Observational Supervision for Medical Image Classification using Gaze Data

Khaled Saab[1] (✉), Sarah M. Hooper[1], Nimit S. Sohoni[2], Jupinder Parmar[3], Brian Pogatchnik[4], Sen Wu[3], Jared A. Dunnmon[3], Hongyang R. Zhang[5], Daniel Rubin[6], and Christopher Ré[3]

[1] Department of Electrical Engineering, Stanford University, Stanford, USA
ksaab@stanford.edu
[2] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, USA
[3] Department of Computer Science, Stanford University, Stanford, USA
[4] Department of Radiology, Stanford University, Stanford, USA
[5] Khoury College of Computer Sciences, Northeastern University, Boston, USA
[6] Department of Biomedical Data Science, Stanford University, Stanford, USA

Abstract. Deep learning models have demonstrated favorable performance on many medical image classification tasks. However, they rely on expensive hand-labeled datasets that are time-consuming to create. In this work, we explore a new supervision source to training deep learning models by using gaze data that is passively and cheaply collected during a clinician's workflow. We focus on three medical imaging tasks, including classifying chest X-ray scans for pneumothorax and brain MRI slices for metastasis, two of which we curated gaze data for. The gaze data consists of a sequence of fixation locations on the image from an expert trying to identify an abnormality. Hence, the gaze data contains rich information about the image that can be used as a powerful supervision source. We first identify a set of gaze features and show that they indeed contain class-discriminative information. Then, we propose two methods for incorporating gaze features into deep learning pipelines. When no task labels are available, we combine multiple gaze features to extract weak labels and use them as the sole source of supervision (Gaze-WS). When task labels are available, we propose to use the gaze features as auxiliary task labels in a multi-task learning framework (Gaze-MTL). On three medical image classification tasks, our Gaze-WS method without task labels comes within 5 AUROC points (1.7 precision points) of models trained with task labels. With task labels, our Gaze-MTL method can improve performance by 2.4 AUROC points (4 precision points) over multiple baselines.

Keywords: Medical Image Diagnosis · Eye Tracking · Weak Supervision.

1 Introduction

A growing challenge in medical imaging is the need for more qualified experts to read an increasing volume of medical images, which has led to interpretation delays and reduced quality of healthcare [25]. Deep learning models in radiology [5], dermatology [6], and other areas [7] can increase physician throughput to alleviate this challenge. However, a major bottleneck to developing such models is the need for large labeled datasets [7].

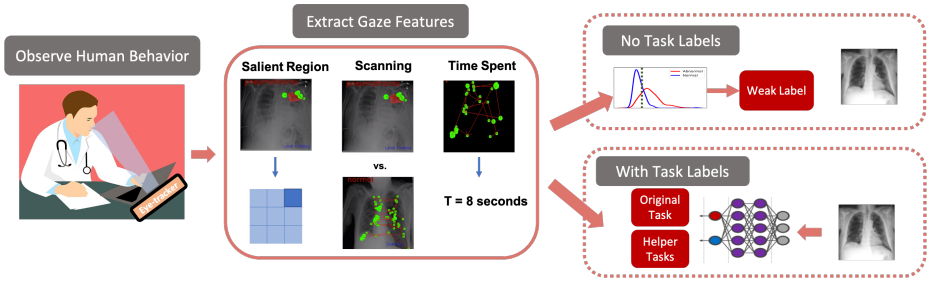

Fig. 1: Illustration of the observational supervision framework. After passively collecting gaze data from experts, we extract informative gaze data statistics which we use in two settings: we turn multiple gaze data statistics into weak labels and combine them to train a ResNet-50 CNN without task labels (top-right). We write helper tasks that predict gaze statistics in a multi-task learning framework along with task labels (bottom-right).

Fortunately, studies in psychology and neurophysiology have shown that human gaze data contains task-related information [38]. Specifically, past studies have shown eye movement is driven by a reward-based mechanism, which induces a hidden structure in gaze data that embeds information about the task [11]. Moreover, with recent advances in eye tracking technology and the rise of commodity augmented reality, gaze data is likely to become both ubiquitous and cheap to collect in the coming years [33,22]. Since collecting gaze data would require no additional human effort in many real-world image analysis workflows, it offers a new opportunity for cheap model supervision, which we term *observational supervision*.

