



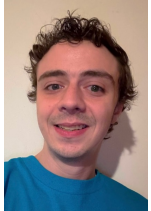
Welcome
to
CSE 331

Let's do some introductions



http://www.zazzle.com/warning_teaching_assistant_bag-149882665435161818

TAs first



Andrew



Vincent



Daniel



Junyan



Mary



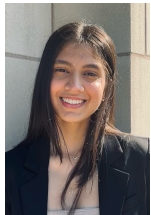
Rico



Wei



Jack



Shreya



Mingi



Vipassana



Rachel



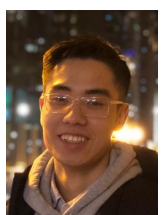
John



ester



Arnav



Mark

About Me

Atri Rudra



Just Atri please!

atri@buffalo.edu

Office: See syllabus for location

Office hours: Mon, Fri, 1:00-1:50pm

OH starts today in Clemens 17

Contact us all at



Or use piazza!

cse-331-staff@buffalo.edu

TAs will not respond to individual emails (except for re-grading requests)

Questions/Comments?




Handouts for today

Syllabus (online)

Syllabus Walkthrough video (online)

Homework Policy document (online)


Read the syllabus CAREFULLY!

CSE 331 Syllabus Piazza Schedule Homeworks ▾ Autolab Project ▾ Support Pages ▾  channel Sample Exams ▾

CSE 331 Syllabus

Algorithms and Complexity

Fall 2024

Time and location: **Mondays, Wednesdays** and **Fridays, 11:00-11:50am**, [KNOX](#)  104.



Under Construction

This page is still under construction. In particular, nothing here is final while this sign still remains here.

Please note

It is **your responsibility** to make sure you read and understand the contents of this syllabus. If you have any questions, please contact the instructor.

Acknowledgment

Once you have read the syllabus carefully, please fill in the Syllabus quiz on [Autolab](#). As an incentive for you to fill in this form, **you will not receive any feedback on your assignments till you successfully answer AT LEAST 18 out of the 20 questions in the quiz.** (You can attempt the quiz as many times as you want.) Note that in addition to this syllabus, the quiz will also ask questions based on the [homework policies](#).

In spirit of trust but verify

Syllabus Quiz

Options

[View handin history](#)



Due: December 10th 2024, 11:59 pm EST (UTC -05:00)



Last day to hand in: December 10th 2024, 11:59 pm EST (UTC -05:00)



No graded material will be handed back until you pass the syllabus quiz!

Academic Integrity

Question 1: Sharing my answers to this syllabus quiz with other 331 students

- Is OK if I do it to help out a friend
- It does not matter since there is no grade attached with it
- Is an academic integrity violation and should not be done
- Is an academic integrity violation but I can take the chance


If you registered by 5pm on Sunday, Aug 25 you will be on Autolab + piazza

Accessibility Resources

Information included in the syllabus

In short, let me know and consult with Accessibility Resources

One Stop Shop for the Course

CSE 331 Syllabus Piazza Schedule Homeworks ▾ Autolab Project ▾ Support Pages ▾  channel Sample Exams ▾

CSE 331

Fall 2024

<http://www-student.cse.buffalo.edu/~atri/cse331/fall24/index.html>

Under Construction

This page is still under construction. In particular, nothing here is final while this sign still remains here.

CSE 331 events

Today   Aug 18 – 24, 2024 ▾

 Print **Week** Month Agenda ▾

| | Sun 8/18 | Mon 8/19 | Tue 8/20 | Wed 8/21 | Thu 8/22 | Fri 8/23 | Sat 8/24 |
|------|----------|----------|--|----------|----------|----------|----------|
| 5am | | | | | | | |
| 6am | | |  | | | | |
| 7am | | | | | | | |
| 8am | | | | | | | |
| 9am | | | | | | | |
| 10am | | | | | | | |
| 11am | | | | | | | |

