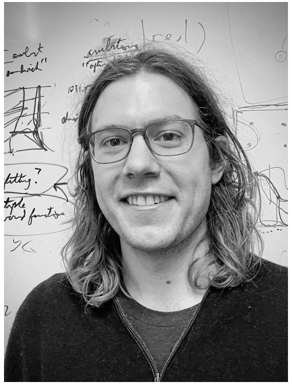
Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics

Shoaib Ahmed Siddiqui [1] **Nitarshan Rajkumar** [1] **Tegan Maharaj** [2] **David Krueger** [1] **Sara Hooker** [3]




UNIVERSITY OF CAMBRIDGE

UNIVERSITY OF TORONTO

Cohere For AI

Problem motivation

Standard benchmark datasets are validated by dataset curators

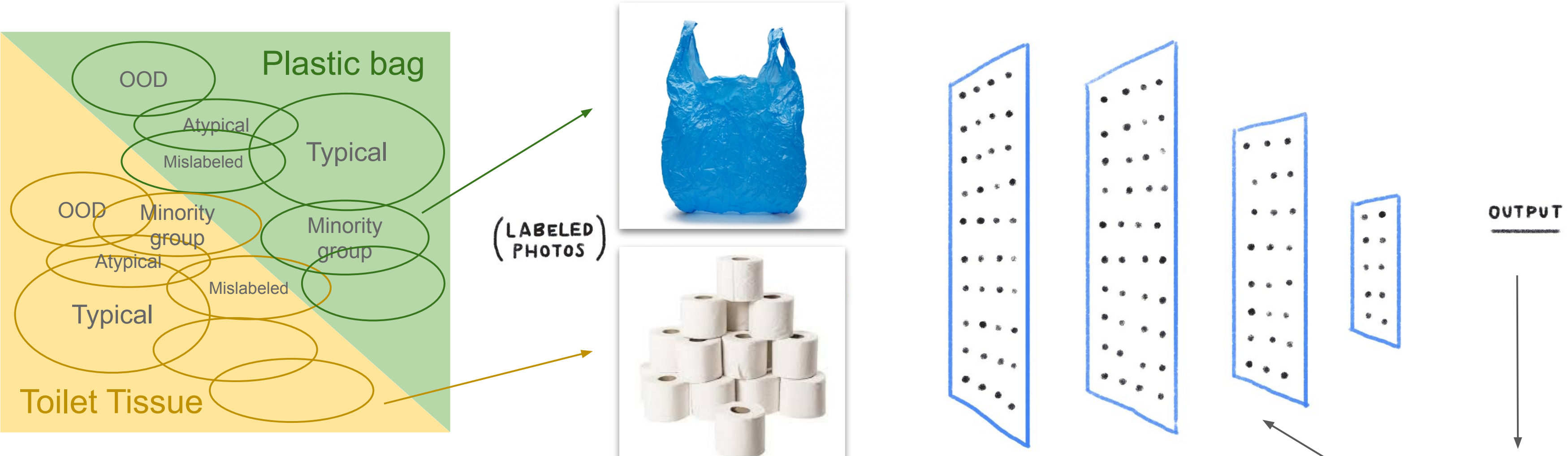
- Dataset curators spend a significant amount of time looking at the dataset for potential issues
- Dataset validation is assumed to be the responsibility of the dataset curator

How can we validate internet-scale data?

- Mostly used for training *"foundation models"*
- Data is randomly crawled from the internet
- Billions of examples
- Data quality is low
- Manual inspection / validation is simply impossible
- ***Can we leverage our models to analyze and audit large-scale datasets?***

Concrete example of classification

Conventional classification setups are indifferent to different sub-groups present in a dataset



$$\mathcal{D} := \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$$

Input-output pairs

Loss Function
L(f(x; W), y)

Understanding these sub-groups is essential to understanding your data

Typical

Corrupted

Toilet Tissue

Atypical

Multi-Label / Mislabeled

We are interested in inferring hidden metadata representing these sub-groups

- This metadata is ***relational*** in nature i.e. considers the whole population rather than a single instance
- An example of such metadata can be: whether an example is **typical vs. atypical, clean vs. mislabeled, in-distribution vs. out-of-distribution, majority vs. minority group** etc.

