Deep Learning from Crowds

Filipe Rodrigues, Francisco C. Pereira

Dept. of Management Engineering, Technical University of Denmark

AAAI-2018

Contents

Introduction

- Other methods requires aggregating labels from multiple noisy contributors with different levels of expertise.

- An EM algorithm for jointly learning the parameters of the network and the reliabilities of the annotators.

- Crowdlayer allows us to train network directly from the crowd labels, and can internally capture the reliability and biases of annotators.

The EM algorithm for crowds

- Dataset $\mathcal{D}$

$$\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N} \quad \mathbf{y}_n = \{y_n^r\}_{r=1}^{R}$$

$$p(y_n^r | z_n, \Pi^r) = Multinomial(y_n^r | \boldsymbol{\pi}_{z_n}^r) \quad \Pi^r = (\boldsymbol{\pi}_1^r, ..., \boldsymbol{\pi}_C^r)$$

- The complete-data likelihood

$$p(\mathcal{D}, \mathbf{z} | \Theta, \{\Pi^r\}_{r=1}^{R}) = \prod_{n=1}^{N} p(z_n | \mathbf{x}_n, \Theta) \prod_{r=1}^{R} p(y_n^r | z_n, \Pi^r)$$

The EM algorithm for crowds

- E-step

$$\mathbb{E}\left[\ln p\left(\mathcal{D},\mathbf{z}|\Theta,\Pi^1,...,\Pi^R\right)\right] = \sum_{n=1}^{N}\sum_{z_n} q\left(z_n\right)\ln\left(p(z_n|\mathbf{x}_n,\Theta)\prod_{r=1}^{R} p\left(y_n^r|z_n,\Pi^r\right)\right)$$

$$q(z_n=c)\propto p\left(z_n=c|\mathbf{x}_n,\Theta_{\text{old}}\right)\prod_{r=1}^{R} p\left(y_n^r|z_n=c,\Pi_{\text{old}}^r\right)$$

- M-step

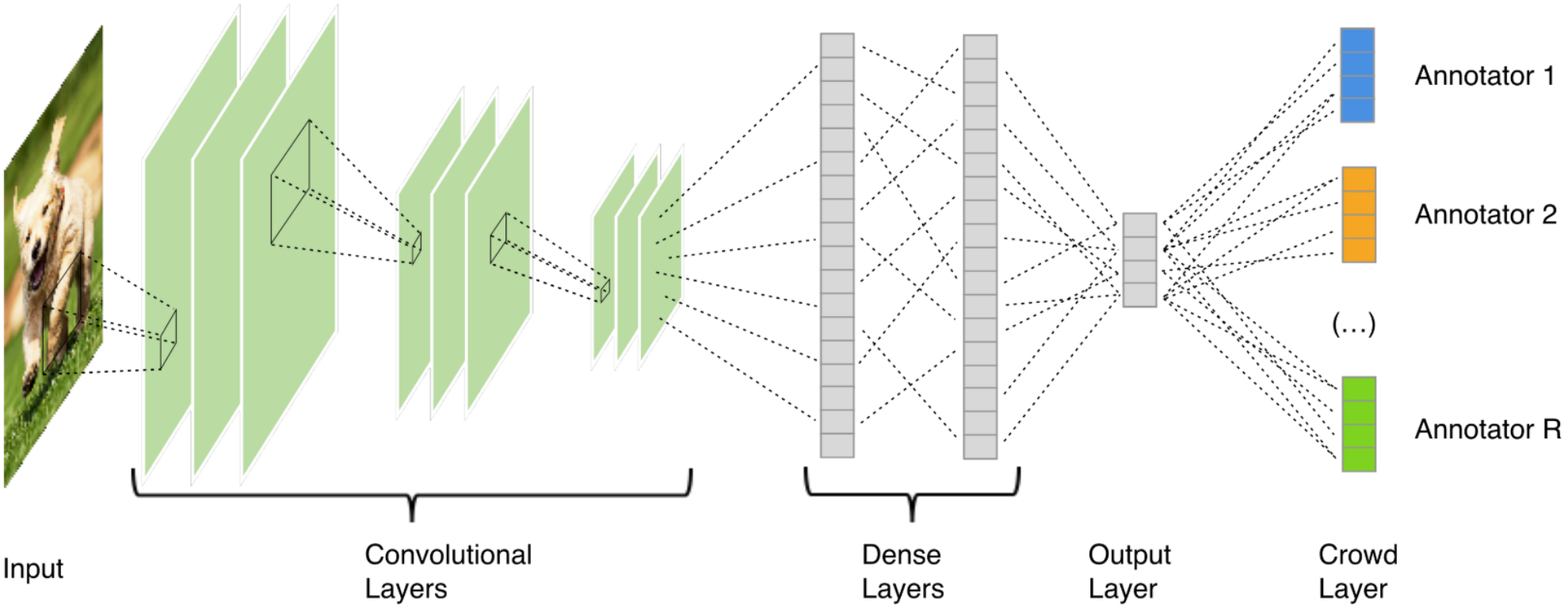
$$\pi_{c,l}^r = \frac{\sum_{n=1}^{N} q\left(z_n=c\right)\mathbb{I}(y_n^r=l)}{\sum_{n=1}^{N} q\left(z_n=c\right)}$$

