Evidence Humans Provide When Explaining Data-Labeling Decisions

Judah Newman, Bowen Wang, Valerie Zhao, Amy Zeng, Michael L. Littman, Blase Ur

Evidence Humans Provide When Explaining Data-Labeling Decisions

Judah Newman[1], Bowen Wang[1], Valerie Zhao[1], Amy Zeng[1], Michael L. Littman[2], and Blase Ur[1]

[1] University of Chicago, Chicago, IL 60637, USA
{jgnewman, bowenwang1996, vzhao, amyzeng, blase}@uchicago.edu
[2] Brown University, Providence, RI 02912, USA
mlittman@cs.brown.edu

Abstract. Because machine learning would benefit from reduced data requirements, some prior work has proposed using humans not just to label data, but also to explain those labels. To characterize the evidence humans might want to provide, we conducted a user study and a data experiment. In the user study, 75 participants provided classification labels for 20 photos, justifying those labels with free-text explanations. Explanations frequently referenced concepts (objects and attributes) in the image, yet 26% of explanations invoked concepts *not* in the image. Boolean logic was common in implicit form, but was rarely explicit. In a follow-up experiment on the Visual Genome dataset, we found that some concepts could be partially defined through their relationship to frequently co-occurring concepts, rather than only through labeling.

Keywords: Machine teaching · ML · Explanations · Data labeling.

1 Introduction

Supervised learning is the paradigm in which algorithms are trained with instances of data matched with carefully assigned labels. Based on these pairings of training instances and labels, a typical supervised-learning algorithm produces a classifier—a function that maps input examples (an image, a snippet of speech) to a binary label indicating whether the input is an instance of the target class. The ability to take such a dataset and produce an accurate classifier has improved dramatically over the years, finding success in domains including image classification [7], machine translation [21], and speech recognition [17].

Nonetheless, this technology is still limited to domains where labeled data is naturally plentiful or where there are strong incentives to make labeled data plentiful. To reduce the need for huge sets of training data, HCI work on machine teaching has begun to consider how to enable richer interactions between humans and machines and how to better support data labeling. Many of these efforts aim to capture extra information from humans to improve the algorithms since humans learn new concepts with much less data than current algorithms [11], presumably because of the extra background information they possess.

Explanations for why a human applied a particular label to a data instance is a promising type of extra information humans could provide. Stumpf et al. [19] investigated rich explanations for classifying email, highlighting how user feedback has the potential to significantly improve machine learning (*ML*). If explanations can successfully reduce the number of instances that must be labeled, there are three main benefits for classifying whether images contain given objects or attributes, which we collectively term *concepts*. First, classifiers could be efficiently defined ex post facto for concepts overlooked in initial data labeling by relating the overlooked concept to concepts that had already been labeled. Second, new concepts could be defined with less effort from humans by relating new concepts to those existing algorithms can already recognize. Third, this approach could enable personalized ML in building classifiers to recognize subjective concepts like "my house," rather than only "a house" in general.

To characterize the types of evidence humans could provide when explaining and justifying data-labeling decisions, we performed a formative user study and a companion experiment on an existing dataset. While prior work has focused on text classification [3, 10], we examine the more complicated domain of image classification. To cast a broad net in eliciting evidence humans might provide in their explanations, participants in our user study typed explanations in unconstrained natural language. In total, 75 participants labeled whether or not twenty images represented a given *target concept* (e.g., "crossroads," "old") and spent at least one minute for each explaining their classification in prose.

We centered our analysis on answering the following five research questions:

- **RQ 1**: What broad types of evidence do participants use to justify a label?
- **RQ 2**: How did participants structure explanations?
- **RQ 3**: How did the evidence and structure vary by person, task, and label?
- **RQ 4**: How often did explanations include ambiguous language, and how often did participants neglect to explicitly make logical connections?
- **RQ 5**: How did participants perceive their teaching and the overall process?

Explanations frequently referred to objects and attributes visible in the image. Surprisingly, 26% of explanations invoked objects and attributes *not* visible in the image. Participants often described spatial relationships (e.g., "next to") when explaining labels for an object ("nightstand") and functional relationships (e.g., X "uses" Y) when explaining labels for an action ("eating"). Many explanations also referred to abstract concepts (e.g., "style", "technology"), which existing object-recognition algorithms struggle to identify.

We observed a number of common structures in explanations. Overall, 26% of explanations included a generalized definition of the target concept before the participant explained why the image did or did not represent that target concept. While only 15% of explanations contained explicit Boolean logic, many other responses implicitly relied on Boolean logic. We also observed a number of ambiguities in explanations that would impair their direct application.

Based on the types of connections between co-occurring concepts that participants referenced in their explanations, we further explored whether target concepts could be defined in terms of their relationship to other concepts through an

experiment on the Visual Genome data set [8]. We used heuristics to automatically generate potential *definitions* for each of the 2,243 target concepts that appeared at least 100 times in Visual Genome. Each definition was a statement in Boolean logic containing up to five auxiliary concepts (e.g., "wetsuit" was defined as likely to occur in images containing "water" and a "surfboard" and the color "black"). We imagined that all images for which that logical statement was true could be classified as containing the target concept, and all images containing none of those auxiliary concepts would be classified as not containing the target concept. Doing so, which notably does not require any additional human labeling of images, we found that 4.9% of the 2,243 target concepts could already be classified with $F1 \geq 0.5$, while 29% could be classified with $F1 \geq 0.25$. While such accuracy is insufficient for training current algorithms, this experiment demonstrates that these co-occurring concepts can be used to partially define new concepts, bootstrapping future interactions.

