2021 IEEE/CVF International Conference on Computer Vision (ICCV) | 978-1-6654-2812-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCV48922.2021.01053

Low-Shot Validation: Active Importance Sampling for Estimating Classifier Performance on Rare Categories

Fait Poms*[1] Vishnu Sarukkai*[1] Ravi Teja Mullapudi[2] Nimit S. Sohoni[1]

William R. Mark[3] Deva Ramanan[2,4] Kayvon Fatahalian[1]

Abstract

For machine learning models trained with limited labeled training data, validation stands to become the main bottleneck to reducing overall annotation costs. We propose a statistical validation algorithm that accurately estimates the F-score of binary classifiers for rare categories, where finding relevant examples to evaluate on is particularly challenging. Our key insight is that simultaneous calibration and importance sampling enables accurate estimates even in the low-sample regime ($<$ 300 samples). Critically, we also derive an accurate single-trial estimator of the variance of our method and demonstrate that this estimator is empirically accurate at low sample counts, enabling a practitioner to know how well they can trust a given low-sample estimate. When validating state-of-the-art semi-supervised models on ImageNet and iNaturalist2017, our method achieves the same estimates of model performance with up to 10$\times$ fewer labels than competing approaches. In particular, we can estimate model F1 scores with a variance of 0.005 *using as few as* 100 *labels.*

1. Introduction

As model training techniques become increasingly label efficient, model validation stands to become a dominant fraction of overall data annotation costs. For example, state-of-the-art semi-supervised [3, 9, 2], weakly supervised [19], few-shot [7, 15], and active learning [25, 4, 8, 24] techniques all offer the promise of training models using a small number of human-labeled examples, but validating the resulting models typically uses large, human-annotated datasets. As a result, the cost of annotating validation sets is a significant factor limiting rapid model development.

In this paper we focus on the challenge of efficiently validating *binary image classifiers for rare categories* (positive instances are $\leq 0.1\%$ of the dataset). Building binary classification models for rare categories is common in real-world settings—wildlife preservation monitoring requires identifying rare flora and fauna species; autonomous vehicles must recognize rare categories, like baby strollers, to avoid collisions. The validation problem is particularly difficult for rare categories: while it is easy to collect a large amount of unlabeled data, finding even a small number of positives via uniform random sampling can require labeling thousands of images.

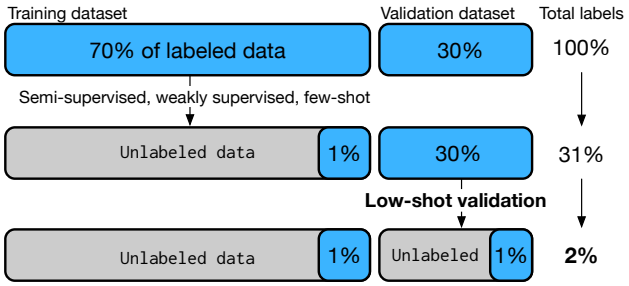

Figure 1: Recent model training techniques such as self-supervised learning, few-shot learning, and weakly supervised learning have made it possible to train models with a fraction of the traditional fully supervised training set. However, these methods still mostly evaluate using a large validation set. In this paper, we focus on *low-shot validation*, which addresses the high relative cost of collecting labeled validation data for models trained using label-efficient techniques.

Given a binary classification model to validate and a large unlabeled dataset, our goal is to estimate the model's F-score [29] on the data using a small number of annotated data samples. The F-score of a model depends on the distribution of the model's predicted labels, which are known, and the dataset's ground-truth labels, which require data annotation. *Importance sampling* [27] is a powerful theoretical tool for stochastically sampling the most important points in a dataset to label, but the efficiency of estimating F-scores using importance sampling depends on accurate knowledge of the likelihood that a given sample is a positive. Therefore, the key challenge in using importance sampling for efficiently computing F-scores is *model calibration*, the task of predicting the likelihood that a given sample is a positive, conditioned on the model scores. Given this observation, we propose an active sampling algorithm that alternates between acquiring labels used to train an iso-

* Both authors contributed equally to this paper
[1] Stanford University
[2] Carnegie Mellon University
[3] Google
[4] Argo AI

978-1-6654-2812-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCV48922.2021.01053

tonic regression model [34] for calibrating model probabilities, then using the calibrated model scores to importance-sample batches of data for metric estimation. Using this alternating strategy, our scheme generates progressively better estimates of the model's F-score.

We demonstrate that, particularly in the low-sample (< 300 labeled samples) regime, our algorithm can estimate F1 with significantly lower error than a variety of baselines, including semi-supervised Gaussian Mixture Models [13], prior importance-sampling approaches [22], and "herding" algorithms [32]. Not only are we able to estimate F1 efficiently, we are also able to estimate the *variance* of our estimate accurately, even in low-sample regimes. This contribution has important practical ramifications in that it allows a practitioner to know if they should trust the estimate generated by a small set of labeled validation data. Our contributions are as follows:

1. An algorithm for joint active calibration and importance sampling-based F-score estimation. Our algorithm produces accurate and reliable estimates of a model's F-score, and it significantly outperforms baseline methods in low-sample regimes (< 300 labeled samples).
2. A single-trial estimator of variance for our method. We demonstrate that our variance estimator is empirically accurate, even for low-sample counts, offering a valuable diagnostic tool when using our algorithm in real-world settings.
3. A study that demonstrates that validation sets chosen specifically for a given model can also efficiently validate *other models* trained for the same task.

