Crowdsourcing label noise simulation on image classification tasks

Tanguy Lefort
IMAG, Univ Montpellier, CNRS
INRIA Montpellier

Benjamin Charlier
IMAG, Univ Montpellier, CNRS
Joseph Salmon
IMAG, Univ Montpellier, CNRS, IUF
Alexis Joly
INRIA Montpellier

Supervised setting: loss $\mathcal{L}$, n_t tasks (x_i, y_i) and predictors family $\mathcal{H}$

$$\arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n_t} \mathcal{L}(f(x_i), y_i)$$

Size of n_t? **The bigger the better...**

- CIFAR-10[1]: $60K$
- MNIST[2]: $70K$
- Pl@ ntNet300K[3]: $+300K$
- ImageNet[4]: $+14.000K$

Each of these needs a label!

---

[1] A. Krizhevsky (2009). *Learning multiple layers of features from tiny images*. Tech. rep.

[2] L. Deng (2012). "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.

[3] C. Garcin et al. (2021). "Pl@ ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution". In: *NeurIPS 2021-35th Conference on Neural Information Processing Systems*.

[4] J. Deng et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Supervised setting: loss $\mathcal{L}$, n_t tasks (x_i, y_i) and predictors family $\mathcal{H}$

$$\arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n_t} \mathcal{L}(f(x_i), y_i)$$

Size of n_t? **The bigger the better...**

- CIFAR-10[1]: $60K$
- MNIST[2]: $70K$
- Pl@ ntNet300K[3]: $+300K$
- ImageNet[4]: $+14.000K$

Each of these needs a label! $\Longrightarrow$ Put humans back in the loop

(1) A. Krizhevsky (2009). *Learning multiple layers of features from tiny images*. Tech. rep.

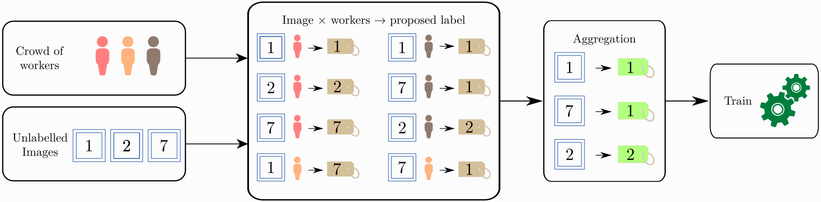
(2) L. Deng (2012). "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.

(3) C. Garcin et al. (2021). "Pl@ ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution". In: *NeurIPS 2021-35th Conference on Neural Information Processing Systems*.

(4) J. Deng et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Crowdsourcing labels

- Participative labelling (CIFAR-10, eyewire[5], Pl@ntNet,...)



(5) `https://eyewire.org/explore`

Problems

- **Label noise**
  - Multiple non-experts workers: who do we trust?
  - How do we aggregate the labels?

(6) J. Peterson, R. Battleday, and T. G. O. Russakovsky (2019). "Human Uncertainty Makes Classification More Robust". In: ICCV, pp. 9617–9626.

(7) https://www.mturk.com/

- **Label noise**
  - Multiple non-experts workers: who do we trust?
  - How do we aggregate the labels?
- **Data access**
  - Only a few such datasets are available freely (CIFAR-10H[6]): Need a crowdsourced data simulator

(6) J. Peterson, R. Battleday, and T. G. O. Russakovsky (2019). "Human Uncertainty Makes Classification More Robust". In: ICCV, pp. 9617–9626.

(7) https://www.mturk.com/

Problems

- **Label noise**
  - Multiple non-experts workers: who do we trust?
  - How do we aggregate the labels?
- **Data access**
  - Only a few such datasets are available freely (CIFAR-10H[6]): Need a crowdsourced data simulator
- **Ethics**
  - invisible and underpaid workers, blurry rights with the law (Amazon Mechanical Turk[7]),
  - Weigh people answers with very little information on them.

[6] J. Peterson, R. Battleday, and T. G. O. Russakovsky (2019). "Human Uncertainty Makes Classification More Robust". In: ICCV, pp. 9617–9626.
[7] https://www.mturk.com/

- **Images** belonging to classes (*e.g.* colors)
- **Workers** with **different abilities**

(8) Y. Qin et al. (2019). "A Multi-class Classification Algorithm Based on Hypercube". In: 2019 IEEE DDCLS, pp. 406–409.

What do we need to simulate?

- **Images** belonging to classes (*e.g.* colors)
- **Workers** with **different abilities**
- Consider that the **difficulty** also comes from the task!

[8] Y. Qin et al. (2019). "A Multi-class Classification Algorithm Based on Hypercube". In: 2019 IEEE DDCLS, pp. 406–409.

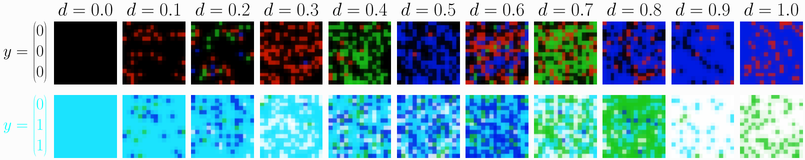
What do we need to simulate?

- **Images** belonging to classes (*e.g.* colors)
- **Workers** with **different abilities**
- Consider that the **difficulty** also comes from the task!

Simulator: Tasks as simple visual experiments

- Simple tasks: RGB images
- Labels $\mathcal{Y} = \{0, 1\}^3$ (vertices of unit hypercube[8])
- Each vertex has 3 neighbors *e.g.* $\mathcal{N}_{(1,0,0)} = \{(1, 1, 0), (0, 0, 0), (1, 0, 1)\}$

[8] Y. Qin et al. (2019). "A Multi-class Classification Algorithm Based on Hypercube". In: 2019 IEEE DDCLS, pp. 406–409.