Can gaze data provide supervision signals for medical imaging? While prior work from natural image classification have used gaze data for object recognition applications [15,20,35], using gaze data in the context of medical image diagnosis introduces new questions. For instance, the visual scanning process of medical image diagnosis is often more protracted than in typical object recognition tasks [17], resulting in longer gaze sequences containing richer task-related information. Therefore, it is not clear which features of the gaze data are useful for supervising image diagnosis models, and how much information can be extracted from them. To explore the above questions, we collected gaze data from radiologists as they performed two medical imaging tasks. The first is chest X-ray diagnosis for pneumothorax (i.e., a collapsed lunch), which is a life threatening event. The second is MRI diagnosis for brain metastases, which is a spreading cancer in the brain that requires quick detection for treatment. Our two novel datasets, along with a public dataset on abnormal chest X-ray detection [14], represent challenging real-world medical settings to study observational supervision.

In this work, we propose an observational supervision framework that leverages gaze data to supervise deep learning models for medical image classification. First, we use our three medical datasets to identify several critical statistics of gaze data (summarized in Table 1) such as the number of uniquely visited regions. Second, we use these gaze data statistics to weakly supervise deep learning models for medical image diagnosis (Gaze-WS)—without relying on any task labels. Finally, we propose

to combine gaze data statistics and task labels using a multi-task learning framework to inject additional inductive bias from gaze data along with task labels (Gaze-MTL). See Figure 1 for an illustration. The key intuition behind our approach is that gaze data provides discriminative information through differences in scanning patterns between normal and abnormal images. Interestingly, such signals can be explained using reward-based modeling from the neuroscience literature [29]. Through a quantitative analysis, we theoretically show that the discriminative power of a gaze sequence scales with the size of the abnormal region and how likely the fixation occurs in the abnormal region.

We evaluate our framework on the three medical imaging tasks.[7] Using only gaze data, we find that Gaze-WS comes within 5 AUROC points (or 1.7 precision points) of models trained using task labels. We achieve these results using attention statistics of gaze data such as the number of uniquely visited regions in an image, which is smaller for abnormal images than normal images. Using gaze data and task labels, we find that Gaze-MTL improves performance by up to 2.4 AUROC points (or 4 precision points) compared to previous approaches that integrate gaze data [18,28,20,15]. We observe that measuring the "diffusivity" (see Table 1 for the definition) of an expert's attention as a statistic transfers most positively to the target task in multi-task learning.

2 Related Work

Deep learning models such as convolutional neural networks (CNNs) are capable of extracting meaningful feature representations for a wide variety of medical imaging tasks, ranging from cancer diagnosis [6] to chest radiograph abnormality detection [5]. Since clinicians do not log structured task labels in their standard workflows, these works often involve an expensive data labeling process. We instead explore a setting where we have access to gaze data collected passively from clinicians during their workflows, which we believe will become ubiquitous and cheap with recent augmented reality eyewear. Our work is a first exploration in using gaze data as the main source of supervision for medical image classification models.

Our work is closely related to object detection research that uses gaze data as auxiliary information [15,35,31,30,37]. A common approach from these works is to turn gaze data into attention heatmaps [18,28,20]. For example, Karessli et al. demonstrates that features derived from attention heatmaps can support class-discriminative representations that improve zero-shot image classification [15]. Wang et al. [35] integrate gaze data into the optimization procedure as regularization for training a support vector machine (see also [8]). However, our work investigates gaze data collected from radiologists, which is distinct from gaze collected during natural objection detection tasks. For example, the gaze data that we have collected consists of long sequences of fixations as a result of a protracted search process. By contrast, the gaze data from object detection tasks consists of much shorter sequences [21]. This distinction brings a new challenge of how to extract meaningful features from gaze data for medical imaging. There has also been recent work investigating the integration of gaze data to improve medical diagnosis systems for lung cancer screening [1,16] and abnormality detection in chest X-rays [14]. However,

[7] Our two novel datasets and code are available at https://github.com/HazyResearch/observational.

these methods are hindered by their need for gaze at test time. In our work, we propose two novel approaches for using gaze data to weakly supervise deep learning pipelines using tools from both data programming [24] and multi-task learning [26,36,39].

A different approach to deal with the high cost of labeling in medical imaging is to is to extract labels from clinician reports. NLP models are trained to classify which abnormalities are described in the text and can therefore act as automated annotators when reports are present [23,34]. Such annotation tools, however, are expensive to create and are application-specific. A cheaper and more flexible alternative is data programming, where experts manually write heuristic functions that label data, which has been applied to medical reports [4,27]. We view our approach as a complement to NLP-based approaches. In scenarios where class labels are not found in medical reports [2] or when NLP-based annotators are trained for a different language or task, gaze is a viable alternative (via Gaze-WS). In scenarios where NLP-based annotators are applicable, gaze may be used alongside the labels to further improve performance (via Gaze-MTL). We believe it is a promising direction to explore combining gaze data and NLP-based labeling for medical imaging in future work.

3 Data Collection and Methods

We start by introducing the datasets and gaze data collection process. Then, we present a set of gaze features and describe a theoretical model for studying these features. Finally, we present two approaches to supervising deep learning models using gaze data.

