온디바이스 소형언어모델 기술개발 동향

Technical Trends in On-device Small Language Model Technology Development

김근용 (G. Kim, gykim@etri.re.kr) 엣지컴퓨팅응용서비스연구실 책임연구원
윤기하 (K. Yoon, kiha@etri.re.kr) 엣지컴퓨팅응용서비스연구실 연구원
김량수 (R. Kim, rskim@etri.re.kr) 엣지컴퓨팅응용서비스연구실 선임연구원
류지형 (J. H. Ryu, jihyoung@etri.re.kr) 엣지컴퓨팅응용서비스연구실 선임연구원
김성창 (S. C. Kim, sungchang@etri.re.kr) 엣지컴퓨팅응용서비스연구실 책임연구원/실장

ABSTRACT

This paper introduces the technological development trends in on-device SLMs (Small Language Models). Large Language Models (LLMs) based on the transformer model have gained global attention with the emergence of ChatGPT, providing detailed and sophisticated responses across various knowledge domains, thereby increasing their impact across society. While major global tech companies are continuously announcing new LLMs or enhancing their capabilities, the development of SLMs, which are lightweight versions of LLMs, is intensely progressing. SLMs have the advantage of being able to run as on-device AI on smartphones or edge devices with limited memory and computing resources, enabling their application in various fields from a commercialization perspective. This paper examines the technical features for developing SLMs, lightweight technologies, semiconductor technology development trends for on-device AI, and potential applications across various industries.

KEYWORDS LLM, SLM, Transformer, 온디바이스 AI

Ⅰ. 서론

인공지능 기술의 발전은 이제 단순한 ICT 기술 동향을 넘어 사회와 문화 그리고 경제에 미치는 영향을 고려할 때, 인류 발전의 궤적과 함께하고 있다고 할 수 있다. 2017년 구글 연구진에 의해 Sequence-to-Sequence 작업을 위해 탄생한 트랜스포머 모델(Transformer Model)[1]은 2022년 11월 OpenAI가 발표한 초거대언어모델(LLM: Large Language Model)인 ChatGPT(Chat + Generative Pre-trained Transformer)를 통해 인공지능 기술이 한 단계 진화했음을 입증했다.

* DOI: https://doi.org/10.22648/ETRI.2024.J.390409
* 본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다[No. 2021202090053B].

ChatGPT가 다양한 지식 분야에 걸쳐 상세하고 정교한 응답을 제공하면서 세계적인 주목을 받는 가운데 전 세계 빅테크 기업들은 새로운 LLM 모델을 연이어 발표하거나 기능을 꾸준히 강화하고 있다. 다양한 분야의 데이터로 사전 학습된 LLM을 파운데이션 모델이라 부르며, 이를 기반으로 특정 도메인 데이터로 추가 학습한 모델을 파인튜닝된 모델(Fine-tuned Model)이라 한다. OpenAI를 비롯해 Google, Meta, Naver 등 주요 빅테크 기업들이 파운데이션 모델 개발에 힘쓰고 있는 가운데, 최근 다양한 사업 모델에 적합한 소규모 언어모델(SLM: Small Language Model)이 주목받고 있다. 삼성전자가 최근 출시한 휴대폰 S24에는 자체 개발한 온디바이스 SLM인 가우스를 탑재하였다고 발표하여 화제가 되었다[2].

온디바이스AI는 인터넷 연결 없이 사용자 단말에서 AI 기능이 수행되는 것을 의미하는데, SLM 모델이 주목을 받으며 온디바이스AI의 전형적인 형태가 SLM으로 인식되고 있다. 온디바이스 SLM을 개발하기 위해서는 언어모델 구조의 경량화는 물론 모델에 최적화된 AI 반도체 개발 역시 뒷받침되어야 한다. 본고에서는 온디바이스 SLM 기술개발 동향에 대해 분석하고, 호남권연구본부 엣지컴퓨팅응용서비스연구실에서 추진하고 있는 다양한 산업과 연계한 온디바이스AI 기술개발 계획에 대해 소개한다.

본고는 Ⅱ장에서 SLM 연구개발 동향을 소개하고, Ⅲ장에서는 온디바이스AI를 위한 반도체 기술개발 동향을 소개한다. Ⅳ장에서는 온디바이스 SLM을 활용한 분야별 응용서비스개발 계획에 대해 소개하고, Ⅴ장에서 결론으로 마무리한다.

Ⅱ. SLM 연구개발 동향

1. LLM과 SLM

거대언어모델은 트랜스포머 모델을 기반으로 개발되어 학습되고 있으며, 최초 모델(Vanilla Transformer Model)은 encoder와 decoder가 조합된 encoder-decoder 모델이었다. 그러나 encoder-only 모델과 decoder-only 모델도 사용되며 최근의 언어모델들은 모두 decoder-only 모델 형태를 가지고 있다. LLM과 SLM의 분류를 위해 정확히 정의한 규격은 존재하지 않지만, LLM 대비 SLM은 매개변수(Parameter)가 적은 모델로 일반적으로 수십만에서 수십억 개 정도의 매개변수를 가진다. 현재 Meta사의 최신 모델인 Llama3 8B가 80억 개, Mistral AI사의 Mistral 7B가 70억 개의 매개변수를 가지고 있으며, 다른 경쟁사들의 SLM들이 그 이하의 매개변수를 가지는 것을 볼 때, 80억 개 이하의 개수를 가지는 SLM 모델들의 개발이 주를 이룰 것으로 예상된다. 그러나 온디바이스AI 형태로 클라우드와의 통신 없이 모델을 구동시키는 AI SoC칩이나 메모리와 프로세서를 병합하는 기술이 개발되고 있어 추후에는 수십억 개 이상의 매개변수를 갖는 모델도 온디바이스AI 형태로 구동될 수 있을 것이다.