$$\mathcal{D}_m := \{(x_1, y_1, m), \ldots, (x_k, y_k, m)\}$$

Input-output pairs w/ additional hidden metadata

A bird's eye view of prior work

Prior work

- Prior work provides siloed treatment of these metadata properties
- Only ranks an example along one axis i.e.
    - Typical vs. Atypical [1]
    - Clean vs. Mislabeled [2]
    - In distribution vs. Out-of-Distribution [3]
    - Majority vs. Minority group [4]
- We are interested in a consolidated framework to deal with all these metadata categories simultaneously

[1] Jiang, Z., Zhang, C., Talwar, K. and Mozer, M.C., 2020. Characterizing structural regularities of labeled data in overparameterized models. arXiv preprint arXiv:2002.03206.
[2] Arazo, E., Ortego, D., Albert, P., O'Connor, N. and McGuinness, K., 2019, May. Unsupervised label noise modeling and loss correction. In International conference on machine learning (pp. 312-321). PMLR.
[3] Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J. and Song, D., 2019. Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132.
[4] Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P. and Finn, C., 2021, July. Just train twice: Improving group robustness without training group information. In International Conference on Machine Learning (pp. 6781-6792). PMLR.

Our approach

We coin the term ***Metadata Archaeology***

Metadata Archaeology refers to the task of inferring characteristics of different data subsets

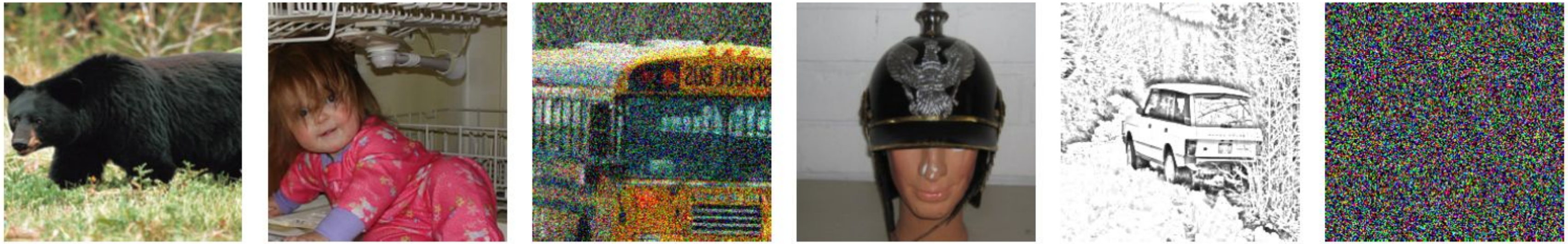
- How can we perform metadata archaeology?

Metadata Archaeology via Probe Dynamics (MAP-D)

- We posit that such metadata can be discovered via monitoring the network's loss throughout the training process on each individual example
- Loss function naturally takes into account both the data population as well as the target label present in the dataset
- This makes loss values suitable for relational metadata
- How to convert loss trajectories into metadata categories? ***Probe suites***

Probe Suites

- Curate a very small set of examples where this metadata is known
- Enables users to focus on properties that they are interested in surfacing
- We simulate this metadata using automated techniques
    - This can be done by a human annotator for a very small number of examples
- We define simple probes such as typical, atypical, out-of-distribution, mislabeled etc. using automated curation or scoring techniques



(a) Typical (b) Atypical (c) Corrupted (d) Rand Label (e) OOD (f) Rand Input

Metadata Archaeology via Probe Dynamics (MAP-D)

- Compute the trajectory of loss values for each of the examples in the dataset as well as probe categories
- Enables us to define a k-Nearest Neighbor (k-NN) classifier using the loss trajectories from the different probe categories

$$\mathbf{s}_i^t := (\ell(x_i, y_i; \theta_1), \ell(x_i, y_i; \theta_2), \ldots, \ell(x_i, y_i; \theta_t) \mid (x_i, y_i) \in \mathcal{D})$$

$$\mathbf{g}_j^t(m) := (\ell(x_j, y_j; \theta_1), \ell(x_j, y_j; \theta_2), \ldots, \ell(x_j, y_j; \theta_t) \mid (x_j, y_j) \in \mathcal{D}_m)$$

$$\mathcal{D}_\mathbf{g} := \Big((\mathbf{g}_1^t(m_1), m_1), \ldots, (\mathbf{g}_{|m_1|}^t(m_1), m_1), (\mathbf{g}_1^t(m_2), m_2), \ldots, (\mathbf{g}_{|m_{|\mathcal{M}|}|}^t(m_{|\mathcal{M}|}), m_{|\mathcal{M}|}) \Big)$$