3.1 Gaze Data Collection and Features

Since medical imaging datasets with gaze data are not readily available, we collected gaze data by collaborating with radiologists. We consider three datasets: classifying chest X-rays for pneumothorax (CXR-P), classifying chest X-rays for a general abnormality (CXR-A), and classifying brain MRI slices for metastasis (METS) (all binary image classification). Positive and negative samples from each task can be found in Figure S.2.

For CXR-P, we use the SIIM-ACR Pneumothorax dataset [19], which contains 5,777 X-ray images (22% contain a pneumothorax). We took a stratified random sample of 1,170 images to form our train and validation sets, and collected task labels and gaze data from three board-certified radiologists. We used the remaining 4,607 images as a held-out test set. For CXR-A, Karagryis et al. [14] collected the gaze data of a radiologist reading 1,083 chest X-ray images taken from the MIMIC-CXR Database (66% abnormal) [13]. We reserved 216 random images as a held-out test set, and used the rest for train and validation sets. Importantly, this dataset also gives us the advantage of evaluating our methods on a different eye tracker. For METS, after receiving training from a board-certified radiologist, we analyzed 2,794 MRI slices from 16 cases, comprising our train and validation sets (25% contain a lesion). The held-out test set has 1,664 images.

To collect gaze data, we built custom software to interface with a screen-based Tobii Pro Nano eye tracker. This state-of-the-art eye tracker is robust to head movements, corrective lenses, and lighting conditions. At the start of each session, each radiologist went through a 9-point calibration process. While in use, the program displays a single

Table 1: A summary of gaze data statistics used in our framework, along with their average difference in value between classes across our three tasks (class gap).

Feature	Description	Class Gap
Time on maximum patch	Maximum time dedicated to a patch	0.09
Diffusivity	Maximum time spent on any local region, where each local region is an average of neighboring patches	0.21
Unique visits	Number of uniquely visited patches	0.12
Time spent	Total time spent looking at the image	0.10

image to the user and collects gaze coordinates. Once the user has analyzed the image, they press a key to indicate the label given to the image. The program then saves the set of gaze coordinates that overlapped with the image and displays the next image.

Gaze data statistics. By analyzing random samples of gaze sequences for multiple tasks, we find that the "scanning behavior," e.g., the amount of time the labeler spends scanning over the image versus fixating on a specific patch, correlates strongly with task-related information, such as the existence of an abnormal region in the image. We derive three quantitative features for scanning behavior: time on maximum patch, diffusivity, and unique visits (described in Table 1). We also consider the amount of time spent on a task, since it has been shown to be indicative of task difficulty [3]. We also considered other features such as total distance, velocity, and fixation angles, but found that they provide less signal than those in Table 1. We provide a full list and detailed descriptions of gaze features considered in Table S.4, and visualize the class-conditional distributions of our key gaze data statistics in Figure S.1.

Modeling scanning behavior using a reward-based search model. We consider a labeler actively searching for salient image features that contribute to their classification. Drawing inspiration from neuroscience literature [29], we assume that the scanning behavior is directed by a latent reward map with high rewards in task-informative regions (e.g. an abnormal region in a radiograph) and low rewards in less informative regions (e.g. background). Suppose that a reward of Q_i is obtained by fixating at region i and that there are p regions in total. We consider the gaze sequence of a labeler as a random sequence, where the probability of visiting the i-th region is given by $\Pr(i) = \frac{\exp(Q_i)}{\sum_{i=1}^{p} \exp(Q_i)}$.

We show that the discriminative power of these gaze statistics scales with an interesting quantity that we term the "attention gap." Informally, the attention gap captures the differences of experts' scanning behaviors between normal and abnormal images. Let $Q_{no} > 0$ denote the reward of visiting a normal region and $Q_{ab} > Q_{no}$ denote the reward of an abnormal region. Let p denote the total number of regions and $s \in [0, 1]$ denote sparsity—the fraction of abnormal regions. Therefore, for a random visit, the probability that the visit lands in a particular abnormal region is equal to

$$\frac{1}{p} \cdot \frac{\exp(Q_{ab} - Q_{no})}{s \cdot \exp(Q_{ab} - Q_{no}) + (1 - s)}. \tag{1}$$

The above quantity is larger than $\frac{1}{p}$ as long as $Q_{ab} > Q_{no}$, resulting in an "attention gap" between abnormal and normal regions. Equation 1 reveals that the discriminative signal in gaze features increases as the attention gap increases and the sparsity decreases.

3.2 First Observational Supervision Method: Gaze-WS

We propose a method that combines the gaze data statistics from Table 1 to compute posterior probabilities of task labels, enabling model supervision using gaze data alone. Given training pairs consisting of an image and a gaze sequence, denoted by $\{(x_i, g_i)\}_{i=1}^n$, our goal is to predict the labels of test images—without being provided gaze sequences at test time.