일반적으로 SLM 모델은 LLM과 비교하여 훈련을 위한 데이터셋의 규모가 작으며, 컴퓨팅 자원 요구사항이 낮기 때문에 메모리와 컴퓨팅 자원이 제한적인 휴대폰이나 엣지 단말 등에서 온디바이스AI 형태로 구동이 가능하다. 언어모델의 성능을 측정하는 지표는 여러 개가 있으며, 대표적으로 광범위한 주제에 대한 AI의 이해 능력과 그 이해를 기반으로 한 태스크 수행 능력을 측정하는 MMLU(Massive Multitask Language Understanding)와 여러 가지 자연어 처리 태스크들을 통해 언어모델의 일반적인 언어 이해 및 생성 능력을 종합적으로 평가하는 MT-Bench(Multitask Benchmark)가 있다[3,4]. 본고에서는 SLM들의 성능 비교를 위해 해당 논문이나 웹페이지에서 공개하는 MMLU와 MT-Bench 점수를 사용한다.

표 1 주요 SLM의 매개변수 비교

모델명 / 매개변수	Llama3 8B(1)	Mistral 7B	Gemma 2B	Gemma 7B	Phi-3-mini(3) (3.8B)
트랜스포머 모델 형태	Decoder-only				
최대 위치 임베딩 크기 (max_position_embeddings)	8,192	131,072(2) (4096*32)	8,192	8,192	4,096
토크나이저 어휘 개수(vocab_size)	128,000	32,000	256,128	256,128	32,064
토큰 벡터 표현 차원 (token_hidden_size)	-	4,096	2,048	3,072	3,072
어텐션 헤드 수 (n_heads)	-	32	8	16	32
Key-Value 어텐션 헤드 수 (n_KV_heads)	-	8	1	16	32
MLP의 은닉층 차원 (hidden_dim)	-	14,336	32,768	49,152	8,192
트랜스포머 블록 개수 (num_hidden_layers)	-	32	18	28	32

* 매개변수 설명

① 트랜스포머 모델 형태: 최초의 트랜스포머 모델(Vanilla Transformer)은 encodor-decoder 형태로 되어 있으나, 최근의 SLM들은 모두 decoder-only 모델 기반임

② 최대 위치 임베딩 크기: 입력 시퀀스의 최대 길이로 토크나이저를 거쳐 입력되는 최대 토큰 수를 의미. 이 값이 크면 더 긴 시퀀스를 입력할 수 있으나, 더 큰 메모리가 필요함

③ 토크나이저 어휘 개수: 토크나이저가 인식할 수 있는 고유 토큰 개수. 이 값이 크면 새로운 표현에 대한 고유 토큰을 가져 단어 간 작은 의미도 잡아낼 수 있는 장점이 있으나, 매핑 테이블의 증가로 더 큰 메모리가 필요함

④ 토큰 벡터 표현 차원: 토큰을 나타내는 벡터의 차원. 이 값이 크면 토큰 간 더 깊은 상관관계를 포착할 수 있으나, 더 큰 메모리가 필요함

⑤ 어텐션 헤드 수: 멀티헤드 어텐션에서 사용되는 헤드의 개수

⑥ Key-Value 어텐션 헤드 수: Key/Value 어텐션 헤드의 개수

⑦ MLP의 은닉층 차원: 트랜스포머 모델의 MLP(Multi-Layer Perception) 부분에서 사용되는 은닉 표현 차원 수

⑧ 트랜스포머 블록 개수: 트랜스포머의 인코더나 디코더의 레이어 개수. Decoder-only 모델에서는 쌓여있는 decoder 개수를 의미함. 모델의 크기와 복잡도에 가장 크게 영향을 미치는 것으로, 일반적으로 이 값이 커지면 모델의 성능이 향상됨

(1): 2024년 5월 현재 기술 규격이 공개되지 않음. 메타에서 곧 논문을 통해 세부규격을 발표할 예정

(2): 최대 4096 길이의 슬라이딩 윈도우 사용 시 입력 시퀀스 최대 길이. 슬라이딩 윈도우 미사용 시는 4,096임

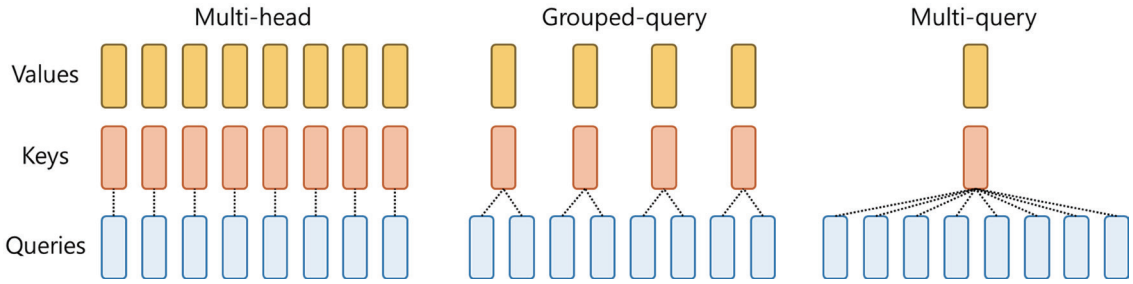
(3): Hugging face에 공개되어 있는 Phi-3-mini-4k-instruct 모델의 설정값을 참고하였음

