Deep Multi-Task and Meta Learning

CS 330
The Plan for Today

1. Course goals & logistics
2. Why study multi-task learning and meta-learning?

Key learning today: what is multi-task learning??
Introductions

Chelsea Finn  Instructor
Amelie Byun  Course Coordinator
Ansh Khurana  Head TA
Yoonho Lee  TA
Alex Sun  TA
Max Sobol Mark  TA
Welcome!

First question: How are you doing?
(answer by raising hand)
Information & Resources

We have put a lot of info here  
Please read it. :)

Ed: Connected to Canvas

Staff mailing list: [cs330-staff-aut2324@lists.stanford.edu](mailto:cs330-staff-aut2324@lists.stanford.edu)

Office hours: Check course website & Canvas, *start today*.

OAE letters can be sent to staff mailing list or in private Ed post.
What will you learn in this course?

1. The foundations of modern deep learning methods for learning across tasks
2. How to implement and work with practical multi-task & transfer learning systems (in PyTorch)
3. A glimpse into the scientific and engineering process of building and understanding new algorithms
Topics

1. Multi-task learning, transfer learning basics
2. Meta-learning algorithms
   (black-box approaches, optimization-based meta-learning, metric learning)
3. Advanced meta-learning topics
   (meta-overfitting, unsupervised meta-learning, Bayesian models)
4. Unsupervised pre-training for few-shot learning
5. Relation to foundation models & in-context learning
6. Domain adaptation & generalization
7. Lifelong learning
8. Open problems

Case studies of important & timely applications
- Multi-task learning in recommender systems
- Meta-learning for land cover classification, education
- Few-shot learning in large language models

Emphasis on deep learning techniques.

Zhao et al. Recommending What Video to Watch Next. 2019
Brown et al. Language Models are Few-Shot Learners. 2020
New last year: No reinforcement learning

What if I want RL?

- **New course** in Spring quarter (CS224R: Deep Reinforcement Learning)
- Removing RL makes the course **more accessible**.
- You can still explore RL topics in final project.
Lectures & Office Hours

Lectures
- In-person, livestreamed, & recorded
- Two guest lectures (TBD)
- Three TA-led tutorial sessions (Thursdays)

Ask questions!
- by raising your hand

Office hours
- mix of in-person and remote

Participation
- Opportunity for up to 2% extra credit (joining lectures, helping on Ed)
Pre-Requisites

**Machine learning**: CS229 or equivalent.
- e.g. we’ll assume knowledge of SGD, cross-val, calculus, probability theory, linear algebra

Some familiarity with deep learning:
- We’ll build on concepts like backpropagation, recurrent networks
- Assignments will require training networks in **PyTorch**.
- Ansh will hold a PyTorch review session on Thursday, Sep 28, 3-4:20 pm in Gates B3.
Assignments

Homework 0: Multi-task learning basics  
5% of grade

Homework 1: Multi-task data processing, black-box meta-learning

Homework 2: Gradient-based meta-learning & metric learning

Homework 3: Fine-tuning pre-trained models

Homework 4 (optional): Bayesian meta-learning & meta-overfitting (replaces 15% of HW or project)

Grading: 50% homework, 50% project

6 late days total across: homeworks, project-related assignments  
maximum of 2 late dates per assignment

Collaboration policy: Please read course website & honor code.  
Document collaborators & write up HW solutions on your own. (incl. no AI tools)
Final Project

Research-level project of your choice

- in groups of **1-3 students**
- if applicable, encouraged to **use your research!**
- can share with other classes, with slightly higher expectation
- same late day policy as HWs
  (but no late days for poster)

Poster presentation on December 6th.

We’ll provide **~$75 in Azure credits** for HWs & project
Initial Steps

1. Homework 0 is out — due next Wednesday at 11:59 pm PT
2. Start forming final project groups if you want to work in a group
The Plan for Today

1. Course goals & logistics

2. Why study multi-task learning and meta-learning?
Some of Chelsea’s Research
(and why I care about multi-task learning and meta-learning)
How can we enable agents to learn a breadth of skills in the real world?

Robots can teach us things about intelligence.

Robots faced with the real world must generalize across tasks, objects, environments, etc.

need some common sense understanding to do well

supervision can’t be taken for granted
The robot had its eyes closed.

Levine et al. ICRA ’15
Our Method
autonomous execution

real-time
Finn et al. ICRA ‘16
Learn one task in one environment, starting from scratch
Behind the scenes…

Yevgen is doing more work than the robot!

It’s not practical to collect a lot of data this way.
Learn **one task** in **one environment**, starting from scratch using detailed supervision.

Not just a problem with reinforcement learning & robotics.

More diverse, yet still **one task, from scratch, with detailed supervision**
Humans are **generalists**.

Source: [https://youtu.be/8vNxjwt2AqY](https://youtu.be/8vNxjwt2AqY)
Why should we care about multi-task & meta-learning?

…beyond the robots and general-purpose ML systems
Why should we care about multi-task & meta-learning?