We conclude by discussing how our characterization of the evidence participants provided when explaining image-classification labels suggests design directions for user interfaces that collect similar information in systematic and structured ways, enabling the information to be used directly by algorithms. We further discuss how the results of our Visual Genome experiment suggest new interactions for minimizing human image-labeling effort. To spur further research on explanatory machine learning, we are publicly releasing our anonymized dataset for the user study and the code from our experiment.[3]

2 Related Work

Crowdsourcing is a primary method of gathering labels for ML algorithms. However, incorporating human input can often introduce variability. For example, Kulesza et al. [9] identified how users' notions of the target concept evolves as they complete labeling tasks, resulting in inconsistent labels. New collaboration methods use crowdsourcing to address unclear label guidelines. For example, in the Revolt platform, Chang et al. [5] created a group workflow where users label items, discuss conflicts, and make revisions. Revolt presents the labels from these stages to a worker who makes the final decisions. Motivated by this work's findings around ambiguity in labeling, we included "unsure" as an option for labels. Uncertainty may also come from the task itself. Laput et al. [12] used crowdsourced answers to simple questions from sensor data (e.g. "how many drinks are on the table?") to train classifiers. Tasks that required personal judgment or additional context led to poor performance.

The broad research area of machine teaching has focused on enabling richer interactions between humans and algorithms, allowing humans to teach machines concepts through mechanisms other than simply labeling data. Allowing users to provide explanations for labels in supervised learning builds on work around dividing problems into smaller parts [18]. By emphasizing concept decomposition, machine teaching can be useful for applications with abundant unlabeled

[3] Available at: https://github.com/UChicagoSUPERgroup/interact19

data where contextual information is necessary. Prior work has identified best practices for helping humans train machines. Amershi et al. [2] noted three elements of effective machine teaching: (1) illustrating the current state of the learned concept; (2) helping users select higher-quality training examples; and (3) presenting multiple learning models. Data labeling and classification are the most popular ways for users to interact with ML systems, but people naturally want to provide more feedback than just labels [20]. Amershi et al. [1] found that richer user interactions and increased transparency can improve model accuracy.

A relatively small literature has investigated human-provided explanations in the context of training ML algorithms, our core aim. Stumpf et al. [19] proposed leveraging user feedback in the form of rich explanations to improve email classification. In subsequent work, Kulesza et al. [10] proposed explanations for improving debugging within end-user programming, again related to email classification. Brooks et al. [3] focused more broadly on using interactive feedback to improve text classifiers, finding particular benefits from visual summaries.

Rather than attempting to parse free-text explanations, Ratner et al. [15] let users define logical labeling functions based on arbitrary heuristics. In follow-up work, Hancock et al. [6] proposed applying techniques from natural-language processing to automatically translate free-text explanations into logical labeling functions. In contrast, we take a step back and examine broader types of explanatory information free-text explanations contain. Further, most prior work on explanatory ML was on text classification; we examine image classification.

While we focus on using explanations to improve ML training, a burgeoning literature has begun to explore the opposite problem of explaining existing algorithms. For example, Stumpf et al. [20] explored generating explanations for a naïve Bayes classifier. They found that the explanation paradigm influences user feedback. With the recent success of deep learning, there has been increasing concern about the interpretability of neural networks. Among the many recent attempts to explain deep learning, Park et al. [14] used an attention model pointing at features influencing classification.

3 User-Study Methodology

The goal for our user study was to identify the types of evidence humans provide in explaining and justifying data labels, as well as to characterize how they structured their presentation of this evidence. This understanding can inform the design of future interfaces that elicit the same types of information in a more structured and directly actionable form. This section describes our data sources, participant recruitment, and study protocol.

3.1 Terminology

We tasked participants with "teaching a computer" new concepts. We define a **concept** to be any noun, verb, adjective, or adverb that could plausibly appear in an image. We distinguish among the following concept types:

– **Object**: Noun (e.g., *crossroads* or *plane*),
– **Attribute**: Adjective or Adverb (e.g., *old* or *fast*),
– **Action**: Verb (e.g., *eating* or *smiling*).

We also divide concepts into abstract and concrete concepts. We defined abstract concepts as those that are not generalizable from viewing a single instance, such as "decor," "weather," and "technology."

Following a study introduction, we presented participants with a series of photos that either did or did not contain the concept. The participant was asked to *label* whether or not a concept was present in the image. We adopt the following terminology, in which an image a participant saw is termed an *instance*:

– **Positive instance**: A photo that *did* contain a concept.
– **Negative instance**: A photo that *did not* contain a concept.

We included both positive and negative instances to characterize how explanations differed based on whether the participant was identifying how they recognized a concept or noting which aspects essential to a concept were missing.

3.2 Source Data

Participants labeled the twenty images shown in Table 1. These images encompassed four target concepts, with five different instances (photos) for each concept. To disperse learning effects, we randomized the order of the four concepts. Within a concept, the order of the five instances was also randomized. The four target concepts included two objects (*nightstand* and *crossroad*), one action (*eating*), and one attribute (*old*). We chose these concepts to represent different levels of abstractness, ranging from the concrete (*nightstand*) to the abstract (*old*). These different target concepts also allow us to examine how explanations vary across slightly different tasks. Because we hypothesized that participants' explanations would differ for positive and negative instances, we selected three positive and two negative instances per concept.