2. 주요 Foundation SLM

구글, 마이크로소프트, 메타 등 빅테크뿐만 아니라 Mistral AI와 같은 스타트업 등 많은 회사가 SLM 파운데이션 모델을 개발하고 있으며, 성능 개선을 위해 노력하고 있다. 본 절에서는 주요 SLM 모델들의 특징 및 성능에 대해 알아보도록 한다. 표 1은 각 모델의 주요 기술 규격을 비교한 표로 모델 간 경량화 비교를 위한 매개변수를 중심으로 나타내었다. 삼성전자 가우스는 세부규격을 발표하지 않아, 이 표에 포함하지 않았다.

가. Llama3 8B

메타의 거대언어모델인 Llama는 OpenAI사의 ChatGPT와 비교되는 대표적인 오픈소스 LLM이다. Llama3는 Llama의 최신 모델로, 2024년 5월 현재 메타에서는 Llama3에 대한 논문을 발표하지 않은 상태이지만, 홈페이지를 통해 주요 기술적 특징과 성능에 대해 밝히고 있으며 가까운 시일 내에 논문을 발표할 것이라고 말하고 있다[5]. Llama3는 8B



출처 Reprinted from [6]. CC BY.

그림 1 Multi-head, Grouped-query, Multi-query 어텐션

와 70B 매개변수를 가지는 두 개의 모델이 발표되었다. Llama3는 그림 1과 같이 복수의 Query 그룹이 공통된 Key를 공유하는 GQA(Grouped Query Attention) 기법[6]을 통해 추론 효율을 높였다고 밝히고 있다. 또한 128K 토큰의 어휘를 가지는 토크나이저를 사용했으며, 최대 입력 토큰의 길이는 8,192이다. 메타의 홈페이지에 따르면 Llama3 8B의 Instruct 모델은 MMLU 벤치마크에서 68.4%, Pre-trained 모델은 66.6% 성능을 보이며. MT-bench 점수는 공개하지 않았다.

나. Mistral 7B

메타 Platform사와 구글 Deepmind사의 과거 직원들이 2023년 4월에 프랑스에서 설립한 Mistral AI 사에서 개발한 SLM이다. Mistral 7B 모델의 가장 큰 기술적 특징은 슬라이딩 윈도우 어텐션(Sliding Window Attention) 기법으로 긴 문맥의 의존성을 효과적으로 포착하기 위해 사용하는 어텐션 기법이다. 입력 시퀀스를 고정된 크기의 윈도우로 분할하여, 각 윈도우 내에서는 전체 어텐션을 계산하고, 윈도우 간의 어텐션은 윈도우의 시작 부분 간 어텐션을 계산하여 컨텍스트의 장기 의존성을 포착한다. Mistral 7B 참고문헌 [7]에 따르면 슬라이딩 윈도우 어텐션을 사용하면 입력 시퀀스의 모든 토큰 간 어텐션을 계산하지 않기 때문에 계산 복잡도를 낮추고 메모리 사용량도 줄이는 효과를 얻을 수 있는 장점이 있다. 논문에 따르면 Mistral 7B 모델은 MMLU 벤치마크에서 60.1%, MT-bench에서 6.84를 달성하였다.

다. Gemma

Gemma는 2024년 2월에 구글이 발표한 SLM으로 2B 매개변수 모델과 7B 매개변수 모델이 있다[8]. 구글은 Gemini라는 명칭의 멀티모달 LLM을 개발하여 출시하였으며, Gemini nano라는 온디바이스 서비스를 위한 경량화된 모델도 있다. 이에 반해 Gemma는 언어만을 처리하며 GPU와 TPU가 장착된 시스템에 적합한 7B 모델과 CPU와 온디바이스 어플리케이션 구동에 적합한 2B 모델이 있다. Gemma 모델의 주요 기술적 특징은 MQA(Multi-Query Attention), RoPE(Rotary Position Embedding), GeGLU(Gated Gated Linear Units), RMSNorm 등이 있다. MQA는 n개의 서로 다른 쿼리 벡터를 생성하여 이들의 가중치를 결합하여 최종 어텐션을 출력하는 메커니즘이다. 구글은 7B 모델에서 MHA(Multi-Head Attention) 기법을 사용하였고, 2B 모델은 MQA를 사용하였다고 밝혔는데, 이는 2B 모델의 경량화를 위해 MHA에 비해 계산량이 적은 MQA

를 채용한 것으로 파악된다. RoPE는 싸인/코싸인 함수를 사용하여 상대적 위치 정보를 벡터로 표현한 매핑이고, GeGLU는 모델에서 사용되는 활성화 함수로 두 개의 게이트를 가지고 있다. RMSnorm은 입력값을 정규화(Normalization)하는 것으로 학습의 안정화를 위해 사용되는 기법이다. 구글은 이런 일련의 기술들과 정제된 데이터가 Gemma의 성능 향상에 기여했다고 밝혔다. 논문에 따르면 MMLU 벤치마크에서 Gemma 2B 모델은 42.3%, 2B 모델은 42.3%를 달성하였으며, MT-Bench 성능은 공개되어 있지 않다.

