Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment

YUYANG GAO, Emory University, USA
TONG STEVEN SUN, George Mason University, USA
LIANG ZHAO, Emory University, USA
SUNGSOO RAY HONG†, George Mason University, USA

While Deep Neural Networks (DNNs) are deriving the major innovations through their powerful automation, we are also witnessing the peril behind automation as a form of *bias*, such as automated racism, gender bias, and adversarial bias. As the societal impact of DNNs grows, finding an effective way to steer DNNs to align their behavior with the human mental model has become indispensable in realizing fair and accountable models. While establishing the way to adjust DNNs to "think like humans" is in pressing need, there have been few approaches aiming to capture how "humans would think" when DNNs introduce biased reasoning in seeing a new instance. We propose ***Interactive Attention Alignment*** (IAA), a framework that uses the methods for visualizing model attention, such as saliency maps, as an interactive medium that humans can leverage to unveil the cases of DNN's biased reasoning and directly adjust the attention. To realize more effective human-steerable DNNs than state-of-the-art, IAAintroduces two novel devices. First, IAAuses ***Reasonability Matrix*** to systematically identify and adjust the cases of biased attention. Second, IAAapplies ***GRADIA*** , a computational pipeline designed for effectively applying the adjusted attention to jointly maximize attention quality and prediction accuracy. We evaluated Reasonability Matrixin Study 1 and GRADIAin Study 2 in the gender classification problem. In Study 1, we found applying Reasonability Matrixin bias detection can significantly improve the perceived quality of model attention from human eyes than not applying Reasonability Matrix. In Study 2, we found using GRADIAsignificantly improves (1) the human-assessed perceived quality of model attention and (2) model performance in scenarios where the training samples are limited. Based on our observation in the two studies, we present implications for future design in the problem space of social computing and interactive data annotation toward achieving a human-centered steerable AI.

CCS Concepts: • **Computing methodologies** → **Learning settings**; **Machine learning**; • **Human-centered computing** → **Human computer interaction (HCI)**; **Interaction paradigms**.

Additional Key Words and Phrases: Explainable AI, Alignable AI, Steerable AI, IAA, Reasonability Matrix, GRADIA

ACM Reference Format:
Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong†. 2022. Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 489 (November 2022), 28 pages. https://doi.org/10.1145/3555590

Code available at:https://github.com/YuyangGao/GRADIA, † indicates the corresponding author.

Authors' addresses: Yuyang Gao, yuyang.gao@emory.edu, Emory University, Atlanta, Georgia, USA; Tong Steven Sun, tsun8@gmu.edu, George Mason University, Fairfax, Virginia, USA; Liang Zhao, liang.zhao@emory.edu, Emory University, Atlanta, Georgia, USA; Sungsoo Ray Hong†, shong31@gmu.edu, George Mason University, Fairfax, Virginia, USA.

2573-0142/2022/11-ART489 $15.00
https://doi.org/10.1145/3555590

	Reasonable Attention	Unreasonable Attention
Accuracte Prediction	**RA:** Reasonable Accurate	**UA:** Unreasonable Accurate
Inaccurate Prediction	**RIA:** Reasonable Inaccurate	**UIA:** Unreasonable Inaccurate

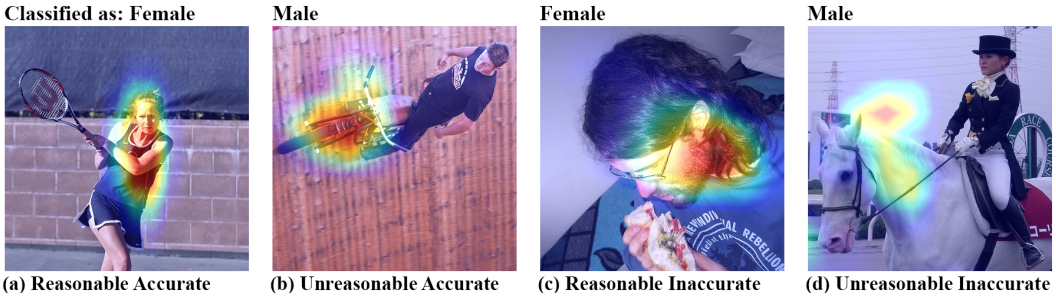

Fig. 1. **Reasonability Matrix** at the top with the four examples in a gender classification problem: (a) **Reasonable Accurate**: the attention given to an image is reasonable while prediction is also accurate, (b) **Unreasonable Accurate**: a substantial amount of attention is given to "contextual" features which make the attention unreasonable while the prediction is accurate, (c) **Reasonable Inaccurate**: despite the reasonable attention given to gender-intrinsic features, the prediction is not accurate, and (d) **Unreasonable Inaccurate**: the attention is unreasonable and the prediction not accurate.

1 INTRODUCTION

Deep Neural networks (DNNs) are becoming the powerhouse of innovation in our society [42]; they drive vehicles on behalf of humans [83], replace repetitive tasks for comic artists [89], and provide powerful support for border control agencies to boost security [51]. While we are witnessing how DNNs' powerful automation can benefit humanity, we are also experiencing the peril behind the automation as a form of *bias* [26]. As understanding how DNNs make predictions when they fail is the first step to remove the bias and retain the model fairness and accountability, recent years have seen an explosion of interest in model interpretability and model failure analysis [28, 32–34, 41]. However, DNNs offer limited transparency regarding the logical structure for prediction compared to other white-box models. Such limited transparency imposes challenges when humans attempt to determine the patterns when DNNs make predictions based on biased reasoning [42] and adjust the DNN's behavior based on the patterns [40, 41].

Although the notion of "bias" can be broadly defined [20, 79], bias is mainly caused by the bias already encoded in data used for training models [4, 13, 20]. In an image classification task, for example, using a training set where the class distribution is highly skewed towards a particular class [7, 13, 20] can decrease a model performance for a minority class [52, 87]. Even when using techniques for handling such imbalanced training-set problem [17, 38], DNNs can still be vulnerable to ***contextual bias*** [86] which arises when images in the training set have some contextual objects commonly co-occurring with a particular class [13]. For example, in a gender classification problem, model explanation results can give attention to contextual objects, such as baseball bats, snowboards, or kitchenware, rather than intrinsic features directly related to classifying a gender [86]. The negative consequences caused by such biased model attention are well represented in recent studies through their catchphrase, such as Hendricks et al.'s "women also snowboard" [13], or Zhao et al.'s "men also like shopping" [92]. Steering DNNs to a direction that decouples "spurious correlation" has been an important and open research topic [78].

Modeling mathematical or computational solutions to detect such "spurious correlations" can be challenging, complex, and hard to be used [13, 75]. However, from the human eyes, detection of such biased reasoning can be "just noticeable". To make DNNs think "more like humans" in the case of biased reasoning, we design and evaluate the framework of Interactive Attention Alignment(IAA, hereinafter). IAAenables a "direct" feedback loop between humans and realizes detection and removal of biased reasoning than state-of-the-art with the two novel features. First, ***Reasonability Matrix*** helps a model builder systematically investigate the cases based on biased reasoning. In doing so, Reasonability Matrixuses a human-based Boolean model attention quality assessment based on model attention explanation methods, such as a heatmap [80] (i.e., a given attention quality is reasonable/not reasonable) on the top of accuracy (i.e., a given prediction is accurate/not accurate). Combining the two criteria, Reasonability Matrixcategorizes unseen into four, which is explained in Fig. 1 While there have been some studies that explained how the model bias is related to the degree to which the model attention has common ground with human-reasoning attention [13, 23, 71, 74], few attempts to systematically collect human assessment in improving the model. Next, we present ***GRADIA***, a novel DNN fine-tuning pipeline that jointly maximizes the attention quality and model accuracy via back-propagation on both prediction and attention loss based on the extension of Grad-CAM [80]. While a few former works, such as Attention Branch Network (ABN) [68] focused on demonstrating the technical feasibility of applying a human's adjusted attention in improving DNNs, there has been no model agnostic approach that attempts to handle contextual bias through systemic detection, adjustment, and update.

We conducted two studies to evaluate Reasonability Matrixand GRADIArespectively in a gender classification problem. In our evaluations, we used 5,000 images collected from Microsoft COCO dataset [59] that half show females and the rest show males. In Study 1 (S1), we compared the following four conditions in selecting instances for fine-tuning a DNN model: (1) the first baseline that only uses inaccurate instances and does not use attention loss, (2) the second basline uses inaccurate instances but applies both the prediction and attention losses, (3) the first experimental condition that uses unreasonable instances but excludes inaccurate instances in fine-tuning, and (4) the last experimental condition that fully uses both inaccurate and unreasonable instances using GRADIA. The results show that the quality of attention assessed computationally and by humans has increased significantly from the order of (1), (2), (3), and (4). The results show how using Reasonability Matrixcan significantly reduce the chance of reducing unbiased reasoning than baselines. In Study 2 (S2), we measured how GRADIAcan improve ABN, the state-of-the-art technique for using human-adjusted attention. In S2-1, we focused on the attention quality measured by humans and computationally. We found that the quality of attention based on the GRADIAsignificantly outperformed ABN. In S2-2, we measured how using the GRADIAcan affect a model accuracy performance where the training samples are limited. We found the predictive power of DNNs can be improved by applying the GRADIAthan ABN. While we observed that both GRADIAand ABN-based approach can effectively leverage the adjusted attention in accelerating a model performance, GRADIAcan achieve a significantly larger improvement than the ABN-based approach under the few sample training scenarios.

The studies shed light on the potential and need for more deeply understanding approaches that can enable humans to directly indicate the way model thinks to align. The studies also draw the need of understanding the intelligent design that can efficiently and effectively capture subtle human reasoning when aiming to adjust the ways DNNs think. Finally, we discuss how IAAcan be applied in different application areas, which span subjective, objective, normative, atypical, personalized, or social tasks.

This work offers the following contributions:

- **Methodological contribution**: We propose a novel methodological framework of Interactive Attention Alignmentthat leverages Reasonability Matrixto (1) systematically detect biased reasoning and (2) effectively remove it through a direct human intervention.
- **Technical contribution**: We present GRADIA, a novel technique that strikes the balance between prediction accuracy and attention accuracy in fine-tuning DNNs. GRADIAcan be readily applied to any existing DNN model without modifications for its model agnostic property.
- **Empirical contribution**: Through the two studies, we present the results of the two studies that indicate the effect of IAAand GRADIAin steering DNNs.
- **Implications for Design**: We reflect on our observations through the two studies and draw the potential new designs that can further facilitate efficient and effective steering of DNNs.

2 RELATED WORK

Debiasing Machine Learning (ML) models and more specifically DNNs is closely related to gaining insights regarding how those models work and connecting the insights to steer their behavior. In this review, we introduce how human factor researchers have devised the interactive design for facilitating ML engineers' reasoning about ML models. Then we discuss how the broader ML communities have evolved human-in-the-loop and computational modeling toward changing the way DNNs work.

