Measuring Model Biases in the Absence of Ground Truth

Osman Aka*
Google

Ken Burke*
Google

Alex Bäuerle†
Ulm University

Christina Greer
Google

Margaret Mitchell‡

ABSTRACT

Recent advances in computer vision have led to the development of image classification models that can predict tens of thousands of object classes. Training these models can require millions of examples, leading to a demand of potentially billions of annotations. In practice, however, images are typically sparsely annotated, which can lead to problematic biases in the distribution of ground truth labels that are collected. This potential for annotation bias may then limit the utility of ground truth-dependent fairness metrics (e.g., Equalized Odds).

To address this problem, in this work we introduce a new framing to the measurement of fairness and bias that does not rely on ground truth labels. Instead, we treat the model predictions for a given image as a set of labels, analogous to a "bag of words" approach used in Natural Language Processing (NLP) [15]. This allows us to explore different association metrics between prediction sets in order to detect patterns of bias. We apply this approach to examine the relationship between identity labels, and all other labels in the dataset, using labels associated with *male* and *female*) as a concrete example. We demonstrate how the statistical properties (especially normalization) of the different association metrics can lead to different sets of labels detected as having "gender bias". We conclude by demonstrating that pointwise mutual information normalized by joint probability (nPMI) is able to detect many labels with significant gender bias despite differences in the labels' marginal frequencies. Finally, we announce an open-sourced nPMI visualization tool using TensorBoard.

CCS CONCEPTS

• **Computing methodologies** → *Machine Learning*.

KEYWORDS

datasets, neural networks, visual recognition, fairness, bias

*Equal contribution.
†Work conducted during internship at Google.
‡Work conducted while author was at Google.

n/a, in review, 2021

https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Osman Aka, Ken Burke, Alex Bäuerle, Christina Greer, and Margaret Mitchell. 2021. Measuring Model Biases in the Absence of Ground Truth. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The impact of algorithmic bias in computer vision models has been well-documented [c.f., 6, 21]. Examples of the negative fairness impacts of machine learning models include decreased pedestrian detection accuracy on darker skin tones [25], gender stereotyping in image captioning [7], and perceived racial identities possibly impacting unrelated labels [23]. Many of these examples are directly related to currently deployed technology, which highlights the urgency of solving these fairness problems as adoption of these technologies continues to grow.

Many common metrics for quantifying fairness in machine learning models, such as Statistical Parity [10], Equality of Opportunity [13] and Predictive Parity [8], rely on datasets with a significant amount of ground truth annotations for each label under analysis. However, modern applications of computer vision models often rely on datasets with relatively sparse ground truth . One reason for this is the significant growth in the number of predicted labels: the benchmark challenge dataset PASCAL VOC introduced in 2008 had only 20 categories [11], while less than 10 years later, the benchmark challenge dataset ImageNet provided hundreds of categories [19], and as systems have rapidly improved, it is now common to use the full set of ImageNet categories, which number more than 20,000 [1].

While the benefits of a large label space are clear, such as a more fine-grained ontology of the visual world, it also increases the complexity of implementing groundtruth-dependent fairness metrics. This concern is compounded by the common practice of collecting training datasets from different online resources [18]. This can lead to patterns where specific labels are omitted in a biased way, either through human bias (e.g., crowdsourcing where certain valid labels aren't "salient" enough to be tagged) or through algorithmic bias (e.g., selecting labels for human verification based on the predictions of another model). If the ground truth annotations in a sparsely labelled dataset are *themselves* potentially biased, then the premise of a fairness metric that "normalizes" to groundtruth patterns may be incomplete. In light of this, we argue that it is important to develop bias metrics that do not explicitly rely on "unbiased" ground truth labels.

In this work, we introduce a novel approach for measuring problematic model biases, focusing on the behavior of the model directly. This has several advantages compared to common fairness approaches in the context of large label spaces, making it possible

to identify skews as a function of the regular practice of running a model over a dataset. We study several different metrics that measure associations between labels, building upon work in Natural Language Processing [9] and information theory. We perform experiments on these association metrics using the Open Images Dataset [17] which has a large enough label space to illustrate how this framework can be generally applied, but we note that the focus of this paper is on introducing the relevant techniques, and do not require a specific dataset in particular. We demonstrate that normalized pointwise mutual information (nPMI) is particularly useful for understanding the correlations between labels and identity attributes in this setting, and can uncover stereotype-aligned correlations that the model has learned. This metric is particularly promising because:

- It requires no ground truth annotations.
- It can be used to provide insight into per-label correlations between model predictions and identity attributes.
- It is robust to low- and high-frequency labels.