라. Phi-3

마이크로소프트사에서 2024년 4월 발표한 Phi-3 모델은 Phi-2 모델에 사용한 데이터셋을 가공하여 더욱 성능을 높인 모델이다[9]. 마이크로소프트는 학습 데이터셋의 품질이 SLM의 성능 향상에 매우 큰 영향을 끼친다고 지속적으로 강조하고 있다[10]. Phi-3 모델은 Phi-3-mini, Phi-3-small, Phi-3-medium 시리즈가 있는데, 차례대로 파라미터 개수가 3.8B, 7B, 14B이다. 시리즈 중 Phi-3-mini는 전형적인 온디바이스AI 모델로 논문에서 'Highly capable language model running locally on a cell-phone'으로 소개되고 있다. 표 1에서 보는 바와 같이 Phi-3-mini 모델은 최대 위치 임베딩 크기와 토크나이저의 최대 어휘 개수가 다른 모델에 비해 적음에도 불구하고, 논문에 따르면 MMLU 벤치마크에서 68.8%, MT-bench에서 8.38을 달성하였다[9]. 마이크로소프트사는 Phi-3-mini의 성능 향상은 Phi-2에서 사용한 데이터를 기반으로 잘 정제한 웹 데이터와 LLM을 활용하여 신규로 생성한 데이터에 기인한다고 말하고 있다. 그러나 작은 모델의 크기로 인해 사실적 지식(Fractual Knowledge)에 대한 부족을 약점으로 인정하고 있으며, 대안으로 외부의 검색엔진을 이용한 데이터 증강(RAG: Retrieval-Augmented Generation)을 제시하고 있다.

3. 모델 경량화 기술

AI에서 모델 경량화 기술이란 학습이 완료된 딥러닝 모델의 성능은 최대한 유지하면서 작고 가벼운 형태로 모델을 변환하는 기술과 방법을 의미한다. 즉, 모델 경량화 기술은 학습이 완료된 딥러닝 모델의 저장 공간을 줄이고, AI 모델 추론 시 필요한 메모리 사용량과 복잡한 연산량을 감소시켜 서버가 아닌 임베디드 보드나 휴대폰과 같은 제한적 하드웨어 환경에서도 딥러닝 모델 기반 서비스를 제공할 수 있게 해주는 기술이다. 본 절에서 소개하는 대표적인 딥러닝 경량화 기술과 함께 SLM 구조를 경량화하는 기술을 소개한다.

가. 가지치기

대부분의 딥러닝 모델의 경우, 입력 데이터가 출력되기까지 다양한 구조로 설계된 연산 계층의 순차적인 연산을 통해 데이터 분석이 수행된다. 각 연산 계층에서는 이전 계층 연산의 결과 데이터에 대하여 선형 또는 비선형적 연산을 수행한 결과 데이터를 다음 계층으로 전달하는 구조로 설계되어 있다. 이때, 각 계층에서는 사전에 학습이 완료된 모델 가중치(Weights)를 활용해 연산을 수행하는데, 모델의 가중치에서 상대적으로 연산 결과에 영향이 적은 가중치를 삭제하는 방법이 가지치기(Pruning)이다. 가지치기를 사용하면 모델의 데이터 분석 정확도 손실은 최소화하면서 모델을 구성하는 가중치 수를 줄임으로써 모델 저장에 필요한 메모리 용량 절감 효과뿐만 아니라 각 계층 연산에 필요한 메모리 사용량 및 연산량을 줄이는 효과를 얻을 수 있다. 최근 가지치기 기술을 LLM에 적용하는 논문들이

발표되고 있다[11,12].

나. 양자화

PyTorch와 같은 딥러닝 모델 학습 플랫폼에서는 딥러닝 모델을 구성하는 가중치에 Floating Point 32bit(FP32) 메모리를 할당하여 변수를 선언하고 학습을 수행한다. 이때, 모델의 각 계층 연산의 무결성(Integrity)은 유지하면서 모델의 가중치 계수를 FP16 또는 INT8과 같이 메모리 할당 크기를 재구성하여 모델 저장에 필요한 저장용량 및 연산 메모리 사용량을 줄이는 방법이 양자화(Quantization) 방법이다. 앞서 언급한 가지치기 방식과 달리 양자화 방법은 완전히 삭제되는 가중치는 없으므로 연산의 무결성은 유지된다고 할 수 있지만, 가중치 값이 양자화되는 과정에서 양자화 오차가 발생하여 연산 결과가 달라지는 단점이 있다. 이를 극복하기 위해 가중치 양자화를 수행하면서 모델 학습에 활용되었던 데이터셋을 이용해 가중치 캘리브레이션을 수행하면 모델 양자화에서 발생하는 모델 정확도 손실을 보상할 수 있다. SLM 양자화의 예로, 마이크로소프트사는 Phi-3-mini 모델은 4-bit 양자화 시에 약 1.8GB 가량의 메모리를 차지하여 iPhone 14의 A16 바이오닉 칩에서도 초당 12토큰을 생성할 수 있다고 밝히고 있다[9].