2.1 Understanding How DNNs work through Interactive Tools

Since the early release of Weka [36] and Fail et. al's proposal of the notion of Interactive ML [25], research communities in Human Factors have designed, built, and deployed interactive tools to improve the interaction between humans and models [45, 73]. For example, one of the early works that have been introduced in the CHI community is *EnsembleMatrix*, which presents a way to linearly combine multiple models to create a new [82]. Subsequently, *iVisClassifier* [19], Alshallakh et. al's system [2], *Squares* [76], and *RegressionExplorer* [24] have applied visual analytic approaches for investigating ML models. FeatureInsight presented a design for ideating features in building a model [12] while *ModelTracker* proposed a system that support a life cycle of model building process through interactive visualization [3]. *MLCube Explorer* presents a way to compare multiple models [46].

With the advent of DNNs, *Prospector* [56], *RuleMatrix* [67], and *VINE* [11] suggest interaction techniques or systems that can increase transparency of black-box-based models' prediction patterns. Another line of research focused on understanding a specific type of DNN's behavior with interactive tools [45, 73]. *LSTMVis* [81] and *RNNVis* [66] proposed visual analytic systems for RNNs, improving the way people perform sequence modeling. Liu et. al [63] and Bilal et al. [9] presented a visual analytics tool dedicated to examining CNNs. Liu et. al [62] and GAN Lab [47] examine the visual analytic approaches to better understand the process of building generative models. A more model-agnostic interactive design includes the TensorFlow Graph Visualizer which is designed for advanced users to understand the architecture of DNNs [88]. Finally, the community is in the early stage of exploring fairness and discrimination issues in ML; a few notable approaches, such as FairVis [14] and Silva [90] started to investigate the new form of design that can help us to handle issues of Fairness, Accountability, and Transparency (FAccT).

The approaches we covered demonstrate a diverse set of designs proposed to improve the way we work with ML models. With some exceptions, the work typically focused on understanding how ML models work rather than how to adjust. Such an approach for DNN revision is more scarce when focusing on the domain of FAccT.

2.2 Computational Approaches: Relying on Humans as an Oracle vs. Modeling Bias

The core mission of ML research is to build a "better" model. While the notion of "better" was predominantly model performance in the past, as the societal impact of DNNs grows, the metric has been diversified towards accommodating more advanced, more "human-like" values, such as unbiasedness, fairness, accountability, transparency, and beyond [20]. In the computational venues, the approaches can be categorized into two—Human-in-the-loop or computational modeling—depending on how the approach involves the human perspective in model improvement.

Human-in-the-loop approach actively uses human input to more directly inject human perspective into DNNs. One school of approach in Human-in-the-loop allows humans to embed predefined human-driven principles and rules, such as desirable distribution and attribution priors, as inductive bias into the models. For instance, logic rules or rationales are augmented to the training process of DNNs [43, 91]. Another approach within this category leverages feature attribution methods into defining the objective function that can be used for users to incorporate human priors in building DNN [61]. While principle-level of "deductive" embedding can help us to interact with ML models from a more global perspective, people's prior level of view can be misaligned with the reality of the world. To mitigate this shortcoming, some approaches started to explore the way to incorporate human input in a more "inductive" way when they identify some cases not meet their viewpoints. Typically, this new type of human annotation-based approach presents a way to regularize some undesirable samples toward correcting attention [30, 68] and salience patterns [5, 29, 61] to enhance the reasonableness of the model prediction process. For example, Balayn et. al's work proposed a bottom-up method where humans can attain a global level of interpretability using the local level of explanation in an image classification task [6]. Some recent work also put emphasis on discovering the unknown unknown—the case that the ML model makes a false prediction but with high confidence—can fall into this category [57, 60]. While the former approach does not directly aim at steering DNNs' behavior based on humans' undesirable observations, one notable approach of Attention Branch Network (ABN) enables an adjustment of a model's prediction behavior through a direct human intervention [68]. Very recently, there have been some recent attempts to align humans and the model attention maps using gaze information in medical image analysis applications, such as X-Ray image assessment [50, 85, 95], to help build more explainable Computer-Aided Diagnosis systems. However, the methods proposed in these works can be very domain-specific and thus cannot be applied to general image data. Besides, few studies have started applying similar ideas to tackle other data types, such as texts [44, 77], attributed data [84], and graph-structured data [29].

Since bias and discrimination have increasingly become critical in ML [20], a fast-increasing number of computational approaches have been proposed for modeling bias for debias [72, 96]. The majority of computational modeling approach tackles inherent bias present in data as well as the complex interaction between data in the three stages of before (i.e., pre), while (i.e., in), and after (i.e., post) training ML models. Pre-processing adjusts the data distribution to guarantee a "fair" representation of the different classes in the training set [52, 75]. The assumption is that if a model is trained on discrimination-free and "balanced" data, its predictions will not be discriminatory [15, 48]. Despite the effectiveness of such approaches, the pre-processing-based method may still be vulnerable to some bias types [1]. For instance, one notable challenge is contextual biases that arise in a data-balanced setting due to latent, unlabeled, and "spurious" correlations between contextual features (e.g., kitchenware such as dishes) closely associated with a particular class (e.g., female class) [86]. In-processing methods improve existing learning algorithms to account for fairness as well, instead of merely the predictive performance [16, 37]. Different from model-agnostic pre-processing approaches, in-processing fall into an algorithm-aware case,

and methods devised for DNNs are rather scarce. Post-processing approaches adjust the resulting models by "correcting" the decision boundaries that lead to redlining for a fair representation of different subgroups in the final decision process [49]. Although there is a breadth of approaches discussed in this area, this line of research focuses relatively less on directly eliciting human knowledge in refining ML models based on the weaknesses observed in a validation set [70].

Our review reveals that the annotation-based human-in-the-loop approach can be an effective way to more directly embed human knowledge in steering DNNs. To make a more practical design belong to this category, the review also identifies the two utmost challenges that need to be tackled. First, formalization of methodological guidance—which helps humans to systemically detect the biased reasoning in reviewing a model's behavior and adjust the biased cases—is missing. Second, the technical side of improvement can be necessary to make the technical pipeline more broadly adoptable while not penalizing the model accuracy. For instance, one of the state-of-the-art approaches, ABN requires modification of base model architecture which makes the solution to be not model-agnostic [68]. Also, ABN directly generates the attention maps using a separate network branch which increases the risk of overfitting when applying human-adjusted attention maps. In short, this review identifies the challenges that belong to the both human factors and technical sides.

3 METHODOLOGICAL FRAMEWORK OF INTERACTIVE ATTENTION ALIGNMENT

We elaborate on our framework of IAAdevised for steering the way DNNs "think" based on human knowledge. Our framework has two novel components: (1) What to adjust: building of the Reasonability Matrixto systemically detect predictions made based on unreasonable/biased reasoning and adjust, and (2) How to adjust: applying GRADIAto leverage the adjusted attention maps in improving DNNs. Our framework is depicted in Fig. 2.

3.1 What to Adjust: Reasonability Matrix

The first stage in our framework aims at identifying the instances made based on biased reasoning. Based on the former work that demonstrates the benefit of considering model's reasoning via model explanation methods [13, 86], Reasonability Matrixelicits from human annotators regarding the *attention accuracy*: whether the model explanation given to an instance is *reasonable* for classifying the instance into a particular class. Specifically, we postulate that a human annotator can determine the whole, or some part of, attention given to an image is either *intrinsic attention*–the attention directly relevant for a classification–or *contextual attention*–the attention that shows "spurious correlation" between the object and a specific class (e.g., kitchenware and female, or a baseball bat and male). To help annotators to decide as to whether attention given to an instance is reasonable, we use the following two-step validation.

- **Q1. Intrinsic attention**: Is the attention given to an image presents sufficient details for a human annotator to classify the instance?
- **Q2. Contextual attention**: Using the attention given to an image, can a human annotator recognize any contextual objects?

We consider the given attention is reasonable when a human annotator answers positive for Q1 and negative for Q2. Combining the attention accuracy with conventional model accuracy, Reasonability Matrixleads to the four cases as follows:

- **RA. Reasonable Accurate**: The attention only focuses on intrinsic features without containing contextual features while the prediction result is also accurate (e.g., see Fig. 1 (a)).

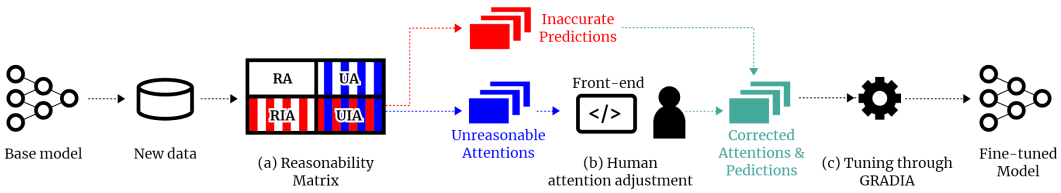

Fig. 2. Overview of our methodological framework of interactive attention alignment. (a) Building Reasonability Matrix, (b) adjusting attention maps of inaccurate predictions & unreasonable instances, (c) fine-tune the model using GRADIA.

- **UA. Unreasonable Accurate**: The prediction itself is accurate. But non-trivial amount of attention is given to contextual features, presumably due to contextual bias embedded in a training set (e.g., see Fig. 1 (b)).
- **RIA. Reasonable Inaccurate**: While the attention is reasonable, the prediction is inaccurate. This might be caised by the lack of data points similar to this type in a training set (e.g., Fig. 1 (c) shows that attention is given to a man's beard but the model's prediction is inaccurate).
- **UIA. Unreasonable Inaccurate**: The attention is not reasonable and the prediction is also not accurate (e.g., see Fig. 1 (d)).

With the rise of the FaccT research, a broader ML community started to establish the consensus that heavily relying on a single performance metric, such as model accuracy, error score, or confusion matrix can be detrimental for a comprehensive capturing of a model's "crucial shortcomings" [70]. The capability of having the attention accuracy in structuring the Reasonability Matrixmeans that we can use the quality of attention as a new way to evaluate DNN's performance. On top of the widely used model prediction accuracy metric, our framework proposes the following metrics as additional ways to add more rigor in evaluating DNN:

- **P1. Reasonable Accurate Performance**: The metric that indicates the proportion of the "right answer based on the right reasoning" (i.e., $\frac{RA}{RA+UA+RIA+UIA}$) which is more rigorous than the commonly used model accuracy performance (i.e., $\frac{RA+UA}{RA+UA+RIA+UIA}$).
- **P2. Attention Accuracy Performance**: The metric explains the proportion of instances with accurate attention (i.e., $\frac{RA+RIA}{RA+UA+RIA+UIA}$). This metric can be a proxy that shows the quality of model attention.

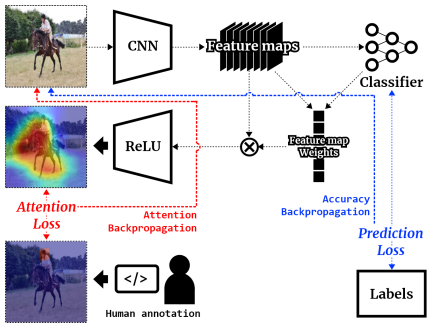

Fig. 3. The computational pipeline of GRADIA.

