Shared Interest: Large-Scale Visual Analysis of Model Behavior by Measuring Human-AI Alignment

Angie Boggust [1] **Benjamin Hoover** [2] **Arvind Satyanarayan** [1] **Hendrik Strobelt** [2 1]

Abstract

Saliency methods — techniques to identify the importance of input features on a model's output — are a common first step in understanding neural network behavior. However, interpreting saliency requires tedious manual inspection to identify and aggregate patterns in model behavior, resulting in ad hoc or cherry-picked analysis. To address these concerns, we present Shared Interest: a set of metrics for comparing saliency with human-annotated ground truths. By providing quantitative descriptors, Shared Interest allows ranking, sorting, and aggregation of inputs thereby facilitating large-scale systematic analysis of model behavior. We use Shared Interest to identify eight recurring patterns in model behavior including focusing on a sufficient subset of ground truth features or being distracted by contextual features. Working with representative real-world users, we show how Shared Interest can be used to rapidly develop or lose trust in a model's reliability, uncover issues that are missed in manual analyses, and enable interactive probing of model behavior.

1. Introduction

As machine learning continues to be deployed in real-world applications, it is increasingly important to understand the reasoning behind model decisions. A common first step for doing so is to compute the model's *saliency*. We define saliency to be the output of any function that, given an input instance (e.g., an image), computes a score of how important each input feature (i.e., pixel) is for the model's output. Example saliency methods range from vanilla gradients (Simonyan et al., 2013; Erhan et al., 2009), where scores represent the amount a change in an input feature would have on the model's output, to black-box methods like LIME (Ribeiro et al., 2016) that use interpretable surrogate models trained to mimic the original model's decision boundary. By analyzing saliencies, people can identify important features for the model's decision and determine how aligned these features are with human decision-making.

While saliency methods provide much-needed introspection into model behavior, making sense of them can still present analysts with a non-trivial burden. In particular, saliencies are often visualized as solitary heatmaps, which do not provide any additional structure or higher-level visual abstractions to aid analysts in interpreting them. As a result, analysts are forced to rely solely on their visual perception and priors to generate hypotheses about model behavior. Similarly, by operating on individual instances, saliency methods make it difficult to conduct large-scale analyses of model behavior and uncover recurring patterns. As a result, analysts must choose between the time-consuming, often infeasible manual analysis of all instances, and the ad hoc, often biased selection of meaningful subsets of instances.

In response, we introduce Shared Interest: a method for comparing model saliencies with human-generated ground truth annotations. Shared Interest quantifies the alignment between these two components by measuring three types of coverage: Ground Truth Coverage (GTC), or how many of the ground truth features are important to the model's decision; Saliency Coverage (SC), or how many of the saliency features are found in the ground truth; and IoU Coverage (IoU), the similarity between the ground truth and saliency feature sets. These coverage scores are agnostic to model architecture, input modality, saliency method, and the mechanism by which ground truth annotations are provided. They also enable a richer and more structured interactive analysis process by allowing analysts to sort, rank, and aggregate input instances based on model behavior. The scores can also be composed together (e.g., a high SC and low GTC) to identify recurring patterns in alignment between model and human decision-making.

We demonstrate how Shared Interest enables structured large-scale analysis of model behavior across multiple domains and saliency methods. By applying it to computer vision and natural language classification and regression tasks, and using a variety of common saliency methods, we identify 8 recurring patterns of interesting model behav-

[1]CSAIL, MIT, Cambridge, Massachusetts, USA [2]IBM Research, Cambridge, Massachusetts, USA. Correspondence to: Angie Boggust <aboggust@mit.edu>.

iors or dataset features: HUMAN-ALIGNED, SUFFICIENT SUBSET, SUFFICIENT CONTEXT, CONTEXT DEPENDANT, CONFUSER, INSUFFICIENT SUBSET, DISTRACTOR, and CONTEXT CONFUSION. Through representative case studies of real-world interactive visual analysis workflows, we explore how Shared Interest helps a board-certified dermatologist and an ML researcher conduct more systematic analyses of model behavior. These users find that, in contrast to their prior approaches which require tedious ad hoc exploration, Shared Interest rapidly surfaces reasons to question a model's reliability, opportunities where a model's learned representations might further their understanding of their domain, and issues they had missed during prior manual analyses they had conducted.

We further demonstrate that Shared Interest is not only valuable to understanding a model's predictive performance, but can also be used to *query* model behavior. Leveraging the Shared Interest metrics alongside interactive human annotation enables a question-and-answer process where analysts probe input features and Shared Interest identifies model decisions whose saliency feature sets are most aligned. In an example human annotation workflow with an image classification task, we show how Shared Interest can reveal insights about the set of input features necessary and sufficient to trigger particular predictions as well as the model's understanding of secondary objects or background features.

2. Related Work

Machine learning systems are increasingly designed for high-stakes tasks such as cancer diagnosis, and as these systems achieve human-calibre or super-human accuracy (Esteva et al., 2017), the temptation to deploy them correspondingly increases. In tandem, a body of work has identified dangerous pitfalls in commonly used models and their underlying training data (Prabhu & Birhane, 2020; Carter et al., 2020). To protect against the repercussions of deploying biased or ungeneralizable models, there has been a growing effort to understand and interpret model decisions (Doshi-Velez & Kim, 2017; Rai, 2020). In this paper, we focus on post hoc saliency methods, also known as feature attribution methods (Sturmfels et al., 2020), that allow us to observe model reasoning (Simonyan et al., 2013; Erhan et al., 2009; Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017; Kapishnikov et al., 2019; Ribeiro et al., 2016; Carter et al., 2019a; Springenberg et al., 2014).

While saliency methods offer much needed introspection into deep learning models, they do so only at the instance-level. Providing one interpretation at a time may be sufficient to answer questions about model behavior for a small collection of instances, but it does not scale to answering questions about global model behavior or dataset characteristics. Moreover, the output saliency maps, require careful visual assessment to determine if the model's decisions were based on human-salient features. Together these drawbacks often result in the tedious inspection of only a few examples selected in an ad hoc or cherry-picked manner. By quantifying instances based on the agreement between model and human reasoning, Shared Interest offers a more a comprehensive overview of model behavior across all instances and enables systematic evaluation of model behavior.

A recent body of work has questioned whether saliency methods are a reliable instrument for interpreting ML models (Adebayo et al., 2018; Tomsett et al., 2020; Adebayo et al., 2020; Yang & Kim, 2019; Kindermans et al., 2019). These papers propose saliency "tests" to measure each method's ability to faithfully represent model behavior. While confirming the fidelity of saliency methods is a critical area of research, this is an orthogonal issue to the focus of our paper as even the most faithful saliency method will still exhibit the instance-wise limitations described above.

