Learning Hyper Label Model for Programmatic Weak Supervision

Renzhi Wu[*] Shen-En Chen[*] Jieyu Zhang[&] Xu Chu[*]

[*]Georgia Tech [&]University of Washington

Data is the Bottleneck for ML

ML ≈ Model + Data

Model is gradually commoditized (e.g. transformers for “all” tasks)

Data is the bottleneck

January 20, 2023

OpenAI Outsourced Data Labeling to Kenyan Workers Earning Less than $2 Per Hour: TIME Report

Jaime Hampton

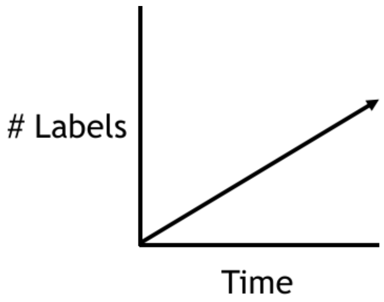
OpenAI has hired an army of contractors

to do what’s called “data labeling”

Sources:
https://www.semafor.com/article/01/27/2023/openai-has-hired-an-army-of-contractors-to-make-basic-coding-obsolete
https://www.datanami.com/2023/01/20/openai-outsourced-data-labeling-to-kenyan-workers-earning-less-than-2-per-hour-time-report/

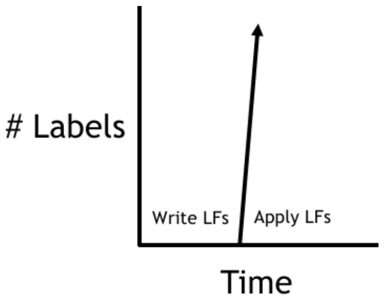
Manual v.s. Programmatic Supervision

Labeling individual data points



Exact Labels

Writing Labeling Functions (LFs) where each LF abstracts a supervision source (e.g. heuristics, existing models, external KBs, …)



Weak Labels

Challenge: noisy and conflicting weak labels from LFs

Label Model

	LF1	LF2	LF3	LFX
Data point 1	1	-1	0	…
Data point 2	0	1	0	…
Data point 3	1	1	-1	..
Data point 4	1	1	-1	…
Data point 5	1	1	-1	…
Data point x	…	…	…	…

Weak label matrix *X*

Label model

y
1
1
-1
1
-1
…

Inferred ground-truth labels *y*

Hyper Label Model

Existing methods (e.g. graphical models) all require ad-hoc parameter learning for each dataset

1. Assume an underlying distribution $\mathbf{p}(y|X, \Theta)$
2. Learn parameter Θ according to some objective function
3. Predict y using $\mathbf{p}(y|X, \Theta)$

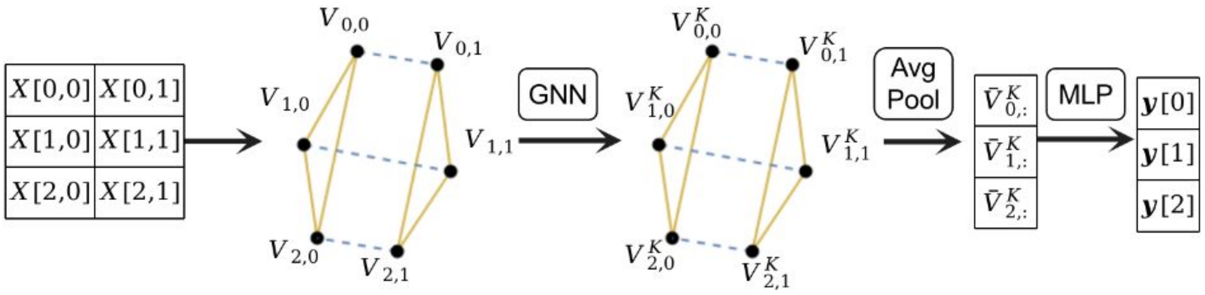
Hyper label model predicts y in a single forward pass $y = \mathrm{net}(X)$
Features: Pre-trained once and works for all datasets. No dataset-specific learning required.

Idea 1: neural network as label model

Weak label matrix X → Neural Network h → labels y

Basic requirements:

1. Arbitrary size of X
2. Switching order of LFs (columns in X) should not affect y
3. Switching order of data points (rows in X), then rows in y should be switched accordingly



Idea 2: learning to be an optimal solution

Pretrain *h* to mimic an optimal solution.

