Probablistic learning to defer

Handling missing expert's annotations and controlling workload distribution

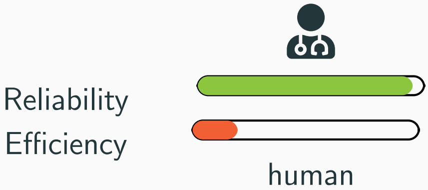
Cuong Nguyen[1] Thanh-Toan Do[2] Gustavo Carneiro[1]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey, UK

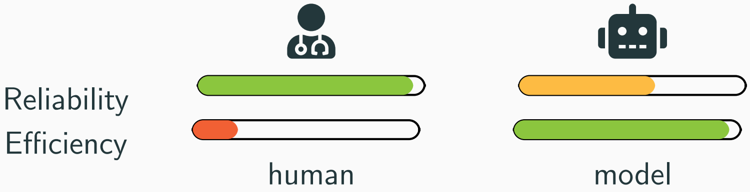
[2]Department of Data Science and AI, Monash University, Australia

Table of contents

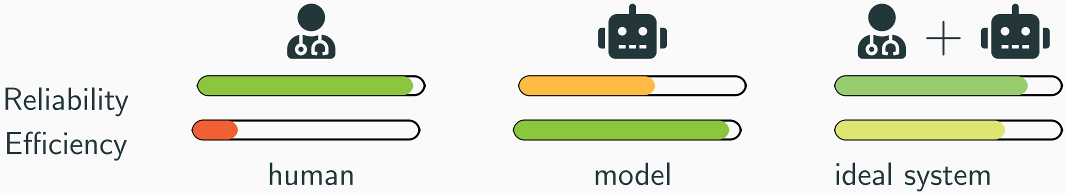
Introduction



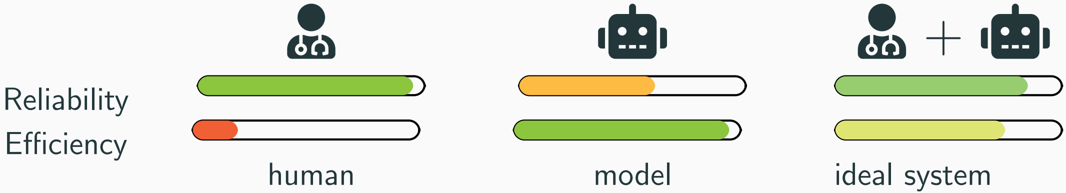
Introduction



Introduction

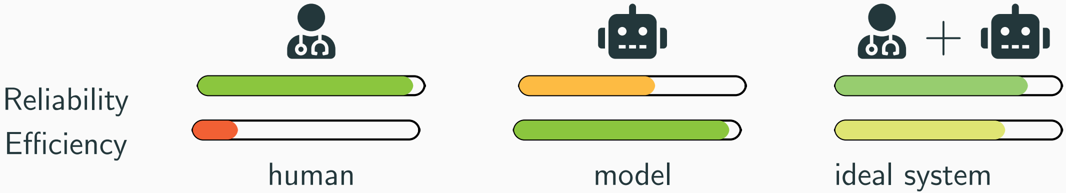


Introduction



Learning to defer (L2D) aims to leverage:

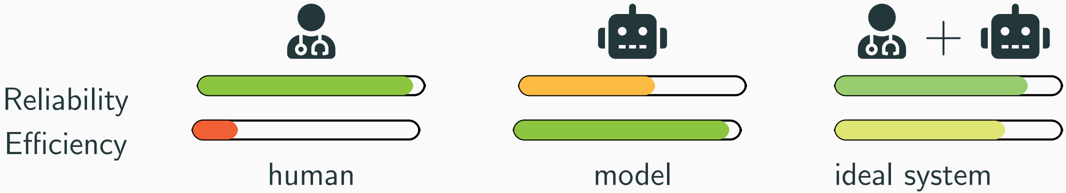
Introduction



Learning to defer (L2D) aims to leverage:

- high *reliability* of human, and

Introduction



Learning to defer (L2D) aims to leverage:

- high *reliability* of human, and
- high *efficiency* of machine learning models.

