Active Data Discovery: Mining Unknown Data using Submodular Information Measures

Suraj Kothawade SURAJ.KOTHAWADE@UTDALLAS.EDU
University of Texas at Dallas

Shivang Chopra T-SCHOPRA@MICROSOFT.COM
Microsoft Research

Rishabh Iyer RISHABH.IYER@UTDALLAS.EDU
University of Texas at Dallas

Abstract

Active Learning is a very common yet powerful framework for iteratively and adaptively sampling subsets of the unlabeled sets with a human in the loop with the goal of achieving labeling efficiency. Most real world datasets have imbalance either in classes and slices, and correspondingly, parts of the dataset are rare. As a result, there has been a lot of work in designing active learning approaches for mining these rare data instances. Most approaches assume access to a seed set of instances which contain these rare data instances. However, in the event of more extreme rareness, it is reasonable to assume that these rare data instances (either classes or slices) may not even be present in the seed labeled set, and a critical need for the active learning paradigm is to efficiently discover these rare data instances. In this work, we provide an active data discovery framework which can mine unknown data slices and classes efficiently using the submodular conditional gain and submodular mutual information functions. We provide a general algorithmic framework which works in a number of scenarios including image classification and object detection and works with both rare classes and rare slices present in the unlabeled set. We show significant accuracy and labeling efficiency gains for unknown classes ($\approx 10\% - 15\%$) and unknown slices ($\approx 5\% - 7\%$) with our approach compared to existing state-of-the-art active learning approaches for actively discovering these unknown classes and slices.

Keywords: Data Discovery, Submodular Information Measures

1. Introduction

Machine learning based predictions have been widely used in critical real-world domains like medical imaging and autonomous driving. Their success relies on availability of suitable supervised training data that can represent potential scenarios during test time. Unfortunately, real-world datasets are imbalanced and contain rare instances of data. These rare instances could represent a class (*e.g.* cat) or a data slice (*e.g.* brown animals). Models that are trained by using these imbalanced datasets are biased and perform poorly on these rare instances. A common practice to mitigate this imbalance is to iteratively acquire more *labeled* data by sampling from an *unlabeled* dataset by using human-in-the-loop active learning (AL) strategies. However, these techniques require a few exemplars of rare instances and assume knowledge about the total number of classes. In this paper, we study a new scenario of extreme imbalance such that the rare instances are completely absent from the labeled training dataset. Furthermore, we do not know if they even exist in the unlabeled dataset. Hence, we call such instances as *unknown* instances and the problem of finding them as the *data discovery* problem. This problem addresses the following question: *Can data points of unknown instances be discovered from a large unlabeled dataset in order to train a robust machine learning model?*

Instantiations of different submodular functions for active data discovery: To address this problem, we use different submodular functions for active data discovery that are presented in (Iyer et al., 2021; Kothawade et al., 2021). Particularly, the formulations for facility location (Fl), graph cut (Gc) and log determinant (Logdet) are as in (Iyer et al., 2021; Kothawade et al., 2021), and we adapt them as AL based acquisition functions for data discovery. Each function is named as the underlying submodular function, followed by Mutual Information (mi)/ Conditional Gain(cg)/ Conditional Mutual Information (cmi). For instance, the Fl based Smi function is denoted as Flmi, the Scg function is denoted as Flcg, and Scmi as Flcmi. The Gc and Logdet functions are denoted similarly. These functions are instantiated using a pairwise similarity matrix S, where $S_{\mathcal{A},\mathcal{B}}$ denotes similarity between items in $\mathcal{A}$ and $\mathcal{B}$, while S_{ij} denotes the entry (i, j) in S.

2. Add: Our framework for Active Data Discovery

In this section, we propose Add, a novel active learning based framework for data discovery. We show how Add uses a combination of conditioning via Scg functions and mutual information via Smi functions to effectively acquire data points of unknown instances.

The main idea behind the data discovery framework follows a *conditioning and targeting* strategy. Assuming that the unlabeled set $\mathcal{U}$ contains some unknown instances, we find them by conditioning on $\mathcal{P}$ that contains data points of only the known instances. Note that $\mathcal{P}$ is initialized as $\mathcal{L}$, since the initial labeled set has only known instances. Intuitively, by conditioning on $\mathcal{P}$, we are trying to find data points that are *dissimilar* to the known instances, thereby potentially finding unknown instances. We perform conditioning by maximizing the Scg function using a greedy algorithm (Mirzasoleiman et al., 2015): $\max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} f(\mathcal{A}|\mathcal{P})$.

