Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models

SWATI MISHRA, Cornell University, USA
JEFFREY M RZESZOTARSKI, Cornell University, USA

An important challenge in building explainable artificially intelligent (AI) systems is designing interpretable explanations. AI models often use low-level data features which may be hard for humans to interpret. Recent research suggests that situating machine decisions in abstract, human understandable concepts can help. However, it is challenging to determine the right level of conceptual mapping. In this research, we explore granularity (of data features) and context (of data instances) as dimensions underpinning conceptual mappings. Based on these measures, we explore strategies for designing explanations in classification models. We introduce an end-to-end concept elicitation pipeline that supports gathering high-level concepts for a given data set. Through crowd-sourced experiments, we examine how providing conceptual information shapes the effectiveness of explanations, finding that a balance between coarse and fine-grained explanations help users better estimate model predictions. We organize our findings into systematic themes that can inform design considerations for future systems.

CCS Concepts: • **Human-centered computing** → **Empirical studies in visualization**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Concepts, Machine Learning, Explanations, Classification

ACM Reference Format:
Swati Mishra and Jeffrey M Rzeszotarski. 2021. Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 139 (April 2021), 26 pages. https://doi.org/10.1145/3449213

1 INTRODUCTION

In order to work efficiently with their human counterparts in critical decision-making activities such as healthcare [29], finance [58], and security [40] while remaining transparent and accountable, Machine Leaning (ML) models should not only be accurate but also must explain their decisions effectively. This has motivated recent research in *explainable AI*[27] (xAI), with one main thread focusing on visualizing model properties or data features that are crucial in completing the task. For instance, when identifying objects in satellite imagery [2] using Convolutional Neural Networks [63], which have complex mathematical representations, an explanation system may highlight the pixels or properties (composition, position, etc.) that were used by the model. In this approach, low-level data features inferred from the model structure form the *language* of the explanation. On the other hand, human users do not reason based on low-level features. Instead, they rely on higher-level ones (green grasslands, vegetation, livestock, etc.) to make sense of the scene. As a result of this mismatch, explanations focusing on low-level features can be hard to interpret [3].

Authors' addresses: Swati Mishra, Cornell University, USA, swati@infosci.cornell.edu; Jeffrey M Rzeszotarski, Cornell University, USA, jeffrz@cornell.edu.

2573-0142/2021/4-ART139 $15.00
https://doi.org/10.1145/3449213

ML explanations that utilize low-level data features such as pixels, n-grams, and data attributes are advantageous because they are mathematically faithful to the model, providing extremely detailed, fine-grained, information about model activity (e.g. Grad-CAM [60] showing pixel response). However, this *granularity* can be overwhelming and it requires the analyst to make inferences in order to interpret the explanation in meaningful ways. Highly granular explanations lack overarching structures that help to structure an analyst's interpretation. For instance, pixel importance values highlighted in Figure 1(v) show a model's focus on the *striped area* of the image, but do not offer any concrete evidence that *stripes* were the true differentiating feature.

Additionally, these explanations are often specific to a particular data instance and may lack information about how other similar data points were treated by the model. A lack of sufficient, relevant *context* may also lead to misinterpretations. For instance, in Figure 1(v), it is important for the analyst to see examples which were interpreted *similarly* by the model so that they can verify that stripes are indeed a deciding factor. Yet, providing extraneous information increases processing time and might confound rather than reinforce analyst observations.

As an alternative to granular, low-level model explanations, researchers have proposed techniques for constructing explanations that use higher-level concepts that are familiar to users [61, 74], with an intuition that familiar concepts may be more usable. For example, techniques can summarize properties of regions across a corpus of images [34] or phrases in text [47](e.g. 'stripes' for zebra as shown in Figure 1(i)). However, this process is not lossless. Concept features are abstract in comparison to highly granular feedback, risking potentially hiding outliers and reducing specificity to exact model features (which could lead to user interpretation errors or bias [65]). A trade-off exists between providing model-faithful, highly granular explanations and simplified, conceptual explanations that might be more interpretable and bridge to contextual information.

In this paper, we explore how granularity, faithfulness to model features, context, and simplification affect user understanding and confidence in machine learning models. In particular, we focus on hybridized, human-centered explanation strategies which are faithful to machine's learned representations along with contextual information (presented through high-level concepts). We explore how these explanation strategies apply to image classification using Convolutional Neural Networks [31] and affect user assessments. We select this task because of its inherent complexity resulting from images' varied interpretations and ambiguities, and its applications to a wide range of scenarios that remain challenging to the ML community. In the discussion, we identify how our findings can apply to a wider range of applications. We evaluate a spectrum of explanation strategies by gathering conceptual information for a model's training data, exploring ways to visualize this information to explain model behavior, and conducting a controlled study to examine different levels of contextual information and granularity, exposing potential trade-offs in explanation design. The main contributions of this research are as follows:

(1) Formulation of explanation designs at varying levels of *granularity* (ranging from pixel-level to concept-level) and *context* (ranging from information about single instance to information about a local neighborhood of samples).
(2) A crowdsourcing pipeline to gather a rich data set of conceptual information that can power explanations based on above formulations.
(3) An empirical study that measures users' confidence on estimating model performance when these explanations are presented in an application context.

Our findings suggest that, while providing context helps to improve the interpretability of an explanation, fine-grained explanations do not yield the best results. Further, these attributes offer different levels of improvement in interpretability depending on model outcome (false positives versus true positives). We suggest that well-tuned explanations which a) align their granularity

