The Spotlight: A General Method for Discovering Systematic Errors in Deep Learning Models

Greg d'Eon
gregdeon@cs.ubc.ca
University of British Columbia
Vancouver, Canada

Jason d'Eon
jndeon@dal.ca
Dalhousie University
Halifax, Canada

Kevin Leyton-Brown
kevinlb@cs.ubc.ca
University of British Columbia
Vancouver, Canada

James R. Wright
james.wright@ualberta.ca
University of Alberta
Edmonton, Canada

ABSTRACT

Supervised learning models often make systematic errors on rare subsets of the data. When these subsets correspond to explicit labels in the data (e.g., gender, race) such poor performance can be identified straightforwardly. This paper introduces a method for discovering systematic errors that do not correspond to such explicitly labelled subgroups. The key idea is that similar inputs tend to have similar representations in the final hidden layer of a neural network. We leverage this structure by "shining a spotlight" on this representation space to find contiguous regions in which the model performs poorly. We show that the Spotlight surfaces semantically meaningful areas of weakness in a wide variety of existing models spanning computer vision, NLP, and recommender systems, and we verify its performance through quantitative experiments.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**.

KEYWORDS

deep learning, auditing, fairness, distributional robustness

ACM Reference Format:
Greg d'Eon, Jason d'Eon, Kevin Leyton-Brown, and James R. Wright. 2022. The Spotlight: A General Method for Discovering Systematic Errors in Deep Learning Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3531146.3533240

1 INTRODUCTION

Despite their superhuman performance on an ever-growing variety of problems, deep learning models that perform well on average often make systematic errors, performing poorly on semantically coherent subsets of the data. A landmark example is the Gender Shades study [3], which showed that vision models for gender recognition tend to exhibit abnormally high error rates when presented with images of black women. AI systems have also been shown to perform poorly for marginalized groups in object recognition [12], speech recognition [25], mortality prediction [7], and recruiting tools [7]. Other systematic errors can be harder for practitioners to anticipate in advance. Medical imaging classifiers can be sensitive to changes in the imaging hardware [6, 11]; essay scoring software can give high scores to long, poorly-written essays [36]; and visual question-answering systems can fail when questions are rephrased [41].

Recognizing and mitigating such errors is critical to avoid designing systems that will exhibit discriminatory or systematically unreliable behaviour. These issues have led the community to develop better tools for testing model performance, clearer standards for reporting model biases, and a plethora of methods for training more equitable or robust models. Even when making repairs is difficult or infeasible, identifying and flagging edge cases where systems fail can also help expert users work around an algorithm's flaws [5]. However, these methods require practitioners to recognize and label well-defined groups in their datasets ahead of time, necessarily overlooking semantically related sets of inputs that are not identified in advance. While practitioners certainly should explicitly assess model performance on sensitive subpopulations, it is extremely difficult to anticipate all of the sorts of inputs upon which models might systematically fail: for example, vision models could perform poorly on a particular age group, pose, background, lighting condition, etc.

In this work, we introduce *the Spotlight*, a method for finding systematic errors in deep learning models even when the common feature linking these errors was not anticipated by the practitioner and hence was not surfaced via an explicit label. Our key idea is that similar inputs tend to have similar representations in the final hidden layer of a neural network. We leverage this similarity by "shining a spotlight" on this representation space, searching for contiguous regions in which the model performs poorly. Using the final layer in this manner makes the Spotlight agnostic to most details of the neural network, making it easily applicable to a wide range of datasets and model architectures.

We demonstrate the Spotlight's broad applicability through qualitative experiments on a variety of otherwise dissimilar models and datasets, including image classifiers, language models, and

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533240

recommender systems. These spotlights identified several kinds of systematic errors. Some were unexpected to us (for example, failures of language models on Spanish text in an otherwise monolingual English dataset), showing that spotlights can discover novel failure modes. Others found known issues (for example, failures of facial recognition systems on black faces), which are well documented in prior work and are readily confirmed by existing group labels, establishing that the Spotlight does not overlook important existing issues. While some spotlights are more difficult to interpret or might require more domain expertise to understand, on balance, our findings provide evidence that the Spotlight can identify systematically meaningful areas of weakness in many disparate domains. Additionally, we validate the Spotlight's performance through two quantitative experiments, showing that the Spotlight is more effective than a standard clustering baseline at finding high-loss groups in several datasets, and that it can consistently find synthetically-generated systematic errors, even when they span multiple labels.

We hope that practitioners will add the Spotlight to their model development pipelines to complement their existing auditing and training tools. Rather than replacing existing methods for measuring biases in trained models, the Spotlight augments these auditing practices, identifying failure modes that do not correspond to known labels. Further, our results demonstrate that the Spotlight's findings are complementary to other error discovery methods, and the issues that the Spotlight discovers can often be addressed outside of the training loop by collecting higher-quality data, adjusting the model architecture, limiting the model's use cases, or flagging the problems to expert users – all solutions that avoid making performance tradeoffs with robust optimization methods. In support of these goals, we provide an open-source implementation of the Spotlight.[1]

2 RELATED WORK

Systematic errors on known groups. A standard approach for auditing a machine learning model is to create a dataset partitioned by group information, (e.g., demographics, lighting conditions, hospital ID, …), and to check whether the model exhibits poor performance on any of these groups. With one or two particularly sensitive group variables, it is straightforward to check the model's performance on each group; e.g., the NLP community advocates for including such disaggregated evaluations in model cards [31]. When there are more group variables—inducing exponentially many intersectional groups—there exist a variety of computational methods for efficiently identifying groups for which performance is poor [8, 27, 35] and model dashboards for interactively exploring groups [1, 4, 44].

Once a systematic model weakness has been identified, a substantial literature proposes methods for making repairs. Fair machine learning methods can incorporate group information into shallow models, requiring that the model perform similarly on each group [see, e.g., 9, for a review] or evaluating the model on its worst group [29]. There also exist generalizations to support many intersectional groups [24]. Additionally, two recent, robust training methods exist to repair deep models that exhibit such biases. First, distributionally robust language modelling [34] allows an adversary to change the distribution over groups during training, requiring the model to do well on each group. Second, invariant learning [2] requires the model to learn a representation of the data that induces the same classifier for each group, protecting against spurious correlations. The Spotlight differs from all of these approaches by aiming to identify systematic failure modes that go beyond existing group labels. Our approach is thus complementary to those just surveyed: after using the Spotlight to uncover a new failure mode, practitioners can augment their datasets with appropriate labels and turn to one of these existing methods to monitor or repair their model's performance.

Systematic errors on unknown groups. When it is difficult to predict in advance which subgroups of inputs may be problematic for a model, practitioners are faced with the task of examining a model's errors and searching for regularities—a process that Oakden-Rayner et al. [32] refer to as *error auditing*. In some cases, this process is entirely manual: e.g., some medical applications rely on experts to dig deeply into a model's false positives and false negatives [28]. To alleviate this manual effort, Oakden-Rayner et al. also discuss the possibility of directing the error discovery process with algorithmic tools, such as clustering algorithms; accordingly, we compare against a Gaussian mixture model baseline in several of our experiments. The Spotlight aims to serve as another such algorithmic tool, finding semantically coherent groups of inputs to flag to experts, but it differs from standard clustering algorithms in that it searches more directly for groups where a model performs poorly.

The existing method most similar to our approach is GEORGE [42], a robust optimization method that does not rely on group labels; indeed, we evaluate it in Section 4. GEORGE infers "subclasses" within the dataset by clustering points within a trained neural network's representation space, then allows an adversary to modify the distribution over subclasses. While Sohoni et al. focused on the more difficult problem of training robust models, they observe that the clusters identified in their first stage tend to correspond to semantically meaningful subsets of the data (for instance, images of birds on land vs. on water). They also observe that their reliance upon a superlinear-time clustering method limits its applicability to large datasets. The Spotlight exploits the same underlying insight as GEORGE: that semantic similarity will correlate with proximity in the embedding space. However, the Spotlight has several advantages: it avoids partitioning the entire embedding space, searching only for contiguous, high-loss regions; it is able to identify issues that involve examples from multiple classes; and it runs in linear time.

Another notable method is Errudite [45], an interactive system for analyzing errors made by NLP models. It allows a user to query a subset of their dataset using a domain-specific language, reporting the model's performance on this query set and proposing related queries to help the user dig deeper into their model. This interactive query system can help developers discover and confirm systematic issues in their models without the need for pre-existing group labels. However, Errudite is very specifically designed for use with models for NLP tasks; in contrast, the Spotlight is domain agnostic, applying to deep models designed for many different domains.

Finally, the spotlight is inspired by related methods for training robust models that perform well across the entire dataset; such

[1] https://github.com/gregdeon/spotlight

a model is guaranteed to achieve similar performance on any semantically meaningful subset of the data. For example, distributionally robust optimization (DRO) methods allow an adversary to change the relative importance of each data point during training [14, 19, 26], and invariant learning algorithms can infer groups during training [10]. While these methods sometimes leverage information from a trained model's representation space, they are required to carefully constrain the set of possible groups in order to keep the optimization problem tractable. For example, the standard DRO algorithm [19] assigns equally high importance to all examples on which losses are high. In contrast, the Spotlight only aims to surface model biases to a human expert, and hence focuses on semantically similar subsets of high-loss inputs.

3 THE SPOTLIGHT

We would like to identify subsets of the data that emphasize poor performance of a model. However, sets of data points upon which the model performs badly may have little semantic similarity, making them ineffective tools for model auditing. We propose instead to search the model's final layer embedding space to identify one or more contiguous sets of points of limited size ("spotlights") that maximize loss (Figure 1). We allow for "soft assignment" of points into the spotlight, making a spotlight a kind of soft clustering; however, note that our method is driven by a supervised objective function and that it pays attention only to the loss of points that fall inside the spotlight rather than seeking to partition the entire dataset.

