Learning from Samples of Variable Quality

Mostafa Dehghani
Motivation

- Why we need to know how to learn from samples of variable quality?
- Deep neural networks are data hungry!
- For many tasks, the more data you have, the better your model will work!
- This amount of data is not always available for many tasks
  - Unsupervised/Semi-supervised methods
  - Weak/ Noisy Supervision
Learning from weak labels

For a large class of problem, we can reduce the (un/)semi-supervised learning to learning from weak labels using a weak annotator.
Sources of Weak Supervision

- Distant supervision setup
  - a heuristic labeling rule or function which can be relying on external knowledge

- Weaker classifiers
  - trained on e.g. non-expert crowd-sourced data

- Biased classifier
  - trained on e.g. data from different domains that are related
Sources of Weak Supervision

- **Indirect supervision**
  - is employed in particular in the structured learning
  - a companion binary task is defined for which obtaining training data is easier

- **Response-based supervision**
  - the model receives feedback from interacting with an environment in a task, and converts this feedback into a supervision signal to update its parameters
Sources of Weak Supervision

● Constraint-based supervision
  ○ Constraints that are represented as weak label distributions are taken as signals for updating
    the model parameters
  ○ For instance, physics-based constraints on the output, or output constraints on execution of
    logical forms
“All labels are equal, but some labels are more equal than others.”

Inspired by George Orwell, Animal Farm, 1945
Treating weakly-labeled samples uniformly (i.e. each weak sample contributes equally to the final classifier) ignores potentially valuable information of the label quality.
Fidelity-Weighted Learning
General Setup

- We assume we are given:
  - A large set of data samples annotated by a labeling function called the *weak annotator*, called the weak dataset:
    \[ D_w = \{(x_i, \tilde{y}_i)\} \]
  - A small set of high-quality samples labeled by experts, called the strong dataset:
    \[ D_s = \{(x_i, y_i)\} \]

- Note that we can generate a large amount of weak training data almost no cost using the weak annotator. In contrast, we have only a limited amount of observations from the true function.
General Setup

The general setting consisting of two main modules:

- One is called the **student** and is in charge of learning a suitable data representation and performing the main prediction task.
- The other is the **teacher** which modulates the learning process by modeling the inaccuracies in the labels.
The procedure has three main steps:

**Algorithm 1** Fidelity-Weighted Learning.

1. Train the student on samples from the weakly-annotated data $D_w$.
2. Freeze the representation-learning component $\psi(.)$ of the student and train teacher on the strong data $D_s = (\psi(x_j), y_j)$. Apply teacher to unlabeled samples $x_t$ to obtain soft dataset $D_{sw} = \{(x_t, \bar{y}_t)\}$ where $\bar{y}_t = T(x_t)$ is the soft label and for each instance $x_t$, the uncertainty of its label, $\Sigma(x_t)$, is provided by the teacher.
3. Train the student on samples from $D_{sw}$ with SGD and modulate the step-size $\eta_t$ according to the per-sample quality estimated using the teacher.
Step 1

Pre-train the student on weak dataset using weak labels generated by the weak annotator.

The main goals of this step are:

- To learn a **task-dependent** representation of the data:
  - Learning a representation function $\psi(.)$

- To **pretrain** the student.
Step 2

Train the teacher on the strong data represented in terms of the student representation, and then use the teacher to generate a soft dataset consisting of \( \langle \text{sample, predicted label, confidence} \rangle \) for all data samples.

The GP is trained on the samples from Ds to learn the posterior mean (used to generate soft labels) and posterior covariance (which represents label uncertainty).

Soft labels: a new set of (hopefully less weak) labels for which we know the certainty.
Step 3

Fine-tune the weights of the student network on the soft dataset, while modulating the magnitude of each parameter update by the corresponding teacher-confidence in its label.