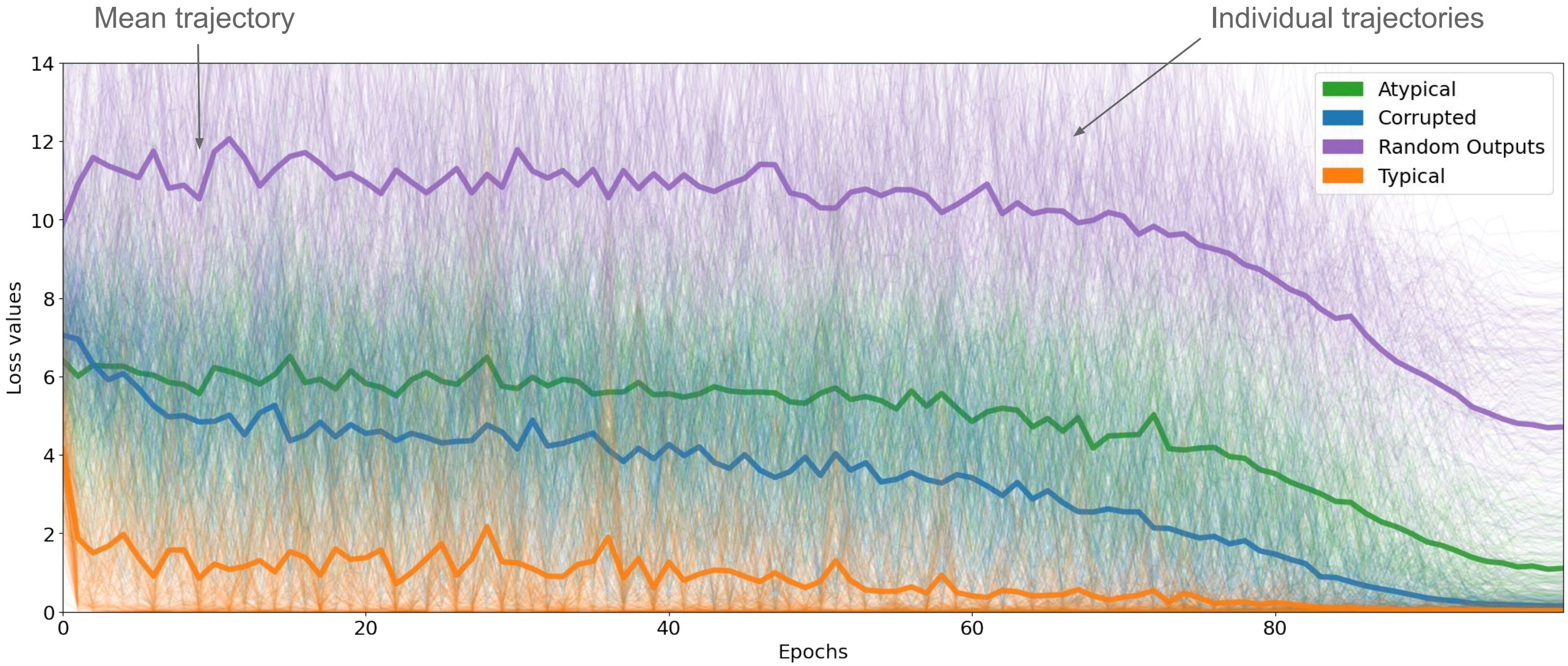
Step # 1: Compute the loss trajectory over probe categories as well as other examples in the dataset

$$p(m \mid \mathbf{s}_i^t) = \frac{1}{k} \sum_{(\mathbf{g}, \hat{m}) \in \mathrm{NN}(\mathbf{s}_i^t, \mathcal{D}_\mathbf{g}, k)} \mathbb{1}_{\hat{m}=m}$$

$$m_i' = \arg\max_{m \in \mathcal{M}} p(m \mid \mathbf{s}_i^t)$$

Step # 2: Use k-NN to assign metadata based on nearest neighbors from the probe suite (loss trajectories)