...beyond the robots and general-purpose ML systems
Standard computer vision: hand-designed features

Modern computer vision: end-to-end training

Deep learning allows us to handle *unstructured inputs* (pixels, language, sensor readings, etc.) without hand-engineering features, with less domain knowledge
Deep learning for object classification

ImageNet competition results

Deep learning for machine translation

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui.schuster.zhifengc.qv1.mnorouzi@google.com

Table 10: Mean of side-by-side scores on production data

<table>
<thead>
<tr>
<th>Source Language → Target Language</th>
<th>PBMT</th>
<th>GNMT</th>
<th>Human</th>
<th>Relative Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>English → Spanish</td>
<td>4.885</td>
<td>5.428</td>
<td>5.504</td>
<td>87%</td>
</tr>
<tr>
<td>English → French</td>
<td>4.932</td>
<td>5.295</td>
<td>5.496</td>
<td>64%</td>
</tr>
<tr>
<td>English → Chinese</td>
<td>4.035</td>
<td>4.594</td>
<td>4.987</td>
<td>58%</td>
</tr>
<tr>
<td>Spanish → English</td>
<td>4.872</td>
<td>5.187</td>
<td>5.372</td>
<td>63%</td>
</tr>
<tr>
<td>French → English</td>
<td>5.046</td>
<td>5.343</td>
<td>5.404</td>
<td>83%</td>
</tr>
<tr>
<td>Chinese → English</td>
<td>3.694</td>
<td>4.263</td>
<td>4.636</td>
<td>60%</td>
</tr>
</tbody>
</table>

Human evaluation scores on scale of 0 to 6

PBMT: Phrase-based machine translation
GNMT: Google’s neural machine translation (in 2016)

Why deep multi-task and meta-learning?
Large, diverse data → Broad generalization (+ large models)

Russakovsky et al. ‘14

Wu et al. ‘16

Vaswani et al. ‘18

What if you don’t have a large dataset?

medical imaging, robotics, personalized education, translation for rare languages

Impractical to learn from scratch for each disease, each robot, each person, each language, each task
What if your data has a long tail?

This setting breaks standard machine learning paradigms.
What if you need to quickly learn something new?
about a new person, for a new task, about a new environment, etc.
training data

Braque

Cezanne

test datapoint

By Braque or Cezanne?
What if you need to quickly learn something new? about a new person, for a new task, about a new environment, etc.

“few-shot learning”

How did you accomplish this? by leveraging prior experience!
What if you want a more general-purpose AI system?
Learning each task from scratch won’t cut it.

What if you don’t have a large dataset?
- medical imaging
- robotics
- translation for rare languages
- personalized education, medicine, recommendations

What if your data has a long tail?

What if you need to quickly learn something new?
- about a new person, for a new task, about a new environment, etc.

This is where elements of multi-task learning can come into play.
Why now?

Why should we study deep multi-task & meta-learning now?
Multitask Learning

RICH CARUANA

Multitask Learning (MTL) is an inductive transfer mechanism whose principle goal is to improve generalization performance. MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks. It does this by training tasks in parallel while using a shared representation. In effect, the training signals for the extra tasks serve as an inductive bias. Section 1.2 argues that inductive transfer is important if we wish to scale tabula rasa learning to complex, real-world tasks. Section 1.3 presents the simplest method we know for doing multitask inductive transfer, adding extra tasks (i.e., extra outputs) to a backpropagation net. Because the MTL net uses a shared hidden layer trained in parallel on all the tasks, what is learned for each task can help other tasks be learned better. Section 1.4 argues that it is reasonable to view training signals as an inductive bias when they are used this way.

Caruana, 1997

On the Optimization of a Synaptic Learning Rule

Samy Bengio  Yoshua Bengio  Jocelyn Cloutier  Jan Geesei

Université de Montréal, Département IRO

This paper presents a new approach to neural modeling based on the idea of using an automated method to optimize the parameters of a synaptic learning rule. The synaptic modification rule is considered as a parametric function. This function has local inputs and is the same in many neurons. We can use standard optimization methods to select appropriate parameters for a given type of task. We also present a theoretical analysis permitting to study the generalization property of such parametric learning rules. By generalization, we mean the possibility for the learning rule to learn to solve new tasks. Experiments were performed on three types of problems: a

Bengio et al. 1992

Is Learning The n-th Thing Any Easier Than Learning The First?

Sebastian Thrun

They are often able to generalize correctly even from a single training example [2, 10]. One of the key aspects of the learning problem faced by humans, which differs from the vast majority of problems studied in the field of neural network learning, is the fact that humans encounter a whole stream of learning problems over their entire lifetime. When faced with a new thing to learn, humans can usually exploit an enormous amount of training data and experiences that stem from other, related learning tasks. For example, when learning to drive a car, years of learning experience with basic motor skills, typical traffic patterns, logical reasoning, language and much more precede and influence this learning task. The transfer of knowledge across learning tasks seems to play an essential role for generalizing accurately, particularly when training data is scarce.