The photos and metadata were taken from Visual Genome [8], which contains 108,077 images. On average, each image contains 35 objects and 26 attributes labeled by Mechanical Turk workers. We used Visual Genome because each photo includes labels for objects and attributes, providing us with a rich list of concepts. After selecting the four target concepts for the study, we chose positive instances by searching Visual Genome for that concept, randomly selecting three. We chose two negative instances for each by randomly selecting two photos from among those that contained related concepts, but not the target, according to the Visual Genome labels.

3.3 Procedure and Study Structure

We recruited participants on Amazon's Mechanical Turk for "a research study on teaching computers." Workers aged 18+ who lived in the United States and had completed 100+ HITs with a 95%+ approval rating were eligible. Through pilot studies, we adjusted the number of tasks so that the study would average 30 minutes to minimize fatigue [4]. We compensated participants $5 (USD).

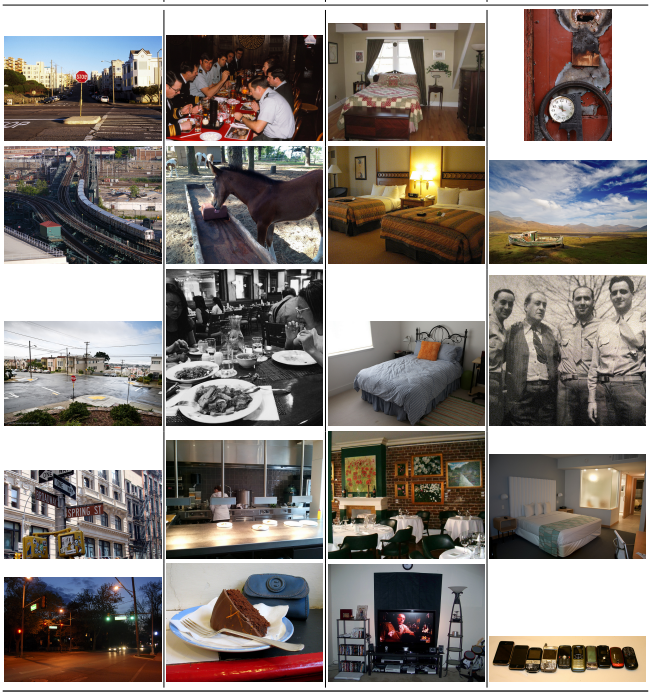
Crossroads	Eating	Nightstand	Old
			

Table 1. The twenty images participants labeled (and explained) in our user study. Each concept includes three positive instances and two negative instances.

The study began with an introduction emphasizing the importance of detailed explanations and that participants were teaching a computer, not a human. We then introduced the first target concept and presented the five instances of that target concept sequentially. For each instance, the participant selected "yes," "no," or "unsure" to "is there a *target concept* in the photo above?"

After the participant chose a label, we asked them to explain their classification decision in a multi-line text box. To encourage detailed explanations, participants could not proceed until one minute had elapsed.

After the participant labeled and explained all five instances for a target concept, we asked five *reflection questions* to evaluate their self-perceptions of their teaching. To gauge perceptions of generalizability, we asked about perceptions of the thoroughness of their teaching for images similar to the five study instances and all future images. At the end of the study, we asked three *process-reflection questions* about how participants approached teaching a computer.

3.4 Analysis Methods and Metrics

To answer our research questions, we both quantitatively and qualitatively analyzed the explanation text. As a first step, members of the research team read

all free-text explanations and informally noted types of evidence and structures they observed. Explanations that any member of the research team identified as especially representative or unique were discussed at a series of full-group research meetings. Following this exploratory process, the members of the research team formally developed a codebook based on these notes. A coder would read an explanation, identify all concepts the explanation referenced, and then answer eleven numerical or true/false questions about the explanation's semantics and structure pinpointed in our exploratory process. Some examples of the numerical or true/false questions are as follows: "How many spatial relationships does this explaination contain?" "How many objects did the explanation reference that are not in the photo?"

Four members of the research team were in charge of the coding process. To ensure consistent understanding of the codebook, all members of the coding team used the codebook to code 60 random explanations. The coding team then met to review those 60 explanations to ensure that their understanding of the codebook was aligned. After discussing differences in this set, two members of the research team were assigned to independently code each explanation. The results of the coding were our main data set for answering RQ 1 through RQ 4. The mean Cohen's κ across the characteristics the team coded was 0.681. We analyzed participants' responses to reflection questions (RQ 5) separately following an analogous process. The mean Cohen's κ for reflection questions was 0.824.

4 User Study Results

We had 75 participants in our user study. Each participant labeled and explained 20 instances. Thus, our data comprises 1,500 explanations. Participants mentioned 19,749 unique concepts across these 1,500 explanations. As shown in Table 2, each image had between two and nine concepts that were mentioned in at least 20 different participants' explanations for that image.

RQ 1 (Evidence in Explanations) Explanations connect the target concept to other concepts. Participants' explanations of why they classified an image as containing a target concept or not often referred to other concepts present in the image. We refer to these other concepts as *auxiliary concepts*. Participants referenced an average of 3.65 auxiliary concepts per explanation. Our coding revealed that 60% of the concepts referenced were concrete concepts visible in the photo, 7% were concrete concepts *not* visible in the photo, and 33% were abstract concepts. These three categories impose different requirements for future interfaces and ML algorithms. While current computer-vision systems recognize concrete objects well [16], handling concepts not visible in a photo and abstract concepts requires new methods.