Closest to our contribution is Olah et al. (2018) which argues saliency methods perform attribution at the wrong level of abstraction — feature-level saliency is not semantically meaningful enough and hidden layer representations should be used instead. To combat the scalability limitations of instance-wise interpretation, they suggest decomposing activations through matrix factorization (Olah et al., 2018) or activation atlases (Carter et al., 2019b). Shared Interest shares its underlying motivation with their work — a lack of semantically-meaningful structure in saliency methods and supporting scalable interpretability — but offers an alternate way forward. In particular, although we compute attribution back to input features, we do so to compare salient features to human-provided ground truth. In doing so, Shared Interest brings structure and scale to the task of reading model saliencies and provides a more direct expression of the alignment between human and model reasoning.

Aside from saliency methods, a growing number of techniques have been developed to help users visually interpret models (Hohman et al., 2018; Wexler et al., 2019); however, these tools often focus on understanding patterns learned by intermediate nodes (Bau et al., 2017; Zeiler & Fergus, 2014; Hohman et al., 2019) or are architecture-specific (Kahng et al., 2018; Hoover et al., 2020; Strobelt et al., 2016). In contrast, Shared Interest is agnostic to the model architecture, saliency method, and dataset and can be incorporated into existing visual analytic workflows.

3. Shared Interest Method

Shared Interest is a method for computing the alignment between model saliencies and human-generated ground truth annotations. Mathematically, we use S to represent the set of input features important for a model's decision as

determined by a saliency method and G for the set of input features annotated as ground truth. For example, in a computer vision classification task, G might represent the pixels within an object-level bounding box and S might represent the set of pixels salient to the model's decision as determined by a saliency method. Similarly, in an NLP sentiment classification task, G might be the set of input tokens annotated as indicative of sentiment while S is the set of tokens determined to be important to the model's prediction.

We compute three metrics: IoU Coverage (IoU), Ground Truth Coverage (GTC), and Saliency Coverage (SC). They each take G and S as inputs and output a score between 0 and 1, inclusive.

$$\text{IoU} = \frac{|G \cap S|}{|G \cup S|} \tag{1}$$

$$\text{GTC} = \frac{|G \cap S|}{|G|} \tag{2}$$

$$\text{SC} = \frac{|G \cap S|}{|S|} \tag{3}$$

IoU (Eq. 1) is the strictest metric and represents the similarity between the ground truth and saliency feature sets. It is the number of features in both the ground truth and saliency sets divided by the number of features in at least one of the ground truth and saliency sets. In machine learning terms, it is analogous to the Jaccard index. GTC (Eq. 2) measures how strictly the model relies on *all* ground truth features — the proportion of the ground truth feature set, G, that is also part of the saliency feature set, S. It is analogous to concepts of recall or sensitivity in machine learning: the fraction of true positives (saliency features that are also ground truth features) successfully identified. SC (Eq. 3) measures how strictly the model relies on *only* ground truth features — the proportion of the saliency feature set, S, that is also part of the ground truth feature set, G. In machine learning terms, it is analogous to precision: the fraction of true positives (saliency features that are also ground truth features) successfully identified among all detected positives.

A score of zero under all three metrics means that an instance's saliency and ground truth feature sets are disjoint, which often indicates the model is relying on background information. A high IoU score indicates the explanation and ground truth feature sets are very similar ($\text{IoU} = 1 \implies S = G$), meaning the model is focused on the salient features. High GTC indicates the model is using most of the ground truth features to make its decision ($\text{GTC} = 1 \implies G \subseteq S$), and high SC indicates the model only relies almost exclusively on ground truth features to make its prediction ($\text{SC} = 1 \implies S \subseteq G$). Shared Interest metrics can also be combined to yield interesting insights. For example, instances with high SC and low GTC indicate the model is focused on a strict subset of the ground truth region, whereas high GTC and low SC indicate the model is relying on the ground truth and contextual features to make its prediction.

To employ set-based metrics, we define S and G to be discrete feature sets. As a result, Shared Interest can be straightforwardly applied to saliency methods like SIS (Carter et al., 2019a). However, to apply Shared Interest on methods that output a continuous score (e.g., Integrated Gradients (Sundararajan et al., 2017)), we compute S by discretizing these scores. We demonstrate Shared Interest is robust to discretization procedure by employing both score-based and model-based thresholding. Score-based thresholding, used in the computer vision examples, creates discrete feature sets using only the saliency. For example, we threshold vanilla gradients at one standard deviation above the mean to allow for variance in the number and value of salient features across instances, and we select LIME's top n positively contributing features to demonstrate even naive thresholding can be effective. Model-based thresholding, used in the NLP examples, creates discrete feature sets containing features directly correlated with the model's prediction. In these examples, we iteratively select features positively correlated with the model's prediction until the model can confidently predict the correct class using only those features.

4. Identifying Patterns in Model Behavior

We apply Shared Interest to models trained on computer vision (CV) and natural language processing (NLP) tasks, and identify eight recurring patterns in model behavior. Each pattern is defined as a function of the Shared Interest metrics and the correctness of the model's output[1]. We find these patterns occur in models regardless of domain or applied saliency method. In Figure 1, we show an example of each pattern in a CV setting (ImageNet classification with LIME saliencies) and an NLP setting (BeerAdvocate aroma sentiment prediction (McAuley et al., 2012; Lei et al., 2016; Carter et al., 2019a) with Integrated Gradients (Sundararajan et al., 2017)). These patterns rapidly surface dataset limitations (e.g., multiple, equally plausible labels) and suggest avenues for further exploration.

Human Aligned Instances that fall into the HUMAN ALIGNED category are predicted correctly and have high IoU, thus indicating that the model is making a correct prediction and its rationale for that prediction aligns with a person's. For example, in the CV setting of Figure 1a, the model relies on almost every pixel in the ground truth bounding box to make the correct prediction of *trailer truck*. HUMAN ALIGNED instances indicate cases when the model

[1]In regression tasks, we define correctness as whether the model's output is within $\pm\Delta$ of the true value ($\Delta = 0.05$ in the NLP examples)

