YANGI O‘ZBEKISTON, YANGI TADQIQOTLAR JURNALI
Volume 2 Issue 9 18.05.2025
https://phoenixpublication.net/ Online ISSN: 3030-3494

XOTRANING SEGMENT-SAHIFALI TAQSIMLASH TIZIMI SUN’IY INTELLEKT VA CNN MODELLARIDA QO‘LLANILISHI

Umarov Begzodbek Azizovich
Farg‘ona davlat universiteti amaliy matematika va informatika kafedrasi o’qituvchisi ubaumarov@mail.ru
Rahimaliyev Mardonali Zokirzoda
Farg`ona Davlat Universiteti, talabasi
rahimalievmardonali@gmail.com

Annotation. *This article explores the application of segmented-paged memory management in artificial intelligence and machine learning, particularly in convolutional neural networks (CNN). It analyzes the memory demands of CNN models, GPU architecture, CUDA Unified Memory technology, and the efficiency of segmentation and paging in AI systems. Practical examples demonstrate how memory management approaches can serve as effective optimization tools in real-world projects.*

Keywords: *Artificial intelligence, machine learning, convolutional neural networks, segmentation, paging, GPU, CUDA, Unified Memory, memory management*

Аннотация . *В данной статье рассматривается применение сегментно-страничного распределения памяти в системах искусственного интеллекта и машинного обучения, в частности в сверточных нейронных сетях (CNN). Проанализированы требования к памяти моделей CNN, архитектура GPU, технология CUDA Unified Memory, а также эффективность сегментации и страничного распределения в ИИ-системах. Приведены практические примеры использования этих подходов в реальных проектах для оптимизации работы с памятью.*

Ключевые слова: *Искусственный интеллект, машинное обучение, сверточные нейронные сети, сегментация, страничная адресация, GPU, CUDA Unified Memory, управление памятью*

Anotatsiya. *Ushbu maqolada xotiraning segment-sahifali taqsimlash usulining sun’iy intellekt va mashinaviy o‘rganish, xususan konvolyutsion neyron tarmoqlar (CNN)da qo‘llanilishi yoritilgan. Maqolada CNN modellarining xotira talablari, GPU arxitekturasi, CUDA Unified Memory texnologiyasi hamda amaliy tajribalar asosida segmentatsiya va sahifalashning AI tizimlarida samaradorligi tahlil qilingan. Real loyihalardagi misollar orqali xotirani boshqarish usullari samarali optimallashtirish vositasi sifatida ko‘rsatilgan.*

Kalit so‘zlar: *Sun’iy intellekt, mashinaviy o‘rganish, konvolyutsion neyron tarmoqlar, segmentatsiya, sahifalash, GPU, CUDA Unified Memory, xotira boshqaruvi*

Kirish

Sun’iy intellekt (AI) va mashinaviy o‘rganish (ML) sohalaridagi konvolyutsion neyron tarmoqlar (CNN) yirik hajmli ma’lumotlar bilan ishlaydi va murakkab arxitekturalarni talab qiladi. Bu esa tizim xotirasiga bo‘lgan ehtiyojni sezilarli darajada oshiradi. Masalan, mashhur CNN modellaridan AlexNet va VGG(Visual Geometry Group) tarmoqlarini trening qilish uchun bir nechta GPU(Graphics Processing Unit) va ularning xotiralari ishlatilgan: AlexNet 3GB hajmli ikkita GPUda, VGG esa to‘rtta GPUda o‘qitilgan. Shuning uchun AI tizimlarida xotirani samarali boshqarish muhim ahamiyat kasb etadi. Operatsion tizimlarda xotirani bo‘linish (segmentation) va sahifalash (paging) kabi texnikalar yordamida boshqaradi. Bu usullar xotirani moslashuvchan bo‘laklarga ajratib, bo‘sh joydan oqilona foydalanishni ta’minlaydi.