Θ is updated through the network

The EM algorithm for crowds

- **Difficult to generalize it to regression problems**
  - Rely on the probabilistic interpretation of the softmax output layer
- **Difficult to generalize it to sequence labeling problems**
  - the number of possible label sequences to sum over grows exponentially with the length of the sequence.

Crowd Layer



Bottleneck structure for a CNN for classification with 4 classes and R annotators

Crowd Layer

- The activation of the crowd layer for each annotator r

$$\mathbf{a}^r = f_r(\boldsymbol{\sigma}) \quad f_r(\boldsymbol{\sigma}) = \mathbf{W}^r \boldsymbol{\sigma}$$

$$\frac{\partial E}{\partial \boldsymbol{\sigma}} = \sum_{r=1}^{R} \mathbf{W}^r \frac{\partial E}{\partial \mathbf{a}^r}$$

- Adding parameters beyond necessary can make the output of the bottleneck layer σ lose its interpretability

Experiments

Image classification

Table 1: Accuracy results for classification datasets: Dogs vs. Cats and LabelMe.

Method	Dogs vs. Cats	LabelMe (MTurk)
MA-sLDAc (Rodrigues et al. 2017)	-	78.120 (± 0.397)
DL-MV	71.377 (± 1.123)	76.744 (± 1.208)
DL-DS (Dawid and Skene 1979)	76.750 (± 1.282)	80.792 (± 1.066)
DL-EM (Albarqouni et al. 2016)	80.184 (± 1.454)	82.677 (± 0.981)
DL-DN (Guan et al. 2017)	79.005 (± 1.347)	81.888 (± 1.114)
DL-WDN (Guan et al. 2017)	76.822 (± 2.838)	82.410 (± 0.783)
DL-CL (VW)	79.534 (± 1.064)	81.051 (± 0.899)
DL-CL (VW+B)	79.688 (± 1.406)	81.886 (± 0.893)
DL-CL (MW)	80.265 (± 1.230)	83.151 (± 0.877)
DL-TRUE	84.912 (± 1.248)	90.038 (± 0.652)