Finally we announce an open-sourced visualization tool in TensorBoard that allows users to explore patterns of label bias in large datasets using the nPMI metric.

2 RELATED WORK

In 1990, Church and Hanks [9] introduced a novel approach to quantifying associations between words based on mutual information [12, 20] and inspired by psycholinguistic work on word norms [24] that catalogue words that people closely associate. For example, subjects respond more quickly to the word *nurse* if it follows a highly associated word such as *doctor*. Church and Hanks' proposed metric applies mutual information to words using a *pointwise approach*, measuring co-occurrences of distinct word pairs rather than averaging over all words. This enables a quantification of the question, "How closely related are these words?" by measuring their co-occurrence rates relative to chance in the dataset of interest. In this case, the dataset of interest is a computer vision evaluation dataset, and the words are the labels that the model predicts.

This information theoretic approach to uncovering word associations became a prominent method in the field of Natural Language Processing, with applications ranging from measuring topic coherence [2] to collocation extraction [5] to great effect, although often requiring a good deal of preprocessing in order to incorporate details of a sentence's syntactic structure. However, without preprocessing, this method functions to simply measure word associations regardless of their order in sentences or relationship to one another, treating words as an unordered set of tokens (a so-called "bag-of-words") [14].

As we show in this paper, this simple approach can be newly applied to an emergent problem in the machine learning ethics space: The identification of problematic associations that an ML model has learned. This approach is comparable to measuring correlations, although the common correlation metric of Spearman Rank [22] operates on assumptions that are not suitable for this task, such as linearity and monotonicity. The related correlation metric of the Kendall Rank Correlation [16] does not require such behavior, and we include comparisons with this approach.

Additionally, many potentially applicable metrics for this problem rely on simple counts of paired words, which does not take into consideration how the words are distributed with other words (e.g., sentence syntax or context); we will elaborate on how this information can be formally incorporated into a bias metric in the Discussion and Future Works section.

This work is motivated by recent research on fairness in machine learning (e.g., [13]), which at a high level seeks to define criteria that result in equal outcomes across different subpopulations. The focus in this paper is complementary to previous fairness work, honing in on ways to identify and quantify the specific problematic associations that a model may learn rather than providing an overall measurement of a model's unfairness. It also offers an alternative to fairness metrics that rely on comprehensive ground truth labelling, which is not always available for large datasets (e.g., the Open Images Dataset, which we experiment with in this work). We now turn to a formal description of the problem we seek to solve.

3 FAIRNESS AND ASSOCIATIONS METRICS

3.1 Problem Definition

We have a dataset $\mathcal{D}$ which contains image examples and labels generated by a image classifier. This classifier takes one image example and predicts "Is label y_i relevant to the image?" for each label in $\mathcal{L} = \{y_1, y_2, ..., y_n\}$. We infer $P(y_i)$ and $P(y_i, y_j)$ from $\mathcal{D}$ such that for a given random image in $\mathcal{D}$, $P(y_i)$ is the probability of having y_i as positive prediction and $P(y_i, y_j)$ is the joint probability of having both y_i and y_j as positive predictions. We further assume that we have identity labels $x_1, x_2 \in \mathcal{L}$ that belong to some protected subgroup for which we wish to compute a bias metric (e.g., *male* and *female*)[1]. We will measure bias with respect to these identity labels for all other labels $y \in \mathcal{L}$.

For ease of discussion in the rest of this paper, we notate any generic association metric as $A(x_j, y)$, where x_j is an identity label and y is any other label. We define a *gap* for label y between two identity labels $[x_1, x_2]$ with respect to the association metric $A(x_j, y)$ as $G(y|x_1, x_2, A(\cdot)) = A(x_1, y) - A(x_2, y)$. For example, the association between the labels *female* and *bike*, $A(female, bike)$ can be compared to the association between the labels *male* and *bike*, $A(male, bike)$. The difference between them is the gap for the label *bike*,
$G(bike|male, female, A(\cdot))$ = $A(female, bike)$ - $A(male, bike)$.