2. Related Work

Approaches to label-efficient validation include statistical importance sampling-based methods [22], indirect techniques that estimate precision-recall curves [13, 31], active learning adapted for validation [18], and stratified sampling techniques [1, 12, 33]. While many methods attempt to solve the validation problem, very few do so for highly imbalanced rare categories with a very small labeling budget. We compare against a representative subset of methods which tackle this problem in our evaluation section (Sec. 4). We delineate these methods below.

Importance sampling. Importance sampling allows for Monte Carlo estimation of metrics using samples drawn from arbitrary distributions. Sawade et al. [22] propose an importance-sampling algorithm to actively estimate F-measures [29], deriving an importance distribution based on the model's predicted probabilities and labels. Their method is statistically consistent but relies on assumptions of good model calibration. Our importance-sampling distribution is based on Sawade et al., and we compare against theirs in our evaluation.

Estimating precision-recall curves. Instead of estimating F-score directly, learning the shape of the precision-recall curve can help calculate a variety of validation metrics indirectly. Miller et al. [13] fit the distributions of positive and negative samples across the score distribution with Gaussian mixture models (**GMM**), while Welinder et al. [31] train a generative Bayesian model on the classifier's confidence scores. We evaluate against the **GMM** method of Miller et al. in order to compare against this class of techniques. Other methods for estimating PR curves make strong assumptions which only apply once hundreds of samples have already been labeled [21]. It is interesting future work to combine our low-sample-count validation with such methods which focus on validation with a budget of thousands of samples.

Covering the data distribution. "Herding" attempts to reconstruct the sufficient statistics of the dataset from a set of pseudo-random samples [32]. These samples can then be used to estimate F-score. We evalute against Herding in our evaluation.

Active learning and stratified sampling for validation. Active learning techniques use a partially-trained model to select samples to label that will maximize the trained performance of that model, and have recently been applied to validation [18]. Stratified sampling techniques produce low-variance estimates by subdividing the domain into "strata" of samples that are similar to one another, sampling from each stratum to ensure that the domain is covered [1, 12, 33]. However, none of these techniques focus on rare categories and either apply only after an initial seed set of hundreds to thousands of labels have been collected or assume labeling budgets in the thousands.

2.1. Other validation settings.

Other validation settings are distinct from the one we study here, but share the challenge of label-efficient validation. **Validation under domain shift:** When faced with domain shift in production settings, Taskazan et al. [26] show that measuring shifts in production data distributions and training models to measure the uncertainty in model prediction both help guide label-efficient and accurate validation. **Validation under noisy annotation:** Nguyen et al. [14] explore the problem of estimating AP from noisy labels, and they introduce an active algorithm for choosing samples to label.

2.2. Calibration.

With a perfectly calibrated model, it is possible to statistically estimate the number of true positives, false positives, etc. in the dataset, and thus the F-score. Guo et al. [10] illustrate that larger and better-performing neural networks are often poorly-calibrated. Platt scaling [17] and isotonic regression [34] provide two well-studied [16] methods of calibrating models, and we compare against both techniques in

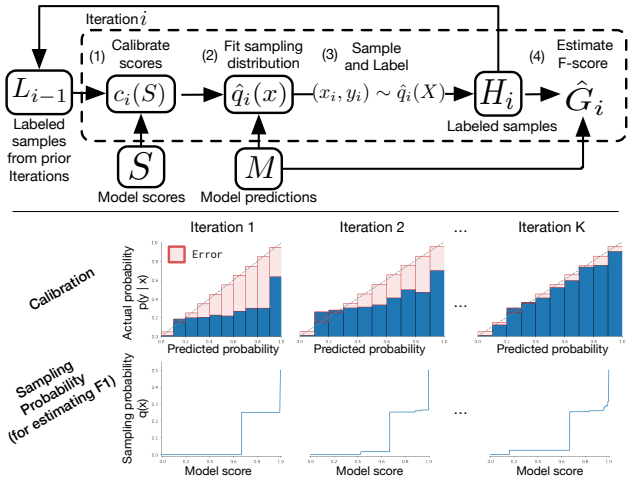

Figure 2: We estimate a model's F-score with a small labeling budget using an iterative algorithm which combines classifer calibration and importance sampling. **Top:** each iteration of our algorithm (1) calibrates the model's scores using the labels from the prior iteration, (2) uses these calibrated scores and the models' predictions to fit an importance sampling distribution, (3) samples a batch from this distribution to label, and (4) uses these samples to estimate the model's F-score. **Bottom:** visualizations of the intermediate iterations of our method on step (1) calibration and step (2) the importance sampling distribution. As our labeling budget increases, the model's scores gradually converge to a calibrated model (linear gray dotted line) and the sampling distribution becomes increasingly refined. The sampling distribution is optimized to compute the model's F1.

our evaluation. We utilize isotonic regression for calibration in our iterative algorithm.

2.3. Classifier training with few manual labels.

Recent advances in semi-supervised training can produce models trained with just 1% of the ImageNet training set labels that are competitive with models trained on the fully supervised ImageNet dataset [2, 11, 9, 6, 3]. Semi-supervised methods produce models using a blend of manual and *automatic* labels [23], However, their performance is evaluated using the *entire* ImageNet validation split (50,000 labeled images), tempering the actual reduction in overall annotation budget achieved. Our goal is label-efficient validation of these methods, and we validate three state-of-the-art semi-supervised models in Section 4.