(i) We first compute m gaze features $h_{i1} \in \mathbb{R}^{a_1}, \ldots, h_{im} \in \mathbb{R}^{a_m}$ from each gaze sequence g_i. Specifically, we use four gaze data statistics from Table 1: time on max patch, diffusivity, unique visits, and time spent. We use these features to compute labels $\{\hat{y}_i\}_{i=1}^n$ that approximate the true (unobserved) class labels $\{y_i\}_{i=1}^n$.

(ii) Using a small validation set, we fit two Gaussians to each feature—one each for the positive and negative classes—which are used to estimate the likelihoods $p(h_{im}|y)$ for each unlabeled training sample. We assume the features are conditionally independent and compute the posterior probability $\hat{y}_i = P(y_i = 1 \mid h_{i1}, \ldots, h_{im})$ using Bayes' theorem. We convert them to binary labels with a threshold selected via cross validation.

3.3 Second Observational Supervision Method: Gaze-MTL

We next consider a second setting, where we have both task labels and gaze data, and propose a multitask learning (MTL) framework known as hard parameter sharing [26]. The idea is that gaze data contains fine-grained information such as task difficulty and salient regions [10], which complements class labels. Specifically, we are given training tuples consisting of an image, a gaze sequence, and a label, denoted by $\{(x_i, g_i, y_i)\}_{i=1}^n$. Our goal is to train an image classification model that predicts the labels of test images. (Again, we do not assume access to gaze at test time.) Denote the domain of $\{x_i\}$ by $\mathcal{X}$.

(i) For each sequence g_i, we compute m gaze features $h_{i1} \in \mathbb{R}^{a_1}, \ldots, h_{im} \in \mathbb{R}^{a_m}$.

(ii) We train a feature representation model (e.g., a CNN) $f_\theta(\cdot) : \mathcal{X} \to \mathbb{R}^d$ with feature dimension d parameterized by θ, along with the target task head $A_0 \in \mathbb{R}^{k \times d}$ and *helper task heads* $A_1 \in \mathbb{R}^{a_1 \times d}, \ldots, A_m \in \mathbb{R}^{a_m \times d}$. We minimize the following loss over the training data:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\ell_0(A_0 f_\theta(x_i), y_i) + \sum_{j=1}^{m} \alpha_j \ell_j(A_j f_\theta(x_i), h_{ij}) \right) . \tag{2}$$

Here ℓ_0 denotes the prediction loss for the main task. $\ell_1, \ldots, \ell_m$ denote prediction losses for the m helper tasks, and $\alpha_1, \ldots, \alpha_m$ are hyperparameters that weight each helper task. In our experiments, we used the soft cross-entropy loss [24] to predict the normalized gaze features as the helper tasks.

In the above minimization problem, a shared feature representation model $f_\theta(x_i)$ is used for the target task as well as all helper tasks. This is also known as hard parameter sharing in MTL, and works as an inductive transfer mechanism from the helper tasks to the target task. Importantly, at inference time, given an image, our model predicts the label using the target task head alone. Hence, we do not require gaze data for inference.

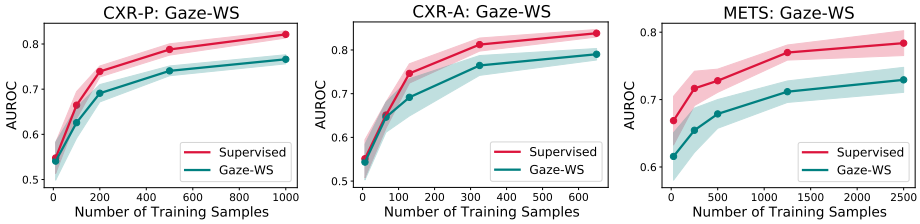

Fig. 2: Mean AUROC of Gaze-WS scales competitively to supervised learning on the same CNN. The results are averaged over 10 random seeds and the shaded region indicates 95% confidence intervals. We find that using gaze data as the sole supervision source achieves within 5 AUROC points to supervised learning on the same CNN model.

4 Experimental Results

We validate that gaze data alone can provide useful supervision via Gaze-WS, and improves model performance using Gaze-MTL. For all models, we train a ResNet-50 CNN architecture [12] pretrained on ImageNet for 15 epochs using PyTorch v1.4.0. Common hyperparameters were chosen by random search to maximize validation accuracy. Models were trained on two Titan RTX GPU's, where a single epoch took approximately 8 seconds. More details are in our code. We choose the operating points that achieve the same recall scores as previously published models: 55 for CXR-P [32], 90 for CXR-A [5], and 53 for METS [9].