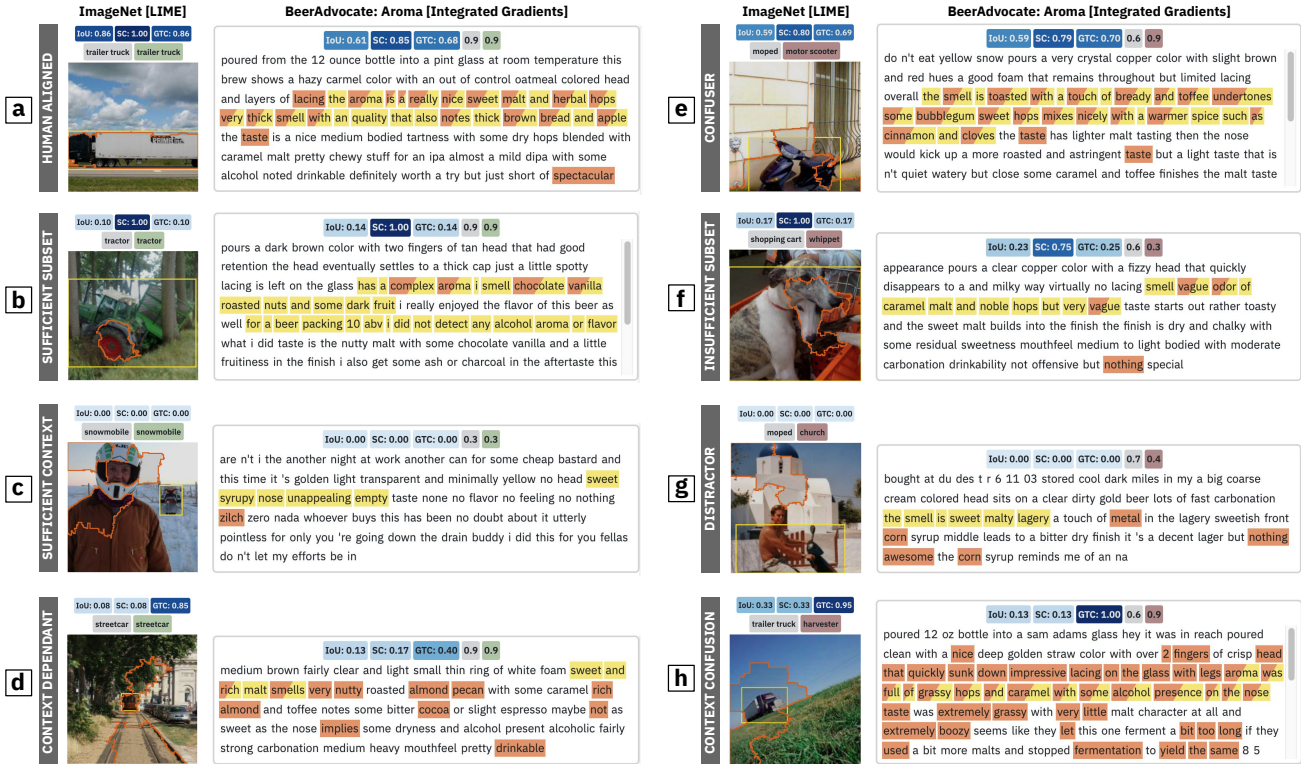

Figure 1. Using Shared Interest we identify eight patterns in model behavior across two domains and saliency methods: an ImageNet (Deng et al., 2009) classification model with LIME (Ribeiro et al., 2016), and a BeerAdvocate sentiment regression model (Carter et al., 2019a; McAuley et al., 2012; Lei et al., 2016) with Integrated Gradients (Sundararajan et al., 2017). The IoU, GTC, SC, label, prediction, ground truth features (yellow), and saliency features (orange) are shown for each example.

is faithful to human decisions, and ideally, all instances would fall into this category.

Sufficient Subset The SUFFICIENT SUBSET category contains instances with high SC and low GTC, revealing where the model relies on a subset of the human-annotated features to make a correct prediction. For example, in the NLP case in Figure 1b, the salient regions indicate the model considers just the words "complex", "aroma", "chocolate", and "vanilla" sufficient to predict strong positive sentiment.

Sufficient Context The SUFFICIENT CONTEXT category includes correctly predicted instances with low IoU, indicating there is sufficient information in non-ground truth features to make the correct prediction. Analyzing instances in this category can validate if contextual features are indeed meaningful and not the result of spurious correlation. The CV example in Figure 1c shows the model uses the helmet to predict *snowmobile*. While the presence of a snowmobile helmet is correlated with the existence of a snowmobile, this model would not be robust to real-life scenarios. This result might inspire further exploration of instances of snowmobiles and snowmobile helmets to confirm there is not a rigid dependence between the two objects.

Context Dependant The CONTEXT DEPENDANT category identifies correctly classified instances with high GTC and low SC, meaning the model relies on ground truth and contextual features to make a correct prediction. In Figure 1d, the CV example shows the model relies not only on the streetcar (the labeled class) but also on the train tracks to predict *streetcar*. While the context is semantically correlated with the ground truth, these instances may indicate nongeneralizable correlations and may require further exploration to uncover whether it is reasonable for the model to use context in its prediction.

Confuser Confusers are instances where the model relies on human-salient features but still makes an incorrect prediction. In Shared Interest terms, members of the CONFUSER case are incorrectly classified instances with high IoU. In the CV example in Figure 1e, the CONFUSER case identifies an ambiguous label — the image is labeled as *moped*, but the model predicts *motor scooter* — a known problem with ImageNet (Beyer et al., 2020; Tsipras et al., 2020). Using Shared Interest, instances of this failure case immediately rise to the forefront of the analysis process, encouraging further exploratory analysis on the dataset or additional preprocessing to resolve ambiguities.

Insufficient Subset INSUFFICIENT SUBSET identifies incorrectly classified instances with high SC and low GTC, meaning the model uses a subset of the ground truth features to make its prediction, but this subset is not sufficient to make a correct prediction. In the NLP example in Figure 1f, the saliency method indicates the model relies upon the words "vague" and "odor" in the aroma sentence to predict negative sentiment (0.3). However, other words in the sentence not used by the model do contain positive sentiment (i.e., "caramel malt" and "noble hops") and likely contributed to the true label of 0.6.

Distractor Distractors are instances where the model does not rely on ground truth features (low IoU) and makes an incorrect prediction. In the NLP example in Figure 1g, the saliency method indicates the model relies on the words "metal", "corn", and "nothing awesome" to predict negative sentiment. While the known aroma words ("the smell is sweet malty lagery") have positive sentiment, the model is distracted by negative sentiment words elsewhere in the review. Such instances may indicate the model is overfitting to the overall sentiment of the review, rather than the specific sentiment associated with the aroma.

Context Confusion The CONTEXT CONFUSION case contains instances where the model is using ground truth features, but is confused by other features and, thus, makes an incorrect prediction. In Shared Interest terms, these instances have high GTC and low SC. For example, in the CV setting in Figure 1h, the saliency indicates the presence of the field next to the trailer truck caused the model to predict *harvester*.

5. Interactive Visual Analysis Workflows

We demonstrate how Shared Interest can be used for real-world analysis through case studies of three interactive visual analysis workflows of ML models. The first case study follows a domain expert (a dermatologist) using Shared Interest to determine the trustworthiness of a melanoma prediction model. The second case study follows a machine learning expert analyzing the faithfulness of their model and saliency method. The final case study examines how Shared Interest can be used to analyze model behavior even without pre-existing ground truth annotations.

