901 ~20210831
노OO 수원대학교	딥 뉴럴 네트워크 설계를 위한 기반 기술 개발	이공학개인기초	20180601 ~20210531
이OO 울산과학기술원	저전력 멀티코어 범용 모바일 딥 러닝 프로세서 구조에 관한 연구	신진연구	20190301 ~20220228
김OO 포항공과대학교	고성능 연산 가속기를 위한 이기종 적층 메모리 설계 연구	신진연구	20190301 ~20220228
전OO 서울대학교	자가 학습이 가능한 초저전력 혼성신호 뉴로모픽 프로세서 설계	신진연구	20190301 ~20220228
홍OO 경북대학교	딥 러닝 하드웨어 가속기의 메모리 병목 문제 해결을 위한 소프트웨어 및 하드웨어 기법 연구	생애 첫 연구	20190301 ~20220228
공OO 성균관대학교	브레인 동작에 기초한 Memristor-CMOS 하이브리드 컴퓨팅 아키텍처	중견연구	20190301 ~20220228
조OO 성균관대학교	고효율 적응형 딥러닝 학습 플랫폼	기본연구	20190601 ~20210228
김OO 경북대학교	엣지 컴퓨팅을 위한 에너지 고효율 어프록시메이트 뉴로모픽 코어 연구	학문균형발전지원	20190601 ~20230531
최OO 한국과학기술원	적층 및 집적도 높은 인메모리 컴퓨팅 시스템 구현을 위한 다공성 물질을 이용한 신개념 저항 변화 소자 제작	기본연구	20190601 ~20200531
이OO 서울과학기술대학교	고성능컴퓨팅(HPC)을 위한 멀티코어 하드웨어 가속기 핵심 기술 개발	기본연구	20190601 ~20220228
하OO 서울대학교	지능형 임베디드 시스템을 위한 소프트웨어 설계 및 코드 생성 기술	중견연구	20190601 ~20220228
김OO 포항공과대학교	뉴로허브: 인메모리 뉴럴 네트워크 하드웨어 공동연구를 위한 웹기반 개방형 시뮬레이션 플랫폼	중견연구	20200301 ~20230228

연구책임자	과제명	사업명	총 연구기간
이OO 광운대학교	실시간 신경신호 분석을 위한 시냅스소자 기반 뉴로모픽 칩	신진연구	20200301 ~20230228
김OO 한성대학교	자율주행차를 위한 소프트웨어 기반 인공지능 컴퓨팅 가속 기술	생애 첫 연구	20200301 ~20230228
정OO 고려대학교	온도를 고려한 3D 적층 메모리 기반 인메모리 가속기	중견연구	20200301 ~20250228
김OO 한국과학기술원	차세대 메모리 내부 프로세싱을 활용한 초저에너지 심층 신경망 가속 프로세서 개발	중견연구	20200301 ~20230228
정OO 광운대학교	기계학습 알고리즘의 저전력 고속 구현을 위한 컴퓨테이션 인 메모리 컴파일러 개발	생애 첫 연구	20200301 ~20230228
박OO 고려대학교	온-디바이스 개인화를 위한 에너지 효율적인 딥러닝 가속기 설계	중견연구	20200301 ~20240229
이OO 서울대학교	NAND 플래시 기반 심층신경망 학습 시스템	중견연구	20200301 ~20230228
김OO 홍익대학교	새로운 확률적 연산기법(Stochastic Computing)을 활용한 저전력 고효율 Spiking Neural Network 회로 설계	기본연구	20200601 ~20230228
박OO 인천대학교	에지 컴퓨팅 장치에서의 에너지 효율적 데이터 처리를 위한 하드웨어 가속기 활용 근사컴퓨팅	기본연구	20200601 ~20230228
김OO 서울대학교	차세대 인공지능 알고리즘 구현을 위한 뉴로모픽 신소재 및 신소자 기반 저전력 고집적 하드웨어 연구	기본연구	20200601 ~20220228
김OO 충북대학교	생체모방 산화물 용액 공정 기반의 고성능 트랜지스터와 저항 변화형 메모리를 이용한 웨어러블 플랫폼 적응형 뉴로모픽 시스템 집적화 연구	학문균형발전지원	20200601 ~20230531
이OO 고려대학교	차세대 메모리의 특성을 고려한 인메모리 가속기	기초연구기반구축	20200601 ~20220531
이OO 울산과학기술원	RRAM 기반 뉴럴넷 시스템 설계를 위한 고속 시뮬레이션 기술 연구	중견연구	20200301 ~20230228
이OO 서울과학기술대학교	딥러닝 기반 핵심 사용자 경험 요소 도출 및 설계 요소 우선 순위화	기본연구	20200601 ~20230228
박OO 숙명여자대학교	차세대 영구 메모리 기반 고성능 인메모리 빅데이터 처리 및 딥러닝 프레임워크 설계와 개발	기본연구	20200601 ~20230228
이OO 연세대학교	지능형 메모리 기반 저비용 분산 딥러닝 시스템 설계	기본연구	20200601 ~20220228