4.1 Using Gaze Data as the Sole Supervision Source

We empirically validate our hypothesis that gaze features alone can be used to supervise well-performing medical image classification models. We compare the test performance of Gaze-WS to that of a supervised CNN trained with task labels. We also measure performance as the number of training samples varies (Figure 2).

This scaling analysis shows that Gaze-WS models improve with more weakly labeled data and approach the supervised CNN performance, ultimately coming within 5 AUROC points for CXR-P and CXR-A. Moreover, we find that Gaze-WS for CXR-P has higher recall on small abnormalities (which are more difficult to detect in practice) compared to the supervised CNN, but misses more of the large abnormalities. Intuitively, a labeler may spend more time examining the smaller abnormalities, making them more easily identifiable from the gaze features.

We next inspect the weak labels estimated by the gaze data for METS and CXR-P. We find that the weak labels achieve a mean AUROC of 80 and 93 on CXR-P and METS, respectively. This performance closely matches the intuition we develop in Section 3. Recall that we expect the separation strengths to scale with dataset sparsity and attention gap. For CXR-P and METS, we find that METS' estimated sparsity ($s = 0.03$) is lower and attention gap ($z = 6.5$) is higher than CXR-P's ($s = 0.07$, $z = 1.5$), which indicates we should expect larger separation strengths for METS.

Table 2: Gaze-MTL improves upon supervised learning and multiple baseline methods by up to 2.4 AUROC points on three medical imaging datasets. The results are averaged over 10 random seeds with 95% significance.

Dataset	Metric	Image-only	HM-REG	CAM-REG	Template	ZeroShot	Gaze-MTL
CXR-P	AUROC	81.5 ± 0.6	78.9 ± 1.8	78.3 ± 1.4	78.6 ± 1.4	82.1 ± 0.8	$\mathbf{83.0} \pm 0.5$
	F1-score	56.6 ± 0.7	50.2 ± 3.4	51.0 ± 3.3	53.5 ± 1.2	56.0 ± 0.8	$\mathbf{57.5} \pm 0.7$
	Precision	58.5 ± 1.4	48.7 ± 4.3	49.7 ± 4.2	52.2 ± 2.2	57.0 ± 1.7	$\mathbf{60.2} \pm 1.4$
CXR-A	AUROC	83.8 ± 0.9	83.1 ± 0.6	82.9 ± 1.0	82.3 ± 1.2	83.2 ± 0.8	$\mathbf{84.3} \pm 1.6$
	F1-score	84.1 ± 0.9	83.9 ± 0.9	82.6 ± 1.8	81.7 ± 0.7	84.2 ± 0.9	$\mathbf{84.5} \pm 0.7$
	Precision	78.9 ± 1.5	78.5 ± 1.5	76.0 ± 3.2	74.7 ± 1.3	79.1 ± 1.6	$\mathbf{79.4} \pm 1.2$
METS	AUROC	78.4 ± 1.8	77.6 ± 1.3	65.3 ± 1.8	56.0 ± 1.1	76.8 ± 1.9	$\mathbf{80.8} \pm 1.1$
	F1-score	53.4 ± 1.4	52.4 ± 1.2	49.5 ± 3.6	38.9 ± 1.4	34.7 ± 10.1	$\mathbf{55.0} \pm 1.2$
	Precision	53.5 ± 2.7	51.5 ± 2.7	48.1 ± 4.4	32.0 ± 1.1	52.7 ± 4.1	$\mathbf{57.5} \pm 2.6$

Due to the noise in the weak labels, there is a clear tradeoff between the number of labels, the ease with which those labels are collected, and model performance. For instance, in CXR-P, Gaze-WS achieves the same performance as the supervised CNN model using about $2\times$ as many training samples. These results suggest that Gaze-WS may be useful for passively collecting large quantities of noisy data to achieve the same performance as a model trained with fewer (but more expensive) task labels.

4.2 Integrating Gaze Data along with Task Labels

We empirically validate our hypothesis that gaze data provides additional information beyond the task labels and can be injected into model training to improve model performance via multi-task learning. We train a CNN for each dataset using Gaze-MTL, and compare its performance to a CNN with the same architecture trained with only task labels, or trained by incorporating gaze data through the following existing methods: CAM-REG [28], HM-REG [18], Template [20], and ZeroShot [15].

Table 2 shows that Gaze-MTL results in a performance improvement over each baseline for our three medical tasks. We also find that different gaze features result in larger performance boosts when used as auxiliary tasks for different datasets (Table S.2).

To further investigate which gaze features are most useful, we compute a task similarity score between each gaze feature and the target task by measuring the impact of transfer learning between the tasks. We find that the gaze features that have higher task similarity scores with the target tasks are the same auxiliary tasks with which Gaze-MTL achieved the largest gains (details in Table S.2).