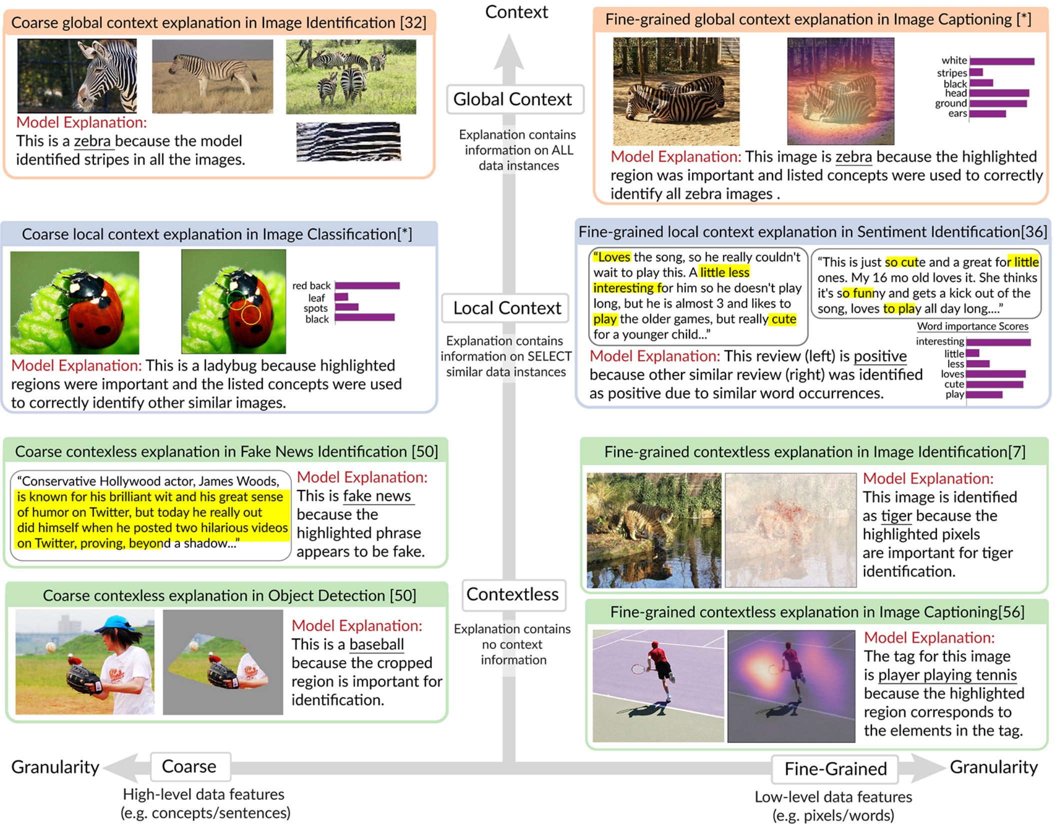

Context
Coarse global context explanation in Image Identification [32]
Model Explanation:
This is a zebra because the model identified stripes in all the images.
Global Context
Explanation contains information on ALL data instances
Fine-grained global context explanation in Image Captioning [*]
white
stripes
black
head
ground
ears
Model Explanation: This image is zebra because the highlighted region was important and listed concepts were used to correctly identify all zebra images .
Coarse local context explanation in Image Classification[*]
red back
leaf
spots
black
Model Explanation: This is a ladybug because highlighted regions were important and the listed concepts were used to correctly identify other similar images.
Local Context
Explanation contains information on SELECT similar data instances
Fine-grained local context explanation in Sentiment Identification[36]
"Loves the song, so he really couldn't wait to play this. A little less interesting for him so he doesn't play long, but he is almost 3 and likes to play the older games, but really cute for a younger child..."
"This is just so cute and a great for little ones. My 16 mo old loves it. She thinks it's so funny and gets a kick out of the song, loves to play all day long...."
Word importance Scores
interesting
little
less
loves
cute
play
Model Explanation: This review (left) is positive because other similar review (right) was identified as positive due to similar word occurrences.
Coarse contexless explanation in Fake News Identification [50]
"Conservative Hollywood actor, James Woods, is known for his brilliant wit and his great sense of humor on Twitter, but today he really out did himself when he posted two hilarious videos on Twitter, proving, beyond a shadow..."
Model Explanation:
This is fake news because the highlighted phrase appears to be fake.
Fine-grained contextless explanation in Image Identification[7]
Model Explanation:
This image is identified as tiger because the highlighted pixels are important for tiger identification.
Coarse contexless explanation in Object Detection [50]
Model Explanation:
This is a baseball because the cropped region is important for identification.
Contextless
Explanation contains no context information
Fine-grained contextless explanation in Image Captioning[56]
Model Explanation:
The tag for this image is player playing tennis because the highlighted region corresponds to the elements in the tag.
Granularity
Coarse
High-level data features (e.g. concepts/sentences)
Fine-Grained
Granularity
Low-level data features (e.g. pixels/words)

Fig. 1. Explanation space can vary in both context (information about similar data instance) and granularity (data features used). Context can vary from local to global, while granularity varies from coarse to fine-grained. Shown above are examples of explanations techniques for (a) image identification [7, 34], (b) sentiment identification [39], (c) object detection [53], (d) fake news identification [53], (e) image captioning [66] and (f) prototype explanations generated using our crowd sourcing methodology(*). All the above tasks can be framed as classification tasks with image and text data.

with the size of the most important features in performing the task, and b) provide context that balances effort and ability to test inferences can offer the best potential payoff for analysts.

2 CONTEXT AND GRANULARITY IN MACHINE EXPLANATIONS

In order to successfully assist users in debugging, auditing, or making sense of ML models, explanations must reflect machine behavior reliably (i.e. fidelity) and should be comprehensible by users (i.e. usability). While machines process images solely based on feature computation, human adjudicators, on the other hand, make sense of images using both bottom-up (photons to brain response) and top-down (knowledge shaping perception) processes [45]. This makes it challenging to design usable visual explanations of model behavior that remain faithful to the underlying mathematics. Explanations based on low-level features, such as pixel-level [7] and region-level [60] (Figure 1v) importance scores prioritize accuracy through absolute fidelity to model features, presenting a single data instance thoroughly. However, they may be difficult to interpret generally, because these *granular*, highly detailed explanations (which rely on low-level data features) lack grounding

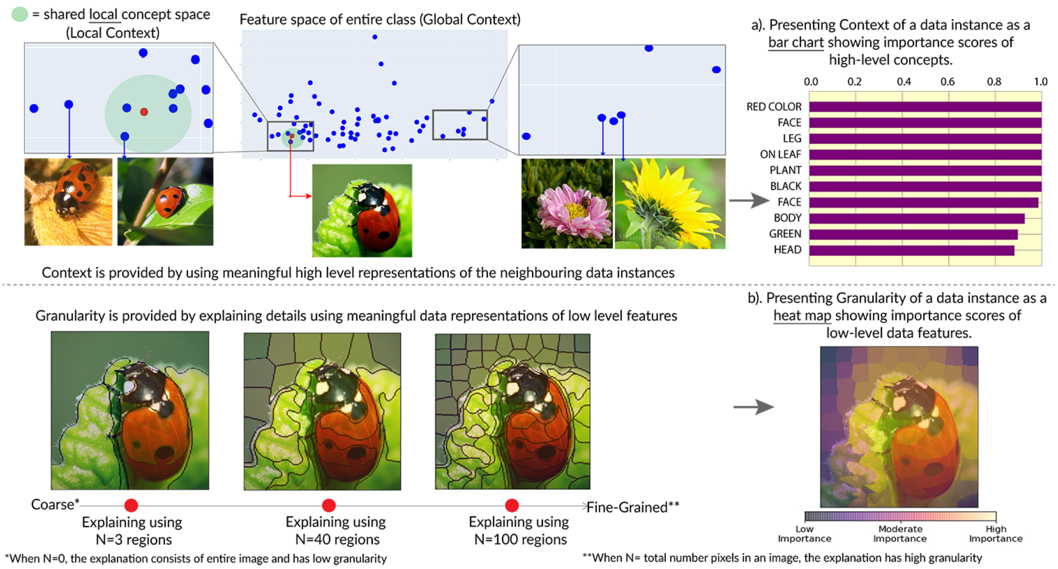

= shared local concept space
(Local Context)
Feature space of entire class (Global Context)
a). Presenting Context of a data instance as a bar chart showing importance scores of high-level concepts.
0.0
0.2
0.4
0.6
0.8
1.0
RED COLOR
FACE
LEG
ON LEAF
PLANT
BLACK
FACE
BODY
GREEN
HEAD
Context is provided by using meaningful high level representations of the neighbouring data instances
Granularity is provided by explaining details using meaningful data representations of low level features
b). Presenting Granularity of a data instance as a heat map showing importance scores of low-level data features.
Coarse*
Explaining using N=3 regions
Explaining using N=40 regions
Explaining using N=100 regions
Fine-Grained**
Low Importance
Moderate Importance
High Importance
*When N=0, the explanation consists of entire image and has low granularity
**When N= total number pixels in an image, the explanation has high granularity