Segment-sahifali xotira boshqaruvi

Operatsion tizimlarda xotira boshqaruvi uchun bo‘linish va sahifalash keng qo‘llaniladi. Bo‘linish xotirani segment deb ataluvchi mantiqiy bo‘laklarga ajratadi: har bir jarayon uchun kod segmenti, ma’lumot segmenti, stek segmenti kabi bo‘laklar yaratiladi. Har bir segmentga baza manzil va hajm ajratilib, uning ichidagi manzil ofseti bilan tizim xotirasi bog‘lanadi. Boshqa tomondan, sahifalash xotirani teng o‘lchamdagi sahifa bloklariga bo‘lib, sahifa jadvallari yordamida har bir sahifani fizik ramka bilan moslaydi. Bo‘linish va sahifalash birgalikdagi yondashuv segment-sahifali xotira boshqaruvi – jarayon manzil fazosini avval segmentlarga, keyin esa har bir segment ichidagi sahifalarga ajratishni nazarda tutadi . Bunday usul xotirani moslashuvchan taqsimlashni imkonini beradi (har bir segment o‘ziga xos hajmga ega bo‘ladi) va sahifalash orqali parchalanishni kamaytiradi, xotirani samarali ishlatishni yaxshilaydi.

CNN arxitekturalarida segment-sahifali xotira boshqaruvi

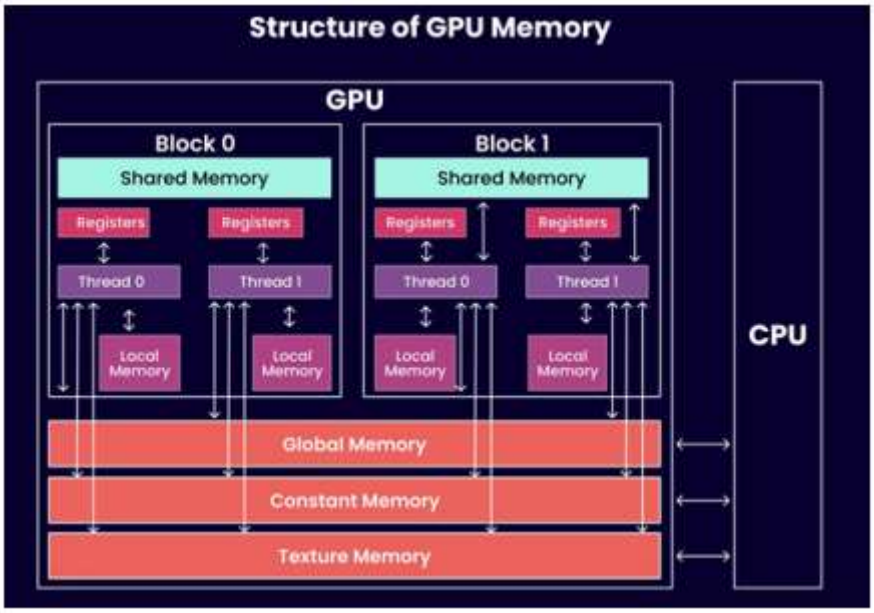
CNN modellarida ayniqsa katta o‘lchamli kirish tasvirlari bilan ishlaganda xotira talablari ortadi. Bunday holatda input ma’lumotlarini kichik parchalar (patch)larga bo‘lish – kirish segmentatsiyasi – qulay yechim bo‘lishi mumkin. Masalan, Lee va boshqa. 2024-yilda taqdim etgan usulda kirish tasviri “patch”larga bo‘linib, ularning chegaralari bir-biriga qisman kesishgan holda tarmoq boshida ko‘proq xotira talabini kamaytirilgan. Har bir patch alohida sub-tarmoqda qayta ishlanadi va oxirida natijalar birlashtiriladi, shunda har bir yo‘nalish xotira byudjeti ichida qoladi.