It is common to use the class activation map (CAM), which highlights the areas of an image that are most responsible for the model's prediction, to reveal additional localization information [27]. For CXR-P, we found the CAMs of Gaze-MTL to overlap with the ground-truth abnormality regions 20% more often than the Image-only model. This suggests that models trained with gaze provide more accurate localization information.

The performance boost we see with Gaze-MTL suggests that it is a promising method for integrating gaze data into ML pipelines. Particularly in high-stakes settings where gaze data can be readily collected in conjunction with class labels, Gaze-MTL may be used to integrate additional information from the expert labeler to boost model performance, without requiring additional labeling effort.

5 Conclusions

This work introduced an observational supervision framework for medical image diagnosis tasks. We collected two eye tracking datasets from radiologists and presented two methods for incorporating gaze data into deep learning models. Our Gaze-WS results showed that using gaze data alone can achieve nearly comparable performance to fully supervised learning on CNNs. This result is rather surprising and suggests that gaze data provides promising supervision signals for medical imaging. Furthermore, our Gaze-MTL results showed that gaze data can provide additional inductive biases that are not present in human labels to improve upon the performance of models supervised with task labels alone. We hope that our novel datasets and encouraging results can inspire more interest in integrating gaze data into deep learning for medical imaging.

References

1. Aresta, G., Ferreira, C., Pedrosa, J., Araújo, T., Rebelo, J., Negrão, E., Morgado, M., Alves, F., Cunha, A., Ramos, I., et al.: Automatic lung nodule detection combined with gaze information improves radiologists' screening performance. IEEE journal of biomedical and health informatics **24**(10) (2020) 3
2. Bosmans, J.M., Weyler, J.J., Parizel, P.M.: Structure and content of radiology reports, a quantitative and qualitative study in eight medical centers. European journal of radiology **72**(2) (2009) 4
3. Cole, M.J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N.J., Zhang, X.: Task and user effects on reading patterns in information search. Interacting with Computers **23**(4) (2011) 5
4. Dunnmon, J.A., Ratner, A.J., Saab, K., Khandwala, N., Markert, M., Sagreiya, H., Goldman, R., Lee-Messer, C., Lungren, M.P., Rubin, D.L., et al.: Cross-modal data programming enables rapid medical machine learning. Patterns (2020) 4
5. Dunnmon, J.A., Yi, D., Langlotz, C.P., Ré, C., Rubin, D.L., Lungren, M.P.: Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology **290**(2) (2019) 1, 3, 7
6. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639) (2017) 1, 3
7. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nature medicine **25**(1) (2019) 1
8. Ge, G., Yun, K., Samaras, D., Zelinsky, G.J.: Action classification in still images using human eye movements. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2015) 3
9. Grøvik, E., Yi, D., Iv, M., Tong, E., Rubin, D., Zaharchuk, G.: Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri. Journal of Magnetic Resonance Imaging **51**(1) (2020) 7

10. Hayhoe, M.: Vision using routines: A functional account of vision. Visual Cognition **7**(1-3) (2000) 6
11. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in cognitive sciences **9**(4) (2005) 2
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 7
13. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr database (2019). https://doi.org/https://doi.org/10.13026/C2JT1Q, https://physionet.org/content/mimic-cxr/1.0.0/ 4
14. Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., et al.: Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. Scientific data **8**(1) (2021) 2, 3, 4
15. Karessli, N., Akata, Z., Schiele, B., Bulling, A.: Gaze embeddings for zero-shot image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017) 2, 3, 8
16. Khosravan, N., Celik, H., Turkbey, B., Jones, E.C., Wood, B., Bagci, U.: A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning. Medical image analysis **51** (2019) 3
17. Klein, J.S., Rosado-de Christenson, M.L.: A Systematic Approach to Chest Radiographic Analysis. Springer International Publishing (2019) 2
18. Lai, Q., Wang, W., Khan, S., Shen, J., Sun, H., Shao, L.: Human vs. machine attention in neural networks: A comparative study. arXiv preprint arXiv:1906.08764 (2019) 3, 8
19. for Imaging Informatics in Medicine (SIIM), S.: Siim-acr pneumothorax segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation (2019) 4
20. Murrugarra-Llerena, N., Kovashka, A.: Learning attributes from human gaze. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2017) 2, 3, 8
21. Papadopoulos, D.P., Clarke, A.D., Keller, F., Ferrari, V.: Training object class detectors from eye tracking data. In: European conference on computer vision. Springer (2014) 3
22. Qiao, X., Ren, P., Dustdar, S., Liu, L., Ma, H., Chen, J.: Web ar: A promising future for mobile augmented reality—state of the art, challenges, and insights. Proceedings of the IEEE **107**(4) (2019) 2
23. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017) 4
24. Ratner, A., De Sa, C., Wu, S., Selsam, D., Ré, C.: Data programming: Creating large training sets, quickly. Advances in neural information processing systems **29** (2016) 4, 6
25. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal (Online) **359** (2017) 1
26. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017) 4, 6
27. Saab, K., Dunnmon, J., Goldman, R., Ratner, A., Sagreiya, H., Ré, C., Rubin, D.: Doubly weak supervision of deep learning models for head ct. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2019) 4, 8
28. Saab, K., Dunnmon, J., Ratner, A., Rubin, D., Re, C.: Improving sample complexity with observational supervision. International Conference on Learning Representations, LLD Workshop (2019) 3, 8
29. Samson, R., Frank, M., Fellous, J.M.: Computational models of reinforcement learning: the role of dopamine as a reward signal. Cognitive neurodynamics **4**(2) (2010) 3, 5

30. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE International Conference on Computer Vision (2019) 3
31. Stember, J., Celik, H., Krupinski, E., Chang, P., Mutasa, S., Wood, B., Lignelli, A., Moonis, G., Schwartz, L., Jambawalikar, S., et al.: Eye tracking for deep learning segmentation using convolutional neural networks. Journal of digital imaging **32**(4) (2019) 3
32. Taylor, A.G., Mielke, C., Mongan, J.: Automated detection of moderate and large pneumothorax on frontal chest x-rays using deep convolutional neural networks: A retrospective study. PLoS medicine **15**(11) (2018) 7
33. Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., et al.: Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nature communications **11**(1) (2020) 2
34. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017) 4
35. Wang, X., Thome, N., Cord, M.: Gaze latent support vector machine for image classification improved by weakly supervised region selection. Pattern Recognition **72** (2017) 2, 3
36. Wu, S., Zhang, H., Ré, C.: Understanding and improving information transfer in multi-task learning. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SylzhkBtDB 4
37. Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.H., Kim, G.: Supervising neural attention models for video captioning by human gaze data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 3
38. Yun, K., Peng, Y., Samaras, D., Zelinsky, G.J., Berg, T.L.: Exploring the role of gaze behavior and object detection in scene understanding. Frontiers in psychology **4** (2013) 2
39. Zhang, H.R., Yang, F., Wu, S., Su, W.J., Ré, C.: Sharp bias-variance tradeoffs of hard parameter sharing in high-dimensional linear regression. arXiv preprint arXiv:2010.11750 (2020) 4

Supplementary materials contain:

- Fig S.1: Visualization of class gaps in gaze features.
- Table S.1: AUROC, F1-score, and precision for Gaze-WS.
- Fig S.2: Positive and negative samples from each dataset.
- Table S.2: Ablation studies for Gaze-MTL.
- Table S.3: Additional details for Gaze-MTL.
- Table S.4: A comprehensive list of gaze features for reference.

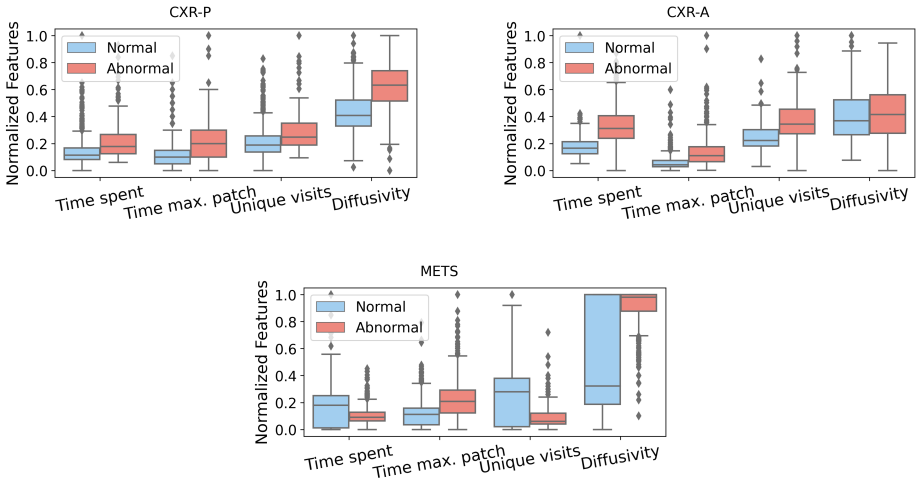

Fig. S.1: **Visualization of gaze features by class.** The gaze feature distributions differ significantly between normal and abnormal classes, indicating the class-discriminative signal present in gaze features.