Explanations revealed it was not just the presence or absence of certain concepts, but rather the way they connect that influenced labeling decisions. For example, in the explanation below, the participant identifies specific relationships

Image	Consensus label	# frequent concepts	% frequent concepts visible
Crossroads-1	Yes	6	100%
Crossroads-2	No	5	60%
Crossroads-3	Yes	6	100%
Crossroads-4	Yes	4	25%
Crossroads-5	Yes	7	100%
Eating-1	Yes	9	88%
Eating-2	Yes	9	88%
Eating-3	Yes	8	100%
Eating-4	No	5	60%
Eating-5	No	8	50%
Nightstand-1	Yes	7	72%
Nightstand-2	Yes	6	100%
Nightstand-3	Yes	7	85%
Nightstand-4	No	6	50%
Nightstand-5	No	5	40%
Old-1	Yes	7	57%
Old-2	Yes	5	60%
Old-3	Yes	8	75%
Old-4	No	6	67%
Old-5	Yes	2	50%

Table 2. A summary of the responses for the twenty images in our study, including the majority of participants' *consensus label* for whether the image depicted the concept. *Frequent concepts* are the number of distinct concepts mentioned by ≥ 20 participants each, and we note how many of these frequent concepts were visible in the image.

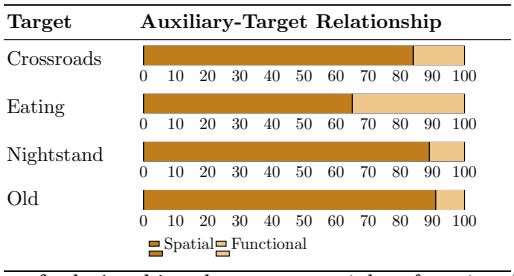


Table 3. Proportion of relationships that were *spatial* or *functional* by target concept.

between concepts that define eating. The mere presence of concepts like food, a table, or people is insufficient.

> "There are plates with food on a table with people sitting around it. There are utensils such as a knife and fork on the plates that people use to eat the food with."

Within participants' explanations, we frequently observed target-auxiliary relationships that can be characterized as either spatial or functional relationships. We defined spatial relationships to be those that can be identified by the relative position of pixels in an image, whereas a functional relationship requires a more complex understanding of interactions. As shown in Table 3, 81% of target-auxiliary relationships in participants' explanations were spatial, while the remaining 19% were functional.

We were also interested in how participants connected the concepts to which their explanations referred. Thus, we examined how often participants explicitly used Boolean logic, as well as how many steps their reasoning encoded.

We searched for the use of logical "and" and "or" connectors as evidence of explicit Boolean logic. The use of Boolean logic indicates more complicated reasoning than direct correlations, echoing the types of reasoning work in weakly supervised learning has begun to explore [15]. The following explanation is an example of using Boolean logic in an explanation because the classification decision depends on a logical combination:

> "This is not eating. Eating involves someone actively putting food into their mouth and swallowing it."

Few explanations contained explicit Boolean logic. Roughly 15% of explanations contained explicit Boolean logic. Many explanations, however, appeared to contain Boolean logic implicitly, such as the following example:

> "A nightstand is a small table placed beside a bed for people to place items on. So, the small table in the corner is a nightstand."

One might conclude that such an explanation implicitly requires both that there is a small table *and* that the table be located beside a bed *and* that the purpose of the table is for holding items. To accurately characterize our data, we chose not to code such implicit examples as exemplifying Boolean logic. However, implicit examples appeared commonly.

A fraction of explanations employed multi-step reasoning. Another metric to characterize the complexity of reasoning is the number of logical jumps made in an explanation. The more steps, the harder it could be for a computer to learn from their explanation. The following explanation contains only one logical step:

> "I based my decision on the fact that I saw a bed. Usually a bed has a night stand beside it."

In contrast, the following explanation instead exemplifies two-step reasoning:

> "People are sitting at a table with plates of food in front of them. Some of them are holding a fork, indicating they are, or expect to be, eating the food on their plate."

The participant first identifies people at a table, augmenting this with the fact that they can use the forks they have to eat. It is difficult to communicate multi-step reasoning to a computer; relative to simple correlation, it requires greater understanding of entities and their relationships.

Multi-step reasoning was not common in participants' explanations. 53% of explanations used single-step reasoning, directly connecting all evidence (auxiliary concepts) to the target concept. 6% of explanations involved two logical steps, while the remaining 1% included at least three logical steps. Surprisingly, 40% of the explanations never explicitly connected the evidence provided to the target concept. We discuss such ambiguities further below as part of RQ4.

RQ 2 (Explanation Structure) Abstract, generalized definitions of target concepts were common. In addition to, or in place of, evidence from the photo itself, some explanations contained a generalized, abstract definition of the target concept. We termed these *definitional structures*. The following explanation is one such example because it defines the general class "nightstand," rather than commenting directly on specific evidence in a data instance:

> "A nightstand is a small table placed beside a bed for people to place items on. So, the small table in the corner is a nightstand."

We found that 26% of explanations included such a definitional structure. The proportion of explanations that included a definitional structure varied across target concepts: 42% of "nightstand" explanations, 28% of "crossroad" explanations, 20% of "eating" explanations, and 11% of "old" explanations contained a definitional structure. This variation suggests that participants may have used definitional structures more frequently for more concrete concepts, such as "nightstand."