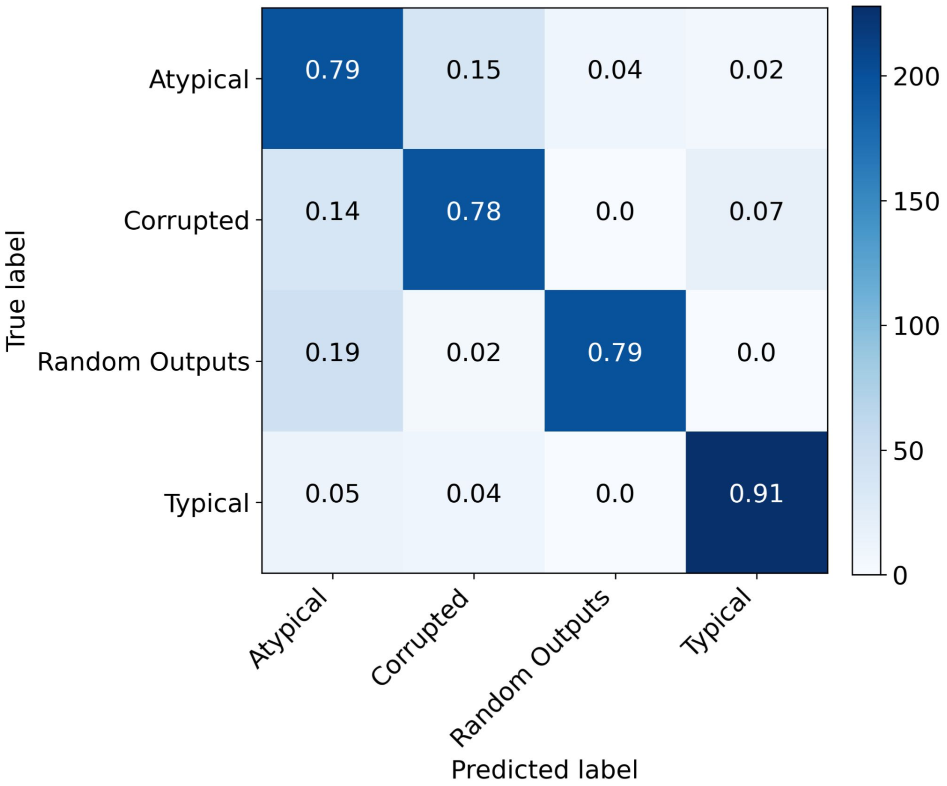
Probe categories have distinct loss profiles



Loss trajectories computed using ResNet-50 on ImageNet

Validating MAP-D

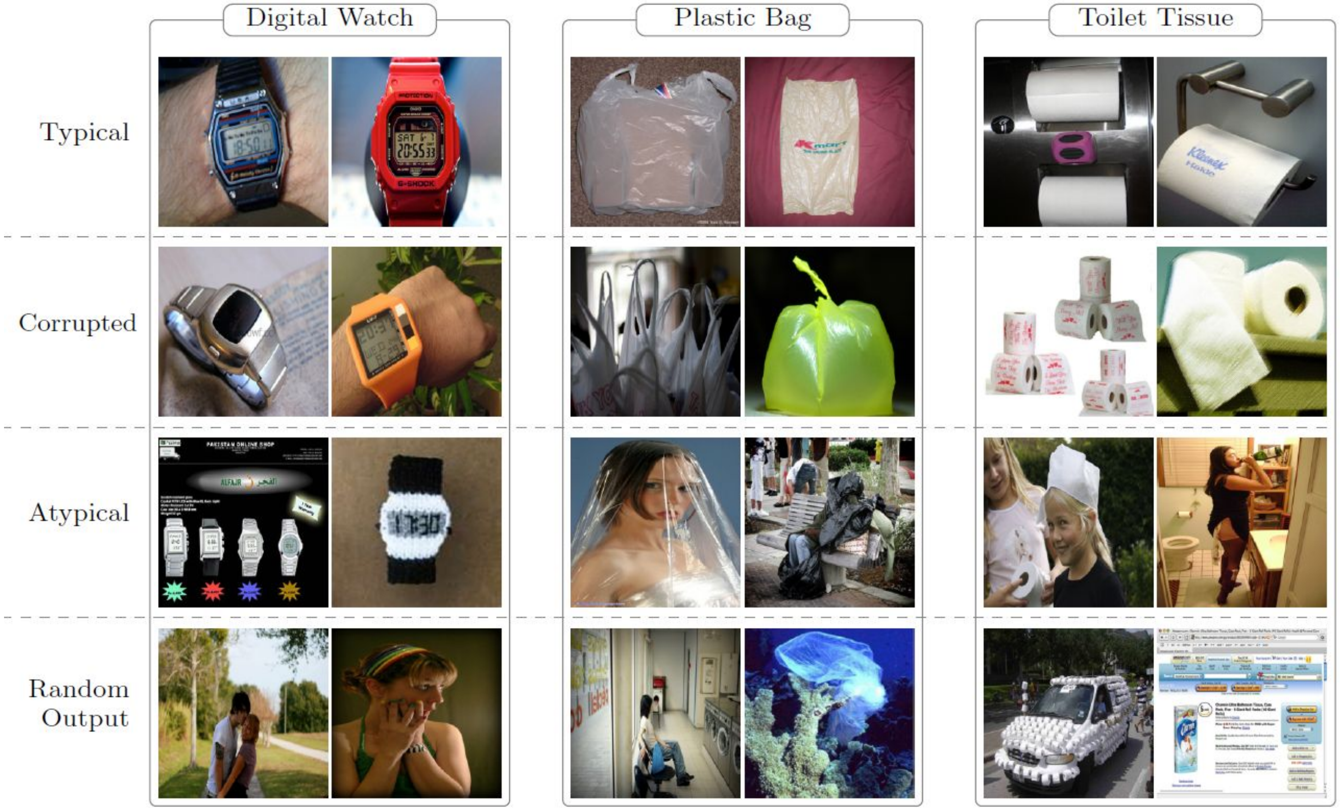
- Validate the metadata assignment on a probe test set
- True underlying metadata is known (not used for training)
- MAP-D achieves high accuracy on this test set, highlighting the effectiveness of this approach