The first objective we are interested in is "Is the prediction of label y biased with respect to x_1 or x_2?". In some contexts, collecting ground truth labels for x_1 and x_2 is possible because it is not as costly as collecting ground truth labels for all labels in $\mathcal{L}$. In that setting, all formulations are equivalent except x_1 and x_2 can be based on the ground truth directly rather than based on classifier predictions.

The second objective is determining which labels have the largest discrepancy in this bias for y between x_1 and x_2. If x_1 and x_2 both belong to a common identity attribute (e.g., *male* and *female* both being labels related to gender), then one may consider them to fall

[1]For the rest of the paper, we focus on only two identity labels with notation x_1 and x_2 for simplicity, but it is straightforwardly extended to any number of identity labels by using pairwise comparison of all identity labels, or by using a one-vs-all gap (e.g. $A(x, y) - \mathbb{E}\left[A(x', y)\right]$ where x' is the set of all other x)

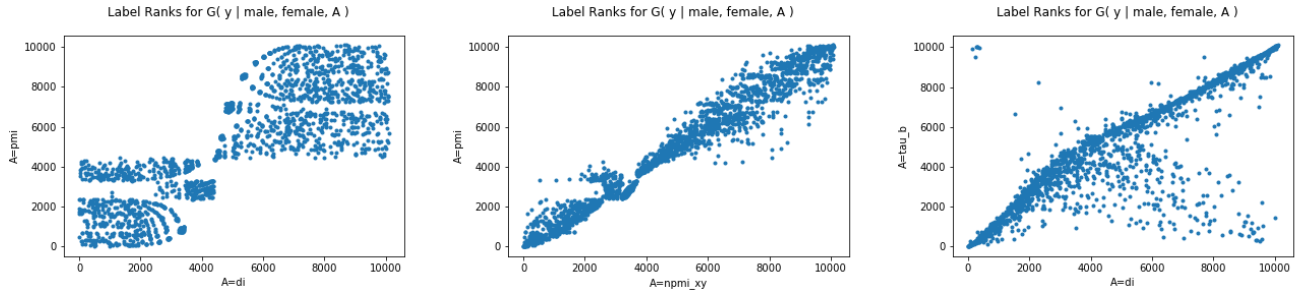

Figure 1: Label ranking shifts for metric-to-metric comparison.
Each point represents the rank of a single label when sorted by $G(y|x_1, x_2, A)$. The coordinates represent the rankings for different metrics A on the x and y axes. A highly-correlated plot along $y = x$ would imply that the two metrics lead to very similar bias rankings according to G.

along a shared identity dimension. In this context, one might consider the skew of y along this implicit dimension to be approximated by the difference in associations, as described in the *bike* example above. We choose this gender example because of the abundance of these specific labels in the Open Images Dataset, however this choice should not be interpreted to mean that gender representation is one-dimensional, nor that paired labels are required for the general[1] approach. Nonetheless, this simplification is important because it allows us to demonstrate how a single per-label approximation of "bias" can be measured between paired labels, and we leave details of further expansions to the Discussion in Section 5. As the count of labels $|\mathcal{L}|$ approaches tens of thousands and increases year by year for modern machine learning models, it is helpful to have a simple metric that can be computed and reasoned about at scale.

3.2 Fairness and Association Metrics

To ground our examination within a variety of potentially applicable approaches for this task, we consider several sets of related metrics $A(\cdot)$ that can be applied given the constraints of the problem at hand – limited groundtruth, non-linearity, and no assumptions about the underlying distribution of the data.

We first consider fairness metrics, where one of the most common fairness metrics, Demographic (or Statistical) Parity [4, 10, 13], a quantification of the legal doctrine of Disparate Impact [3], can be applied directly for the given task constraints.[2] Other metrics that are possible to adopt for this task include those based on Intersection-over-Union (IOU) measurements, and metrics based on correlation and statistical tests. We next describe each of these metrics in further detail and their relationship to the task at hand. In summary, we compare the following metrics:

- Fairnesss: Demographic Parity
- Entropy: Pointwise Mutual Information (PMI), Normalized Pointwise Mutual Information (nPMI).
- IOU: Sørensen-Dice Coefficient (SDC), Jaccard Index (JI).
- Correlation and Statistical Tests: Kendall Rank Correlation (τ_b), Log-Likelihood Ratio (LLR), T-test.

