Deep Multi-Task and Meta-Learning

CS 330
Introductions

Chelsea Finn
Instructor

Karol Hausman
Co-Lecturer

Rafael Rafailov
TA

Dilip Arumugan
TA

Mason Swofford
TA

Albert Tung
TA

More TAs coming soon.
We're here
The Plan for CS330 in 2020

Live lectures on zoom, as interactive as possible
- Ask questions!
  - By raising your hand (preferred)
  - By entering the question in chat
- Camera use encouraged when possible, but not at all required
- Lectures from Karol, Matt Johnson, Jane Wang to mix things up
- Project proposal spotlights, project presentations
- Options for students in far-away timezones, conflicts, zoom fatigue

Assignments & Project
- Short project spotlight presentations
- Less time for project than typical (no end-of-term period)
- Making fourth assignment optional

Case studies of important & timely applications
- Multi-objective learning in YouTube recommendation system
- Meta-learning for few-shot land cover classification
- Few-shot learning from GPT-3

Rußwurm et al. Meta-Learning for Few-Shot Land Cover Classification. 2020
Brown et al. Language Models are Few-Shot Learners. 2020
Zhao et al. Recommending What Video to Watch Next. 2019
First question: How are you doing?

(answer in chat)
The Plan for Today

1. Course logistics
2. Why study multi-task learning and meta-learning?
Course Logistics
Course website: http://cs330.stanford.edu/

Piazza: Stanford, CS330

Staff mailing list: cs330-aut2021-staff@lists.stanford.edu

Office hours: Check course website & piazza, start on Weds.
Pre-Requisites and Enrollment

**Pre-requisites:** CS229 or equivalent, previous or concurrent RL knowledge highly recommended.

**Lectures are recorded,**
- will be internally released on Canvas after each lecture
- will be edited & publicly released after the course
Assignments will require training networks in TensorFlow (TF) in Colab notebook.

TF Review section:
- Rafael will hold a TF 2.0 review session on Thursday, September 17, 6 pm PT.
- You should be able to understand the overview here: https://www.tensorflow.org/guide/eager
- If you don’t, go to the review session & ask questions!
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)
4. Hierarchical Bayesian models & meta-learning
5. Multi-task RL, goal-conditioned RL
6. Meta-reinforcement learning
7. Hierarchical RL
8. Lifelong learning
9. Open problems

Emphasis on deep learning techniques.
Emphasis on reinforcement learning domain (6 lectures)
Topics We Won’t Cover

Won’t cover AutoML topics:
- architecture search
- hyperparameter optimization
- learning optimizers

Though, many of the underlying techniques will be covered.
Assignments & Final Project

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

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

**Homework 3:** Multi-task RL, goal relabeling

**Homework 4 (optional):** Meta-RL

**Project:** Research-level project of your choice

Form groups of 1-3 students, you’re encouraged to start early!

**Grading:** 45% homework (15% each), 55% project

HW4 either replaces one prior HW or part of project grade (whichever is better for grade).

**6 late days** total across: homeworks, project-related assignments

maximum of 2 late dates per assignment
Homework Today

1. Sign up for Piazza
2. Start forming final project groups if you want to work in a group
3. Review this: https://www.tensorflow.org/guide/eager
The Plan for Today

1. Course logistics

2. Why study multi-task learning and meta-learning?
Some of My 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.

Why robots? Robots can teach us things about intelligence.

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
Beginning of my PhD

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, with detailed supervision and guidance.

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
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

*Slide adapted from Sergey Levine*
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, qvl, mnorouzi@google.com


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

<table>
<thead>
<tr>
<th>Language Pair</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 (+ large models) \[\rightarrow\] Broad generalization

What if you don’t have a large dataset?

Impractical to learn from scratch for each disease, each robot, each person, each language, each task

Russakovsky et al. ‘14

Wu et al. ‘16

Vaswani et al. ‘18
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.
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  personalized education, medicine, recommendations
translation for rare languages

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.
What is a task?
What is a task?

For now: $\text{dataset } D \rightarrow \text{model } f_\theta$

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”
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!

Even if the tasks are seemingly unrelated:

- 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.

This leads to far greater structure than random tasks.
Informal Problem Definitions

We’ll define these more formally next time.

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

The meta-learning problem: Given data/experience on previous tasks, 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?

\[ D = \bigcup D_i \quad \text{and} \quad L = \sum L_i \]

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, we can often do better! Exploit the fact that we know that data is coming from different tasks.
Why now?

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

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 Gecsei
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:

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 algorithms are continuing to play a fundamental role in machine learning research.

**Multilingual machine translation**

*Massively Multilingual Neural Machine Translation*

Roei Aharoni*
Bar-Ilan University
Ramat-Gan
Israel
roee.aharoni@gmail.com

Melvin Johnson and Orhan Firat
Google AI
Mountain View
California
melvinp,orhanf@google.com

while supporting up to 59 languages. Our experiments on a large-scale dataset with 102 languages demonstrate that from English and up to one million examples per direction also show promising results, surpassing strong bilingual baselines and encouraging future work on massively multilingual NMT.

NAACL, 2019

**Text-to-Text Transformer**

Raffel et al. JMLR 2020

**One-shot imitation learning from humans**

DAML Yu et al. RSS 2018

**Multi-domain learning for sim2real transfer**

CAD²RL Sadeghi & Levine, RSS 2017

**YouTube recommendations**

Recommending What Video to Watch Next: A Multitask Ranking System

Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sethia, Shwetha, Xinyang Yi, Ed Chi

Google, Inc.

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
These algorithms are playing a fundamental, and **increasing** role in machine learning research.

**Interest level via Google search queries**

- **How transferable are features in a deep neural network?**
  Yosinski et al. ‘15
- **Learning to learn by gradient descent by gradient descent**
  Andrychowicz et al. ‘15
- **Model-agnostic meta-learning for fast adaptation of deep networks**
  Finn et al. ‘17
- **An overview of multi-task learning in neural networks**
  Ruder ‘17

Graph sources: Google scholar, Google trends
Its success will be critical for the **democratization** of deep learning.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</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 dataset</td>
<td>35K labeled images</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 still have many open questions and challenges!
Reminder: Homework Today

1. Sign up for Piazza
2. Start forming final project groups if you want to work in a group
3. Review this: https://www.tensorflow.org/guide/eager

Next time (Weds): Multi-Task Learning & Transfer Learning Basics