Background - Learning to defer

L2D can be seen under 2 different models:

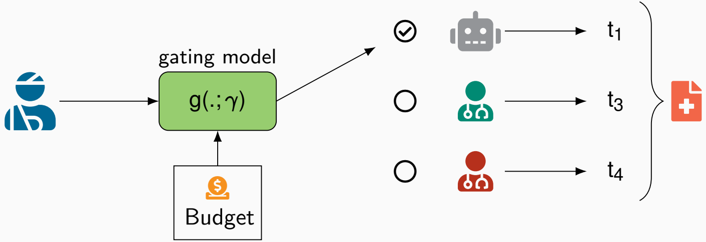
- **Mixture of experts** consists of 2 phases: expert selection and prediction,
- **“Unified” L2D** outputs both expert selection and classifier’s prediction.

Background - Learning to defer

Mixture of experts[1]

The model consists of:

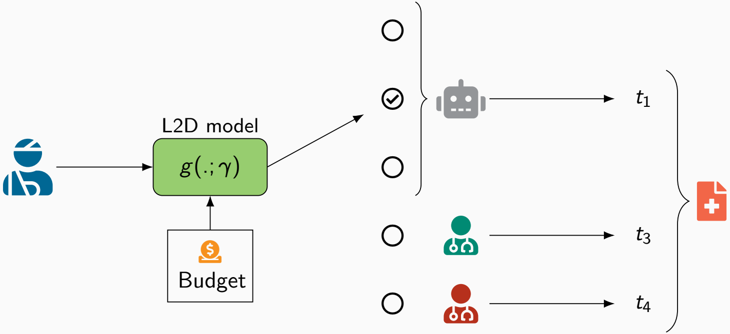
- a gating model,
- M human experts, and
- a classifier.



[1] David Madras, Toni Pitassi, and Richard Zemel. **"Predict responsibly: Improving fairness and accuracy by learning to defer".** In: *Advances in Neural Information Processing Systems*. 2018.

Background - Learning to defer

"Unified" L2D[2] integrates both the classifier's prediction and expert selection into a single model.



[2] Hussein Mozannar and David Sontag. **"Consistent estimators for learning to defer to an expert"**. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7076–7087.

Limitations of current L2D

$$
\underbrace{\begin{array}{c|cccc}
 & & & \dots & \\
\hline
X_1 & t_1^{(1)} & t_1^{(2)} & \dots & t_1^{(M)} \\
X_2 & t_2^{(1)} & t_2^{(2)} & \dots & t_2^{(M)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
X_N & t_N^{(1)} & t_N^{(2)} & \dots & t_N^{(M)}
\end{array}}_{\text{Complete annotation}}
$$

requires *all* human experts must annotate every training sample

Limitations of current L2D

$$\underbrace{\begin{array}{c|cccc} & & & \cdots & \\ \hline X_1 & t_1^{(1)} & t_1^{(2)} & \cdots & t_1^{(M)} \\ X_2 & t_2^{(1)} & t_2^{(2)} & \cdots & t_2^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & t_N^{(1)} & t_N^{(2)} & \cdots & t_N^{(M)} \end{array}}_{\text{Complete annotation}} \quad \rightarrow \quad \underbrace{\begin{array}{c|cccc} & & & \cdots & \\ \hline X_1 & \times & t_1^{(2)} & \cdots & \times \\ X_2 & t_2^{(1)} & \times & \cdots & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & \times & \times & \cdots & t_N^{(M)} \end{array}}_{\text{In-complete annotation}}$$

⚠ requires *all* human experts must annotate every training sample

Limitations of current L2D

$$\underbrace{\begin{array}{c|cccc} & & & \dots & \\ \hline X_1 & t_1^{(1)} & t_1^{(2)} & \dots & t_1^{(M)} \\ X_2 & t_2^{(1)} & t_2^{(2)} & \dots & t_2^{(M)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & t_N^{(1)} & t_N^{(2)} & \dots & t_N^{(M)} \end{array}}_{\text{Complete annotation}} \quad \rightarrow \quad \underbrace{\begin{array}{c|cccc} & & & \dots & \\ \hline X_1 & \times & t_1^{(2)} & \dots & \times \\ X_2 & t_2^{(1)} & \times & \dots & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & \times & \times & \dots & t_N^{(M)} \end{array}}_{\text{In-complete annotation}}$$