The Scg function is instantiated using a similarity kernel $\mathcal{S}^{\mathcal{P}} \in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{U}|}$ that contains pairwise similarities between data points in $\mathcal{P}$ and $\mathcal{U}$ in the feature space. In every round of AL, we obtain labels via a human-in-the-loop for the selected subset $\mathcal{A}$ and add it to the labeled set $\mathcal{L}$. We augment the conditioning set $\mathcal{P}$ with $\mathcal{A}^{\mathcal{P}} \subseteq \mathcal{A}$ containing newly selected known instances, and $\mathcal{Q}$ is augmented with $\mathcal{A}^{\mathcal{Q}} \subseteq \mathcal{A}$ containing newly selected unknown instances. Note that $\mathcal{A} = \mathcal{A}^{\mathcal{P}} \cup \mathcal{A}^{\mathcal{Q}}$.

We keep a track of the unique known instances throughout AL rounds using a concept coverage set $\mathcal{K}$. Typically, $\mathcal{K}$ contains unique concepts like class indices if the goal is to discover classes, or $\mathcal{K}$ may contain attributes (*e.g.* color, shape), if they goal is to discover slices, or a combination of class and attribute (*e.g.* brown cat). If there are no new concepts in the newly discovered subset $\mathcal{A}^C$, *i.e.* $\mathcal{A}^C \cap \mathcal{K} == \emptyset$, we conclude the conditioning phase and start the targeting phase.

Algorithm 1 ADD: Active Data Discovery

```
Require: Initial Labeled set of data points: $\mathcal{L}$, containing $\mathcal{K}$ unique known instances. Large unlabeled dataset: $\mathcal{U}$. Initial conditioning set $\mathcal{P} \leftarrow \mathcal{L}$, query set $\mathcal{Q} \leftarrow \emptyset$. Model $\mathcal{M}$, batch size: $B$, number of selection rounds: $N$, $unknown$ = True
for selection round $i = 1 : N$ do
    Train model $\mathcal{M}$ with loss $\mathcal{H}$ on the current labeled set $\mathcal{L}$ and obtain parameters $\theta$
    if $unknown$ == True then
        Compute $S^{\mathcal{P}} \in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{U}|}$ such that: $S_{pu} \leftarrow$ COSINE_SIM$(\mathcal{M}_{\theta_i}, p, u)$, $\forall p \in \mathcal{P}, \forall u \in \mathcal{U}$
        Instantiate a SCG function $f(\mathcal{A}|\mathcal{P})$ based on $S^{\mathcal{P}}$.
        $\mathcal{A}_i \leftarrow \text{argmax}_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} f(\mathcal{A}|\mathcal{P})$ {Maximize SCG if unknown instances may exist}
    else
        Compute $S^{\mathcal{Q}} \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{U}|}$ such that: $S_{qu} \leftarrow$ COSINE_SIM$(\mathcal{M}_{\theta_i}, q, u)$, $\forall q \in \mathcal{Q}, \forall u \in \mathcal{U}$
        Instantiate a SMI function $I_f(\mathcal{A}; \mathcal{Q})$ based on $S^{\mathcal{Q}}$.
        $\mathcal{A}_i \leftarrow \text{argmax}_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} I_f(\mathcal{A}; \mathcal{Q})$ {Maximize SMI if no more unknown instances exist can be found}
    end if
    Get labels $L(\mathcal{A}_i)$ for batch $\mathcal{A}_i$ and $\mathcal{L} \leftarrow \mathcal{L} \cup L(\mathcal{A}_i)$, $\mathcal{U} \leftarrow \mathcal{U} - \mathcal{A}_i$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{A}_i^{\mathcal{P}}$, $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{A}_i^{\mathcal{Q}}$ {Add newly selected known instances to $\mathcal{P}$ and unknown instances to $\mathcal{Q}$}
    if $\mathcal{A}_i^C \cap \mathcal{K} == \emptyset$ then
        $unknown$ = False {No new unknown instances discovered}
    end if
    $\mathcal{K} \leftarrow \mathcal{K} \cup \mathcal{A}_i^C$ {Add newly discovered unique instances, if any}
end for
Return trained model $\mathcal{M}_{\theta_N}$ and labeled set $\mathcal{L}$ augmented with newly discovered instances.
```

In the targeting phase, we use a query set $\mathcal{Q}$ containing unknown instances that were accumulated in the conditioning phase. We maximize the mutual information with $\mathcal{Q}$ to find more semantically similar unknown instances from the unlabeled set $\mathcal{U}$. We do so by maximizing the SMI function using a greedy algorithm (Mirzasoleiman et al., 2015): $\max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} I_f(\mathcal{A}; \mathcal{Q})$.