3. Method

Our goal is to estimate the F-score [29] for a target model on a (potentially infinite) test distribution given a small labeling budget. Figure 2 provides an overview of our method: given the model's predictions and scores, we iteratively improve our estimate of the model's F-score by repeatedly (1) calibrating the model's raw scores S to produce better estimates $c(s)$ of the probability that a sample is a positive, (2) using the calibrated scores to compute an importance sampling distribution $\hat{q}(x)$ that prioritizes samples most important for accurately estimating the F-score, (3) labeling samples drawn according to $\hat{q}(x)$ and (4) using those samples to estimate the F-score using importance sampling. In this section, we first introduce our problem setup more formally and provide the necessary background on importance sampling for estimating F-scores. Then, we describe our method and how to estimate the method's variance accurately with a single trial.

3.1. Problem Statement

Given a model M and a dataset $(x_j, y_j) \sim X, Y, j \in \{1...n\}$ with features x_j and unknown true labels y_j, our goal is to estimate M's performance at predicting the true labels on X given a finite labeling budget. In this paper, we focus on estimating the F-score, G, of the model, a measure of the deviation between labels predicted by the model $\hat{Y}$ and the true labels Y. Parameterized by $\alpha \in [0, 1]$, The F_α-score G [29] is defined:

$$G = \frac{tp}{\alpha(tp + fp) + (1 - \alpha)(tp + fn)} \tag{1}$$

where $tp = E[\mathbb{1}[Y = 1 \wedge \hat{Y} = 1]]$ is the model's true positive rate, $fp = E[\mathbb{1}[Y = 0 \wedge \hat{Y} = 1]]$ the false positive rate, and $fn = E[\mathbb{1}[Y = 1 \wedge \hat{Y} = 0]]$ the false negative rate. We obtain F1 by setting $\alpha = 0.5$, and obtain precision and recall by setting $\alpha = 1$ and $\alpha = 0$ respectively.

We aim to understand the deviation between our predicted F-score $\hat{G}$ and the true F-score G. We do so by measuring the mean-squared error (MSE) of our estimation method $E[(\hat{G} - G)^2]$ and by studying its bias $E[\hat{G} - G]$ and the variance $E[\hat{G}]^2 - E[\hat{G}^2]$.

We assume the model M generates predicted labels $\hat{y}_j \in \{0, 1\} \sim \hat{Y}, j \in \{1...n\}$, and model scores $s_j \in S, j \in \{1...n\}$.

3.2. Background: Importance Sampling for Consistent Estimation of F-Score

When estimating a metric, importance sampling is a technique that non-uniformly selects samples according to a sampling distribution q. The technique aims to choose a q that is optimized to produce good estimates of the metric, and it corrects for the unequal probabilities of selection in order to produce accurate estimates. Sawade et al. [22] introduce an importance sampling method for the consistent (asymptotically unbiased) estimation of F-measures. Suppose $p(x, y)$ defines the probability distribution across the population of samples $x \in X$ and labels $y \in Y$. Let $v(x, y, \hat{y}) = \alpha\hat{y} + (1 - \alpha)y$, with $\alpha \in [0, 1]$, and $\ell = 1 - \ell_{0/1}$, the zero-one loss. With any distribution $q(x, y)$ where q is nonzero across the domain of p and

$(x_1, y_1), (x_2, y_2)... \sim q(x, y)$, Sawade et al. approximate the population F-score G as $\hat{G}_{n,q}$:

$$\hat{G}_{n,q} = \frac{\sum_{j=1}^{n} w(x_j, y_j, \hat{y}_j)\ell(\hat{y}_j, y_j)}{\sum_{j=1}^{n} w(x_j, y_j, \hat{y}_j)} \tag{2}$$

$$w(x_j, y_j, \hat{y}_j) = \frac{p(x_j)}{q(x_j)} v(x_j, y_j, \hat{y}_j)$$

They prove that $\hat{G}_{n,q}$ is a consistent estimator of G i.e. $\hat{G}_{n,q} \xrightarrow[n\to\infty]{} G$. Therefore, as the sample size n increases, choosing the distribution q^* that minimizes the variance of $\hat{G}_{n,q}$ becomes inceasingly effective at minimizing the MSE of the estimate.

Optimal sampling distribution Sawade et al. derive the theoretically optimal variance-minimizing sampling distribution, q^*, for their estimator:

$$q^*(x) \propto \begin{cases} p(x)\big(p(y=1|x)(1-G)^2 + \\ \quad \alpha^2(1-p(y=1|x))G^2\big)^{0.5} & : \hat{y} = 1 \\ p(x)(1-\alpha)\big(p(y=1|x) * G^2\big)^{0.5} & : \hat{y} = 0 \end{cases} \tag{3}$$

However, this formula assumes knowledge of G, the very metric that we aim to estimate. In addition, it assumes knowledge of $p(y|x)$, which is unknown as well. Sawade et al. work around this limitation by substituting model scores S for $p(y = 1|x)$, which assumes that the model being evaluated is perfectly calibrated. However, many neural networks are not well-calibrated [10], in particular when training classifiers for rare categories [30]. We address these limitations in our proposed algorithm.