[2]Other common fairness metrics, such as Equality of Opportunity, require both an estimated and a groundtruth label, which makes the right way to apply them to this task less clear. We flag this for further work.

One of the important aspects of our problem setting is the counts of images with labels and label intersections, i.e., $C(y)$, $C(x_1, y)$, and $C(x_2, y)$. These values can span a large range for different labels y in the label set $\mathcal{L}$. Some metrics are theoretically more sensitive to the frequencies/counts of the label y as determined by their nonzero partial derivatives with respect to $P(y)$ (see Table 2). However, as we further discuss in Sections 4 and 5, our experiments indicate that in a real dataset, metrics with non-zero partial derivatives are best able to capture biases across a range of label frequencies. Differential sensitivity to label frequency could be problematic in practice for two reasons:

(1) It would not be possible to compare $G(y|x_1, x_2, A(\cdot))$ *between* labels y with different marginal frequencies (counts) $C(y)$.
(2) The alternative, bucketizing labels by marginal frequency and setting distinct thresholds per bucket, would add significantly more hyperparameters and essentially amount to manual frequency-normalization.

The following sections contain basic explanations of these metrics for a general audience, with the running example of *bike, male, female*. We leave further mathematical analyses of the metrics to the Appendix. However, integral to the application of nPMI in this task is the choice of normalization factor, and so we discuss this in further detail in Section 3.3.

Demographic Parity

$$G(y|x_1, x_2, DP) = P(y|x_1) - P(y|x_2)$$

Demographic Parity focuses on differences between the conditional probability of y given x_1 and x_2: How likely *bike* is for *male* vs $female$.

Entropy

$$G(y|x_1, x_2, PMI) = ln\left(\frac{P(x_1, y)}{P(x_1)P(y)}\right) - ln\left(\frac{P(x_2, y)}{P(x_2)P(y)}\right)$$

Pointwise Mutual Information, adapted from information theory, is the main entropy-based metric studied here. In this form, we are analyzing the entropy difference between $[x_1, y]$ and $[x_2, y]$.

This essentially examines how *bike* patterns in two different distributions: *male and female.*

There are several commonly used normalization techniques, the precise choice of which is crucial in our problem setting because of the imbalances in count values $C(y)$ for different labels across $\mathcal{L}$.

And so, for example, it can be harder to figure out how rare labels (like *libbertigibbet*) are behaving with respect to the chosen identity labels.

Remaining Metrics

We use the Sørensen-Dice Coefficient (SDC), which has the commonly-used F1-score as one of its variants; the Jaccard Index (JI), a common metric in Computer Vision also known as Intersection Over Union (IOU); Log-Likelihood Ratio (LLR), a classic flexible comparison approach; Kendall Rank Correlation, which is also known as τ_b-correlation, and is the particular *correlation* method that can be reasonably applied in this setting; and the t-test, a common statistical significance test that can be adapted in this setting [15]. Each of these metrics have different behaviours, however, we limit our mathematical explanation to the Appendix, as we found these metrics are either less useful in practice or behave similarly to other metrics in this use case (further described in the Appendix).

3.3 Normalizing PMI

As mentioned previously, one major challenge in our problem setting is the sensitivity of these association metrics to the frequencies/counts of the labels in $\mathcal{L}$. Some metrics are weighted more heavily towards labels with large marginal frequencies, $C(y)$, in spite of differences in joint probabilities along identity dimensions ($P(x_1, y), P(x_2, y)$). The same is true for other metrics that are weighted towards smaller marginal frequencies. Because of this problem, several different normalization techniques have been applied to PMI for years. Common normalizations include:

- $nPMI_y$: Normalizing each term by $P(y)$.
- $nPMI_{joint}$: Normalizing the two terms by $P(x_1, y)$ and $P(x_2, y)$, respectively.
- PMI^2: Using $P(x_1, y)^2$ and $P(x_2, y)^2$ instead of $P(x_1, y)$ and $P(x_2, y)$, the normalization effects of which are further illustrated in the Appendix.

