



NXP, 업계 최초의 엣지 머신 러닝용 MCU 기반 글로우 신경망 컴파일러 발표

- 비전(vision) 및 음성 기반 머신 러닝 애플리케이션 용도로 글로우 NN 컴파일러의 광범위한 장점 증명
- NXP는 반도체 벤더 중 최초로 글로우 기본 버전에 비해 MCU 성능을 2-3배 향상
- 원래 페이스북이 개발한, 오픈 소스 글로우 컴파일러가 이제 NXP의 eIQ™ 머신 러닝 소프트웨어 개발 환경에서 이용할 수 있게 되어 NXP의 i.MX RT 시리즈 크로스오버 MCU에 고성능 추론 성능 제공
- NXP의 글로우 실행은 i.MX RT 시리즈 크로스오버 MCU를 위한 플랫폼별 최적화로 Arm® 코어텍스®-M 코어와 카덴스® 텐실리카® HiFi 4 DSP를 대상으로 실행

2020년 7월 29일 – NXP 반도체는 글로우 신경망(NN: neural network) 컴파일러를 위한 eIQ 머신 러닝(ML) 소프트웨어를 지원한다고 발표했다. 이로써 업계 최초의 NN 컴파일러 실행을 통해 NXP의 i.MX RT 크로스오버 MCU상에서 적은 메모리 사용량(footprint)으로 우수한 성능을 구현하게 되었다. 페이스북(Facebook)이 개발한 글로우(Glow)는 타깃 별 최적화를 통합할 수 있다. NXP는 NN 오퍼레이터 라이브러리(operator library)를 사용해서 이 기능을 Arm 코어텍스(Cortex)-M 코어와 카덴스 텐실리카(Cadence Tensilica) HiFi 4 DSP에 적용하여 i.MX RT685, i.MX RT1050, RT1060의 추론 성능을 최적화했다. 이 기능은 NXP MCUXpresso SDK에서 무료로 제공되는 NXP eIQ 머신 러닝 소프트웨어 개발 환경(Machine Learning Software Development Environment)에 통합됐다.

글로우(Glow) 이용해서 MCU 아키텍처 특징 활용

2018년 5월, 파이토치(PyTorch)의 주요 선도기업인 페이스북은 다양한 하드웨어 플랫폼에서 신경망 성능을 가속화하기 위해 최적화를 제공하고자 오픈 소스 커뮤니티 프로젝트로 글로우(Glow: Graph Lowering NN compiler)를 도입했다. NN 컴파일러로 글로우는 최적화되지 않은 신경망을 가져와 최적화가 잘된 코드를 생성한다. 그렇기 때문에 더 높은 성능을 요구하고, 메모리 오버헤드를 추가하는 JIT(just-in-time:적시) 컴파일레이션이 활용되는 일반 신경망 모델 프로세싱과는 다르다. 글로우에서 가능한 것처럼 직접 최적화된 코드를 돌리면 프로세싱 및 메모리 요건이 현격히 줄어든다. NXP는 글로우 오픈 소스 커뮤니티에서 적극적인 역할을 맡아 새로운 글로우 기능이 포괄적으로 채택되도록 지원하고 있다.

드와락 라자고팔(Dwarak Rajagopal), 페이스북의 소프트웨어 엔지니어링 매니저는 “깃허브(GitHub)에서 나온 바로 사용 가능한 글로우 기본 버전은 RISC-V뿐만 아니라 Arm 코어텍스-A 와 코어텍스-M 코어 등을 비롯해 사용자들이 관심이 있는 기본 아키텍처에 신경망 모델을 컴파일(compile) 할 수 있도록 유연성을 제공해, 특정 기기에 종속적이지 않다”면서, “MCU의 컴퓨트 엘리먼트(compute element)를 활용하는 특수 목적 소프트웨어 라이브러리를 이용하고 2-3배 성능을 향상함으로써, NXP는 고급사



양 클라우드 기반 기계에서 저가 임베디드 플랫폼까지 다양한 머신 러닝 애플리케이션에 글로우 NN 컴파일러를 사용함으로써 얻을 수 있는 여러 가지 장점을 입증했다"고 말했다.

최적화된 머신 러닝 프레임워크로 경쟁 우위 확보

향후 수 년간 ML 애플리케이션 수요가 상당히 증가할 것으로 예측된다. TIRIAS 리서치는 2025년까지 전체 엣지 디바이스의 98%가 어떤 형태이든 머신 러닝/인공 지능을 이용하게 될 것으로 전망한다. 시장 예측에 따르면, 2025년까지 전용 ML 가속기 없이도 180억-250억 개의 디바이스에 ML 기능이 포함될 것이라고 한다. 소비자 기기 제조업체들과 임베디드 IoT 개발자들은 MCU를 이용하는 저전력 엣지 임베디드 애플리케이션용으로 최적화된 ML 프레임워크를 필요로 하게 될 것이다.

론 마티노(Ron Martino), NXP 반도체의 수석부사장은 "NXP는 eIQ ML 소프트웨어 프레임워크로 통합도가 높은 i.MX 애플리케이션 프로세서와 고성능 i.MX RT 크로스오버 MCU의 강력한 기능을 활용해서 엣지 디바이스 상에서 머신 러닝의 기능 구현을 추진하고 있다"면서, "i.MX RT 시리즈 크로스오버 MCU에 글로우 지원이 추가되면서 본사 고객들은 심층 신경망을 컴파일하고 자사 애플리케이션에 경쟁 우위를 부여할 수 있게 되었다"고 밝혔다.