3.3. Active Calibration and Importance Sampling

We propose **A**ctive **C**alibration and **I**mportance **S**ampling (ACIS, Algorithm 1), an iterative importance sampling algorithm for simultaneously estimating $p(y = 1|x)$ (calibrating the model) and estimating G. Each iteration i first calculates $\hat{q}_i$, our current best estimate of q^*, using the samples that are already labeled. $\hat{q}_i$ is estimated by training an isotonic regression [34] model $c_i(s)$ on the model scores $s \in S$ to predict $p(y = 1|x)$, using the samples that are already labeled to train. Applying equation 3, substituting $c_i(s)$ in place of $p(y = 1|x)$ and $\hat{G}_{i-1}$ in place of G:

$$\hat{q}(x) \propto \begin{cases} p(x)\big(c_i(s)(1-\hat{G}_{i-1})^2 + \\ \quad \alpha^2(1-c_i(s))\hat{G}_{i-1}^2\big)^{0.5} & : \hat{y} = 1 \\ p(x)(1-\alpha)\big(c_i(s) * \hat{G}_{i-1}^2\big)^{0.5} & : \hat{y} = 0 \end{cases} \tag{4}$$

Isotonic regression outperforms Platt scaling [17] in the setting with extremely imbalanced classes and classification models of varying quality (see Supplemental for details). The isotonic model is trained by reusing samples L that are initially labeled in order to estimate $\hat{G}$. In the first iteration,

Algorithm 1: ACIS

Data: $\hat{Y}, S, B, \alpha$
Result: $\hat{G}_1, \hat{G}_2, ..., H_1, H_2..., W_1, W_2, ...$
Function `Estimate`($\hat{P}_y$, *budget*, $\hat{G}$)**:**
  $\hat{q}$ = **ImportanceDistribution**$(\hat{P}_y, \hat{Y}, \hat{G}, \alpha)$
  H, W = **WeightedSample**$(q, budget)$
  $\hat{G}$ = **FScore**$(H, \hat{Y}, W)$
  return $H, W, \hat{G}$
c_0 = **IsotonicRegression**$(S, \hat{Y})$, $\hat{G}_0 = 0.5$
$B_1 = 10$, $i = 1$, $L_0 = \{\}$
while $|L_{i-1}| < B$ **do**
  $H_i, W_i\hat{G}_i$ = `Estimate`$(c_{i-1}(S), B_i, \hat{G}_{i-1})$
  $L_i = L_{i-1} \cup H_i$
  c_i = **IsotonicRegression**$(\hat{P}_y, L_i)$
  $B_{i+1} = 2 * B_i$, $i = i + 1$
end
return $\hat{G}_1, \hat{G}_2, ..., H_1, H_2..., W_1, W_2, ...$

when there are no existing labeled examples, the algorithm samples from the calibration prior $c_0(s)$, which is an isotonic model trained on the predicted labels $\hat{Y}$ rather than the true labels Y. For all iterations, the range of $c_i(s)$ is linearly rescaled from $[0, 1]$ to $[\epsilon, 1 - \epsilon]$ in order to ensure that $\hat{q}_i$ is nonzero for all samples. Next, a new batch of samples H_i drawn from $\hat{q}_{i-1}$ is first labeled, then used to compute $\hat{G}_i$, the current estimate of G. At the end of the process, the algorithm returns a sequence of estimates $\hat{G}_1, \hat{G}_2, ...$ of G. Our estimates of G progressively improve as our estimate of q^* improves.

Estimate averaging. By default, $\hat{G}_k$, the last iteration returned by Algorithm 1, can be used as the estimate of F-score. However, in order to utilize the earlier samples, the final ℓ estimates of G are aggregated as follows:

$$\hat{G} = \frac{\sum_{i=k-\ell+1}^{k} \hat{G}_i * |W_i|_1}{\sum_{i=1}^{k} |W_i|_1} \tag{5}$$

By Sawade et al. [22], each of the ℓ $\hat{G}_i$'s are consistent estimators. Since $\hat{G}$ is a weighted average of a finite number of consistent estimators, $\hat{G}$ is also consistent. By averaging across the ℓ $\hat{G}_i$'s, we aim to improve the estimate of G by incorporating information from earlier iterations of sampling, not just the last iteration.

Reusing prior iteration samples In each iteration i, sample-efficiency is improved by making use of all previously-labeled samples L_{i-1}. Since these points have already been labeled in a previous iteration, they do not add to our overall labeling budget. To account for deterministically labeling $|L|$ points in a dataset of size $|X|$, their importance weights are set to $\frac{p(x)}{\hat{q}_i} = \frac{|L|/|X|}{|L|/|L|} = \frac{|L|}{|X|}$.

Weighted average of $c_i(s)$. In practice, the $c_i(s)$ learned in the early iterations of the model can be unstable. We

compensate for this effect by calibrating using $\beta * c_0(s) + (1-\beta) * c_i(s)$, a weighted combination of $c_i(s)$ and the calibration prior $c_0(s)$, and decreasing the weight β linearly over the first few iterations of the algorithm.

Adaptive top-K prior for rare categories. When estimating F-score for rare categories, the vast majority of ground-truth positives are often high in the sorted ordering of model scores S. This means that sampling the rest of the ordering largely yields negatives. Therefore, when estimating model F1 for rare categories the sampling domain is restricted to the $3 * (i+1) * n_{pos}$ examples with the highest score s, where n_{pos} is the number of samples labeled positive by the model in the dataset. There are $n_{pos} + fn$ samples in the dataset that are relevant for estimating F-score, and it is difficult to estimate fn accurately, so our sampling range heuristic is dependent on n_{pos}. In addition, by making the sampling range dependent on i, the set of potential samples that can be labeled every iteration is expanded, slowly "relaxing" the heuristic. While limiting the sampling domain introduces bias by potentially undersampling false negatives, in practice the reduction in variance far outweighs the slight bias introduced, and in low-sample regimes the domain restriction significantly reduces the MSE of our algorithm.