⚠ requires *all* human experts must annotate every training sample

- 🚫 impractical (e.g., each sample is annotated by few human experts), and

Limitations of current L2D

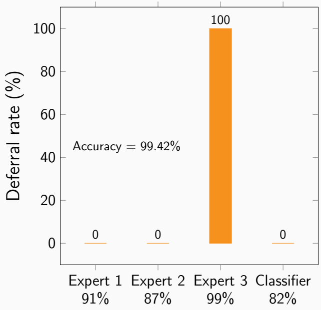
$$
\underbrace{\begin{array}{c|ccccc}
 & & & \cdots & \\
\hline
X_1 & t_1^{(1)} & t_1^{(2)} & \cdots & t_1^{(M)} \\
X_2 & t_2^{(1)} & t_2^{(2)} & \cdots & t_2^{(M)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
X_N & t_N^{(1)} & t_N^{(2)} & \cdots & t_N^{(M)}
\end{array}}_{\text{Complete annotation}}
\quad \rightarrow \quad
\underbrace{\begin{array}{c|ccccc}
 & & & \cdots & \\
\hline
X_1 & \times & t_1^{(2)} & \cdots & \times \\
X_2 & t_2^{(1)} & \times & \cdots & \times \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
X_N & \times & \times & \cdots & t_N^{(M)}
\end{array}}_{\text{In-complete annotation}}
$$

- requires *all* human experts must annotate every training sample
  - impractical (e.g., each sample is annotated by few human experts), and
  - costly, time-consuming, and even infeasible (e.g., radiology[3]),

[3] Leonard Berlin. **"Liability of interpreting too many radiographs"**. In: *American Journal of Roentgenology* 175.1 (2000), pp. 17–22.

Limitations of current L2D (cont.)

- most likely selects the *best human expert* all the time
  - unfair workload assignment, and
  - fatigue, burnout $\rightarrow$ misdiagnosis.

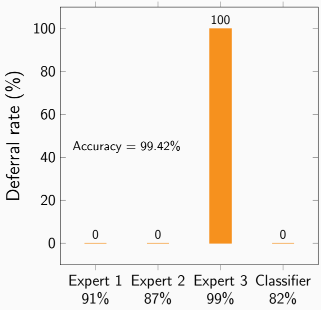


(a) Imbalanced workload

Limitations of current L2D (cont.)

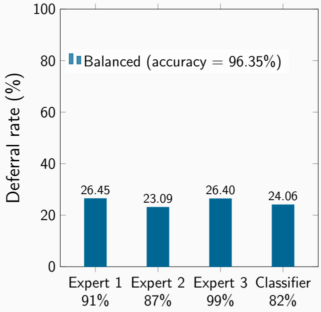
⚠ most likely selects the *best human expert* all the time

- unfair workload assignment, and
- fatigue, burnout $\rightarrow$ misdiagnosis.



(a) Imbalanced workload

$\longrightarrow$



(b) Balanced workload

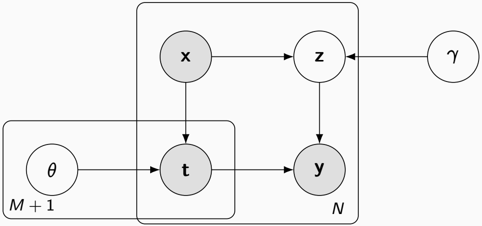
Probabilistic L2D - Complete annotations

Probabilistic L2D is based on the *mixture of experts* modelling approach.

Probabilistic L2D - Complete annotations

Probabilistic L2D is based on the *mixture of experts* modelling approach.

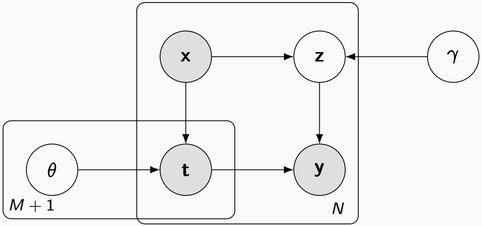
With "complete" annotation data, L2D can be modelled as a graphical model:



Probabilistic L2D - Complete annotations

Probabilistic L2D is based on the *mixture of experts* modelling approach.

With "complete" annotation data, L2D can be modelled as a graphical model:

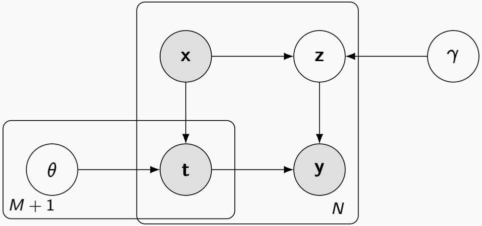


1. draw a sample: $\mathbf{x} \sim \Pr(\mathbf{x})$,

Probabilistic L2D - Complete annotations

Probabilistic L2D is based on the *mixture of experts* modelling approach.