○ 성과현황

(단위 : 건)

구분	SCI 논문 수	JCR 상위 25%	기술실시 계약	특허등록	특허출원
2017	7	2	1	0	2
2018	17	4	1	0	9
2019	28	11	4	2	20
2020	6	4	0	0	6

※ 2017~2020년 지원과제가 발표한 성과

○ 기초연구사업 주요 연구 결과 ※ 종료과제 중심으로 연구결과 조사

★ Domain Wall Motion 시냅스 기반의 On-Chip 지도-자율 통합학습 뉴로모픽 SoC 개발

<연구책임자>
■ 정OO(연세대학교)
<성과내용>
■ 멀티 MAC 동작을 수행하는 뉴로모픽 시스템 및 그 방법
- 기존 대비 MAC 연산 에너지를 줄인 스파이킹 뉴럴 네트워크 SoC 구현
- All-to-all STDP 학습방법을 간소화한 on-chip 학습 방법 개발

★ 근사 컴퓨팅과 DRAM 내부 프로세싱을 활용한 초저에너지 심층 신경망 가속 프로세서 개발

<연구책임자>
■ 김OO(한국과학기술원)
<성과내용>
■ 초저에너지 근사 컴퓨팅과 DRAM 내부 프로세싱 기법
- 초저에너지 CNN을 위한 이종 데이터 표현방식 및 근사 컴퓨팅 연산기 개발
- DRAM 내부 연산 구조 및 내부 연산 가능한 embedded DRAM 프로토타입 개발

★ 인공지능 및 빅 데이터 처리를 위한 메모리-가속기 구조 개발

<연구책임자>
■ 송OO(연세대학교)
<성과내용>
■ 인공지능 및 빅데이터 연산과 메모리 접근 패턴 기반 효율적 데이터 관리 기법
- 워크로드의 특징 분석 및 메모리-가속기 구조의 모델링 및 시뮬레이터 개발
- 시뮬레이터를 기반으로 인공지능 및 빅데이터 워크로드의 성능 개선 방안 제시

★ 내장형 기계학습을 위한 스파이크 기반 인공신경망 연구

<연구책임자>
■ 이OO(울산과학기술원)
<성과내용>
■ 고집적 인공지능 하드웨어를 위한 스파이크 펄스 기반의 시냅스 회로
- 1 비트 스파이크 펄스로 동작하는 인공지능 모델과 하드웨어 회로를 개발
- 기존 디지털 회로 대비 수백 배 이상 집적도의 인공지능 하드웨어를 제작 가능

★ 모바일 인공지능 플랫폼을 위한 에너지 효율적인 근사연산 기반의 뉴로모픽 로직 연구

<연구책임자>
■ 김OO(광운대학교)
<성과내용>
■ 근사연산기를 활용한 에너지 효율적인 이미지 프로세스 개발
– 8개의 트랜지스터를 이용한 새로운 XNOR 기반의 근사연산 덧셈기를 개발
– 개발한 연산기를 JPEG 엔코더에 적용하여 면적 15%, 에너지 24%를 절감

★ 빠른 전력 분석 속도를 가지는 자동생성이 가능한 전력 모델 개발

<연구책임자>
■ 이OO(광운대학교)
<성과내용>
■ 에뮬레이션 기반의 전력 분석의 느린 속도에 대한 문제를 해결
– 기존 연구에서는 속도 증가를 위해 정확도나 칩의 면적을 희생하는 문제가 있음.
– 정확도와 면적의 희생을 최소화한 자동생성 가능한 전력 모델을 제시

