Hello!

CS 839: ADVANCED MACHINE LEARNING SYSTEMS

Shivaram Venkataraman
Spring 2022
WHO AM I?

Assistant Professor in Computer Science

PhD at UC Berkeley: System Design for Large Scale Machine Learning

Industry: Google, Microsoft Research
Open source: Apache Spark committer

Call Me: Shivaram or Prof. Shivaram
COURSE LOGISTICS

Instructor: Shivaram Venkataraman
Office hours: TBD, CS 7367 (or Zoom?)

Discussion, Questions: Use Piazza!

In-person lectures, discussion
What is this course about?

What will you do in this course?
MACHINE LEARNING SUCCESSES
UNREASONABLE EFFECTIVENESS OF DATA

[Halvey et. al, IEEE Intelligent Systems 2009]
IMAGENET ACCURACY

What next??

99%

1. Jump from SIFT + FVs → AlexNet
2. Need for new benchmarks?

Deep learning models

density of points

Florence-CoSwin-H

110
100
90
80
70
DEEP LEARNING

ResNet18

Convolution
ReLU
MaxPool
Fully Connected
SoftMax
DATA PARALLEL MODEL TRAINING

SGD to train
part 0
part 1
compute gradients
aggregate
next iteration
SCALING CHALLENGES

Large Models

GPT-3: Language Models are Few-Shot Learners

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3’s few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

Large Datasets
MACHINE LEARNING SYSTEMS
FROM A SYSTEMS VIEW

Abstract operations that go on in ML training
- Accelerate or Improve utilization

Read
Compute
Aggregate
Broadcast

Communication stage

Data flow:
- Data flows into the system
- Read
- Compute
- Aggregate
- Broadcast
- Gradient
Developing libraries that make it easier and more efficient for users

```python
class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = nn.Conv2d(1, 10, kernel_size=5)
        self.conv2 = nn.Conv2d(10, 20, kernel_size=5)
        self.conv2_drop = nn.Dropout2d()
        self.fc1 = nn.Linear(320, 50)
        self.fc2 = nn.Linear(50, 10)
```
MACHINE LEARNING IN ENTERPRISES

Hyper parameter tuning

Inference

Collective goal across users or across models

Deploy models on devices

Hey Siri
HARDWARE EVOLUTION

Commodity CPUs
Lots of disks
Low bandwidth network
(2001 Google)

GPUs – Graphics Cards
Lots of parallelism
Bigger power footprint
Expensive!
(~2010)

How can we make ML operations (training + inference) work well on new hardware?

TPUs, FPGAs, ASICS
ML specialized hardware
(~2020)
FLOPS OVER TIME

THIS CLASS

Hyper parameter tuning / Scheduling

TRAINING

Compute

Communication

INFERENCEn

Data-center

Deploy on Edge devices
COURSE GOALS

• Learn in depth about ML Systems
• Prepare for research in this area
• State-of-the-art topics
  – No textbook
  – First time class offered: Learn together
COURSE SYLLABUS
LEARNING OBJECTIVES

At the end of the course you will be able to

• Critique and evaluate the design of machine learning systems
• Develop and utilize tools to profile and understand the performance of ML systems
• Propose new research ideas in topics related to ML systems
• Design and implement new ML systems.
LEARNING OBJECTIVES

At the end of the course you will be able to

• Critique and evaluate the design of machine learning systems
• Develop and utilize tools to profile and understand the performance of ML systems
• Propose new research ideas in topics related to ML systems
• Design and implement new ML systems.
COURSE FORMAT

Schedule: http://cs.wisc.edu/~shivaram/cs839-sp22
Reading: ~1 paper per class

Five broad themes
  Compute
  Communication
  Serving
  Hyperparameter tuning
  Applications
COURSE THEMES

Each theme will have
- First lecture by professor. Overview, background of area
- **Three** sessions of student led presentations
  - In depth discussion of state-of-the-art systems
- Compute, Communication also have a tools lecture
  - Led the professor to discuss how to profile, understand
  - Assignment to do this on different models / HW
What do you need to do?

- Form a group of two students and sign up for a slot (Shared later today)
- Prepare a ~35 min presentation
  - Include some background
  - Technical details
  - Reviews from other students (9pm)
- Prepare 3 discussion questions
  - ~20 mins for discussion in small groups
  - ~20 mins discussion as a class (Professor)
HOW TO MAKE A KILLER PRESENTATION
Goals

Educate audience
Promote discussion

ORGANIZE
ORGANIZE

Goals

Structure

Problem Statement  3 slides
Approach  8 slides
Comparison  4 slides
...

ORGANIZE

Goals
Structure
Outline
Key
Melp
Narrative
Content
Draft

BUILD A STORY
DESIGN
The iPhone

• LCD touch screen
• ultra thin: 115 x 61 x 12 mm
• works as a widescreen iPod
• 2 megapixel camera
• Safari browser
• conference calling
DELIVERY

Exercise Your Voice
COURSE FORMAT

Schedule: http://cs.wisc.edu/~shivaram/cs839-sp22
Reading: ~1 paper per class

Review: Fill out review form (link posted on Piazza) by 9pm night before!
Discussion: In-class group discussion, submit responses within 24 hours

What if you cannot attend?
   Best 15 responses out of 20 or so sessions
ASSESSMENT

• Paper Reviews: 10%
• Class Participation: 10%
• Paper presentation: 20%
• Assignments (10% each): 20%
• Final Project (in groups): 40%
ASSIGNMENTS

Two homework assignments
- Assignment 1: Compute
- Assignment 2: Communication

Short assignments in profiling, understanding existing models, systems
Work in groups of two
COURSE PROJECT

Main grading component in the course!

Explore new research ideas in Machine Learning systems

Work towards workshop/conference paper
COURSE PROJECT

Project Selection:
- Form groups of two (discuss with instructor for groups of 3)
- Propose project title, abstract
- Instructor feedback

Assessment:
- Prepare project introduction pitch (5 mins)
- Mid-semester check-in
- Final project presentation (15 mins)
- Final project report
WAITLIST

Class size is limited to \(~30\)
Focus on research projects, active discussion

If you are enrolled but don’t want to take, please drop ASAP!
If you are on the waitlist, we will admit students as spots open up

If you want to audit the class: you are welcome to attend, submit reviews, participate in discussions.
BEFORE NEXT CLASS

Join Piazza: https://piazza.com/wisc/spring2022/8393

Sign up for a presentation slot!

Paper Reading: The GPU computing era

"Shivaram 839 UW Madison"