Introduction to Weak Supervision

Chris Ré
CS229
Various techniques for limited labeled data

- **Active learning:** Select points to label more intelligently
- **Semi-supervised learning:** Use unlabeled data as well
- **Transfer learning:** Transfer from one training dataset to a new task
- **Weak supervision:** Label data in cheaper, higher-level ways

This lecture.

https://www.snorkel.org/blog/weak-supervision
Messages for Today

• Biased by our on-going work.
  • This is one of many approaches (but you may have used it!)

• 1\textsuperscript{st} high-level vision about the area.
  • Why supervision is so critical in this age.
  • Missing: Why pretraining is awesome!

• Guts of lecture: Mathematical details of a problem that allows me to introduce two key concepts in (hopefully) simple way:
  • Latent probabilistic variable models with provable solutions
    • Continuing our EM-like thread.
  • Probability distributions on graphs (graphical models)
    • Fun facts about Gaussians that are good for your soul (Inverse covariance matrix structure and graphs)
Biased by active on-going work.
ML Application =

Model + Data + Hardware

State-of-the-art models and hardware are available. Training data is not
But supervision comes from god herself....
... but training data usually comes from a dirty, messy process.

Can we provide **mathematical** and **systems structure** for this messy process?
Supervision is where the action is...

Model differences *overrated*, and supervision differences *underrated*.
Automated Chest X-ray Triage

Optimizing Workflows with Automated Prioritization, Radiology 19

We spent a year on this challenge
• Created large dataset of clinical labels
• Evaluated effect of label quality
• Work published in a *clinical journal*

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW + KSVM</td>
<td>0.88</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.87</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.89</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td><strong>0.91</strong></td>
</tr>
</tbody>
</table>

**Often**: Differences in models ~ 2-3 points.

Label quality & quantity > model choice.
Data augmentation by specifying invariances

Images
- Rotations
- Scaling / Zooms
- Brightness
- Color Shifts
- Etc...

Text
- Synonymy
- Positional Swaps
- Etc...

Medical
- Synonymy
- Positional Swaps
- Etc...

Domain-specific transformations. Ex:
1. Segment tumor mass
2. Move
3. Resample background tissue
4. Blend

How do we choose which to apply? In what order?
Simple Benchmarks: Data Augmentation is Critical

Ex: 13.4 pt. avg. accuracy gain from data augmentation across top ten CIFAR-100 models—difference in top-10 models is less!
Training Signal is key to pushing SotA

New methods for gathering signal leading the state of the art

- **Google AI** AutoAugment: Using learned **data augmentation policies**
  - **Augmentation Policies** first in Ratner et al. NIPS ’17

- **Facebook** Hash tag weakly supervised pre-training
  - Pre-train using a massive dataset with *hashtags*

*Sharon Y. Li (to: Wisconsin)*
*Henry Ehrenberg (to: Washington)*
*Alex Ratner*
Automating the Art of Data Augmentation
Part I Overview

The Stanford AI Lab Blog

Sharon Y. Li
(to: Wisconsin)

http://ai.stanford.edu/blog/data-augmentation/
Training data: the new bottleneck

Slow, expensive, and static
Manual Labels

Slow

Expensive

Static

$10 - $100/hr

Manual Labels

Fast

Programmatic Labels

Cheap

Dynamic

write programs

run programs

Trade-off: programmatic labels are noisy...
Snorkel: Formalizing Programmatic Labeling

Observation: Weak supervision applied in *ad hoc* and isolated ways.
Snorkel: Formalizing Programmatic Labeling

Goal: Replace *ad hoc* weak supervision with a formal, unified, theoretically grounded approach for programmatic labeling
Running Example: NER

Dr. Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

Let’s look at labeling “Person” versus “Hospital”

Goal: Label training data using *weak supervision* strategies for these tasks
Dr. Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

**Problem:** These noisy sources conflict and are correlated

```python
def existing_classifier(x):
    return off_shelf_classifier(x)

def upper_case_existing_classifier(x):
    if all(map(is_upper, x.split())) and off_shelf_classifier(x) == 'PERSON':
        return 'PERSON'

def is_in_hospital_name_DB(x):
    if x in HOSPITAL_NAMES_DB:
        return 'HOSPITAL'
```
The Snorkel Pipeline

Users write *labeling functions* to generate noisy labels.

Snorkel *models and combines* the noisy labels into probabilities.

The resulting *probabilistic* labels train a model.

**KEY IDEA:** Probabilistic training point carries accuracy. No hand labeled data needed.
“Snorkel DryBell” collaboration with Google Ads. Bach et al. SIGMOD19.

Used in production in many industries, startups, and other tech companies!
Collaboration Highlight: Google + Snorkel

- **Snorkel DryBell** is a production version of Snorkel focused on:
  - Using *organizational knowledge resources* to train ML models
  - Handling *web-scale* data
  - Non-servable to servable feature transfer.

