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

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

Jason d'Eon
Vector Institute/Dalhousie University, Canada
jndeon@dal.ca

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

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

Abstract—**Supervised learning models often make systematic errors on relatively rare subsets of the data. However, such performance problems can be difficult to identify: model performance can be broken down across sensitive groups, but only when these groups are known and explicitly labelled. This paper introduces a method for discovering systematic errors, which we call *the spotlight*. 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 where the model performs poorly. We show that the spotlight surfaces semantically meaningful areas of weakness in a surprisingly wide variety of model architectures, including image classifiers, language models, and recommender systems.**

I. 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 [2], which showed that vision models for gender recognition tend to exhibit far higher error rates when presented with images of black women. AI systems have also been shown to perform poorly for marginalized groups in object recognition [7], speech recognition [15], mortality prediction [4], and recruiting tools [4]. 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]; essay scoring software can give high scores to long, poorly-written essays [18]; and visual question-answering systems can fail when questions are rephrased [21].

Recognizing and mitigating such systematic errors is critical to avoid designing systems that will exhibit discriminatory or unreliable behaviour. In response to these issues, the community has begun to advocate for a deeper understanding of where deep learning models perform poorly. Pretrained models are often released with model cards [16], which include descriptions of the model's performance across relevant sensitive groups, and interactive tools have been developed to explore areas where models tend to make errors [3, 1, 24]. Further, new optimization methods intend to produce models with more balanced performance across predefined groups (see, e.g., [5] for a review).

All such methods require practitioners to recognize and label well-defined groups in their dataset ahead of time, necessarily overlooking semantically related sets of inputs that the practitioner failed to identify in advance. While practitioner should certainly continue to assess model performance on sensitive subpopulations such as marginalized groups, such approaches do not constitute a fully satisfying solution because it is extremely difficult to anticipate all of the sorts of inputs upon which models might systematically fail: e.g., 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 semantic link between these errors was not anticipated by the practitioner. The 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. We show that the spotlight surfaces semantically meaningful areas of weakness in a surprisingly wide variety of otherwise dissimilar models and datasets, including image classifiers, language models, and recommender systems.

Our approach is closely related to a recent literature on distributionally robust optimization (DRO) that aims to train models to perform well across a range of test distributions. These methods define an adversary that has the ability to reweight the dataset, aiming to select a reweighting where the model performs poorly. The earliest work in this literature defined adversaries with the power to reweight each example in the dataset separately [10], or to select a new distribution over group labels [17]. Most relatedly, Sohoni et al. presented GEORGE [22], a method for applying DRO without group labels, which takes an approach similar to our own. 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 training robust models, they do observe that these clusters 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 about inputs that embed together sharing semantic structure but focuses on auditing models rather than robust training; dramatically lowers computational cost to linear time; relies upon model loss rather than an unsupervised clustering algorithm to identify such inputs; and avoids partitioning the entire embedding space, searching only for contiguous, high-loss regions.

Rather than replacing existing methods for measuring biases in trained models, we believe the spotlight is a tool that practitioners will want to add to their model-building pipelines. Integrating the spotlight into a model development process can help developers recognize when their model has failure modes that need to be addressed by augmenting the dataset, adjusting the model architecture, or using more robust optimization methods. We hope that informing these feedback loops can help the community build models with fewer systematic errors, making for more equitable and reliable machine learning systems in the real world.

II. THE SPOTLIGHT: FINDING SYSTEMATIC BIASES IN TRAINED MODELS

Our method aims to explore potential distribution shifts by reweighting examples in the dataset. In particular, we examine reweightings induced by a Gaussian-like parameterized kernel function that emphasizes examples centred around a given point in the model's final layer representation space. We refer to the reweighting associated with a particular choice of the kernel function as a *spotlight*.

We wish to avoid spotlights that are highly focused on only a few points, so we restrict the kernel functions of interest by imposing a minimum constraint on the total weight. The methodology of describing distribution shifts according to a kernel function both enforces reweightings that are continuous over the representation space, and allows us to easily optimize for the kernel's parameters for desirable spotlights.

In order to capture contiguous regions of high loss in the representation space of a model's final layer, we optimize the parameters of the kernel function to find a configuration that maximizes the expected weighted loss over the dataset. Formally, if we suppose the data points have representation vectors $x_1, \ldots, x_N$ in $\mathbb{R}^d$, we compute the weights $k_i = k(x_i, \mu, \tau)$ by

$$k(x_i, \mu, \tau) = \exp\left(-\frac{1}{2}(x_i - \mu)^\top \tau (x_i - \mu)\right)$$

where $\mu \in \mathbb{R}^d$ is the center of the spotlight and and $\tau \in \mathbb{R}^{d \times d}$ is a positive semidefinite precision matrix. Note that the range of k_i is $(0, 1]$, and 1 is achieved if and only if $x_i = \mu$. If the respective losses the model achieves on the data points are $l_1, \ldots, l_N$, then we define our objective as:

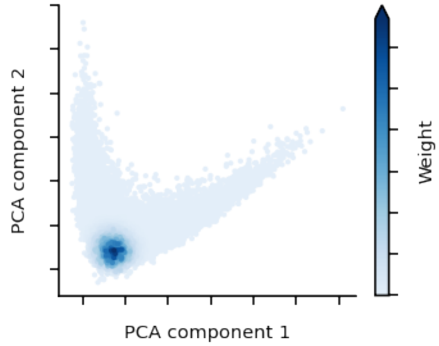


Fig. 1: An example of a spotlight in a model's representation space.

$$\max_{\mu,\tau} \quad \sum_i \left(\frac{k_i}{\sum_i k_i}\right) l_i$$
$$\text{s.t.} \quad \sum_i k_i \geq S,$$

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. Since the sum of the weights, k_i, has an upper bound of N, a typical setting of S is some fixed fraction of the number of data points. In development, we considered $S \in [0.001N, 0.1N]$.