Results

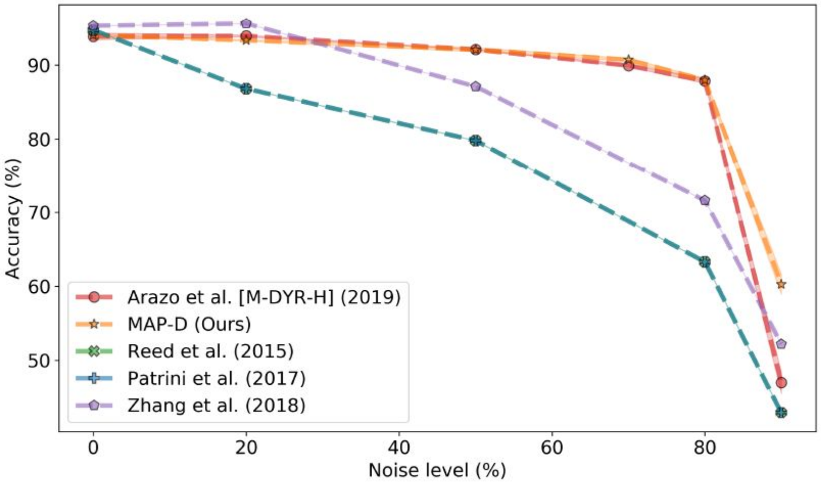
Application: Surfacing interesting examples from the dataset

Examples surfaced from ImageNet training set when defining 4 different probe categories

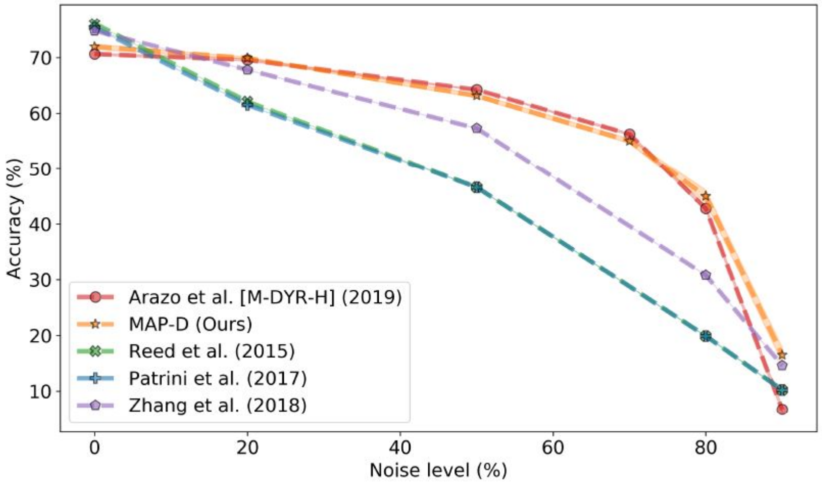


Application: Label correction

- MAP-D is on-par with more sophisticated approaches developed specifically for label noise correction
- Online construction of loss trajectories