5.1. Model Analysis by a Domain Expert

Our first case study follows a use case of a domain expert, a dermatologist, who wishes to evaluate the trustworthiness of an ML model that could assist them in diagnosing melanoma. Accurately diagnosing melanoma early is a critical task that can have a large impact on patient outcome, and ML models could assist dermatologists in making more accurate decisions. In order to do so, however, our participant noted it would be extremely important for dermatologists to be able to personally evaluate how the model operates.

We evaluate Shared Interest in this context to understand how its ability to convey model behavior may help a domain expert determine whether or not they should trust a model. To do so, we applied Shared Interest to a Convolutional Neural Network trained on the ISIC Melanoma dataset (Codella et al., 2019; Tschandl et al., 2018) to classify images of lesions as either *malignant* (cancerous) or *benign*. We used lesion segmentations from the dataset as the ground truth feature set and the output of LIME (Ribeiro et al., 2016) towards the predicted class as the saliency feature set. We display the results in a prototype visual interface (Figure 2) designed to enable interactive analysis of Shared Interest.

Using the HUMAN ALIGNED, CONTEXT DEPENDANT, and SUFFICIENT SUBSET categories, the dermatologist surfaced insight into cases where the model was trustworthy. Analyzing *malignant* lesions in the HUMAN ALIGNED case surfaced examples where the model correctly classified cancerous lesions by relying on features of lesion. The dermatologist agreed with the model on these images and began to build trust with the model, noting *"obviously it does a pretty good job"*. CONTEXT DEPENDANT images identified cases where the model relied not only on the lesion but also on surrounding skin. While there was potential for the dermatologist to distrust the model, they actually found these instances especially interesting because cancerous cells can lie beyond the pigmented lesion boundary. Thus, the dermatologist wondered if *"there's really subtle changes that we're not picking up that [the model] is able to."* Images in the SUFFICIENT SUBSET case showed cases where the model only relied on a subset of the lesion. While the dermatologist agreed with the model, they expressed some concern that it was not using the complete lesion, especially when there were meaningful cancerous features in the unused regions.

Shared Interest was also able to quickly reveal cases where the model was not trustworthy. The SUFFICIENT CONTEXT and DISTRACTOR cases showed images where the model relied on contextual features such as peripheral skin regions or the presence of artifacts (see Figure 2). While the dermatologist was tolerant to a few instances where the model relied on non-salient features, seeing the number of images in these cases led the dermatologist to distrust the model in all cases, stating *"I would discard the model."*

By classifying inputs into cases where the model was or, more importantly, was not aligned with human reasoning, Shared Interest enabled the dermatologist to rapidly and confidently decide whether or not to trust the model. If the dermatologist evaluated the model by randomly selecting images, they might not have identified that the model repeatedly made decisions based on background informa-

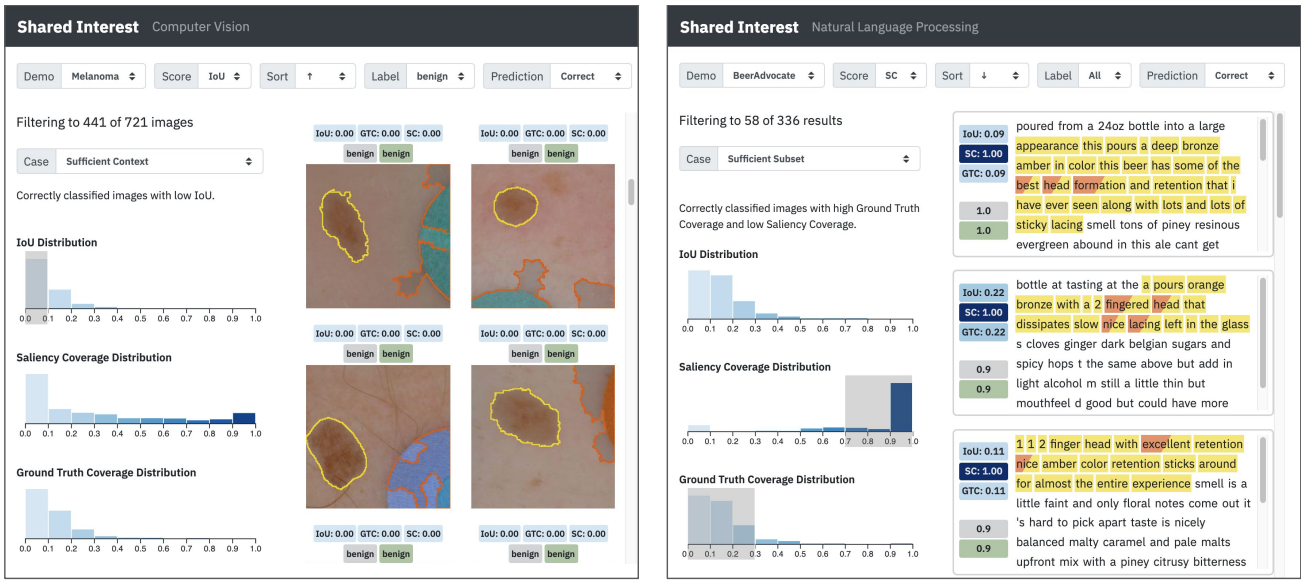

Figure 2. Shared Interest can assist a range of users perform large-scale analysis of machine learning models. We evaluate Shared Interest with a dermatologist using a melanoma prediction model and LIME (left) and with a machine learning expert using a sentiment analysis model and SIS (right). The prototypes display instances with ground truth features (yellow), saliency features (orange), true labels, model predictions, and Shared Interest scores. The SUFFICIENT CONTEXT case (left) surfaces images where the model has latched onto artifacts to make a *benign* prediction. Since these artifacts only occur in *benign* dataset images, the are sufficient to make a prediction; however, this model would not generalize to clinical cases where the artifacts may be included in *malignant* images. The SUFFICIENT SUBSET case (right) identifies reviews where the saliency method indicates the model relied on meaningful features, such as "best head formation", and where it overfits to general positive sentiment words such as "excellent". The demos are available at: `http://www.shared-interest.csail.mit.edu/computer-vision/` and `http://www.shared-interest.csail.mit.edu/nlp/`.

tion, and they would not have known how frequently that case occurred. As the dermatologist said, Shared Interest is *"helpful [as a way to] see how the computer is thinking and allow me to understand if I should trust it."*

5.2. Saliency Method Analysis by a ML Researcher

Our second case study is representative of use cases where an ML expert wants to analyze a model or saliency method they are developing. To evaluate Shared Interest's value in the ML development pipeline, we worked with an author of the Sufficient Input Subset (SIS) interpretability method (Carter et al., 2019a) whose goal is to understand how well SIS explains model decisions. During development of the SIS method, one of the ways the researchers analyzed the method was by applying it to the BeerAdvocate dataset and comparing the SIS saliencies (called "rationales" by the researchers) to the ground truth annotations. This process enabled them to evaluate whether the rationales *"fell within the ground truth"* and represented a *"compact set"* of meaningful features.