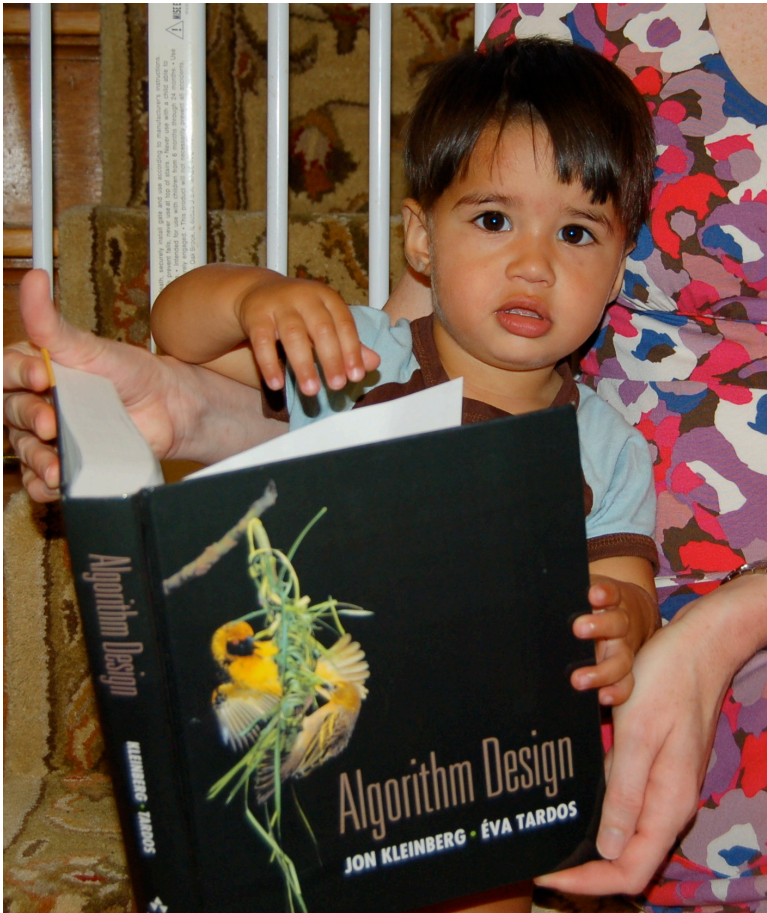
Three things to remember

WORK HARD!

DO NOT CHEAT!

READ CAREFULLY!

Wait.. What???



Make sure you follow submission instructions

Two most common ways
of losing points

Make sure you read problem statements carefully

Academic Dishonesty

All your submissions must be your own work

Penalty:

Minimum: A **grade reduction in course**

Possible: **F** (or higher penalty) if warranted

YOUR responsibility to know what is cheating, plagiarism etc.

If not sure, come talk to me

Excuses like “I have a job,” “This was OK earlier/in my country,” “This course is hard,” etc. **WON’ T WORK**

I DO NOT HAVE ANY PATIENCE WITH ANY CHEATING :
YOU WILL GET A GRADE REDUCTION IN THE COURSE
FOR YOUR FIRST MISTAKE

Withdrawing a suspect submission

Withdrawing a submission for academic integrity violation

Sometime mistakes can happen so you have the **option of withdrawing any of your Homework submission with 24 HOURS of the assignment deadline**. You can do this by sending Atri an email, e.g. by using the following template (thanks to [Oliver Kennedy](#) for providing us the template):

Email template for withdrawing submission

Dear Dr. Rudra/Atri,

I wish to inform you that I have violated CSE 331 policies on my submission for Question X on Homeworks/Assignment N. I wish to withdraw my submission to preserve academic integrity.

J.Q. Student
Person #12345678
UBIT: jqstuden

Sincerely, J

On receiving the above email, I will assign J a 0 on Question X on Homeworks/Assignment N but disregard any Academic Integrity issues with the problematic submission. Note that J is not required to present any details on how they violated academic integrity.

- - - - -

Use of ChatGPT is not allowed

AT

Why should I not use you to do my algorithms course homework?



While I can certainly help you understand concepts and provide explanations related to your algorithms course, there are several reasons why you shouldn't solely rely on me to do your homework:

1. **Academic Integrity:** Using external sources to complete your assignments without understanding the material yourself violates the principles of academic integrity. It's important to learn and demonstrate your knowledge honestly.
2. **Learning:** The primary goal of your course is to learn and understand the material. Relying solely on my answers without engaging with the material will hinder your learning and your ability to apply the concepts in real-world scenarios.
3. **Long-Term Retention:** Active engagement with the material, including struggling through problems and solving them on your own, enhances your understanding and retention of the content. Passive reading or copying solutions does not provide the same benefits.
4. **Exams and Practical Application:** If your course includes exams or projects that test your problem-solving skills, simply copying answers won't prepare you for those assessments. You need to be able to apply your knowledge independently.
5. **Conceptual Understanding:** Algorithms often involve complex concepts. Simply having an answer isn't enough; you need to understand why a particular approach works and how it relates to the broader context of the subject.