With "complete" annotation data, L2D can be modelled as a graphical model:

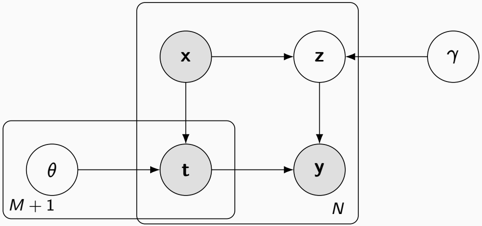


1. draw a sample: $\mathbf{x} \sim \Pr(\mathbf{x})$,
2. draw expert's annotation: $\mathbf{t}^{(m)} \sim \Pr(\mathbf{t}|\mathbf{x}, \theta_m) = \mathrm{Categorical}(\mathbf{t}|f(\mathbf{x}; \theta_m))$,

Probabilistic L2D - Complete annotations

Probabilistic L2D is based on the *mixture of experts* modelling approach.

With "complete" annotation data, L2D can be modelled as a graphical model:

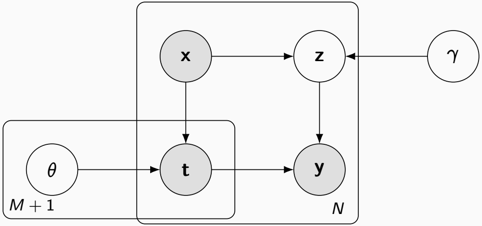


1. draw a sample: $\mathbf{x} \sim \Pr(\mathbf{x})$,
2. draw expert's annotation: $\mathbf{t}^{(m)} \sim \Pr(\mathbf{t}|\mathbf{x}, \theta_m) = \text{Categorical}(\mathbf{t}|f(\mathbf{x}; \theta_m))$,
3. draw an expert: $\mathbf{z} \sim \Pr(\mathbf{z}|\mathbf{x}, \gamma) = \text{Categorical}(\mathbf{z}|g(\mathbf{x}; \gamma))$,

Probabilistic L2D - Complete annotations

Probabilistic L2D is based on the *mixture of experts* modelling approach.

With "complete" annotation data, L2D can be modelled as a graphical model:



1. draw a sample: $\mathbf{x} \sim \Pr(\mathbf{x})$,
2. draw expert's annotation: $\mathbf{t}^{(m)} \sim \Pr(\mathbf{t}|\mathbf{x}, \theta_m) = \mathrm{Categorical}(\mathbf{t}|f(\mathbf{x}; \theta_m))$,
3. draw an expert: $\mathbf{z} \sim \Pr(\mathbf{z}|\mathbf{x}, \gamma) = \mathrm{Categorical}(\mathbf{z}|g(\mathbf{x}; \gamma))$,
4. draw the ground truth: $\mathbf{y} \sim \Pr(\mathbf{y}|\mathbf{z}, \mathbf{t}) = \mathrm{Categorical}(\mathbf{y}|\mathbf{t}^{(z)})$,

Probabilistic L2D - Missing annotations

Probabilistic L2D - Missing annotations

Relax the assumption of annotation availability

			$\dots$	
X_1	$t_1^{(1)}$	$t_1^{(2)}$	$\dots$	$t_1^{(M)}$
X_2	$t_2^{(1)}$	$t_2^{(2)}$	$\dots$	$t_2^{(M)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_N	$t_N^{(1)}$	$t_N^{(2)}$	$\dots$	$t_N^{(M)}$

Complete annotation

$\rightarrow$

			$\dots$	
X_1	$\times$	$t_1^{(2)}$	$\dots$	$\times$
X_2	$t_2^{(1)}$	$\times$	$\dots$	$\times$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_N	$\times$	$\times$	$\dots$	$t_N^{(M)}$