Thank you, Google! (More soon)

[Bach et. al., SIGMOD 2019]
You may have *used something based on it...*

Overton: A Data System for Monitoring and Improving Machine-Learned Products

Christopher Ré
Apple

Feng Niu
Apple

Pallavi Gudipati
Apple

Charles Srisuwananukorn
Apple

---

**Migrating a Privacy-Safe Information Extraction System to a Software 2.0 Design**

Ying Sheng
Google
Mountain View, CA, USA
yingsheng@gmail.com

Nguyen Vo
Google
Mountain View, CA, USA
nguyenvo@gmail.com

James B. Wendt
Google
Mountain View, CA, USA
jwendt@gmail.com

Sandeep Tata
Google
Mountain View, CA, USA
tata@gmail.com

Marc Najork
Google
Mountain View, CA, USA
najork@gmail.com
It has changed use real systems...

A couple of highlights

- Used by multiple teams with good error reduction over production.
- Take away: many systems are almost entirely weak supervision based.

<table>
<thead>
<tr>
<th>Resourcing</th>
<th>Error Reduction</th>
<th>Amount of Weak Supervision</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>65% (2.9×)</td>
<td>80%</td>
</tr>
<tr>
<td>Medium</td>
<td>82% (5.6×)</td>
<td>96%</td>
</tr>
<tr>
<td>Medium</td>
<td>72% (3.6×)</td>
<td>98%</td>
</tr>
<tr>
<td>Low</td>
<td>40% (1.7×)</td>
<td>99%</td>
</tr>
</tbody>
</table>

CIDR2020
Weak Supervision in Science & Medicine

Cross-Modal Weak Supervision

"Indication: Chest pain. Findings: No focal consolidation or pneumothorax."

A. Callahan et al., NPJ Dig Med, 2020

V. Kuleshov et al., Nat Comms, 2019


Blog: http://hazyresearch.stanford.edu/ws4science

Text & Extraction

Imaging & Diagnostics

J. Fries et al., Nat Comms, 2019

J. Dunnmon et al., Radiology, 2019

K. Saab et al., NPJ Dig Med, 2020
Let’s look under the hood and take a peak at some math (to the whiteboard soon..)
Users write labeling functions for multiple related tasks

We model the labeling functions' behavior to de-noise them

We use the probabilistic labels to train a multi-task model

How can we do anything without the ground truth labels?
How to learn the parameters of this model (accuracies & correlations) without $Y$?

Later: We will define what this picture means precisely.
def existing_classifier(x):
    return off_shelf_classifier(x)

def upper_case_existing_classifier(x):
    if all(map(is_upper, x.split())) and \
    off_shelf_classifier(x) == 'PERSON':
        return PERSON

def is_in_hospital_name_DB(x):
    if x in HOSPITAL_NAMES_DB:
        return HOSPITAL

Intuition: Learn from the Overlaps

Key idea: We observe agreements (+1) and disagreements (-1) on many points! (More later!)
Changes how you iterate...
Automated Chest X-ray Triage

Optimizing Workflows with Automated Prioritization, Radiology 19

Cross-Modal Chest X-ray Classification

![ROC Curve]

- Months
- FS, 5K (AUC=0.88)
Cross-Modal Chest X-ray Classification

![ROC Curve](image)

- FS, 50K (AUC=0.95)
- FS, 5K (AUC=0.88)
Applying Weak Supervision Across Modalities

We can leverage data programming across modalities to make weak supervision of complex tasks easier!
Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

Cross-Modal Chest X-ray Classification

def LF_pneumothorax(c):
    if re.search(r'pneumo.*', c.report.text):
        return "ABNORMAL"

def LF_pleural_effusion(c):
    if "pleural effusion" in c.report.text:
        return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.report.words)) > thresh:
        return "NORMAL"

20 Labeling Functions
Months

Cross-Modal Chest X-ray Classification

Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

20 Labeling Functions
Related Work in Weak Supervision


- **Crowdsourcing**: Dawid & Skene 1979, Karger et. al. 2011, Dalvi et. al. 2013, Ruvolo et. al. 2013, Zhang et. al. 2014, Berend & Kontorovich 2014, etc.

- **Co-Training**: Blum & Mitchell 1998

- **Noisy Learning**: Bootkrajang et. al. 2012, Mnih & Hinton 2012, Xiao et. al. 2015, etc.

- **Indirect Supervision**: Clarke et. al. 2010, Guu et. Al. et. al. 2017, etc.


- **Boosting & Ensembling**: Schapire & Freund, Platanios et. al. 2016, etc.

- **Constraint-Based Supervision**: Bilenko et. al. 2004, Koestinger et. al. 2012, Stewart & Ermon 2017, etc.

- **Propensity SVMs**: Joachims 17
More Related work

• So much more! Work was inspired by classics and new Cotraining, GANs, capsule networks, semi-supervised learning, crowd-sourcing and so much more!

• Please see blog for summary. https://www.snorkel.org/blog/weak-supervision