Bundan tashqari, kattaroq model yoki tasvirlar bilan ishlaganda grafik protsessor (GPU)ning o‘ziga xos xotirasi yetishmovchiligi masalasi tug‘iladi. Buni bartaraf etish uchun

NVIDIA kabi platformalarda birlashtirilgan xotira (CUDA Unified Memory) texnologiyasi qo‘llaniladi. CUDA Unified Memory GPU va CPU xotiralarini yagona manzil fazosida birlashtirib, agar GPU kerakli sahifaga ega bo‘lmasa, sahifa xatosi yuzaga keladi va tegishli ma’lumot CPU xotirasidan GPU xotirasiga ko‘chiriladi . Shu orqali GPU xotirasi “sun’iy” ravishda kengaytiriladi va yirik CNN modellari o‘qitilishi mumkin bo‘ladi. NVIDIA ma’lumotlariga ko‘ra, CUDA 6.0 dan boshlab birlashtirilgan xotira texnologiyasi orqalik GPU va CPU manzil fazolarini birlashtirish mumkinligi ta’kidlangan.

Resurslardan samarali foydalanish va GPU optimizatsiyasi

Neyron tarmoqlarni hisoblashda GPU resurslaridan oqilona foydalanish muhimdir. CUDA kabi muhitlarda GPUda bir nechta xotira turi mavjud: har bir ip uchun (thread) registrlar va mahalliy (local) xotira, ip-bloklar uchun shared xotira, umumiy foydalaniladigan global xotira, shuningdek constant va texture xotira modullari.



GPU arxitekturasi xotira tuzilishini ko‘rsatuvchi diagramma (har bir blok uchun registrlar, shared memory, global, constant, texture xotiralari). Misol uchun, texnik maqolalarga ko‘ra, global xotira barcha iplar tomonidan ishlatilishi mumkin va katta hajmli bo‘lsa, registrlar har bir ip uchun juda tezkor lekin kichik xotira bo‘lib xizmat qiladi. Shared memory esa bitta blokdagi iplar orasida parallel hisoblashlar uchun tezkor almashinuvni ta’minlaydi.

GPU xotirasini bo‘shatish va taqsimlash uchun ko‘plab optimallashtirish usullari mavjud. Masalan, Zhang va boshqa. (2019) GPU xotirasi uchun “smart pool” yondashuvini taklif qiladiki, u tarmoq o‘zgaruvchilarining hayot davrini hisobga olib, bir-biriga to‘qnashmaydigan o‘zgaruvchilarni bir xil xotira bo‘lagiga joylashtiradi. Bundan tashqari,

o‘rganish jarayonida hozirda faol bo‘lmagan o‘zgaruvchilarni vaqtincha CPU xotirasiga ko‘chirib olish (swapping) usuli GPU xotirasi yukini sezilarli kamaytirishi aniqlangan. Ular ko‘rsatishicha, to‘g‘ri almashish rejalari yordamida GPU xotirasi talabi 34.2% gacha pasaytirilishi mumkin . Shu bilan birga, Unified Memory kontekstida ma’lumotlarni turkumlash va ularning foydalanish naqshlariga mos ‘memory advise’ burchak so‘zlarini qo‘llash mumkin. Masalan, faqat o‘qiladigan ma’lumotlar uchun “Read Mostly”, ma’lumotni asosan ma’lum qurilmaga joylashtirish uchun “Preferred Location” ko‘rsatmalari sahifa xatolarini kamaytiradi.