In-complete annotation

Probabilistic L2D - Missing annotations

Relax the assumption of annotation availability

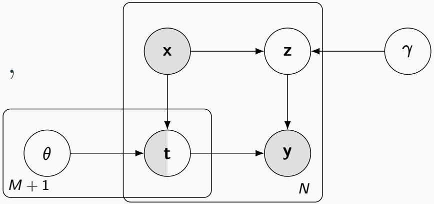
$$
\underbrace{\begin{array}{c|cccc}
 & & & \dots & \\
\hline
X_1 & t_1^{(1)} & t_1^{(2)} & \dots & t_1^{(M)} \\
X_2 & t_2^{(1)} & t_2^{(2)} & \dots & t_2^{(M)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
X_N & t_N^{(1)} & t_N^{(2)} & \dots & t_N^{(M)}
\end{array}}_{\text{Complete annotation}}
\quad \rightarrow \quad
\underbrace{\begin{array}{c|cccc}
 & & & \dots & \\
\hline
X_1 & \times & t_1^{(2)} & \dots & \times \\
X_2 & t_2^{(1)} & \times & \dots & \times \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
X_N & \times & \times & \dots & t_N^{(M)}
\end{array}}_{\text{In-complete annotation}}
$$

$\mathbf{t}_i^{(m)} \forall m \in \{1, \dots, M\}$ is observed. $\rightarrow$ Some $\mathbf{t}_i^{(j)}$ are missing (e.g., latent).

Probabilistic L2D - Missing annotations

Objective

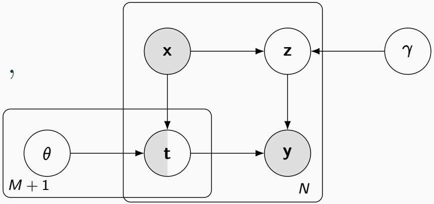
$$\max_{\gamma, \{\theta_m\}_{m=1}^{M+1}} \sum_{i=1}^{N} \ln \Pr \left(\mathbf{y}_i, \prod_{m \in \mathcal{D}_i^{\text{obs.}}} \mathbf{t}_i^{(m)} \middle| \mathbf{x}_i, \gamma, \{\theta_m\}_{m=1}^{M+1} \right),$$



Probabilistic L2D - Missing annotations

Objective

$$\max_{\gamma, \{\theta_m\}_{m=1}^{M+1}} \sum_{i=1}^{N} \ln \Pr \left(\mathbf{y}_i, \prod_{m \in \mathcal{D}_i^{\text{obs.}}} \mathbf{t}_i^{(m)} \middle| \mathbf{x}_i, \gamma, \{\theta_m\}_{m=1}^{M+1} \right),$$

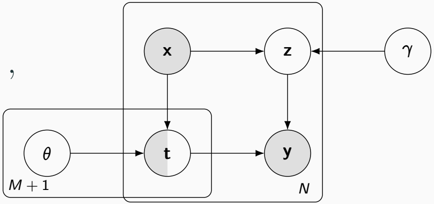


where latent variables are:

Probabilistic L2D - Missing annotations

Objective

$$\max_{\gamma, \{\theta_m\}_{m=1}^{M+1}} \sum_{i=1}^{N} \ln \Pr \left(\mathbf{y}_i, \prod_{m \in \mathcal{D}_i^{\text{obs.}}} \mathbf{t}_i^{(m)} \middle| \mathbf{x}_i, \gamma, \{\theta_m\}_{m=1}^{M+1} \right),$$



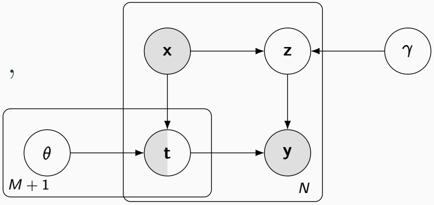
where latent variables are:

- $\mathbf{z}_i$ denoting the r.v. of expert selection,

Probabilistic L2D - Missing annotations

Objective

$$\max_{\gamma, \{\theta_m\}_{m=1}^{M+1}} \sum_{i=1}^{N} \ln \Pr\left(\mathbf{y}_i, \prod_{m \in \mathcal{D}_i^{\text{obs.}}} \mathbf{t}_i^{(m)} \middle| \mathbf{x}_i, \gamma, \{\theta_m\}_{m=1}^{M+1} \right),$$



where latent variables are:

- $\mathbf{z}_i$ denoting the r.v. of expert selection,
- $\prod_{j \in \mathcal{D}_i^{\text{unobs.}}} \mathbf{t}_i^{(j)}$ denoting the r.v. of missing annotations.