Fig. 2. Context is determined by summarizing data features that were important to the model to correctly classify select neighbouring instances (*local*) or all instances (*global*). Granularity is determined by using data features through low-level pixels (*fine-grained*) or high-level regions (*coarse*). Highest granularity in this case is where all pixels are used, and lowest level granularity is when no pixel importance is shown.

in human familiar concepts, making them potentially abstract and reducing the avenues analysts have to employ their own domain knowledge. This absence of implicit guidance also makes it difficult to identify patterns when examining many instances (e.g. given heat-maps of two different tiger images, it is hard to conclude whether the model responds to stripes in both). On the other hand, explanations based on high-level, conceptual features (e.g. a visual summary showing stripes along with a natural language summary, "stripes were used to identify the image as tiger") may be easier to comprehend and provide a general understanding of the behavior. Further, a high-level explanation can also offer analysts a *contextualized* understanding of how a model identifies a group of like images (e.g. "most images *like this one* were classified using stripes"). Contextual information, which naturally lends itself to these high level explanations, may vary based on its scope, for instance incorporating closely similar data instances or data instances of the entire class. In this paper, we explore how variations in the explanatory structure (or *granularity*) and contextual information provided (or *context*) in model explanations can be used to improve user's confidence on the model and their assessment of its activities. In the following section we define granularity and context for the purpose of exploring the design space of possible explanations.

2.1 Granularity

Granularity of explanations refers to the *level of detail provided by the data features used to explain the decision*. Low-level data features (e.g. visualizing weights of each pixel that is important for classification in Figure 2(b)) provide a high degree of detail with regard to models' mathematical representations, generating a *fine-grained* explanation. A low granularity explanation, on the other hand, is comprised of high-level data features that offer more specificity (e.g specific well defined regions showing the tail of the tiger), and thus can be referred to as *coarse* in nature. When varied on a scale (Figure 1), different levels of granularity lead to detailed or abstract explanations.

There is risk in assuming that the more detailed an explanation, the better. A recent study [46] measured how factors like the length of the explanation, introduction of new concepts and repeating input terms impacted how users perceived explanations in a text classification task. The results suggest that highly detailed explanations may not necessarily improve user confidence. While the detail in fine-grained explanations risks high effort and forcing inferences on ambiguous but detailed feedback, coarse explanations demand greater inference and interpretation in order to apply the feedback, posing risks of bias (both analyst and algorithmic) and ambiguous or poorly generalized inferences.

In this paper, we build on existing work by generating 4 different explanations of an image classification task, having varying levels of granularity ranging between coarse to fine-grained, by systematically varying the number of segments (N) used in the explanation. At the highest level of detail, we present an activation heatmap of model responses over the image sample using all pixels. As granularity decreases, we segment the image into meaningful regions, presenting increasingly schematic representations of machine response. We use segmentation as a tool for varying granularity as in the case of image classification. We find that different amounts of segmentation match well to different scopes of image features. While a large number of segments readily unitize small details, a coarse amount of segments draws attention to larger features in general.

2.2 Context

Context in an explanation refers to the *breadth of data instances used to explain the model's general behavior, and the method used to display any additional instances.* Information regarding how the model processes similar data instances provides some context to the explanations, helping analysts to verify their inferences for generalizability and identify subtle patterns across samples. Different amounts of data can be sampled using similarity metrics and other filtering techniques, producing explanations that include varying levels of context. Context-less explanations present information about a single instance of training or test data. Much as in high granularity cases, these presentations are closest to the raw behavior of the model, with minimal summarization over sample space. A *local context* explanation presents information about an instance's local neighborhood (e.g. showing information about images that are closely similar to one ladybug image, Figure 2a) and a *global context* explanation presents summary information about all data points that belong to that class (also see 1i). Recent studies have explored how instance-level explanations might actually do little to improve an analyst's understanding of a model [3] as compared to global explanations of a model's behavior across classes of samples [53]. By incorporating neighborhood information using higher-level concepts, an explanation might provide a bird's eye view of the model's performance that is more usable and actionable.

In this research, we consider 3 different explanation strategies providing varying levels of context by constructing neighborhoods of related image samples at increasing scale and presenting information about them. We consider *context-less* (no neighborhood data instance information is presented), *local context* (information from only select highly similar samples is presented), and *global-context* (information from all similar data instances is presented) explanations in our study. We present contextual information to the users using a bar chart visualizing importance scores of features used by the model (see Figure 6). Explanations tell the analyst what kind of features neighboring samples presented (either very similar points or all points belonging to the class). These summaries of model explanations use concepts that are elicited through a crowdsourcing technique we describe in the later sections. Conceptual data about images act as a bridge between samples that are semantically similar but visually dissimilar. This dimension operates separately from granularity. Furthermore, a concept-based approach to presenting context is more stable in scenarios where images are semantically similar but have visual elements that generate significant dissimilarity scores unrelated to semantic differences when compared using statistical image similarity metrics.

In general, the dimension of context will be highly influenced by the kind of display chosen to represent the contextual information. We chose to use a bar chart metaphor because it provided a separate, distinct visual summary, with the trade-off that specific instances are not shown. We valued the schematic view this provided and the way it encouraged participants reason on a conceptual level. Another way to surface these examples would be to show them as additional samples of interest. We hope to investigate additional displays in future studies.

One might also argue that contextual information can be *derived* by examining several low-level feature based explanations in the local neighborhood (such as, a model's response on similar images). However, this approach can only work when the explanation is low fidelity (coarse granularity), less prone to open interpretation and is easily distinguishable across all data samples. For instance, when investigating a model that uses facial features to identify emotions, the user should be able to confidently judge that the region of importance being shown in all samples is eye and not forehead.