To search for an optimum, we apply the Adam optimizer with an adaptive learning rate, halving the learning rate each time the objective reaches a plateau. Since the feasible region is non-convex and the optima tend to lie close to the constraint, we struggled to find good optima in preliminary experiments. We therefore employ a barrier method, which adds a penalty term that increases as we move closer to the constraint. Over time, we apply this penalty in smaller regions around the constraint, so that the optimization path is primarily through the interior of the feasible region rather than along the constraint. We found that this significantly improved optimization of the spotlights.

To get the most out of the spotlights, we developed a method of finding multiple distinct spotlights on the same dataset without changing any hyperparameters. This method iteratively subtracts spotlight weight from its internal accounting of each example's loss and then finds another spotlight. More formally, using the weights provided by spotlight, k_i, we update the losses as

$$l_i' := \left(1 - \frac{k_i}{\max_i k_i}\right) l_i;$$

we then perform the same optimization as above and repeat as many times as desired. In all of our experiments, we present multiple spotlights obtained in this way.

We considered two approaches for parameterizing the precision matrix. In the first, we allow the spotlight to use any positive semidefinite precision matrix $\tau \succ 0$, allowing it to form "elliptical" spotlights with a different length scale along each axis. We found that these flexible spotlights sometimes failed to find semantically meaningful groups, particularly in very high-dimensional embedding spaces. In our experiments, we instead focus on "spherical" spotlights, requiring that $\tau = cI$ for some precision parameter $c > 0$. We found that these spherical spotlights were generally sufficient to find semantically meaningful problem areas, and were much faster to optimize with far fewer precision parameters.

In preliminary tests we tried spotlight sizes ranging from 0.1% to 10%, optimized to maximize cross entropy. Overall, spotlights on the smaller end of this spectrum were too selective to observe any cohesion, whereas the largest spotlights were too inclusive. For vision models, we could simply scan the images in the spotlight for cohesion, while for the other models we found it necessary to describe the spotlights using summary statistics. Taking this into account, we settled on a spotlight size of 2% for vision models and a spotlight size of 5% for non-vision models, to allow a larger sample for higher-level summary statistics.

III. EXPERIMENTS

The spotlight is relatively model-agnostic: it only requires knowledge of the final layer representations and the losses for each input. We demonstrate this flexibility by using the spotlight to evaluate a broad range of pretrained models from the literature, spanning image classification (faces; objects; x-rays), NLP (sentiment analysis; question answering), and recommender systems (movies). In each case, we show that the spotlight recovers systematic issues that would have otherwise required group labels to uncover. We note that report full results for every dataset we tested (most in Appendix **??**), showing the method's surprising reliability. We present the first 5 and 3 iterative spotlights for vision and non-vision models respectively. We ran our experiments on a compute cluster having NVIDIA Tesla V100 GPUs. Using this hardware, each spotlight presented in this section took under 1 minute to optimize, emphasizing the computational tractability of our approach even on very large datasets.

A. FairFace

We first study FairFace [13], a collection of 100,000 face images annotated with crowdsourced labels about the perceived age, race, and gender of each image. 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 [2]. 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 for the representation space.

Our results are shown in Figures 2 and A2. We found that each of the spotlights discovered a strikingly different set of faces, each representing a problem area for the model. The first shows a set of profile (i.e., side) views, where it is difficult to see many of the facial features; the second shows several gray or discolored images; the third consists mostly of young children whose genders are relatively harder to discern. The fourth and fifth spotlights consist of black faces in poor lighting and Asian faces, respectively. Overall, our spotlights identified both age and racial groups that the model performs poorly on—without access to these demographic labels—and 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.

B. ImageNet

For a second vision dataset, we study the pretrained ResNet-18 model from the PyTorch model zoo [19], 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.

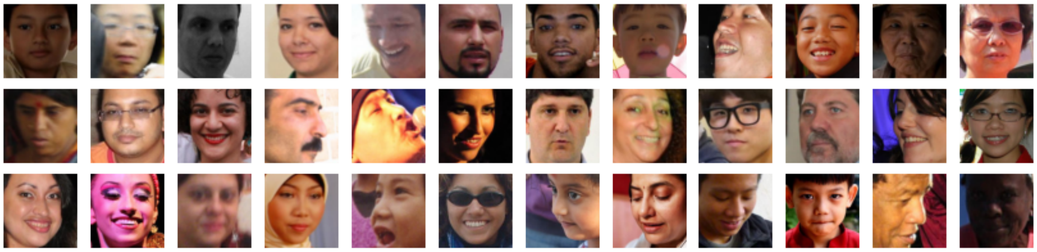
Our results are shown in Figures 3 and A3. Each spotlight found a set of images that have a clear "super-class", but are difficult to classify beyond this super-class. The first spotlight contains a variety of images of people working, where it is difficult to tell whether the label should be about the person in the image, the task they're performing, or another object. The second and fourth spotlights consist of cluttered countertops, where it is tough to decide which object in the image should be labelled. The remaining spotlights show a variety of animals in natural settings, where recognizing a high-level class such as "dog" might be simple, but predicting the correct breed of dog is not. 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.