Formally, suppose that we have N data points with final-layer representations $x_1, \ldots, x_N \in \mathbb{R}^d$ and losses $\ell_1, \ldots, \ell_N \in \mathbb{R}$. A *spotlight* is a set of weights $k_1, \ldots, k_n \in [0, 1]$, calculated using the kernel

$$k_i = \max(1 - \tau \|x_i - \mu\|_2^2, 0), \tag{1}$$

where $\mu \in \mathbb{R}^d$ is the spotlight's center in the model's embedding space, and $\tau \in \mathbb{R}$ is the precision of the spotlight, with large precisions producing small spotlights and vice versa. Notice that k_i has a maximum of 1 when $x_i = \mu$ and a minimum of 0 when x_i is sufficiently far from μ; intermediate values of k_i allow for "soft assignment" of points into the spotlight. Then, we wish to solve the optimization problem

$$\max_{\mu,\tau} \quad \sum_{i=1}^{N} \left(\frac{k_i}{\sum_j k_j} \right) \ell_i \tag{2}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} k_i \geq S, \tag{3}$$

for some choice of the hyperparameter S. We interpret S as the "spotlight size," as this setting ensures a lower bound on the total weight that the spotlight assigns across the dataset.

To make optimization tractable, we replace the hard constraint $\sum_i k_i \geq S$ with a soft penalty term in the objective. In preliminary tests, we considered using penalty terms that are positive only when the constraint is violated, but we found that the optimization process frequently got stuck in suboptimal local maxima. Instead, we apply a quadratic penalty for coming close to the constraint, penalizing reweightings that include fewer than $S + w$ points, and shrink the value of w throughout the optimization. Specifically, we define the penalty term as

$$p(k) = C \cdot \max \left(\frac{(\sum_{i=1}^{N} k_i - (S + w))^2}{w^2}, 0 \right) \tag{4}$$

and optimize the unconstrained objective

$$\sum_{i=1}^{N} \left(\frac{k_i}{\sum_j k_j} \right) \ell_i - p(k). \tag{5}$$

We fix C to be large relative to typical losses in the dataset (e.g., $C = 1$ for binary classification; $C = 10$ for problems with thousands of classes).

We optimized Equation (5) by beginning with a large, diffuse spotlight containing the entire dataset, initializing to $\mu = 0$ and $\tau = 10^{-4}$. We ran the Adam optimizer for 5000 steps with an adaptive learning rate, halving the learning rate each time the objective reached a plateau, and shrinking the width of the barrier geometrically from $w = N - S$ to $w = 0.05S$.

Optimizing multiple spotlights. In practice, models often have several distinct failure modes. So far we have only shown how to find a single spotlight, highlighting a single systematic error. We can find additional spotlights by decreasing the losses of points to the extent that they participated in previous spotlights and then running the same procedure again. More formally, after computing spotlight weights k_i, we update the losses to

$$\ell_i' := \left(1 - \frac{k_i}{\max_j k_j} \right) \ell_i, \tag{6}$$

and then compute a new spotlight as above using these updated losses. This process can be repeated until it starts finding spotlights with relatively low error or semantic coherence. Each of our experiments presents three to five spotlights obtained in this way.

Choosing a spotlight size. Our procedure has one free parameter, the spotlight size S. Since it is not possible to quantify whether a spotlight has captured a semantically meaningful subset of a dataset, this parameter cannot be set automatically. However, it is straightforward to optimize spotlights of several different sizes and compare them qualitatively. During testing, we generally found that very small spotlights (around 0.1% of the dataset) were too selective to identify sets of points linked by a meaninful semantic trait, whereas very large spotlights (around 10% of the dataset) were too inclusive to focus on high-loss data points. For example, Figure D1 shows spotlights ranging in size from 0.1% to 10% on FairFace (see Section 4.1); the smallest spotlight is difficult to describe, while the largest spotlight has a substantially lower average loss. Taking this balance into account, in our experiments, we settled on a spotlight size of 2% for vision models, where images can be simply be scanned for cohesion, and a spotlight size of 5% for non-vision models, where we found it necessary to describe spotlights using summary statistics.

4 EXPERIMENTS

The Spotlight is model-agnostic: it can be applied to any deep model that exposes its final layer representations and for which per-input losses are available. We demonstrate this flexibility by using the Spotlight to qualitatively evaluate a broad range of classification

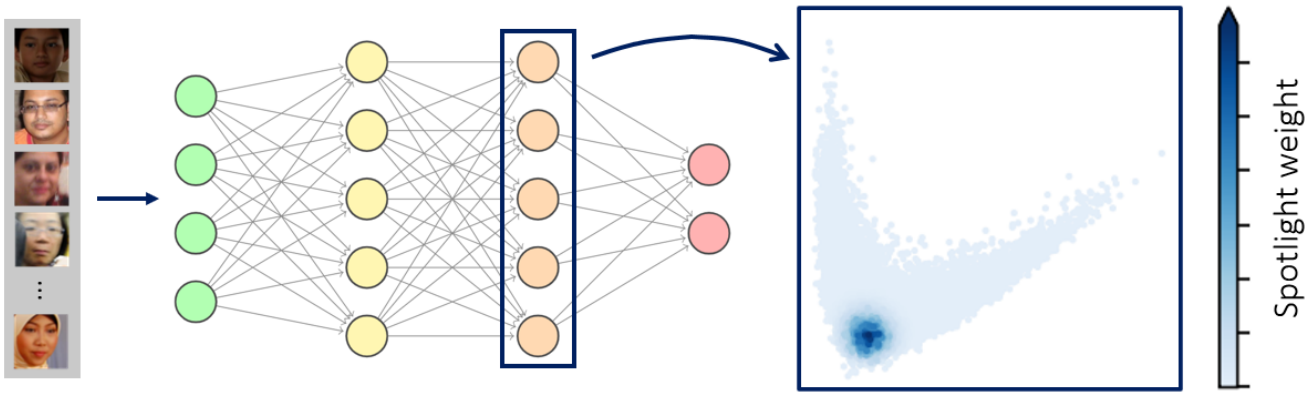


Figure 1: An example of a spotlight in a image classification model's representation space.

models from the literature, spanning image classification (faces; objects; x-rays), NLP (sentiment analysis; question answering), and recommender systems (movies). Our goal was to investigate how useful the Spotlight would be for auditing models used in production, and so we we focused on popular, publicly available pre-trained models whenever possible and tested the models' performance on datasets that were not seen during training. In each domain, we found one or more spotlights providing evidence of distinct, systematic failure modes, without leveraging any information about existing group labels. Full results can be found for every dataset we tested (most in the appendix), showing the method's broad applicability. We ran all experiments using a single NVIDIA Tesla V100 GPU; each spotlight presented in this section took under 1 minute to identify, emphasizing the computational tractability of our approach even on very large datasets.

In these qualitative experiments, we compared our spotlights to two baselines. First, we found the examples with the highest losses on each dataset, demonstrating the issues that DRO's reweightings would uncover. On most of our datasets, these high loss examples are difficult to interpret, highlighting quirks of the datasets more than systematic issues with the models. Second, because spotlights are contiguous (weighted) sets of points in a model's representation space, clustering algorithms offer a sensible baseline. We thus compared our spotlights with clusters identified by GEORGE. In particular, we use the publicly available implementation of GEORGE[2], which separately clusters examples from each class, automatically selecting the number of clusters using a Silhouette-based heuristic. Wherever feasible, we identified the three highest-loss clusters, summarized the examples in each of these clusters, and compared them to our spotlights. Overall, we found that our spotlights often differed from GEORGE's clusters, identifying both more granular problem areas within classes and systematic errors that span multiple classes.

Finally, while these qualitative experiments test the Spotlight's ability to identify semantically meaningful areas of weakness, they do not explicitly test whether it identifies the most egregious systematic errors in each model. We round out this section with two quantitative experiments, where we compared the Spotlight, GEORGE, and a standard Gaussian mixture model [32].[3] In Section 4.6, we test each method's ability to identify synthetically-generated issues that span multiple classes, showing that the Spotlight can find these issues more consistently than either baseline. We follow up with Appendix B, where we provide a quantitative comparison between the clusters found by the Spotlight and the clustering baseline. These results confirm that the Spotlight's optimization algorithm successfully finds systematic issues, and that these issues would often be difficult to identify using an off-the-shelf clustering algorithm.

4.1 FairFace

We first studied FairFace [22], a collection of 100,000 face images annotated with crowd-sourced labels about the perceived age, race, and gender of each face. FairFace is notable for being approximately balanced across 7 races and 2 genders. In particular, we trained a model to predict the perceived gender label as a proxy for the gender prediction systems studied in prior work [3]. Our model was a ResNet-18, trained using Adam with cross-entropy loss and a learning rate of 3e-4; we stopped training after 2 epochs when we found that the validation loss began increasing. We ran the Spotlight on the validation set, using the final 512-dimensional hidden layer as the representation space.

The spotlights and highest-loss examples are shown in Figures 2 and D2. We found that each of the spotlights discovered a strikingly different set of faces. The first shows a set of profile (i.e., side) views; the second consists mostly of young children; the third contains a preponderance of faces that are shadowed or partially occluded. The fourth and fifth spotlights consist of black faces in poor lighting and Asian faces, respectively. (These demographic disparities are summarized in Figure D3, which illustrates the distribution of ages and races on each spotlight.) Overall, our spotlights identified that the model performs poorly on pictures of very young and old people

[2]https://github.com/HazyResearch/hidden-stratification

[3]We chose to only compare to a clustering baseline on these quantitative terms. Like the Spotlight, a clustering method finds contiguous sets of points in representation space, so it is likely to produce sets of inputs that are semantically coherent. These experiments test whether the resulting clusters can also identify high-loss regions.

