MCU-NPU 기반 Vision Transformer 모델 구현 및 성능 평가

김태훈, 김태구, 조용훈, 신기훈, 곽도균, 구동한, 나상진, 조은영, 황우찬, 백윤주*

부산대학교

{bigteach0508, tbg8577, kchoyh95, kihunshin, kwakdg, ehdgksrn, sktkdwls1222, ciara0124, woochan629}@pusan.ac.kr,

*yunju@pusan.ac.kr

Implementation and Performance Evaluation of Vision Transformer Model Based on MCU-NPU

Kim Tae Hun, Kim Tae Gu, Cho Yong Hun, Shin Ki Hun, Kwak Do Gyun, Koo Dong Han, Na Sang Jin, Cho Eun Young, Hwang Woo Chan, Baek Yun Ju*

Pusan National Univ.

요 약

AI 모델을 MCU에서 직접 추론하는 온디바이스 AI는 외부 자원 의존도 감소, 개인정보 보호 측면에서 장점을 가지지만, 제한된 연산 성능으로 AI 모델을 구동하는데 한계가 존재한다. 이러한 한계를 극복하고자 MCU에 NPU를 통합하는 사례가 증가하고 있으나, 일반적인 NPU의 경우 CNN에 최적화 되어있어 Transformer 모델을 활용한 사례는 드물다. 본 논문에서는 Vision Transformer 모델을 NPU에 실행하기 위해 모델 구조를 NPU에 최적화하였으며 경량화를 수행하였다. NPU와 CPU의 연산 성능 및 메모리의 차이를 고려하여 다양한 크기의 Vision Transformer 모델을 제작하고 탑재하여 비교 실험을 수행하였으며, NPU에서 실행 시 CPU 대비 추론 시간이 약 91% 감소하여 NPU를 통한 Transformer 모델의 가속화 유효성을 입증하였다.

Ⅰ. 서론

최근 IoT 기기와 엣지 컴퓨팅 장치의 증가로 엣지 디바이스의 MPU(Micro Processor Unit)나 MCU(Micro Controller Unit) 내부에서 직접 추론을 실행하는 온디바이스 AI(On-device AI)의 필요성이 증가하고 있다[1]. 온디바이스 AI는 지연시간 단축, 클라우드 컴퓨팅 리소스 및 네트워크 의존성 완화, 시스템 안정성 향상 및 개인정보 보호 측면에서 이점을 가진다.

MCU는 CPU(Central Processing Unit), 메모리, I/O 포트가 단일 칩에 통합된 저전력 프로세서로, 운영체제 없이 구동되어 자원 제약적인 시스템에 최적화되어 있다. 그러나 제한된 연산 성능과 메모리 용량으로 인해, 높은 계산량과 복잡한 연산 구조를 요구하는 AI 모델을 실행하는 데 근본적인 한계가 있다. 특히 Transformer 같은 대규모 모델을 MCU의 연산만으로 실시간 동작시키기에는 많은 제약이 있다.

이러한 하드웨어 제약을 극복하기 위해 전용 신경망 처리 장치(Neural Processing Unit, NPU)를 내장한 MCU가 주목받고 있다. MCU에 통합된 NPU는 CNN(Convolutional Neural Network)과 같은 복잡한 신경망 모델을 고속으로 처리할 수 있도록 지원한다. 이를 통해 기존 MCU 기반 시스템보다 더 높은 수준의 AI 성능을 기대할 수 있게 되었다.

하지만 현재까지 출시된 MCU에 내장된 NPU는 CNN 관련 연산을 가속하는데 초점을 맞추고 있으며, Transformer 기반 모델을 구동하고 검증한 사례는 매우 한정적이다. Transformer는 비전, 오디오 및 자연어 처리 등 다양한 분야에서 활용되고 있지만, 많은 연산량을 요구하는 self-attention 구조로 인해 MCU 환경에 적용하기 어렵다. 본 연구에서는 NPU를 내장한 MCU에서 Transformer 기반 모델을 실제로 구동하고, 그 성능과 한계를 분석하고자 한다.

Ⅱ. 본론

1. NPU 탑재를 위한 ViT 모델 설계

본 논문에서 사용한 NPU를 내장한 MCU는 ST Microelectronics(STM) 사에서 출시한 STM32N6 MCU이다. 해당 MCU에 내장된 NPU는 양자화된 CNN 모델을 가속하도록 설계되어 있어 신경망 모델 관련 모든 연산을 가속할 수 있는 것은 아니며, 가속할 수 없는 것은 CPU에서 실행된다. 또한 해당 NPU에서 신경망 모델을 실행하기 위해 STM 사에서 제공하는 STEdgeAI 도구를 통해 모델을 컴파일하는 것이 필요하다. 해당 도구는 NPU 설계에 따라 컴파일을 지원하는 연산이 제한 되어있어, 지원되는 연산으로 Transformer를 구성하는 것이 필요하다.