C. Sentiment analysis: Amazon reviews

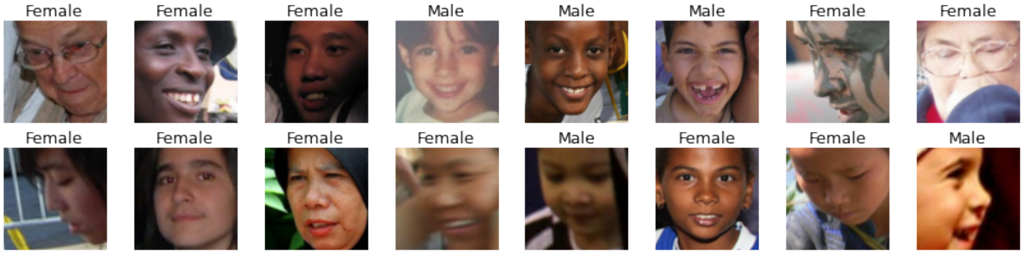
Next, we turn to the Amazon polarity dataset [25], a collection of 4 million plain-text Amazon reviews labelled as "positive" (4-5 stars) or "negative" (1-2 stars). We used a popular pretrained checkpoint of a DistilBERT model from Huggingface [12], which was fine-tuned on SST-2, a set of English sentences labelled with binary sentiment labels. 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 classifies as negative: it was only trained on English sentences, and 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

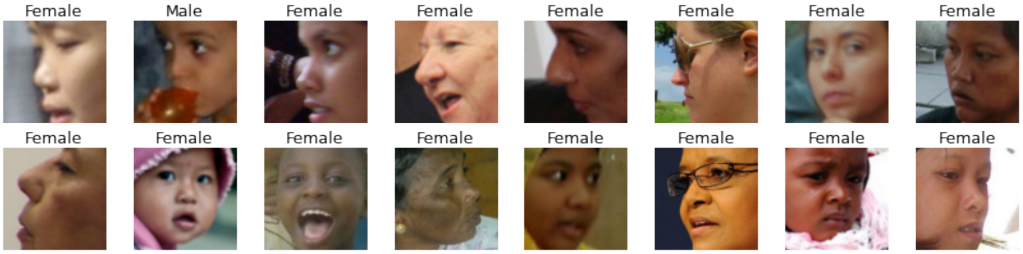
Random sample:



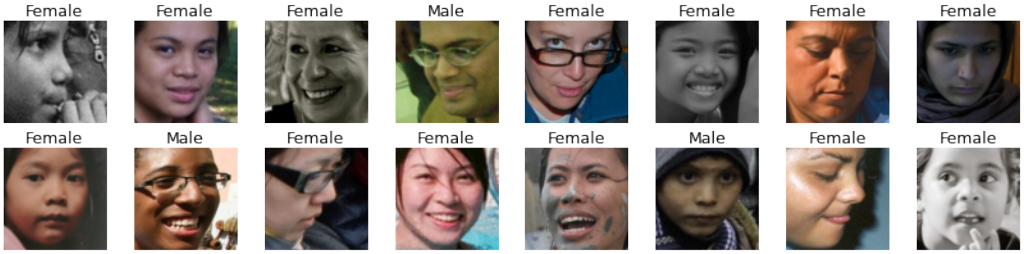
Highest losses: a diffuse set



Spotlight 1: side profile views/poor framing



Spotlight 2: greyscale/discolored images



Spotlight 3: young children

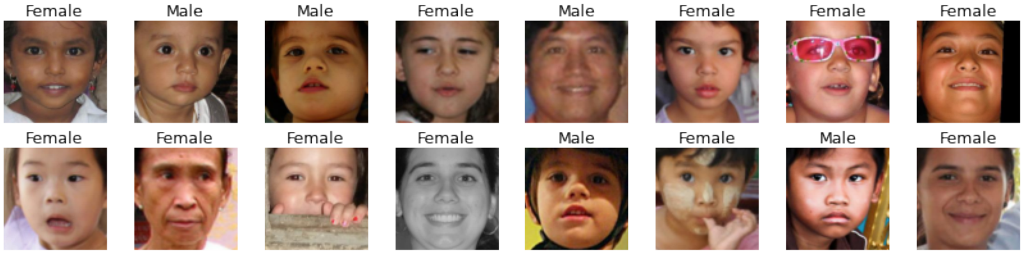
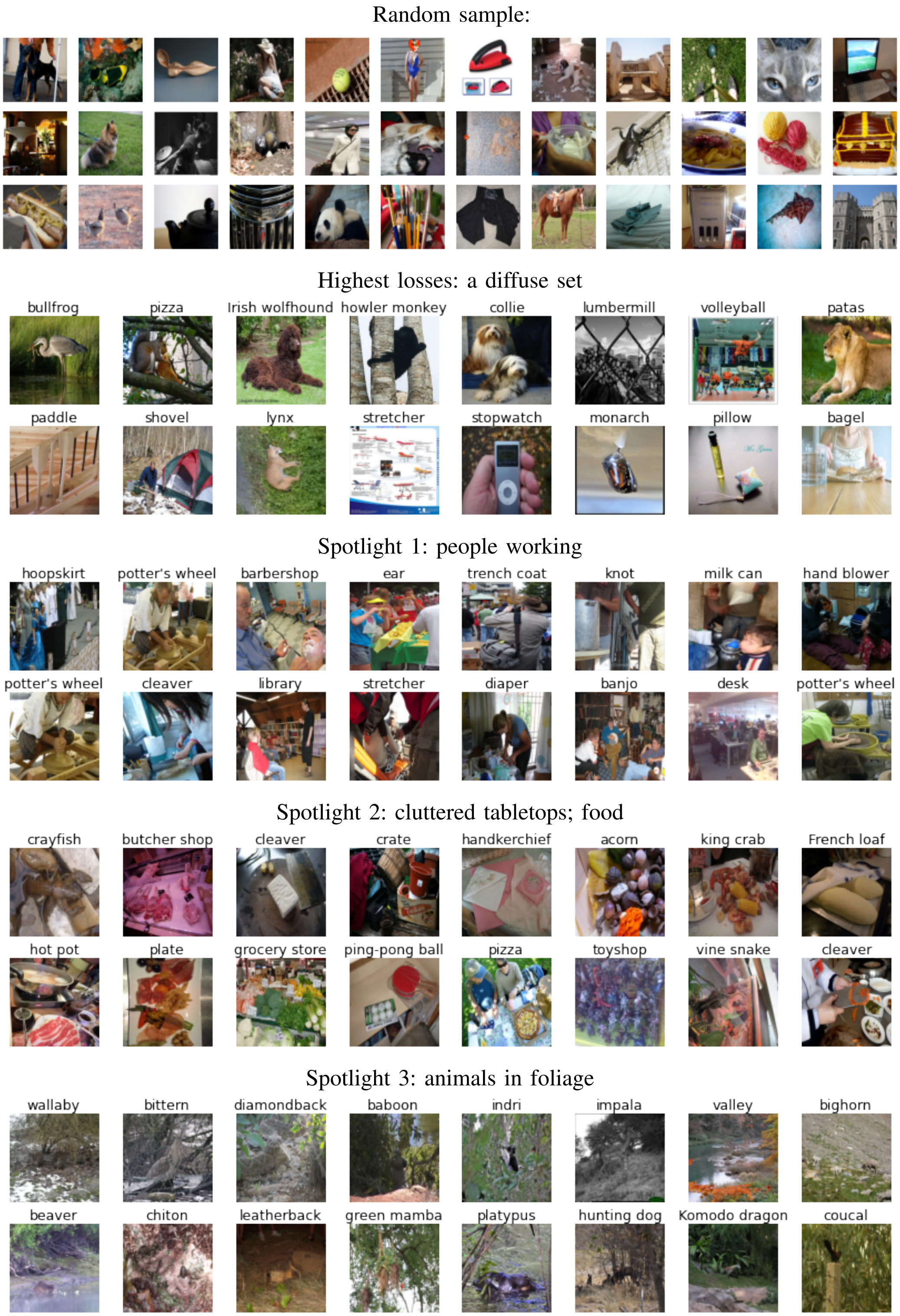


Fig. 2: Spotlights on FairFace validation set. Image captions list true label.



Random sample:

Highest losses: a diffuse set

Spotlight 1: people working

Spotlight 2: cluttered tabletops; food

Spotlight 3: animals in foliage

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

that mention aspects of customer service, such as product returns, which the model classifies as extremely negative, incurring very high losses when these reviews are positive.

The highest-loss reviews in the dataset are quite different, consisting almost entirely of mislabeled 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 distributionally robust optimization methods, which could insist that the model learn to memorize these outliers rather than focusing on other important portions of the dataset.

D. SQuAD

To further explore language models, we analyzed spotlights using a pretrained DistilBERT model [11] fine-tuned on the Standford question answering dataset (SQuAD) [20]. The 100,000+ examples in the dataset were constructed from 536 Wikipedia articles. The dataset also includes article titles, which we use for our analysis, but were not visible to the model. For all experiments, we used the test set, excluding long examples where the sum of the context and answer sequence lengths was greater than 384, leaving 10,386 question-answer pairs.

Unlike all other classification models in our experiments, this question-answering model did not have a single, clear representation space. Instead, it produces a pair of output probabilities for every token in the input sequence, with each token using its own 768-dimensional representation space. To create a single representation for each question, we concatenated all of these hidden vectors together, producing one representation space with 300,000 dimensions; then, we applied a random projection onto 1,000 dimensions to make it feasible to optimize the spotlight directly on the flattened representation.

Summaries of the spotlights are shown in Figure 5. We applied the same token analysis as in our sentiment analysis experiments, but additionally show the most common article titles, relative to their frequency across the entire dataset. The first spotlight highlighted words related to internet packets, as well as selecting many questions from the "packet switching" category; breaking down the model's losses across these categories reveals that "packet switching" is indeed the category where the model gets the highest losses. We generally found a high level of similarity between spotlight iterations, which could be a side effect of the random projection. In comparison, high loss examples correlated with the "super bowl 50" category, which has a moderately high average loss.