3.4. Calculating Variance

The variance of a randomized estimator is a powerful diagnostic tool for understanding its potential error. We derive the following consistent estimator of sampling variance for our active method:

$$S^2_{n,q} = \frac{C^{-1} * \sum_{j=1}^{n} w(x_j, y_j, f_\theta)^2 \left(\ell(f_\theta(x_j), y_j) - \hat{G}_{n,q}\right)^2}{\frac{1}{n}\left(\sum_{j=1}^{n} w(x_j, y_j, f_\theta)\right)^2} \tag{6}$$

$$C = 1 - \frac{\sum_{j=1}^{n} w(x_j, y_j, f_\theta)^2}{(\sum_{j=1}^{n} w(x_j, y_j, f_\theta))^2}$$

Our derivation leverages the Delta Method [5] to obtain a consistent estimator of sampling variance, then applies a Bessel-style correction C to improve performance in low-sample regimes. The derivation is included in the Supplemental.

When applying Equation 5 to combine the estimates $\hat{G}_i$'s, the variance of the weighted average $\hat{G}$ is estimated by taking a weighted average of the sampling variances of each iteration. The implied assumption that there is no covariance between the $\hat{G}_i$'s is a reasonable assumption in practice, and yields better estimates of the variance of $\hat{G}$ than the worst-case estimator (which assumes a covariance of 1). Experimental evidence is included in the Supplemental.

4. Evaluation

Our evaluation compares the sample efficiency of our F-score estimation algorithm to baseline approaches. We also provide an analysis of our method's ability to provide bounds on the estimated metric's error by predicting the variance from just a single trial. We refer to our method as ACIS (**A**ctive **C**alibration and **I**mportance **S**ampling).

4.1. Experimental setup

Datasets. We evaluate ACIS on the ImageNet [20] and iNaturalist [28] large-scale image classification datasets. ImageNet is an image classification dataset with 1000 categories, 1.2 million training images, and 50,000 validation images. To investigate the semi-supervised setting, we follow [3] by restricting training labels to the same 1% split of the training dataset. We measure validation performance on two datasets: 1) the remaining 99% of ImageNet train dataset (which we refer to as *ImageNet1M*) and 2) the ImageNet validation dataset (*ImageNet50K*). In addition, we evaluate on iNaturalist (*iNat100K*) an image classification dataset with 5089 categories, 579,000 training images, and 95,986 validation images. Taking inspiration from [3], we construct a 10% split of the training dataset for semi-supervised learning.

Models. To test our algorithm on models trained with limited labeled data, we estimate the F1 of three self-supervised learning methods which provide state-of-the-art semi-supervised performance: SwAV [2], SimCLRv2 [3], and BYOL [9]. We treat each of the 1000-way outputs of the classifier as a binary classifier through one-vs-all classification. Results are presented in terms of average performance across these 1000 binary classification tasks. (We train the SwAV and BYOL models ourselves as off-the-shelf weights are not provided; details in the supplemental.)

While our experiments evaluate binary classifiers constructed from multi-class classifiers, our algorithm is capable of validating any arbitrary binary classifier that produces class scores. Our method only requires class scores and predicted labels from a target model.

Baselines. We evaluate our approach against: TOP-K, an approach common in information retrieval [18] which draws the top K samples from the model's ranked scores; GMM [13], which also labels the top K samples from the model's ranked scores, then fits a two-component Gaussian mixture model to the model's score distribution to predict labels on the unlabeled examples; and HERDING [32], which attempts to approximate the metric using samples that can reconstruct the sufficient statistics of the dataset.

4.2. Validating semi-supervised models

4.2.1 Comparison to baselines

Validation on *ImageNet1M*. We first compare the performance of the different validation methods on *ImageNet1M*,

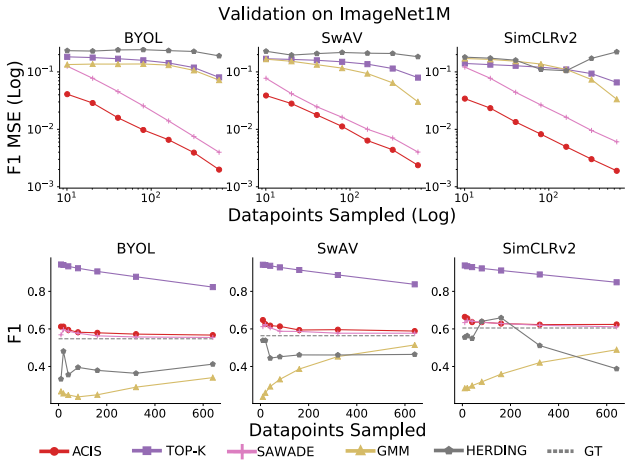

Figure 3: ACIS estimates the F1 scores of semi-supervised models to MSE < 0.01 using less than 100 samples, even when sampling from the large *ImageNet1M* dataset. **Top:** the mean squared error (MSE) of the estimated F1, averaging across a single trial for each of the 1000 ImageNet categories. ACIS has consistently lower MSE than all baselines. **Bottom:** the predicted F1 score, averaged across a single trial for each of the 1000 ImageNet categories. In all cases, ACIS estimates the F1 score in expectation to within 0.1 of the true value, even for as few as 10 samples. Other than SAWADE, the other baselines exhibit more bias.