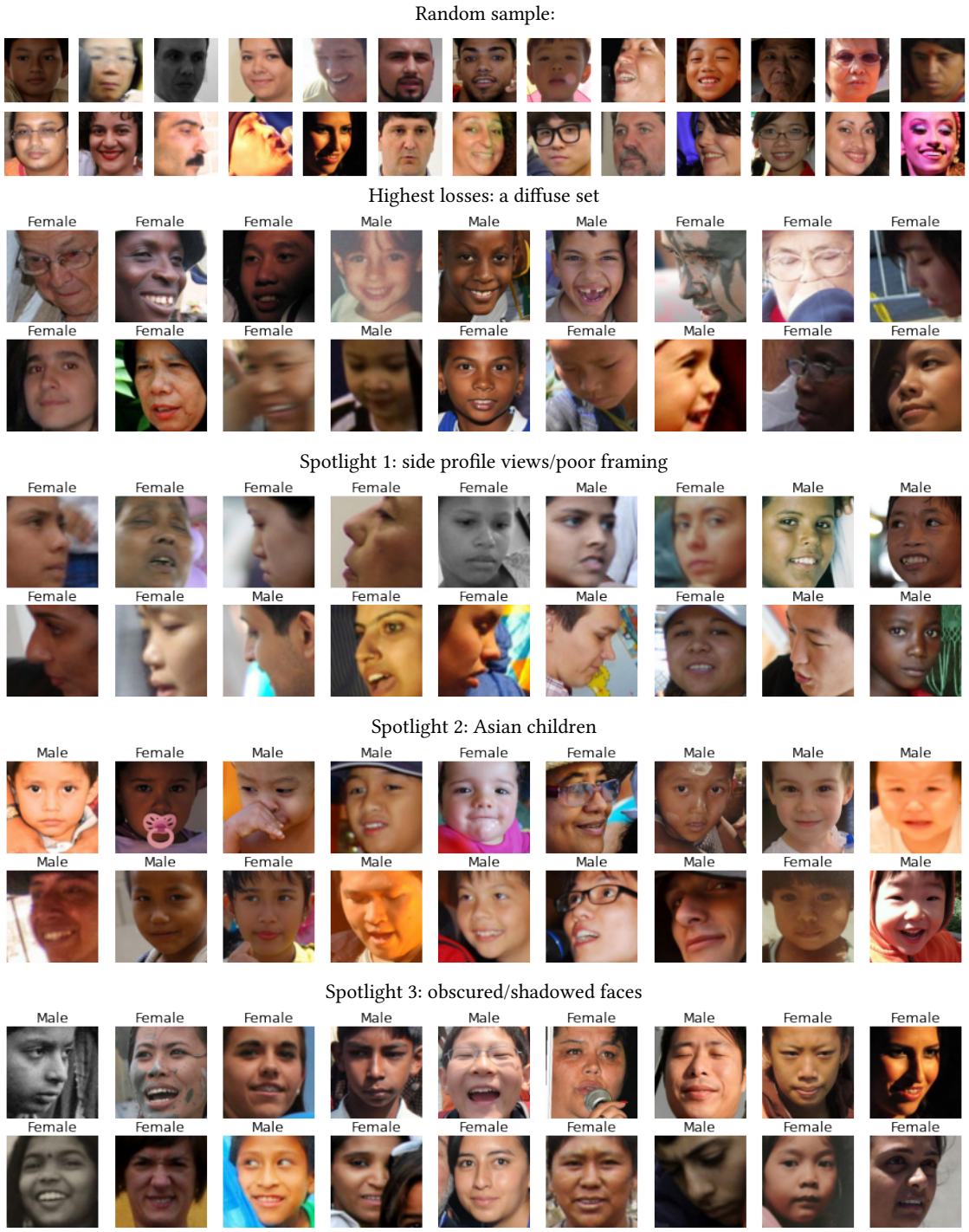


Figure 2: Spotlights on FairFace validation set. Image captions list true label.

and of Black people without access to these demographic labels; it also identified additional, semantically meaningful groups for which labels did not exist.

In comparison, the high-loss images are an unstructured set of examples that include occluded faces, poor lighting, blurry shots, and out-of-frame faces.

GEORGE identified a total of six clusters. A random sample of the images from the three clusters with the highest average losses are shown in Figure D2. Notably, all of the images in each cluster have the same label, but share little in common beyond their labels. While several images in cluster 2 include people wearing glasses, most of GEORGE's clusters display a combination of camera angles, lighting conditions, and occlusions that made the images difficult for us to classify.

4.2 ImageNet

For a second vision dataset, we study the pre-trained ResNet-18 model from the PyTorch model zoo [37], running the Spotlight on the 50,000 image validation set. As in FairFace, we used the final 512-dimensional hidden layer of the model as the representation space.

Our results are shown in Figures 3 and D5. Each spotlight identified a set of images having a wide variety of distinct but semantically similar labels. The first spotlight contains a variety of images of people working, where it is difficult to tell whether the label should be about the person in the image, the task they're performing, or another object; the second shows a variety of tools; the third shows a variety of green plants, where there is often an animal hiding in the image; the fourth identifies some food and people posing; and the fifth shows a variety of dogs.

In contrast, the high-loss images appear to have little structure, with many of them having unexpected labels, such as "pizza" for an image of a squirrel in a tree holding a piece of pizza.

We did not run GEORGE on this model, as the clustering algorithm failed when attempting to split the 50 images in each class into even smaller clusters. We are nevertheless able to conclude that GEORGE's clusters would be much different from the spotlights just described. The reason is that GEORGE's "subclass" clusters deaggregate existing classes, whereas the Spotlight tended to reveal semantically similar groups of classes that the model had trouble distinguishing.

4.3 Sentiment analysis: Amazon reviews

Next, we turn to the Amazon polarity dataset [46], a collection of 4 million plain-text Amazon reviews labelled as "positive" (4-5 stars) or "negative" (1-2 stars). We used a popular pre-trained checkpoint of a DistilBERT model from Huggingface [21], which was fine-tuned on SST-2. We ran the Spotlight on a sample of 20,000 reviews from the validation set, using the final 768-dimensional hidden layer as the representation space.

We found it more difficult to spot patterns in the spotlights on this dataset by simply reading the highest-weight reviews, so we instead summarized each spotlight by identifying the tokens that appeared most frequently in the spotlight distributions, relative to their frequencies in the validation set. These results are shown in Figure 4. Remarkably, the first spotlight surfaced reviews that were written in Spanish, which the model consistently classified as negative. We determined that the model was only trained on English sentences; its tokenizer appears to work poorly on Spanish sentences. The second spotlight highlighted long-winded reviews of novels, which the model has difficulty parsing. The third found reviews that mention aspects of customer service, such as product returns, which the model confidently classifies as negative; these predictions lead to high losses on positive reviews that describe customer service.

The highest-loss reviews in the dataset are quite different, consisting almost entirely of mislabelled reviews. For example, one review reads "The background music is not clear and the CD was a waste of money. One star is too high.", but has a 4-5 star rating; dozens of high-loss outliers follow this pattern, where the rating clearly contradicts the review text. We note that this type of label noise would pose a problem for many robust optimization methods, which could insist that the model learn to memorize these outliers rather than focusing on other important portions of the dataset.

GEORGE found a total of 11 clusters; the three with the highest losses are also summarized in Figure 4. The first cluster is similar to our second spotlight, containing many negative and wordy reviews for novels and movies that are misleading or difficult to parse. The second consists entirely of positive reviews, including many written in Spanish. The third is small, only containing 0.5% of the dataset, and we found it difficult to summarize. Overall, we found the spotlights more coherent, but observed more overlap between spotlights and GEORGE clusters than in other datasets.

4.4 MovieLens 100k

We investigated a third domain, recommender systems. Specifically, we considered the MovieLens 100k dataset [17], a collection of 100,000 movie reviews of 1,000 movies from 1,700 different users. It also includes basic information about each movie (titles, release dates, genres) and user (age, gender, occupation), which we use during the analysis, but did not make available to the model. For our model, we used a deep factorized autoencoder [18], using the final 600-dimensional hidden layer for our representation space.

The highest-weight movies in each spotlight are shown in Figure D7. The first spotlight mostly identifies 3–4 star action and adventure films rated by prolific users, where the model is highly uncertain about which review they will give. The second finds reviews of highly rated drama films from a small group of users with little reviewing history. The third shows unpopular action and comedy films, where the model is nonetheless optimistic about the rating. In comparison, the highest-loss predictions consist mostly of 1-star ratings on movies with high average scores.

GEORGE identified 21 clusters; we show the highest-loss predictions from three in Figure D8. The first two consist of a variety of 1- or 2-star ratings respectively, where the model confidently makes 4- or 5-star predictions for both categories. Both clusters have similar genre distributions to the entire dataset. The third cluster instead contains many 5-star ratings on comedy and drama films where the model is skeptical about these high ratings. The GEORGE clusters differ from the spotlights, which tend to have more consistent movies or genres, but less consistent ratings.