6. **Feedback and Improvement:** Making mistakes is an essential part of learning. When you attempt problems on your own, you can receive feedback from instructors or peers, which helps you identify areas for improvement.
7. **Career Preparedness:** Algorithms and problem-solving skills are often crucial in technical interviews and real-world applications. If you haven't practiced solving problems on your own, you might struggle in job interviews or when faced with practical challenges.
8. **Uniqueness of Assignments:** Homework assignments can vary in their requirements. If you rely on pre-generated solutions, you might miss out on tailoring your answers to the specific instructions of your assignment.
9. **Ethical Considerations:** While seeking help is acceptable, outsourcing the entire problem-solving process to an AI raises ethical concerns, especially if it's done to a degree that undermines the learning process.

It's okay to seek help and clarification on concepts, but make sure you're actively engaging with the material and attempting problems on your own. This will not only benefit your learning but also your overall academic and professional growth.

In spirit of trust but verify

TAs will have access to a database of ChatGPT answers to cross check

It's no fun failing any of you but **use of ChatGPT will result in an F**

Use of ChatGPT is not allowed



Why should I not use you to do my algorithms course homework?



While I can certainly help you understand concepts and provide explanations related to your algorithms course, there are several reasons why you shouldn't solely rely on me to do your homework:

1. **Academic Integrity:** Using external sources to complete your assignments without understanding the material yourself violates the principles of academic integrity. It's important to learn and demonstrate your knowledge honestly.
2. **Learning:** The primary goal of your course is to learn and understand the material. Relying solely on my answers without engaging with the material will hinder your learning and your ability to apply the concepts in real-world scenarios.
3. **Long-Term Retention:** Active engagement with the material, including struggling through problems and solving them on your own, enhances your understanding and retention of the content. Passive reading or copying solutions does not provide the same benefits.
4. **Exams and Practical Application:** If your course includes exams or projects that test your problem-solving skills, simply copying answers won't prepare you for those assessments. You need to be able to apply your knowledge independently.
5. **Conceptual Understanding:** Algorithms often involve complex concepts. Simply having an answer isn't enough; you need to understand why a particular approach works and how it relates to the broader context of the subject.

6. **Feedback and Improvement:** Making mistakes is an essential part of learning. When you attempt problems on your own, you can receive feedback from instructors or peers, which helps you identify areas for improvement.
7. **Career Preparedness:** Algorithms and problem-solving skills are often crucial in technical interviews and real-world applications. If you haven't practiced solving problems on your own, you might struggle in job interviews or when faced with practical challenges.
8. **Uniqueness of Assignments:** Homework assignments can vary in their requirements. If you rely on pre-generated solutions, you might miss out on tailoring your answers to the specific instructions of your assignment.
9. **Ethical Considerations:** While seeking help is acceptable, outsourcing the entire problem-solving process to an AI raises ethical concerns, especially if it's done to a degree that undermines the learning process.

It's okay to seek help and clarification on concepts, but make sure you're actively engaging with the material and attempting problems on your own. This will not only benefit your learning but also your overall academic and professional growth.

Coding jobs will be done by AI



stacksort

In a [recent xkcd's](#) alt text, Randall Munroe suggested **stacksort**, a sort that searches StackOverflow for sorting functions and runs them until it returns the correct answer. So, I made it. If you like running arbitrary code in your browser, try it out.

Like (or hate) it? Comment on HackerNews

stackoverflow_sort(
Try a list of numbers, a string, a
list of words or json.

);

Sort

var output =
Output from the function.

;

output console

Coding jobs will be done by AI

MIT News

ON CAMPUS AND AROUND THE WORLD

Browse

or

Search



 FULL SCREEN

Researchers have developed a flexible way of combining deep learning and symbolic reasoning to teach computers to write short computer programs. Here, Armando Solar-Lezama (left), a professor at CSAIL, speaks with graduate student Maxwell Nye.

Photo: Kim Martineau

Toward artificial intelligence that learns to write code

Researchers combine deep learning and symbolic reasoning for a more flexible way

It happened sooner than expected

Support the Guardian

Fund independent journalism with \$5 per month

Support us →

The Guardian

News Opinion Sport Culture Lifestyle More ▾

The Guardian view Columnists Letters Opinion videos Cartoons

The Observer

ChatGPT

Sat 1 Apr 2023 11.00 EDT

934

• This article is more than 4 months old

Programmers, beware: ChatGPT has ruined your magic trick

John Naughton



The generative AI tool can write code on request, making the specialist skill of programming open to everyone



Advertisement




Complete a HouseCalls visit and earn up to a **\$15 reward**

Learn more

Must be enrolled in a qualifying UnitedHealthcare plan. HouseCalls may not be available in all areas. Rewards restrictions apply.