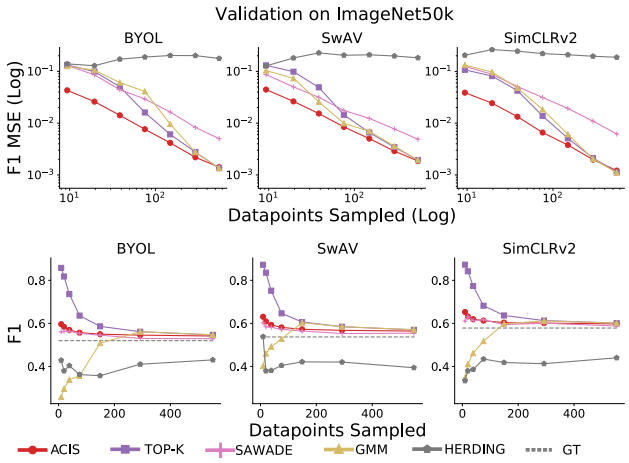

Figure 4: Validation accuracy on *ImageNet50K* following the experimental setup of Fig. 3. Again, ACIS generates lower MSE estimates in low-sample regimes (< 300 samples). Since *ImageNet50K* is $20\times$ smaller than *ImageNet1M* there are fewer relevant samples to find for a given category, so the MSE for most methods converges by 500 samples.

a dataset with over a million images. Figure 3 gives the estimated F1 score (bottom) and mean squared error (MSE) (top) for the F1 estimate (compared to the metric computed on the full dataset) for all methods, running a single

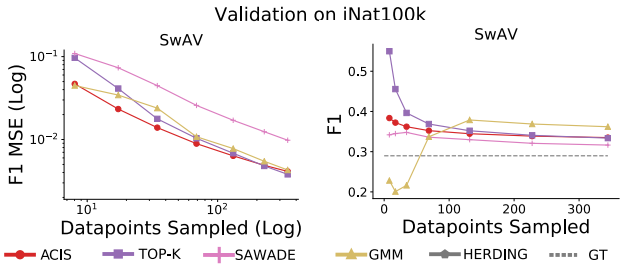

Figure 5: Validation accuracy on *iNat100K* following the experimental setup of Fig. 3. In low-sample regimes (< 100 samples), F1 estimates by ACIS on the *iNat100K* dataset have lower MSE than baselines. *iNat100K* has fewer images than *ImageNet1M* and fewer images per class than *ImageNet50K*, so there are fewer relevant samples to find, so most methods converge in MSE by 200 samples.

trial of each method on every ImageNet class, then averaging across the 1000 classes. In low-sample regimes (labeling < 600 samples), ACIS generates estimates with significantly lower MSE than the baseline approaches. Estimates by SAWADE have slightly lower bias than ACIS, but SAWADE performs worse in terms of MSE since it constructs its sampling distribution using uncalibrated probabilities. TOP-K is not competitive because it fails to sample false negatives when sampling a small fraction of a million-image dataset. Similarly, GMM appears to lack sufficient signal in the tail of the distribution in order to fit the mixture model for the distribution of positive and negatives samples. HERDING performs poorly, perhaps because the task of generating pseudo-random samples to estimate the moments of a high-dimensional distribution does not translate well to the task of estimating F-score in low-sample regimes.

Validation on *ImageNet50K*. We also compare the validation methods on the smaller *ImageNet50K* dataset (50,000 unlabeled samples). As seen in Figure 4, ACIS consistently outperforms baseline approaches when labeling less than 300 samples for a binary classifier. This low-sample regime is particularly significant because the classifiers being evaluated are only trained on 10 positive samples per class. TOP-K and GMM have high bias in the low-sample regime due to consistently overestimating or underestimating the model's F1, respectively (Fig. 4-bottom). As the labeling budget increases, TOP-K and GMM converge to ACIS because most of the relevant samples for the F1 score have been labeled. ACIS performs similarly in terms of MSE on both *ImageNet50K* and *ImageNet1M*, as seen in a comparison of Figures 3 and 4.

Validation on *iNat100K*. We also compare the validation methods on the *iNat100K* dataset, validating a model trained with SwAV. The typical class in *iNat100K* has approximately 20 instances of each class, whereas *ImageNet50K*, a dataset of 50,000 images, has $\frac{50000}{1000}$=50 im-

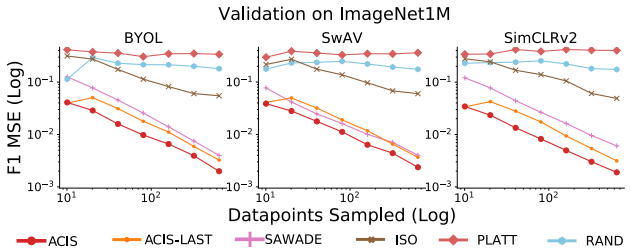

Figure 6: Ablation analysis: ACIS, which jointly performs calibration and importance sampling, performs better than both calibration-only approaches (ISO, PLATT) and importance sampling without calibration (SAWADE). Taking a weighted average of our iterative estimates of the F1 also outperforms just taking the last (ACIS vs ACIS-LAST respectively). All approaches except PLATT significantly outperform uniform random sampling (RAND).

ages per class. As a result, in Figure 5, we observe trends similar to *ImageNet50K*, but methods converge after a smaller number of samples (100, as opposed to 300 for *ImageNet50K*). Similar to the *ImageNet50K* experiments, TOP-K and GMM compare the most favorably to ACIS.