Probabilistic L2D - Missing annotations

Parameter inference

Leverage the variational Expectation - Maximisation algorithm:

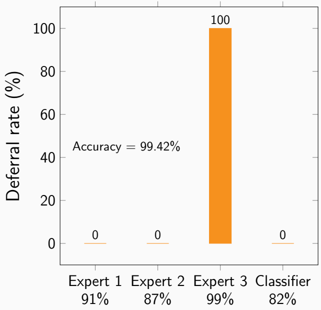
- *E-step:* approximate posterior $q(\mathbf{z}, \prod_{j \in \mathcal{D}_i^{\text{unobs.}}} \mathbf{t}^{(j)})$ via variational inference:

$$q\left(\mathbf{z}, \prod_{j \in \mathcal{D}_i^{\text{unobs.}}} \mathbf{t}^{(j)}\right) = q(\mathbf{z}) \prod_{j \in \mathcal{D}_i^{\text{unobs.}}} q\left(\mathbf{t}^{(j)}\right)$$

$$q^* = \operatorname*{argmin}_q \mathrm{KL}\left[q\left(\mathbf{z}, \prod_{j \in \mathcal{D}_i^{\text{unobs.}}} \mathbf{t}^{(j)}\right) \middle\| \Pr\left(\mathbf{z}_i, \prod_{j \in \mathcal{D}_i^{\text{unobs.}}} \mathbf{t}_i^{(j)} \middle| \mathbf{x}_i, \mathbf{y}_i, \prod_{m \in \mathcal{D}_i^{\text{obs.}}} \mathbf{t}_i^{(m)}, \gamma^{(k)}, \{\theta_m^{(k)}\}_{m=1}^{M+1}\right)\right]$$

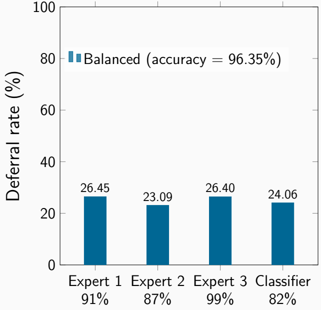
- *M-step:* maximise the "complete"-data log-likelihood w.r.t. $\gamma, \{\theta_m\}_{m=1}^{M+1}$.

Probabilistic L2D - Control workload assignment



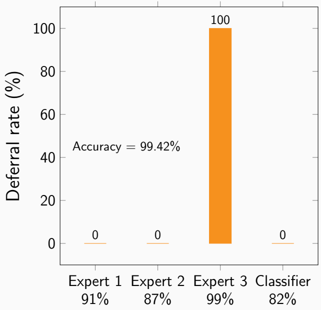
(a) Imbalanced workload

$\longrightarrow$

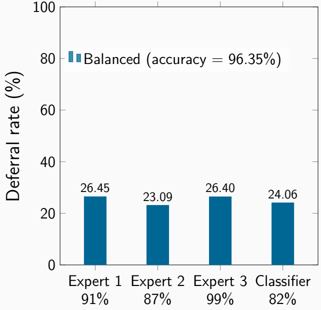


(b) Balanced workload

Probabilistic L2D - Control workload assignment



(a) Imbalanced workload



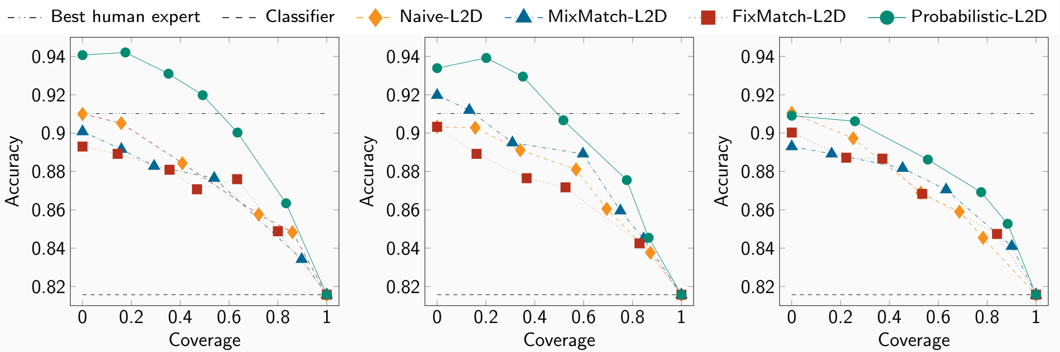
(b) Balanced workload

An additional E-step is introduced to calculate the constrained posterior:

$$\widetilde{q}^* = \operatorname*{argmin}_{\widetilde{q}} \mathrm{KL}\left[\widetilde{q}(\mathbf{z}) \| q^*(\mathbf{z})\right], \forall i \in \{1, \dots, N\} \quad \text{s.t.: } \varepsilon_{\mathrm{l}} \preceq \frac{1}{N} \sum_{i=1}^{N} \widetilde{q}(\mathbf{z}) \preceq \varepsilon_{\mathrm{u}},$$

where q^* and $\widetilde{q}$ denote the unconstrained and constrained posteriors of $\mathbf{z}$.

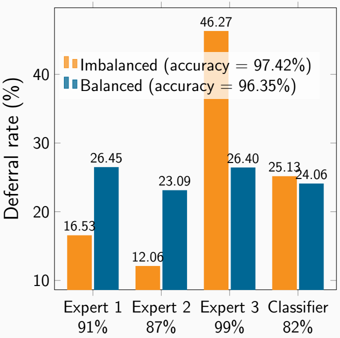
Evaluation - Coverage-accuracy curve on Chaoyang



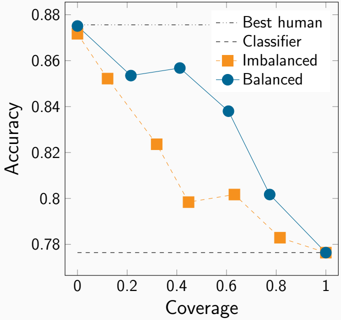
(a) Missing rate of 0.3 **(b)** Missing rate of 0.5 **(c)** Missing rate of 0.7

Figure 3: Comparison of coverage - accuracy curves between different L2D methods on Chaoyang with 2 human experts, each at a different missing rate.

Evaluation - Controllable workload



(a) Chaoyang - coverage of 0.25.



(b) MiceBone - missing rate of 0.3

Figure 4: ((a)) shows comparisons of two different workload constraints on Chaoyang dataset with 50% missing annotations per expert, where in the *imbalanced* setting, $\varepsilon_l = 0$ and $\varepsilon_u = 1$ for each human expert, while in the *balanced* setting, $\varepsilon_l \approx \varepsilon u = (1 - \mathrm{coverage})/M$ for each human expert, and ((b)) coverage - accuracy curve on MiceBone at 30% missing rate.

Limitations

Limitations

- *De-skilling* human by deferring one type of samples,

Limitations

- *De-skilling* human by deferring one type of samples,
- *Scalability* w.r.t. the number of human experts (current is $\mathcal{O}(n_{\text{human}})$)

Limitations

- *De-skilling* human by deferring one type of samples,
- *Scalability* w.r.t. the number of human experts (current is $\mathcal{O}(n_{\text{human}})$)
- *Static performance* assumption on human experts

Limitations

- *De-skilling* human by deferring one type of samples,
- *Scalability* w.r.t. the number of human experts (current is $\mathcal{O}(n_{\text{human}})$)
- *Static performance* assumption on human experts
- *Weak cooperation* between human and machine.

Future work

Future work

- "Reverse" L2D design for education purpose (e.g. train radiologists),

Future work

- "Reverse" L2D design for education purpose (e.g. train radiologists),
- To address the scalability issue:

Future work

- "Reverse" L2D design for education purpose (e.g. train radiologists),
- To address the scalability issue:
    - use a shared model conditioned on "human representation", or

Future work

- "Reverse" L2D design for education purpose (e.g. train radiologists),
- To address the scalability issue:
    - use a shared model conditioned on "human representation", or
    - cluster humans into groups

Future work

- "Reverse" L2D design for education purpose (e.g. train radiologists),
- To address the scalability issue:
    - use a shared model conditioned on "human representation", or
    - cluster humans into groups
- Integrate dynamic performance of human (collaborate with psychologists and radiologists)

Future work

- “Reverse” L2D design for education purpose (e.g. train radiologists),
- To address the scalability issue:
    - use a shared model conditioned on “human representation”, or
    - cluster humans into groups
- Integrate dynamic performance of human (collaborate with psychologists and radiologists)
- Enhance further the human - machine cooperation.