TOP 10 RESEARCH ACHIEVEMENTS

GANPU: 생성적 적대 신경망을 위한 온 디바이스 학습 프로세서

GANPU: An On-Device Training Processor for Generative Adversarial Networks

소속학과 전기및전자공학부 연구책임자 유회준 홈페이지 http://ssl.kaist.ac.kr

생성적 적대 신경망 (Generative Adversarial Network, GAN)의 추론 및 학습을 모바일 기기 상에서 에너지 효율적으로 처리하는 인공지능 반도체에 대한 연구이다. 생성적 적대 신경망은 기존 인공지능과 달리 존재하지 않는 사람의 사진을 생성해내는 등, 새로운 데이터를 생성해낼 수 있다는 점에서 주목을 받는다. 그러나 많은 연산이 필요하고 신경망의 구조가 복잡하여 최적화된 가속기 설계가 어렵다는 문제가 있었다. 본 연구에서는 복잡한 신경망 구조에 맞춰 프로세서 내부 구조를 재정렬 하는 가변형 프로세서 아키텍처와 희소 연산을 효율적으로 처리하는 방식을 제안하였다. 제안된 기술을 통해 기존 연구대비 4.8 배 증가한 에너지 효율을 달성하였으며, 속도와 전력소모가 제한된 모바일 기기 상에서 온 디바이스 AI의 구현을 가능케 하였다.

1. 연구배경

기존의 인공지능 반도체 영역에서 주로 연구된 인공지능 기술인 분류형 모델(Discriminative Model)은 주어진 질문에 대한 답을 하도록 학습된 인공지능 모델로, 물체 인식 및 추적, 음성 인식, 안면 인식 등에 활용된다. 이와 달리 생성적 적대 신경망(Generative Adversarial Network, GAN)은 새로운 이미지를 생성, 재생성 할 수 있어 이미지의 스타일 변환, 영상 합성, 손상된 이미지의 복원 등 광범위한 분야에 활용된다. 동영상에 다른 사람의 얼굴을 입힐 수 있어 큰 이슈가 되었던 딥페이크 (Deepfake) 기술도 GAN을 기반으로 한 기술이다. GAN은 많은 영상, 사진 콘텐츠가 생산되고 소비되는 모바일 기기의 다양한 응용 프로그램에 활용될 가치가 높아 학계뿐만 아니라 산업계에서도 큰 주목을 받고 있다.

하지만 GAN은 기존의 네트워크와는 달리 복수 개의 심층 신경망으로 이루어진 복잡한 구조로, 개별 심층 신경망마다 서로 다른 특성을 보여 가속기의 최적화가 어렵다. 계속되는 디스플레이, 이미지센서의 고해상도화로 요구되는 고해상도 이미지 생성을 위해 기존 AI 모델보다 수십 배 많은 연산량을 요구한다. 따라서 속도와 전력이 제한된 모바일 장치 (스마트폰, 태블릿 등)에서 GAN을 구현하는데 제약이 있으며, 온 디바이스 학습을 구현하는 데는 더 큰 한계가 있다.

2. 연구내용

기존의 연구들은 추론 단계만을 지원하거나 단일 신경망에 한정되어있었지만 GAN과 같은 다중 심층 신경망을 처리할 수 있도록 본 연구에서는 세계최초로 모바일 플랫폼에서 추론 및 학습이 가능한 저전력

인공지능 반도체 GANPU (Generative Adversarial Networks Processing Unit)를 개발하여, 모바일 장치가 스스로 자연스러운 그림을 그릴 수 있는 새로운 인공지능 시대를 열었다.
모바일 기기에서 저전력으로 다중 심층 신경망을 가속하기 위해서는 적응형 워크로드 할당을 통해 다중 신경망에 맞춰 프로세서 아키텍처를 가변적으로 재정렬하는 기술이 제안되었다. 이를 통해 외부 메모리 대역폭, 연산 자원등의 제한된 자원의 효율적 할당을 가능케 한다. 또한, 심층 신경망에는 비선형 활성화 함수로 인해, 데이터에 0이 다량 존재한다. 불필요한 0 연산을 뛰어넘는 프로세서 아키텍처를 설계하여 추론 및 학습 과정에서의 속도와 에너지 효율을 극대화 하였다.
위의 기술을 활용하여 제작된 인공지능 반도체 GANPU는 기존 기술 대비 4.8 배 높은 에너지 효율을 달성하였다. 추가적으로, 본 연구를 통해 설계한 GANPU를 집적하여 모바일 시스템을 구현하였으며, 이를 통해 GAN 응용 기술을 시연하였다. 태블릿 카메라로 찍은 사진을 사용자가 직접 수정할 수 있으며, 사진 상의 얼굴에서 머리, 안경, 눈썹 등 17가지 특징에 대해 추가, 삭제 및 수정사항을 입력하면 GANPU를 통한 연산으로 이 결과를 실시간으로 보여주는 시스템이다. 이것에 더해, 새로운 사용자 데이터가 입력되었을 때 온 디바이스 학습으로 개선된 신경망을 통해 정확한 결과를 도출해낼 수 있다.
응용 기술 시연 영상 - https://www.youtube.com/watch?v=HnNWsgqkEU0

3. 기대효과

세계최초로 하나의 칩에서 추론만이 아니라 학습까지 가능하며, 여러 개의 심층 신경망을 동시에 가속하는 저전력 인공지능 반도체를 개발하여 모바일 기기에서 새로운 인공지능 응용을 가능하게 했다는 점에서 의미가 크다.
또한, 본 연구에서 개발한 인공지능 반도체는 서버로 데이터를 보내지 않고 모바일 장치 내에서 생성적 적대 신경망(GAN)을 스스로 학습할 수 있어 사생활 보호를 가능하게 하며 화면의 스타일 변환, 영상 합성, 손상된 이미지의 복원하는 프로세서라는 점에서 그 활용도가 주목받는다. 이를 통해 모바일 기기가 스스로 자연스럽게 그림을 그리는 새로운 인공지능 시대를 여는 발판이 될 수 있다.

그림 1. setup

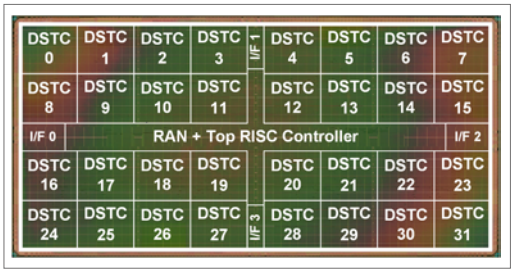

그림 2. 칩사진

연 구 성 과

[논문] S. Kang, D. Han, J. Lee, D. Im, S. Kim, S. Kim, H-J. Yoo, "GANPU: A 135TFLOPS/W Multi-DNN Training Processor for GANs with Speculative Dual-Sparsity Exploitation", 2020 IEEE International Solid-State Circuits Conference (ISSCC 2020).
S. Kang, D. Han, J. Lee, D. Im, S. Kim, S. Kim, J. Ryu, H-J. Yoo, "GANPU: A Versatile Many-Core Processor for Training GAN on Mobile Devices with Speculative Dual-Sparsity Exploitation", 2020 IEEE Hot Chips Symposium.
[특허] 국내 특허 출원 완료, 미국 특허 출원 완료
[수상] 제 26회 삼성전자 휴먼테크 논문대상 금상 수상
2020 Microsoft Research Asia Fellowship 수상
[홍보] YTN 뉴스 보도, "스스로 그림 그리는 AI 반도체 칩 개발" / 문화일보, 동아사이언스 등 국내 30 여개 언론 보도