Table S.1: **AUROC, F1-score, and precision for Gaze-WS.** The results are averaged over 10 random seeds with 95% confidence intervals. The operating points were chosen such that the recall was 53, 90, and 55 for CXR-P, CXR-A, and METS, respectively.

Dataset	Metric	Hand-Label Supervision	Gaze-WS
CXR-P	AUROC	81.5 ± 0.6	76.6 ± 1.0
	F1-score	56.6 ± 0.7	50.3 ± 0.8
	Precision	58.5 ± 1.4	46.4 ± 1.3
CXR-A	AUROC	83.8 ± 0.9	79.0 ± 1.3
	F1-score	84.1 ± 0.9	82.8 ± 0.6
	Precision	78.9 ± 1.5	76.5 ± 1.0
METS	AUROC	78.4 ± 1.8	72.9 ± 1.9
	F1-score	53.4 ± 1.4	50.4 ± 1.4
	Precision	53.5 ± 2.7	48.3 ± 2.6

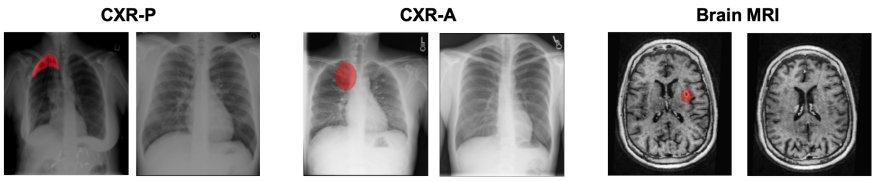

Fig. S.2: **Images from the three medical tasks.** For each pair, the left image is abnormal, and the right is normal.

Table S.2: **Ablation studies for Gaze-MTL.** We find that transfer learning – i.e. training a model only on the helper task (Gaze-Model) then finetuning the classification head on the target task (Gaze-TL) – is indicative of multi-task learning performance. The results are averaged across 5 random seeds with 95% confidence intervals. Accuracy is reported for (multi-class) salient region, while AUROC is reported for diffusivity and time.

Task	Gaze Feature	Gaze-Model	Gaze-TL	Gaze-MTL Boost
CXR-P	**Salient Region** (Acc)	83.7 ± 0.9	82.3 ± 0.5	$+1.4$
	Diffusivity	54.8 ± 4.5	57.5 ± 4.7	$+0.4$
	Time	61.6 ± 4.4	62.2 ± 4.6	$+0.7$
CXR-A	Salient Region (Acc)	59.6 ± 2.2	59.5 ± 2.2	-0.3
	Diffusivity	62.2 ± 3.8	63.3 ± 5.7	$+0.5$
	Time	73.2 ± 1.7	75.0 ± 1.0	$+0.1$
METS	Salient Region (Acc)	90.2 ± 0.6	81.4 ± 0.7	$+0.3$
	Diffusivity	94.1 ± 0.5	89.7 ± 0.7	$+2.4$
	Time	92.0 ± 0.5	82.6 ± 0.8	-4.9

Table S.3: **Addtional details for Gaze-MTL.** Helper task performance and parameters.

Task	Learning Rate	Weight Decay	Gaze Feature	Helper Task Weight (α)	Helper Task Performance
CXR-P	0.0001	0.0001	Salient Region	0.5	84.2 ± 1.5 (Acc)
CXR-A	0.0001	0.01	Diffusivity	1.0	61.5 ± 8.1 (AUROC)
METS	0.0001	0.00001	Diffusivity	0.5	92.3 ± 0.6 (AUROC)

Table S.4: **Additional gaze features.** A comprehensive list of gaze features that we have analyzed in the experiments (for reference). Here, we denote a gaze sequence by $g = \{(g_x^j, g_y^j, t^j), j = 1, .., n\}$ and the total time by $\sum_1^n t^j = T$.

Feature	Description
Time	Mean and variance of $[t^1, t^2, ..., t^n]$
Distance	Mean and variance of $[d_2, d_3, ..., d_n]$, d_i is the distance between consecutive fixations
Velocity	$\frac{\sum d_i}{T}$
Fixations	Total number of fixations n, and the fixation rate $\frac{n}{T}$
Alpha angle	Mean and variance of $[\alpha_2, \alpha_3, ..., \alpha_n]$, where α_i is the angle between the segment connecting (g_x^{i-1}, g_y^{i-1}) and (g_x^i, g_y^i), and the horizontal plane
Beta angle	Mean and variance of $[\beta_2, \beta_3, ..., \beta_{n-1}]$, where β_i is the angle between three fixations (g_x^{i-1}, g_y^{i-1}), (g_x^i, g_y^i), and (g_x^{i+1}, g_y^{i+1})
Side bias	Percentage of time spent on the left, right, top, and bottom halves