Each of these normalization methods have different impacts on the PMI metric. The main advantage of these normalizations is the ability to compare metric gaps *between* label pairs $[y_1, y_2]$ (e.g., comparing the gender skews of two labels like *long hair* and *dido flip*) even if $P(y_1)$ and $P(y_2)$ are very different. In the Experiments section, we discuss which of these is most effective and meaningful for the fairness and bias use case motivating this work.

4 EXPERIMENTS

In order to compare these metrics, we use the Open Images Dataset (OID) [17]. According to the dataset description, the images are diverse and often contain complex scenes with several objects, with machine-generated labels (see [17] for details on the model). This dataset is very useful for demonstrating realistic use cases as well because of the number of labels, which is nearly 20,000. The label space is also diverse, including objects (*cat*, *car*, *tree*), materials (*leather*, *basketball*), moments/actions (*blowing out candles*, *woman playing guitar*), and scene descriptors and attributes (*beautiful*, *smiling*, *forest*). Because the spectrum of labels is broad, and exhaustive ground truth collection is impractical, there is significant potential for bias and fairness concerns.

Metrics	Min/Max $C(y)$	Min/Max $C(x_1, y)$	Min/Max $C(x_2, y)$
PMI	15/10551	1/1059	8/7755
PMI^2	15/10551	1/1059	8/7755
LLR	15/10551	1/1059	8/7755
DP	6104/785045	628/239950	5347/197795
JI	4158/562445	399/144185	3359/183132
SDC	2906/562445	139/144185	2563/183132
$nPMI_y$	35/562445	1/144185	9/183132
$nPMI_{xy}$	34/270748	1/144185	20/183132
τ_b	6104/785045	628/207723	5347/183132
T-test	960/562445	72/144185	870/183132

Table 1: Minimum and maximum count values for top100 labels for metrics gaps

In our experiments, we apply each of the association metrics $A(x_i, y)$ on these machine-generated predictions for identity labels $x_1 = male, x_2 = female$ and all other labels y' in the Open Images Dataset $\mathcal{L}$. We then compute the metric gap for each label y, for the different metrics. Finally, we use these values to compare and contrast the metrics.

4.1 Label Ranks

The first experiment we performed is to compute the metrics, and the gaps – $A(x_1, y)$, $A(x_2, y)$ and $G(y|x_1, x_2, A)$ – over the OID dataset. We then sorted the labels by G and studied how the distribution of ranks differed between different A. Figure 1 shows examples of these metric-to-metric label rank distribution comparisons (all other metric-to-metric comparisons can be found in the Appendix). Here, we plot how a single label can have quite a different ranking when using two different metrics for computing "bias".

When comparing metrics, we found that they grouped together in a few clusters based on similar ranking patterns when sorting by $G(y|male, female, A(\cdot))$. In the first cluster, pairwise comparisons between PMI, PMI^2 and LLR show linear relationships when sorting labels by G. Indeed, while some labels show modest changes in rank between these metrics, they share *all* of their top 100 labels, and >99% of label pairs maintain the same relative ranking between metrics. By contrast, there are only 7 labels in common between the top 100 labels of PMI^2 and SDC. Due to the similar behavior of this cluster, we therefore focus on PMI moving forward, and see the Appendix for further details on these relationships.

Similar results were obtained for another cluster of metrics: DP, JI, and SDC. All pairwise comparisons generated linear plots, with about 70% of the top 100 labels shared in common between metrics when sorted by $G(\cdot)$. Furthermore, about 95% of pairs of those overlapping labels maintained the same relative ranking between metrics. We chose to focus on Demographic Parity moving forward due to its mathematical simplicity and prominence in fairness literature.

