Learning with Less Labels (LwLL)

Wade Shen
Goal

1. Create AI/ML algorithms that can learn with $10^6+$ less labels and adapt to new environments with $<10^2$ labels

2. Find the theoretical limits on learning and adaptation given tasks and data
The labeling problem

1. Need lots of labels

2. Domain change = catastrophic failure

3. Labeling USG data is expensive

4. You need them again!

Crowd Sourcing

<table>
<thead>
<tr>
<th>Task</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object recognition</td>
<td>100+ m objs</td>
</tr>
<tr>
<td>Speech recognition</td>
<td>35k+ hrs (10^10 frames)</td>
</tr>
<tr>
<td>Translation</td>
<td>10^10 words</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Google</th>
<th>USG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available labor</td>
<td>10^5-10^6</td>
<td>100-1000</td>
</tr>
<tr>
<td>Cost of labor</td>
<td>$1.25/hr</td>
<td>$34.00/hr</td>
</tr>
<tr>
<td>Cost to translation model</td>
<td>$41m</td>
<td>$~1bn</td>
</tr>
<tr>
<td>Time to translation model</td>
<td>0.5 yrs</td>
<td>17 yrs</td>
</tr>
</tbody>
</table>

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)
Learning with Less Labels (LwLL)

Today: Costly training and retraining

- Unable to adapt
- Lots of labels to learn
- Lots of labels to adapt

Tomorrow: Learning with less labels

- $10^6+$ reduction in labels
- $10^2$ labels to adapt

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)
LwLL: Supervision via generalized objectives

- Learning boils down to optimization

\[
\arg \min_w E(w) = \frac{1}{N} \sum_{i=1}^{N} (w \cdot x_i - y_i)^2
\]

- Problems and opportunities:

1. Observations aren’t ideal

Nuisance variables

Observation (data)

Ideal observation

2. Learning requires data paired with labels

Labeled (finite and expensive)

Unlabeled (infinite and free)

LwLL will:
1) Automatically find a generalized objective \( J(w) \) that eliminates nuisance and efficiently utilizes unlabeled data
2) Formally define the limits on \( J(w) \) for adaption and training given data and labels
TA1: Autonomously learn what matters

- **Goal:** discover the data that matters

- **Approaches:**
  1. **Automated** transfer learning that discovers similar problems and learns to factor out nuisance variables
  2. Learn what matters via **automated** and optimal experimentation

ML Corpus
- ImageNet
- Pascal2
  - Material
  - Spacenet
  - LDC

### Diagram:
- Automatically find related problem
- Jointly embed related problems
- Nuisance dimension

### Algorithms:
- Hypothesize “Background color doesn’t matter”
- Experiment
- Same?
- Initial data
- Model

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)
TA1: Automatically derive learning objectives

- **Goal:** Learn correlated objective function with small numbers of examples

- **Approaches:**
  1. **Automated** discovery of correlated reward
     - **Goal:** object detection
       - Find correlated reward problem
       - Labeled examples: 104 human + 108 generated 72%
       - Unlabeled examples: 109 human labels 76.4%

  2. **Automatically discover** and learn low-resource generators to train DNNs

     - **Generative Model**
       - Object DNN 64.3%
     - **Discriminative Model**
       - Deep Q Network 81.4%


A. Ratner et al, "Snorkel: Rapid Training Data Creation with Weak Supervision", VLDB 2018
**Goal:** Analytical model of optimal training and adaptation efficiency in terms of labeled data

1. What is the “true” rank of a classification problem? How complex is an ML problem?
2. How much labeled data is required to discover a problem’s true rank?
3. How complex is the data?
4. What are the discoverability limits for optimal manifolds given sampled data?