4.5 Additional Datasets: SQuAD; Chest X-Rays

We ran the Spotlight on two additional datasets: SQuAD, an NLP question-answering benchmark; and a chest x-ray image dataset. Our results here were more ambiguous, but we describe these experiments regardless to emphasize the Spotlight's generality and to reassure the reader that we have presented all of our findings

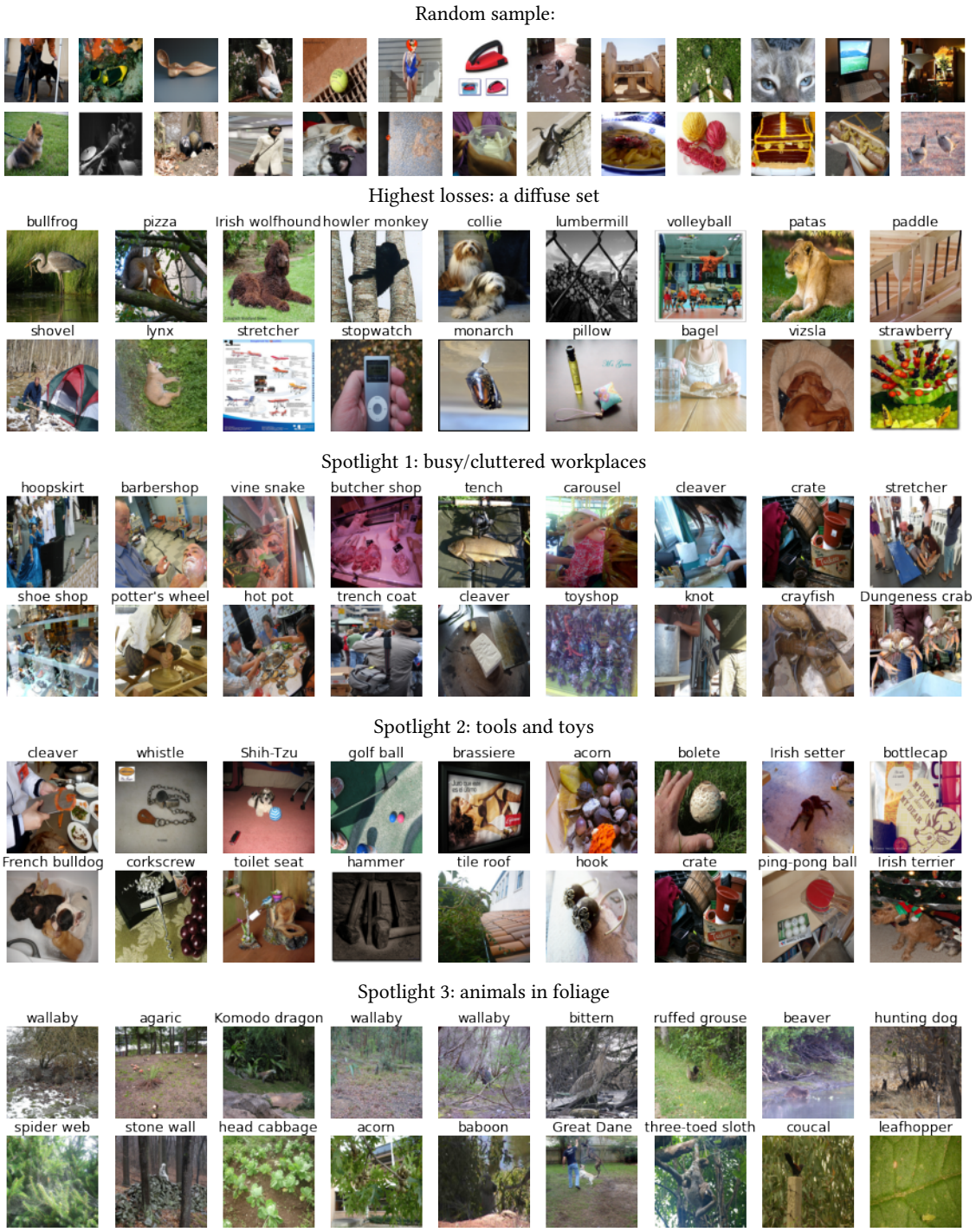


Figure 3: Spotlights on ImageNet validation set. Image captions list true label.

rather than cherry-picking favourable results. Full details can be found in the appendix.

SQuAD. The Stanford question answering dataset (SQuAD) [38] is a benchmark of question–answer pairs constructed from 536 Wikipedia articles. We analyzed a pre-trained DistilBERT model [20] fine-tuned on this dataset, running spotlights on the test set. We excluded long examples where the sum of the context and answer

Subset	Avg Length	Frequent words
High loss	68.8	length, outdated, potter, bubble, contact, cinematography, adjusting, functions, stock, versus
Spotlight 1: Spanish	79.1	que, est, como, y, las, tod, es, la, si, por
Spotlight 2: novels	88.7	super, wearing, job, prefer, bigger, hang, discover, killing, slip, source
Spotlight 3: customer service	80.7	problem, returned, que, hoping, ok, unfortunately, returning, however, las, maybe
GEORGE cluster 1	77.6	ok, moore, fiction, okay, above, potter, thank, cinematography, usa, jean
GEORGE cluster 2	75.9	que, para, y, est, como, es, tod, las, installation, la
GEORGE cluster 3	87.7	visuals, sugar, investment, score, study, surfer, dimensional, era, dune, scarlet

Figure 4: Spotlight on Amazon reviews.

sequence lengths was greater than 384, leaving 10,386 question–answer pairs. It was unclear which representation to use for the SQuAD spotlights, as the model's last layer has a representation for each token in the context rather than a single representation for the entire example. We chose to discard the representations of all tokens except for the first one—a special `[CLS]` token prepended to each example—because BERT's representation of this token is trained to summarize the entire text through a "next sentence prediction" task [13].

The results are summarized in Figure D11. The spotlights particularly highlighted questions from the "packet switching" and "civil disobedience" categories, which were the two categories with the highest loss, despite not having had access to category labels; explicit consideration of individual questions with the highest losses identified the latter but not the former. We found little semantic structure beyond these high-level categories; a richer representation space is likely required to get more insight into this dataset.

Chest X-Rays. The chest x-ray dataset consists of 6,000 chest x-rays labelled as "pneumonia" or "healthy" [23]. Using the Spotlight, we were able to identify at least two semantically meaningful failure modes in this domain: images with a text label "R" on their sides, and images with very high contrast. However, such images were also relatively easy to identify in the set of high-loss inputs, so we were unconvinced that the Spotlight offered a decisive benefit in this domain. More broadly, it became clear to us that our lack of expertise in radiology made it impossible for us to determine whether data points in other spotlighted clusters shared more fundamental semantic relationships, reminding us that model auditing requires enough domain knowledge to assess the coherence of failure modes. Separately, we were also unsure of whether our model gave rise to a meaningful embedding space: the dataset is small, leading to a risk of overfitting, and in this case we were unable to leverage an existing, pre-trained model.

4.6 Synthetic Evaluation

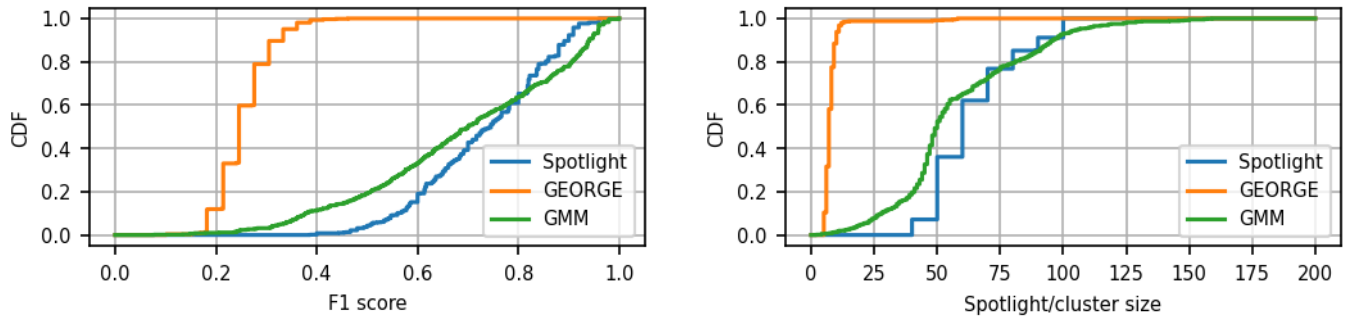
While the qualitative results in this section demonstrate the Spotlight's ability to identify areas of weakness in a variety of domains, these findings are subjective. To back up these claims, we ran a brief quantitative experiment, where we generated synthetic datasets with known systematic errors and evaluated the methods' ability to recover these errors.

Specifically, we began with a superclassed version of ImageNet [15] containing 3000 images from 10 superclasses (dog, bird, insect, monkey, car, feline, truck, fruit, fungus, boat), each consisting of images from 6 subclasses (e.g., the dog superclass consists of Chihuahua, Japanese spaniel, Maltese dog, Pekinese, Shih-Tzu, and Blenheim spaniel). We chose an arbitrary subclass and randomized the (superclass) labels of all 50 images in this subclass, creating a group of semantically coherent examples where our ImageNet model performs poorly. Then, without retraining the model, we ran the Spotlight, GEORGE, and a standard Gaussian mixture model (GMM) clustering algorithm on the model's representations.

The Spotlight depends on a hyperparameter: the number of data points to include. We could simply have set this value to 50, but this may have biased results in favor of our methods. We thus tuned the spotlight size on this dataset, running spotlights ranging from 10 to 100 points in size and keeping the spotlight that was the best predictor of the model's misclassified examples.[4] For the GMM, we fit 60 clusters, keeping the cluster with the highest average loss. For GEORGE, also we kept the cluster with the highest loss. We evaluated each cluster on its F1 score—the harmonic mean of its precision and recall—at detecting the erroneous subclass.

The results from 600 runs of this experiment are shown in Figure 5. The Spotlight tended to achieve F1 scores around 0.7, detecting most of the examples from the subclass while including relatively few points from other subclasses. Some GMM clusters had comparable performance to the Spotlight, but many had much lower F1 scores. Investigating these clusters revealed that the GMM produced clusters varying wildly in size; small clusters typically had low precision while large clusters typically had low recall. GEORGE

[4]We computed each spotlight's F1 score on predicting which examples the model classified incorrectly and kept the spotlight with the highest score. Note that this differs from the F1 score reported in the results, and does not require any knowledge of an existing systematic error.