Experiments

Image classification

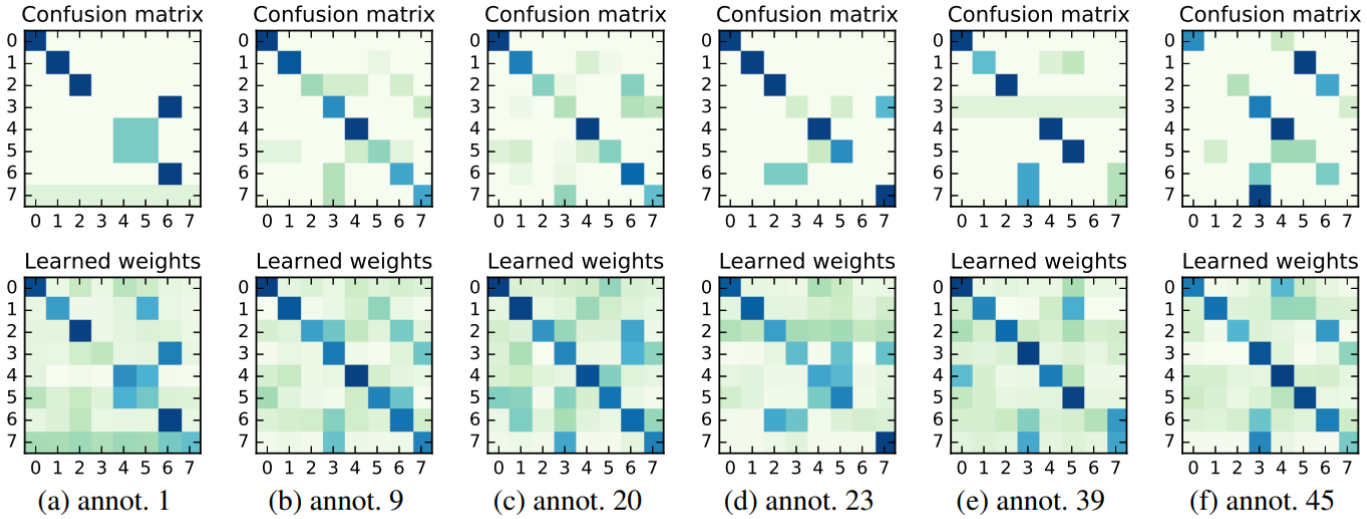
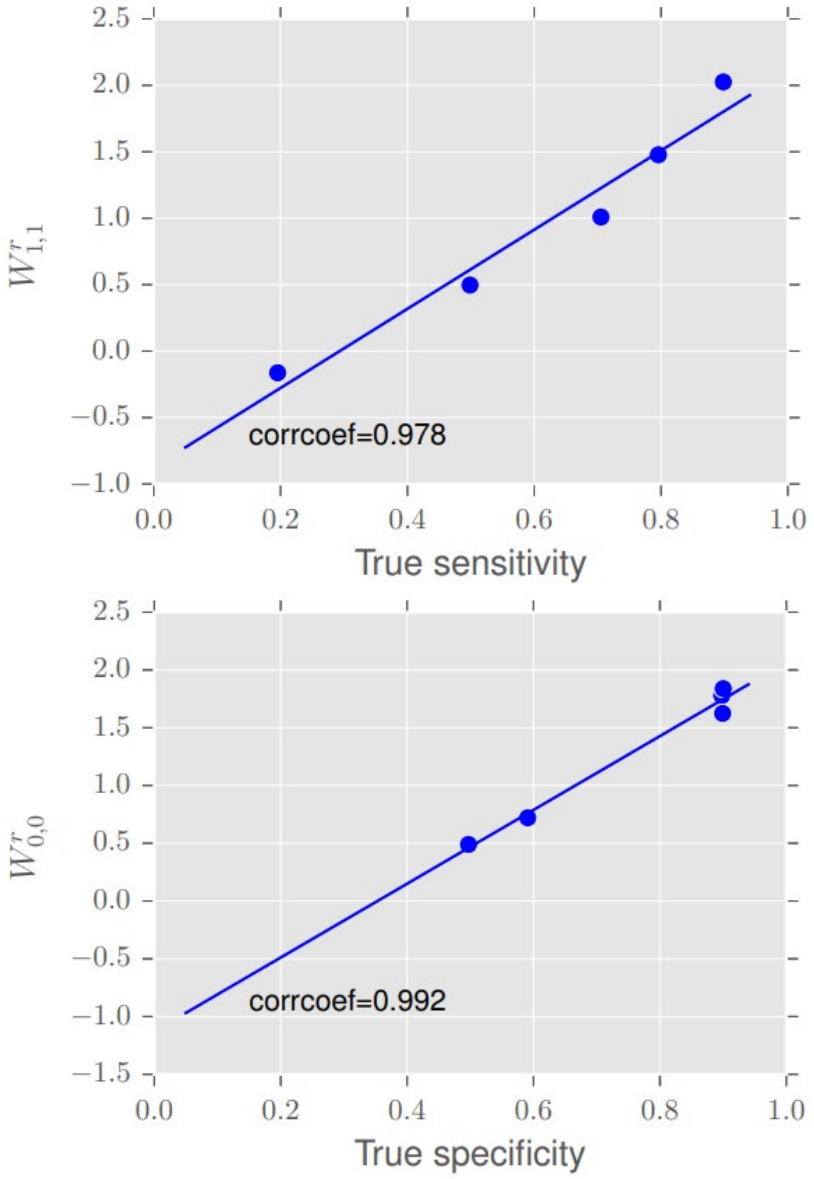


Figure 3: Comparison between the learned weight matrices $\mathbf{W}^r$ and the corresponding true confusion matrices.