2.3 Hypotheses

Through an experimental study we aimed to explore the trade-offs mentioned for granularity (e.g. fidelity vs. ease of use) and context (e.g. generalizability vs. effort) and expose any potential interactions between them. We use several different outcome measures. First, we ask participants to report their satisfaction with the model's explanation and their confidence in the model's overall performance. Second, we ask participants to "predict the prediction" of the model, based on the explanation it provides. This measure has been used in prior work [49] as a proxy for participants' ability to understand the model and the predictions it makes. We compute the difference between the participants' predicted confidence and the actual output confidence of the model to measure the efficacy of the explanations. A low difference implies accurate estimation of model performance. Based on these measures, we hypothesize:

- **H1**: **Coarse granularity should help participants estimate model predictions more accurately**. This is because these explanations are low in fidelity and should align better with participants' concepts.
- **H2**: Likewise, **higher degrees of context should help participants estimate the model predictions more accurately**, as these explanations will give them more comparison points for making an informed judgment.
- **H3**: Considering self-report measures, **coarse granularity explanations will result in higher self-reported participant confidence**, as their simplified feedback will be easier for participants to interpret.
- **H4**: **For poor quality data instances, there will be higher differences in the participant's ability to estimate predictions and their self-reported confidence**, as compared to high quality, reliably classified ones because universally high confidence (or skepticism) of the model's performance may be disrupted by samples that are difficult to classify.
- **H5**: **Explanations incorporating local and global context will report higher participant confidence on model decisions**, as compared to context-less explanations.
- **H6**: **Presence of context in high granularity explanations will have stronger effect on participant's confidence than weaker ones**, because providing context ought to help participants place a sample into a broader space of data.

3 RELATED WORK

Some machine learning (ML) classification models have structural features that are inherently explainable such as Decision Tree splits [52] or Logistic Regression [51] coefficients, while others require post-hoc explanation techniques to extract meaningful representations from the models' complex architectures (e.g. Convolutional Neural Networks (CNNs) [31]). Some of these post-hoc techniques rely on visualizing the model's internal operations, for instance, visualizing layer-wise activations in a neural network [70]. A comprehensive survey of visualization techniques for CNNs can be found in [71]. They vary widely, including layer-by-layer views, overlays on data samples, diagrams of information flow, and depictions of changes during training. However, these explanations often require that the user has sufficient knowledge of model architecture and hence can be difficult to interpret by ML non-experts. As model explanations reach broader audiences (both because ML is increasingly being integrated into everyday life [20] and there is increasing societal focus on model accountability [24]), researchers have explored new, more accessible techniques for explaining complex models.

One way the research community has explored for making more accessible explanations is by utilizing causal relationships between input instances and the model's response. For instance, techniques such as identifying images [25] or parts of images (using occlusion) [62] that maximally activate hidden regions of the model. These techniques allow users to *estimate* regions of importance to the model at a broader level. On the other hand, heatmap visualizations produced via CAM [73] and Grad-CAM[60] algorithms generate more fine-grained visual explanations by *explicitly* providing importance scores at a per-pixel and region-based levels. Other techniques like saliency maps generated by layer-wise relevance propagation (LRP) [7] also explicitly visualize pixel-level information about features that have high relevancy scores, providing a sense of the most important features belonging to a predicted class. These approaches prove very useful in explaining the results of a single instance of data on a very fine-grained, *granular* level. However, they do not connect to broader information required for a model's performance evaluations and auditing (e.g. determining whether a sample is exceptional or typical) and their fine-grained, pixel-level feedback may be noisy and hard to interpret, as there is an assumption that human-meaningful features will emerge from analysis of a performant model, which may not be guaranteed.

One thread of research [72] attempts to add structure by assigning meaningful labels to the individual units (or layers) of deep learning models, providing a reference point for adjudicators to compare their visual semantic interpretations. In this work Zhou et al. use labels belonging to colors, materials, textures, parts, objects and scenes. Explanations for single data instances are a combination of these concept labels. A similar approach is adopted by IBD [74], where parts and object labels are accompanied by heatmaps [23] to explain the predictions of various models. These explanations provide more schematic feedback, presenting outlines of regions that were important. This reduces the granularity of the feedback, with potential benefits in ease of understanding. Approaches like TCAV [34] follow a parallel, low-granularity strategy, generating concepts to explain an entire class of data instances. A user can access model performance based on whether or not it employs a specified concept in identifying a given class (e.g. one can investigate whether "red" was used in the identification of "firetrucks"). This approach moves beyond instance-level explanation and only employ high contextual information to measure a model's performance for a given class. While these approaches improve the explanations of a model's decisions, it is unclear to what degree we should make use of conceptual information in explanations. Is the gain in clarity offered by these techniques sufficient to exceed the potential loss in specificity? Should we present one concept or a combination of concepts? Do these techniques work best on a sample-by-sample level or with class-level context? In our lab study, we investigate these questions.

In this paper we make use of crowdsourcing as a tool for gathering human conceptual data. Crowdsourcing platforms find application in many ML tasks such as dataset generation [8, 22, 30, 48]. In particular, crowdsourcing can be leveraged in places where human adjudication outperforms that of a machine [11], as is the case with the kind of conceptual data we seek. However, quality and work structure might limit achievable results [35]. While gathering ground truth image classification data on a crowdsourcing platform is relatively straightforward, gathering human conceptual and attentional data may be much more challenging because tasks must allow for high expressabilty on the part of workers (so that concepts are faithfully gathered) but also maintain high consistency and work quality. Despite these challenges, researchers have gathered large-scale datasets that incorporate information about how individuals perceive and make sense of stimuli. Salicon [32] is a large-scale dataset containing information about user attention as inferred from mouse movement. While attention is a bit different from concept evocation, this provides evidence of the feasibility of gaining access to the cognition of participants on crowdsourcing platforms. Further work has investigated how behavioral logging [57], a game with a purpose [17], and blurring techniques [15] can provide reliable information about user's cognitive state. While in this work we do not directly seek attentional data, we do need workers to reliably report where they are directing their attention when conceptualizing an object. We also make use of traditional majority vote quality control schemes [8] as data validation in our workflow.

Just as important as it is to design new explanation techniques for ML models, it is also necessary to understand how best to evaluate explanations. Generally speaking, the goal of an explanation is to give its intended users the *rationale* behind the model's decision making process. There are a variety of strategies for evaluating this. Some metrics include users' trust and confidence [28], their satisfaction [6], and verbalizations of their mental representations of the model [36]. Expertise often plays a role, as an explanation designed for an ML expert may not make sense for the ML novices [43]. Doshi-Velez and Kim categorize evaluations as being application-grounded, human-grounded and functionally-grounded based on the type of subjects and tasks [19]. We employ human-grounded metrics to evaluate how user confidence in the model's performance is impacted by explanation design using a *predict the prediction* task [50]. We also include measures of user confidence and satisfaction in our evaluation. While, we do not explicitly measure trust, prior work [34, 49, 53] suggests that humans tend to trust ML models if they find model explanations reliable.

We specifically focus on CNN based image classification tasks due to their widespread applications [5] and the unique challenge images pose for explanations. Though image classification of animals and objects, as examined in this paper, may seem trivial, the methods can be transferred to practical applications like agricultural auditing [56], and facial recognition [64]. Selecting a neutral problem of image classification with Imagenet dataset, provides a better guarantee that our experimental participants will possess necessary domain expertise to participate and also avoid potential confounds from socially or culturally sensitive tasks.