To recreate the researcher's original workflow, we applied Shared Interest to the BeerAdvocate reviews annotated on the appearance aspect, trained Recurrent Neural Network model, and SIS rationales from Carter et al. (2019a). We populate our visual prototype with the results (Figure 2).

Using Shared Interest, the researcher surfaced a number of insights that inspired confidence in the SIS algorithm. For example, the researcher immediately identified that most reviews have high SC, indicating most of the SIS rationales were contained almost entirely within the ground truth. Since *"ideally, the model is learning the right set of features and thus the rationales live within the correct set of features"*, the researcher found the distribution of scores indicative that the SIS procedure was capturing meaningful information. In their original analysis, the researchers had even computed a metric equivalent to SC as a quantitative way to analyze their method. So, seeing the same metric populated by Shared Interest validated the use of Shared Interest and the SUFFICIENT SUBSET and INSUFFICIENT SUBSET categories. The researcher found it reassuring to find HUMAN ALIGNED and SUFFICIENT SUBSET instances that matched their expectations, such as rationales that contained appearance-specific words (i.e., "red", "copper", and "head"), but did not include uninformative ground truth words like stop words. The SUFFICIENT SUBSET category was particularly meaningful to the researcher since it aligned with SIS's goal to find minimal rationales. Seeing all of these examples at once helped the researcher identify cases where the rationale was truly a meaningful sufficient

subset of words such as "lovely looking".

Shared Interest also helped the researcher uncover previously unknown pitfalls in the model and the data. Looking at instances in the SUFFICIENT SUBSET category, the researcher identified common cases of model overfitting such as a correct prediction using only the word "beautiful". As the researcher put it, *"These are positive words, so it makes sense they are correlated with positive appearance, but I don't think they should be sufficient for separating [the appearance] aspect from others."* Looking at reviews in the INSUFFICIENT CONTEXT case, exposed instances where the model was again overfitting to positive sentiment, but now the researcher was even more concerned since it caused an incorrect prediction. Although the researcher had *"previously observed that the model had associated single tokens that were general positive sentiment words with predicting high sentiment"*, they did not *"as quickly notice particular words like 'beautiful' that were immediately surfaced [via Shared Interest]."* Finally, looking at SUFFICIENT CONTEXT reviews, where SIS rationales are disjoint from the ground truth annotation, the researcher uncovered reviews with incomplete or incorrect annotations. Until using Shared Interest, the researcher had previously never identified an incorrectly annotated review, saying *"in the past, I did not note any cases where I thought the annotations might have been incomplete. I think that's a pretty interesting insight."*

Overall, the researcher found that grouping and aggregating via Shared Interest helped them *"see all of the [reviews] grouped together by the various cases"* which categorized and *"clearly described what the various patterns are"*. In the researchers' original analysis they had *"skimmed through a big file of [reviews] not sorted in anyway"*, and, while they *"were noting patterns, it was harder to keep track of these different cases."* If they would have had access to Shared Interest at the time of their original analysis, this researcher thought it would have *"more quickly exposed some of the patterns and behaviors that we identified and also led to additional discoveries."*

5.3. Interactive Probing of Model Behavior

For our final case study we demonstrate a workflow where Shared Interest can be used as a mechanism to *query* model behavior. For any given input instance, rather than computing the saliency for only the predicted class, we instead do so for all possible classes. Moreover, rather than relying on pre-existing ground truth, users can interactively annotate the instance to designate a "ground truth" region of interest. By calculating all three Shared Interest metrics between these two sets of features, and returning classes with the highest Shared Interest scores, we enable a user to engage in a style of "what if" reasoning: interactively probing the model to understand what the necessary and sufficient set of input features are to trigger particular predicted classes.

In Figure 3, we show an example of this style of "what if" analysis on an ImageNet classification model. By interactively re-specifing the "ground truth" on a single image, we repeatedly probe the model and surface insights about its behavior. Since the model was trained to predict the *otterhound* in the image, we can use Shared Interest to validate that the model has indeed learned the salient features of the dog. By selecting the pixels associated with the dog's face and body (Figure 3a), we find that although none of the top 3 returned classes are *otterhound*, they are all dog breeds and the salient feature sets are focused primarily on the dog. This result may suggest that the model has learned generalizable features associated with dogs — a positive characteristic if we plan to deploy this model.

Since the model learned to associate the entire dog region with dog classes, this prompts a follow-up question: how much of the dog do we have to annotate before the model no longer associates it with dog breed classes? Brushing over just the dog's head (Figure 3b) or even just the dog's snout (Figure 3c) still return dog breeds as the top classes, suggesting this model has learned to correlate even small characteristic features (e.g., black noses) with dog breeds.

This style of analysis also enables us to ask questions about other objects in the image. Although the model was trained to classify this image as *otterhound*, it was also trained to classify 1,000 ImageNet objects. Thus, the model may know salient information about other objects in the image as well. In Figure 3d we validate this claim by brushing the person's hat and observing the top returned classes are types of hats: *sombrero*, *cowboy hat*, and *bonnet*. Similarly, we select the person's hand (Figure 3e) and, as *hand* is not a class our model was trained to detect, observe classes associated with hands such as *cleaver*, *notebook*, and *space bar*. This result is interesting as a hand is often, but not always, present in images of these objects. Thus, further analysis is warranted to determine if the model is overly-reliant on the presence of hands to make predictions for these classes.

We can also probe the model to see if it has learned anything about image backgrounds or textures, despite only being trained on foreground objects. In Figure 3f, we select a region of the stone wall. Interestingly, the model returns classes associated with rocks such as *cliff*, suggesting that training on images with foreground labels may still impart information to the model about background scenes.

As we have seen, Shared Interest allows us to probe model behavior in new ways, enabling exploration into what the model has learned and where it might fail. Users can identify subsets of features important to classification, explore how well a model can identify secondary objects, and even the extent to which a model has learned about objects it has

Figure 3. Shared Interest enables users to probe the model with different instance features to understand model behavior. By comparing the user's annotation (yellow) to the saliency feature set (orange) for every ImageNet class, Shared Interest surfaces the classes most related to the annotated features. Probing with smaller and smaller sets of features (a-c) shows the model has learned characteristic features of dogs. Probing with secondary objects (d) demonstrates the model learns about objects even when they are not the main label. Probing the model with features it has not learned to classify (e) indicates the model learns to related these features to associated objects (e.g., hand and cleaver). Finally, probing with background features (f) demonstrates the model has learned related features despite only being trained on foreground classification. This demo is available at: http://www.shared-interest.csail.mit.edu/human-annotation/

never classified. Using this procedure can help a user test hypotheses about what the model has learned and identify information that could help them improve model behavior.