Similar to the targeting phase, we augment $\mathcal{L}, \mathcal{P}$ and $\mathcal{Q}$ in every round of AL. Note that the purpose of the conditioning phase is to find a few points that represent the unknown instances and can serve as exemplars for the targeting phase. Note that one can continue conditioning throughout all AL rounds using, SCG or simultaneously perform conditioning and targeting using SCMI $\max_{\mathcal{A} \subseteq \mathcal{U}, |\mathcal{A}| \leq B} I_f(\mathcal{A}; \mathcal{Q}|\mathcal{P})$.

However, we empirically find the combination of SCG for conditioning and SMI for targeting to be most effective (see Sec. 3) and scalable. We present our active data discovery framework in Algo. 1. The ADD framework is generic and can be applied for any task. The only step that changes across tasks is the computation of pairwise similarity kernels $\mathcal{S}^{\mathcal{P}}$ and $\mathcal{S}^{\mathcal{Q}}$. In this paper, we apply ADD for image classification and object detection. Next, we discuss the similarity kernel computation for these tasks.

Similarity kernel for image classification: In order to represent each data point, we extract features from the penultimate layer of the model $\mathcal{M}$ that is trained using $\mathcal{L}$. Next, the cosine similarity kernels for an image classification discovery task can be computed easily by computing the dot product between the feature vectors. One can efficiently use off-the-shelf functions[1] for efficiently computing a vectorized dot product.

3. Experiments

In this section, we empirically evaluate the effectiveness of Add on a diverse set of datasets for image classification Sec. 3.1. Using these experiments, we show that existing AL approaches are not efficient for data discovery and that a *conditioning and targeting* approach as proposed in Add is essential. In addition to standard bench-marking datasets, we also conduct experiments on a real-world medical dataset for discovering unknown classes (Sec. 3.1.1), and for a realistic unknown slices setting (Sec. 3.1.2). Sec. 3.1 For all experiments using the Add framework, we use the same underlying submodular function f for Scg and Smi and is denoted as Scg+mi in the legend. For *e.g.*, we use facility location based variants Flcg and Flmi for one experiment and denote it as Flcg+mi. Note that we use the same f for simplicity of the experiments, and this is not a requirement for the Add framework. To ensure a fair treatment for all the methods, we use a common training procedure and set of hyper-parameters. We run all experiments 3× on a V100 GPU and provide error bars (standard deviation).

Baselines in all experiments: We compare Add with several uncertainty and diversity based baselines since they are the most intuitive solutions for the data discovery problem. Particularly, we compare against three uncertainty based baselines: Entropy (Settles, 2009), Margin (Roth and Small, 2006) and Least-Conf (Wang and Shang, 2014), and a recent diversity based baseline, Badge (Ash et al., 2020). Lastly, we compare with Random sampling.

3.1 Image Classification

In this section, we present the results for data discovery in the context of image classification tasks. We evaluate the performance of Add with existing AL acquisition functions for discovering unknown classes (Sec. 3.1.1) and unknown slices (Sec. 3.1.2). We do so by: 1) comparing the cumulative number of unknown data points selected by each acquisition function, and 2) comparing the mean accuracy obtained for the unknown classes or slices by training a model with the selected data points. For all the classification experiments, we train a ResNet-18 (He et al., 2016) model using an SGD optimizer with an initial learning rate of 0.01, momentum of 0.9 and weight decay of 5e-4. In every round of AL, we reinitialize the weights of our model using Xavier initialization and train the model till 99% accuracy is reached, or 200 epochs are complete. For evaluation, we use the default test set which contains both, the known and unknown instances.

3.1.1 Unknown Classes

Datasets and Experimental Setup: For discovering unknown classes, we apply our framework to the standard MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky et al., 2009) bench-marking datasets. We also conduct experiments on Path-MNIST (Yang et al., 2021; Kather et al., 2019), a real-world medical imaging dataset for colorectal cancer clas-

