CS 598: Transfer Learning

Lecture 1: Overview and Introduction

Han Zhao
08/25/2021

Please keep your face covering on!
This fall, the university will require students, faculty and staff who are able to do so to be fully vaccinated for COVID-19.

FACE COVERINGS

When and where do you have to wear a face covering?

- All university students, faculty, staff and visitors must wear a face covering in university spaces indoors, regardless of vaccination status.
  - People who are NOT fully vaccinated are required to wear a face covering in university spaces indoors, as well as outdoors when they cannot practice social distancing. This applies to all students, faculty, staff and visitors.

A face covering does not have to be worn when eating or drinking.

- Additionally, face coverings do not need to be worn when alone in an office or room with the door closed, alone in a vehicle—essentially, when you are alone in a private space.
  - A “private room or space” is a room where a person is alone with the door closed. A cubicle is not considered a private space unless it has a door that is closed and cubicle walls of sufficient height that are above full head height. If the occupant’s head will extend beyond the top of the cubical, a face covering must be worn. Regardless of the presence of a door or cubicle wall height, a face covering is not required in one’s office or cubicle while eating or drinking, but the occupant must be seated.

If a student is not wearing a face covering in your class, ask them to put one on. If they refuse to put on a face covering, ask the student to leave. If the student refuses to leave, dismiss the class and report the student to the Office for Student Conflict Resolution for further discipline by filling out this form.
Covid-19 Policy

More details about Covid-19 policy at the university: https://covid19.illinois.edu/

Following University policy, all students are required to engage in appropriate behavior to protect the health and safety of the community. Students are also required to follow the campus COVID-19 protocols.

Students who feel ill must not come to class. In addition, students who test positive for COVID-19 or have had an exposure that requires testing and/or quarantine must not attend class. The University will provide information to the instructor, in a manner that complies with privacy laws, about students in these latter categories. These students are judged to have excused absences for the class period and should contact the instructor via email about making up the work.

Students who fail to abide by these rules will first be asked to comply; if they refuse, they will be required to leave the classroom immediately. If a student is asked to leave the classroom, the non-compliant student will be judged to have an unexcused absence and reported to the Office for Student Conflict Resolution for disciplinary action. Accumulation of non-compliance complaints against a student may result in dismissal from the University.

TL; DR:

- Please always wear a face covering in classrooms;
- If you feel ill, please do not come to class;
- If you test positive for Covid-19 or have had an exposure that requires testing and/or quarantine, please do not come to class;
Brief Bio

- Name: Han Zhao
- Current position: Assistant Professor @ CS
- Previous positions:
  - Machine Learning Researcher @ D. E. Shaw
  - Ph.D. in MLD @ CMU
  - MMath @ Waterloo
  - BEng in CS @ Tsinghua
- Research Interests:
  - Domain adaptation / generalization
  - Multitask learning / Meta-learning
  - Algorithmic fairness
  - Probabilistic circuits (e.g., arithmetic circuits, sum-product networks)
Logistics

- Course website: https://hanzhaoml.github.io/courses/cs598-haz/index.html
- Piazza: UIUC, CS 598 - HAZ
- Office Hour: Friday, 3:30 pm - 4:30 pm
- Email: hanzhao@illinois.edu
- My office: 3320 Siebel Center

- TA: Ruicheng Xian: rxian2@illinois.edu
- Office Hour: Thursday, 3:00 pm - 4:00 pm
- Ruicheng’s office: 3405 Siebel Center
Course Overview

• Mostly on topics related to transfer learning:
  - generalization theory of supervised learning in iid setting
  - domain adaptation / generalization
  - multitask learning
  - meta-learning
  - invariant representations

• Often the case, we will cover these topics from a perspective of learning proper representations
Course Overview

We will mostly talk about the existing theory and algorithms of these topics. Prerequisites include:

- Probability and statistics
- Linear algebra
- Mathematical analysis
- Information theory (basic)
- Comfortable with programming in Python (TensorFlow, PyTorch, etc)

Basically, a certain level of mathematical maturity + some basic knowledge about machine learning (CS 446)
Course Overview

What we will NOT cover:
- Practical implementation / training of the introduced algorithms
- Reinforcement learning
- Causality
- Empirical tricks used in optimizing neural networks
- ……
Course Overview

This course is graduate-level seminar course (discussion & project heavy), it will contain two parts:

- Lectures (Tentative: Week 1 ~ Week 7)
- Paper presentation (Tentative: Week 8 ~ Week 13)
- Course project (Group of 1 ~ 3)

Grading scheme:

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper Presentation</td>
<td>30%</td>
<td>Presenting one research paper (related to transfer learning) in class</td>
</tr>
<tr>
<td>Course Project</td>
<td>60%</td>
<td>This includes a break down of 10% for project proposal, 20% for final presentation and 30% for the project report</td>
</tr>
<tr>
<td>Participation</td>
<td>10%</td>
<td>This includes both in-class discussion and Piazza participation</td>
</tr>
</tbody>
</table>
Course Project

- Literature review or research project
- Project proposal (10%) + final presentation (20%) + project report (30%)
- Tentative timeline:
  - Proposal (due on Sep. 29th, 2 pages)
  - Final representation (the last three lectures, ~15 mins)
  - Project report (due on Dec. 10th, 9 pages + reference + appendix)
- Format: pdf in NeurIPS LaTeX template (https://neurips.cc/Conferences/2021/PaperInformation/StyleFiles)
- At the end of the project report, please clearly list the contributions of each team member
- I want to learn from you!
Course Project

Literature Review:

• Any topic broadly related to transfer learning, e.g., recent advances in meta-learning / multitask learning, out-of-domain generalization etc.

• Group = 1

• Must be very clear and include your own thoughts about the potential future directions in the corresponding sub-field
Course Project

Research Project:

- Any topic broadly related to transfer learning, e.g., recent advances in meta-learning / multitask learning, out-of-domain generalization etc.

- Group of 2 ~ 3

- You can choose to work on either a theory or application-oriented project

- I will provide a list of references and potential topics to work on within the first month (hopefully sooner), but you are free to work on any topic that interests you.
Course Project

Project proposal:
- <= 2 pages (due on Sep. 29th)
- Should cover:
  - For both literature review and research project, explain what topics you want to work on
  - Motivation: why do you choose the topic?
  - In the case of a research project: what’s the goal of this project? Will it be mostly theory-focused or application-oriented?
  - List of group members
  - ...

Paper Presentation

- I will provide a list of papers that you can choose from, but you are also free to choose any paper that you find interesting (as long as it is broadly related to transfer learning)
- Again, group presentation (1~3, the same group for your course project)
- More details on the guideline will be posted on the course website
- I will provide a google sheet link for signup
Course Project

Homework today:

- Sign up for Piazza
- Start to form your group for both the paper presentation and course project

Questions?
Introduction

The success of large-scaled supervised learning in acoustic speech processing:

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD-HMM [26]</td>
<td>27.3%</td>
</tr>
<tr>
<td>AUGMENTED CONDITIONAL RANDOM FIELDS [26]</td>
<td>26.6%</td>
</tr>
<tr>
<td>RANDOMLY INITIALIZED RECURRENT NEURAL NETS [27]</td>
<td>26.1%</td>
</tr>
<tr>
<td>BAYESIAN TRIPHONE GMM-HMM [28]</td>
<td>25.6%</td>
</tr>
<tr>
<td>MONOPHONE HTMS [29]</td>
<td>24.8%</td>
</tr>
<tr>
<td>HETEROGENEOUS CLASSIFIERS [30]</td>
<td>24.4%</td>
</tr>
<tr>
<td>MONOPHONE RANDOMLY INITIALIZED DNNs (SIX LAYERS) [13]</td>
<td>23.4%</td>
</tr>
<tr>
<td>MONOPHONE DBN-DNNs (SIX LAYERS) [13]</td>
<td>22.4%</td>
</tr>
<tr>
<td>MONOPHONE DBN-DNNs WITH MMI TRAINING [31]</td>
<td>22.1%</td>
</tr>
<tr>
<td>TRIPHONE GMM-HMMs DT W/ BMMI [32]</td>
<td>21.7%</td>
</tr>
<tr>
<td>MONOPHONE DBN-DNNs ON FBANK (EIGHT LAYERS) [13]</td>
<td>20.7%</td>
</tr>
<tr>
<td>MONOPHONE MCRBM-DBN-DNNs ON FBANK (FIVE LAYERS) [33]</td>
<td>20.5%</td>
</tr>
<tr>
<td>MONOPHONE CONVOLUTIONAL DNNs ON FBANK (THREE LAYERS) [34]</td>
<td>20.0%</td>
</tr>
</tbody>
</table>

