

Pytorch-ის ბიბლიოთეკისთვის იშვიათი მატრიცების ოპერაციების იმპლემენტაცია Apple Silicon-ის GPU-ებზე Metal Performance Shader-ების გამოყენებით

ირაკლი სალია

სამაგისტრო ნაშრომი წარდგენილია ილიას სახელმწიფო უნივერსიტეტის ბიზნესის,
ტექნოლოგიისა და განათლების ფაკულტეტზე პროგრამული ინჟინერიის
მაგისტრის აკადემიური ხარისხის მინიჭების მოთხოვნების შესაბამისად

პროგრამული ინჟინერია

ნაშრომის ხელმძღვანელი: შოთა ცისკარიძე, ილიას სახელმწიფო უნივერსიტეტის
ასოცირებული პროფესორი

პროექტის ხელმძღვანელი: ლელა მირცხულავა



**ილიას სახელმწიფო უნივერსიტეტი
თბილისი, 2026**

განაცხადი

როგორც წარდგენილი სამაგისტრო ნაშრომის ავტორი, ვაცხადებ, რომ ნაშრომი ჩემი ორიგინალური ნამუშევარია და არ შეიცავს სხვა ავტორების მიერ აქამდე გამოქვეყნებულ, გამოსაქვეყნებლად მიღებულ ან დასაცავად წარდგენილ მასალებს, რომლებიც არ არის მოხსენიებული ან ციტირებული სათანადო წესების შესაბამისად.

ირაკლი სალია

19 ივნისი 2026

მადლობა

განსაკუთრებული მადლობა მინდა გადავუხადო ჩემი სამაგისტრო ნაშრომის ხელმძღვანელს, ბატონ შოთა ცისკარიძეს, კვლევის მთელი პერიოდის განმავლობაში გაწეული თავდადებადი ხელმძღვანელობისა და მხარდაჭერისთვის. მიუხედავად მისი დატვირთული გრაფიკისა, ყოველკვირეულად, ხშირად გვიან საღამოს საათებზე, ახერხებდა ჩემთან სამუშაოდ დროის გამონახვას, ონლაინ თუ ადგილზე შეხვედრებისას. ნაშრომზე მუშაობის პროცესში წარმოშობილ ჩემს არაერთ კითხვასა და ეჭვს გულისხმიერებითა და მოთმინებით უპასუხა, ხოლო რჩევებისა და მიმართულების მოცემის გარდა, გამიზიარა საკუთარი გამოცდილება და მიდგომები, რამაც მნიშვნელოვანი გავლენა მოახდინა როგორც მიღებულ ტექნიკურ გადაწყვეტილებებზე, ისე ნაშრომის საბოლოო ხარისხზე. მისი ხელშეწყობისა და ერთგული მონდომების გარეშე ეს ნაშრომი წარმოდგენილი ფორმით ვერ ჩამოყალიბდებოდა.

ასევე მადლობა მინდა გადავუხადო პროექტის ხელმძღვანელს, ქალბატონ ლელა მირცხულავას, პროექტის ტექნიკური მოთხოვნების ჩამოყალიბებასა და ტესტირებისთვის.

რთულია მადლობის წერილში არ მოვიხსენიო ის ლექტორები, რომლებთანაც ამ წლების განმავლობაში მქონდა სწავლის საშუალება. თითოეულმა მათგანმა გადმომცა საკუთარი ცოდნა და გამოცდილება, რამაც განამტკიცა ჩემი თეორიული საფუძვლები და მნიშვნელოვანი გავლენა იქონია ჩემს პროფესიულ ზრდასა და განვითარებაზე.

და ბოლოს, მადლობა მინდა გადავუხადო ჩემს ოჯახს, ნაშრომზე მუშაობის პერიოდში გაწეული მუდმივი მორალური მხარდაჭერისა და გულშემატკივრობისთვის.