본 논문에서 Transformer 모델 탑재 및 구동 가능성을 검증하기 위해 ViT(Vision Transformer)[2] 모델을 기반으로 하되, NPU에 탑재하기 위해 구조를 개선하였다. PE(Patch Embedding)를 모델 입력 전 전처리 과정에서 수행하도록 구성하였으며, 이는 모델에 PE와 self-attention이 같이 있을 경우 STEdgeAI를 활용한 모델 구조 해석이 불가하기 때문이다.

또한, 2차원의 입력을 받는 Fully Connected(FC) 레이어(layer)만 지원하는 NPU에 ViT에서 사용되는 $(1, patch, \dim)$ 형태의 3차원 입력을 받는 FC 레이어를 적용하기 위해 2가지 방법을 설계하였다. 첫 번째로, FC에서 bias 파라미터를 제거하였다. 이는 컴파일 시 STEdgeAI가 3차원의 입력에서 기존 batch 차원을 제거하고 patch 부분이 배치 차원으로 설정되어 $(patch, \dim)$ 형태가 되는데, NPU에서 batch 차원으로 bias를 브로드캐스팅(broadcasting)하는 것이 지원되지 않기 때문이다. 두 번째로 FC 레이어를 동일한 역할을 하는 1x1 Conv2D로 변경하였다. 즉 첫 번째 방법으로 생성된 모델은 bias가 포함 되어있지 않으며, 두 번째 방법으로 생성된 모델은 bias가 포함 되어있다.

마지막으로, feed-forward 시 사용하는 MLP(Multi-Layer Perceptron)에 ReLU(Rectified Linear Unit) 활성화 함수를 사용하였다. 이는 NPU가

TFLite(TensorFlow Lite)의 GELU(Gaussian Error Linear Unit) 연산을 지원하지 않기 때문이다.

2. 모델 경량화

STM32N6에 내장된 NPU는 8비트 정수(int8)로 양자화된 모델만 지원하기에, TensorFlow 2.7.0에 탑재된 TFLiteConverter를 사용하여 훈련 후 양자화 방식인 PTQ(Post-Training Quantization)를 수행하였다. 모델의 가중치를 int8로 양자화하였으며, 입력은 비부호 8비트 정수(uint8), 출력은 32비트 부동소수점(float32)으로 설정하여 양자화를 수행하였다.

3. 모델 컴파일

모델을 STEdgeAI를 활용하여 컴파일을 수행하였다. 그림 1은 ViT에 사용된 연산이 NPU와 CPU 중 어디에서 실행되는지를 나타낸다. 컴파일 과정에서 QKV projection, feed-forward 등에 사용되는 FC와 1x1 Conv2D는 NPU에 할당되었다. 한편, NPU는 두 입력이 모두 변수인 행렬 연산을 지원하지 않아 ViT의 self-attention에 사용되는 matmul 연산은 CPU에 할당되었다.

Ⅲ. 실험

1. 실험 설정

ViT의 NPU 가속 유무에 따른 추론 시간을 평가하기 위해 총 4종류의 ViT를 설계하였다. 파라미터 개수가 적은 순으로 ViT-T(Tiny), ViT-S(Small), ViT-B(Base), ViT-L(Large)을 설계하였으며, bias 파라미터 여부에 따른 자세한 ViT의 파라미터 개수와 MAC(Multiply-ACcumulate) 연산량은 표 1과 같다. 표 1에서 'w/o bias'는 bias가 없는 모델, 그리고 'w/ bias'는 bias가 있는 모델을 의미한다.

모델 추론 시간을 평가하기 위해 MCU 내부에 CPU와 NPU를 내장한 STM32N6의 개발 보드 STM32N6570-DK와 CPU만 내장한 STM32H7의 개발 보드 NUCLEO-H753ZI에 탑재하여 실험을 수행하였다. 모델의 학습은 PC에서 진행하였으며, 사용된 데이터셋은 AI Planet에서 진행된 Data Sprint #25: Flower Recognition 챌린지에 사용된 데이터로 5종류의 생화로 구성된 이미지 데이터셋이다. 평가 보드의 제원은 표 2와 같다.