[Hinton et al. 12]
The success of large-scaled supervised learning in computer vision:

ImageNet: ~1M images, ~1K classes

[Deng et al. 09]
The success of large-scaled supervised learning in natural language understanding:

Machine Translation on WMT2014 English-French

Machine Translation, ~3M parallel sentences

[Cho et al. 2014; Devlin et al. 2014]
Introduction

Key assumption underlying the success: large-scale labeled data from stationary domains

\[
\text{Source (Train)} = \text{Target (Test)}
\]
Introduction

But, often the case, such an assumption does not hold.
Introduction

But, often the case, such an assumption does not hold

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Corpora size</th>
<th>BLEU Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>French</td>
<td>~3M</td>
<td>~40</td>
</tr>
<tr>
<td>English</td>
<td>German</td>
<td>~1.92M</td>
<td>~35</td>
</tr>
<tr>
<td>Finnish</td>
<td>English</td>
<td>~1.96M</td>
<td>~34</td>
</tr>
<tr>
<td>Romanian</td>
<td>English</td>
<td>~400K</td>
<td>~30</td>
</tr>
</tbody>
</table>

WMT ’16-19, Europarl Parallel Corpus
Introduction

Domain adaptation: given unlabeled data from the target domain + labeled data from the source domain, can we do better?

Note: closely related to the setting of semi-supervised learning, but with a key difference:

Semi-supervised learning:

Training distribution \(\equiv\) Test distribution

Domain adaptation:

Training distribution \(\neq\) Test distribution
Introduction

Domain adaptation: Training phase

Source domain:

\[(4, 8, 5) \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad (7, 3, 0)\]

Target domain:
Introduction

Domain adaptation: Training phase

Source domain:

(4, 8, 5)  (7, 3, 0)

Target domain:

Classifier
Introduction

Domain adaptation: Testing phase

Target domain:

(4, 0, 1)

✓ ✓ ✓

(0, 4, 2)

✗ ✓ ✓
Introduction

Domain generalization: but, what if I cannot even collect unlabeled data from the target domain? Instead, I can collect multiple labeled data for training the task of interest

Source domains:

Target domain:
Introduction

Domain generalization:
- There is no free lunch of course (why?)
- What if we have some prior knowledge about the unseen target domain?
  - Target = convex combination of the sources?
  - Target = (restricted) affine combination of the sources?
  - Target ~ the same meta-distribution where the sources are sampled from?
- ......
Multitask learning: What if I have multiple related tasks, will learning them together simultaneously better than learning them separately?
Introduction

Multitask learning:
- What is called a “task”?
- Is it always better to train multiple tasks simultaneously?
- Quantitatively, what’s the advantage of multitask learning in terms of sample complexity? e.g., how many samples does the learner need to reach a pre-specified accuracy on each task?
- Learning shared representations?
- How to measure task-similarity?
Introduction

Meta-Learning: no standard interpretation about the problem setup, but a collection of algorithms that aim for learning to learn.

- Similar to multitask learning, but the test set contain novel tasks
- Relationship with domain generalization? Very flexible setting, novel tasks do not necessarily share the same input / output spaces
### Introduction

**Supervised Learning:** predict an outcome $y$ from some set $\mathcal{Y}$ of possible outcomes, on the basis of some observation $x$ from a feature space $\mathcal{X}$

- Given $\{(x_i, y_i)\}_{i=1}^{n}$, learn $f : \mathcal{X} \rightarrow \mathcal{Y}$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>words in a document</td>
<td>topic</td>
</tr>
<tr>
<td>image of a digit in a zipcode</td>
<td>(sports, music, tech, ...)</td>
</tr>
<tr>
<td>email message</td>
<td>the digit</td>
</tr>
<tr>
<td>sentence</td>
<td>spam or ham</td>
</tr>
<tr>
<td>patient medical test results</td>
<td>correct parse tree</td>
</tr>
<tr>
<td>gene expression levels of a tissue sample</td>
<td>patient disease state</td>
</tr>
<tr>
<td></td>
<td>presence of cancer</td>
</tr>
</tbody>
</table>
Introduction

Supervised Learning: predict an outcome $y$ from some set $\mathcal{Y}$ of possible outcomes, on the basis of some observation $x$ from a feature space $\mathcal{X}$

- Given $\{(x_i, y_i)\}_{i=1}^n$, learn $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Goal: the learned classifier $f$ maximizes # of correct predictions for subsequent inputs
- Key question (generalization): how do I know whether my classifier will be good in general on unseen data from the same distribution?
Introduction

How to measure the goodness of a predictor \( f \)?