ირაკლი სალია

აბსტრაქტი

იშვიათი მატრიცები წარმოადგენს ფუნდამენტურ მონაცემთა სტრუქტურას სამეცნიერო გამოთვლებში, გრაფულ ნეირონულ ქსელებსა და თანამედროვე მანქანური სწავლების მოდელებში. PyTorch ბიბლიოთეკა, როგორც ერთ-ერთი წამყვანი მანქანური სწავლების ბიბლიოთეკა, უზრუნველყოფს იშვიათ ტენზორებზე ოპერაციების მხარდაჭერას CPU-სა და NVIDIA GPU-ებზე CUDA-ს მეშვეობით, თუმცა Apple Silicon-ის Metal Performance Shaders (MPS) ბექენდზე აღნიშნული ფუნქციონალი არ იყო ხელმისაწვდომი, მიუხედავად პლატფორმის მზარდი პოპულარობისა და უნიფიცირებული მეხსიერების არქიტექტურის (UMA) უპირატესობებისა.

წინამდებარე ნაშრომში წარმოდგენილია იშვიათ ტენზორებზე ოპერაციების სრულფასოვანი იმპლემენტაცია PyTorch ბიბლიოთეკაში Apple Silicon-ისთვის MPS ბექენდის მეშვეობით. იმპლემენტაცია მოიცავს 161 ოპერატორის ვარიანტს: მატრიცულ ოპერაციებს (SpMM, SpGEMM, bmm, mv), ელემენტურ და უნარულ ოპერაციებს, რედუქციებსა და ნორმებს, softmax-ს backward pass-ებით, ასევე ინდექსირებისა და სტრუქტურული დამუშავების უტილიტებს. არსებული CUDA-ზე ორიენტირებული ოპერატორული დიზაინები ადაპტირებულია Metal Compute Shader-ებზე და ინტეგრირებულია PyTorch-ის ბიბლიოთეკის დისპეტჩერისა და ოპერატორების სტეკში.

იმპლემენტაციის ვალიდაციისა და პრაქტიკული გამოყენებადობის შესაფასებლად ჩატარებულია წარმადობის ანალიზი, რომელიც ფოკუსირებულია იშვიათ-მკვრივ მატრიცულ გამრავლებაზე (SpMM). MPS ბექენდზე დაფიქსირდა $16\times$ -ზე მეტი აჩქარება მკვრივ გამრავლებასთან შედარებით დიდი მატრიცებისთვის ($n = 16384$) დაბალ იშვიათობის დონეებზე (0.1% NNZ). შედარებითი ანალიზი NVIDIA H100 GPU-ზე CUDA-ზე დაფუძნებულ იმპლემენტაციებთან ადასტურებს თვისებრივად მსგავს მასშტაბირების ქცევას, crossover წერტილებით, რომლებიც ერთდროულად არის განსაზღვრული მატრიცის ზომითა და იშვიათობის დონით.

ძირითადი საძიებო სიტყვები: იშვიათი მატრიცები, PyTorch ბიბლიოთეკა, Apple Silicon, MPS, CUDA, იშვიათ-მკვრივი მატრიცული გამრავლება, SpMM, წარმადობის ანალიზი, GPU გამოთვლა.

Abstract

Sparse matrices represent a fundamental data structure in scientific computing, graph analytics, and modern machine learning pipelines, arising naturally in Graph Neural Networks (GNNs), recommendation systems, and finite element methods. PyTorch library provides sparse tensor support across multiple backends, including CPUs and NVIDIA GPUs via CUDA, but GPU acceleration on Apple Silicon through the Metal Performance Shaders (MPS) backend lacked sparse tensor operation support — a significant limitation given Apple Silicon’s widespread adoption and its unified memory architecture (UMA).

This thesis addresses this gap by implementing a comprehensive suite of 161 sparse tensor operator variants in PyTorch library for Apple Silicon via the MPS backend, covering matrix operations (SpMM, SpGEMM, bmm, mv), elementwise and unary operations, reductions, softmax with backward passes, and indexing utilities. Existing CUDA-oriented operator designs are ported to the Metal/MPS execution model through custom Metal Compute Shaders and integrated into PyTorch library’s dispatcher and operator stack.

Performance evaluation on SpMM demonstrates that the MPS implementation achieves speedups exceeding $16\times$ over dense multiplication for large matrices ($n = 16384$) at low sparsity levels (0.1% NNZ). Comparative analysis with CUDA-based implementations on NVIDIA H100 confirms qualitatively similar scaling behavior, with crossover points determined jointly by matrix size and sparsity level.

Keywords: Sparse Matrices, PyTorch library, Apple Silicon, MPS, CUDA, Sparse-Dense Matrix Multiplication, SpMM, Performance Analysis, GPU Computing.