★ 뉴로모픽 집적화 디바이스를 위한 생체모방 산화물 용액 공정 개발

<연구책임자>
■ 김OO(충북대학교)
<성과내용>
■ 뉴로모픽 고성능 전자소자의 특성을 분석하여 생체모방적 패터닝 기술을 개발
– 생체모방적 나노 원자층 클러스트 웨어러블 보호막 박막을 이용한 용액형 고성능 산화물 트랜지스터와 저항 변화형 메모리의 전기적/환경적 안정성을 규명
– 유연하고 투명한 감성 UI/UX 기판에서 생체모방적 패터닝/보호막 기반의 용액형 고성능 뉴로모픽 전자소자의 웨어러블 집적 시스템을 개발

★ 화면 내 예측연산의 복잡도 감소를 위한 AFMD 알고리즘 고안

<연구책임자>
■ 선우OO(아주대학교)
<성과내용>
■ 블록 분할 연산의 복잡도를 감소시키기 위한 새로운 알고리즘을 제안
– Adaptive fast mode decision 알고리즘을 제안하여 계층적 모드를 예측하고 skip

★ 모바일 및 IoT을 위한 초저전력 딥러닝 하드웨어 및 프레임워크 개발

<연구책임자>
■ 강OO(포항공과대학교)
<성과내용>
■ 초저전력으로 동작하는 딥러닝 기반의 인공지능 프로세서를 개발
– 근산 연산 기술, 이상치 시분할 다중화 단위 모듈 연산법, 채널 루프 타일링 최적화 등 다양한 기법을 사용하여 초저전력 딥러닝 프레임워크를 개발
– 인공지능 기술의 에너지 문제를 해결하고, IoT 기기 연구 개발이 가속화 될 수 있다는 점에서 실용화 가능성이 매우 높고 전자산업의 핵심으로써 발전 가능

6. 향후 기초연구사업에서 어떤 연구들이 필요한가?

연구자 인터뷰 결과 ※ 이영주 교수 (포항공대)	■ 심층신경망(Deep Neural Network, DNN) 기반의 다양한 응용 시스템이 학계를 중심으로 개발되고 있으나, 대부분 **알고리즘이 edge-level 에서 동작하기 어려운 상황으로 파급력에 한계가 존재**함. - 기존 심층신경망은 학습 및 추론 과정에서 서버(혹은 클라우드) 컴퓨팅 플랫폼에서 제공하는 강력한 연산능력을 활용하기에, 에너지 및 연산 성능에서 제한이 존재하는 엣지 수준 디바이스 (edge-level device)의 특성을 반영하지 못하고 있음. - **온 디바이스 AI 처리(On-device AI processing)는 데이터를 서버/클라우드로 전송하지 않기에 보안 및 처리시간에서 큰 장점**을 보이고 있어, 음성 및 카메라 신호처리와 같은 일부 알고리즘을 중심으로 사용화 사례가 2019년 말부터 보고되고 있으나, **여전히 낮은 에너지 효율로 제한적인 용도로만 사용된다는 한계점**이 보고되고 있음. - 따라서, 산업계에 파급력이 높은 응용 시스템을 중심으로 저전력 심층신경망(DNN) 알고리즘 개발, 기존 상용 엣지 수준 시스템과의 호환을 고려한 시스템 아키텍처 구조, 높은 효율을 지원하는 가속기 시스템 설계 등이 융합되는 **"애플리케이션에 특화된 온 디바이스 (application-specific on-device) AI반도체" 연구의 지원이 단/중기적인 관점에서 시급하게 요구**됨. - 또한, 연합학습 (federated learning), 센서융합 (sensor fusion) 및 지능형 바이오/웨어러블 시스템 등과 같이 인접 학문분야와의 융복합 연구로 **엣지 수준에서 보안성이 높은 초저전력 심층신경망 (DNN) 연산을 제공, 차세대 먹거리를 제공할 수 있는 "선도적 AI 반도체 기초연구"에 대한 중/장기적 지원이 요구**되고 있음.
연구자 인터뷰 결과 ※ 궁재하 교수 (대구경북과기원)	■ 최근 AutoML (automated machine learning) 즉, 기계 학습의 자동화에 관한 연구가 활발히 진행되고 있으며, 그 중 특히 **네트워크 구조를 자동으로 탐색해주는 신경망 구조 탐색(Neural Architecture Search, NAS), DARTS (Differentiable ARchiTecture Search) 등의 연구가 주목받고 있음.** - 사람이 만든 네트워크 구조보다 NAS **및** DARTS **등을 활용하여 자동화된 네트워크 탐색을 통해 찾은 네트워크 구조가 좋은 성능을 보임**. 특히 최근 양자화(quantization), pruning 등이 적용된 모바일용 네트워크 생성에서도 좋은 성능을 보이는 등 앞으로도 수많은 응용 분야에 활용될 것으로 예상됨. - 기존 연구들은 주로 NAS 및 DARTS를 이용해 좋은 성능의 네트워크 구조를 탐색하는 방향으로 진행되어 왔지만, NAS 및 DARTS 탐색 과정에 막대한 연산량과 시간이 소모됨. 예를 들어 최근 개발된 FBNet 또는 One-shot NAS의 경우에도 그래픽카드 한 개로 학습 시 12-20일이 소요됨. 이를 줄이는 것이 절실하지만 관련 하드웨어/시스템 연구가 미비한 상황