4 ELICITING CONCEPTUAL DATA FROM THE CROWD

To construct our explanation samples, we use conceptual features to summarize model performance across a variety of data instances. We use concepts as a means to connect across samples and model behavior characteristics through meaningful shared properties. A concept is a mental representation of classes or categories of things which help individuals reason about the world [45]. For instance, conceptual schema help people to understand what precisely makes an entity a bird and to describe how a single bird is different or similar to other birds. Because they are universally employed by individuals, and form an important part of our understanding of the real world [9], they have been considered a prime target for making more usable machine explanations [42]. These mental representations, however, vary widely across people (as a result of, but not limited to, experiential, cultural, and social factors). Common theory holds that there is no such thing as an ideal prototype of a class of things [54], so humans reason about classes using a mixture of different conceptual elements. In light of this, we must gather a wide sample of human judgments in order to identify common patterns among individuals. We then use these concepts to produce explanations of varying levels of granularity and context.

Prior work on concept based explanations for deep learning models [4, 34, 74] provides several approaches for connecting pieces of ground truth from different datasets to generate descriptive labels. For instance, in [4], densely annotated datasets like Pascal-Part[12], Pascal-Context[44] and Describable Textures[14] were used to assemble a dataset with pixel-level labels of each image describing colors, materials, textures, parts, objects and scenes. Adopting this approach or borrowing from existing datasets is promising but has several limitations. The underlying assumption in assembling descriptive datasets from others is that these label categories can correctly represent higher-level concepts used by humans. If the data set we are exploring does not have well defined categories representing human conceptualizations of entities (perhaps because it is appropriating an existing schema rather than starting from base human responses), then this approach would not work. Likewise, the labels available in different datasets form an exhaustive, but shallow vocabulary for conceptual data. Because single class membership and assigning the most accurate label are prioritized, the size and heterogeneity of labels is limited (e.g. while a feature detector may help decompose a face into [eyes, nose, mouth, forehead], it may miss that many human adjudicators think of "hairline" and "lips" when they conceptualize a face, limiting expressability). Finally, it should be feasible to assemble pixel-level annotations from existing samples in different datasets, which is often not tractable when borrowing/applying features across different models.

For this reason, instead of inferring conceptual data for explanations from existing datasets, we designed a crowdsourcing pipeline to sample conceptual information directly from humans in their own terms, providing unadulterated feedback on how humans perceive categories of images. It also gives us flexibility for gathering data that will be useful in varying context and granularity.

4.1 Crowdsourcing Approach

Our elicitation process begins with a source corpus of images from which seeds are selected for concept generation. For the purposes of this work, we selected 10 random classes from ImageNet [16], an image recognition dataset commonly used in the ML community. We make use of Mobilenet [31], one of the many performant models trained on Imagenet dataset, owing to its lightweight architecture (4,253,864 parameters, 88 layers), and high performance (top-5 accuracy of 89.5 percent). Using this dataset–model pair offers simplicity in tasks ("what is in this image?") with familiar object categories (e.g. animals, stationary, sporting goods), and moderate accuracy ensures that false positives will be generated which are equally useful to evaluate explanations. In the discussion, we revisit this consideration and examine alternative applications.

4.1.1 Selecting Seed Concept Images. We selected a corpus of 431 images belonging to 10 different ImageNet classes. We then sampled 146 *seed concept images* in total (M=14 seed images per class) from each of the selected classes. As random selection of these seed concept images risks cornering and may not guarantee even coverage of all patterns present in class (e.g. there maybe an image subset where the black dots on a ladybug may not be clearly visible), we employed a clustering approach. This approach is based on the notion of feature similarity commonly used in computer vision applications. Images that contain similar objects have similar mathematically derived feature vectors. We first extracted the feature vectors of all images using MobileNet model as feature extractor and then performed K-Means clustering [18] over the vectors. We specified K (number of clusters=6) to be small owing to the small size of our dataset (M=43 images per class). Finally, we

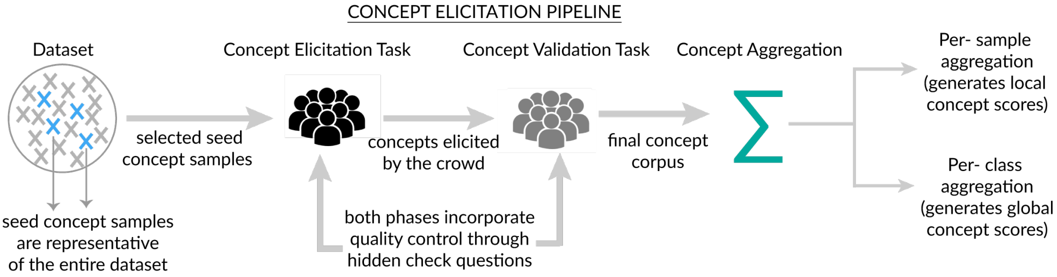
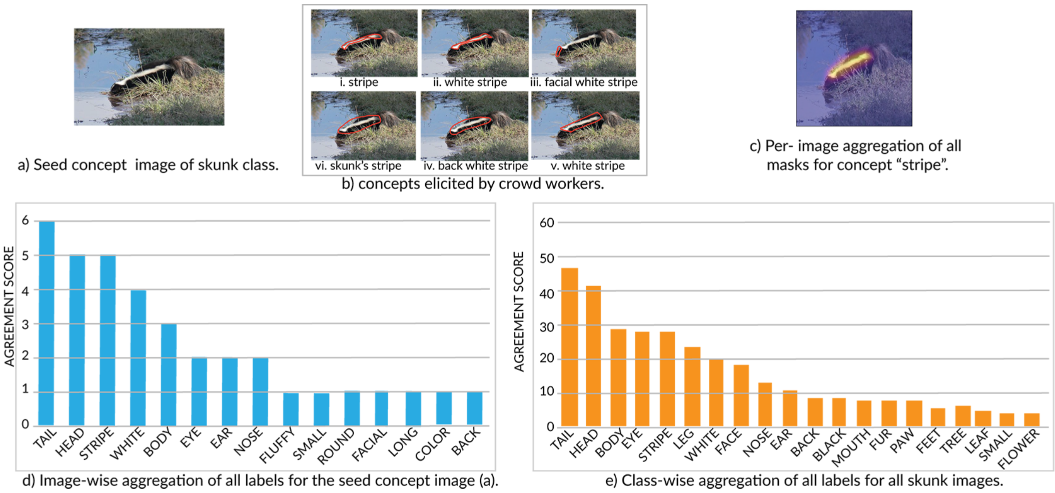

CONCEPT ELICITATION PIPELINE
Dataset
Concept Elicitation Task
Concept Validation Task
Concept Aggregation
Per- sample aggregation (generates local concept scores)
Per- class aggregation (generates global concept scores)
selected seed concept samples
concepts elicited by the crowd
final concept corpus
seed concept samples are representative of the entire dataset
both phases incorporate quality control through hidden check questions

Fig. 3. The concept elicitation pipeline. Groups of crowdworkers generate and validate concepts.

sampled the most central feature vectors from each cluster to be the seed concept image representing that cluster. One limitation of this approach is that it is not effective when the feature space is organized into sparse clusters. To overcome this limitation, we selected more than one seed concept image from such classes. However, in future work we intend to refine our sampling technique.