다. 지식 증류

지식 증류(Knowledge Distillation) 방법은 앞서 언급한 가지치기/양자화 방법처럼 학습이 완료된 모델의 가중치를 직접적으로 수정하여 모델을 경량화하는 방법이 아니라 기존 모델(Teacher Model)보다 크기가 작은 모델(Student Model)을 설계하고, 해당 모델이 기존의 규모가 크고 잘 학습된 모델의 입/출력 데이터 연산 특성을 유지할 수 있도록 학습시켜 작은 모델이 기존 모델의 성능을 최대한 유지하도록 학습하는 방법이다. 이를 위한 방법으로 Teacher Model과 Student Model의 최종 계층 결과 데이터의 확률값을 출력하도록 Soft Label을 추가하고, 같은 입력 데이터에 대한 두 모델 사이의 출력 데이터 차이를 KL Divergence Loss를 이용해 측정한 뒤, 이를 최소화하도록 Student Model의 가중치를 업데이트하는 방식의 응답 기반 지식 증류(Response-based Knowledge Distillation) 방법이 대표적이다. 최근 사전학습된 LLM으로부터 SLM으로 지식 증류를 통해 지식을 전이하는 방법이 발표되고 있다[13,14].

라. 모델 구조 경량화

SLM은 LLM에 비해 작은 임베딩 벡터 차원, 제한된 어휘 집합의 토크나이저, 그리고 Multi-Query 어텐션과 같은 기법들을 사용하기 때문에 경량화된 구조 가진다. 트랜스포머 모델의 특성에 기반한 경량화로써 제한된 매개변수 수에도 불구하고 높은 추론 성능을 내기 위한 대표적인 최신 기술로 MoE(Mixture of Experts)가 주목받고 있다[15]. Mistral AI가 발표한 Mixtral of Experts는 한정된 모델 크기에서 성능을 높이고자 복수의 MLP 레이어(전문가 레이어, Experts)와 이들을 선택하는 라우터로 구성된다. 구체적으로, Mixtral of Experts 모델은 8개의 MLP 전문가 레이어와 하나의 라우터로 이루어져 있다. 각 타임스탬프에서 질의(Query), 키(Key), 값(Value) 어텐션 결과가 라우터를 통해 8개 전문가 레이어 중 2개에 매핑되고, 이 두 전문가의 출력이 병합되어 최종 결과를 산출한다. Mistral AI사는 논문을 통해 Mistral 7B 모델에 8개의 Experts를 사용한 Mixtral 8x7B 모델이 Llama2 70B와 GPT-3.5 모델과 비교하여 모든 평가 지표에서 우수한 성능을 보인다고 밝히고 있다[15].

Ⅲ. 온디바이스AI 반도체 기술 동향

Ⅱ장에서 설명한 바와 같이 언어모델 경량화를 위한 기술개발이 지속되고 있음에도 불구하고 실제 제품이나 서비스에 적용하기 위해서는 저전력으로 동작하는 고성능 컴퓨팅 자원이 필수적이다[16]. 특히 언어모델의 핵심 알고리즘인 Query, Key, Value 어텐션 기법에서는 메모리와 CPU 간 데이터 교환이 매우 빈번하게 일어나기 때문에, 이러한 병목 현상을 해결할 수 있는 온디바이스AI 반도체 기술개발은 기술적 측면뿐만 아니라 사업화 측면에서도 매우 중요한 의미를 지닌다.

최근 PC · 노트북은 물론 태블릿과 스마트폰에 적용되는 AP(Application Processor) SoC(System on a Chip)에 CPU와 GPU 외에 AI를 위한 NPU(Neural Processing Unit)를 탑재하는 것이 보편화되고 있으며, AI 모델 구동을 위한 NPU 성능 · 전력효율 개선과 AI 서비스를 구동할 때 발생하는 처리장치와 메모리 간 병목현상을 줄이기 위한 메모리 반도체 기술개발이 진행 중이다. 본 장에서는 AI SoC 반도체와 메모리 내에서 데이터 처리를 지원하는 구조인 메모리 반도체 기술개발 동향에 대해 소개한다.

1. AI SoC 반도체 기술개발 동향

최근 구글과 삼성전자 등 글로벌 업체에서 출시한 최신 스마트폰들은 다양한 생성형 AI서비스를 별도의 서버(클라우드)와 연결 없이 제공할 수 있는 점을 적극적으로 홍보하고 있다. 실시간 대화 번역(음성), 텍스트 번역, AI 사진 편집 등 사용자가 직관적으로 체감할 수 있는 AI 서비스를 구동하기 위해 하드웨어는 초미세 반도체 공정기반의 SoC를 통해 제작된다. 이와 같이 최근 생성형 AI서비스를 대대적으로 홍보하기 이전인 2015년부터 퀄컴은 Snapdragon 프로세서에 영상, 음성 및 센서 동작을 위한 AI Engine을 적용했으며, 이후 사진 편집, 동영상 등 각종 서비스 분야별 AI처리를 위한 엔진들을 추가하고 있다[17]. 애플도 2017년 출시된 아이폰 8시리즈에 탑재된 A11 SoC부터 약 600GOPS 성능의 2코어 NPU를 적용하여 Face ID 처리나 머신러닝 작업을 처리하도록 했으며, 최근 2023년 출시된 아이폰 15 Pro시리즈에 탑재된 A17 Pro SoC는 35TOPS 성능의 16코어 NPU를 적용하는 등 AI 처리성능을 지속적으로 향상시키고 있다[18]. 이와 같이 모바일프로세서 분야에서는 각 사용자가 주로 사용하는 콘텐츠나 보안 처리에 보다 효율적인 서비스 제공을 위한 NPU를 적용시켜 왔으며 최근 ChatGPT를 필두로 하는 생성형 AI 관련 서비스와 고성능 및 고효율의 NPU 기술개발로 PC · 노트북은 물론 태블릿, 스마트폰에 이르기까지 다양한 컴퓨팅 디바이스에 온디바이스AI가 확산되는 추세이다.