So am I doomed?

There will still be room for high level *algorithmic* thinking!



European Association for
Theoretical Computer Science

HOME ABOUT SEARCH CURRENT ARCHIVES

Home > No 125: June 2018 > **Lamport**

If You're Not Writing a Program, Don't Use a Programming Language
Leslie Lamport, Distributed Computing & Education Column by Juraj Hromkovic, Stefan Schmid

*Today, programming is generally equated with coding. It's hard to convince students who want to write code that they should learn to think mathematically, above the code level, about what they're doing. Perhaps the following observation will give them pause. **It's quite likely that during their lifetime, machine learning will completely change the nature of programming. The programming languages they are now using will seem as quaint as Cobol, and the coding skills they are learning will be of little use. But mathematics will remain the queen of science, and the ability to think mathematically will always be useful.***

Questions/Comments?



Any question on course policies?

note @9

stop following

1 view

Actions

Homework for BEFORE the Monday lecture

To optimize the first two lectures I have put everything that I used to talk about the syllabus (in FA 22 and earlier) into a video:

- [Walk through video](#)

Please review the above video before you come to lecture on Monday.

I will have a short Q&A for any questions y'all might have on the syllabus/course policies during the lecture on Monday but my expectation is that unless you ask a question, y'all have understood the syllabus and homework policies.

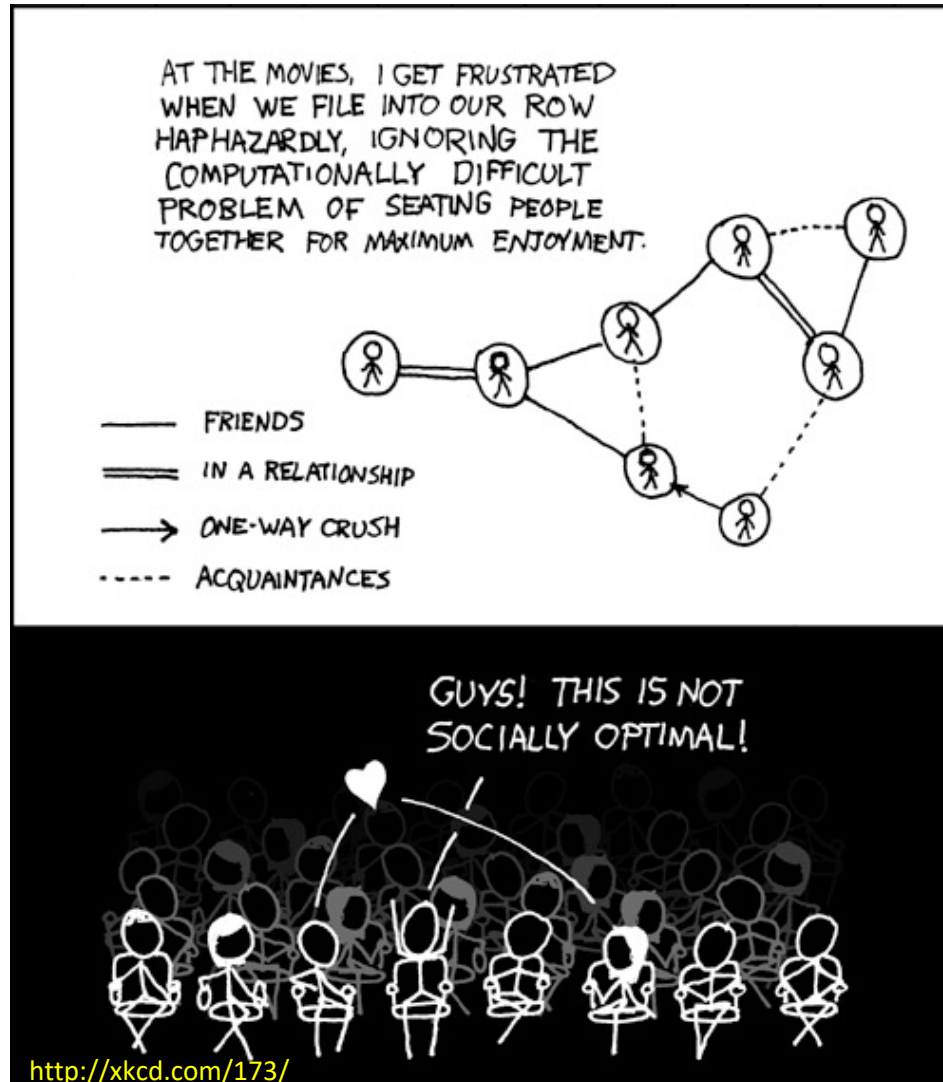
The [schedule page](#) also has links to the above videos as well as link to the slides (see the Notes column) in case that is of interest.

lectures

Edit good note | 0

Updated 41 seconds ago by Atri Rudra

This course: how to solve problems!