4.2.2 Ablation analysis

We perform an ablation analysis to understand the benefits of different components of ACIS. To ablate the averaging of F1 estimates from several iterations, ACIS-LAST uses the F1 estimate from only the last iteration of ACIS. Unlike ACIS, SAWADE [22] ablates model calibration. We also ablate importance sampling in three configurations: ISO samples at uniform random, then applies isotonic regression [34] to infer labels on unlabeled data points. PLATT samples at uniform random, then applies Platt scaling [17] to infer labels on unlabeled data points; and RAND samples at uniform random, then estimates the metric using only the selected samples.

Figure 6 shows the MSE of the various methods on *ImageNet50K*. Jointly calibrating (ACIS) is more sample-efficient than a pure importance sampling method (SAWADE) because the importance sampling distribution improves with well-calibrated models. Only using model calibration (ISO, PLATT) performs no better than uniform sampling (RAND) in the low-data regime (< 100 samples) because it does not actively select samples to improve the estimate of the metric. Jointly calibrating and performing importance sampling to estimate F1 is more effective than using either technique in isolation. Combining estimates from multiple iterations has a small benefit–ACIS outperforms ACIS-LAST in the regime with fewer than 100 samples, though the methods converge for larger sample sizes.

4.2.3 Computational cost

Compared to training, inference, and labeling costs of the model development process, the computational costs of ACIS are trivial. The entirety of the computational cost of iterative calibration and sampling is less than a minute per model in our experiments on a single CPU. This small computation cost can potentially yield a significant reduction in the number of images that must be hand-annotated to achieve a target validation accuracy.

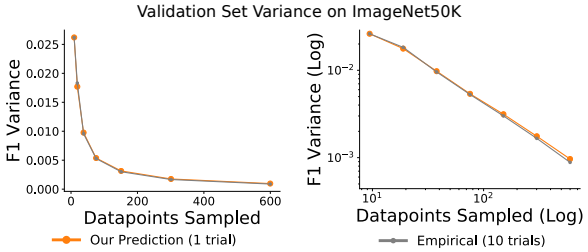

Figure 7: Our single-trial predictions of the variance of ACIS (orange) closely mimic empirical estimates of its variance (gray) when evaluating SwAV. We slightly underestimate the empirical variance in the very low-sample regimes, but quickly converge to the empirical estimates as sample size increases. Reliable estimates of empirical variance make our estimates of F-score more practically useful. For example, if aiming for an variance of 0.01 in an F1 estimate, we know that the variance objective has been met after 40 samples.

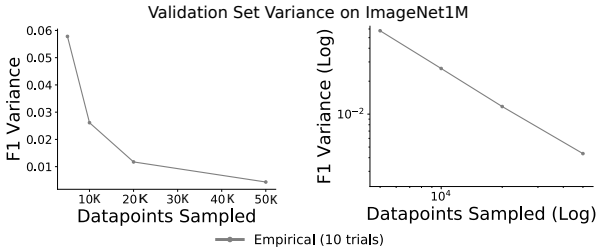

Figure 8: When validating SwAV on large, randomly sampled validation datasets, we observe significant variance in predicted F1 even when sampling up to 50,000 images. This suggests that even validation estimates calculated from traditional large datasets have notable uncertainty when validating binary classifiers trained for rare categories.

4.3. Variance Diagnostics for Estimating F-score

4.3.1 Variance Estimation for ACIS

We compare our estimate (Formula 4) of the variance of ACIS's F1 estimate to the method's empirical variance on *ImageNet1M* (Fig. 7). Our method (orange line) closely approximates the empirical variance (gray line) of ACIS even when computed from a small number of samples. The empirical variance is computed by performing 10 independent trials of ACIS.

4.3.2 Finite-Dataset Variance

Computing F-score using a large, but finite, dataset will often yield a very good estimate of the F-score for the full

test population. However, due to their finite size, even these estimates have variance, in particular when classifying rare categories. We can estimate this variance using Formula 4.

Figure 8 illustrates that notable variance in estimates remain present even when using large, randomly-sampled validation datasets. To assess the accuracy of our variance estimate, for each of the 1000 ImageNet categories we randomly sample subsets (of up to 50,000 images) of the *ImageNet1M* dataset, and we evaluate the F1 of SwAV on these validation sets.

The variance of the estimate of F1 on a 50,000-sample dataset (the size of the ImageNet validation dataset) is 0.003, implying a standard devation of $\sqrt{0.003} \simeq 0.055$. *Therefore, when evaluating binary classifiers trained on rare categories, many datasets commonly used to compute ground-truth estimates of model performance themselves contain notable uncertainty.*

4.4. Sharing Validation Sets Across Models

The previous results demonstrate that validation efficiency can be significantly improved by curating a validation set for the *specific model under evaluation*. However, these labeled samples can be used to estimate F-scores for other models as well. (Our estimator remains consistent when evaluating other models.) To understand how validation sets transfer to other models, for each of SwAV, SimCLRv2, BYOL we curate validation sets for estimating F1 performance of the model, then we use these validation sets to estimate the performance of all three models (Figure 9).

For all three models, the F1 estimates have the lowest MSE when curated specifically for the desired model. However, F1 estimates generated from datasets curated for a given model are surprisingly effective for validating other models. ACIS datasets curated for different models obtain F1 estimates with MSE values competitive with the best baseline techniques that are actually tailored to the desired model (brown lines in Figure 9), sometimes even outperforming the baseline techniques in the low-sample regime. For instance, when validating SimCLRv2, an ACIS dataset curated for SwAV performs better than the best baseline trained on SimCLRv2 itself when sampling less than 80 points. The surprising effectiveness of using a model-specific validation set to validate other models may be due in part to similarity between the models validated. Nevertheless, our results suggest that, for a similar family of models, actively curated validation sets can effectively estimate F-score across different models.