3.2 How to adjust: GRADIA

Using the proposed Reasonability Matrix, our framework elicits adjusted attention from human annotators. In this section, we introduce how GRADIAuses the adjusted attention maps in fine-tuning DNNs. In addition to minimize the error in the original training set, our major goal is to also minimize the losses from the three terms UA, RIA, and UIA in the Reasonability Matrix, which directly leads to our objective:

$$\min \ \mathcal{L}_{\text{Train}} + \mathcal{L}_{\text{UA}} + \mathcal{L}_{\text{UIA}} + \mathcal{L}_{\text{RIA}} \tag{1}$$

where $\mathcal{L}_{\text{Train}}$ denotes the model prediction loss on the original training set; $\mathcal{L}_{\text{UA}}$, $\mathcal{L}_{\text{UIA}}$, and $\mathcal{L}_{\text{RIA}}$ measure the errors on Unreasonable Accurate (UA), Unreasonable Inaccurate (UIA), and Reasonable Inaccurate (RIA) samples in the Reasonability Matrixof validation set, respectively.

For each term in Equation (1), there are two types of losses, namely *prediction loss*, denoted by $\mathcal{L}^{(p)}$, and *attention loss*, denoted by $\mathcal{L}^{(a)}$. Considering that different term (from different quadrant in Reasonability Matrix) requires different focus and balance between prediction and attention, we further introduce the balance factors for each term to give the model the flexibility to better weight between the attention and prediction loss in different cases. Specifically, Equation (1) can be expanded into the following one:

$$\min \ \mathcal{L}_{\text{Train}} + (\alpha\mathcal{L}^{(p)}_{\text{UA}} + (1-\alpha)\mathcal{L}^{(a)}_{\text{UA}}) + (\beta\mathcal{L}^{(p)}_{\text{UIA}} + (1-\beta)\mathcal{L}^{(a)}_{\text{UIA}}) + (\gamma\mathcal{L}^{(p)}_{\text{RIA}} + (1-\gamma)\mathcal{L}^{(a)}_{\text{RIA}}) \tag{2}$$

where the parameters α, β, and $\gamma \in [0, 1]$ are the tunable factors for controlling the balance between the prediction loss and attention loss for UA, UIA, and RIA samples, respectively.

This way, the first term $\mathcal{L}_{\text{Train}}$ can also be expanded as a special case $\mathcal{L}_{\text{Train}} = \mathcal{L}^{(p)}_{\text{Train}}$ where the weight for $\mathcal{L}^{(p)}$ is set to 1 and the weight for $\mathcal{L}^{(a)}$ is set to 0, such that the attention map labels are not required. Finally, by further expanding the first term and rearranging the terms for prediction losses and attention losses, the final objective of GRADIAcan be written as:

$$\mathcal{L}_{\text{GRADIA}} = \underbrace{\mathcal{L}^{(p)}_{\text{Train}} + \alpha\mathcal{L}^{(p)}_{\text{UA}} + \beta\mathcal{L}^{(p)}_{\text{UIA}} + \gamma\mathcal{L}^{(p)}_{\text{RIA}}}_{\text{prediction loss}} + \underbrace{(1-\alpha)\mathcal{L}^{(a)}_{\text{UA}} + (1-\beta)\mathcal{L}^{(a)}_{\text{UIA}} + (1-\gamma)\mathcal{L}^{(a)}_{\text{RIA}}}_{\text{attention loss}} \tag{3}$$

where $\mathcal{L}^{(p)}$ can be calculated by applying the Cross-entropy loss on the corresponding samples of each terms; and $\mathcal{L}^{(a)}$ is the newly proposed attention loss that measure the attention quality of the samples.

Notice that since both the original training set and the new data samples are considered as a whole for the fine-tuning of the model (as shown by the prediction losses inside the 'prediction loss' bracket in Equation (3)), the above fine-tuning setup can naturally ensure the previously learned knowledge to be preserved and does not require freezing of the model parameters. Concretely, the prediction loss in Equation (3) consists of the prediction loss on the original training samples (i.e. $\mathcal{L}^{(p)}_{\text{Train}}$) as well as new samples introduced in the fine-tuning stage (i.e. $\mathcal{L}^{(p)}_{\text{UA}}$, $\mathcal{L}^{(p)}_{\text{UIA}}$, and $\mathcal{L}^{(p)}_{\text{RIA}}$); while the attention loss consists of only the new samples introduced in the fine-tuning stage that has the human-adjusted attention labels available (i.e. $\mathcal{L}^{(a)}_{\text{UA}}$, $\mathcal{L}^{(a)}_{\text{UIA}}$, and $\mathcal{L}^{(a)}_{\text{RIA}}$).

Therefore, by introducing $\mathcal{L}^{(a)}$ into the fine-tuning step with GRADIA, the base DNN model can be jointly optimized both to generate higher quality attention maps and to make better and unbiased predictions on the original task. Our assumption is that this attention de-biasing process will also enhance the generalizability of the model to unseen data. As a result, GRADIAwill ultimately not only improve the model prediction accuracy, but also yield a more interpretable model.

To quantify the attention quality of the model, we propose a general attention loss for estimating the discrepancy between the model-generated attention maps and the human-annotated attention

labels of the selected samples from the validation set. Concretely, the attention loss can be computed as the following:

$$\mathcal{L}^{(a)} = \text{dist}(M, M') \tag{4}$$

where M and M' are the model-generated attention maps and the ground truth attention maps provide by the human annotators on those samples that require attention adjustment; the function $\text{dist}(x, y)$ can be a common divergence metric such as absolute difference or square difference. In practice, we found that absolute difference is more robust to the labeling noise from the annotator, while square difference can be more sensitive and yield a high loss on the border areas of the labels that could not actually be related to the object.

To generate the model attention maps on images, several existing works have been proposed. Response-based methods such as CAM [93] and ABN [27] typically require substantial modification on the DNN architectures that either hurt the model's performance and extensibility or over-decouple the generation process of attention and prediction. For example, to handle the performance issue, ABN proposed to add another module called 'attention branch' onto the model architecture that is specialized for generating the attention maps. However, this incurs much more parameters and hence more samples and time to train the model. Moreover, over-decoupling the components for producing attention and prediction substantially decreases the reliability that the attention is indeed the explanation for the prediction. In contrast, gradient-based methods such as Grad-CAM [80] does not require changes of the base model and hence is applicable to a wide range of various DNN models. Moreover, it does not incur additional model parameters and hence can be more computationally cheap. Furthermore, its attention and prediction are tightly coupled and hence maintain a strong dependency and reliability between the prediction and its attention map.

Therefore, we propose to build our pipeline by extending Grad-CAM which uses the gradient of the feature maps with respect to the target class to generate the attention maps. Mathematically, suppose the penultimate layer produces K feature maps, $A^k \in \mathbb{R}^{u \times v}$ where u and k are the width and height of the image of each feature map. The attention maps $M_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ for target class c can be computed as:

$$M_{\text{Grad-CAM}} = \text{ReLU}\left(\frac{1}{uv} \sum_k \sum_i \sum_j \frac{\partial Y^c}{\partial A^k_{i,j}} \cdot A^k\right) \tag{5}$$

where Y^c denotes the output of the model for predicting class c, and $\sum_i \sum_j \partial Y^c / \partial A^k_{i,j}$ denotes the weight of the feature map k for class c as also illustrated by Figure 3. To ensure the generated and labeled attention maps are in the same scale, we further normalize $M_{\text{Grad-CAM}}$ to the values between 0 and 1, as:

$$M = \frac{M_{\text{Grad-CAM}} - \min(M_{\text{Grad-CAM}})}{\max(M_{\text{Grad-CAM}}) - \min(M_{\text{Grad-CAM}})} \tag{6}$$

where the function $min(\cdot)$ and $max(\cdot)$ return the element-wise min and max of the input, respectively.

Notice that there are two major differences that distinguish the proposed GRADIAfrom the previous works, such as in [68] and [61]. First, instead of only correct the model attention on the misclassified instances, GRADIAbest leverages the Reasonability Matrixto identify which sample's attention needs to be adjusted based on both the prediction accuracy as well as attention accuracy. Second, GRADIAoffers the flexibility to control the balance between attention accuracy and prediction accuracy for different instances from different quadrants, which enables us to easily find the 'sweet spot' of the model that can produce more reasonable attention without scarifying the model accuracy. For example, for the instances in quadrant 'Unreasonable Accurate (UA)' of the Reasonability Matrix, we can set α in the range of $(0, 0.5)$ to force the model to pay less attention

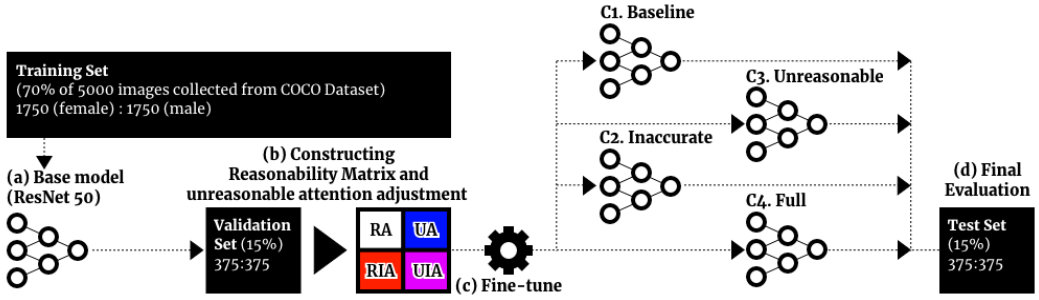

Fig. 4. Study 1 Methodological process: (a) Using a training set to build a base model, (b) using a base model to build a reasonability matrix using a validation set, (c) fine tune four different model that each represents a study condition, (d) use each model to validate using a test set.

to prediction loss while paying more attention to the attention loss, as the prediction was accurate but the attention was biased.

3.3 Directionality of Evaluation

In all, we introduced the two novel devices that establish IAA. The first device is the Reasonability Matrixthat helps users to systematically identify the instances with biased attention that need to be adjusted. Building the Reasonability Matrixrequires humans to scan the quality of attention of instances and adjust those when required. This process requires additional human-based computation costs. Therefore, in Study 1 (S1), we investigate if using the Reasonability Matrixin fine-tuning DNNs can counterbalance such costs by improving the model's quality in both computational and human-assessment-based ways. Meanwhile, the second device of GRADIAis devised for optimally utilizing the human-adjusted attention maps to strike the balance between prediction accuracy and attention accuracy in fine-tuning DNNs. To understand the effect of GRADIAin fine-tuning DNNs, we compare GRADIAto ABN [68], the best performing human-in-the-loop approach that applies human's direct data annotation in revising DNNs in Study 2 (S2).

4 S1: EFFECT OF REASONABILITY MATRIX

S1 aims at understanding the effect of using Reasonability Matrixin fine-tuning DNNs—i.e., how additional annotation effort required for building Reasonability Matrixcan specifically change the way DNNs think and behave?

4.1 Methodology

4.1.1 Task & Dataset: We chose gender classification in a binary scheme because the binary classification can be easy to understand the effect of IAAdue to its simplicity compared to other multi-class problems. In addition, since the gender classification problem is one of the widely used tasks in the research of fairness in ML [7, 92], well-annotated binary gender labels (including boundary) are available in public [59] which can be beneficial for reproducibility of results. Despite the aforementioned benefits, we are aware that classifying gender using the binary schema does not reflect the diverse viewpoint of gender in the real world. We state that our choice of a binary gender classification task does not represent our viewpoint on gender.