The first analytical optimal solution (of exponential complexity to directly use):

Theorem 1. *$\forall X$, $h^*(X) = \frac{1}{|\mathcal{U}_{\boldsymbol{y}}(X)|} \sum_{\boldsymbol{y} \in \mathcal{U}_{\boldsymbol{y}}(X)} \boldsymbol{y}$ is an optimal estimator of the ground-truth in the sense that it minimizes $\epsilon'(X, h)$.*

Synthetic training data generation: ensure the trained model is asymptotically close to the optimal solution:

Theorem 2. *$\forall X \in \mathcal{D}$, if the corresponding $\boldsymbol{y}$ is uniformly sampled and valid, when $|\mathcal{D}| \to +\infty$, then $\arg\min_h \mathcal{L}(h, \mathcal{D}) \to h^*(X) = \frac{1}{|\mathcal{U}_{\boldsymbol{y}}(X)|} \sum_{\boldsymbol{y} \in \mathcal{U}_{\boldsymbol{y}}(X)} \boldsymbol{y}$.*

Experiments: accuracy and efficiency

1.4 points better

Table 2: Performance (F1 or acc score depending on the dataset) on all datasets

Dataset	Census	IMDB	Yelp	Youtube	SMS	Spouse	CDR	Commercial	Tennis	Basketball	AGNews	TREC	SemEval	ChemProt	AVG.
MV	22.2	**75.0**	**74.4**	80.3	84.0	**51.6**	63.3	**85.9**	85.0	18.9	81.4	49.9	**84.2**	53.7	65.0±0.0
DP	11.1	74.4	71.9	84.5	83.8	50.3	33.9	77.5	**85.1**	17.1	81.7	47.2	73.5	**56.2**	60.6±0.1
FS	17.1	74.5	74.0	83.7	74.4	49.9	69.6	82.5	84.0	17.1	81.3	50.1	23.8	52.4	59.6±0.0
MeTaL	51.1	**75.0**	**74.4**	86.0	57.7	49.9	67.9	83.7	80.9	**19.0**	**82.2**	52.1	**84.2**	52.9	65.5±0.2
NPLM	0.0	55.2	68.3	45.2	0.0	34.3	0.0	76.5	85.0	0.0	81.3	36.5	30.2	48.4	40.1±0.0
DS	0.0	74.4	68.3	45.2	65.0	34.3	0.1	77.8	85.0	17.1	26.6	20.9	73.5	35.1	44.5±0.0
EBCC	0.0	74.4	69.6	45.2	0.0	34.3	8.7	77.5	85.0	17.1	27.8	20.8	30.2	35.0	37.6±0.1
CLL	53.6	72.7	72.0	86.1	**84.2**	50.0	64.9	84.8	83.5	17.5	80.7	59.0	**84.2**	53.1	67.6±0.0
HLM	**56.1**	**75.0**	**74.4**	**91.4**	84.1	**51.6**	**71.0**	83.6	84.3	17.1	81.4	**59.8**	**84.2**	52.3	**69.0**±0.2

6 times faster

Table 3: Running time (seconds) of label aggregation on all datasets

Dataset	Census	IMDB	Yelp	Youtube	SMS	Spouse	CDR	Commercial	Tennis	Basketball	AGNews	TREC	SemEval	ChemProt	AVG.
MV	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
DP	147.8	18.8	40.5	2.5	14.4	8.4	29.5	8.5	10.0	14.9	225.0	100.8	190.2	213.0	73.2
FS	21.1	1.7	3.7	0.2	3.2	0.8	3.7	0.6	0.6	14.9	22.1	16.3	69.0	26.4	12.2
MeTaL	0.5	0.3	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.5	3.6	4.6	3.6	1.2
NPLM	15.7	4.0	5.7	0.4	2.2	1.8	6.3	11.2	1.5	3.4	27.9	5.4	3.4	12.1	7.2
DS	2.4	79.8	116.1	0.2	3.6	0.9	29.7	267.7	4.6	2.1	16.3	78.3	36.6	255.9	63.9
EBCC	3.9	5.1	52.5	2.2	2.8	2.3	5.8	3.0	2.5	6.0	18.0	9.0	9.8	84.8	14.8
CLL	33.7	2.9	6.6	0.5	3.8	1.4	6.0	7.4	1.1	2.0	28.5	12.4	20.5	21.3	10.6
HLM	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.3	0.4	0.2	0.3	0.2	0.2

Summary

1. A hyper label model
   a. Only needs to be pretrained once on synthetic data, works for all datasets
   b. Faster: obtaining inferred labels in one single forward pass
   c. Better
2. Technical innovations:
   a. GNN-based model architecture design
      i. Supporting arbitrary input matrix size
      ii. Invariance/equivariance to permutations of columns/rows in input matrix
   b. The first analytical optimal method
      i. but cannot be directly used due to its exponential complexity
   c. Principled training data generation
      i. The trained model is asymptotically close to the optimal solution