	– 이러한 측면에서 **기초연구사업에서는** NAS, DARTS **등 자동 탐색의 하드웨어 연산량과 시간 소모를 줄이기 위한 소프트웨어 및 하드웨어 연구를 활발히 수행할 수 있도록 하는 것이 바람직**함. 소프트웨어에서는 효율적 연산분산기법 및 NAS 관련 API 개발, 하드웨어에서는 메모리의 효율적 제어 및 서치 후보군을 모두 지원 가능한 하드웨어 연구가 필요함.

7. 향후 중점적으로 추진하여야 할 연구주제는?

연구자 인터뷰 결과 ※ 박종선 교수 (고려대), 이영주 교수 (포항공대)	① **파급력이 높은 애플리케이션(application)에 최적화된 온 디바이스 AI 처리를 지원하는 초저전력 edge-level AI 가속기 시스템** ■ 모바일 및 edge-level IoT 디바이스에 탑재되어 다양한 서비스를 제공하는 온 디바이스 AI 처리의 실현을 위해서는 알고리즘 수준부터 하드웨어 구조 및 고효율 회로설계 전반에 걸친 융합적 연구가 필수적임. ■ 자동 음성인식(Automatic speech recognition, ASR), 자연어 처리(natural language processing, NLP) 및 다채널 카메라용 이미지 신호처리(image signal processing, ISP) 등과 같이 **edge-level DNN 솔루션에 대한 현재 산업계의 수요를 반영하는 애플리케이션에 특화된(application-specific) AI 시스템 개발이 필요**함. ■ 이와 동시에, 연합학습, 센서융합(sensor fusion) 및 biomedical wearable 시스템과 같은 온 디바이스 AI 처리를 다양한 인접 분야로 확장시켜 차세대 기술발전의 동력을 선점할 수 있는 융합적 연구에 대한 적극적인 지원이 필요함. ② **사생활 정보 보호를 위한 하드웨어 친화적인 딥러닝 보안 알고리즘 및 가속기 개발** ■ 미래에 AI 반도체는 의료·금융 정보와 같은 민감한 개인 정보에 대한 공격을 방어하기 위한 딥러닝 보안 기술 적용이 필수적임. ■ 딥러닝 보안에 사용되는 대표적인 알고리즘인 동형 암호화(homomorphic encryption)는 암호화된 데이터로 직접 연산하므로, 모듈로(modulo) 연산 비용을 줄이기 위한 알고리즘 연구가 필요 하고, 개발된 알고리즘에 최적화된 딥러닝 가속기 개발이 필요함. ③ **신경구조망탐색(NAS, Neural Architecture Search) 및 DARTS (Differentiable ARchiTecture Search) 탐색 비용 감소 알고리즘 및 하드웨어 가속기 개발** ■ Network search space의 redundancy 파악 및 제거를 통해 NAS 및 DARTS의 탐색 시간 및 연산량을 감소시키는 하드웨어 친화적인 알고리즘 연구가 요구됨. ■ 또한, NAS **및** DARTS **연산에 최적화된 하드웨어 가속기 개발을 통해 이러한 연산 시간과 연산량을 획기적으로 줄이는 연구가 필요**함.