표 1 NPU 탑재용 Vision Transformer 구조

	ViT-T	ViT-S	ViT-B	ViT-L
Image	96	160	160	224
Patch	16	16	16	16
Hidden size D	128	128	144	176
Layers	4	6	6	6
Heads	4	4	4	8
MLP size	128	256	576	704
Params(w/o bias) (M)	0.494	0.889	1.608	2.371
MACs(w/o bias) (M)	20.3	113.4	188.0	609.5
Params(w/ bias) (M)	0.498	0.894	1.616	2.380
MACs(w/ bias) (M)	20.4	113.6	188.4	610.5

표 2 디바이스 제원

	STM32N6570-DK	NUCLEO-H753ZI
CPU	Arm Cortex M55@800MHz	Arm Cortex M7@480MHz
	1352DMIPS@800MHz	1027DMIPS@480MHz
NPU	ST Neural-ART Accelerator@1GHz	-
	600GOPS@1GHz	
Flash memory	128MB Octo-SPI flash memory@200MHz	2MB@240MHz
memory	NPU RAM 448kB x4 @900MHz	1MB@240MHz
	System RAM 1MB+624kB+400kB@400MHz	
	32MB Hexadeca-SPI PSRAM@200MHz	

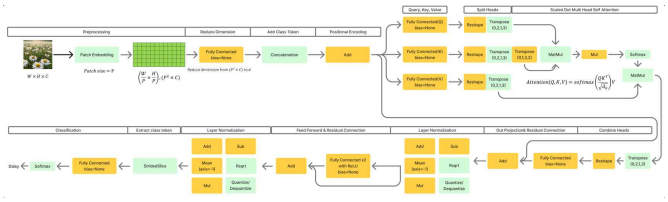

그림 1 ViT 모델에 사용된 연산의 CPU와 NPU 할당 내역. PE 부분은 전처리 과정이다. 주황색으로 표시된 연산은 NPU에서, 연녹색으로 표시된 연산은 CPU에서 실행된다. bias를 포함하는 모델은 classification에 사용된 FC를 제외한 FC가 bias 파라미터가 포함된 1x1 Conv2D로 변경된다.

2. 실험 결과

본 논문에서는 각 장치별 모델 추론 시간을 비교 분석하였다. 표 3은 각 장치에서 PE 처리 시간을 제외한 ViT의 추론 시간을 비교한 결과이다. NPU를 탑재하지 않은 MCU에 비해 NPU를 탑재한 MCU에서 추론 시 추론 시간이 평균 약 91% 감소하였다. 이는 ViT에 사용된 FC, 1x1 Conv2D 등의 연산이 NPU에서 병렬적으로 가속 처리되어, 연산 속도가 향상되었기 때문으로 판단된다.

ViT-L이 ViT-B 대비 증가한 연산량에 비해, NPU 보드에서 추론 시간이 더 증가한 것을 확인할 수 있다. 이는 CPU에서 실행되는 self-attention 연산량은 패치 개수의 제곱에 비례하므로, 패치 개수 증가에 따라 CPU 연산량이 크게 증가하였기 때문이다. 이는 향후 NPU가 self-attention 연산을 지원하게 되면 추론 시간이 더 감소할 수 있음을 보여준다. 메모리 사용량이 1MB를 초과하는 ViT-L 모델의 경우 메모리 용량이 제한적인 NUCLEO-H753ZI 보드에 탑재가 불가능하였다.

표 3 Vision Transformer 추론 시간 비교

	보드	ViT-T	ViT-S	ViT-B	ViT-L
추론 시간(ms) (w/o bias)	STM32N6	12	72	82	417
	H753ZI	142	748	968	-
추론 시간(ms) (w/ bias)	STM32N6	9	61	71	373
	H753ZI	108	632	826	-

Ⅵ. 결론

본 논문은 NPU를 내장한 MCU에서 ViT의 가속 가능성을 검증하였다. 이를 위해 기존 ViT 구조를 NPU에 최적화하기 위해 모델 구조 개선 및 경량화를 수행하였다. 실험을 통해 ViT가 NPU에서 실행 시 CPU에서 실행될 때보다 추론 시간이 평균 약 91% 감소함을 검증하였다. 향후 연구로는 ViT 모델의 self-attention 연산을 지원하는 Ethos U85 NPU를 내장한 Alif Ensemble E8 MCU와의 연산 속도 비교를 통해 온디바이스 ViT 모델의 실시간성을 검증한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임 (IITP-2026-RS-2023-00260098)

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2026-RS-2023-00254177)

참 고 문 헌

[1] Wang, Xubin, Zhiqing Tang, Jianxiong Guo, Tianhui Meng, Chenhao Wang, Tian Wang, and Weijia Jia. "Empowering edge intelligence: A comprehensive survey on on-device ai models." ACM Computing Surveys 57, no. 9 (2025): 1-39.

[2] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).