Amaliy misollar va tadqiqotlar

Segment-sahifali xotira boshqaruvi va unga o‘xshash usullar real loyihalarda ham qo‘llanilmoqda. Lee va boshqa. (2024) taqdim etgan yondashuvda kichik xotirali mobil qurilmalar uchun input segmentation texnikasi qo‘llanilib, faqat 63 KB xotira bilan ImageNet sinflovida 61.58% aniqlikka erishilgan. Bu misol o‘z navbatida xotira cheklangan tizimlarda konvolyutsion tarmoqlarning arxitekturasini qayta tashkillashtirish orqali resurs talabini sezilarli kamaytirish mumkinligini ko‘rsatadi. Shuningdek, ko‘p yadroli akseleratorlar tizimlarida modelni segmentlarga bo‘lish xotira va hisoblash yukini taqsimlashga yordam beradi. Masalan, bitta tadqiqotda sakkiz ta Edge TPUli PLC tizimida model pipeline segmentlarga ajratilib, har bir TPUga vaznlar bo‘lingan holda ish bo‘lishi natijasida umumiy xotira yukini kamaytirganlik kuzatilgan. Robototexnika va kiritilgan tizimlarda ham optimallashtirish talablariga e’tibor qaratilmoqda: TinyCNN loyihasi FPGAda CNN tezlashtiradi, ammo ularda ham cheklangan xotira hajmi sababli unumdorlik masalalari tug‘ilmoqda. Ushbu misollar xotira boshqaruvi yondashuvlarining tasvirni aniqlash, video tahlil, robototexnika kabi real vazifalarda ijobiy ta’sirini ko‘rsatadi.

Xulosa

Xotraning segment-sahifali taqsimlash usullari xotirani samarali boshqarishni ta’minlab, sezilarli afzalliklarni beradi. Bu yondashuvlar xotira bo‘shlig‘ini yaxshilaydi, parchalanishni kamaytiradi va moslashuvchan taqsimlash imkonini yaratadi. Ayniqsa, GPU arxitekturasi va CUDA Unified Memory kabi texnologiyalar yordamida AI sistemalari o‘zlarining cheklangan GPU xotirasini kengaytirib, katta hajmdagi ma’lumot bilan ishlash imkoniyatini oshirmoqda. Misol uchun, yuqorida ko‘rib o‘tilgan tadqiqotlarda bir nechta optimallashtirish orqali neyron tarmoq xotira talabini yirik hajmdan kichik RAM darajasigacha kamaytirish namoyon etildi. Kelajakda ko‘p GPUli tizimlar, ilg‘or virtual xotira usullari va apparat yordami xotira boshqaruvini yanada takomillashtirishi kutilmoqda. Umuman olganda, segment-sahifali xotira boshqaruvi AI va mashinaviy o‘rganish sohasida chuqur tadqiq va amaliy qo‘llanilishini davom ettiradi, bu esa texnologiyaning yanada keng imkoniyatlarini ochadi.

Foydalanilgan adabiyotlar:

1. GeeksforGeeks. (2022). *Paged Segmentation and Segmented Paging*. https://www.geeksforgeeks.org/paged-segmentation-and-segmented-paging/

2. NVIDIA Developer Blog. (2021). *Unified Memory for CUDA Beginners*. https://developer.nvidia.com/blog/unified-memory-cuda-beginners

3. NVIDIA Developer Blog. (2022). *Maximizing Unified Memory Performance in CUDA*. https://developer.nvidia.com/blog/maximizing-unified-memory-performance-cuda

4. arXiv.org. (2024). *Memory-Efficient Training of CNNs with Input Segmentation*. https://arxiv.org/abs/2408.03663

5. Madumarov U. A., Abdurahmonov A. O. (2019). Operatsion tizimlar. – Toshkent: “Fan va texnologiya” nashriyoti.

6. Abdullayev X. M., Karimov A. A. (2020). Sun’iy intellekt va mashinaviy o‘qitish. – Toshkent: “Innovatsiya” nashriyoti.

7. Qurbonov B. Q., Eshqobilov A. T. (2021). Parallel hisoblash tizimlari. – Samarqand: SamDU nashriyoti.

8. Tutorialspoint. (2023). *Memory Management in Operating Systems*. https://www.tutorialspoint.com/operating_system/os_memory_management.htm

9. NVIDIA CUDA Toolkit Documentation. (2024). *Unified Memory Programming Guide*. https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#um-overview

10.O‘zbekiston Respublikasi Oliy va o‘rta maxsus ta’lim vazirligi. (2018). Kompyuter grafikasi va dasturlash: o‘quv qo‘llanma. – Toshkent: “Yangi asr avlodi”.

11.Umarov, B., G'ulomjonova, S. (2024). BULUT TEXNOLOGIYASI VA ULARDAN FOYDALANISH.