The general flow of S1 is explained in Fig. 4. For S1, we constructed our dataset from the Microsoft COCO dataset [59], which has been widely used in ML research [92]. We extracted images that had the word "men" or "women" in their captions. We then filtered out instances that (1) contain

both words, (2) include more than two people, or (3) human appear in the figure is nearly not recognizable from human eyes. For the scope of our problem, we manually selected images until we reach to collecting 2,500 female and 2,500 male images. In our settings, we used the data split of 70%:15%:15% for the training set, validation set, and test set. The three black boxes in Fig. 4 indicate the three splits.

4.1.2 Conditions. Using the the training set (3,500 images), we built a baseline model based on the ResNet50 architecture [39] (see Fig. 4 (a)). Specifically, for the training stage, the base model was trained for 50 epochs using the ADAM optimizer [53] with a fixed learning rate of 0.0001. To determine what to adjust for each condition, we asked a human annotator to assess the attention accuracy (i.e., asking if the attention given to an image is reasonable) in the explanation and built the Reasonability Matrix(see Fig. 4 (b)). In eliciting their assessment, they answered **Q1.** Does the focus area contains necessary details that enable you to classify a gender? (Yes/No) **Q2.** Does the focus area contains unnecessary details not related for you to classify a gender? (Yes/No) Note that the two questions are an easier version of what we paraphrased using Q1 and Q2 in 3.1. The instances where the annotator answered "Yes" in Q1 and "No" in Q2 were considered reasonable/accurate in terms of attention. Among 750 instances in a validation set, the annotator answered 232 unreasonable. For the instances assessed either as unreasonable, the annotator used a line drawing interface we provided to draw a binary mask that shows the areas (s)he felt the attention should be given to be reasonable.

Using the Reasonability Matrixand the adjusted binary attention maps that the annotator provided, we fine-tune the four different models that correspond to the four conditions following the same model training setups as described in the training stage, except in an open-loop manner without validation, though an additional validation set can be beneficial if available. The two conditions represent the baseline that can be built without building Reasonability Matrix. The rest of two needs Reasonability Matrixto be implemented (see Fig. 4 (c)):

- **C1. Baseline**: We fine-tuned the model only using the prediction loss; no human-labeled attention maps were used. Notice that this condition can be treated as a special case of GRADIAwhere all explanation losses are disabled, i.e. the hyperparameters α, β, and γ as shown in Equation (2) were all set to 0.
- **C2. Inaccurate**: In C2, we fine-tuned the model via the prediction loss and attention loss only using inaccurate instances (i.e. RIA and UIA) but excluding accurate instances (i.e., RA and UA)—which we can build this condition without assessing attention accuracy. To implement the condition, annotators must still need adjusted boundary of unreasonable instances within inaccurate (i.e., UIA). In total, we used 119 attention map labels. The hyperparameters α , β, and γ as shown in Equation (2) were set to 0, 0.5, and 0.8 respectively, for adjusting the model only based on the UIA and RIA samples.
- **C3. Unreasonable**: C3 applies the prediction and attention loss using unreasonables (i.e. UA and UIA) while excluding reasonables (i.e., RA and RIA). In total used 232 attention map labels. The hyperparameters α, β, and γ as shown in Equation (2) were set to 0.2, 0.5, and 0 respectively, for adjusting the model only based on the UA and UIA samples.
- **C4. Full**: This condition fully utilizes both prediction and attention loss from instances that are inaccurate and/or unreasonable (i.e. UA, RIA, and UIA) as shown in Equation (3), which in total used 295 attention map labels. The hyperparameters α, β, and γ as shown in Equation (2) were set to 0.2, 0.5, and 0.8 respectively. The values were selected based on a grid research [58] via the model performance on the validation set.

4.1.3 Measures & Study Apparatus. We evaluate the quality of the four models using the test set (The last black box in Fig. 4 (d)). In particular, we deployed the following measures to quantitatively/qualitatively evaluate the performance of the models.

- **M1. Prediction Accuracy Performance**: A measure that shows a model prediction accuracy.
- **M2. Reasonable Accurate Performance**: A measure that shows the proportion of Reasonable Accurate (RA) out of every instance, (i.e., P1 in 3.1).
- **M3. Intersection over Union (IoU)**: A measure that quantitatively assesses the quality of attention. Following the work on network dissection [8], we collect the ground truth attention from a human annotator for every instance in a test set. Then we make bit-wise intersection and union operations with the ground truth attention maps and each model's attention maps to measure how well the two attention masks overlap.
- **M4. Attention Accuracy Performance**: A measure that shows the proportion of the instances that human annotators assessed the attention is accurate/reasonable out of every instance in the test set (i.e., P2 in 3.1). To collect M4, we presented an interface where the annotators can see the four different attention made using C1, C2, C3, and C4 displayed based on random order, as shown in Fig. 5, top. The annotator answered the same two questions of Q1 and Q2 we used for assessing the attention accuracy of the validation set.
- **M5. Perceived Attention Quality**: A measure that shows the perceived quality of attention using a five-level Likert scale starting from "Very Poor (1)" to "Excellent (5)" instead of measuring the Boolean level of attention accuracy in M4. The question was: "Overall, please rate the quality of the focus." To collect M5, we prepared another interface where the annotators can see the four attentions displayed in a random order as shown in Fig. 5, bottom.

We explain in detail how we elicited the annotation results from participants for M4 and M5. M1 and M3 can be yielded computationally, and M2 can be calculated based on the results of M4. To acquire assessment results ran based on a consistent standard, we aimed at asking each annotator to assess the whole attention results of 750 images multiplied by 4 conditions (3,000 model attention in total). If we assume each annotator uses 10 seconds to assess one attention without resting, assessing the whole attention will take more than 8 hours. Such a level of commitment may not be suitable for crowdsourcing [21].

To recruit our participants, we used on-campus flyers in a public university in the United States and through word-of-mouth. As a result, we recruited 3 participants for M4 and 7 participants for M5. Upon the agreement of their participation, we sent a 15-minutes video that explains the concept of (1) "attention" in DNNs, (2) "contextual bias", and (3) the two criteria for assessing the quality of attention, (3-a) inclusion of intrinsic attention—whether the attention includes enough information for a human to classify a gender in an image, and (3-b) exclusion of contextual attention—whether the attention excludes contextual objects that are not directly relevant for a classifying a gender in an image. After they watch the video, we presented a quiz that asks three questions related to attention quality assessment. Once participants complete the quiz, we gave a web link to the M4 interface (Fig. 5, top) or the M5 interface (Fig. 5, bottom). We advised participants to take enough rest to refresh their attention. We also allowed them to perform their annotation across multiple days. On average, participants spent 5 days completing their assessment. Upon their completion, we compensated our participants with a gift card of $60 value.

4.2 Results

Overall, the most notable aspect that we observed in S1 is that all the conditions that adopted the attention loss (i.e. C2, C3, and C4) significantly decreased the contextual bias assessed by human eyes (M2, M4, and M5) and computationally (M3). This pattern implies that applying the attention

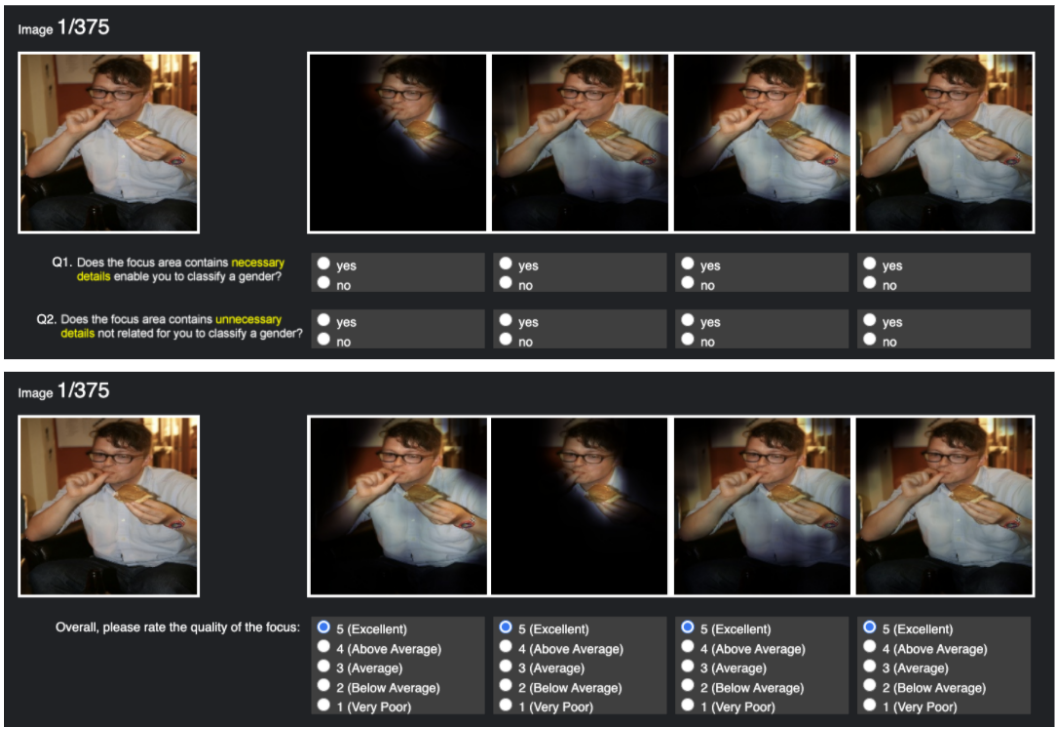

Fig. 5. Study apparatus used for eliciting M4. Attention Accuracy (top), and M5. perceived attention quality (bottom).

loss in fine-tuning DNNs can be effective in decreasing DNNs' biased reasoning caused by the contextual bias embedded in a training set. Further, for the four attention quality-related measures of M2, M3, M4, and M5, we identified that the ranking between the conditions was consistently C4 (the best performing condition), C3, C2, and C1 (the worst-performing condition). This pattern explains that putting more effort into detecting and adjusting the cases with biased reasoning can predict more effective in removing the contextual bias. The amount of effort we put into each

Conditions	Adjusted Attention #	Reasonability Matrix	M1	M2	M3	M4	M5
C1. Baseline	0 (0%)	$\begin{bmatrix} 306 & 310 \\ 33 & 101 \end{bmatrix}$	82.13%	40.80%	0.23 ± 0.19	45.20%	2.82 ± 1.13
C2. Inaccurate	119 (3%)	$\begin{bmatrix} 456 & 163 \\ 87 & 44 \end{bmatrix}$	82.53%	60.80%	0.32 ± 0.19	72.40%	3.68 ± 1.16
C3. Unreasonable	232 (5%)	$\begin{bmatrix} 497 & 117 \\ 99 & 37 \end{bmatrix}$	81.86%	66.27%	0.34 ± 0.18	79.47%	3.81 ± 1.13
C4. Full	295 (6%)	$\begin{bmatrix} 518 & 104 \\ 94 & 34 \end{bmatrix}$	**82.93%**	**69.07%**	**0.36±0.19**	**81.60%**	**3.97±1.08**

Table 1. S1 Results that show how measures (M1: Prediction accuracy performance, M2: Reasonable Accurate performance, M3: IoU, M4: Attention accuracy performance, and M5 Perceived attention quality) change across the conditions (C1: Baseline, C2: Inaccurate, C3: Unreasonable, and C4. Full. The best performing condition is bolded.