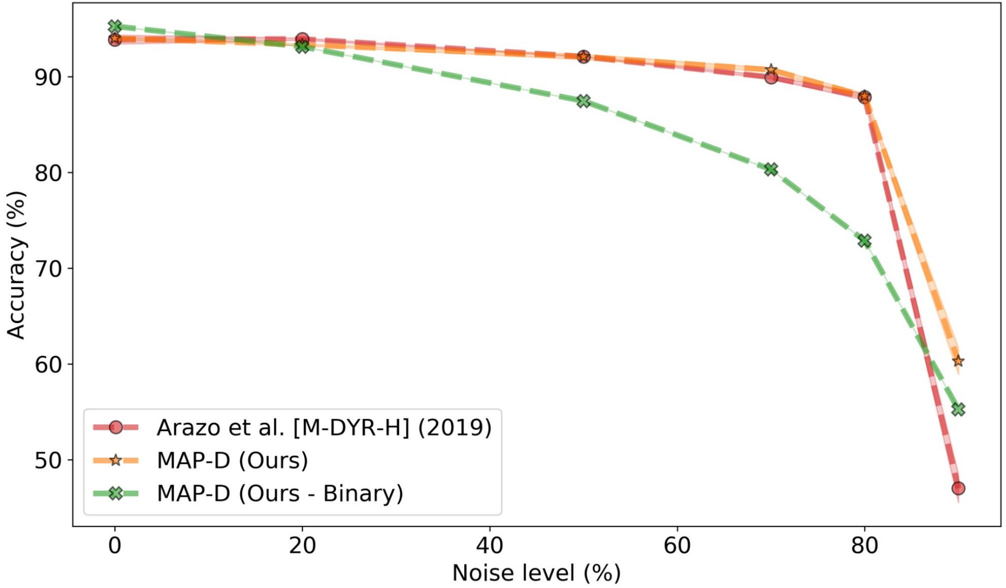


(a) CIFAR-10



(b) CIFAR-100

Validity of uncertainty estimates from MAP-D

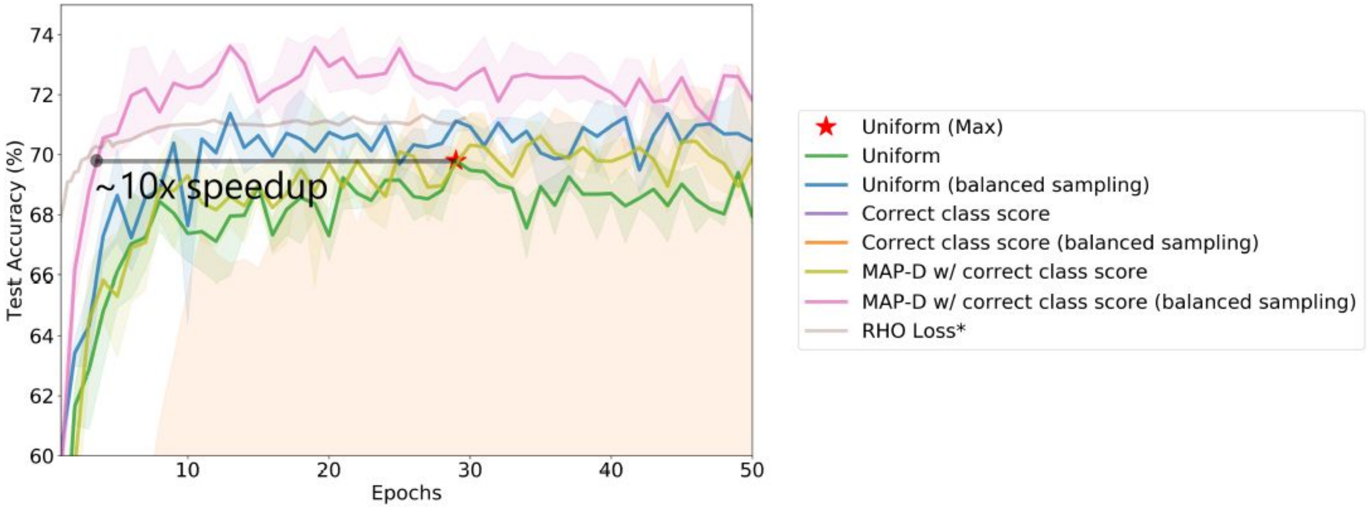


Ablation for label correction on CIFAR-10, where we use a binary prediction instead of probability estimates returned by MAP-D. This highlights the utility and effectiveness of the uncertainty estimates computed by MAP-D.