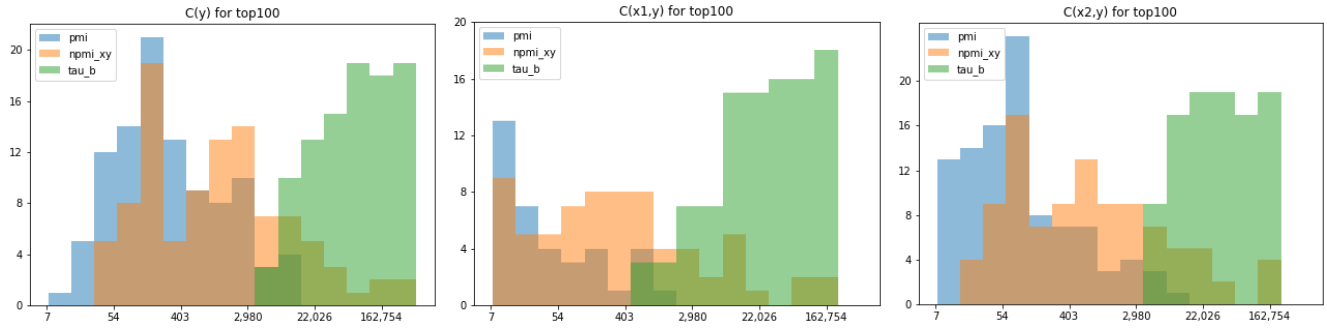

Figure 2: Top 100 count distributions for PMI, nPMI$_{xy}$, and τ_b

The distribution of $C(y)$, $C(x_1, y)$, and $C(x_2, y)$ for the top 100 biased labels for each metric. The x-axis is the logarithmic-scaled bins and y-axis is the number of labels which have the corresponding count values in that bin. Each metric sums up to 100.

	$\partial p(y)$	$\partial p(x_1, y)$
∂DI	0	$\frac{1}{p(x_1)}$
∂PMI	0	$\frac{p(x_1)}{p(x_1,y)p(y)}$
$\partial nPMI_y$	$\frac{ln(\frac{p(x_2,y)}{p(x_1,y)})}{ln^2(p(y))p(y)}$	$\frac{1}{ln(p(y))p(x_1,y)}$
$\partial nPMI_{xy}$	$\frac{ln(\frac{p(x_2,y)}{p(x_1,y)})}{ln(p(x_2,y))ln(p(x_1,y))p(y)}$	*check App-3.2
∂PMI^2	0	$\frac{p(x_1)}{p(x_1,y)p(y)} + \frac{1}{ln(p(x_2,y))p(x_1,y)}$
∂SDC	$\frac{p(x_1,y)}{(p(x_1)+p(y))^2} - \frac{p(x_2,y)}{(p(x_2)+p(y))^2}$	$\frac{1}{(p(x_1)+p(y))^2}$
∂JI	*check App-3.2	$\frac{p(x_1)+p(y)}{(p(x_1)+p(y)-p(x_1,y))^2}$
∂LLR	0	$\frac{1}{p(x_1,y)}$
$\partial\tau_b$	*check App-3.2	*check App-3.2
∂T-test	*check App-3.2	*check App-3.2

Table 2: Metric orientations.

This table shows the partial derivatives of the metrics with respect to $P(y)$ and $P(x_1, y)$. It is useful for understanding sensitivity of the metrics for different probability values of $P(y)$, $P(x_1, y)$, and $P(x_2, y)$.

We next sought to understand how incremental changes to the counts of the labels, $C(y)$, affect these rankings in a real dataset. To achieve this, we added a fake label to the real labels of OID, setting initial values for its counts and co-occurences in the dataset, $P(y)$, $P(x_1, y)$, and $P(x_2, y)$. Then we incrementally increased or decreased the count of label y, and measured whether its bias ranking in G would change relative to the other labels in OID. We repeated this procedure for different orders of magnitude of label count $C(y)$ while maintaining the ratio $P(x_1, y)/P(x_2, y)$ as constant.

This experiment allowed us to determine whether the theoretical sensitivities of each metric to label count $C(y)$ (as determined by partial derivatives) would hold in the context of real-world data, where the underlying distribution of label frequencies may not be uniform. If certain subsets of the distribution of the label distribution are relatively sparse, for example, then the *rank* of our hypothetical label may not change even if the metric $A(\cdot)$ is itself dependent on the label count. However, in practice we do not observe this behavior in the tested settings (see Appendix for plots of these experiments), where label rank moves with label count roughly as predicted by the partial derivatives in Table 2. In fact, we observed that metrics with larger partial derivatives for x_1, x_2, or y often led to a larger change in rank. For example, slightly increasing $P(x_1, y)$ when y always co-occurs with x_1, $P(y) = P(x_1, y)$ affects ranking more for $A = nPMI_y$ compared to $A = PMI$ (see Appendix).