We update the parameters of the student by training on soft dataset using SGD:

\[
\begin{align*}
    w^* &= \arg\min_{w \in W} \frac{1}{N} \sum_{(x_t, \bar{y}_t) \in D_{sw}} l(w, x_t, \bar{y}_t) + R(w), \\
    w_{t+1} &= w_t - \eta_t (\nabla l(w, x_t, \bar{y}_t) + \nabla R(w))
\end{align*}
\]

where:

\[
\eta_t = \eta_1(t) \eta_2(x_t) \quad \text{and} \quad \eta_2(x_t) = \exp[-\beta \Sigma(x_t)]
\]
Experiments
A one-dimensional toy problem:

- Let $f(x) = \sin(x)$ be the true function from which a small set of observations is provided.
  - These observations might be noisy, in the same way that labels obtained from a human labeler could be noisy.
- A weak annotator function $f(x) = 2\text{sinc}(x)$ is provided, as an approximation to the true function.
- The task is to obtain a good estimate of the true given the small set strong observations and the weak annotator function.
Tox problem

(a) Training student on 100 examples from the weak function.

(b) Fitting teacher based on 10 observations from the true function.

(c) Fine-tuning the student based on observations from the true function.

(d) Fine-tuning the student based on label/confidence from teacher.
Two real world tasks

**Document Ranking**

**Sentiment Classification**

Learning from Samples of Variable Quality
Table 1: Performance of FWL approach and baseline methods for ranking task. \(^*\) indicates that the improvements with respect to the baseline \(i\) are statistically significant at the 0.05 level using the paired two-tailed t-test with Bonferroni correction.

<table>
<thead>
<tr>
<th>Method</th>
<th>Robust04</th>
<th>ClueWeb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>nDCG@20</td>
</tr>
<tr>
<td>1 (WA_{BM25})</td>
<td>0.2503(^*)(^{37})</td>
<td>0.4102(^*)(^{37})</td>
</tr>
<tr>
<td>2 (NN_{W}) (Dehghani et al., 2017c)</td>
<td>0.2702(^*)(^{137})</td>
<td>0.4290(^*)(^{137})</td>
</tr>
<tr>
<td>3 (NN_{S})</td>
<td>0.1790</td>
<td>0.3519</td>
</tr>
<tr>
<td>4 (NN_{S+\rightarrow W})</td>
<td>0.2763(^*)(^{1237})</td>
<td>0.4330(^*)(^{1237})</td>
</tr>
<tr>
<td>5 (NN_{W\rightarrow S})</td>
<td>0.2810(^*)(^{1237})</td>
<td>0.4372(^*)(^{1237})</td>
</tr>
<tr>
<td>6 (NN_{W\rightarrow S})</td>
<td>0.2899(^*)(^{123457})</td>
<td>0.4431(^*)(^{123457})</td>
</tr>
<tr>
<td>7 (FWL_{unsupprep})</td>
<td>0.2211(^*)(^{37})</td>
<td>0.3700(^*)(^{37})</td>
</tr>
<tr>
<td>8 (FWL \setminus \Sigma)</td>
<td>0.2980(^*)(^{123457})</td>
<td>0.4516(^*)(^{123457})</td>
</tr>
<tr>
<td>9 (FWL)</td>
<td><strong>0.3124</strong>(^*)(^{12345678})</td>
<td><strong>0.4607</strong>(^*)(^{12345678})</td>
</tr>
</tbody>
</table>
Table 2: Performance of the proposed FWL approach and baseline methods for sentiment classification task. * indicates that the improvements with respect to the baseline are statistically significant, at the 0.05 level using the paired two-tailed t-test, with Bonferroni correction.