2. 메모리 반도체 기술개발 동향

AI 기술개발에서 컴퓨팅 구조에 의해 발생되는 CPU, GPU와 메모리 간 병목현상은 해결하기 어려운 문제로 손꼽힌다. 이런 병목현상은 처리속도 저하는 물론 전력소모에도 많은 영향을 끼치기 때문에 이를 해결하기 위한 반도체 기술개발은 클라우드를 구성하는 서버에서 동작하는 LLM은 물론, 온디바이스 SLM의 성능 향상을 위해서 매우 중요하다. 특히 토큰 간의 상관도를 학습하는 Q, K, V 어텐션 기법은 매우 빈번한 메모리 액세스가 필요하기 때문에 LLM을 위한 반도체 기술개발이 더욱 중요해지고 있다[6]. 이와 같은 병목현상을 개선하기 위한 대표적인 기술로 CXL(Compute Express Link)은 물론 그림 2에서 요약한 바와 같이 HBM(High Bandwidth Memory), PNM(Processing Near Memory),

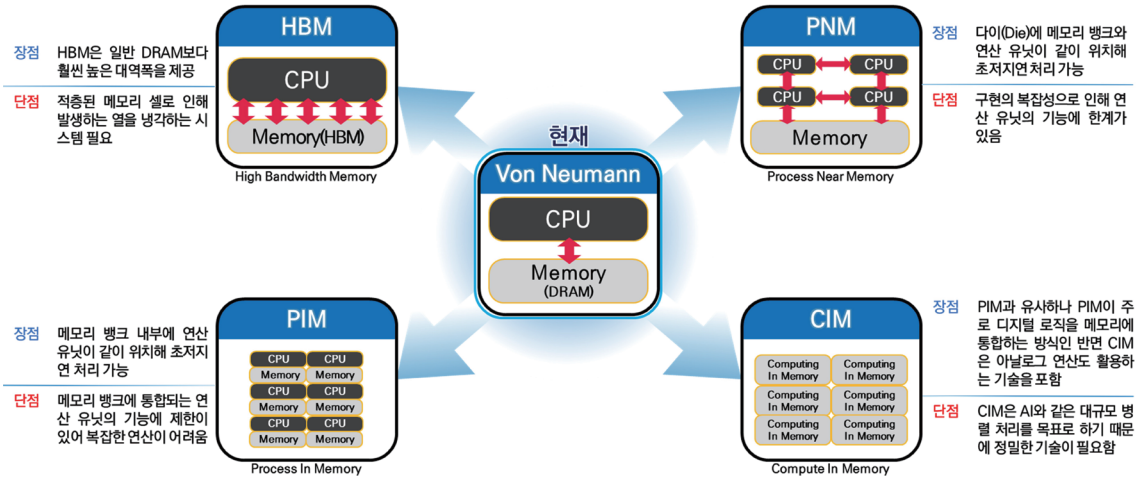


그림 2 CPU와 메모리가 융합된 구조의 메모리 반도체 기술

PIM(Processing In Memory), CIM(Compute In Memory) 등이 있다.

CXL은 현재 CPU와 메모리 사이의 Bus 구조에서 사용되고 있는 PCIe(Peripheral Component Interconnect Express)의 전기적, 물리적 인터페이스를 그대로 활용하는 프로토콜로써 기존 DRAM 확장에 따른 개수, 대역폭 등 제한된 확장성 문제를 개선하고 CPU를 비롯한 GPU나 NPU 같은 여러 장치가 CXL 메모리를 효율적으로 공유할 수 있도록 하는 기술이다. 2023년 5월에는 삼성전자가 세계 최초로 CXL 2.0을 적용한 128GB DRAM 개발에 성공했다고 발표했다[19].

HBM은 여러 개의 DRAM을 수직으로 쌓아 연결하여 기존의 DDR(Double Data Rate) 메모리보다 데이터 전달 속도를 대폭 증가시킨 방식으로 2013년 SK하이닉스에서 1세대 HBM을 최초 개발하여, 2015년 AMD Fiji GPU에 처음 적용되었다. 현재는 5세대 HBM인 HBM3E 메모리가 양산되고 있으며, 선두 주자인 SK하이닉스에서는 HBM3E 16단 기술을 2024년 2월 국제고체회로학회(ISSCC) 컨퍼런스에서 세계 최초로 공개했다.

PNM은 메모리 뱅크와 연산 유닛이 동일한 다이(Die)에 집적되어 초저지연 메모리 액세스가 가능한 메모리 반도체 기술이다. PIM은 메모리 뱅크 내부에 연산 기능을 더해 메모리 내부에서 연산하는 특징을 갖는다. 메모리 반도체 내부저장공간별 전용 데이터 대역을 활용하여 연산 속도를 향상시키고, 소량의 연산 결과 데이터만 CPU나 GPU로 전달할 수 있어 데이터 전달량을 획기적으로 줄일 수 있다. 또한, 이에 따라 전력 소모량 절감 효과까지 얻을 수 있다. 2023년 9월 SK하이닉스는 GDDR6기반으로 PIM기술을 적용한 GDDR6-AiM(Accelerator in Memory) 및 AiMX(AiM based Accelerator) 시제품을 공개했으며, AiMX는 'OPT(Open Pre-trained Transformer) 13B' 모델시연을 통해 GPU기반 컴퓨팅 시스템 대비 10배 이상 빠른 반응속도를 갖는 반면, 소모전력은 20% 수준임을 강조했다[20].