6. Conclusion

In this paper, we present Shared Interest, a method for large-scale visual analysis of machine learning (ML) model behavior via metrics that quantify instances based on the model's alignment with human reasoning. Shared Interest enables instances to be sorted, ranked, and aggregated based on this alignment, and allows us to identify eight patterns in model behavior that recur across multiple domains (computer vision and natural language processing), model architectures (convolutional and recurrent neural networks), and saliency methods (gradient based and black-box). These patterns range from incorrectly classified instances despite the model relying on ground truth features (CONFUSER) to correctly classified instances where the model does not rely on a single ground truth feature (SUFFICIENT CONTEXT). And, through three case studies, we demonstrate how Shared Interest helped representative real-world more systematically explore model behavior, and how it can be used to interactively probe and query model behavior.

While the Shared Interest methodology can help users efficiently and comprehensively understand model behavior, it requires data paired with ground truth annotations. Research datasets, such as the ones used in this paper, may include such annotations, but real world data rarely does due to the time and effort required in the collection process. While this limits Shared Interest's applicability, we believe that understanding model behavior is critical enough to warrant the collection of human annotations. This may range from annotating just a few instances (i.e., via the model probing interface) for general research analysis to annotating entire datasets when deciding to deploy a model on a critical task.

Shared Interest opens the door to several promising future work directions. One straightforward path is to apply Shared Interest to tabular data. Since tabular data is a common format for sensitive data (e.g., health data), our interactive probing prototype could be used to systematically analyze if a model perpetuates bias in the data. Future work may also consider using Shared Interest to compare the fidelity of different saliency methods. Using Shared Interest metrics during saliency method faithfulness experiments (Adebayo et al., 2018) may help distinguish whether saliency map sensitivity represents semantically-meaningful signal. Finally, Shared Interest could be used during training to identify the most challenging instances for the model. These insights could then be used to inform future training procedures or augment the dataset with more informative examples.

Acknowledgments

This work was supported by MIT-IBM Watson AI project "Towards Intuitive AI" and the first author was supported by the John W. Jarve (1978) Fellowship.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Carter, B., Mueller, J., Jain, S., and Gifford, D. What made you do this? understanding black-box decisions with sufficient input subsets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 567–576. PMLR, 2019a.

Carter, B., Jain, S., Mueller, J., and Gifford, D. Overinterpretation reveals image classification model pathologies. *arXiv preprint arXiv:2003.08907*, 2020.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., and Olah, C. Activation atlas. *Distill*, 4(3):e15, 2019b.

Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S. W., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. A., Kittler, H., and Halpern, A. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1902.03368, 2019. URL http://arxiv.org/abs/1902.03368.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning, 2017.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hohman, F., Kahng, M., Pienta, R., and Chau, D. H. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics*, 25(8):2674–2693, 2018.

Hohman, F., Park, H., Robinson, C., and Chau, D. H. P. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26 (1):1096–1106, 2019.

Hoover, B., Strobelt, H., and Gehrmann, S. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 187–196, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.22. URL https://www.aclweb.org/anthology/2020.acl-demos.22.

Kahng, M., Thorat, N., Chau, D. H., Viégas, F. B., and Wattenberg, M. Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE transactions on visualization and computer graphics*, 25(1):310–320, 2018.

Kapishnikov, A., Bolukbasi, T., Viégas, F., and Terry, M. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4948–4957, 2019.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lei, T., Barzilay, R., and Jaakkola, T. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

McAuley, J., Leskovec, J., and Jurafsky, D. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pp. 1020–1025. IEEE, 2012.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Prabhu, V. U. and Birhane, A. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

Rai, A. Explainable ai: from black box to glass box. *Journal of the Academy of Marketing Science*, 48, 01 2020. doi: 10.1007/s11747-019-00710-5. URL https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Strobelt, H., Gehrmann, S., Huber, B., Pfister, H., and Rush, A. M. Visual analysis of hidden state dynamics in recurrent neural networks. *CoRR*, abs/1606.07461, 2016. URL http://arxiv.org/abs/1606.07461.

Sturmfels, P., Lundberg, S., and Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6021–6029, 2020.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018. URL https://doi.org/10.1038/sdata.2018.161.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From imagenet to image classification: Contextualizing progress on benchmarks, 2020.

Vedaldi, A. and Soatto, S. Quick shift and kernel methods for mode seeking. In *European conference on computer vision*, pp. 705–718. Springer, 2008.

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

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994*, 2020.

Yang, M. and Kim, B. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Supplementary Information
Shared Interest: Large-Scale Visual Analysis of Model Behavior by Measuring Human-AI Alignment

S1. Live Demos

Live demos of the computer vision interface (§ 5.1, Figure S1), natural language processing interface (§ 5.2, Figure S2), and human annotation interface (§ 5.3, Figure S3) are available:

CV: http://www.shared-interest.csail.mit.edu/computer-vision/
NLP: http://www.shared-interest.csail.mit.edu/nlp/
Probing: http://www.shared-interest.csail.mit.edu/human-annotation/

