Heungkuk
흥국증권

반도체

2026.03.27

KV Cache 압축은 이미 존재했다

반도체, 연구원
손인준
02) 6260-2470
sij96@
heungkuksec.co.kr

KV Cache 압축에 대한 연구는 새롭지 않다

NVIDIA의 KVTC(KV Cache Transform Coding), Google의 Turbo Quant 논문이 시장에서 회자되며 메모리 업종 주가 급락의 배경으로 거론되고 있다. 두 논문 모두 KV Cache를 압축해 대규모 추론 과정에서 발생하는 메모리 병목, 그리고 이에 따른 AI 워크로드 처리 속도 저하 문제를 완화하는 데 목적이 있다. KVTC는 JPEG와 유사하게 중요한 정보만 선별적으로 남기는 압축 방식을 사용했고, Turbo Quant의 경우 KV Cache 벡터를 랜덤 회전한 뒤 각 좌표를 양자화하는 방식을 제안했다. 논문에 따르면, 큰 성능 저하 없이 KVTC는 약 20배, Turbo Quant는 약 6배 수준의 KV Cache 압축이 가능하다.

다만 해당 논문의 공개 시점에 주목할 필요가 있다. KVTC는 2025년 11월, Turbo Quant는 2025년 4월에 이미 공개되었던 논문이다. 두 논문 모두 4월 23일부터 개최 예정인 AI 학회 ICLR 2026에 채택됐고, 학회 일정을 앞둔 일반적인 홍보 과정에서 관련 기술에 대한 주목도가 과도하게 높아진 측면이 있다고 판단된다. 더욱이 KV Cache 압축을 다룬 연구는 새롭게 등장한 주제가 아니다. KVQuant(2024년 1월), KIVI(2024년 2월) 등 관련 연구는 과거부터 존재해왔고, 일부 기술은 실제 AI 워크로드에서 적용되고 있다. KVTC와 Turbo Quant가 기존 KV Cache 압축 기술 대비 혁신적인 방식이었다면, 작년부터 AI 업계에서 높은 관심을 받았을 가능성이 높다. 따라서 해당 기술이 단기간 내 광범위하게 배포될 가능성은 제한적일 것이라는 점을 알 수 있다. 결국 KV Cache 압축은 이미 활발히 연구돼 온 분야이며, AI 성능 개선이 본질적으로 컴퓨팅 파워, 데이터, 알고리즘이라는 3가지 축에 기반해 이루어져 왔다는 점에서 이러한 기술 발전의 방향성 자체는 새롭지 않다. 이러한 상황 속에서도 Google을 비롯한 하이퍼스케일러 업체들은 메모리 공급 부족에 대응하기 위해 3~5년 단위 장기 공급 계약을 준비 중인 상황으로 파악된다.

일시적 조정 단계, 펀더멘털에 기반해 주가는 장기 우상향 추세 유지할 것

메모리 업종에 대한 시장 기대치가 높아진 상황에서, 최근 투자자들의 관심이 메모리에서 비메모리로 이동할 만한 여러 이슈가 발생했다. NVIDIA의 Groq 3 LPX 공개, Tesla의 Terafab 계획 발표, Arm EveryWhere 행사 등이 대표적이다. Groq 3 LPX는 SRAM 기반 가속기로써 HBM TAM 잠식에 대한 우려를 자극했고, Terafab은 자율주행차·로봇 및 우주용 로직 칩 생산에 초점이 맞춰져 선단 파운드리 쇼티지에 대한 기대를 한층 키웠다. 또한 Arm은 투자자 행사를 통해 자체 서버용 CPU 출시와 함께, 공격적 TAM 전망치(FY2031 기준 $100B 이상)를 제시한 바 있다. 다만 이 모든 이슈들에도 불구하고, AI 하드웨어 업종 내에서 메모리는 가장 높은 수익성과 가장 낮은 벨류에이션을 동시에 보이고 있고, 더불어 최대 5년 단위의 장기 공급 계약 확대를 통해 AI 인프라의 최대 병목 지위 역시 확인되고 있다. 최근 급락은 펀더멘털 훼손이 아닌 시장 수급 및 관심 분산에 따른 일시 조정으로 판단되며, 장기 호황 및 실적 추정치 상향, 벨류에이션 리레이팅에 따른 주가 우상향이 나타날 것이라 판단한다.