CPU와 메모리가 동일한 패키지에 통합되어 있는 PIM과 달리 CIM은 CPU와 메모리가 동일한 다이 안에 통합되어 있는 구조다. 다이란 회로가 제작되어 있는 반도체 물질의 자그마한 사각형 조각을 말한다. 아직 상용화 단계로 진입하지 못한 기술이지

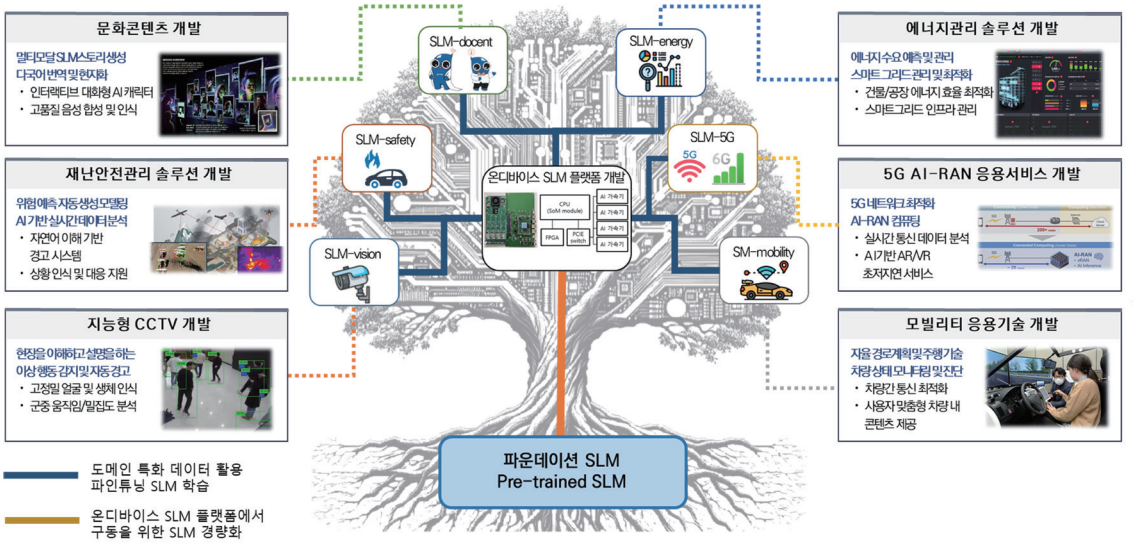


그림 3 도메인 지식 기반 온디바이스 SLM 개발 및 응용서비스 개발 전략도

만, 메모리 셀에서 데이터 이동 없이 연산이 수행되기 때문에 매우 높은 밀도와 에너지 효율을 가질 수 있어 최근 관련 논문들이 발표되고 있다[21].

Ⅳ. SLM 응용서비스 개발

도메인 특화 데이터로 파인튜닝된 SLM 모델은 다양한 산업 분야에 적용이 가능하다. 이를 위해서는 각 분야 고유의 전문 지식과 언어 패턴을 반영한 데이터셋이 구축되어야 한다. 또한, 최근 멀티모달 LLM의 개발 및 학습을 위해 이미지 데이터도 활용되고 있다[22]. 도메인 특화 데이터로 파인튜닝된 SLM은 기존 범용 모델 대비 특정 분야에서 높은 성능을 발휘할 수 있기 때문에 비즈니스 모델 수립에 중요한 역할을 한다.

그림 3은 엣지컴퓨팅응용서비스연구실에서 계획하고 있는 도메인 특화 데이터 기반 SLM 활용 서비스 개발 계획을 나타낸 그림이다. 그림 3에서 보는 바와 같이, 문화콘텐츠 분야에서는 온디바이스 SLM을 활용한 인터랙티브 대화형 AI 캐릭터, 고품질 음성 기반 콘텐츠 생성 등의 서비스가 가능해진다. 재난안전 솔루션에서는 자연어 이해를 기반으로 상황인지, 재난 피해 상황 분석 등에 활용될 수 있다. 지능형 CCTV에서는 실시간 영상 속 대화 인식, 비정상 행위 모니터링 등에 사용되며, 에너지관리 솔루션에서는 건물/공장 에너지 효율 최적화나 스마트그리드 인프라 관리에 도입될 수 있다. 뿐만 아니라 5G AI-RAN(Radio Access Network) 분야[23]에서는 5G 초저지연 서비스 구현을, 모빌리티 기술 분야에서는 자율주행 내 차량 내 맞춤형 콘텐츠 제공 등에 온디바이스 SLM이 핵심 역할을 할 것으로 기대된다. 이처럼 도메인 지식 기반의 특화 SLM 모델은 다양한 산업 분야에서 혁신적인 서비스를 창출할 수 있기 때문에 호남권연구본부 엣지컴퓨팅응용서비스연구실은 산업계와의 협력을 통해 SLM 생태계를 창출해 기여하도록 노력할 것이다.

Ⅴ. 결론

본고에서는 온디바이스 SLM의 기술개발 동향에

대해 살펴보았다. 앞으로 온디바이스 SLM 개발을 위해 모델 압축, 지식 축약, AI 반도체 친화적 구조 최적화 등 다양한 기술적 접근이 시도될 것으로 예상된다. SLM을 활용한 응용서비스 개발은 다양한 산업 분야에 걸쳐 새로운 비즈니스 기회를 창출하고 혁신을 이끌어낼 수 있는 기회를 제공할 수 있다. 특히 도메인 특화 지식을 효과적으로 학습시킨 언어모델을 구축하는 것은 해당 분야에 특화된 서비스를 제공하는 데 있어 필수적이다. 관련 분야의 전문 지식과 용어, 문맥 등을 깊이 있게 이해하고 적용할 수 있는 언어모델을 통해 보다 정교하고 실용적인 서비스를 구현할 수 있기 때문이다. 따라서 도메인 지식 기반의 SLM 개발은 사업화를 위한 핵심 요소라고 볼 수 있으며, 이를 위한 지속적인 연구와 투자가 이루어져야 할 것이다.