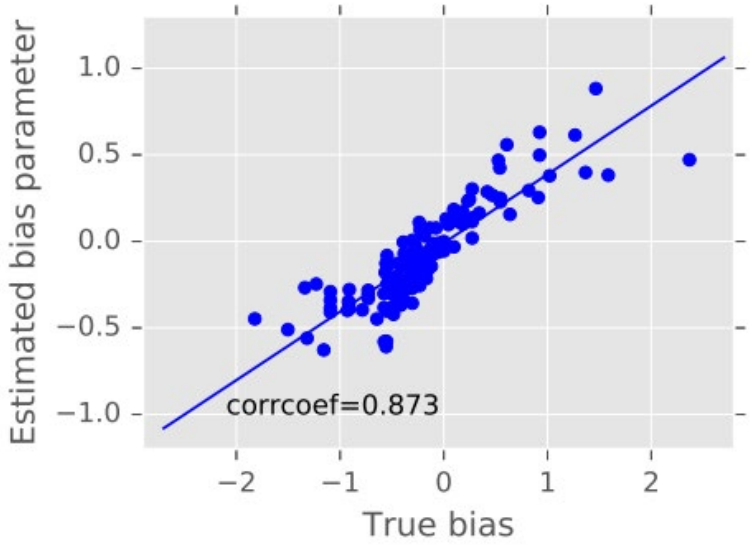


Experiments

Text regression

Table 2: Results for MovieReviews (MTurk) dataset.

Method	MAE	RMSE	R^2
MA-sLDAr (Rodrigues et al. 2017)	-	-	35.553 ($\pm$ 1.282)
DL-MEAN	1.215 ($\pm$ 0.048)	1.498 ($\pm$ 0.050)	31.496 ($\pm$ 4.690)
DL-EM	1.201 ($\pm$ 0.046)	1.482 ($\pm$ 0.048)	32.974 ($\pm$ 4.457)
DL-DN (Guan et al. 2017)	1.270 ($\pm$ 0.021)	1.549 ($\pm$ 0.022)	26.775 ($\pm$ 2.102)
DL-WDN (Guan et al. 2017)	1.261 ($\pm$ 0.016)	1.541 ($\pm$ 0.018)	27.597 ($\pm$ 1.763)
DL-CL (S)	1.228 ($\pm$ 0.041)	1.508 ($\pm$ 0.044)	30.560 ($\pm$ 4.101)
DL-CL (S+B)	1.163 ($\pm$ 0.031)	1.440 ($\pm$ 0.033)	37.086 ($\pm$ 2.407)
DL-CL (B)	1.130 ($\pm$ 0.025)	1.411 ($\pm$ 0.028)	39.276 ($\pm$ 2.374)
DL-TRUE	1.050 ($\pm$ 0.029)	1.330 ($\pm$ 0.036)	45.983 ($\pm$ 2.895)



Experiments

Named entity recognition

Table 3: Results for CoNLL-2003 NER (MTurk) dataset.

Method	Precision	Recall	F1
CRF-MA (Rodrigues, Pereira, and Ribeiro 2013)	0.494	0.856	0.626
DL-MV	0.664 ($\pm$ 0.017)	0.464 ($\pm$ 0.021)	0.546 ($\pm$ 0.014)
DL-EM	0.679 ($\pm$ 0.012)	0.499 ($\pm$ 0.010)	0.575 ($\pm$ 0.008)
DL-DN (Guan et al. 2017)	0.723 ($\pm$ 0.009)	0.459 ($\pm$ 0.014)	0.562 ($\pm$ 0.012)
DL-WDN (Guan et al. 2017)	0.611 ($\pm$ 0.063)	0.480 ($\pm$ 0.058)	0.534 ($\pm$ 0.042)
DL-CL (VW)	0.709 ($\pm$ 0.013)	0.472 ($\pm$ 0.020)	0.566 ($\pm$ 0.016)
DL-CL (VW+B)	0.603 ($\pm$ 0.013)	0.609 ($\pm$ 0.012)	0.606 ($\pm$ 0.007)
DL-CL (MW)	0.660 ($\pm$ 0.018)	0.593 ($\pm$ 0.013)	0.624 ($\pm$ 0.010)
DL-TRUE	0.711 ($\pm$ 0.013)	0.740 ($\pm$ 0.009)	0.725 ($\pm$ 0.008)