Explanations used an identify-explain structure. Explanations often consisted of two parts. First, participants would identify particular aspects of the photo. They would subsequently explain how those aspects connected to their classification decision. The following explanation is one of the many examples of this structure:

> "There are roads in the image. The two roads meet and cross each other. There is a stop sign. There are usually stops signs or traffic lights at a crossroad."

We found that 58% of explanations followed this identify-explain structure. That the majority of explanations did so suggests that future interfaces for eliciting explanations may benefit from explicitly incorporating this two-step process.

Participants occasionally described their process. Finally, some participants explained the process they used to reach a decision. For example, they would talk about where they first looked or what drew their attention. In total, 18% of explanations, including the example below, did so:

> "Well, I looked to the right and saw a desk. Then I looked to the left and saw a lamp. That is a good sign of a nightstand. Then there were books on it. Definitely a nightstand."

RQ 3 (Inter-Participant / Inter-Task Variation) We observed differences in explanations across participants and tasks.

Participants used more functional relationships to explain eating. Relationships were functional or spatial. As shown in Table 3, participants used a higher proportion of functional relationships when explaining eating classifications (35%) than when explaining classifications for the other three target

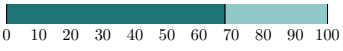
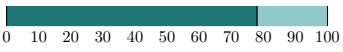
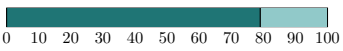
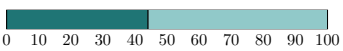
Target	Type of Concepts
Crossroads	
Eating	
Nightstand	
Old	

Table 4. Proportion of *concrete* and *abstract* concepts by target concept.

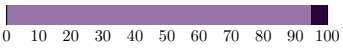
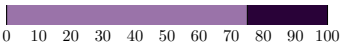
Instance	Visibility of Concepts
Positive	
Negative	

Table 5. Proportion of concepts *visible* for positive ("yes") / negative ("no") instances.

concepts (9%–16%). We speculate this is because eating is an action, which functional relationships lend themselves to describing.

The types of auxiliary concepts referenced varied based on the target concept. As shown in Table 4, participants' explanations contained a comparatively higher proportion of abstract auxiliary concepts when explaining classifications for "old." Notably, "old" itself is an abstract concept. In contrast, "nightstand" is a fairly concrete concept.

We also found that participants referenced concepts not in the photo more frequently for negative instances than for positive instances, as shown in Table 5. This result may seem intuitive since, for negative instances, the target concept itself is not contained in the photo. However, for both positive and negative instances, participants sometimes referred to objects not visible in the photo, which has implications for designing user interfaces for explanatory labeling, which we elaborate on in Section 6. Note that we observed two main reasons explanations for negative instances (those not containing the target) nonetheless referenced auxiliary concepts visible in the image. First, participants sometimes pointed out concepts incongruous with the target concept, such as:

> "This is a dining room. There are no nightstands in dining rooms."

Second, participants would point out auxiliary concepts that, in isolation, might be suggestive of the presence of the target concept, and then highlight missing auxiliary concepts or incorrect relationships between concepts. The following is one such example (for a negative instance of "nightstand"):

> "There are tables in the photo, but they are not beside a bed. There are no lamps on the tables. There are objects on the table, but they are kitchen objects."

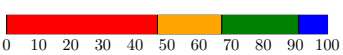
Concept	Part of Speech
Crossroads	
Eating	
Nightstand	
Old	

Table 6. Part-of-speech distribution by target concept.

The usage of parts of speech differed across target concepts. While computer vision systems are particularly adept at recognizing objects in images [16], we observed many parts of speech in participants' explanations, as shown in Table 6. Summing across all explanations, 58% of unique concepts mentioned were objects (nouns), while 20% were actions (verbs). Attributes (adjectives and adverbs) were much less common than objects; only 14% of concepts were adjectives and only 8% were adverbs, though this may be an artifact of our task involving only static images. The non-negligible inclusion of concepts that were verbs or adjectives reinforces the need to account for other parts of speech when eliciting explanations from users. The usage of concepts of different parts of speech differed by target concept. Even though nouns were used the majority of the time, participants tended to use more words that were of the same part of speech as the target concept. For example, 24% of the auxiliary concepts used in the explanations for "old" were adjectives, while 7%–15% of auxiliary concepts were adjectives for the other three target concepts.

Explanations varied in length. The length of explanations ranged widely. While the mean length was 29 words, explanations ranged from a single word to 114 words long. Each participant used between 129 and 1,373 words in total (summed across the 20 instances), with a median of 528 words. Explanation length also varied slightly by target concept. Explanations for *crossroad* were shorter than for the other three concepts. Furthermore, participants gave slightly longer explanations for positive instances than for negative instances.

RQ 4 (Ambiguities in Explanations) Explanations sometimes contained ambiguities that would impair algorithms from using them directly. We observed two key ambiguities: not explicitly connecting the explanation to the target concept and using undefined pronouns.

Some evidence did not connect to the target concept. As mentioned earlier, 40% of explanations never explicitly connected the evidence to the target concept. The example below contains potentially important information. However, because it does not explicitly connect to the target concept ("nightstand"), it would be difficult for a machine to use:

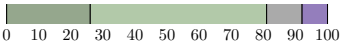
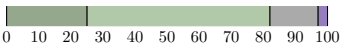
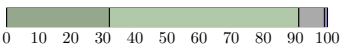
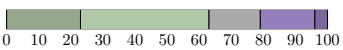
Target	Agreement
Crossroads	
Eating	
Nightstand	
Old	

Table 7. Participants' self-reported agreement with the following statement: "I feel I have thoroughly taught the computer to identify whether or not future images similar to the five examples in this study represent *CONCEPT*."