Thrun, 1998
These methods are continuing to play a major role in AI.

Visual language models can learn many distinct tasks.

Object recognition:

- This is a chinchilla. They are mainly found in Chile.
- This is a shiba. They are very popular in Japan.
- This is a flamingo. They are found in the Caribbean and South America.

Reading & arithmetic:

- 2 + 1 = 3
- 5 + 6 = 11
- 3 × 6 = 18

Counting:

- Pandas: 3
- Dogs: 2
- Giraffes: 4

These methods are continuing to play a major role in AI.

Meta-learning enables automatic feedback on student work on new problems.

Code-in-Place 2021

Humans gave feedback on ~1k student programs. Meta-learning system gave feedback on the remaining ~15k.

Wu, Goodman, Piech, Finn. ProtoTransformer: A Meta-Learning Approach to Providing Student Feedback. 2021
These methods are continuing to play a major role in AI.

**Multilingual machine translation**

*Introducing the First AI Model That Translates 100 Languages Without Relying on English*

Fan et al. JMLR, 2021

- M2M-100 is trained on a total of 2,200 language directions — or 10x more than previous best, English-centric multilingual models. Deploying M2M-100 will improve the quality of translations for billions of people, especially those that speak low-resource languages.

**YouTube recommendations**

*Recommending What Video to Watch Next: A Multitask Ranking System*

Chao Zhao, Lichun Hong, Li Wei, JiLin Chen, Anirudh Nath, Shawn Andrews, Aditya Kunzhekar, Maheswaran Sathiamoorthy, Xinyang Yi, Yi Li

Google, Inc.

[jichao,lichun,jilin,jilin 본인의 이메일 주소를 써주세요. Shawn Andrews and Xinyang Yi (ylxy2018@google.com)

In this paper, we introduce a large scale multi-objective ranking system for recommending what video to watch next on an industrial video sharing platform. The system faces many real-world challenges, including the presence of multiple competing ranking objectives, as well as implicit selection biases in user feedback. To

RecSys 2019

**A Generalist Agent**

Reed et al. TMLR 2022

**Zero-shot robot generalization**

RT-2 Brohan et al. 2023

- move coke can to Taylor Swift
- pick up the bag about to fall off the table
- put strawberry into the correct bowl

**GPT-4-V**

**One-shot imitation from humans**

DAML Yu et al. RSS 2018
Its success is important for the \textit{democratization} of deep learning.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>1.2 million images and labels</td>
</tr>
<tr>
<td>WMT ’14 English - French</td>
<td>40.8 million paired sentences</td>
</tr>
<tr>
<td>Switchboard Speech Dataset</td>
<td>300 hours of labeled data</td>
</tr>
<tr>
<td>Kaggle’s Diabetic Retinopathy Detection</td>
<td>35K labeled images</td>
</tr>
<tr>
<td>dataset</td>
<td></td>
</tr>
<tr>
<td>Adaptive epilepsy treatment with RL</td>
<td>&lt; 1 hour of data</td>
</tr>
<tr>
<td>Guez et al. ‘08</td>
<td></td>
</tr>
<tr>
<td>Learning for robotic manipulation</td>
<td>&lt; 15 min of data</td>
</tr>
<tr>
<td>Finn et al. ‘16</td>
<td></td>
</tr>
</tbody>
</table>
But, we also still have many open questions and challenges!
What is a task?
What is a task?

Informally: dataset $\mathcal{D}$ → model $f_\theta$
loss function $\mathcal{L}$

Different tasks can vary based on:
- different objects
- different people
- different objectives
- different lighting conditions
- different words
- different languages
- ...

Not just different “tasks”
Critical Assumption

The bad news: Different tasks need to share some structure. If this doesn’t hold, you are better off using single-task learning.

The good news: There are many tasks with shared structure!

- The laws of physics underly real data.
- People are all organisms with intentions.
- The rules of English underly English language data.
- Languages all develop for similar purposes.

Even if the tasks are seemingly unrelated: This leads to far greater structure than random tasks.
Informal Problem Definitions
We’ll define these more formally later.

The multi-task learning problem: Learn a set of tasks more quickly or more proficiently than learning them independently.

The transfer learning problem: Given data on previous task(s), learn a new task more quickly and/or more proficiently.

This course: anything that solves these problem statements.
Doesn’t multi-task learning reduce to single-task learning?

\[ \mathcal{D} = \bigcup \mathcal{D}_i \quad \mathcal{L} = \sum \mathcal{L}_i \]

Are we done with the course?
Doesn’t multi-task learning reduce to single-task learning?

Yes, it can! Aggregating the data across tasks & learning a single model is one approach to multi-task learning.

But, what if you want to learn *new* tasks?

And, how do we tell the model what task to do?

And, what if aggregating doesn’t work?
Reminders

1. Homework 0 is out
2. Start forming final project groups if you want to work in a group

Next time (Mon): Multi-Task Learning Basics