4.2 Top 100 Labels by Metric Gaps

When applying these metrics to fairness use cases, model owners may often be most interested in detecting the labels with the largest metric gaps. If one filters results to a "top K" label set, then the normalization chosen could lead to vastly different sets of purportedly "biased" labels (e.g. as mentioned earlier, PMI^2 and SDC only shared 7 labels in their top 100 set for OID).

To further analyze this issue, we calculated simple values for each metric's top 100 labels sorted by G: minimum and maximum values of $C(y)$, $C(x_1, y)$ and $C(x_2, y)$ as shown in Table 1. The most salient point is that the clusters of metrics from Section 4.1 also appear to hold in this analysis as well; PMI, PMI^2, and LLR have low $C(y)$, $C(x_1, y)$ and $C(x_2, y)$ ranges, whereas DP, JI, and SDC have relatively high ranges. Another straightforward observation we can make is that the nPMI$_y$ and nPMI$_{xy}$ ranges include the other metrics' ranges especially for the counts of the identity attributes and a given label, $C(x_1, y)$ and $C(x_2, y)$.

To demonstrate this point more clearly, we plot the distributions of these counts for PMI, NPMI$_y$ and τ_b in Figure 2 (all other combinations can be found in the Appendix). These three metric distributions show that PMI exclusively detects labels with low counts, whereas τ_b almost exclusively detects those with much higher counts. The exception is nPMI$_y$ and nPMI$_{xy}$; these two metrics are cable of capturing labels across a range of marginal frequencies.

5 DISCUSSION

In the previous section, we first showed that some of these association metrics behave very similarly when ranking labels from the Open Images Dataset. We then showed that the mathematical orientations and sensitivity of these metrics align with experimental results from OID. Finally, we showed that the different normalizations affect whether labels with high or low marginal frequencies are likely to be detected as having a significant bias according to $G(y|x_1, x_2, A(\cdot))$ in this dataset.

Of particular note is the difference between PMI and nPMI (both nPMI_y and nPMI_{xy}). The effect of normalizing PMI in practice is that labels with larger marginal counts can achieve high ranks in $G(y|x_1, x_2, nPMI)$ alongside labels with smaller marginal counts. While it is true that $\partial PMI/\partial p(y) = 0$, whereas this derivative for nPMI is non-zero, in practice the labels with smaller counts can achieve very large $P(x_1, y)/P(x_2, y)$ ratios merely by reducing the denominator to a single image example. By contrast, the normalizations we use for nPMI still allow us to capture a significant amount of rare labels in the top 100 labels by $G(y|x_1, x_2, nPMI)$, as indicated by the ranges in Table 1 and Figure 2.

Indeed, if the evaluation set is properly designed to match the distribution of use cases of the model "in the wild", then we argue more common labels that have a smaller $P(x_1, y)/P(x_2, y)$ ratio are still critical to audit for biases. Normalization strategies must be titrated carefully to balance this simple ratio of joint probabilities with the label's rarity in the dataset.

An alternative solution to this problem could be bucketing labels by their marginal frequency. We argue this is a suboptimal solution for two reasons. First, determining even a single threshold hyperparameter is a painful process for defining fairness constraints. Systems that prevent models from being published if their fairness discrepancies exceed a threshold would then be required to titrate this threshold for every bucket. Secondly, bucketing labels by frequency is essentially a manual and discontinuous form of normalization anyways; we argue that building normalization into the metric directly is a more elegant solution.

Finally, to enable detailed investigation of the model predictions, we implemented and open-sourced a tool to visualize nPMI metrics as a TensorBoard plugin for developers[3]. It allows users to investigate discrepancies between two or more identity labels and their pairwise comparisons. Users can visualize probabilities, image counts, sample and label distributions, and filter, flag, and download these results.

6 CONCLUSION AND FUTURE WORK

In this paper we have described association metrics that can measure biases in the large label space of current computer vision classification models. These metrics do not require exhaustive ground truth annotations for each sample, which allows them to be applied in contexts where it is difficult to apply standard fairness metrics such as Equality of Opportunity [13]. According to our experiments, Normalized Pointwise Mutual Information is a particularly useful metric for measuring specific biases in a real-world dataset with a large label space, e.g. the Open Images Dataset.