4.1.2 Task Design. Each concept in our final dataset is a mask – label pair, connecting a selection mask defining a region of pixels to a corresponding label (Figure 4b). We designed a web-based interface for Amazon Mechanical Turk (AMT) which allowed participants to draw on an image using a free-form lasso tool and then "tag" the region with a concept label. Our task first showed users the ground truth label of the image and asked them to draw an outline around the whole entity. This functioned both as a way to train users to use the lasso tool and to validate that they were attentive and understood the task. We then prompted participants to draw an outline around the most important part of the image they used to determine that it was a member of the ground truth class (e.g. to identify a ladybug, users could highlight its spotted wing case and label it accordingly). We avoided showing users any list of possible concepts, drop-down menus, or pre-defined regions in order to avoid them from being primed. This strategy helped us extract human mental models free from external influence. In the future, one might imagine strategies that make use of different layers of elicitation, drawing for example on literature in crowd brainstorming [10] and taxonomy construction [13]. Following this, we prompted users to draw an additional 4 outlines for other features they used to identify that the entity was a member of its class. A thumbnail preview reminded users of the parts they had already submitted. An informal pilot suggested that requesting more than 5 parts resulted in diminishing returns in useful information and data quality. We chose this design pattern to force users to first attend to the image and then declare a concept. By grounding them in the particular sample, we make certain that the concepts provided are bound to actual features in the image.

4.1.3 Quality Control. Even with attention checks [35] and qualifications, not all contributions received were of high quality or submitted in good faith. Manual inspection of our data indicated that about 14 percent of submissions were of poor quality. As a result, we introduced additional quality validation on a worker- and concept-level. In this process, we gathered our corpus of approximately 9000 concepts (mask-label pairs) generated by 309 unique workers and reached out to a new, distinct group of 322 workers to rate concepts based on their relevance to a class, and the relevance of the label to the specific image mask. Again, to validate that the responses produced by the workers of the quality validation task were not random series of clicks, in each bundle of rating tasks we added one task with obviously poor answers – concepts were described using random keyboard characters 'bgdg','sdsdf', etc. We only accepted work that successfully rated these check



a) Seed concept image of skunk class.

b) concepts elicited by crowd workers.

c) Per- image aggregation of all masks for concept "stripe".

d) Image-wise aggregation of all labels for the seed concept image (a).

e) Class-wise aggregation of all labels for all skunk images.

Fig. 4. Per-image and per-class aggregation to generate local and global context information a) The seed concept image. b) The concept of "stripe" as produced by 6 different crowd workers (out of 309 unique workers) showing concept masks and labels for a data sample. c) Per-image aggregation of all masks. d) 15 unique concept labels produced by the 6 workers. e) 20 most frequently used concept labels (out of 75 unique labels) for the "skunk" class. Agreement score shows the number of crowdworkers who gave this label.

tasks as irrelevant (1-point). We further rejected all work by the workers who universally submitted poor quality work and blocked them from further submissions.

4.1.4 Results. We deployed tasks on AMT, iterating over all 146 images in our dataset. Each image was presented to 10 unique workers. We collected all the coordinates corresponding to each drawn image mask and their concept label. For concept elicitation, each worker was compensated \$0.35. For the quality control task, where each worker rated the quality of the concept, we compensated \$0.06 and each task was also completed by 10 workers. Wages were set based on task completion times (as recorded in pilot tests) to achieve a \$15 per hour wage. A total of 309 unique workers participated in the crowdsourced elicitation phase of the study, and 322 unique workers participated in the validation phase. We describe the concept aggregation process in the subsequent sections.

4.2 Aggregating Concepts

The concept elicitation phase produced 7,379 submissions (approximately M=720 concepts per class) with all 146 seed concepts images described by (M=45) mask-label pairs from 309 unique workers. We systematically aggregated this data in multiple ways in order to create an explanation corpus on a per-image and per-class level. Our goal was to construct an ordered set of regions paired with appropriate labels describing salient concepts representing an image and a class. We then use these to construct a variety of concept-driven explanations at different levels of granularity and contextual content.

Our concept corpus consisted of 10 classes (*airplane, baseball, guitar, koala, ladybug, perfume, racket, skunk, tiger, zebra*). Each class contained P seed concept images, encoding Q concepts. Each of the Q concepts were one of the image mask (M) and label (L) pairs from the crowdsourced elicitation stage. We define a concept for a given seed image, S, as C_q = (M_q, L_q), where M_q is the image mask and L_q is its corresponding label for the qth concept of every seed image (where, q∈1,2,...,Q). The goal of aggregation is to find the most salient concepts people use to describe a sample and an entire class.

The set of concepts at a per-image level is the aggregation of all unique concepts C_q that are most frequently agreed upon by the crowdworkers. We aggregate using the following steps 1 through 4:

(1) For a given seed concept image, we extract all labels $\in L_q$, generated by the crowd.
(2) Using stemming [67] and Wu-Palmer word similarity metric [69], we grouped all similar words under 1 representative label. For instance, labels like ‘stripe’, ‘stripes’, ‘stripe of skunk’ and ‘skunks stripes’ were all grouped together under 1 label-‘stripe’.
(3) We sample the top-10 most frequently occurring labels to create a final resulting set L.
(4) Finally, we filter all the masks $\in M_q$ described by using the aggregated labels $\in L$ and then perform mask aggregation by assigning a value 1 to all pixel locations in M, then performing score summation followed by normalization. This gives an aggregate mask M_q of pixel locations described by L (see Figure 4a-c).

We also aggregate over concept labels that represents an entire class. These concepts, however, are only described by labels L_q and not by image masks M_q, because, while aggregating highlighted regions for each image is relatively straightforward, it is more challenging (and less meaningful) across different images. We identify the most frequently occurring labels across different classes. The final concept set C is extracted using steps 1-3 as described in previous section. We found that workers did agree on common concepts. For instance, the most common concepts used to describe a skunk were ’tail’, ’head’, ’stripe’, ’white’ and ’body’ (unordered). This technique might be further improved by using word similarity algorithms such as [26] or ontologies to nest and combine similar words like, leg, foot, etc. (one could also create nested concepts such as head contains face, face contains mouth,etc.). The concepts at both the per-image and per-class level are used to provide local and global context information to users in our intervention.

5 DESIGNING EXPLANATORY PROTOTYPES

Our final dataset after elicitation and aggregation consisted of 146 images belonging to 10 classes; each image had an aggregated pool of concepts that were represented by labels and associated aggregate image regions (that matched with the concept labels, Figure 4). Each class consisted of an average of 14 seed-concept images (that represented most of the images present in that class), and each seed-concept image was represented by the top-5 *dominant concepts* (most frequently used concepts) from its aggregated pool. We design our experimental explanations by making use of different portions of our aggregated data. One part of our explanation visualizes the model’s behavior on a per-image level using traditional activation *heatmaps* segmented into meaningful regions to achieve varying levels of granularity. We then make use of per-image and per-class information to provide increasing amounts of detail about neighboring points as context using a *bar chart.* Combined, each of these hybrid explanations represent one condition in our experiment permuted over the levels of *granularity* and *context* we support.