Why should I care ?



I know twitter ain't that cool any more but..

I'll start off with a tweet...



<https://twitter.com/sytelus/status/1644546251302621184>



Tri



Dan

What is FlashAttention?

FLASHATTENTION: Fast and Memory-Efficient Exact Attention with IO-Awareness

Tri Dao[†], Daniel Y. Fu[†], Stefano Ermon[†], Atri Rudra[‡], and Christopher Ré[†]

[†]Department of Computer Science, Stanford University

[‡]Department of Computer Science and Engineering, University at Buffalo, SUNY

{trid,danfu}@cs.stanford.edu, ermon@stanford.edu, atri@buffalo.edu, chrisrmre@cs.stanford.edu

June 24, 2022

Abstract

Transformers are slow and memory-hungry on long sequences, since the time and memory complexity of self-attention are quadratic in sequence length. Approximate attention methods have attempted to address this problem by trading off model quality to reduce the compute complexity, but often do not achieve wall-clock speedup. We argue that a missing principle is making attention algorithms *IO-aware*—accounting for reads and writes between levels of GPU memory. We propose FLASHATTENTION, an IO-aware exact attention algorithm that uses tiling to reduce the number of memory reads/writes between GPU high bandwidth memory (HBM) and GPU on-chip SRAM. We analyze the IO complexity of FLASHATTENTION, showing that it requires fewer HBM accesses than standard attention, and is optimal for a range of SRAM sizes. We also extend FLASHATTENTION to block-sparse attention, yielding an approximate attention algorithm that is faster than any existing approximate attention method. FLASHATTENTION trains Transformers faster than existing baselines: 15% end-to-end wall-clock speedup on BERT-large (seq. length 512) compared to the MLPerf 1.1 training speed record, 3× speedup on GPT-2 (seq. length 1K), and 2.4× speedup on long-range arena (seq. length 1K-4K). FLASHATTENTION and block-sparse FLASHATTENTION enable longer context in Transformers, yielding higher quality models (0.7 better perplexity on GPT-2 and 6.4 points of lift on long-document classification) and entirely new capabilities: the first Transformers to achieve better-than-chance performance on the Path-X challenge (seq. length 16K, 61.4% accuracy) and Path-256 (seq. length 64K, 63.1% accuracy).

The screenshot shows the GitHub repository for 'flash-attention'. The repository is public and has 88 watchers, 552 forks, and 6.6k stars. It contains 37 tags and 3 branches. The README section is visible, titled 'FlashAttention', and describes the official implementation of FlashAttention and FlashAttention-2. It includes a diagram of the memory hierarchy and a bar chart comparing attention methods on GPT-2.

FlashAttention

This repository provides the official implementation of FlashAttention and FlashAttention-2 from the following papers.

FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness
Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, Christopher Ré
Paper: <https://arxiv.org/abs/2205.14135>
IEEE Spectrum [article](#) about our submission to the MLPerf 2.0 benchmark using FlashAttention.

Memory Hierarchy with Bandwidth & Memory Size

| Memory Type | Bandwidth | Size |
|------------------------|-----------|-------|
| GPU SRAM | 19 TB/s | 20 MB |
| HBM | 1.5 TB/s | 40 GB |
| Main Memory (CPU DRAM) | 12.8 GB/s | >1 TB |

Attention on GPT-2

| Method | Time (ms) |
|----------------|-----------|
| PyTorch | ~15 |
| FlashAttention | ~5 |

<https://github.com/Dao-AILab/flash-attention>

OK, but what does FlashAttention do?

This over simplifying things and then some more but....



You

Write a 50 word essay on FlashAttention



ChatGPT

FlashAttention is a contemporary phenomenon reflecting our culture's decreasing attention spans. In the era of rapid information exchange, individuals are bombarded with stimuli, leading to shortened focus and heightened distractibility. This trend poses challenges to deep thinking and sustained engagement, underscoring the need for mindful strategies in the digital age.

FlashAttention helps ChatGPT (or other similar models) write longer essays!