용어해설

어텐션 기법(Attention Mechanism) 컴퓨터 비전이나 언어처리 분야에서 널리 사용되는 기법으로 입력과 출력 데이터의 연관성을 계산하여 가중치를 부여함으로써 출력 품질을 높이는 기법

온디바이스AI(On-device AI) 휴대폰이나 엣지 단말에서 클라우드와의 통신 연결 없이 AI 서비스를 구동하는 것을 일컫는 용어로 SLM의 부각과 함께 주목받고 있음

Small Language Model 거대언어모델과 비교하여 매개변수 개수가 적은 언어모델을 지칭하는 용어. sLLM이라는 명칭으로도 사용되었으나 최근에는 SLM을 사용하는 것이 일반적임

Transformer 모델 2017년 구글에서 자연어 처리를 위해 발표한 딥러닝 모델로 현재 언어모델의 기반이 되는 모델

약어 정리

AP	Application Processor
CIM	Computing In Memory
CXL	Compute Express Link
DDR	Double Data Rate
GeGLU	Gated Gated Linear Units
GQA	Grouped-Query Attention
LLM	Large Language Model
MHA	Multi-Head Attention
MMLU	Massive Multitask Language Understanding
MoE	Mixture of Experts
MQA	Multi-Query Attention
NPU	Neural Processing Unit
PCIe	Peripheral Component Interconnect Express
PIM	Processing In Memory
PNM	Processing Near Memory
RAG	Retrieval-Augmented Generation
RoPE	Rotary Position Embedding
SLM	Small Language Model
SoC	System on a Chip

참고문헌

[1] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, (Long Beach, CA, USA), Dec. 2017.

[2] Samsung Newsroom, "삼성전자, '삼성 AI 포럼'서 자체 개발 생성형 AI '삼성 가우스' 공개," 2023. 11. 8.

[3] D. Hendrycks et al., "Measuring massive multitask language understanding," in Proc. ICLR, (Virtual Only), May 2021.

[4] L. Zheng et al., "Judging LLM-as-a-judge with MT-bench and chatbot arena," in Proc. NeurIPS, (New Orleans, LA, USA), Dec. 2023.

[5] Meta, Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date, https://ai.meta.com/blog/meta-llama-3/

[6] J. Ainslie et al., "GQA: Training generalized multi-query transformer models from multi-head checkpoints," in Proc. EMNLP, (Singapore, Singapore), Dec. 2023.

[7] A.Q. Jiang et al., "Mistral 7B," arXiv preprint, CoRR, 2023, arXiv: 2310.06825.

[8] Google Gemma Team, "Gemma: Open models based on gemini research and technology," arXiv preprint, CoRR, 2024, arXiv: 2403.08295.

[9] M. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," arXiv preprint, CoRR, 2024, arXiv: 2404.14219.

[10] S. Gunasekar et al., "Textbooks are all you need," arXiv preprint, CoRR, 2023, arXiv: 2306.11644.

[11] X. Ma et al., "LLM-Pruner: On the structural pruning of large language models," in Proc. NeurIPS, (New Orleans, LA, USA), Dec. 2023.

[12] M. Sun et al., "A simple and effective pruning approach for large language models," in Proc. ICLR, (Vienna, Austria), May 2024.

[13] Y. Gu et al., "MiniLLM: Knowledge distillation of large language models," in Proc. ICLR, (Vienna, Austria), May 2024.

[14] W. Liu et al., "Mind's mirror: Distilling self-evaluation capability and comprehensive thinking from large language models," arXiv preprint, CoRR, 2024, arXiv: 2311.09214v3.

[15] A.Q. Jiang et al., "Mixtral of experts," arXiv preprint, CoRR, 2024, arXiv: 2401.04088.

[16] 전원, 여준기, "초거대 인공지능 프로세서 반도체 기술 개발 동향," 전자통신동향분석, 제38권 제5호, 2023.

[17] Qualcomm, Unlocking On-Device Generative AI With an NPU and Heterogeneous Computing, 2024. 2.

[18] CNET, Apple A17 Pro: The New Chip Brain in the iPhone 15 Pro, Pro Max, 2023. 9. 12.

[19] Samsung Newsroom, "삼성전자, 업계 최초 'CXL 2.0 D램' 개발," 2023. 5. 12.

[20] SK하이닉스 뉴스룸, "[2023 AI 메모리 결산] HBM·PIM·CXL 라인업 '탄탄' SK하이닉스, Global No.1 AI Company로 도약한다," 2023. 12. 20.

[21] S. Jung et al., "A crossbar array of magnetoresistive memory devices for in-memory computing," Nature, vol. 601, 2022, pp. 211-216.

[22] H. Liu et al., "Visual instruction tuning," in Proc. NeurIPS, (New Orleans, LA, USA), Dec. 2023.

[23] AI-RAN alliance, https://ai-ran.org/news/industry-leaders-in-ai-and-wireless-form-ai-ran-alliance/