Condition Pairs	M3.IoU	M4.Attention Accuracy	M5.Perceived Quality
C1 vs. C2	**1.02e-16**‡‡	**2.19e-30**‡‡	**1.46e-287**‡‡
C1 vs. C3	**3.04e-22**‡‡	**3.23e-47**‡‡	**0**‡‡
C1 vs. C4	**1.60e-31**‡‡	**4.74e-53**‡‡	**0**‡‡
C2 vs. C3	0.1624	**0.0029**‡	**2.61e-08**‡‡
C2 vs. C4	**0.0007**‡‡	**0.0001**‡‡	**1.13e-36**‡‡
C3 vs. C4	**0.0475**†	0.3689	**1.41e-12**‡‡

(†: $p < 0.05$, ‡: $p < 0.01$, ‡‡: $p < 0.001$)

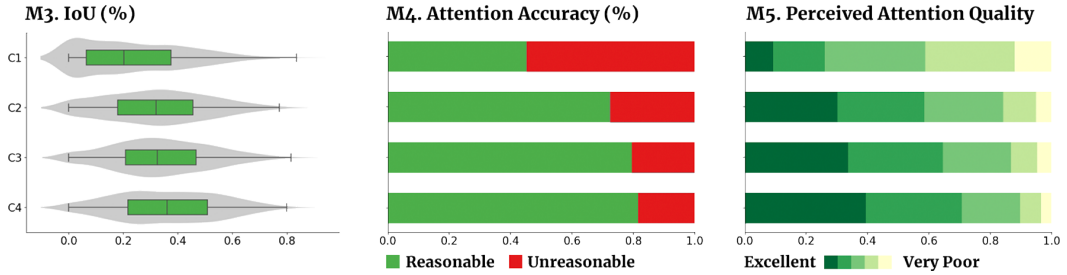

Fig. 6. The pairwise comparison results of M3, M4, and M5 (Top), distribution charts (bottom)

condition is shown in the second column of Table 1. Speaking of the effort vs. effect trade-offs, implementation of C2 requires seeing inaccurate instances and revising their attention boundary, which requires significantly less effort compared to C3 and C4. But even with that relatively less effort, we saw a dramatic increase in M2, M3, M4, and M5 in C2 compared to C1. As assessing every instance in a test set is required for building Reasonability Matrix, C3 & C4 require more effort than C2. We can observe that the gap between C2 & C3 or C2 & C4 has been less than the gap between C1 & C2. However, the gaps between C2 & C3 and C2 & C4 are still substantial.

We note that the focus of our approach is improving the model reasoning rather than increasing the prediction accuracy performance. However, one of our foci is *not* to decrease the model performance because of the human intervention, as observed in a recent study, namely the "accuracy-fairness gap". The gap has been originally observed in Barlas et. al's study which denotes the case where embedding the human's perspective in steering black-box models can result in dropping the prediction accuracy performance [7]. In general, we did not see a significant drop or increase in M1 between the conditions. This pattern indicates that the IAAcan still retain the M1 performance regardless of how many inputs will be used for finetuning. Still, we have used a 70%:15%:15% split for training, validation, and test sets, and increasing the portion for a validation set may cause an effect otherwise. Two implications we can derive from our observations in M1. First, studying how IAAcan robustly retain the M1 measure depending on the proportion between a training set and a validation set needs further research. Second, while not seeing the drop in M1 between C1 and the rest conditions is not unfortunate, to improve the efficacy of IAA, further study dedicated to improving both prediction accuracy and attention accuracy is essential.

We explain the results of our statistical analysis for M3, M4, and M5. As none of the independent variables we collected in M3, M4, and M5 follow the normal distribution, we applied Kruskal-Wallis H-test in identifying the differences between the conditions. In terms of M3 of IoU, we found a significant difference within the four conditions (p=4.13$e-34$, < 0.05). Dunn's test for post-hoc pairwise comparison found that every pair except the pair of C2 and C3 was significantly different. See the second column of the table in Fig. 6, as well as the left chart at the bottom.

In terms of M4, Kruskal-Wallis H-test found that the four conditions are significantly different with a p-value of 1.20e-64 (< 0.05). Dunn's test found a significant difference between every pair

except the pair of C3 and C4. For conducting the post-hoc power analysis for M4, we first calculated the Cohen's d effect sizes for the condition pairs. All the condition pairs that involve C1 have medium to large effect sizes: C1 vs. C2 (-0.57), C1 vs. C3 (-0.76), and C1 vs. C4 (-0.82). According to the power analysis results, almost all the condition pairs have enough statistical power except the condition pair of C3 and C4, which has a power below 80%. The third column in the table of Fig. 6, as well as the center chart at the bottom, shows the results. Since the approach applied the majority vote, it's important to check the degree to which the three annotators' agreed. Based on the Cohen's Kappa Agreement test, their results showed each pair of annotators ranged from 0.75 to 0.8, which are considered as having "Substantial agreement" and "Almost perfect agreement."

Regarding M5, the perceived attention quality received by 7 participants, the p-value yielded using Kruskal-Wallis H-test was close to 0 while Dunn's post-hoc pairwise test found the perceived qualities of all pairs are significantly different. The last column in the table of Fig. 6 and the chart at the bottom right show the results related to M5. The calculated effect sizes (Cohen's d) for the conditions, the collected data has enough power (over 80%). By computing the Cohen's d effect size of the Perceived Attention Qualities for the condition pairs, we found large effect sizes on three pairs of conditions: C1 vs. C2 (-0.75), C1 vs. C3 (-0.87), and C1 vs. C4 (-1.04). So, when comparing condition C1 with any other condition, a large but negative effect was detected. Even though other condition pairs' effect sizes are relatively small, with the large number of observations collected, the post-hoc power analysis tells us that we have enough statistical power (all above 80%) to detect any significant difference between the conditions' qualities, if we set the significance level as 0.05.

Finally, we aimed at "qualitatively" understanding if we could detect some patterns between conditions with our own eyes. To do so, we displayed four attention types generated with C1-C4 for every 750 images on a single wall. Two researchers looked at the patterns of attention between the conditions to understand how the attention varies depending on the conditions change. As a result, they found two notable patterns.

The first pattern we observed is what we named "altered gaze" where the gaze tended to gradually shift from the contextual object(s) to intrinsic objects as the condition moves from C1 to C4. Some examples of such "altered gaze" is presented in Fig. 7 where the initial gaze was posed on some objects such as tennis rackets, motorcycles, doll, or the background was shifted to humans. The next pattern is what we named "concentrated gaze". As we browse the attention from C1 to C2, C3, and C4, we found a tendency that the areas of focus tend to become smaller, focused, and concentrated on intrinsic objects directly related to classifying a gender. This "concentrated gaze" tended to make a positive effect when the baseline gaze includes some contextual objects which would gradually be excluded as the attention is shifted to C2, C3, and C4. Such examples are displayed in Fig. 7 where the earlier attention includes a human with the contextual objects such as the skateboard, animals, and surf excluding the contextual objects in C3 and/or C4.

4.3 Discussion

Through S1, we conclude that the core effect of applying Reasonability Matrixin solving the problem of "what to adjust" can be effective in increasing the quality by aligning what they see with what humans would also see. Also, a sharply increased M2 and M4 measures and an unchanged M1 measure imply that many accurate predictions made by a variety of DNNs can be due to biased reasoning embedded in a training set. As shown at the Reasonability Matrixpresented in the third column in Table 1, the number of "Reasonable Accurate" has gradually increased from 306 (C1) to 518 (C4) while the accurate instance with a biased reason, in our term, unreasonable accurate has decreased from 310 (C1) to 104 (C4). The proportion of accurate instances made based on the right reasoning (i.e., RA) out of every accurate instance (i.e., RA and UA) was merely 49% in C1. Then it dramatically increased to 73% in C2. Then gradually increased to 80% in C3, and reached 83% in C4.

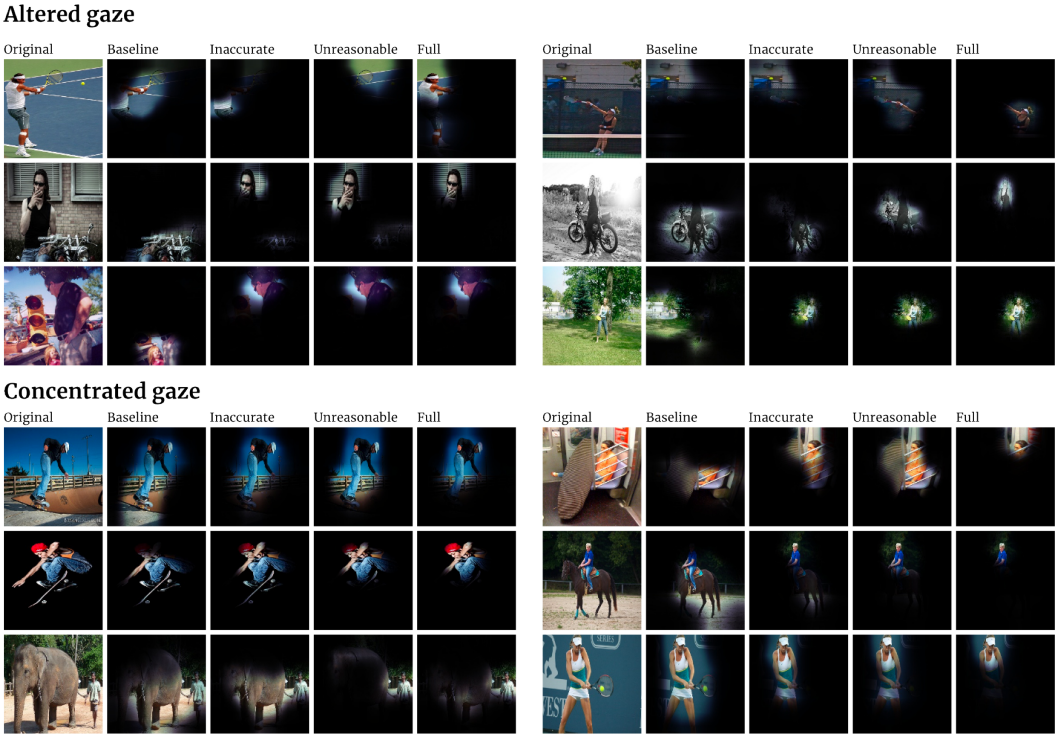

Fig. 7. Observed attention change patterns between conditions: altered gaze (top), concentrated gaze (bottom)

Such a pattern explains the potential vulnerability of existing DNNs when the test set includes a lot of obvious contextual objects. For instance, including some contextual object in an image could make the model predict a new instance to a particular class regardless of what the reality is.

Another pattern that we found noteworthy is the gradual increase of RIA (i.e., cases for making inaccurate predictions with right reasoning) from C1 to C4. We believe this pattern implies the potential drawback of IAAwhere excluding contextual objects from initial attention resulted in inaccurate prediction. As we discussed in our qualitative Post-hoc analysis, applying the GRADIAcan make the attention more focused. In case the initial attention in C1 was posed on the "right spot", we often observed that the focused gaze didn't include anything identifiable from our eyes as moving to C2, C3, and C4. We assume such change resulted in increasing RIA which might have contributed to the drop in M1 performance in C2, C3, and C4. In the future, understanding how to handle such cases of RIA can be a key to increasing M1 performance in improving the IAA. For example, penalizing the case where the attention map doesn't include the recognizable objects in an image can be used in designing an improved objective function.