> "Although this appears to be a hotel room, there is still a small table located between the beds with a lamp on it and a clock."

In contrast, the explanation below did not provide as much information as the one above, but likely would be easier for an algorithm to leverage because the evidence is explicitly connected to the target concept ("crossroad"):

> "This is a crossroad. Two streets intersect or cross, making it a crossroad."

Ambiguous pronouns were used frequently. The other source of ambiguity comes from the use of ambiguous pronouns. About 10% of explanations, including the example below, used ambiguous pronouns. It is difficult to use an explanation when what "this" refers to is unspecified:

> "While this is something that can be eaten, there is nobody doing the eating in this image."

RQ 5 (Reflection) Although participants had no objective basis on which to evaluate the quality of their teaching, we were curious how well they felt they did. If a participant feels they have sufficiently taught the computer how to complete a task, the motivation to continue teaching may decline. Across all four concepts, over 50% of participants agreed or strongly agreed that "I feel I have thoroughly taught the computer to identify whether or not future images similar to the five examples in this study represent the *target concept*." Given current and likely future ML data requirements, these judgments are almost certainly highly overconfident. As shown in Table 7, participants were even more confident about classifying photos similar to those in the study despite labeling and explaining only five instances. Participants felt more confident for concrete concepts (e.g., nightstand) than abstract ones.

In our process-reflection questions, 77% of participants reported they would have changed their explanation if justifying their classification to a human, rather

than a computer. Notably, 19% of participants said they would check the human's understanding of the concept; P27 said, "I would be able to ask them if they had any questions about it. I could not do that with a computer." Further, 11% of participants wanted more physical input modalities, such as gestures.

5 Experiment on Automated Labeling

Participants in our user study often explained classification labels by relating the target concept to other concepts in the image. Buoyed by our findings, we conducted an experiment to estimate the degree to which inter-concept relationships can be used to automatically apply classification labels.

For instance, if a dataset already labeled by either humans or off-the-shelf object-recognition software does not contain a label for a concept (e.g., "wetsuit"), how helpful would it be if a human explained to the system that images containing "water," a "surfboard," and the color "black" likely contain a "wetsuit?" Currently, humans need to label whether every image in a huge training dataset contains a wetsuit. Using a human-provided explanation like the one above, could most images be automatically and very accurately classified as *not* containing that concept? If only a few images remain, human labeling effort could be used far more productively.

5.1 Procedure

For each target concept, we automatically constructed a statement in Boolean logic defining that target concept in terms of up to five auxiliary concepts. We simulated automated labeling by applying that logical definition to predict the presence or absence of that target concept in all images in Visual Genome 1.4 [8]. We treated the presence or absence of the target concept's label in Visual Genome's label set for that image as the ground truth classification. To balance precision and recall, we used the F_1 score as our metric. F_1 is the harmonic mean between precision and recall. If our target concept is "wetsuit," the precision is the percentage of photos we label as containing a wetsuit that actually contain a wetsuit. Recall is the percentage of all photos that contain a wetsuit that we label as containing a wetsuit.

We first investigated this approach by defining the four target concepts from our user study based on the evidence participants provided. In particular, we used the five auxiliary concepts most frequently included in participants' explanations. However, this study only investigated a small number of concepts. To benchmark this conceptual approach more broadly, we also automatically constructed definitions for all 2,243 target concepts that appear at least 100 times in Visual Genome. For each, we selected the five concepts from among those that often co-occur with the target and that had the highest F_1 scores when individually defining the target concept.

We then constructed all possible logical combinations of the five auxiliary concepts, including those that exclude some of the five. We chose the definition

with the highest F_1 score and used it in all further analyses. For example, the auxiliary concepts for "watch" were "wrist," "shirt," "wear," "play," and "man," leading to the following definition: *(play | wrist) & (wrist | man) & (shirt | man | wear)*. This definition alone had $F_1 = 0.34$ classifying "watch" in Visual Genome without any further human labeling. To simulate labeling negative instances (the bulk of any dataset), we labeled all images containing *none* of the auxiliary concepts as negative instances.

Synonyms can affect data accuracy by causing one to wrongly believe an image of a dog to not contain a dog because it contains the synonymous label "canine," but not "dog." To partially account for this, we treated co-occurrence in a WordNet [13] synset as a match. Nonetheless, we manually observed that we may *underestimate* labeling success, such as some nightstand images being labeled "table," but not "nightstand," despite the two not being synonyms.

5.2 Results

We first present results of applying our approach to "nightstand," "eating," "crossroads," and "old" using the explanatory data collected in our user study. We then simulate this approach for all 2,243 target concepts that appeared frequently in Visual Genome.

Definitions relating target to auxiliary concepts can, with minimal human effort, partially label images. Using the auxiliary concepts study participants most frequently referenced for "nightstand," the definition *lamp & bed* could be used to label nightstands in Visual Genome with $F_1 = 0.48$ (0.37 precision, 0.69 recall). While insufficient for immediate ML use, minimal human labeling could make it sufficient. Without this definition, a human would have needed to apply "nightstand" labels to all 108,077 Visual Genome images. In contrast, this simple definition suggested there were nightstands in 873 images, 322 of which actually contained a nightstand. Labeling all images that contained neither "lamp" nor "bed" as negative instances would result in 102,167 true negatives (those without a nightstand) and only 8 false negatives. This automated method left 5,902 images unlabeled, of which 142 contained a nightstand.