5.1 Presenting Context

We implement *context* by presenting information about neighborhoods of related image samples at increasing scale using a bar chart. We hypothesize that increasing amounts of context will help users judge whether the performance of the model on one sample generalizes, and how it may or may not be taking advantage of unique features of a class (e.g. skunk stripes). We relate samples based on shared machine activation, conceptual connections, and class. Conceptual data about images can act as a bridge between samples that are semantically similar but visually dissimilar. This dimension operates orthogonally to granularity, as we can add contextual information regardless of the features used to present information about individual samples.

It is hard to meaningfully connect pixel regions across multiple images. As a result, we make use of the aggregated concept labels. We construct scores which represent the importance attributed by the model to regions defined by the dominant concepts of the class prediction. For instance, given an image classified as "racket," if the dominant concept for the class "racket" is "net" (as defined by the aggregated corpus), then the concept score of "net" would be the maximum score assigned by the algorithm [60] to the region associated with the label "net" for that instance of "racket".

To calculate the concept-based score of an image $\mathrm{I_{w,h}}$ (of width w and height h), we employ the coarse localisation map $\mathrm{L_{w,h}}$ of the image made earlier. Given that $\mathrm{C_j}$ belongs to the set of concepts that can be used to describe the concepts of $\mathrm{I_{w,h}}$, then $\mathrm{A_j}$ is the image highlighting the aggregated regions associated with each concept $\mathrm{C_j}$, and $\mathrm{B_j}$ is the associated label. We set the limit of j to be the number of concepts used to describe an image $\mathrm{I_{w,h}}$ (in this case J=5). We can create an aggregate masked localization map $\mathrm{M_j = T(A,L_{w,h})}$, where T is an image masking operation defined by,

$$\mathrm{T_{x,y}} = \begin{cases} \mathrm{s_{x,y}}, & \text{if } \mathrm{a_{x,y}} \geq \text{thresh} \\ 0, & \text{otherwise} \end{cases}$$

Here $\mathrm{s_{x,y}}$ is the activation score of pixel at (x,y) in $\mathrm{L_{w,h}}$ and $\mathrm{a_{x,y}}$ is the value of pixel at (x,y) in A, the crowd-supplied regions. We set the threshold value to be $w/2$, where w is the number of workers per image, so that we only consider regions of the image where the majority of workers agree. Finally, we compute concept scores for a given concept by finding the maximum model-assigned value for that region. Therefore, for image $\mathrm{I_{w,h}}$, concept score $\mathrm{CS_I = max(m_{x,y})}$, where $\mathrm{m_{x,y} \in M_J}$.

We make use of these concept scores to generate several different kinds of explanations where they were presented to the user using a *bar chart* alongside the decision (see Figure 2a). Focusing on the neighborhood of images that share similar conceptual or feature vector properties to the current sample, we provide summary views (e.g. graphing the prevalence of different concepts) of similar samples (local context) or the prevalence of concepts in the entire class (global context). Class-level or model-level, provide a broader comparison with which the user can distinguish typical or unusual classifications. Though this goes beyond the scope of our initial investigation, one might also introduce concept label weights and model accuracy to further refine the context.

The aforementioned approaches, however, are only possible if a $\mathrm{C_j}$ exists for a given $\mathrm{I_{w,h}}$ − one can only derive concept scores $\mathrm{CS_I}$ for images that were seed concept images in the first place. This limits the scalability of the concept score calculation process. If we do not have data for a specific image, we cannot construct the necessary aggregated concepts. Given machine learning corpora can contain tens of thousands of samples, we must resolve this limitation without requiring an immense amount of crowdsourced labor to cover all points.

5.2 Scalability through Concept Neighbourhoods

As previously mentioned, one thread of research on concepts, theorizes that individuals combine several representations within a shared feature space to build an understanding about a class of objects [45]. These schematic representations (or prototypes, depending on school of thought) help to identify new stimuli and categorize them conceptually. We apply a similar technique to scale up our concept data to larger corpora.

We assert that data instances that share the same model feature space (numerical vectors derived from a data sample) are also likely to share same concept space. Data instances that are closer in the feature space are likely to share more concepts than the ones that are further apart (Figure 2a). In effect, the neighborhoods of data samples we created earlier in order to identify seed images also describe conceptually similar feature spaces. Based on this notion, we use the feature space neighborhoods as a proxy for conceptual neighborhoods used by human adjudicators.

We use our image neighborhoods from the elicitation process to derive a concept for a newly introduced image $I_{w,h}$ where $I_{w,h}$. To identify which concepts are closest to this new image, we create a neighborhood score NS_I that captures the concepts of neighboring seed-concept images that have labeled data. To compute NS_I, we first extract the features of all images belonging to a class using our model as feature extractor. We reduce the dimensionality of these features (from 1024 to 2) using Principle Component Analysis [68] and then compute the neighbours of each data point using nearest neighbour algorithms [37] in the reduced space. Finally, we calculate a threshold value K. For a new image I, all other data points that fall under the radius of distance ≤ K, are considered to be neighbours. If a seed concept image S exists within this cluster, then the concepts for S can be borrowed and used to describe $I_{w,h}$. However, if the there is no labeled seed within the distance ≤ K, then the segmented image features can be compared to feature segments of other labeled images in the corpus to locate the closest labeled concepts.

5.3 Presenting Granularity

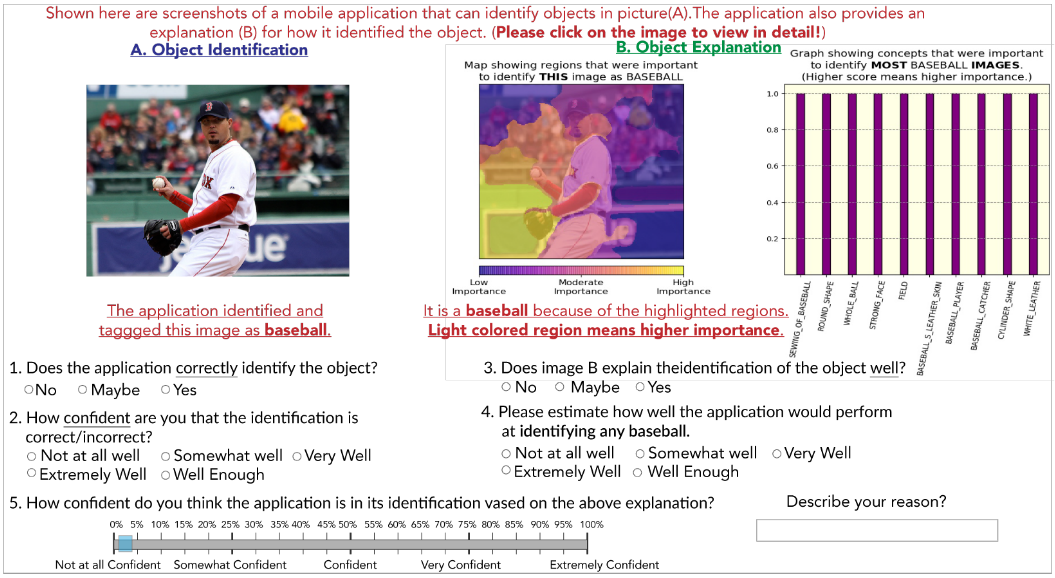
In our explanations, we implement *granularity* by varying the level of detail of our presentation of instance-level (or per-image) model behavior, motivated by the idea that reducing the level of detail will help individuals better parse information that is important to the model. At the highest level of detail we present a standard activation map of model responses over the image sample showing importance scores assigned to all pixels. As granularity decreases and explanations become more coarse, we segment the image into regions that present meaningful object parts, presenting increasingly schematic representations of machine response.