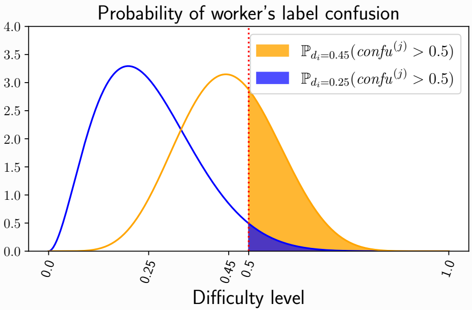
Simulation process

- Start with image of color y and difficulty d
- Sample a distribution over neighbors: $\nu_{\cdot|y} \sim \mathcal{D}irichlet(1/3, 1/3, 1/3)$
- Switch each pixel with probability d and color

$$\arg\max_{y' \in \mathcal{N}_y} \mathcal{D}irichlet(\nu_{\cdot|y})$$

Ask them 2 questions

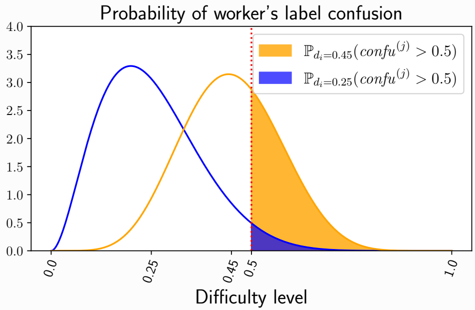
- Is y the true label?
- If not: which color from $\mathcal{N}_y$ is it?



Simulating worker responses

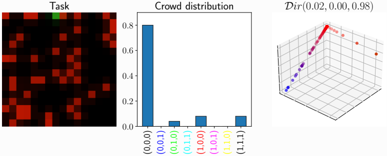
Ask them 2 questions

- Is y the true label?
- If not: which color from $\mathcal{N}_y$ is it?
- Confusion in $[0, 1]$: use Beta distributions (flexible and parametrized) with mean d (step 1) or $(1 - d)\nu_{\cdot|y}$ (step 2)
- Capability as variance levels: $\sigma^{(j)}_{y\leftrightarrow y'}$

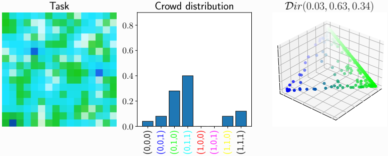


Examples of generated tasks

- 25 workers
- including 6 spammers (answer any label uniformly)



$d_i = 0.2$



$d_i = 0.45$

Difficulty impacts workers' consensus

- We want to remove spammers without removing lower-able workers

Raykar spam score [9]

Let $\hat{u}_j = \arg\min \|\pi^{(j)} - \mathbb{1}_K^\top u_j\|_F^2$ with $\mathbb{1}_K^\top u_j = 1$ and $\mathbb{1}_K = (1, \dots, 1)^\top \in \mathbb{R}^K$:

$$s^{(j)} = \|\pi^{(j)} - \mathbb{1}_K \hat{u}_j^\top\|_F^2 = \frac{1}{K(K-1)} \sum_{c<c'} \sum_{k\in[K]} \left(\pi_{ck}^{(j)} - \pi_{c'k}^{(j)}\right)^2$$

(9) V. Raykar and S. Yu (2012). “Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks”. In: *J. Mach. Learn. Res.* 13, pp. 491–518.

(10) A. Dawid and A. Skene (1979). “Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm”. In: *J. R. Stat. Soc. Ser. C. Appl. Stat.* 28.1, pp. 20–28.

- We want to remove spammers without removing lower-able workers

Raykar spam score[9]

Let $\hat{u}_j = \arg\min \|\pi^{(j)} - \mathbb{1}_K^\top u_j\|_F^2$ with $\mathbb{1}_K^\top u_j = 1$ and $\mathbb{1}_K = (1, \ldots, 1)^\top \in \mathbb{R}^K$:

$$s^{(j)} = \|\pi^{(j)} - \mathbb{1}_K \hat{u}_j^\top\|_F^2 = \frac{1}{K(K-1)} \sum_{c<c'} \sum_{k\in[K]} \left(\pi_{ck}^{(j)} - \pi_{c'k}^{(j)}\right)^2$$

- Use Dawid ans Skene model[10] to get workers' confusion matrices *ie* maximize the likelihood:

$$\prod_{i\in[n_t]} \prod_{k\in[K]} \left\{ \rho_k \prod_{j\in[n_w]} \prod_{\ell\in[K]} \pi_{k\ell}^{(j)} \right\}^{\mathbb{1}_{\{y_i=k\}}}$$

[9] V. Raykar and S. Yu (2012). "Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks". In: *J. Mach. Learn. Res.* 13, pp. 491–518.

[10] A. Dawid and A. Skene (1979). "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *J. R. Stat. Soc. Ser. C. Appl. Stat.* 28.1, pp. 20–28.

Get $s^{(j)}$ and then split spammers / non-spammers using k-means ($k = 2$)

- Simulated crowd: **100 workers** with **88 spammers**
- Logistic regression with $n_t = 500$ and $d_i \in (0, 0.6)$, learning with **smooth labels**

	With spam	Without spam
Accuracy	0.19	**0.81**

Results CIFAR-10H

- CIFAR-10H: out of 2571 workers, only 19 spammers (very curated dataset $\Longrightarrow$ incentives given can temper results)

What's next?

- Difficulty d_i in non-simulated tasks: how can we retrieve it? (theoretically and in practice $\Longrightarrow$ gain of time for experts)
- Introduce the task difficulty in the aggregation process
- Use Pl@ntNet data ($K \gg$, very imbalanced number of answers, very imbalanced number of tasks per class)