If humans manually corrected the labels of the 873 images the automated method classified as having a "nightstand," it would leave 322 true positives, 102,718 true negatives, and 8 false negatives ($F_1 = 0.99$ on 103,048 images). That said, it would exclude the 5,902 images that contain either a "lamp" or a "bed," and these images might represent particularly helpful training data as potential boundary cases.

"Nightstand" was the most concrete of the four concepts, and thus easiest to define. Study participants were more likely to write definitional structures for "nightstand" than the others. In line with this finding, automated labeling of the other three target concepts from the user study yielded worse results, yet still showed some promise. "Eat" (substituted for "eating," which Visual Genome synsets lacked) was automatically defined as *(person & food) | (plate & table) | (plate & mouth) | (food & mouth) | (table & mouth)*; $F_1 = 0.15$. "Intersection" (substituted for "crossroads") was defined as *(road & stop & street) | (road &*

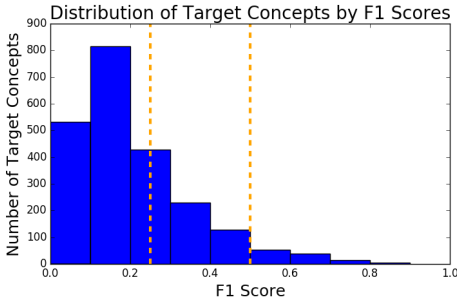


Fig. 1. Histogram of F_1 score ranges (e.g., the leftmost is 0.0–0.1) for the 2,243 target concepts that appeared in ≥ 100 Visual Genome images.

Concept	Source	Most Accurate Definition	F1 Score
nightstand	User study	lamp & bed	0.48
	Simulated	(quilt & bedroom) | (headboard & bed) | (bedspread & bedroom)	0.42
eat	User study	(person & food) | (plate & table) | (plate & mouth) | (food & mouth) | (table & mouth)	0.15
	Simulated	giraffe | zebra | crop	0.29
intersection	User study	(road & stop & street) | (road & direction & sign) | (road & sign & street) | (stop & sign & street)	0.11
	Simulated	(crossing | stopped) & (trafficlight | traffic | crossing) & (trafficlight | traffic | signal | stopped)	0.19
old	User study	look | boat | picture	0.08
	Simulated	(rusty | building | brick) & (rusty | brick | wooden) & (rusty | building | window | wooden) & (building | brick | window | wooden)	0.16

Table 8. Comparison of definitions for the four target concepts generated from the results of the user study and simulated labeling.

direction & sign) | (road & sign & street) | (stop & sign & street); $F_1 = 0.15$. The definition for "old" was *look | boat | picture* ($F_1 = 0.08$), overfitting to an image of an old boat from the user study (see Table 1).

We then applied the same method to all 2,243 target concepts that appear at least 100 times in Visual Genome 1.4. This allowed us to simulate the technique more broadly.

This approach generalizes to many concepts. Overall, we found 29% of these 2,243 target concepts could be classified with $F_1 \geq 0.25$, while 4.9% could be classified with $F_1 \geq 0.5$. Figure 1 shows the full distribution of F_1 scores.

Many targets were defined in terms of auxiliary concepts that make intuitive sense, suggesting humans would likely have volunteered similar ones. For example, "bride" was defined as *groom | bridal gown* with $F_1 = 0.81$. "Melted" was defined as *cheese & (crust | burned)* with $F_1 = 0.47$.

Some of the most effective definitions were very succinct:

– *sofa | bed | headboard* $\Rightarrow$ "pillow" ($F_1 = 0.64$)

- *beach* ⇒ "sand" ($F_1 = 0.61$)
- *cheese* | *pepperoni* | *crust* ⇒ "pizza" ($F_1 = 0.60$)
- *feather* | *beak* ⇒ "bird" ($F_1 = 0.55$)

Many of the target concepts with low F_1 scores were relatively abstract. In particular, based on Visual Genome's categorization of labels, F_1 scores for objects were higher than those for attributes or relationships. In total, 37% of objects could be classified with $F_1 \geq 0.25$, and 6.5% with $F_1 \geq 0.50$.

Automatic classifications of negative instances were highly accurate. We labeled images to be negative instances if they contained *none* of the auxiliary concepts in the automatically generated definition. This heuristic proved $\geq 95\%$ accurate for 98.8% of the 2,243 target concepts. Furthermore, it was $\geq 90\%$ accurate for 99.6% of target concepts.

6 Conclusions and Discussion

In a user study, we elicited and characterized 75 humans' free-text explanations of data labels for the type of image-classification tasks used in supervised learning. Through a follow-up simulation experiment on the Visual Genome dataset, we showed how the types of explanatory information we observed can underpin semi-automated labeling of large datasets for hundreds of concepts. Our protocol, including publicly releasing our anonymized dataset, was approved by our IRB. Participants opted into this data release.