E. MovieLens 100k

We investigated a third domain, recommender systems. Specifically, we considered the MovieLens 100k dataset [8], a collection of 100,000 movie reviews of 1,000 movies from 1,700 different users. Besides the rating matrix, 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 the deep factorized autoencoder from Graham, Hartford et al. [9], using the final 600-dimensional hidden layer for our representation space.

The highest-weight movies in each spotlight are shown in Figures A5-A7. The first spotlight mostly identifies a mixture of action and comedy films where the model is overly confident that users will give 3-4 star ratings, and is surprised when they even give 2- or 5-star ratings. The second finds that the model often does poorly at predicting ratings for Pulp Fiction, reminiscent of the "Napoleon Dynamite" problem in the Netflix challenge [23]. The third shows that the model is very uncertain about unpopular, poorly-rated comedy and drama movies, getting high losses even when it predicts the correct rating. In comparison, the highest-loss predictions consist mostly of 1-star ratings on movies with high average scores.

F. Chest X-rays

Finally, we ran the spotlight on a sixth dataset, consisting of 6,000 chest x-rays labelled as "pneumonia" or "healthy" [14]. 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 rather than cherry-picking favorable results. Full details appear in the appendix. To summarize, the spotlight identified at least two semantically meaningful failure modes in this domain: images with the text labels L and R on their sides; 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 case. Examining other spotlighted sets of images made it quickly apparent to us that we have no expert knowledge in radiology; it is quite possible that other spotlighted clusters were semantically related in more fundamental ways. It is also possible that the small training set size led to a less meaningful embedding space; indeed, this was one of only two datasets for which we were unable to leverage an existing, pretrained model.

IV. 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.

a) 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 DRO methods surveyed earlier. 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 embedding space. Unfortunately, our preliminary tests suggest that optimizing for this objective is not simple. With large spotlights (10% of

Subset	Frequent words
High loss	length, outdated, potter, bubble, contact, cinematography, adjusting, functions, stock, versus
Spotlight 1: Spanish	que, est, las, como, y, tod, si, la, dry, por
Spotlight 2: novels	bigger, super, hang, prefer, killing, job, discover, slip, easy, wearing
Spotlight 3: customer service	problem, returning, hoping, returned, unfortunately, okay, box, unless, sadly, ok

Fig. 4: Spotlight on Amazon reviews.

Subset	Frequent words	Frequent topics
High loss	sacks, tackles, confused, yards, behavior, touchdowns, defendants, protesters, cornerback, interceptions	civil disobedience, 1973 oil crisis, complexity theory, super bowl 50, imperialism
Spotlight 1	packet, networking, pad, packets, switching, communication, messages, ignition, circuit, bandwidth	packet switching, complexity theory, teacher, civil disobedience, climate change
Spotlight 2	why, know, arrest, collective, packet, membrane, wage, happens, might, protesters	civil disobedience, packet switching, chloroplast, french and indian war, prime number
Spotlight 3	packet, why, packets, punishment, know, arrest, messages, switching, circuit, collective, wages	packet switching, civil disobedience, pharmacy, chloroplast, complexity theory

Fig. 5: Spotlights on SQuAD.

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.

b) 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 [22], 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? Understanding the conditions where the spotlight works well is an important practical problem to investigate.

c) More flexible spotlights.: The structure of the representation space relates to our previous discussion of spherical vs elliptical spotlights. In our evaluation, we restricted our attention to the former, requiring that the adversary choose the same variance in each dimension, because we observed that this was sufficient for discovering semantically meaningful problem areas in a variety of different domains. However, we do not claim that this finding will hold across new settings, nor that an expert could not have made further discoveries about a model's failure modes using an elliptical spotlight.

d) Shallow models.: During development, we conducted some preliminary experiments using the spotlight on shallow neural nets trained on the Adult and Wine Quality datasets from the UCI machine learning repository. These datasets differed from the others we considered because their inputs consist of a small number of semantically meaningful features. On both, we found that the spotlight was unable to pick out any specific problem areas: even when we allowed the spotlight to select a small fraction of the dataset, it still put an appreciable amount of weight on more than 20% of the inputs. One explanation is that the shallow models tended to make quite diffuse errors, so that each high-loss point was surrounded by blanket of low-loss examples, making it difficult to pick out a problem area without also finding many correct examples. These findings explain this paper's focus on deep learning models; however, we remain interested in examining whether the spotlight can aid in group discovery in "shallow learning" domains where semantically meaningful features are provided as part of the input.

V. 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, difficult question categories, and specific movies with unpredictable reviews. The spotlight found all of these sets without access to side information such as demographics, topics, or genres.

The spotlight is not a direct solution to the problem of systematic errors in deep learning models, instead fitting into 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. 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.

a) Potential Social Impacts.: 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, 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.

REFERENCES

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

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

[3] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56. IEEE, 2019.

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

[5] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[6] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, and M. D. e. a. Hoffman. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[7] T. de Vries, I. Misra, C. Wang, and L. van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[8] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[9] J. Hartford, D. Graham, K. Leyton-Brown, and S. Ravanbakhsh. Deep models of interactions across sets. In *International Conference on Machine Learning*, pages 1909–1918. PMLR, 2018.

[10] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[11] HuggingFace. Distilbert base uncased distilled squad. https://huggingface.co/distilbert-base-uncased-distilled-squad, 2021. Accessed: 2021-05-28.

[12] HuggingFace. Distilbert base uncased finetuned sst-2. https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english, 2021. Accessed: 2021-05-28.