<table>
<thead>
<tr>
<th>Method</th>
<th>SemEval-14</th>
<th>SemEval-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 WA&lt;sub&gt;Lexicon&lt;/sub&gt;</td>
<td>0.5141</td>
<td>0.4471</td>
</tr>
<tr>
<td>2 NN&lt;sub&gt;W&lt;/sub&gt;</td>
<td>0.6719&lt;sup&gt;∗1&lt;/sup&gt;</td>
<td>0.5606&lt;sup&gt;∗1&lt;/sup&gt;</td>
</tr>
<tr>
<td>3 NN&lt;sub&gt;S&lt;/sub&gt;</td>
<td>0.6307&lt;sup&gt;∗1&lt;/sup&gt;</td>
<td>0.5811&lt;sup&gt;∗12&lt;/sup&gt;</td>
</tr>
<tr>
<td>4 NN&lt;sub&gt;S+/W&lt;/sub&gt;</td>
<td>0.7032&lt;sup&gt;∗1237&lt;/sup&gt;</td>
<td>0.6319&lt;sup&gt;∗1237&lt;/sup&gt;</td>
</tr>
<tr>
<td>5 NN&lt;sub&gt;W→S&lt;/sub&gt;</td>
<td>0.7080&lt;sup&gt;∗1237&lt;/sup&gt;</td>
<td>0.6441&lt;sup&gt;∗1237&lt;/sup&gt;</td>
</tr>
<tr>
<td>6 NN&lt;sub&gt;W→S&lt;/sub&gt;</td>
<td>0.7166&lt;sup&gt;∗12347&lt;/sup&gt;</td>
<td>0.6603&lt;sup&gt;∗123457&lt;/sup&gt;</td>
</tr>
<tr>
<td>7 FWL&lt;sub&gt;unsuprep&lt;/sub&gt;</td>
<td>0.6588&lt;sup&gt;∗13&lt;/sup&gt;</td>
<td>0.6954&lt;sup&gt;∗123&lt;/sup&gt;</td>
</tr>
<tr>
<td>8 FWL &lt;sup&gt;Σ&lt;/sup&gt;</td>
<td>0.7202&lt;sup&gt;∗123457&lt;/sup&gt;</td>
<td>0.6590&lt;sup&gt;∗123457&lt;/sup&gt;</td>
</tr>
<tr>
<td>9 FWL</td>
<td>0.7470&lt;sup&gt;∗12345678&lt;/sup&gt;</td>
<td>0.6830&lt;sup&gt;∗12345678&lt;/sup&gt;</td>
</tr>
<tr>
<td>10 SemEval&lt;sup&gt;Best&lt;/sup&gt;</td>
<td>0.7162</td>
<td>0.6618</td>
</tr>
</tbody>
</table>

(Rouvier & Favre, 2016) (Deriu et al., 2016)
Analysis and Discussions
Handling The Bias-variance Trade-off

- Played with toy problem with different values of $\beta$ in three cases:
  - Having 10 observations from the true function (“Toy Data”)
  - Having only 5 observations from the true function (“Toy Data *”)
  - Having $f(x)=x+1$ as the weak function, which is an extremely bad approximator of the true function (“Toy Data **”)

- Also tried different values of $\beta$ on real world tasks.
Handling The Bias-variance Trade-off
A Good Teacher Is Better Than Many Observations

Models trained on different amount weak data.

Models trained on different amount of strong data.
A Good Teacher Is Better Than Many Observations

- We looked at the rate of learning for the student as the amount of training data is varied.
  - The student learns faster when there is a teacher!
  - A **LUPI** (learning using privileged information) perspective to FWL.
In the ranking task, besides BM25 (our weak annotator) we use three other weak annotators:

- Vector space model with binary term occurrence (BTO) weighting schema (weaker than BM25)
- Vector space model with TF-IDF weighting schema (weaker than BM25)
- BM25+RM3 that uses RM3 as the pseudo-relevance feedback method on top of BM25 (better than BM25)
Sensitivity Of The FWL To The Quality Of The Weak Annotator
We can translate the confidence score as how likely including a sample in the training set for the student model improves the performance.

- Rather than using this score as the multiplicative factor in the learning rate, we can use it to bias sampling procedure of mini-batches so that the frequency of training samples are proportional to the confidence score of their labels.
From Modifying The Learning Rate To Weighted Sampling
THANK YOU