그림 1 KVTC: KV Cache 압축비에 따른 벤치마크 결과 비교. 20배까지 압축해도 안정적임을 주장

	Vanilla	$GEAR_{2\text{-bit}}$	$KIVI_{2\text{-bit}}$	H_2O	TOVA	xKV	FP8	$kvtc_{8\times}$	$kvtc_{16\times}$	$kvtc_{32\times}$	$kvtc_{64\times}$
Llama 3.1 8B											
CR	1	5	5	8	8	1-5	2	9-10	18-22	34-44	60-88
GSM8K	**56.8**	52.8	52.8	54.3	54.5	**56.6**	55.2	**57.0**	**56.9**	**57.8**	**57.2**
MMLU	**60.5**	**59.6**	**59.6**	44.3	44.8	**59.5**	**60.1**	**59.8**	**60.1**	**60.6**	**60.7**
QASPER	**40.4**	**40.4**	39.1	34.3	38.6	35.6	**40.8**	**40.1**	**40.7**	**39.4**	37.8
LITM	**99.4**	96.9	88.8	20.2	1.2	**99.9**	**99.4**	**99.3**	**99.3**	**99.1**	90.2
RULER-VT	**99.8**	**99.8**	**98.9**	50.4	**99.7**	**99.8**	**99.9**	**99.1**	**99.1**	**98.9**	95.9
MN-Minitron 8B											
CR	1	5	5	8	8	1-5	2	10-11	17-21	32-46	53-95
GSM8K	**59.1**	57.9	58.0	55.3	**59.2**	**59.3**	**60.1**	**60.6**	**60.3**	**59.1**	57.8
MMLU	**64.3**	**63.6**	63.2	43.5	48.1	63.1	**64.3**	**64.2**	**64.1**	**63.7**	62.1
QASPER	**38.2**	**38.2**	**38.2**	30.0	33.9	34.5	**38.3**	**39.1**	**38.6**	**37.7**	**38.1**
LITM	**99.8**	96.0	86.3	16.6	0.3	**99.6**	**99.8**	**99.4**	**99.3**	86.9	59.5
RULER-VT	**99.4**	98.3	96.8	39.2	**99.3**	**99.1**	**99.2**	**98.8**	**98.8**	96.0	93.4
Mistral NeMo 12B											
CR	1	5	5	8	8	1-5	2	10-11	17-20	31-43	51-87
GSM8K	**61.9**	59.8	59.7	57.0	60.3	**61.9**	**61.7**	**62.5**	**62.0**	**62.2**	**61.9**
MMLU	**64.5**	**64.0**	**64.3**	45.4	49.0	**63.9**	**64.5**	**64.6**	**64.4**	**63.8**	61.4
QASPER	**38.4**	**38.6**	**38.2**	29.5	36.0	33.5	**37.9**	**37.6**	**37.6**	**37.5**	**38.0**
LITM	**99.5**	96.9	91.9	16.2	8.7	97.9	**99.0**	**99.9**	**99.8**	**99.6**	95.3
RULER-VT	**99.8**	**99.4**	98.3	35.2	**99.6**	**99.4**	**99.8**	**99.5**	**99.5**	**98.9**	98.0

자료: NVIDIA, 흥국증권 리서치센터

그림 2 Turbo Quant: KV Cache 최대 6배(16 → 2.5비트) 압축 시에도 벤치마크 결과 차이 없음을 주장

Method	KV Size	SingleQA	MultiQA	Summarization	Few shot	Synthetic	Code	Average
Llama-3.1-8B-Instruct								
Full Cache	16	45.29	45.16	26.55	68.38	59.54	46.28	50.06
KIVI	3	43.38	37.99	27.16	68.38	59.50	44.68	48.50
KIVI	5	45.04	45.70	26.47	68.57	59.55	46.41	50.16
PolarQuant	3.9	45.18	44.48	26.23	68.25	60.07	45.24	49.78
TurboQuant (ours)	2.5	44.16	44.96	24.80	68.01	59.65	45.76	49.44
TurboQuant (ours)	3.5	45.01	45.31	26.00	68.63	59.95	46.17	50.06
Ministral-7B-Instruct								
Full Cache	16	47.53	49.06	26.09	66.83	53.50	47.90	49.89
TurboQuant (ours)	2.5	48.38	49.22	24.91	66.69	53.17	46.83	49.62