This work also introduces several questions for future work. The first is whether nPMI is also similarly useful as a bias metric in small label spaces (e.g., credit and loan applications). Second, if we were to have exhaustive ground truth labels for such a dataset, how would the sensitivity of nPMI in detecting biases compare to ground-truth-dependent fairness metrics? Finally, in this work we treated the labels predicted for an image as a flat set. However, just like sentences have rich syntactic structure beyond this "bag-of-words" model in NLP, images also have rich structure and relationships between objects that are not captured by mere binary co-occurrence rates. This opens up the possibility that *within-image label relationships* could be leveraged to better understand how concepts are associated in a large computer vision dataset. We leave these questions for future works.

[3] https://github.com/tensorflow/tensorboard/tree/master/tensorboard/plugins/npmi

REFERENCES

[1] 2020. ImageNet. *http://image-net.org/explore* (2020). accessed 6.Oct.2020.

[2] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Association for Computational Linguistics, Potsdam, Germany, 13–22. https://www.aclweb.org/anthology/W13-0102

[3] Solon Barocas and Andrew D. Selbst. 2014. Big Data's Disparate Impact. *SSRN eLibrary* (2014).

[4] Richard Berk. 2016. A primer on fairness in criminal justice risk assessments. *The Criminologist* 41, 6 (2016), 6–9.

[5] G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *(Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[7] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. *CoRR* abs/1803.09797 (2018). arXiv:1803.09797 http://arxiv.org/abs/1803.09797

[8] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1610.07524 [stat.AP]

[9] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29. https://www.aclweb.org/anthology/J90-1003

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. *CoRR* abs/1104.3913 (2011). arXiv:1104.3913 http://arxiv.org/abs/1104.3913

[11] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (Jan. 2015), 98–136. https://doi.org/10.1007/s11263-014-0733-5

[12] Robert M Fano. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics* 29 (1961), 793–794.

[13] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[14] Zellig Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162. https://doi.org/10.1007/978-94-009-8467-7_1

[15] Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Pearson Prentice Hall, Upper Saddle River, N.J. http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y

[16] M. G. KENDALL. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika* 30, 1-2 (06 1938), 81–93. https://doi.org/10.1093/biomet/30.1-2.81 arXiv:https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf

[17] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR* abs/1811.00982 (2018). arXiv:1811.00982 http://arxiv.org/abs/1811.00982

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

	DP		PMI		nPMI_{xy}	
Ranks	Label	Count	Label	Count	Label	Count
0	Female	265853	Dido Flip	140		610
1	Woman	270748	Webcam Model	184	Dido Flip	140
2	Girl	221017	Boho-chic	151		2906
3	Lady	166186		610	Eye Liner	3144
4	Beauty	562445	Treggings	126	Long Hair	56832
5	Long Hair	56832	Mascara	539	Mascara	539
6	Happiness	117562		145	Lipstick	8688
7	Hairstyle	145151	Lace Wig	70	Step Cutting	6104
8	Smile	144694	Eyelash Extension	1167	Model	10551
9	Fashion	238100	Bohemian Style	460	Eye Shadow	1235
10	Fashion Designer	101854		78	Photo Shoot	8775
11	Iris	120411	Gravure Idole	200	Eyelash Extension	1167
12	Skin	202360		165	Boho-chic	460
13	Textile	231628	Eye Shadow	1235	Webcam Model	151
14	Adolescence	221940		156	Bohemian Style	184

Figure 3: Top 15 labels
Top 15 labels according $G(y|x_1, x_2, A(\cdot))$. Identity label names are omitted.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[20] Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 3 (1948), 379–423. http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48

[21] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. (2018).

[22] C. Spearman. 1904. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology* 15 (1904), 88–103.

[23] Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 498–512.

[24] M. P. Toglia and W. F. Battig. 1978. *Handbook of semantic word norms.* Lawrence Erlbaum.

[25] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. *CoRR* abs/1902.11097 (2019). arXiv:1902.11097 http://arxiv.org/abs/1902.11097