5 S2: EFFECT OF GRADIA

The goal of S2 is to understand how using GRADIAin fine-tuning DNNs can improve the quality of DNNs than state-of-the-art. For this benchmark, we chose Attention Branch Network (ABN) [68]. Such a baseline choice has been made because the ABN is the only technique we identified enabling a human user to directly adjust the attention boundary to steer the DNN's model prediction behavior. In S2, we ran the two sub-studies, one for understanding the degree to which the quality of attention is improved (S2-1). Second, understanding if the GRADIAcan improve model performance in the

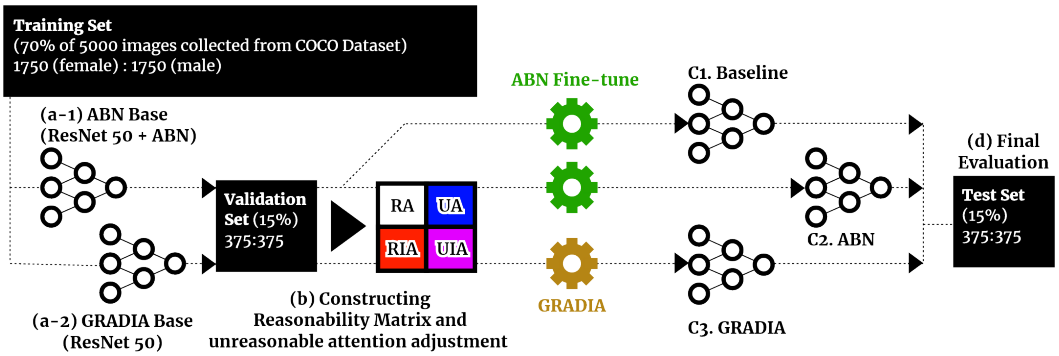

Fig. 8. Study 2 Methodological process: (a) Using a training set to build two base models, (b) using each base model to construct a reasonability matrix using a validation set, (c) fine-tune three different models that each represents a study condition, (d) evaluate the three models using a test set.

scenarios where the size of the sample is scarce (S2-2). Notice that to best control the influence of the Reasonability Matrixin the comparative studies, here in S2 instead of directly following the original workflow in [68] for training ABN which only leverage inaccurate samples, here we adaptively trained the ABN model under the same condition as GRADIAon deciding what sample to adjust, i.e. considering all UIA, UA, and RIA samples via the Reasonability Matrixof the corresponding initial models and fed all the identified unreasonable or inaccurate samples to adjust each model accordingly.

5.1 S2-1: Assessment through Perceived Quality of Attention

5.1.1 Methodology. In S2, we used the same dataset with the same split, as well as the same task we used in S1. Our goal in S2-1 is to compare the three conditions as follows:

- **C1. Baseline**: The base ABN model fine-tuned using prediction loss without using human adjusted attention (Fig. 8 (c-1)).
- **C2. ABN**: Beside the prediction loss, C2 used the additional 433 instances' adjusted attention maps elicited from human annotators to fine-tune the ABN attention maps following the attention loss as shown in Equation (3) in [27] (Fig. 8 (c-1)).
- **C3. GRADIA**: The proposed computational pipeline that fully utilizes both prediction and attention loss from instances that are inaccurate and/or unreasonable. Note that this condition is the same model we used as C4 Full in S1 (Fig. 8 (c-2)).

In preparing our conditions, we first used the training set to build two base models; BM1 with ResNet50 + ABN [27, 68] architecture for building the baseline condition and the ABN condition (see Fig. 8 (a-1)). Then we built BM2 with the base ResNet50 architecture [39] for constructing GRADIA-based approach (see Fig. 8 (a-2)). Same as S1, for the training stage for building base models, the two base models were trained for 50 epochs using the ADAM optimizer [53] with a fixed learning rate of 0.0001. Using inaccurate instances in a validation set identified with BM1, we built the baseline condition of C1. We then used the BM1 and BM2 in building Reasonability Matrixusing the validation set to build C2 and C3. Same as S1, a human annotator answered the Q1 and Q2 to assess 750 images' attention accuracy. Consequently, we found that 433 instances from BM1 and 295 from BM2 for GRADIAneed to be adjusted (Fig. 8 (b)) which we asked the human annotator to adjust every attention map. We applied ABN to build C2 using 433 adjusted maps. Then we applied GRADIAwith 295 adjusted maps to build C3. Similar to S1, here we fine-tune

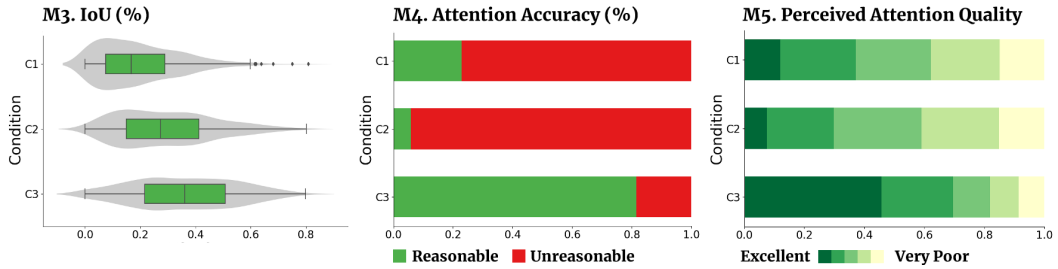

Fig. 9. The pairwise comparison results of M3, M4, and M5 (Top) and distribution charts (Bottom)

the models using the adjusted maps in an open-loop manner, although an additional validation set can be beneficial if available. Finally, we evaluate the quality of the three models of C1, C2, and C3 using the test set (see Fig. 8 (d)). In evaluating the three models, we used the same five performance measures in S1. In S2, we used the same data annotation interface in S1. For eliciting M4 of attention accuracy, we recruited 3 participants. For M5 of perceived quality of attention, we recruited 5 participants. The participants went through the same procedure of watching 15 minutes introduction video, performing a short quiz, and eliciting their perspectives using the study apparatus we provided. Upon their completion, we compensated for their participation with a gift card with the $60 of value.

5.1.2 Results. Similar to S1, there were no significant differences between conditions in terms of M1. On the contrary, attention quality-related measures all presented significant differences between conditions.

M3, the computation-based attention quality measure showed C3 GRADIAas the best performing condition, followed by C2 ABN, and C1 Baseline. Since the distribution of IoU didn't follow the normal distribution, we applied Kruskal-Wallis H-test. The test found that there is a significant difference between the three conditions with a p-value of 1.67e-59. Dunn's test found the p-values of all pairs are below 0.05 (see M3 related metric in Fig. 9). The gaps between the three conditions in M3 were all significant (see the second column of the table in Fig. 11).

Interestingly, M4 and M5, the human assessment-based attention quality measures disagreed with M3 in terms of the ranking between conditions. Same as M3, C3 turned out to be the best performing condition. However, both M4 and M5 showed that C1, the baseline condition outperformed C2.

In terms of M4, The Kruskal-Wallis H-test shows a significant difference in the Attention Accuracy Performance of the three conditions with a p-value of 1.22e-221. The Dunn's test results for the post-hoc pairwise comparison are at the top in Fig. 9. With the number of observations collected, the power analysis result shows a sufficient amount of power (all above 80%) in our data to detect

Conditions	Adjusted Attention #	Reasonability Matrix	M1	M2	M3	M4	M5
C1. Baseline	0 (0%)	$\begin{bmatrix} 147 & 462 \\ 25 & 116 \end{bmatrix}$	81.20%	19.60%	0.19±0.15	22.93%	2.96± 1.25
C2. ABN	433 (10%)	$\begin{bmatrix} 38 & 580 \\ 6 & 126 \end{bmatrix}$	82.40%	5.07%	0.29±0.18	5.87%	2.81± 1.16
C3. GRADIA	295 (6%)	$\begin{bmatrix} 515 & 108 \\ 97 & 30 \end{bmatrix}$	**82.93%**	**68.67%**	**0.36±0.19**	**81.60%**	**3.89±1.31**

Table 2. S2 Results. For each measure, the best performing condition is bolded.

Condition Pairs	M3. IoU	M4. Attention Accuracy	M5. Perceived Quality
C1 vs. C2	1.57e-24‡‡	7.31e-12‡‡	2.47e-07‡‡
C1 vs. C3	1.48e-59‡‡	1.20e-122‡‡	7.91e-213‡‡
C2 vs. C3	1.43e-09‡‡	4.94e-203‡‡	1.89e-288‡‡

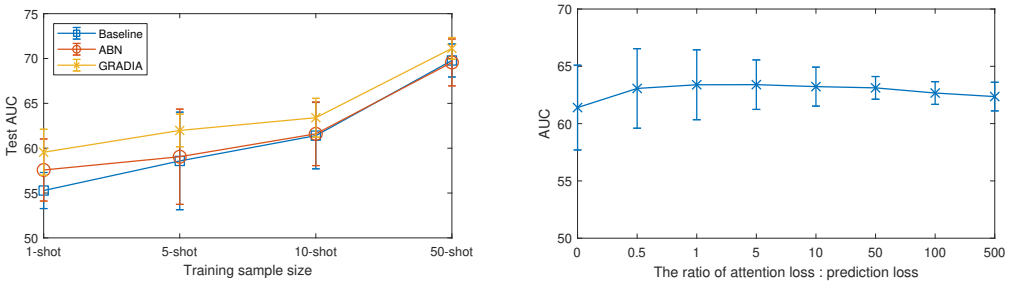

Fig. 10. Model performance under limited training samples. The data point represents the mean value over 10 random sample selection seeds, and the error bar here corresponds to the standard deviation. (a) The test AUC score comparison under different training sample size scenarios. (b) The sensitivity study of GRADIAunder the different ratios of perdition and attention loss on a 10-shot scenario.

any significant difference between the conditions' M4 performance. In addition, three annotators have substantial agreement with kappa coefficients that are at least 0.65.

The Kruskal-Wallis H-test in M5 found there was a significance ($p<0.05$). The post-hoc comparison indicates that the condition's qualities are significantly different, pairwise, with p-values less than 0.05 (see Fig. 9, M5). The pairwise effect sizes for M5 are: C1 vs. C2 (0.12), C1 vs. C2 (-0.72), and C2 vs. C3 (-0.87). So, any condition pair with C3 had a large negative effect size. By conducting the post-hoc power analysis, even with a relatively small effect size between C1 and C2, the number of observations was large enough to maintain the power (above 80%) for further statistical tests.

5.2 S2-2: Assessment through Model Performance Increase

5.2.1 Methodology. In this subsection, we studied how the DNN models can benefit from IAAto gain a better generalization power under the scenarios where the training samples are limited. Specifically, we randomly sample a certain amount of training samples from our original validation set pool and provide the human annotation labels for each selected sample. Next, we use the samples as well as the attention labels as the training set to fine-tune a pre-trained ResNet50 and evaluate the model performance with the ROC-AUC score on the original test set. Similar to the previous study, we made comparative studies among the following three models: 1) Baseline, 2) ResNet50 + ABN, and 3) GRADIA. Notice that this setting does not require the reasonability assessment like the previous studies. Instead, here we assume that all samples have attention labels available, as we have enough human resources under the scenarios where the training samples are limited.