(a) (b)

Figure 5: Results from 600 runs of our synthetic experiment, showing (a) F1 score and (b) cluster size.

had even lower F1 scores. It almost always produced small clusters with high recall but low precision: as the errors span multiple classes, GEORGE cannot capture them in a single subclass cluster. Overall, these results show that the Spotlight can consistently discover systematic errors, even when they span multiple class labels. Additional plots in Appendix C confirm these precision–recall tradeoffs and demonstrate performance after including multiple clusters.

5 FUTURE DIRECTIONS

Our methods give rise to various promising directions for future work, many of which we have begun to investigate. This section describes some of these ideas along with our initial findings.

Using the Spotlight for adversarial training. While this paper advocates for the Spotlight as a method for auditing deep learning models, it also gives rise to a natural, adversarial objective that could be optimized during training in the style of the distributionally robust methods surveyed earlier [14, 26]. That is, model training could iterate between identifying a spotlight distribution, reweighting the input data accordingly, and minimizing loss on this reweighted input. A model that performed well on this objective would have very balanced performance, distributing inputs with poor performance diffusely across the representation space. Unfortunately, our preliminary tests suggest that optimizing for this objective is not simple. With large spotlights (10% of dataset), we found that this method made little difference, with the model improving more slowly than in regular training; with smaller spotlights (1%), the model struggled to learn anything, fluctuating wildly in performance between epochs. We intend to continue investigating approaches for training against this flexible adversary.

Structure in representations. An important assumption that the Spotlight makes is that nearby points in the representation space will tend to correspond to semantically similar inputs. While this assumption is empirically supported both by our results and by prior work [30, 32, 33, 39, 42, 43], it is an emergent property of deep learning models, and we do not currently understand this property's sensitivity to details of the architecture and training method. For instance, does the choice of optimizer (SGD/Adam, weight decay, learning rate, ...) affect the representation space in a way that interacts with the Spotlight? Could we instead leverage representations learned by alternative models, such as autoencoders?

Spurious correlations. In particular, models that have learned spurious correlations could act quite differently from the models we studied in this work. For example, consider a model trained on the Waterbirds dataset [40], where it is possible to reach high training accuracy by learning to recognize land/water backgrounds instead of correctly identifying land/water birds. The model's representation would then focus mostly on details of the backgrounds, and spotlights would be unable to substantially change the distribution of bird types. Investigating a model's representation spaces with tools like the Spotlight could help to understand why a model is failing on a particular distribution shift.

Comparing to Domino. While this paper was under review, we became aware of late-breaking work by several of the authors of GEORGE. Their new method, Domino [16], builds on the Spotlight (citing this paper's arXiv preprint). Like the Spotlight, Domino attempts to identify contiguous regions in a model's embedding space in which a model makes systematically incorrect predictions. A key advantage of Domino is that it aims to find multiple systematic biases in a model simultaneously, whereas the Spotlight identifies such biases sequentially (effectively, corresponding to a greedy algorithm). Domino works by fitting a Gaussian mixture model to a model's entire representation space, augmented with information about the predictions and labels for each input. Given that Domino takes model loss into account and fits Gaussians in the representation space, we expect it to perform well. However, since Domino aims to cluster the entire embedding space, we expect that it could exhibit weaker performance on high-loss regions than the Spotlight, which focuses solely on modeling these regions. A thorough comparison of the two approaches is an important direction for future work.[5]

[5]The Domino paper does already include a comparison to the Spotlight, but the performance they describe is inconsistent with our own experiences. We are concerned that these experiments may have used low learning rates that would cause the Spotlight to perform poorly.

6 CONCLUSIONS

The Spotlight is an automatic and computationally efficient method for surfacing semantically related inputs upon which a deep learning model performs poorly. In experiments, we repeatedly observed that the Spotlight was able to discover meaningful groups of problematic inputs across a wide variety of models and datasets, including poorly modelled age groups and races, ImageNet classes that were difficult to distinguish, reviews written in Spanish, and specific movies with unpredictable reviews. These findings often complemented systematic issues identified by GEORGE's clustering stage, a related auditing method. The Spotlight found all of these sets without access to side information such as demographics, topics, or genres.

The Spotlight's ability to discover systematic errors in deep learning models makes it well-suited for use in a broader feedback loop of developing, auditing, and mitigating models. The Spotlight is useful in the auditing stage of this loop, helping practitioners to discover semantically meaningful areas of weakness that they can then test in more depth and address through changes to their pipeline. For instance, a machine learning engineer armed with our results might seek higher-quality data for poorly represented demographics (on FairFace), switch to a multi-label classification model (on ImageNet), restrict use of their model to English text (on Amazon reviews), or avoid using their model on users with little data (on MovieLens). Such a human-in-the-loop discovery process is critical to identify systematic failure modes in deep learning systems and mitigate them before they are able to cause harmful consequences in deployed systems.

Potential social impacts. The Spotlight can be a useful addition to machine learning practitioners' toolboxes, augmenting their existing robustness tests. However, the Spotlight can only show the existence of a systematic error, not prove that a model has none. It is possible that a practitioner could get a false sense of security if the Spotlight turns up no significant issues—perhaps because their model's biased representation hides an important issue, or because they miss a systematic error on visual inspection. On balance, we believe that the potential to uncover new issues outweighs the risk of believing there are none, especially when the Spotlight is used in concert with other fairness or robustness methods.

Additionally, it is conceivable that the Spotlight could be used for debugging harmful AI systems, such as surveillance technology, to identify regimes under which these technologies fail and to further improve their efficacy. This is unavoidable: the Spotlight is general enough to work on a wide range of model architectures, including those that might cause negative social impacts. Overall, though, we do not see this as a likely use case; the Spotlight's main likely effect would be helping practitioners to increase the fairness and robustness of deployed deep learning systems and to gain confidence that their models to not systematically discriminate against coherent subpopulations of users.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments. This work was supported by Compute Canada, a GPU grant from NVIDIA, an NSERC Discovery Grant, a DND/NSERC Discovery Grant Supplement, a CIFAR Canada AI Research Chair at the Alberta Machine Intelligence Institute, and DARPA award FA8750-19-2-0222, CFDA #12.910, sponsored by the Air Force Research Laboratory. Addtionally, resources used in preparing this research were provided, in part, by NSERC, the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (www.vectorinstitute.ai/#partners).

REFERENCES

[1] Yongsu Ahn and Yu-Ru Lin. 2019. FairSight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1086–1095.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[3] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[4] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[5] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359206

[6] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* 25, 8 (2019), 1301–1309.

[7] Irene Y Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 3543–3554.

[8] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, and Steven Euijong Whang. 2018. Slice finder: Automated data slicing for model validation. *CoRR* abs/1807.06068 (2018). arXiv:1807.06068 http://arxiv.org/abs/1807.06068

[9] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[10] Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. 2020. Exchanging lessons between algorithmic fairness and domain generalization. *CoRR* abs/2010.07249 (2020). arXiv:2010.07249 https://arxiv.org/abs/2010.07249

[11] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, and Matthew D et. al. Hoffman. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[12] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 52–59.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] John C. Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. 2020. Distributionally robust losses for latent covariate mixtures. *CoRR* abs/2007.13982 (2020). arXiv:2007.13982 https://arxiv.org/abs/2007.13982

[15] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. 2019. Robustness (Python Library). https://github.com/MadryLab/robustness

[16] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. 2022. Domino: Discovering Systematic Errors with Cross-Modal Embeddings. https://openreview.net/pdf?id=FPCMqjI0jXN

[17] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[18] Jason Hartford, Devon Graham, Kevin Leyton-Brown, and Siamak Ravanbakhsh. 2018. Deep models of interactions across sets. In *International Conference on Machine Learning*. PMLR, 1909–1918.

[19] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[20] HuggingFace. 2021. DistilBERT base uncased distilled SQuAD. https://huggingface.co/distilbert-base-uncased-distilled-squad. Accessed: 2021-05-28.

[21] HuggingFace. 2021. DistilBERT base uncased finetuned SST-2. https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english. Accessed: 2021-05-28.

[22] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.

[23] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 5 (2018), 1122–1131.

[24] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/3306618.3314287

[25] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[26] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. 2020. Fairness without demographics through adversarially reweighted learning. *CoRR* abs/2006.13114 (2020). arXiv:2006.13114 https://arxiv.org/abs/2006.13114

[27] Jinyang Li, Y. Moskovitch, and H. V. Jagadish. 2021. DENOUNCER: Detection of unfairness in classifiers. *Proc. VLDB Endow.* 14 (2021), 2719–2722.

[28] Vidur Mahajan, Vasantha Kumar Venugopal, Murali Murugavel, and Harsh Mahajan. 2020. The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it. *Academic radiology* 27, 1 (2020), 132–135.

[29] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 6755–6764. https://proceedings.mlr.press/v119/martinez20a.html

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (2013). arXiv:1301.3781 [cs.CL]

[31] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[32] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*. 151–159.

[33] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill* 3, 3 (2018), e10.

[34] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060* (2019).

[35] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Identifying biased subgroups in ranking and classification. *CoRR* abs/2108.07450 (2021). arXiv:2108.07450 https://arxiv.org/abs/2108.07450

[36] Les Perelman. 2014. When "the state of the art" is counting words. *Assessing Writing* 21 (2014), 104–111.

[37] PyTorch. 2021. Torchvision models. https://pytorch.org/vision/stable/models.html. Accessed: 2021-05-28.

[38] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[39] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and Measuring the Geometry of BERT. *Advances in Neural Information Processing Systems* 32 (2019). https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf

[40] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks. In *International Conference on Learning Representations*.

[41] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6649–6658.

[42] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems* 33 (2020).