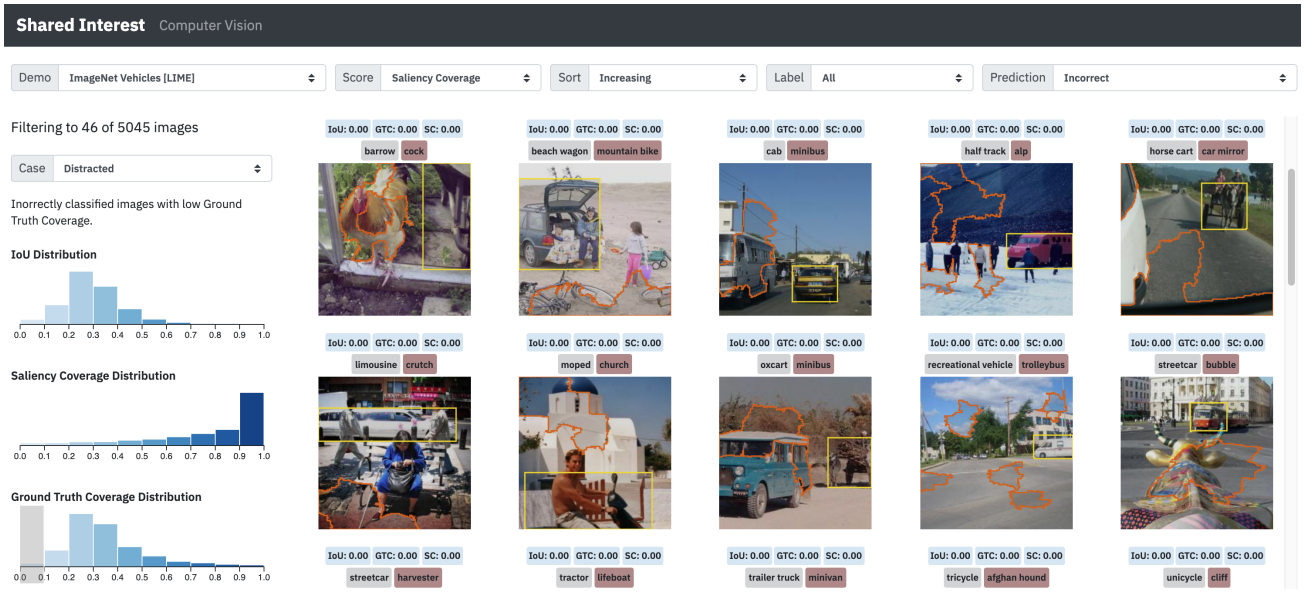

Figure S1. The computer vision interface live demo is available at: http://www.shared-interest.csail.mit.edu/computer-vision/

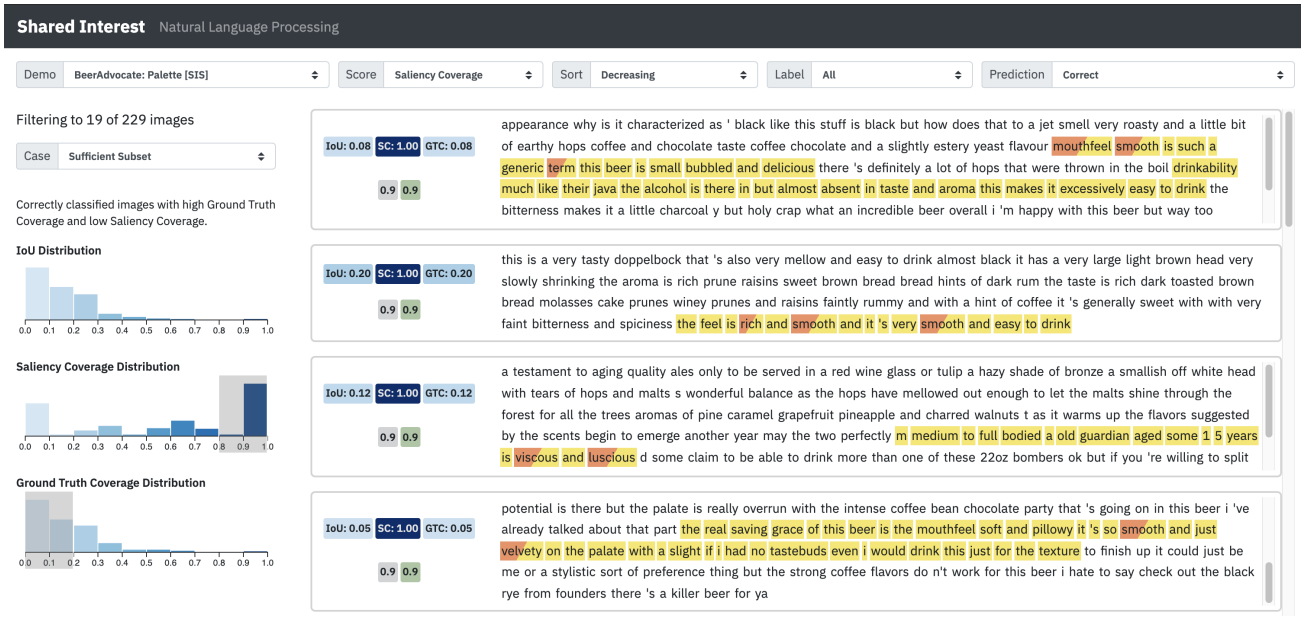

Figure S2. The natural language processing interface live demo is available at: http://www.shared-interest.csail.mit.edu/nlp/

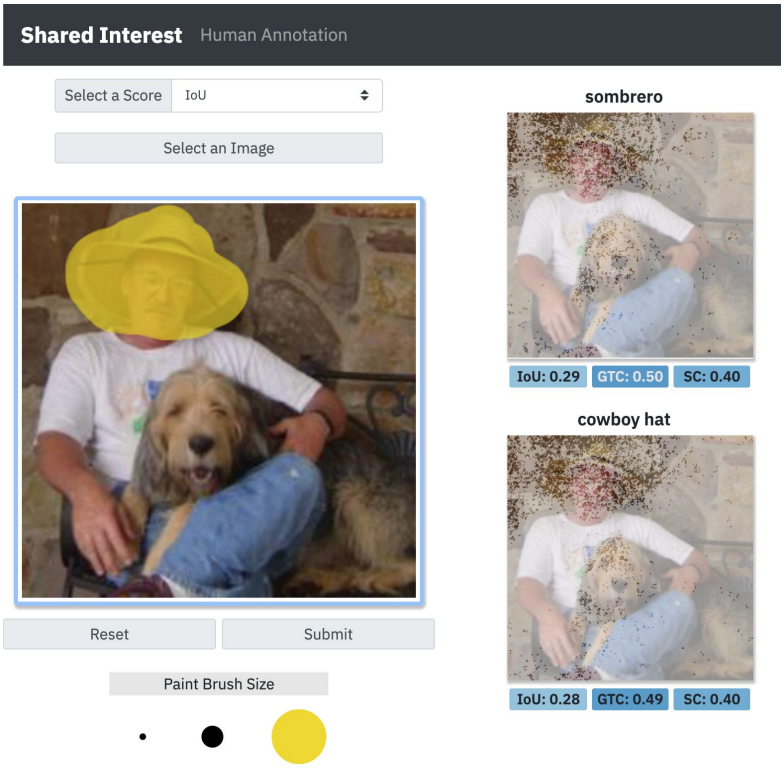

Figure S3. The human annotation interface live demo is available at: http://www.shared-interest.csail.mit.edu/human-annotation/

S2. Experimental Setup

Here we describe the implementation details of the datasets, models, and saliency methods.

S2.1. ImageNet Image Classification

In the ImageNet image classification examples (§ 3, § 4, § 5.3), we use two subsets of the original ImageNet dataset: the dog and vehicle subsets from ImageNet-9 (Xiao et al., 2020). Since ImageNet only provides bounding box annotations for a subset images, we further subset these sets to only contain images with annotations. On the ImageNet data, we use a pretrained ResNet50 (He et al., 2016) provided by PyTorch (Paszke et al., 2019) trained on 1000-way classification on ImageNet. In § 3 and § 4, we compute LIME (Ribeiro et al., 2016) explanations using the open source python package[2]. We compute the saliency feature set using the top 5 features that had a positive impact on the model's prediction, where features are super-pixels defined by QuickShift (Vedaldi & Soatto, 2008). In § 5.3, we compute vanilla gradients (Simonyan et al., 2013; Erhan et al., 2009) for all 1000 ImageNet classes. To discretize the output, we threshold each individual saliency map at $\mu + \sigma$.