Loss function: \( \ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \)

- So \( \ell(\hat{y}, y) \) quantifies the cost of predicting \( \hat{y} \) when the true outcome is \( y \)
- Goal: learn \( f \) so that \( \ell(f(x), y) \) is small
Introduction

Classification problem:

In object classification (e.g., ImageNet), the aim is to classify a given image into one of a finite number (1000 on ImageNet) of classes. If all the classes are equally important, we could define
Classification problem:
In object classification (e.g., ImageNet), the aim is to classify a given image into one of a finite number (1000 on ImageNet) of classes. If all the classes are equally important, we could define

$$\ell(\hat{y}, y) := \mathbb{I}(\hat{y} \neq y) = \begin{cases} 1 & \text{if } \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases}$$

- If $|\mathcal{Y}| = 2$, we call it a binary classification problem
- For convenience, sometimes we consider $\mathcal{Y} = \{0, 1\}$ but we may also use $\mathcal{Y} = \{-1, 1\}$
Introduction

Regression problem:
In stock price prediction, the aim is to predict the future return of a given stock. The return of a stock could take both positive and negative values. In this case, we could define

Squared loss: \( \ell(\hat{y}, y) := (\hat{y} - y)^2 \)

- Other loss functions are possible, e.g., L1 or the Huber loss
Introduction

Some assumptions about our problem:
- There is a joint probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- (i.i.d) The data $\{(x_i, y_i)\}_{i=1}^{n}$ the learner sees are drawn independently and identically from $\mathcal{D}$

The goal is to choose $f$ with small error:

$$\varepsilon_{\mathcal{D}}(f) := \mathbb{E}[\ell(f(X), Y)]$$

For example, for classification problem, this is the misclassification probability

$$\varepsilon_{\mathcal{D}}(f) = \mathbb{E}[\ell(f(X), Y)] = \mathbb{E}[\mathbb{I}(f(X) \neq Y)] = \Pr(f(X) \neq Y)$$
Some notation:

- Capital letters (like $X$) denote random variables

- Lower case letters (like $x$) denote realization of random variables

- $\mathbb{E}[\cdot]$ denotes taking expectation, and we omit the underlying distribution if it’s clear from the context

- The distribution $\mathcal{D}$ is a joint distribution over both the inputs and the labels, i.e., this include both the marginal distribution of $X$ and the conditional distribution of $Y \mid X$

- Often the case, the predictor $f$ depends on the training data. Because the training data $\{(x_i, y_i)\}_{i=1}^{n}$ is random, the predictor $f$ is also random

- Hence when we say $\varepsilon_{\mathcal{D}}(f)$ is small, most of the time we mean $\varepsilon_{\mathcal{D}}(f)$ is small with high probability (whp)
Introduction

The learning process:

- We can choose the predictor \( f \) from some pre-defined class of functions \( \mathcal{F} \), e.g., the class of linear predictors, decision trees, kernel machines, neural networks, etc.

Clearly, the optimal error one can ever hope to achieve depends on the distribution \( \mathcal{D} \). We call this error the Bayes error:

\[
\forall f, \quad \mathcal{D}(f) = \Pr(f(X) \neq Y) \\
= \int \left\{ \Pr(Y = 1 \mid X)\mathbb{I}(f(X) = 0) + \Pr(Y = 0 \mid X)\mathbb{I}(f(X) = 1) \right\} \, d\mu(X) \\
\geq \int \min\{\Pr(Y = 1 \mid X), \Pr(Y = 0 \mid X)\} \, d\mu(X) \\
= \mathbb{E}[\min\{\Pr(Y = 1 \mid X), \Pr(Y = 0 \mid X)\}] \\
= \frac{1}{2} - \frac{1}{2} \mathbb{E}[2\eta(X) - 1]
\]

\( \eta(X) := \Pr(Y = 1 \mid X) \)

\( \mu(\cdot) \) is the corresponding measure of \( \mathcal{D} \)