FlashAttention is used everywhere

FlashAttention adoption

We've been very happy to see FlashAttention being adopted by many organizations and research labs to speed up their training / inference (within 6 months after FlashAttention's release, at the time of writing). This page contains a partial list of places where FlashAttention is being used. If you'd like to add links to your organization / product / codebase, please open a PR or email us. We'd very much like to hear from you!

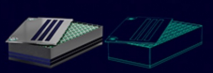
Integrated into machine learning frameworks

- Pytorch: [integrated](#) into core Pytorch in `nn.Transformer`.
- Huggingface's [transformers](#) library. [On-going](#), blogpost coming soon.
- Microsoft's [DeepSpeed](#): FlashAttention is [integrated](#) into DeepSpeed's inference engine.
- Nvidia's [Megatron-LM](#). This library is a popular framework on training large transformer language models at scale.
- MosaicML [Composer library](#). Composer is a library for efficient neural network training.
- EleutherAI's [GPT-NeoX](#). This is a research library for training large language transformer models at scale based on NVIDIA's Megatron-LM and Microsoft's DeepSpeed.
- PaddlePaddle: integrated into the framework with [API](#) `paddle.nn.functional.flash_attention` .

1 Ph.D. student > Multiple Ph.D.s


IEEE Spectrum FOR THE TECHNOLOGY INSIDER

ADVERTISEMENT

SIEMENS Harness the power of digitalization and optimize cell production before it begins  See h

NEWS | ARTIFICIAL INTELLIGENCE

We're Training AI Twice as Fast This Year as Last > New MLPerf rankings show training times plunging

BY SAMUEL K. MOORE | 30 JUN 2022 | 5 MIN READ | 

Samuel K. Moore is IEEE Spectrum's semiconductor editor.

Performing MLPerf benchmarks is no easy task, and often involves the work of many engineers. But a single graduate student, with some consultation, can do it, too. Tri Dao was that graduate student. He's member of **Hazy Research**, the nom de guerre of Chris Re's laboratory at Stanford. (Re is one of the founders of AI giant SambaNova.) Dao, Re, and other colleagues came up with a way to speed up the training of so-called attention-based networks, also called transformer networks.

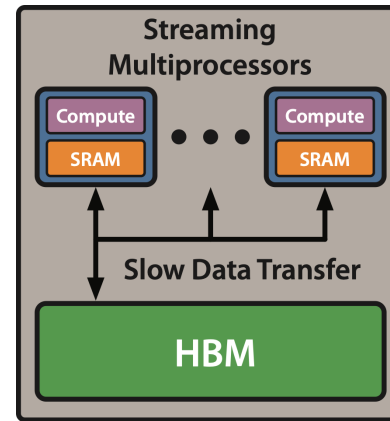
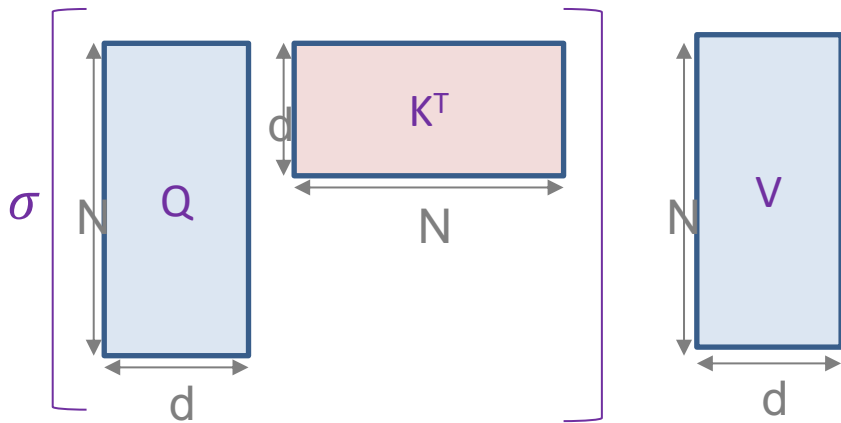
Here's the entire algorithm

Algorithm 1 FLASHATTENTION

Require: Matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d}$ in HBM, on-chip SRAM of size M .