5.2.2 Results. We simulated four limited training sample scenarios, i.e. 1-shot, 5-shot, 10-shot, and 50-shot, where the number of shots means the total number of training samples per class. For example, in our gender classification dataset, 5-shot means we sample a total of 10 image samples, 5 for male class and 5 for female class.

As shown in Fig. 10 (a), we present the test AUC score under the four training sample size scenarios. The data point here represents the mean value of the test AUC scores over 10 random training sample selection seeds, and the error bar here corresponds to the standard deviation. We can see that the proposed GRADIAoutperforms all baseline models by a significant margin under

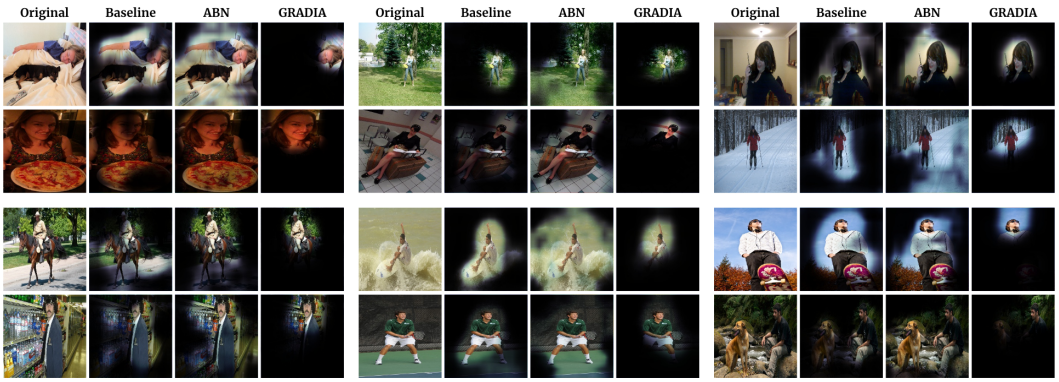

Fig. 11. Twelve sets of attention maps that show how attention varies across the three conditions

all scenarios studied. Specifically, the GRADIAis able to improve the baseline model performance by 7.7%, 5.8%, 3.3%, and 1.9%, respectively under 1-shot, 5-shot, 10-shot, and 50-shot scenarios. Notice that ABN could also improve the model performance by leveraging the additional human attention labels, but generally much less effective than the GRADIA. This is largely due to the additional layers and model parameters that are required in building the attention branch in ABN. This, on one hand, requires a large number of samples to learn how to generate attention, and on the other hand, makes the model more complex and prone to overfitting under a small training sample size.

We also studied how will different balance factors between the prediction and attention loss affect the final model performance. Fig. 10 (b) shows the sensitivity study of GRADIAunder different ratios of perdition and attention loss on the 10-shot scenario. We can obtain two major findings: 1) the variance of the AUC score will be reduced as more weights are put to the attention loss; 2) the improvement of model performance is not very sensitive to the choice of the balance factor, as long as it is not set to 0 (which means we don't use the attention loss, thus this is equivalent to the baseline model). Those observations justified the general effectiveness of the GRADIAunder the scenarios where we don't have enough samples to train a model. Thus, our study can benefit the application domains where a large amount of labeled data are difficult to obtain, such as in the weakly supervised learning scenarios [94].

5.3 Discussion

In general, fine-tuning DNNs using the GRADIAshowed clear improvement on the quality of attention side without penalizing the prediction accuracy performance. Computational and human-assessment-based measures all indicated this pattern. Just like in S1, we did not observe the "accuracy-fairness gap" in S2. Compared to C1, C3's number of instances categorized as Reasonable Accurate grew more than 3 times (147 in C1 to 515 in C3) while Unreasonable shrunk by more than one third (578 in C1 to 138 in C3). Like what we've observed in S1, Reasonable Inaccurate increased from 25 to 97, showing the side effect of having the "concentrated gaze".

What made us rather unexpected was the results of C2. Although M3 showed significantly improved IoU than C1, human-based assessments received lower ratings compared to C1. In order to understand such disagreement between computational and human-assessment-based measures, we posed every attention map result and checked how the gazes vary between the three conditions. As a result, we found two notable patterns. First, ABN tends to add more attention weight to the areas that are necessary to identify gender, which would likely elicit "yes" on the Q1: first question in the reasonability questionnaire. Indeed, C2 had the highest number of yes in Q1 (n=739 out of 750) than C1 (n=725) and C3 (n=670). However, at the same time, we found the second pattern that

the attention would likely disperse in C2 to include more "contextual objects" while C3 strictly concentrates its gaze on intrinsic objects. We assume this aspect of "dispersing gaze" of ABN made human annotators answering the second reasonability question of Q2 also "yes" (i.e., they perceive the attention was given to contextual object(s) not directly related to classifying gender). For C1, 706 instances were rated they include contextual objects in their attention maps while C1 had 576 C3 had only 112. This aspect influenced C2's Reasonability Matrixto include a huge amount of Unreasonable Accurate samples (n=580) which is worse than baseline (n=462).

This misalignment between computational and human-assessment-based measures implies the necessity of diversifying the way we understand the DNN's performance, especially when the model's task encompasses context-dependent and subtle human subjectivity [22]. When Mitsuhara et al. evaluated the ABN in their study, they used the "deletion metric" which measures the decrease of prediction accuracy score change by gradually deleting the high attention area of an attention map [68]. Such an approach may result in a model with high accuracy, but this doesn't guarantee that model made a prediction with a "right" reason [13]. By including humans in the loop of evaluation, we argue that we can design/build a model trustworthy, reasonable, and high-performing.

Besides, in our few sample learning studies, we observed that the GRADIAcan improve both model explanation quality as well as predictive power with the additional supervision of the model's attention. As nowadays DNNs are known to be data-hungry and typically having a large number of training samples is a nontrivial task for data scientists, we hope our findings in S2-2 can provide useful insights and motivate future studies on the directions toward learning more reliable and interpretable DNNs with less amount of labeled data.

6 IMPLICATIONS FOR DESIGN

The overarching insights derived us to investigate IAAis our belief in sharp human intuition. While the detection of bias through human eyes can be just noticeable, achieving the same task may become nontrivial when relying on computational approaches. Despite the potential of applying human intuition in adjusting the way DNNs think for debias, we noticed that rather a few approaches are categorized into this category. In this section, we explain what we have learned through our studies and what can be done to build a more generalizable, efficient, and effective design toward ***alignable DNNs*** where the fundamental goal is to make the way DNNs "think" aligning to our mental models using interactive tools.

6.1 Application Areas for Interactive Attention Alignment

Identifying when IAAcan be useful is crucial for practically adopting our framework in practice. One of the cases we believe IAAwould be useful is the case where model explanation made by DNNs can affect human reasoning in a variety of human-AI collaboration scenarios. In terms of studying contextual bias, we expect that studying IAAin different application areas than the gender classification can broaden the applicability as well as generalizability. One noteworthy application can be public safety. For example, GDXray [64] or SIXray [65] datasets present objects that can harm public safety, such as guns, knives, or hammers combined along with safe objects. Another example can be a car dataset that presents semantic segmentation. For example, A2D2 [31] presents non-sequential 41,277 frames with semantic segmentation image and point cloud labels. Mapillary Vistas dataset [69] presents 25,000 high-resolution images annotated into 66 object categories with additional, instance-specific labels for 37 classes. Meanwhile, the xBD dataset [35] presents the multi-class, multi-level building damage using satellite images where the damage adjuster can take a look at images with AI, and make a collaborative decision.

Another promising application area that we plan to investigate is to apply IAAin adversarial bias to help data scientists for building more robust DNNs. The adversarial attack is a well-known

problem in ML where an attacker can apply the adversarial patch to an image to flip DNN's classification results. Recent studies started to identify the role of model attention [10] in handling adversarial robustness in image classifiers.

Finally, while we limited the influence of contextual bias in our fine-tuning process, the IAAcan work reversely to increase the impact of contextual bias in different application areas. For example, consider we are training a DNN to classify human emotion. The DNN is seeing a blunt-faced lady standing on the podium and receiving a gold medal in an international sports event [55]. In such a case, merely focusing on a human face may not result in an accurate prediction. Contextual objects become indispensable assets in such cases, which is an opposite setting to "de-bias" at least based on our settings. Fundamentally, IAA's major goal is to define a human-usable interaction mechanism that can help humans to easily connect to DNNs and present their perspectives (i.e., what types of attention to augment or dwindle) in steering DNNs.

6.2 Topics in Front-end: Interactive Tools for Data Annotation and DNN Steering

Our studies open a variety of open questions related to the design of interactive user interfaces. The first interface problem is to understand how to efficiently and effectively elicit annotations from humans. More specifically, the current cost for eliciting human annotations for establishing a reasonability and attention boundary is expensive. For instance, we expected that the cost for eliciting reasonability questions to be lightweight, but every participant spent more than a day to complete their answers in our studies. High cost of human elicitation makes IAAless scalable. The design goal of capturing the reasonability aspect from human annotators may have a different design goal depending on the nature of the goal.

When the task can be seen as objective and the ground truth exists, the data annotation interface may seek for improving the human annotators' task efficiency—i.e., how fast an annotator can complete the annotation task and task effectiveness—i.e., how accurately an annotator can complete the annotation task. Ordering of the new instances, grouping them as a batch when presenting, or showing information relevant for supporting an annotator's decision (e.g., attention), and intelligent object selection suggestion [18, 22] could be considered an effective design manipulation.

When the task contains a certain level of subjectivity, capturing human annotators' diverse aspects can be a research challenge. When developing questions to understand the reasonability for a subjective task, we suggest breaking down a single question of "Is the gaze given to the image reasonable?" into more specific, possibly multiple questions (e.g., Q1 and Q2 in section 3). In that way, we can expect high-quality elicitation that could lead to better attention adjustment. After capturing the diverse aspects from several human annotators, modeling and eliciting that diversity in fine-tuning will be also a crucial design challenge.

The second problem is to understand how we can effectively support data scientists and ML engineers who would aim at eliciting human viewpoints and updating the model's future behavior based on the new elicitation. The goal of this research direction is to develop a centralized tool that data scientists can use features to define the directions for how to steer. We believe that even with the collection of new attention boundaries, the way to apply the inputs elicited from human annotators in steering DNNs can be highly diverse. While we added "Interactive" in our framework of Interactive Attention Alignment, we didn't scope the "interactive" design that data scientists can use in applying our framework. Providing an analytic system that helps data scientists to reason about the direction of elicitation and applying elicitation in fine-tuning the DNNs would increase the accessibility of IAAand similar approaches to a broader audience.

6.3 Topics in Back-end: Improving GRADIAand Beyond Images

Fine-tuning DNNs: Based on our experimental study, GRADIAshows the most desirable results from both computers and humans. When designing fine-tuning mechanism, we suggest using our technique. One crucial shortcoming in our approach is its interactivity. Although we named our framework "Interactive" Attention Alignment, the way we used our framework was not interactive, as a fine-tuning process can take multiple hours. Current design cannot show the effect of adjusted attention in real-time, even with the help of state-of-the-art techniques. Even though fine-tuning is a computationally expensive procedure that requires some time, improvement of round-trip speed should one of the important directions for making IAAto be a more viable solution.