6.1 Design Implications and Future Work

Our simulation results showed the initial promise of semi-automated labeling based on relating a target to auxiliary concepts. Besides evaluating the performance of this method on more training sets, building and testing interfaces for eliciting such definitions is key future work. Many participants in the user study defined the target concept abstractly before referencing the image, which suggests that interfaces could empower users to do so without specific examples. Breaking down teaching into multiple steps is supported by prior work in machine teaching [18]. Specific data instances could then help communicate to users the system's current understanding of a concept, as in Revolt [5]. These interfaces could be compared against others grounded in specific instances, which sometimes overfit (e.g., the "old" boat).

Nonetheless, the tendency for participants' explanations to logically combine evidence implicitly, rather than explicitly, highlights the need for designing interactions or interfaces that elicit such logic explicitly. Inspired by visual programming, one could imagine an interface that lets users "wire" concepts together to indicate these connections.

Recent work has sought to improve classifiers by applying NLP techniques to free-text explanations of labels, finding some success even without an HCI focus [6]. That work focused on text classification, not image classification. Nevertheless, our findings suggest best practices for designing user interfaces to minimize ambiguities when capturing text-based explanations. Many of participants'

explanations contained ambiguous pronouns. An interface could automatically detect them and guide the user in clarifying any pronouns they used. Participants also often neglected to directly connect the evidence they presented to the target concept, which a multi-step interaction may be able to correct.

Other work has proposed letting humans define computational functions to automatically label training data [15]. This approach might be unnecessary for labeling concrete objects, given the lack of multi-step reasoning and Boolean logic observed in the user study, as well as the simple, succinct, and accurate definitions that emerged from our Visual Genome simulation experiment. Future work could investigate the usability of such an approach, which may enable definitions of abstract concepts (e.g., "old") our experiments struggled with.

We grounded our task in explanations to "a computer" as we felt that would best capture typical data labeling processes. While we asked particpants to speculate and self-report how they would have changed their explanations if they had been teaching a human, we could also run a study where they actually taught a human. This future work could give insight into how particpants might unwittingly change their explanations when teaching a computer. It could also highlight techniques that are used when teaching another person that could be simulated by a dynamic interface.

Lastly, future work could investigate input modalities. Based on the identify-explain process, users could point at important parts of an image, then connect each to the target.

6.2 Limitations

Study participants were an unrepresentative convenience sample recruited on Amazon's Mechanical Turk, limiting generalizability. A timing minimum different from the 60 seconds we used may have elicited different explanations. Further, we collected rich human-subjects data in our user study, but only for a small number of target concepts. Due to the small number of target concepts and small number of instances tested per concept, our findings may not generalize to different concepts or instances. Explanations may have been different for different concepts. Exploring explanations for a different group of concepts could be an avenue for future work. We partially address this limited generalizability by conducting a simulated experiment on 2,243 Visual Genome concepts.

References

1. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: The role of humans in interactive machine learning. In: Proc. AAAI (2014)
2. Amershi, S., Fogarty, J., Kapoor, A., Tan, D.: Effective end-user interaction with machine learning. In: Proc. AAAI (2011)
3. Brooks, M., Amershi, S., Lee, B., Drucker, S.M., Kapoor, A., Simard, P.: Featureinsight: Visual support for error-driven feature ideation in text classification. In: Proc. VAST (2015)

4. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science (2011)
5. Chang, J.C., Amershi, S., Kamar, E.: Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In: Proc. CHI (2017)
6. Hancock, B., Varma, P., Wang, S., Bringmann, M., Liang, P., Ré, C.: Training classifiers with natural language explanations. In: Proc. ACL (2018)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR (2016)
8. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalanidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (2016)
9. Kulesza, T., Amershi, S., Caruana, R., Fisher, D., Charles, D.: Structured labeling to facilitate concept evolution in machine learning. In: Proc. CHI (2014)
10. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proc. IUI (2015)
11. Lake, B.M., Ullman, T.D., Joshua B. Tenenbaum, S.J.G.: Building machines that learn and think like people. In: Proc. Behavioral and Brain Sciences (2016)
12. Laput, G., Lasecki, W.S., Wiese, J., Xiao, R., Bigham, J.P., Harrison, C.: Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In: Proc. CHI (2015)
13. Miller, G.A.: Wordnet: A lexical database for English. Communications of the ACM (1995)
14. Park, D.H., Hendricks, L.A., Akata, Z., Schiele, B., Darrell, T., Rohrbach, M.: Attentive explanations: Justifying decisions and pointing to the evidence. In: arXiv:1612.04757 (2016)
15. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. Proc. VLDB **11**(3) (2017)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. NIPS (2015)
17. Saon, G., Kuo, H.K.J., Rennie, S., Picheny, M.: The IBM 2015 English conversational telephone speech recognition system. In: arXiv:1505.05899 (2015)
18. Simard, P.Y., Amershi, S., Chickering, D.M., Pelton, A.E., Ghorashi, S., Meek, C., Ramos, G., Suh, J., Verwey, J., Wang, M., Wernsing, J.: Machine teaching: A new paradigm for building machine learning systems. In: arXiv:1707.06742 (2017)
19. Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., Herlocker, J.: Toward harnessing user feedback for machine learning. In: Proc. IUI (2007)
20. Stumpf, S., Rajaram, V., Li, L., Wong, W.K., Burnett, M., Dietterich, T., Sullivan, E., Herlocker, J.: Interacting meaningfully with machine learning systems: Three experiments. International Journal of Human-Computer Studies (2009)
21. Wang, M., Lu, Z., Zhou, J., Liu, Q.: Deep neural machine translation with linear associative unit. In: arXiv:1705.00861 (2017)