- 1: Set block sizes $B_c = \lceil \frac{M}{4d} \rceil$, $B_r = \min(\lceil \frac{M}{4d} \rceil, d)$.
 - 2: Initialize $\mathbf{O} = (0)_{N \times d} \in \mathbb{R}^{N \times d}$, $\ell = (0)_N \in \mathbb{R}^N$, $m = (-\infty)_N \in \mathbb{R}^N$ in HBM.
 - 3: Divide \mathbf{Q} into $T_r = \lceil \frac{N}{B_r} \rceil$ blocks $\mathbf{Q}_1, \dots, \mathbf{Q}_{T_r}$ of size $B_r \times d$ each, and divide \mathbf{K}, \mathbf{V} into $T_c = \lceil \frac{N}{B_c} \rceil$ blocks $\mathbf{K}_1, \dots, \mathbf{K}_{T_c}$ and $\mathbf{V}_1, \dots, \mathbf{V}_{T_c}$, of size $B_c \times d$ each.
 - 4: Divide \mathbf{O} into T_r blocks $\mathbf{O}_1, \dots, \mathbf{O}_{T_r}$ of size $B_r \times d$ each, divide ℓ into T_r blocks $\ell_1, \dots, \ell_{T_r}$ of size B_r each, divide m into T_r blocks m_1, \dots, m_{T_r} of size B_r each.
 - 5: **for** $1 \leq j \leq T_c$ **do**
 - 6: Load $\mathbf{K}_j, \mathbf{V}_j$ from HBM to on-chip SRAM.
 - 7: **for** $1 \leq i \leq T_r$ **do**
 - 8: Load $\mathbf{Q}_i, \mathbf{O}_i, \ell_i, m_i$ from HBM to on-chip SRAM.
 - 9: On chip, compute $\mathbf{S}_{ij} = \mathbf{Q}_i \mathbf{K}_j^T \in \mathbb{R}^{B_r \times B_c}$.
 - 10: On chip, compute $\tilde{m}_{ij} = \text{rowmax}(\mathbf{S}_{ij}) \in \mathbb{R}^{B_r}$, $\tilde{\mathbf{P}}_{ij} = \exp(\mathbf{S}_{ij} - \tilde{m}_{ij}) \in \mathbb{R}^{B_r \times B_c}$ (pointwise), $\tilde{\ell}_{ij} = \text{rowsum}(\tilde{\mathbf{P}}_{ij}) \in \mathbb{R}^{B_r}$.
 - 11: On chip, compute $m_i^{\text{new}} = \max(m_i, \tilde{m}_{ij}) \in \mathbb{R}^{B_r}$, $\ell_i^{\text{new}} = e^{m_i - m_i^{\text{new}}} \ell_i + e^{\tilde{m}_{ij} - m_i^{\text{new}}} \tilde{\ell}_{ij} \in \mathbb{R}^{B_r}$.
 - 12: Write $\mathbf{O}_i \leftarrow \text{diag}(\ell_i^{\text{new}})^{-1} (\text{diag}(\ell_i) e^{m_i - m_i^{\text{new}}} \mathbf{O}_i + e^{\tilde{m}_{ij} - m_i^{\text{new}}} \tilde{\mathbf{P}}_{ij} \mathbf{V}_j)$ to HBM.
 - 13: Write $\ell_i \leftarrow \ell_i^{\text{new}}$, $m_i \leftarrow m_i^{\text{new}}$ to HBM.
 - 14: **end for**
 - 15: **end for**
 - 16: Return \mathbf{O} .
-

Flash Attention [Dao, Fu et al. 2022]



Tri



Dan

Want to avoid accessing slow memory

Better algorithm with little hacking will beat a worse algorithm with tons of hacking

With apologies to

Flash Attention avoids unnecessary writes to slow memory!

Questions/Comments?



Few words on math/proofs

note @10

stop following 28 views

Actions

Three math/proof books

In CSE 331, we will be using a lot of math and specifically proofs. While there are [support pages](#) and [care packages](#) [with more forthcoming for the latter] for y'all to brush up on your proofs, you are expected to do any catch up as needed on your own (though we are always happy to help!)

What we do not get much time covering (either in the above resources or lectures/recitations) is on how one goes about writing proofs/doing math-- more so the human aspects of it. Below, I'll list three books that might be helpful in this regard (they range from the more philosophical to the more concrete):

- [Mathematics for Human Flourishing](#) by [Francis Su](#), which is available online via UB libraries!
 - This book is the best articulation that I have seen of the reasons for why I *personally* love math. When I read the book, it was like someone was literally writing down my thoughts (but I would never have been able to write it so well myself).
 - I highly recommend that you read through the entire book but if you do not have the time, I would recommend at least reading through Chapters 1, 8, 9, 12 and 13.
- [Math Girls](#) by [Hiroshi Yuki](#).
 - I picked up this series of books during my recent trip to Japan.
 - What I like about this book series is that it walks you through the process of doing math.
 - Start off with Volume 1 to see if this is your cup of tea (the rest of my family were rolling their eyes when I bought the books!).
- [How to Prove It](#) by [Daniel J. Velleman](#)
 - This is a no frills book that just gets you down to doing proofs.
 - We have mentioned this book in the [CSE 331 HW policy document](#)