Application: Prioritized Training

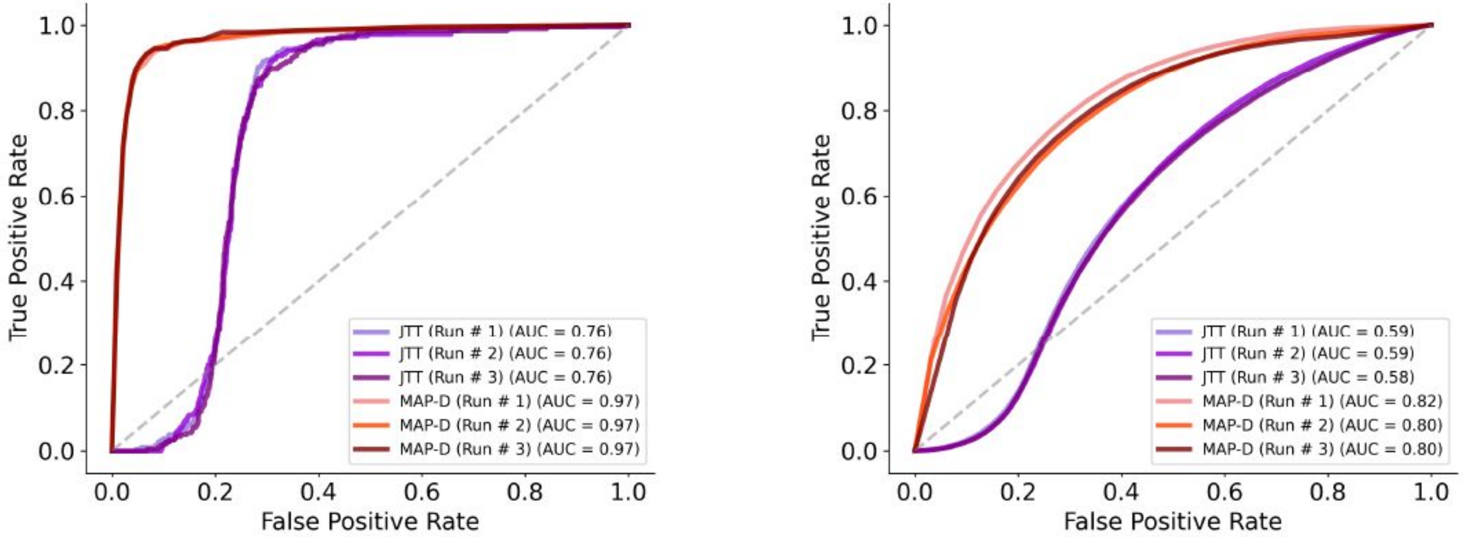
- MAP-D can identify if an example is clean or noisy
- Needs inclusion of metric which takes into account the extent to which an example is already learnt

```
training_score = (clean_score + (1.  - correct_class_confidence)) / 2.
```



Application: Minority group identification

- MAP-D is much more competitive in identifying minority group samples than competing methods relying on sophisticated early-stopping techniques



(a) Waterbirds (b) CelebA

Concluding remarks

Conclusion

- Automated data auditing techniques are essential for internet-scale data
- Prior work presents a siloed treatment of these metadata categories
- MAP-D is a simple and competitive approach for dealing with multiple metadata categories simultaneously based on a small set of reference examples (probe suites)
- MAP-D is capable of surfacing interesting subset of examples for scalable data auditing
- Combines well with metadata-specific interventions such as label correction or prioritized training

Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics

Shoaib Ahmed Siddiqui [1] Nitarshan Rajkumar [1] Tegan Maharaj [2] David Krueger [1] Sara Hooker [3]

Check out our paper / webpage for more details!

metadata-archaeology.github.io