자료: Google, 흥국증권 리서치센터

표 1 KV Cache 압축 관련 주요 논문

논문	발표일	핵심 기법
Scissorhands	2023.03.26	중요도 지속성 가설, 일관된 attention weight 패턴의 pivotal token만 유지
H_2 O	2023.06.24	Heavy Hitter(H_2) 토큰 식별 → 누적 attention score 기반 동적 eviction
Efficient StreamingLLM	2023.09.29	Attention Sink 발견: 초기 토큰 KV + 최근 토큰 KV만 유지 → 무한 길이 스트리밍
KVQuant	2024.01.31	Per-channel Key 양자화, Pre-RoPE 양자화, 비균일 양자화, Dense-and-Sparse 분리
KIVI	2024.02.05	Key per-channel, Value per-token 비대칭 2bit 양자화, Tuning-free
GEAR	2024.03.08	양자화, low-rank 행렬로 오차 근사, sparse 행렬로 outlier 보정
QuaRot	2024.03.30	Hadamard 회전으로 outlier 제거 → weight/activation/KV cache 전체 4bit 양자화
CQ (Coupled Quantization)	2024.05.07	채널 간 상호의존성 활용 → 다채널 결합 양자화; joint entropy 최적화
SnapKV	2024.04.22	Observation window로 head별 중요 토큰 사전 식별 → 선별적 KV 유지
RotateKV	2025.01.25	Outlier-aware rotation + Pre-RoPE grouped-head rotation + Attention-sink-aware 양자화
Turbo Quant	2025.04.28	랜덤 회전 → Beta 분포 유도 → Lloyd-Max 최적 스칼라 양자화 + 1-bit QJL 잔차 보정
KVTC	2025.11.03	PCA 기반 feature decorrelation + 적응적 양자화 + 엔트로피 코딩; 미디어 압축 원리 응용

자료: 흥국증권 리서치센터

Compliance Notice

- 당사는 보고서 제공시점 현재 상기 종목을 1% 이상 보유하고 있지 않으며, 동 보고서를 기관투자가 또는 제3자에게 사전 제공한 사실이 없습니다.
- 동 자료의 조사분석 담당자 및 그 배우자 등 관련자는 상기 종목의 유가증권을 보유하고 있지 않습니다.
- 당사는 보고서 제공시점 기준으로 지난 6개월간 상기 종목의 유가증권 발행에 주간사로 참여한 사실이 없습니다.
- 본 자료에 게재된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭없이 작성되었음을 확인합니다.

본 보고서는 당사 고객들의 투자에 관한 정보를 제공할 목적으로 당사 고객에 한하여 배포되는 자료입니다. 본 보고서의 내용은 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻어진 것이나, 당사가 그 신뢰성이나 완전성을 보증하는 것이 아닙니다. 따라서 어떠한 경우에도 본 보고서가 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 또한 본 보고서의 지적재산권은 당사에 있으므로 당사의 동의없이 무단 복제, 대여, 전송, 변형 및 배포 될 수 없습니다.

흥국씨앗체

본 보고서에 사용된 흥국씨앗체는 흥국의 기업철학 모티브를 반영한 글씨체로서, 세계 3대 디자인상인 독일 '2015 iF 디자인 어워드'에서 커뮤니케이션 분야 브랜드 아이덴티티 부문 본상을 수상하였습니다. 친근하고 희망적인 느낌의 흥국씨앗체는 고객존중과 으뜸을 지향하는 흥국의 아이덴티티를 부각시킵니다.

Heungkuk
흥국증권

□ 주소 (본사) 서울시 영등포구 국제금융로 2길 32 (여의도파이낸스타워 14층)
(리서치센터) 서울시 영등포구 국제금융로 2길 32 (여의도파이낸스타워 6층)
□ 전화번호 영업부 대표 02)6742-3635
□ 팩스 영업부 대표 02)739-6286