Align with this aspect, the current approach is not presented in a visual analytic platform which could introduce some gaps to users who are not familiar with terminal-based interfaces and hidden settings. In the future, we hope there can be commitment towards developing visual analytic tools that a user can easily specify the split ratio, intelligently browse reasonable/unreasonable instances, perform batch adjustment, and see the aftermath of adjustment at a faster round trip speed (ideally in an interactive manner).

Beyond Convolutional Neural Network (CNN) and Image Data: Lastly, our approach has been built based on CNNs which operate on image data. Since the IAA framework can be general and model-agnostic, we believe the proposed framework can also be easily extended to other data types, such as tabular data, text data, and graph-structured data, with the corresponding DNN models. For example, one recent work proposed a very similar approach as IAA to steer the explanation of Graph Convolutional Networks (GCN) [54] on graph data (such as sense graphs and molecule graphs) [29]. The authors found that such an IAA-like framework can also effectively improve the reasonability of the model explanation for graph data and still keep or even improve the backbone GCN model performance. In all, we believe the direction of study on IAA is promising, and it will be beneficial to help DNNs to better align their attention/explanation with humans and consequentially enhance our understanding of the machine learning models as a whole.

7 CONCLUSION

We observe several side-effects behind the DNNs' powerful automation as a form of "bias" every day. We are directly or indirectly influenced by the AI's decisions affected by automated racism, gender bias, lack of considering people with a neurodiverse spectrum, insecurities on adversarial attacks, and many more. CSCW, HCI, and broader ML communities have invested substantial effort into devising straightforward and human-usable solutions for effectively aligning DNNs' behavior with our norms and expectation. However, several empirical studies revealed that steering DNNs as we intended is highly challenging not only for domain experts but also for skilled data scientists.

The overarching motivation behind this work is to devise a human-usable interaction modality that a human can directly see how DNNs think and intuitively modify the cases when needed. To do so, this work aimed at laying the groundwork for establishing a platform that can use IAAto more directly infuse their perspectives in fine-tuning DNNs. As a closing remark, we hope this work can motivate future research in IAAon DNN and more generally devising novel interaction modalities that can realize DNNs that better align with a human mental model.

8 ACKNOWLEDGMENTS

This work was supported by NSF Future of Work grant No. 2026513. We appreciate the additional generous support from NSF Grant No. 1755850, No. 1841520, No. 2007716, No. 2007976, No. 1942594, No. 1907805, a Jeffress Memorial Trust Award, Amazon Research Award, NVIDIA GPU Grant, and Design Knowledge Company (subcontract number: 10827.002.120.04).

REFERENCES

[1] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1418–1426.

[2] Bilal Alsallakh, Allan Hanbury, Helwig Hauser, Silvia Miksch, and Andreas Rauber. 2014. Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1703–1712.

[3] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[4] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[5] Guangji Bai and Liang Zhao. 2022. Saliency-regularized Deep Multi-task Learning. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[6] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*. 1937–1948.

[7] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To" See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–31.

[8] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.

[9] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. 2017. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 152–162.

[10] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. 2020. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*. PMLR, 1014–1023.

[11] Matthew Britton. 2019. VINE: Visualizing Statistical Interactions in Black Box Models. *arXiv preprint arXiv:1904.00561* (2019).

[12] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. 2015. FeatureInsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 105–112.

[13] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*. Springer, 793–811.

[14] 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.

[15] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[16] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[18] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. AILA: Attentive Interactive Labeling Assistant for Document Classification through Attention-Based Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 230, 12 pages. https://doi.org/10.1145/3290605.3300460

[19] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. 2010. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 27–34.

[20] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[21] Chaeyeon Chung, Jungsoo Lee, Kyungmin Park, Junsoo Lee, Minjae Kim, Mookyung Song, Yeonwoo Kim, Jaegul Choo, and Sungsoo Ray Hong. 2021. Understanding Human-Side Impact of Sampling Image Batches in Subjective Attribute Labeling. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 296 (oct 2021), 26 pages. https://doi.org/10.1145/3476037

[22] Chaeyeon Chung, Jung Soo Lee, Kyungmin Park, Junsoo Lee, Jaegul Choo, and Sungsoo Ray Hong. 2021. Understanding Human-side Impact of Sequencing Images in Batch Labeling for Subjective Tasks. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2021).

[23] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[24] D. Dingen, M. van't Veer, P. Houthuizen, E. H. J. Mestrom, E. H. H. M. Korsten, A. R. A. Bouwman, and J. van Wijk. 2019. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 246–255. https://doi.org/10.1109/TVCG.2018.2865043

[25] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[26] Alice Feng and Shuyan Wu. 2019 (accessed August 23, 2020). *The Myth of the Impartial Machine.* https://parametric.press/issue-01/the-myth-of-the-impartial-machine/

[27] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. [n. d.]. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10705–10714.

[28] Yuyang Gao, Giorgio A Ascoli, and Liang Zhao. 2021. BEAN: Interpretable and efficient learning with biologically-enhanced artificial neuronal assembly regularization. *Frontiers in Neurorobotics* 15 (2021), 68.

[29] Yuyang Gao, Tong Sun, Rishab Bhatt, Dazhou Yu, Sungsoo Hong, and Liang Zhao. 2021. GNES: Learning to Explain Graph Neural Networks. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 131–140.

[30] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Liang Zhao. 2022. RES: A Robust Framework for Guiding Visual Explanation. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[31] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. 2020. A2D2: Audi Autonomous Driving Dataset. (2020). arXiv:2004.06320 [cs.CV] https://www.a2d2.audi

[32] Yolanda Gil and Bart Selman. 2019. A 20-Year Community Roadmap for Artificial Intelligence Research in the US. *arXiv preprint arXiv:1908.02624* (2019).

[33] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)* 51, 5 (2019), 93. https://doi.org/10.1145/3236009

[34] Xiaojie Guo, Liang Zhao, Zhao Qin, Lingfei Wu, Amarda Shehu, and Yanfang Ye. 2020. Interpretable deep graph generation with node-edge co-disentanglement. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1697–1707.

[35] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 10–17.

[36] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[37] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[38] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[40] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* (2018). https://doi.org/10.1109/TVCG.2018.2843369

[41] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, III, Miro Dudik, and Hanna Wallach. [n. d.]. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 600:1–600:16. https://doi.org/10.1145/3290605.3300830

[42] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 1–26.

[43] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. *arXiv preprint arXiv:1603.06318* (2016).

[44] Alon Jacovi and Yoav Goldberg. 2020. Aligning Faithful Interpretations with their Social Attribution. *arXiv preprint arXiv:2006.01067* (2020).

[45] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau. 2018. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 88–97. https://doi.org/10.1109/TVCG.2017.2744718

[46] Minsuk Kahng, Dezhi Fang, and Duen Horng Chau. 2016. Visual exploration of machine learning results using data cube analysis. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–6.

[47] Minsuk Kahng, Nikhil Thorat, Duen Horng Polo Chau, Fernanda B Viégas, and Martin Wattenberg. 2018. Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 1–11.

[48] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 1–6.

[49] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.

[50] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. 2021. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific data* 8, 1 (2021), 1–18.

[51] John Kendall. 2018 (accessed August 23, 2020). *How to Boost Border Security While Protecting Privacy*. https://www.nextgov.com/ideas/2018/05/how-boost-border-security-while-protecting-privacy/148288/

[52] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9012–9020.

[53] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[54] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[55] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. EMOTIC: Emotions in Context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 61–69.

[56] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5686–5697. https://doi.org/10.1145/2858036.2858529

[57] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Thirty-first aaai conference on artificial intelligence*.

[58] Petro Liashchynskyi and Pavlo Liashchynskyi. 2019. Grid search, random search, genetic algorithm: A big comparison for NAS. *arXiv preprint arXiv:1912.06059* (2019).

[59] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[60] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards hybrid human-AI workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*. 2432–2442.

[61] Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286* (2019).

[62] Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, and Shixia Liu. 2017. Analyzing the training processes of deep generative models. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 77–87.

[63] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. 2016. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 91–100.

[64] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. 2015. GDXray: The database of X-ray images for nondestructive testing. *Journal of Nondestructive Evaluation* 34, 4 (2015), 1–12.

[65] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. 2019. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2119–2128.

[66] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. 2017. Understanding hidden memories of recurrent neural networks. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 13–24.

[67] Yao Ming, Huamin Qu, and Enrico Bertini. 2018. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

[68] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Embedding Human Knowledge into Deep Neural Network via Attention Map. *arXiv preprint*

arXiv:1905.03540 (2019).

[69] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*. 4990–4999.

[70] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. *arXiv preprint arXiv:1809.07424* (2018).

[71] Badri Patro, Vinay Namboodiri, et al. 2020. Explanation vs attention: A two-player game to obtain attention for vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11848–11855.

[72] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.

[73] Nicola Pezzotti, Thomas Höllt, Jan Van Gemert, Boudewijn PF Lelieveldt, Elmar Eisemann, and Anna Vilanova. 2017. Deepeyes: Progressive visual analytics for designing deep neural networks. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 98–108.

[74] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[75] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8227–8236.

[76] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 61–70. https://doi.org/10.1109/TVCG.2016.2598828

[77] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717* (2017).

[78] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*. PMLR, 8346–8356.

[79] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[80] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[81] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. 2017. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 667–676.

[82] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. 2009. EnsembleMatrix: Interactive Visualization to Support Machine Learning with Multiple Classifiers. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'09)*. ACM, 1283–1292. https://doi.org/10.1145/1518701.1518895

[83] Tesla. 2020 (accessed August 23, 2020). *Tesla Autopilot*. https://www.tesla.com/autopilotAI

[84] Roman Visotsky, Yuval Atzmon, and Gal Chechik. 2019. Few-shot learning with per-sample rich supervision. *arXiv preprint arXiv:1906.03859* (2019).

[85] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2022. Follow My Eye: Using Gaze to Supervise Computer-Aided Diagnosis. *IEEE Transactions on Medical Imaging* (2022).

[86] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*. 5310–5319.

[87] Gary M Weiss, Kate McCarthy, and Bibi Zabar. 2007. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? *Dmin* 7, 35-41 (2007), 24.

[88] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mane, Doug Fritz, Dilip Krishnan, Fernanda B Viégas, and Martin Wattenberg. 2017. Visualizing Dataflow Graphs of Deep learning Models in Tensorflow. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 1–12. https://doi.org/10.1109/TVCG.2017.2744878

[89] Chuan Yan, John Joon Young Chung, Kiheon Yoon, Yotam Gingold, Eytan Adar, and Sungsoo Ray Hong. [n. d.]. FlatMagic: Improving Flat Colorization through AI-driven Design for Digital Comic Professionals.

[90] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[91] Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2016. NIH Public Access, 795.

[92] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).
[93] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.
[94] Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review* 5, 1 (2018), 44–53.
[95] Hongzhi Zhu, Septimiu Salcudean, and Robert Rohling. 2022. Gaze-Guided Class Activation Mapping: Leveraging Human Attention for Network Attention in Chest X-rays Classification. *arXiv preprint arXiv:2202.07107* (2022).
[96] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).

Received January 2022; revised April 2022; accepted August 2022