[1] np.tensordot or torch.tensordot

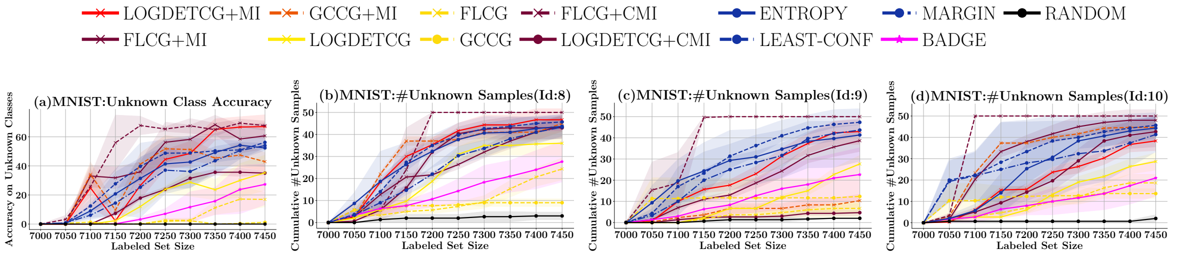

Figure 1: Active Data Discovery for unknown classes on MNIST. We observe that the Scg+mi (Flcg+mi and Logdetcg+mi) and Scg+cmi (Flcg+cmi) variants outperform other methods in terms of the average accuracy on the unknown classes. Flcg+cmi selects all data points from the unknown classes, the fastest, by 5^{th} round of AL.

sification. To consider a realistic scenario for data discovery, we create a labeled set $\mathcal{L}$ containing data points from K randomly chosen known classes. The unlabeled set $\mathcal{U}$ contains data points from X classes, where $X = K + Y$, *i.e.* $\mathcal{U}$ contains Y additional unknown classes. Realistically, $\mathcal{U}$ needs to be imbalanced and have lesser number of data points that belong to the unknown classes. Hence, we create an imbalance in the unlabeled set such that $|\mathcal{U}_k| = \rho|\mathcal{U}_y|$, where k is a class from the K known classes and y is a class from the Y unknown classes, and ρ is an imbalance factor. For MNIST and CIFAR-10 ($X = 10$), we set the first 7 classes as known ($K = 7$) and the last three as unknown classes ($Y = 3$). The number of data points in the labeled dataset $|\mathcal{L}| = 7000$, and unlabeled dataset $|\mathcal{U}| = 7150$ using an imbalance factor $\rho = 20$, and a batch size $B = 50$. For CIFAR10, we use $|\mathcal{L}| = 7000$, and $|\mathcal{U}| = 21900$ using an imbalance factor $\rho = 10$ and $B = 50$. For Path-MNIST, there exist a total of 9 classes ($X = 9$), we set the first 7 classes as known ($K = 7$) and the last two as unknown classes ($Y = 2$). We use $|\mathcal{L}| = 3500$, and $|\mathcal{U}| = 7200$ and an imbalance factor $\rho = 10$.

Results: We present the results for discovering unknown classes on MNIST in Fig. 1. We observe that the conditioning and targeting strategy as in Add using scg+mi and scg+cmi acquisition functions outperforms the uncertainty and diversity based methods by $\approx 5 - 15\%$ in terms of the average accuracy on the unknown classes (see Fig. 1 (a)). They obtain this gain in accuracy quickly, in the early rounds of AL and maintain it till the end. This is due to the fact that Add based strategies are able to select more data points from the unknown classes (see Fig. 1(b,c,d)). Particularly, Flcg+cmi finds all the data points from the unknown classes in early rounds of AL. However, as we discussed in Sec Sec. 2, Flcmi is computationally much more expensive than Flmi, and as we can see in Fig. 1(a), Flmi eventually obtains the same accuracy in the later rounds of AL.

In Fig. 2, we compare the computationally cheaper strategy of Add (scg+mi) for discovering unknown classes on two additional datasets - CIFAR10 (Fig. 2(a,b)) and Path-MNIST (Fig. 2(c,d)). We observe that the best performing methods are scg+mi variants which model representation (Fl) and diversity (Logdet) in addition to relevance. Particularly, Flcg+mi and Logdetcg+mi outperform other methods by $\approx 5 - 15\%$ in terms of the average accuracy on the unknown classes. Importantly, they acquire the best subset of unknown class data points in the early AL rounds (see Fig. 2(b,d)), thereby quickly reaching high accuracy values (see Fig. 2(a,c)).