5. Discussion

We have presented a method for estimating the F-score of binary classification models on a low label budget, as well as a method for predicting the variance in this estimate. Our approach constructs validation sets specifically for the target model to evaluate, but we demonstrate that validation sets constructed for one model can also be used to efficiently validate similar models. This observation suggests that, for a given task, it might be possible to actively construct model-agnostic datasets that enable accurate validation with far fewer labeled samples than datasets typically used today.

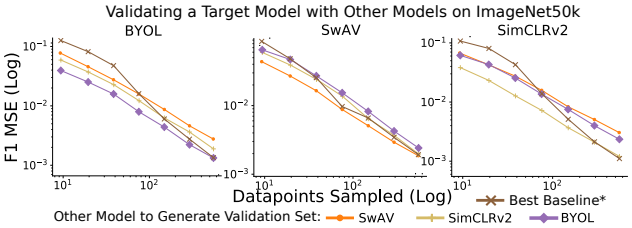

Figure 9: Validation sets curated for each of SwAV, BYOL, and SimCLR can be used to efficiently estimate F1 score of the other two models on *ImageNet50K*. For a given sample size, the MSE of the ACIS estimator tailored to a different model is at most double the MSE of the model-specific estimator. These estimators are competitive with, and sometimes outperform, the best model-specific baseline approaches in the low-sample regime. *The brown lines reflect the best of the model-specific baseline methods from Figure 4.

That said, for models that substantially differ from prior art, our experiments suggest that it may be worth constructing one-off model-*specific* validation sets. For example, safety-critical perception tasks that enable self-driving vehicles might require extremely precise estimates of validation performance of deployed models. Rather than validating models on pre-defined fixed test sets, our method may provide a framework for actively validating models on live on-fleet data streams.

Finally, our algorithm's accurate variance estimates can also be used to construct confidence intervals, and also provide guidance on the sample budget required to obtain an estimate of F-score with an acceptable level of variance. Future analysis analyzing the covariance of F-score estimates across models, combined with per-model variance estimates, might facilitate analysis of whether differences in model performance, potentially estimated on different datasets, are statistically significant or not.

Acknowledgments This work is supported by the National Science Foundation (NSF) under III-1908727 and CCF-1937301, as well as the CMU Argo AI Center for Autonomous Vehicle Research. Vishnu Sarukkai is supported by the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program.

References

[1] Paul N. Bennett and Vitor R. Carvalho. Online stratified sampling: Evaluating classifiers at web-scale. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 1581–1584, New York, NY, USA, 2010. Association for Computing Machinery. 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020. 1, 3, 5

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1, 3, 5

[4] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996. 1

[5] Joseph L Doob. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169, 1935. 5

[6] William Falcon and Kyunghyun Cho. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*, 2020. 3

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 1

[8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017. 1

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 3, 5

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org, 2017. 2, 4

[11] Olivier Henaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020. 3

[12] Zenan Li, Xiaoxing Ma, Chang Xu, Chun Cao, Jingwei Xu, and Jian Lü. Boosting operational dnn testing efficiency through conditioning. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2019, page 499–509, New York, NY, USA, 2019. Association for Computing Machinery. 2

[13] Benjamin A. Miller, Jeremy Vila, Malina Kirn, and Joseph R. Zipkin. Classifier performance estimation with unbalanced, partially labeled data. In *Proceedings of The International Workshop on Cost-Sensitive Learning*, volume 88 of *Proceedings of Machine Learning Research*, pages 4–16, SDM, San Diego, California, USA, 05 May 2018. PMLR. 2, 5

[14] Phuc Nguyen, Deva Ramanan, and Charless Fowlkes. Active testing: An efficient and robust framework for estimating accuracy. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3759–3768. PMLR, 10–15 Jul 2018. 2

[15] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 1

[16] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. 2

[17] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 2, 4, 7

[18] Md Mustafizur Rahman, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. Efficient test collection construction via active learning. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 177–184, 2020. 2, 5

[19] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017. 1

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[21] A. Sabharwal and H. Sedghi. How good are my predictions? efficiently approximating precision-recall curves for massive datasets. In *UAI*, 2017. 2

[22] Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active estimation of f-measures. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NeurIPS'10, page 2083–2091. Curran Associates Inc., 2010. 2, 3, 4, 7

[23] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *arXiv preprint arXiv:2002.08721*, 2020. 3

[24] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 1

[25] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012. 1

[26] Begum Taskazan, Jiri Navratil, Matthew Arnold, Anupama Murthi, Ganesh Venkataraman, and Benjamin Elder. Not your grandfathers test set: Reducing labeling effort for testing, 2020. 2

[27] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010. 1

[28] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 5

[29] C Van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, pages 1–14, 1979. 1, 2, 3

[30] Byron C Wallace and Issa J Dahabreh. Class probability estimates are unreliable for imbalanced data (and how to fix them). In *2012 IEEE 12th international conference on data mining*, pages 695–704. IEEE, 2012. 4

[31] Peter Welinder, Max Welling, and Pietro Perona. A lazy man's approach to benchmarking: Semi-supervised classifier evaluation and recalibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 2

[32] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009. 2, 5

[33] Tiancheng Yu, Xiyu Zhai, and Suvrit Sra. Near optimal stratified sampling. *arXiv preprint arXiv:1906.11289*, 2019. 2

[34] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery. 2, 4, 7