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]

[44] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The What-If Tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[45] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 747–763. https://doi.org/10.18653/v1/P19-1073

[46] Xiang Zhang, Junbo Zhao, and Yann Lecun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems* 28 (2015).

A DATASET DETAILS

In Figure A1, we summarize licensing and content considerations for each of the datasets used in this work.

B QUANTITATIVE RESULTS

While the qualitative results in the main text demonstrate the Spotlight's ability to identify semantically meaningful groups of inputs, one might wonder whether it successfully identifies high loss regions in the model's representation space. As a sanity check, we confirmed that the spotlights have higher average losses than the clusters found by a naive clustering baseline, which does not use information about the model's losses during optimization. In particular, we fit a Gaussian mixture model to each representation space and compared the Spotlight to the cluster with the highest average loss. We fit 50 clusters on FairFace and ImageNet and 20 clusters on Amazon and MovieLens, ensuring that a typical cluster would be comparable in size to the spotlights, and compared the size and average loss of each of these clusters to the spotlights.

The results are shown in Figure B1. On FairFace, ImageNet, and MovieLens, the first spotlight had higher loss than any of the clusters, showing that spotlight effectively identified a high-loss region. On Amazon, two clusters had higher average loss than the first spotlight, but they incorporated considerably fewer points, finding a less widespread error. These results indicate that the spotlights reliably use information about the model's losses to identify a large systematic error, while a naive clustering method tends to split high-loss regions across several clusters or identify smaller failure modes.

C ADDITIONAL SYNTHETIC RESULTS

In this section, we include additional results from our synthetic experiments in Section 4.6:

- Figure C1: the precision and recall of the clusters identified by each method. The Spotlight has both moderate precision and recall; GEORGE has very low precision but high recall; and the GMM shows more variability in both metrics.
- Figure C2: changes in precision and recall as additional clusters are added. The spotlights tend to overlap, producing few new points in each cluster; GEORGE splits the single systematic issue across many clusters; the GMM quickly loses recall as more points from outside of the subclass are added. (Note that spotlights in these two plots have their sizes fixed at 50 points, rather than having their sizes tuned for each dataset.)

D ADDITIONAL SPOTLIGHTS

In this section, we include additional outputs from the spotlights on the datasets that were described in the text. In particular, we include:

- FairFace: examples of spotlights of different sizes in Figure D1; fourth and fifth spotlights and GEORGE clusters in Figure D2; demographic info in Figure D3 and average losses by demographic in Figure D4
- ImageNet: fourth and fifth spotlights in Figure D5
- MovieLens: high loss ratings in Figure D6; spotlight examples in Figure D7; GEORGE clusters in Figure D8.
- Chest x-rays: random sample, high loss examples, and first three spotlights in Figure D9; final two spotlights in Figure D10
- SQuAD: common words and topics from each spotlight in Figure D11

Dataset	License	PII	Offensive content
FairFace	CC BY 4.0	none	none
ImageNet	custom non-commercial	none	none
Amazon reviews	Apache 2.0	none	Offensive words in reviews are censored
SQuAD	CC BY 4.0	none	none
MovieLens 100K	custom non-commercial	none	none
Chest x-rays	CC BY 4.0	none	none
Adult	MIT	none	none
Wine quality	MIT	none	none

Figure A1: Details for each of the datasets used in this paper.

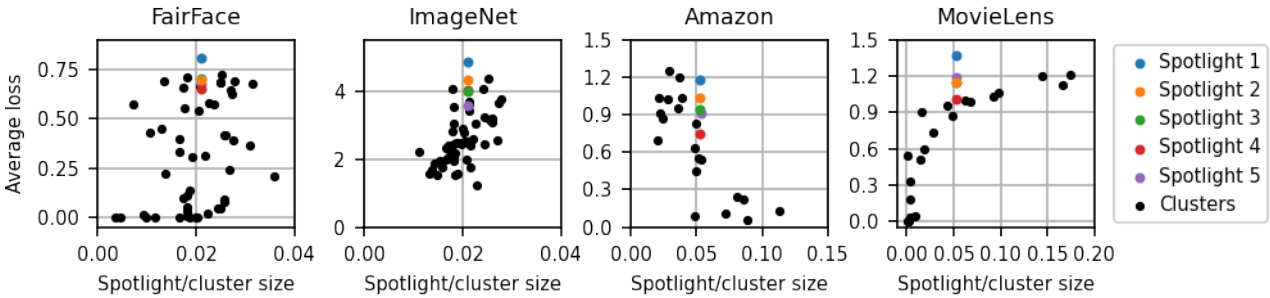
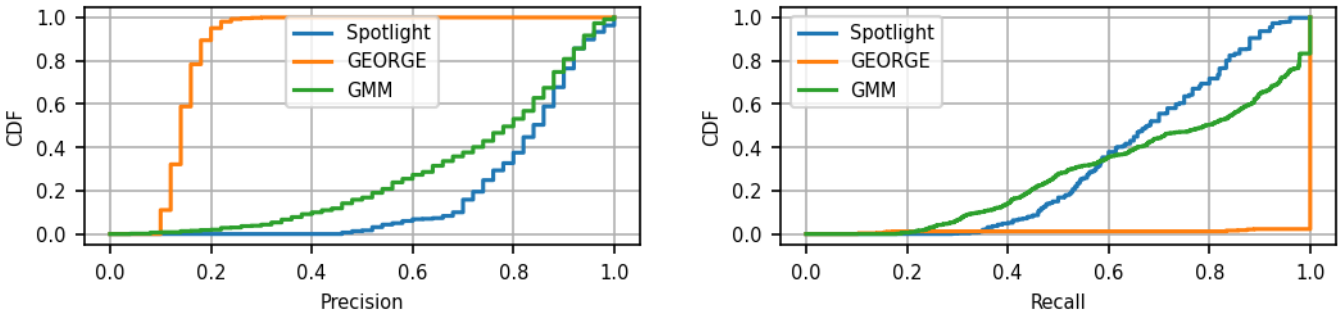
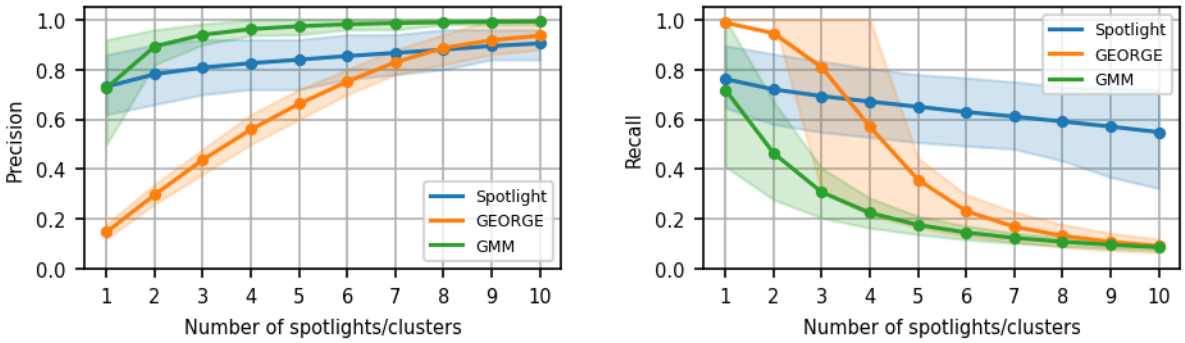


Figure B1: Sizes and average losses for clusters and spotlights. FairFace and ImageNet show spotlights containing 2% of the dataset and 50 clusters; Amazon reviews and MovieLens show spotlights containing 5% of the dataset and 20 clusters.



(a) (b)

Figure C1: Additional results from our synthetic experiment, showing (a) precision and (b) recall of each method's clusters.



(a) (b)

Figure C2: Additional results from our synthetic experiment, showing how precision and recall change as multiple clusters are included. Filled area shows 1 standard deviation.