[13] K. Karkkainen and J. Joo. 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*, pages 1548–1558, 2021.

[14] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

[15] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[16] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[17] Y. Oren, S. Sagawa, T. B. Hashimoto, and P. Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

[18] L. Perelman. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111, 2014.

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

[20] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad:

100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[21] M. Shah, X. Chen, M. Rohrbach, and D. Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[22] N. Sohoni, J. Dunnmon, G. Angus, A. Gu, and C. Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33, 2020.

[23] C. Thompson. If you liked this, you're sure to love that. https://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html, 2008. Accessed: 2021-05-28.

[24] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[25] X. Zhang, J. Zhao, and Y. Lecun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.

APPENDIX

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

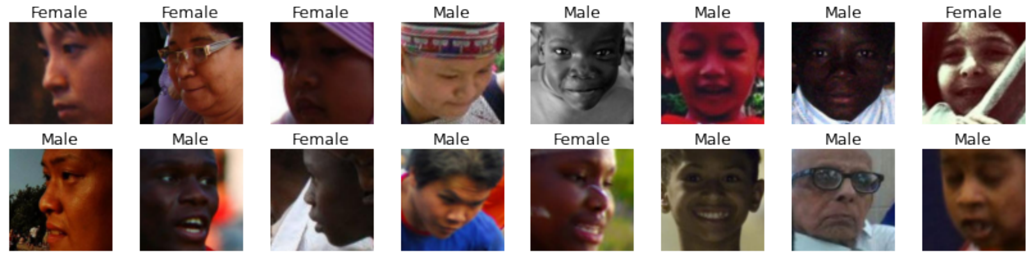
We also include additional outputs from the spotlights on the image and MovieLens datasets that were described in the text. In particular, we include spotlights for:

- FairFace: fourth and fifth spotlights in Figure A2
- ImageNet: fourth and fifth spotlights in Figure A3
- MovieLens: high loss ratings in Figure A4; spotlight examples in Figures A5-A7
- Chest x-rays: random sample, high loss examples, and first three spotlights in Figure A8; final two spotlights in Figure A9

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

Fig. A1: :)

Spotlight 4: dark skin tones; poor lighting



Spotlight 5: Asian faces

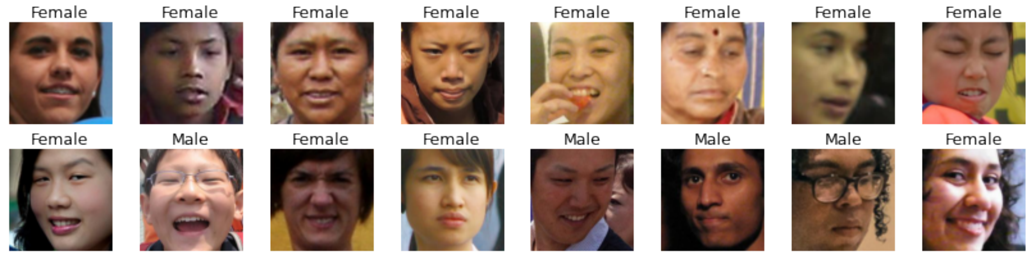
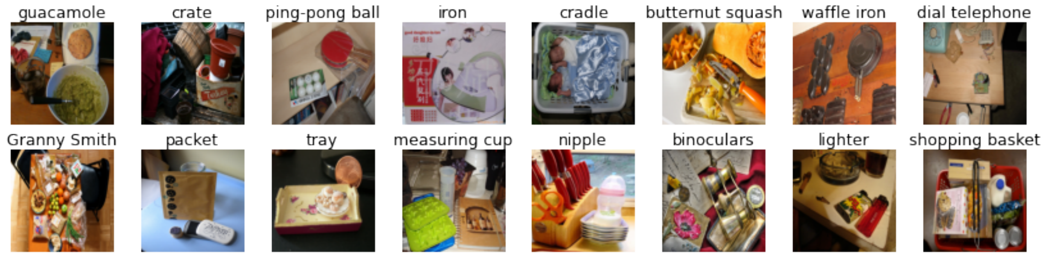


Fig. A2: Additional spotlights on FairFace.

Spotlight 4: cooking tools; food



Spotlight 5: outdoor dogs

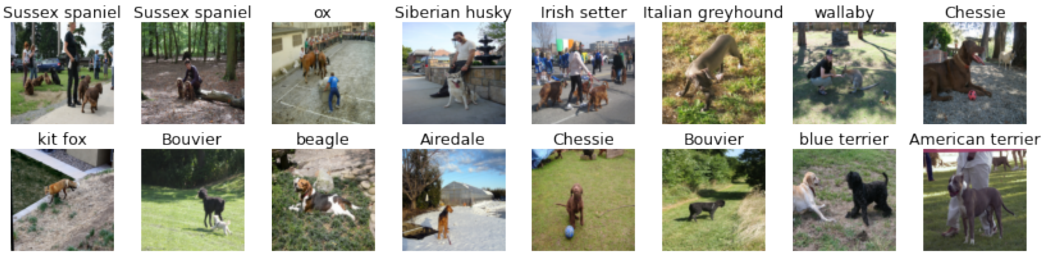


Fig. A3: Additional spotlights on ImageNet.