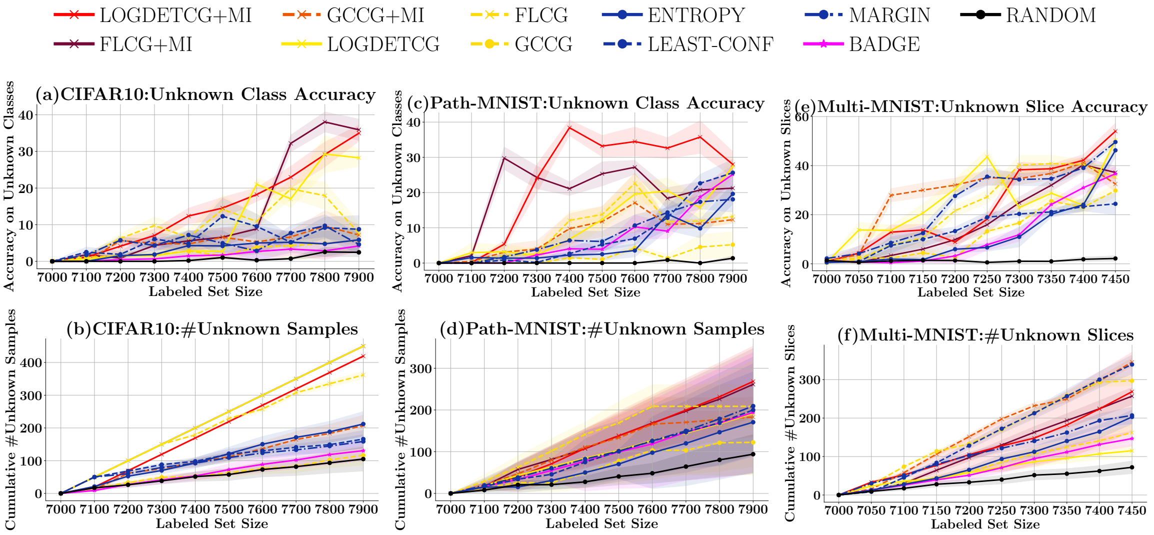

Figure 2: Active Learning for discovering unknown classes on CIFAR-10 (Krizhevsky et al., 2009) (**left** col.) and Path-MNIST (Yang et al., 2021; Kather et al., 2019)(**center col.**), and for discovering unknown slices on Multi-MNIST (Jiang, 2020)(**right** col). FLCG+MI and LOGDETCG+MI discovers the best subset of unknown class data points in early rounds of AL, thereby quickly gaining high accuracy for unknown classes. GCCG+MI discovers the best subset of unknown slices and obtains high accuracy for unknown slices.

3.1.2 UNKNOWN SLICES

Datasets and Experimental setup: For discovering rare slices, we apply our framework to Multi-MNIST, a dataset of images of digits from multiple languages. We consider two languages, English and Kannada, and try to train a single digit classification model for both the languages. In order to simulate unknown slices, we create a labeled set $\mathcal{L}$ which is missing data points for a few digits *only* for the Kannada language - 1,5, and 6 in our experiments. Note that $\mathcal{L}$ contains data points for all digits for the English language. We refer to these missing digits from the Kannada language as the unknown slice of data. Since these Kannada digits are unknown in $\mathcal{L}$, we create an imbalance for the Kannada language digits in the unlabeled set $\mathcal{U}$, as done in Sec. 3.1.1. For the labeled set, we use $|\mathcal{L}^{en}| = 10K$, $|\mathcal{L}^{ka}| = 7K$, and for the unlabeled set , we use $|\mathcal{U}^{en}| = 10K$, $|\mathcal{U}^{ka}| = 7.6K$, an imbalance factor $\rho = 5$, and a batch size $B = 50$. The superscripts en and ka denote the English and Kannada data slices, respectively.

Results: We present the results for discovering rare slices in Fig. 2. We observe that functions that model relevance to the query set $\mathcal{Q}$ outperform other functions in discovering rare slices. Particularly, GCCG+MI followed by LOGDET+MI outperforms other methods by $\approx 5 - 15\%$ in terms of average accuracy on the rare slices (see Fig. 2(e)). This is due to the fact that GCCG+MI selects the most number of data points from the unknown slices (see Fig. 2(f)).

4. Conclusion

In this paper, we propose Add an active learning based framework for data discovery. We show the effectiveness of Add for image classification tasks in discovery of unknown instances across classes and slices for a wide range of diverse datasets. Our experiments show that using a combination of Scg and Smi is the most effective and scalable, and obtains a $\approx 5\% - 15\%$ gain compared to existing baselines. The main limitation of this work is that the known instances need to be well represented in the feature space so that they are not mixed with the unknown instances. A potential negative societal impact of Add is that it can be used to discover and incorporate new biases in the dataset.

References

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pages 722–754. PMLR, 2021.

Weiwei Jiang. Mnist-mix: a multi-language handwritten digit recognition dataset. *IOP SciNotes*, 1(2):025002, 2020.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Prism: A rich class of parameterized submodular information measures for guided subset selection. *arXiv preprint arXiv:2103.00128*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. at&t labs, 2010.

Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, pages 413–424. Springer, 2006.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pages 112–119. IEEE, 2014.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2008*, 2021.