NXP의 ML을 위한 엣지 인텔리전스 환경 솔루션은 개발자들이 엣지 디바이스에서 효율적으로 ML을 실행하는데 필요한 기본 요소를 제공하는 포괄적인 툴킷이다. 글로우가 eIQ 소프트웨어에 통합되어, 이제 ML 개발자들은 i.MX RT 크로스오버 MCU와 i.MX 8 애플리케이션 프로세서와 같은 NXP의 엣지 프로세싱 솔루션 전반에서 확장이 가능한 포괄적인 고성능 프레임워크를 갖게 될 것이다. 고객들은 i.MX RT MCU와 i.MX 애플리케이션 프로세서 상에서 ML 음성 애플리케이션, 객체 인식, 얼굴 인식 등을 개발하기에 더 유리해질 것이다.

NXP의 글로우 신경망 실행으로 성능 가속화

이제 eIQ에 글로우와 텐서플로우 라이트(TensorFlow Lite)에 대한 추론 지원이 포함되었다. NXP는 두 가지 모두에 대해 정례적으로 벤치마킹 활동을 실행해서 성능을 측정하고 있다. MCU 벤치마크에는 CIFAR-10과 같은 표준 NN 모델이 포함되어 있다. 예로 CIFAR-10 모델을 이용한 NXP가 획득한 벤치마크 데이터는 (600MHz Arm 코어텍스-M7로) i.MX RT1060 디바이스, (1GHz Arm 코어텍스-M7로) i.MX RT1170 디바이스 그리고 (600 MHz 카덴스 텐실리카 HiFi 4 DSP로) i.MX RT685 디바이스의 성능 장점을 활용하는 방법을 보여준다.

NXP의 글로우 구현은 카덴스가 4.8GMACs 성능을 구현하는 텐실리카 HiFi 4 DSP에 제공하는 신경망 라이브러리(NNLib)와 밀착하게 결합되어 있다. 동일한 CIFAR-10 예에서, NXP의 글로우 실행은 NN연산을 가속화하는데 DSP를 이용함으로써 성능을 25배 향상한다.

산지브 아가왈라(Sanjive Agarwala), 카덴스의 텐실리카 IP 부문 기업 부사장은 "텐실리카 HiFi 4 DSP는 다양한 오디오 및 음성 처리 애플리케이션을 가속화하기 위해 원래 i.MX RT600 크로스오버 MCU에 통



합되었다. 하지만, 저가, 저전력 MCU 급 애플리케이션을 타겟으로 하는 ML 추론 애플리케이션 수가 늘어나면서, HiFi 4 DSP에 내재된 DSP 연산 성능이 NN 모델을 가속화하는데 이상적인 타겟이 되었다"면서, "NXP가 eIQ ML 소프트웨어에서 글로우를 실행함으로써, i.MX RT600 MCU고객들은 키워드 검출 (KWS), 음성 인식, 잡음 제거, 이상 탐지 등을 포함한 많은 ML 애플리케이션을 다루는데 DSP를 활용할 수 있게 되었다"고 말했다.

데니스 로딕(Dennis Laudick), Arm 머신 러닝 부문 마케팅 부사장은 "NXP가 Arm CMSIS-NN 소프트웨어 라이브러리를 eIQ에 포함한 것은 Arm 코어텍스-M 코어의 신경망 성능을 최대화하고 메모리 사용량을 최소화하기 위함"이라면서, "예로 CIFAR-10 신경망 모델을 이용해서 NXP는 CMSIS-NN으로 성능을 1.8배 개선할 수 있었다. 다른 NN 모델도 비슷한 결과를 나타낼 것이며, 이는 첨단 컴파일러와 NXP의 최적화된 NN 오퍼레이터 라이브러리를 장점을 확실히 증명하는 것"이라고 설명했다.

시장 출시

글로우 NN 컴파일러용 NXP의 eIQ는 현재 구매 가능하며, i.MX RT600 크로스오버 MCU와 i.MX RT1050 및 i.MX RT1060 크로스오버 MCU를 위한 MCUXpresso SDK를 통해 제공된다. 앞으로 그 밖의 NXP MCU에 대해서도 글로우 NN 컴파일러를 위한 eIQ가 출시될 것이다.

i.MX RT 시리즈 크로스오버 MCU 관련 정보

i.MX RT 시리즈는 업계 최초의 크로스오버 MCU 포트폴리오로 저렴한 가격에 고성능 Arm 코어텍스-M코어, 실시간 기능과 MCU 사용성을 제공한다. i.MX RT 시리즈는 저전력 애플리케이션 프로세서와 고성능 마이크로컨트롤러의 융합을 의미한다. i.MX RT 시리즈는 전통적인 MCU와 i.MX 애플리케이션 프로세서 공간의 간극을 이어주어 MCU 고객들이 사용 용이성을 희생하지 않고도 성능과 통합을 상당히 개선할 수 있는 길을 마련해준다.

<http://www.nxp.com/eiq> 및 <http://www.nxp.com/eiq/glow> 방문하면 보다 상세한 정보를 확인할 수 있다.

NXP 반도체 소개

NXP® 반도체(나스닥: NXPI)는 더욱 편리하고 안전하며 더 나은 삶을 위한 첨단 솔루션을 개발하여, 안전하게 연결되는 스마트 월드를 만들고 있다. NXP는 임베디드 애플리케이션용 보안 연결 솔루션의 선도 기업으로서, 자동차, 산업 및 IoT, 모바일, 통신 인프라 시장의 혁신을 주도하고 있다. NXP는 60년 이상의 전문성과 경험을 바탕으로, 전 세계 30개 이상의 국가에서 30,000명의 직원을 고용하고 있다. 2019년 매출은 미화88억 8천만불이다. NXP 관련 뉴스는 www.nxp.com에서 찾아볼 수 있으며, NXP 반도체 블로그 (<http://blog.naver.com/nxpkor>) 에서도 NXP 관련 정보를 확인할 수 있다.