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
4	1	11.2	Pulp Fiction	Crime	4.2 (82)	97
4	5	11.0	Princess Bride, The	Action	4.1 (58)	3
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
3	5	8.7	Wizard of Oz, The	Adventure	4.2 (46)	8
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
3	5	7.6	Addams Family Values	Comedy	3.1 (18)	74
5	1	7.5	Raiders of the Lost Ark	Action	4.3 (76)	156

Fig. A4: Rating predictions with highest losses from MovieLens 100k.

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
3	2	1.8	Crow, The	Action	3.4 (30)	66
3	3	1.2	Crow, The	Action	3.4 (30)	100
4	4	1.1	True Lies	Action	3.2 (40)	78
4	5	1.5	Jurassic Park	Action	3.6 (53)	39
4	2	1.7	Romeo and Juliet	Drama	3.4 (27)	263
4	5	1.3	Jurassic Park	Action	3.6 (53)	78
1	3	1.6	Crow, The	Action	3.4 (30)	39
4	3	1.6	Romeo and Juliet	Drama	3.4 (27)	73
4	5	1.4	Scream	Horror	3.4 (87)	78
4	2	1.7	Pretty Woman	Comedy	3.5 (32)	131
4	3	1.4	True Lies	Action	3.2 (40)	263
4	2	1.8	True Lies	Action	3.2 (40)	66
3	3	1.1	Scream	Horror	3.4 (87)	66
4	4	1.2	MST3K	Comedy	3.3 (30)	263
4	2	2.1	MST3K	Comedy	3.3 (30)	20

Fig. A5: Spotlight 1 from MovieLens 100k.

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
2	5	1.3	Pulp Fiction	Crime	4.2 (82)	79
5	2	1.5	Pulp Fiction	Crime	4.2 (82)	15
5	5	0.8	Pulp Fiction	Crime	4.2 (82)	26
4	5	1.4	Pulp Fiction	Crime	4.2 (82)	171
5	5	0.9	Pulp Fiction	Crime	4.2 (82)	94
4	4	0.9	Pulp Fiction	Crime	4.2 (82)	124
5	4	1.2	Pulp Fiction	Crime	4.2 (82)	54
5	5	0.6	Pulp Fiction	Crime	4.2 (82)	108
5	3	1.6	Citizen Kane	Drama	4.3 (40)	74
5	5	0.7	Pulp Fiction	Crime	4.2 (82)	46
4	5	1.2	Pulp Fiction	Crime	4.2 (82)	143
4	1	11.2	Pulp Fiction	Crime	4.2 (82)	97
5	5	0.6	Pulp Fiction	Crime	4.2 (82)	99
5	4	0.9	Pulp Fiction	Crime	4.2 (82)	173
4	2	2.4	Shine	Drama	4.0 (23)	55

Fig. A6: Spotlight 2 from MovieLens 100k.

Prediction	Rating	Loss	Movie	Genre	Avg (# Reviews)	User reviews
4	4	1.0	Great White Hype, The	Comedy	2.8 (12)	80
1	1	0.7	Black Sheep	Comedy	2.2 (12)	53
4	4	0.8	Great White Hype, The	Comedy	2.8 (12)	179
1	1	1.0	Speed 2: Cruise Control	Action	2.0 (8)	174
4	1	1.6	Georgia	Drama	2.4 (10)	74
4	3	1.2	Georgia	Drama	2.4 (10)	109
3	2	1.6	Casper	Adventure	2.6 (12)	174
4	4	0.8	Transformers: The Movie	Action	2.2 (8)	95
4	4	1.1	Beyond Rangoon	Drama	2.6 (5)	167
1	1	0.8	Black Sheep	Comedy	2.2 (12)	84
3	2	1.4	Georgia	Drama	2.4 (10)	126
4	2	1.8	Beyond Rangoon	Drama	2.6 (5)	121
4	3	1.0	Great White Hype, The	Comedy	2.8 (12)	63
2	3	1.3	Great White Hype, The	Comedy	2.8 (12)	28
3	3	0.9	Transformers: The Movie	Action	2.2 (8)	208

Fig. A7: Spotlight 3 from MovieLens 100k.