Hope these resources help!

society

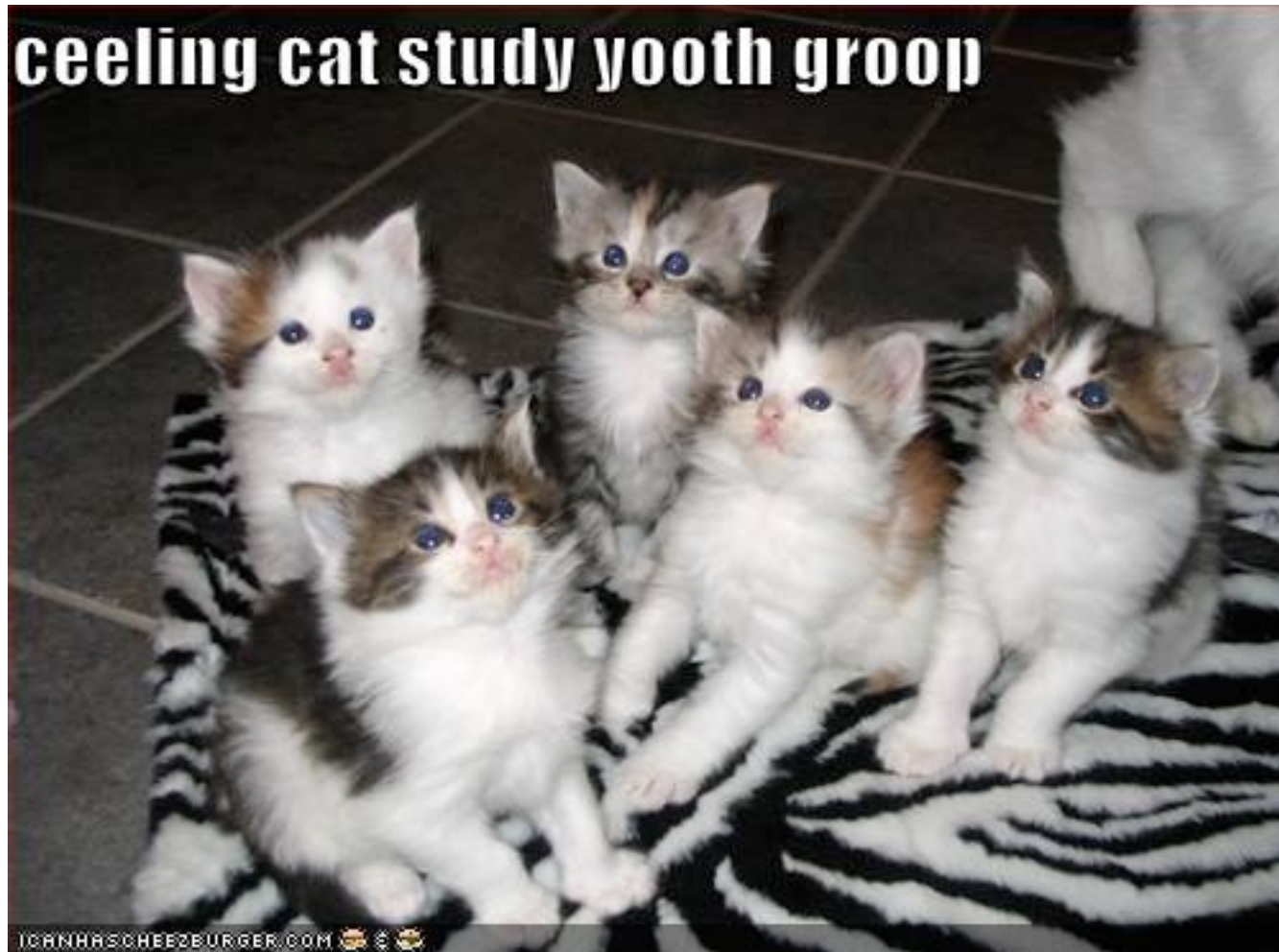
proof_alert

support_pages

Let the fun begin!



Remember: Stick with your group



Halting Problem

Input: A program P

Output: **Yes** if P terminates on all possible inputs
No otherwise

Let A be an algorithm/program that solves the Halting problem on all inputs