**Approach:** Formally prove the limits of learning problems

1. Formally define an approximate problem complexity measure
2. Formally characterize ML problems and domain mismatch
3. Prove complexity and transfer limits for problem classes and domains (sample complexity under adaptation, transfer and training)
4. Prove minimum data rank given $M$ observed samples for chosen accuracy $\varepsilon$ (data complexity)
Evaluation and metrics

- **Evaluation design**
  - 12-month release cycle for challenge problems tied to evaluation
  - Problems: images object detection/recognition, video activity recognition, translation

**Task TA1.1**: Train from scratch

**Task TA1.2**: Adapt to domain

**Task TA2.1**: Apply model of problem classes and domain variance to characterize challenge problem

**Task TA2.2**: Apply theoretical manifold bounds to estimate minimal subspace given # labels and $\varepsilon$

**Empirical results**

**Theoretical limits**

**Estimated analytical bounds for initial training and adaptation**

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)
LwLL’s objective is reduction in labels without loss of state of the art performance; all percentages represent current state of the art.

### Challenges

<table>
<thead>
<tr>
<th>Object detection¹</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train: LSVRC (open)</td>
<td>80% @ 10⁵</td>
<td>80% @ 10³</td>
<td>80% @ 10²</td>
<td>80% @ 10²</td>
</tr>
<tr>
<td>Adapt: TBD</td>
<td>Metric: mAP @ # labels</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Object classification²</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train: LSVRC (open)</td>
<td>97% @ 10⁶</td>
<td>97% @ 10³</td>
<td>97% @ 10³</td>
<td>97% @ 10²</td>
</tr>
<tr>
<td>Adapt: TBD</td>
<td>Metric: mAP @ # labels</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activity recognition³</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train: TRECVID MED task</td>
<td></td>
<td></td>
<td>41% @ 10³</td>
<td>41% @ 10²</td>
</tr>
<tr>
<td>Adapt: TBD</td>
<td>Metric: mAP @ # labels</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Machine translation⁴</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
<th>Train (TA1.1)</th>
<th>Adapt (TA1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train: OpenMT task</td>
<td></td>
<td></td>
<td>47% @ 10³</td>
<td>47% @ 10²</td>
</tr>
<tr>
<td>Adapt: TBD</td>
<td>Metric: BLEU @ # labels</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

[¹] LSVRC 2017 best DET system, unlimited data track
[²] LSVRC 2017 best LOC Task 1b system, unlimited data track
[³] TRECVID 2017 best system MED task
[⁴] OpenMT 2015 best system Arabic-English task
TA2 goals

**Phase 1**
- Formally define an approximate problem complexity measure (minimum learning efficiency)
- Formally characterize ML problems and domain mismatch

**Phase 2**
- Find and prove the properties of a data sample $M$ essential to satisfying the approximate Johnson-Lindenstrauss transform for training and adaptation

Prove approximate data complexity bounds (e.g. approximate Johnson-Lindenstrauss $O(e^{-2 \log N})$ given $M \ll N$)

Prove bounds on problem isomorphism for transfer and adaptation (transfer limits)

Prove complexity limits for problem classes and domains (sample complexity)

**GOAL:** Analytical model of optimal training and adaptation efficiency in terms of labeled data

Use model to estimate bounds

Challenge problem #1
Challenge problem #2
Challenge problem #3

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)
**Schedule**

**Execution and phasing**

---

**Phase 1 (18 months)**
- ML corpus construction via D3M augmentation
- Similarity metrics / problem embedding (problem2vec)
- Problem clustering and generalization
- Automated optimal experimentation for active learning
- Reward discovery for reinforcement learning
- Generalized adversarial proxy discovery

**Phase 2 (18 months)**
- Preliminary TA1 toolkit + TA2 publications

---

**TA1: Learn and adapt with less labels**
- Formal problem complexity measures
- Formal characterization problem/domain attributes
- Initial data complexity bounds
- Initial problem complexity bounds

**TA2: Limits of generalized learning**
- Final data complexity bounds
- Final problem complexity bounds

**GovTeam: Annual evaluation problems**
- Image object recognition
- Video activity recognition
- Machine translation

---

**Annual challenge problem evaluation**

**Intermediate integration event**

**End of phase evaluation**

**Preliminary TA1 toolkit + TA2 publications**

**Final TA1 toolkit + TA2 publications**

---

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)