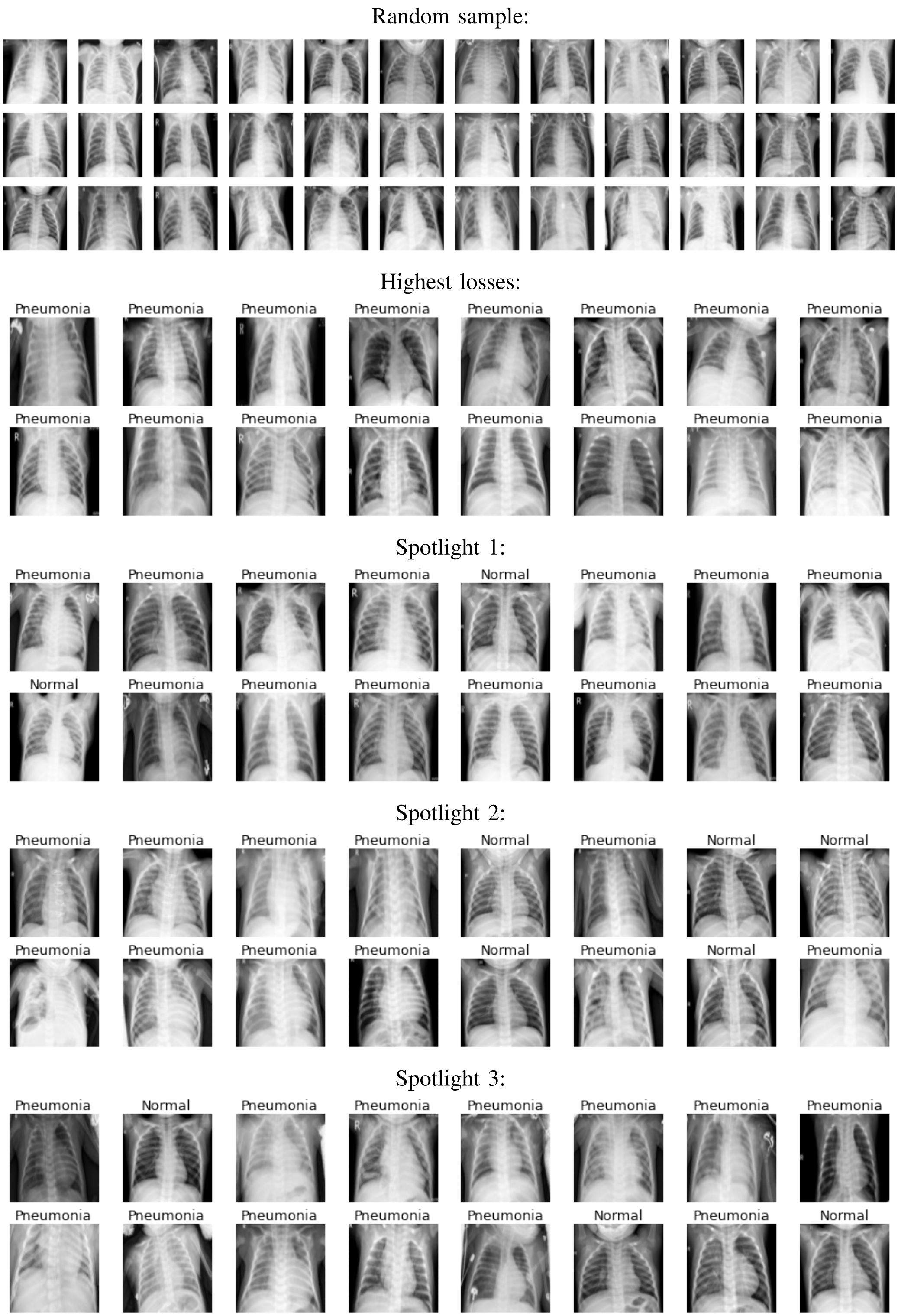
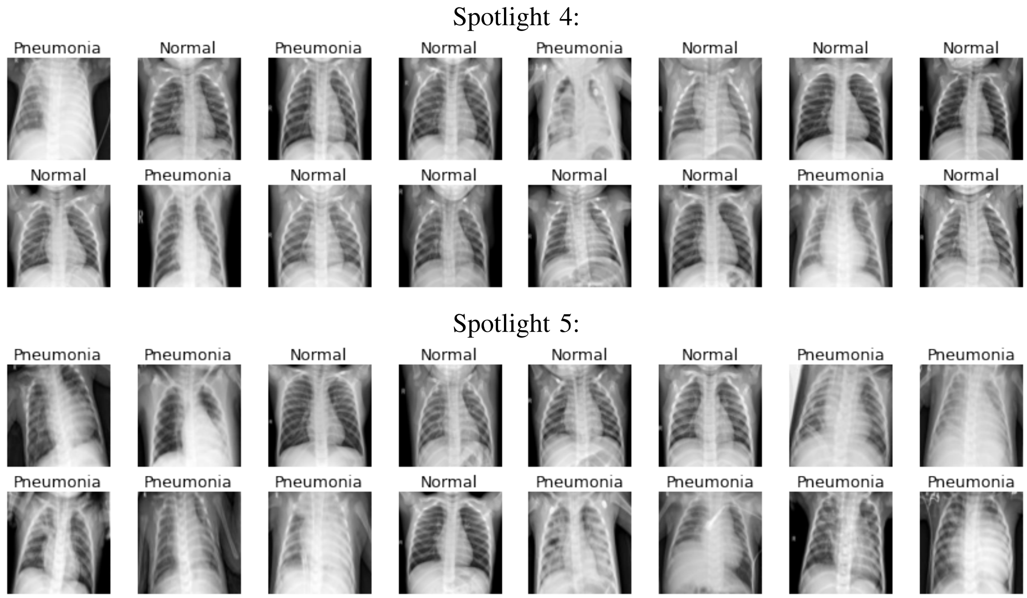

Random sample:
Highest losses:
Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia
Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia
Spotlight 1:
Pneumonia Pneumonia Pneumonia Pneumonia Normal Pneumonia Pneumonia Pneumonia
Normal Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia
Spotlight 2:
Pneumonia Pneumonia Pneumonia Pneumonia Normal Pneumonia Normal Normal
Pneumonia Pneumonia Pneumonia Pneumonia Normal Pneumonia Normal Pneumonia
Spotlight 3:
Pneumonia Normal Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia
Pneumonia Pneumonia Pneumonia Pneumonia Pneumonia Normal Pneumonia Normal

Fig. A8: Chest xray sample images, high loss images, and spotlights.



Spotlight 4:

Spotlight 5:

Fig. A9: Additional chest xray spotlights.