S2.2. Melanoma Classification

In the melanoma classification example (§ 5.1), we use lesion images and segmentations from the ISIC 2016 Challenge (Gutman et al., 2016). Each image is classified as malignant or benign. We trained a ResNet50 (He et al., 2016) model from scratch for 4 epochs using Cross Entropy loss, Adam (Kingma & Ba, 2014) optimization, a learning rate of 0.1, a batch size of 128, and class-weighted sampling. We evaluate on the validation set, since the test set is not public, and achieve 0.822 accuracy on balanced classes. We display LIME (Ribeiro et al., 2016) explanations using the open source python package[1]. We compute the saliency feature set using the top 5 features that had a positive impact on the model's prediction, where features are super-pixels defined by QuickShift (Vedaldi & Soatto, 2008).

S2.3. BeerAdvocate Sentiment Regression

In the beer review regression examples (§ 4, § 5.2), we use beer reviews from the BeerAdvocate dataset processed by Lei et al. (2016). Each review is annotated with scores ranging from 0 to 1 in 0.1 increments representing 0 to 5 star reviews. Each review has scores and sentence level annotation for each aspect (aroma, appearance, palette, taste). We use a recurrent neural network (RNN) used by Carter et al. (2019a) on this dataset and trained on each individual aspect. We evaluate this model using SIS (Carter et al., 2019a) and Integrated Gradients (Sundararajan et al., 2017). The sufficient input subsets were selected via the SIS procedure using an 85% model confidence threshold. For comparison the integrated gradients were also iteratively selected from highest to lowest impact on the predicted class until the model was able to make the original prediction with 85% confidence.

S3. Additional Metric Details

Here we provide supplementary details for each of the Shared Interest metrics via a visual explanation in Figure S4 and additional examples for high and low values of each metric in Figure S5.

As described in §3 and shown visually in Figure S4, Shared Interest takes a set of saliency features S and a set of ground truth features G and outputs metrics for identifying instances of interest. IoU Coverage (IoU) represents the alignment between the model's decision and the human annotation and is maximized when $S = G$. Ground Truth Coverage (GTC) represents how many human-salient features the model uses and is maximized when $G \subseteq S$. Saliency Coverage (SC) represents how much model uses only human-salient features and is maximized when $S \subseteq G$.

A score of zero indicates the ground truth set and saliency feature set are disjoint. When a correctly classified instance has a low score, it often indicates the model was relying on background information such as a cyclist's helmet and uniform to predict *mountain bike* or train tracks to predict *electric locomotive*. When an instance has a low score and is incorrectly classified, it can indicate the model is focusing on a secondary object or incorrectly relying on background context (e.g., using snow to predict *arctic fox*).

A high IoU score indicates the explanation and ground truth feature sets are very similar. Correctly classified images with

[2] https://github.com/marcotcr/lime

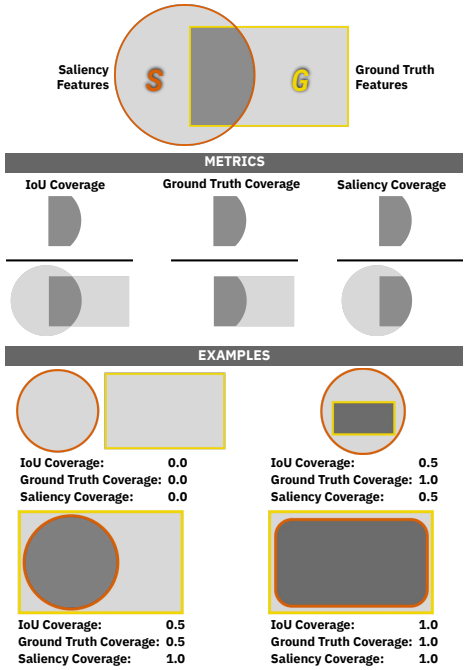

Figure S4. A visual explanation of the Shared Interest metrics.

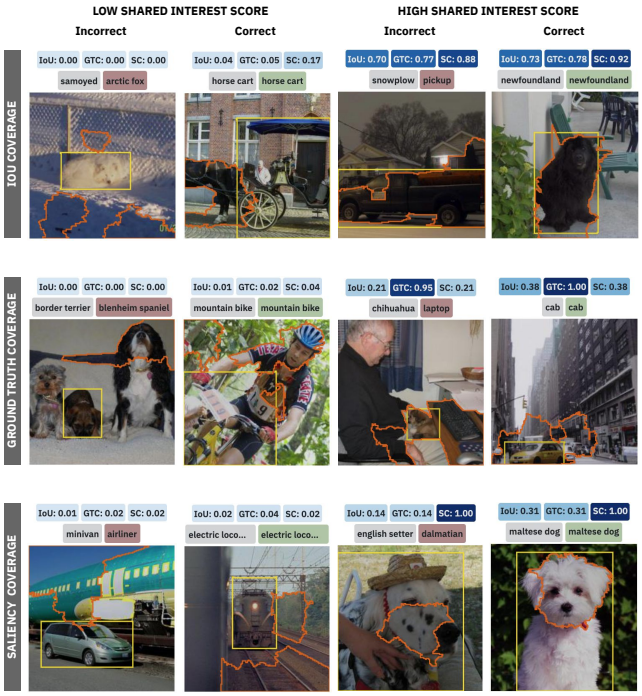

Figure S5. Examples of instances with high and low values for each of the Shared Interest metrics.

high IoU scores can help identify instances where the model was right in ways that tightly align with human reasoning such as relying on the entire body of the dog to predict *newfoundlander*. Incorrectly classified images with high IoU scores, on the other hand, can surface challenging instances, such as the image of a snowplowing truck that is labeled as *snowplow* but predicted as *pickup*.

High GTC indicates that the model is using every ground truth feature to make its decision. Correctly classified images with high GTC are cases where the model relies on the object and relevant background pixels (e.g., the cab and the street), to make a correct prediction. Incorrectly classified instances with high GTC are examples where the model overly relies on local contextual information such as using the keyboard and person's lap to predict *laptop*.

High SC indicates that the model relies on a subset of salient features to make its prediction. Correctly classified instances with high SC are instances where a subset of the object, such as the dog's face, was sufficient to make a correct prediction. Incorrectly classified instances with high SC are instances where the model uses an insufficient portion of the object to make a prediction (e.g., using a small region of black and white spots to predict *dalmatian*).