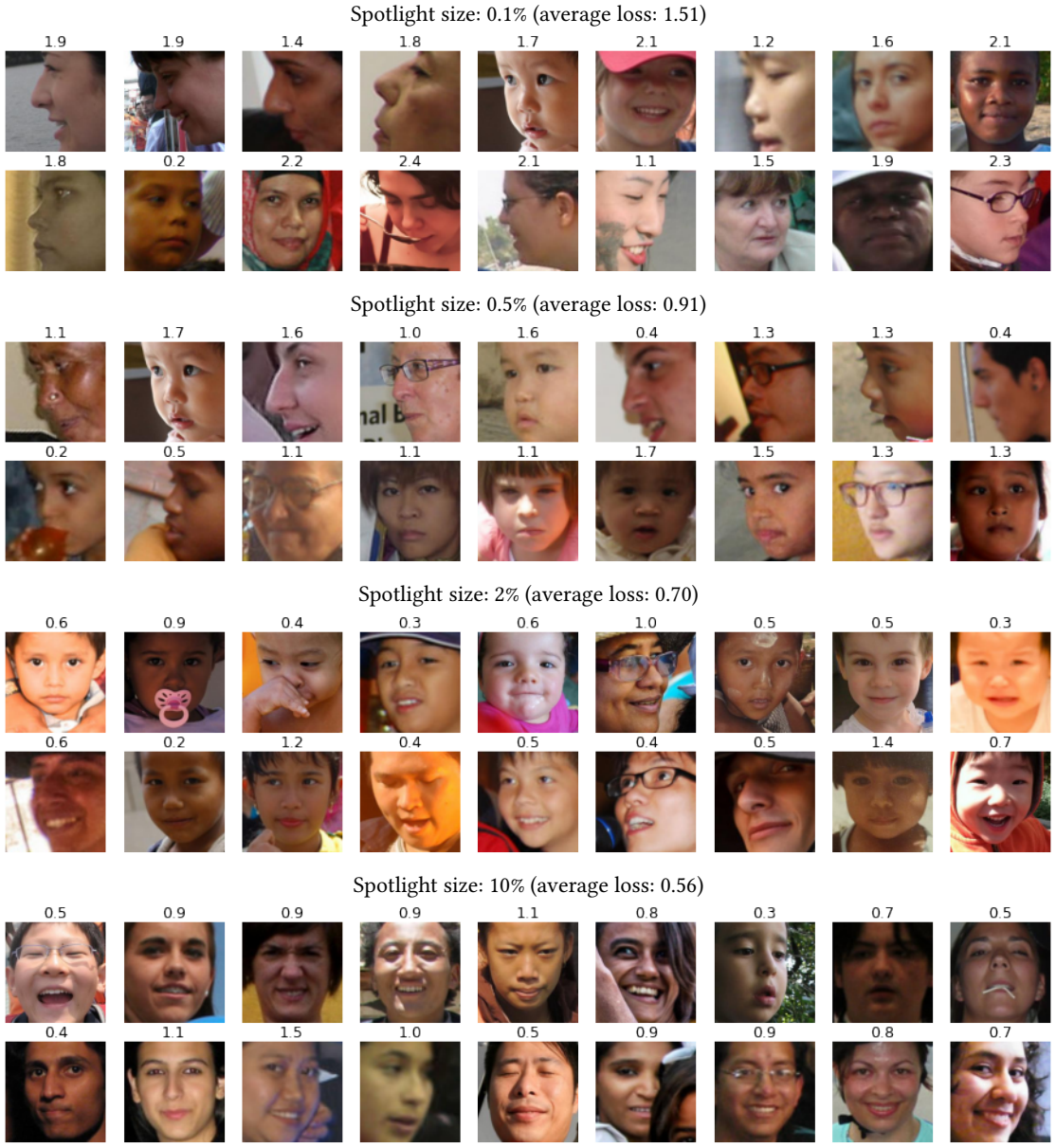


Figure D1: Spotlights on FairFace ranging in size from 0.1% to 10% of the dataset's size.

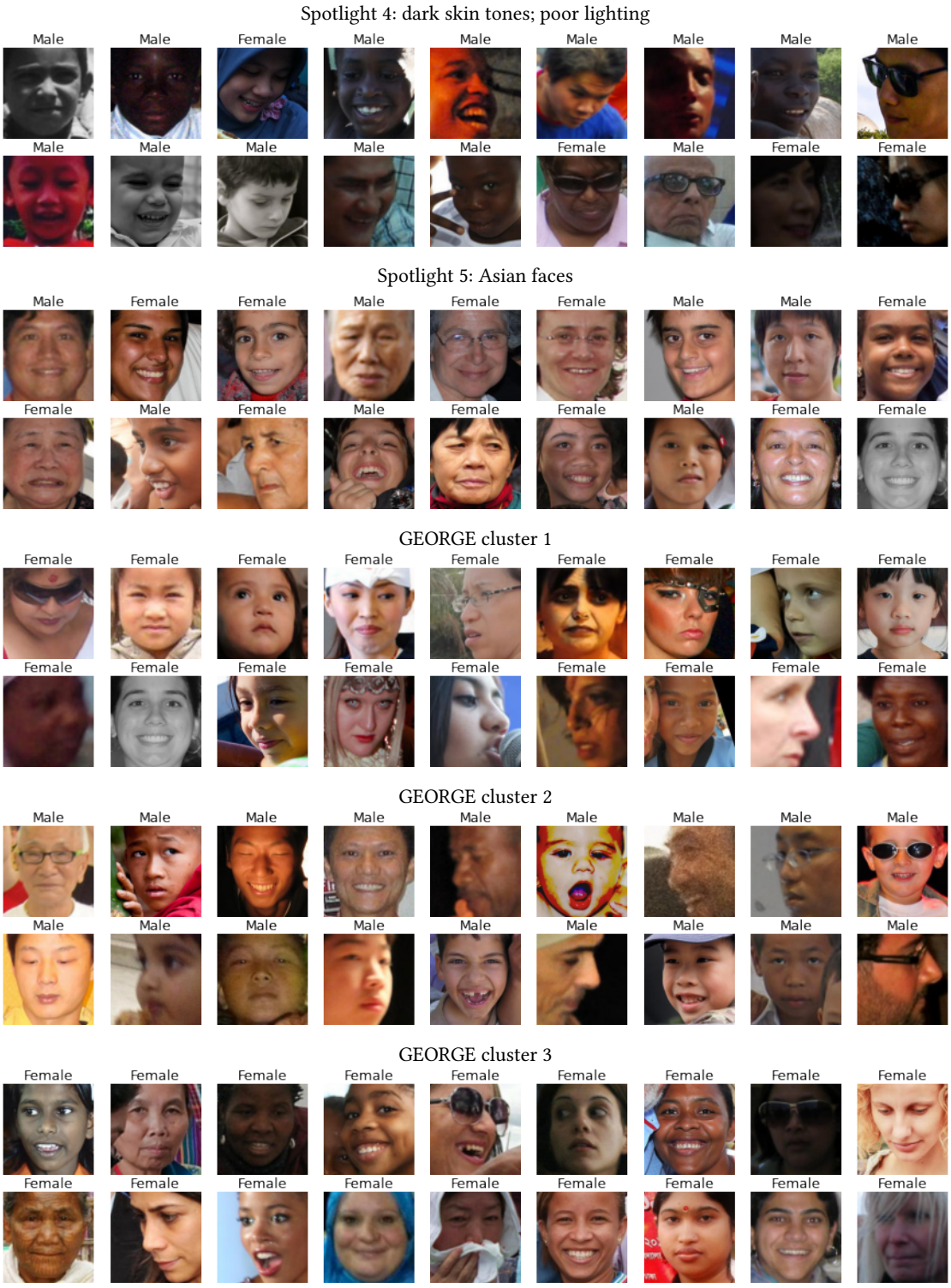


Figure D2: Additional spotlights and GEORGE clusters on FairFace.

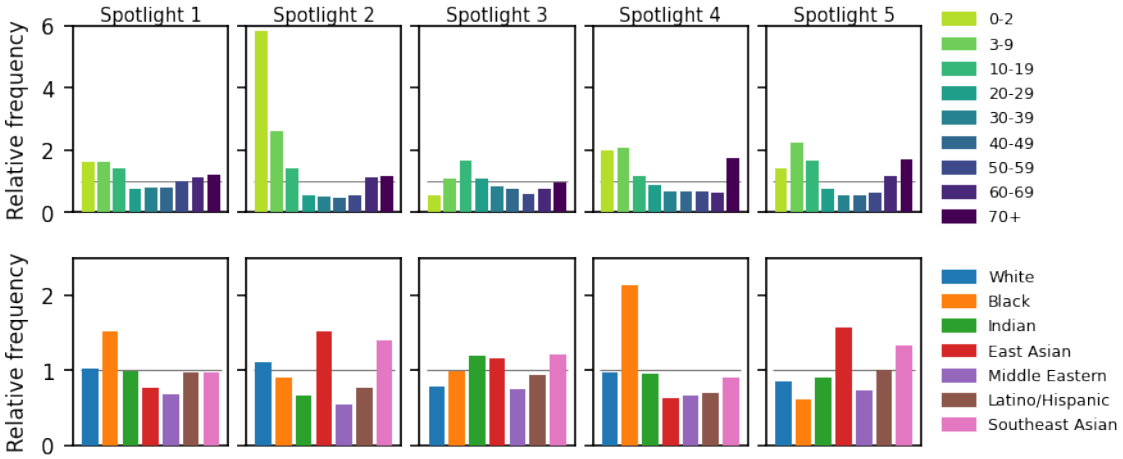


Figure D3: Demographics captured in each spotlight on FairFace.

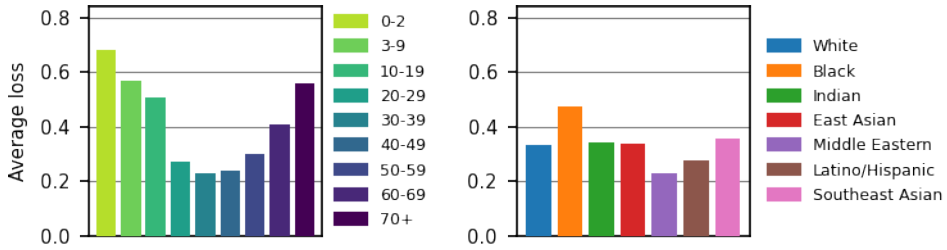


Figure D4: Average losses broken down by age group (left) and race (right) on FairFace.

Figure D5: Additional spotlights on ImageNet.

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User ID (# Reviews)
4	1	11.2	Pulp Fiction	Crime	4.2 (82)	305 (97)
4	5	11.0	Princess Bride, The	Action	4.1 (58)	419 (3)
5	1	10.9	Face/Off	Action	3.9 (42)	296 (73)
4	1	10.5	Usual Suspects, The	Crime	4.3 (56)	234 (202)
4	1	8.8	Fargo	Crime	4.3 (113)	198 (75)
3	5	8.7	Wizard of Oz, The	Adventure	4.2 (46)	358 (8)
5	1	8.1	Alien	Action	4.2 (68)	295 (96)
3	1	8.0	Mother	Comedy	3.2 (34)	100 (25)
4	1	7.9	Boot, Das	Action	4.0 (35)	102 (104)
5	1	7.9	English Patient, The	Drama	3.7 (93)	239 (73)
5	1	7.8	Shallow Grave	Thriller	3.7 (14)	342 (73)
5	1	7.8	Face/Off	Action	3.9 (42)	145 (131)
4	1	7.7	Devil's Advocate, The	Crime	3.7 (31)	15 (44)
3	5	7.6	Addams Family Values	Comedy	3.1 (18)	326 (74)
5	1	7.5	Raiders of the Lost Ark	Action	4.3 (76)	269 (156)

Figure D6: Rating predictions with highest losses from MovieLens 100k.

Spotlight 1: 3-4 star action films; high model uncertainty

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User ID (# Reviews)
4	2	1.7	Romeo and Juliet	Drama	3.4 (27)	13 (263)
5	1	1.6	Lost Highway	Mystery	2.8 (26)	347 (78)
1	3	1.6	Crow, The	Action	3.4 (30)	217 (39)
4	3	1.4	True Lies	Action	3.2 (40)	13 (263)
3	2	1.8	Crow, The	Action	3.4 (30)	197 (66)
4	3	1.7	Jurassic Park	Action	3.6 (53)	363 (102)
5	4	1.4	Happy Gilmore	Comedy	3.2 (19)	145 (131)
3	3	1.2	Crow, The	Action	3.4 (30)	109 (100)
4	1	1.3	Crash	Drama	2.5 (35)	286 (130)
4	3	1.5	Happy Gilmore	Comedy	3.2 (19)	223 (53)
5	5	1.0	Cook the Thief, The	Drama	3.6 (13)	269 (156)
4	4	1.1	True Lies	Action	3.2 (40)	347 (78)
2	2	1.1	Romeo and Juliet	Drama	3.4 (27)	201 (171)

Spotlight 2: highly rated drama films; users with few reviews

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User ID (# Reviews)
3	4	1.2	Shine	Drama	4.0 (23)	382 (20)
4	2	1.9	Big Night	Drama	4.0 (30)	382 (20)
4	5	1.4	Madness of King George	Drama	4.0 (22)	354 (81)
5	3	1.2	Godfather, The	Action	4.4 (73)	382 (20)
4	3	1.3	Bound	Crime	3.8 (30)	329 (28)
4	4	0.8	Shine	Drama	4.0 (23)	214 (65)
5	2	3.4	Fish Called Wanda, A	Comedy	4.0 (50)	370 (19)
3	3	0.7	People vs. Larry Flynt, The	Drama	3.6 (49)	382 (20)
4	2	3.5	Pulp Fiction	Crime	4.2 (82)	370 (19)
4	3	1.3	Singin' in the Rain	Musical	4.2 (38)	370 (19)
4	2	2.4	Shine	Drama	4.0 (23)	116 (55)
3	3	1.0	Alien	Action	4.2 (68)	382 (20)
3	4	1.1	Braveheart	Action	4.2 (67)	370 (19)

Spotlight 3: unpopular action/comedy movies; users with many reviews

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User ID (# Reviews)
4	4	0.4	Drop Zone	Action	2.4 (9)	130 (175)
4	2	1.8	Mouse Hunt	Childrens	2.6 (7)	29 (17)
4	4	0.9	Arrival, The	Action	2.7 (14)	363 (102)
3	4	1.1	Father of the Bride Part II	Comedy	2.7 (22)	222 (174)
4	1	1.8	Father of the Bride Part II	Comedy	2.7 (22)	81 (28)
4	3	1.1	Drop Zone	Action	2.4 (9)	393 (133)
3	3	0.8	Space Jam	Adventure	2.6 (13)	303 (208)
4	3	1.3	Father of the Bride Part II	Comedy	2.7 (22)	223 (53)
1	3	1.2	Disclosure	Drama	2.7 (10)	303 (208)
4	3	1.2	Arrival, The	Action	2.7 (14)	303 (208)
3	3	0.9	Space Jam	Adventure	2.6 (13)	21 (84)
4	4	0.6	Casper	Adventure	2.6 (12)	83 (77)
4	2	1.6	Last Man Standing	Action	2.8 (14)	303 (208)

Figure D7: Spotlights on MovieLens 100k.

GEORGE cluster 1

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
4	1	11.2	Pulp Fiction	Crime	4.2 (82)	97
5	1	10.9	Face/Off	Action	3.9 (42)	73
4	1	10.5	Usual Suspects, The	Crime	4.3 (56)	202
4	1	8.8	Fargo	Crime	4.3 (113)	75
5	1	8.1	Alien	Action	4.2 (68)	96
3	1	8.0	Mother	Comedy	3.2 (34)	25
4	1	7.9	Boot, Das	Action	4.0 (35)	104
5	1	7.9	English Patient, The	Drama	3.7 (93)	73
5	1	7.8	Shallow Grave	Thriller	3.7 (14)	73
5	1	7.8	Face/Off	Action	3.9 (42)	131
4	1	7.7	Devil's Advocate, The	Crime	3.7 (31)	44
5	1	7.5	Raiders of the Lost Ark	Action	4.3 (76)	156
4	1	7.3	Devil's Own, The	Action	2.9 (47)	175

GEORGE cluster 2

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
4	2	7.4	Grosse Pointe Blank	Comedy	3.7 (29)	208
5	2	7.2	Citizen Kane	Drama	4.3 (40)	102
5	2	6.0	Fargo	Crime	4.3 (113)	22
4	2	5.9	Jaws	Action	3.8 (62)	84
5	2	5.8	Schindler's List	Drama	4.4 (61)	184
5	2	5.8	Sense and Sensibility	Drama	4.2 (51)	131
3	2	5.6	Peacemaker, The	Action	3.4 (24)	12
4	2	5.4	Star Wars	Action	4.4 (99)	78
5	2	5.4	Full Monty, The	Comedy	4.0 (63)	263
4	2	5.3	Shawshank Redemption, The	Drama	4.5 (60)	109
4	2	5.2	Groundhog Day	Comedy	3.6 (56)	179
4	2	5.2	Sweet Hereafter, The	Drama	3.3 (9)	12
4	2	5.2	Usual Suspects, The	Crime	4.3 (56)	76

GEORGE cluster 3

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
2	5	4.8	2001: A Space Odyssey	Drama	4.1 (57)	155
4	5	4.2	Cool Hand Luke	Comedy	4.1 (31)	143
2	5	4.2	Birdcage, The	Comedy	3.4 (62)	217
3	5	2.8	Gandhi	Drama	4.0 (38)	155
3	5	2.4	Schindler's List	Drama	4.4 (61)	155
4	5	2.2	Cape Fear	Film-Noir	3.7 (20)	155
4	5	1.7	Blues Brothers, The	Action	3.9 (45)	155
4	5	1.5	Cool Hand Luke	Comedy	4.1 (31)	146
4	5	1.4	Cool Hand Luke	Comedy	4.1 (31)	98
4	5	1.3	Cool Hand Luke	Comedy	4.1 (31)	33
4	5	1.2	Cool Hand Luke	Comedy	4.1 (31)	47
4	5	1.1	Cool Hand Luke	Comedy	4.1 (31)	101
5	5	1.1	Grease	Comedy	3.6 (32)	155

Figure D8: GEORGE clusters on MovieLens 100k.

Figure D9: Chest xray sample images, high loss images, and spotlights.

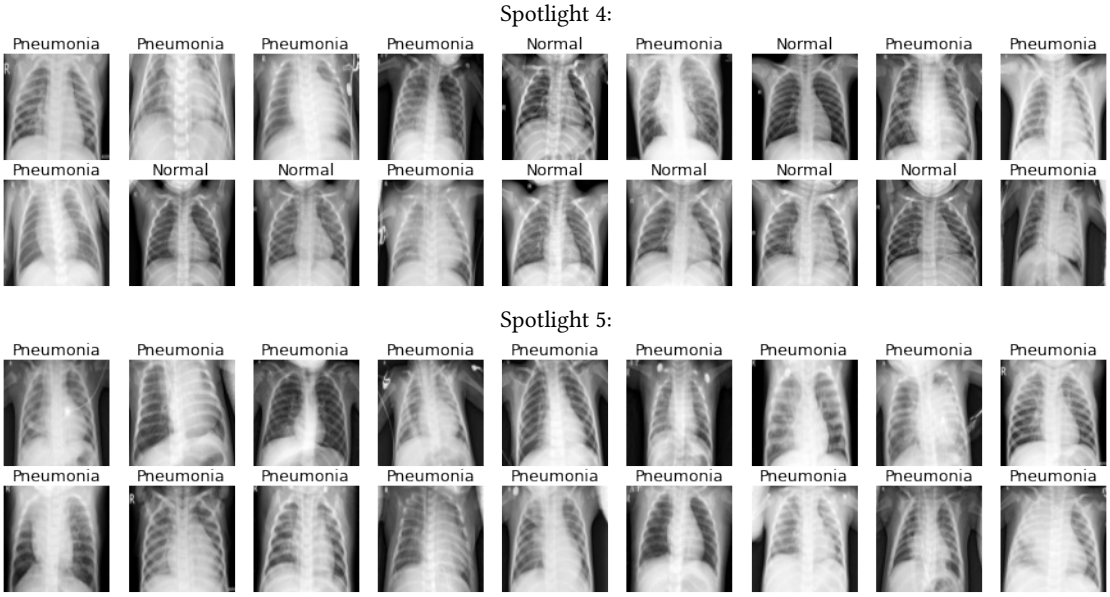


Figure D10: Additional chest xray spotlights.

Subset	Frequent words	Frequent topics
High loss	sacks, tackles, confused, yards, behavior, touchdowns, defendants, protesters, cornerback, interceptions	civil disobedience, 1973 oil crisis, complexity theory
Spotlight 1	packet, packets, switching, circuit, pad, messages, dialogue, aim, networking, why,	packet switching, civil disobedience, computational complexity theory
Spotlight 2	touchdowns, passes, offense, yards, receptions, rating, anderson, receiver, punt, selections	civil disobedience, ctenophora, yuan dynasty
Spotlight 3	networking, alice, capacity, consequence, combining, teach, views, protest, acceleration, switching	packet switching, teacher, force

Figure D11: Spotlights on SQuAD.