We begin our explanations of model behavior by applying a traditional approach for visualizing the model activity using a *heatmap*. To calculate the region-based scores of an image $I_{w,h}$ of width *w* and height *h*, we first use the Grad-CAM[60] algorithm which maps the output features of the model to the input features based on predicted class. The algorithm produces a coarse localization map $L_{w,h}$ in which, each pixel is assigned a score $s_{x,y}$ between 0 and 1. Pixels belonging to the region of high importance are assigned higher scores than the pixels belonging to the region of low importance to the model. This provides us with a high granularity baseline visualization. Using a perceptually uniform color scale, we display scores as an overlay on the image.

To create coarser (less granular) explanations, we segment the image into N regions which vary based on granularity target using the SLIC segmentation technique [1]. We then compute an importance score S_n for each region R_n, by averaging the corresponding scores $s_{x,y}$ of all pixels belonging to the region R_n. This brings about the trade-off mentioned earlier in the paper. As N decreases, the heatmap is simplified with larger outlined regions of solid color. This potentially occludes small outliers with the benefit of highlighting larger, feature-based and concept oriented trends. The interpretation of these regions is still left to the user. For instance, the region of high importance for an image classified as "ladybug" may appear to belong to the "red and black dots" of the ladybug or to the "leaf" or any other part (see Figure 2b). We use the same heatmap technique and color scale to show the scores of segmented regions. At the very lowest level, we eschew the heatmap completely in favor of circle annotations around the centroid of the N highest activating regions in a low segmentation map, pointing the user directly to salient features at the cost of hiding pixel-level details of model feature vectors.

6 EVALUATION

In the previous sections, we outlined the general process we used to gather conceptual data and construct explanation strategies for an image classification model. We created a series of prototypes



Shown here are screenshots of a mobile application that can identify objects in picture(A).The application also provides an explanation (B) for how it identified the object. (**Please click on the image to view in detail!**)

A. Object Identification

B. Object Explanation

Map showing regions that were important to identify **THIS** image as BASEBALL

Graph showing concepts that were important to identify **MOST** BASEBALL **IMAGES**. (Higher score means higher importance.)

The application identified and taggged this image as **baseball**.

It is a **baseball** because of the highlighted regions. **Light colored region means higher importance.**

1. Does the application correctly identify the object?
○No ○ Maybe ○ Yes

2. How confident are you that the identification is correct/incorrect?
○ Not at all well ○ Somewhat well ○ Very Well
○ Extremely Well ○ Well Enough

3. Does image B explain theidentification of the object well?
○ No ○ Maybe ○ Yes

4. Please estimate how well the application would perform at **identifying any baseball.**
○ Not at all well ○ Somewhat well ○ Very Well
○ Extremely Well ○ Well Enough

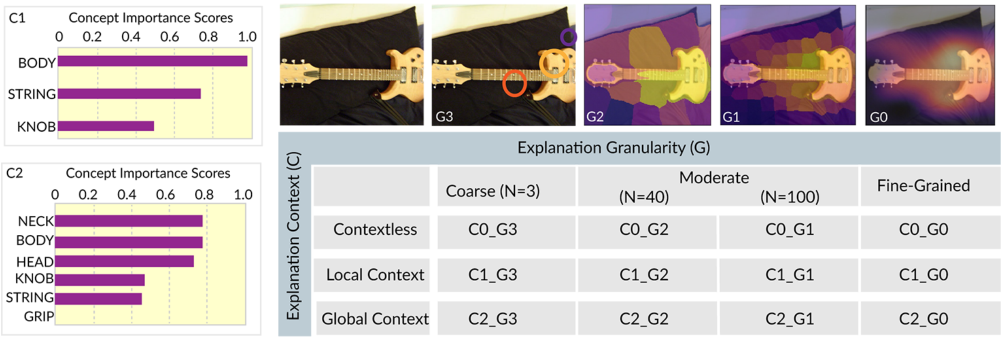
5. How confident do you think the application is in its identification vased on the above explanation?

Describe your reason?

Fig. 5. The evaluation interface (in the global context and moderate granularity condition C2_G2). Participants can click on the images to zoom.

that present model behavior for an instance at different levels of simplification and different ways of integrating conceptual data to provide context about a sample's neighborhood of points as *granularity* and *context*. We designed a crowdsourced study exploring concept-based explanations across a variety of granularity and context levels, investigating the potential benefits and costs of these approaches in terms of participants' ability to understand the model and their confidence in its reliability. Participants were presented with image samples from our dataset and their predicted class. Depending on their experimental condition, they received a specific explanatory visualization describing why the model made a classification decision.

Evaluation Task: In our evaluation task, the users were presented with a stimulus, the model response on that stimulus, and its corresponding explanation (Fig. 5). Participants provided several responses using an explanation. In the first response, they were asked to rate their *confidence* in the model's reported identification (the classification results) on a 5-point scale. Participants then report *their* level of confidence that the model will reliably classify any instance of the class *in the future* on another 5-point scale. In the third response, we asked users to estimate how confident *the model* was in making its prediction using a percentage range scale, following a "predict-the-prediction" pattern from previous work in this area [50]. We realize that it is challenging to predict this numeric value with absolute confidence and were concerned that participants would be overwhelmed by a numeric entry widget or continuous slider. To avoid choice overload and to simplify the interface, we discretized a continuous slider ranging from 0% to 100% with markers for each 5% increment of confidence. We felt that encouraging users to "bin" accuracy rather than estimate a precise number might better match how individuals in our pilot investigations thought of model accuracy (e.g. "very low", "neutral", "somewhat high"). We also provided them with a similar scale that helps them report their confidence on model performance (see question 5 in Figure 5). Users were also required to provide written feedback about why they chose this percentage. This helped us to capture the nuances of user experience while isolating responses that were obviously fraudulent.



Explanation Granularity (G)				
	Coarse (N=3)	Moderate (N=40)	 (N=100)	Fine-Grained
Contextless	C0_G3	C0_G2	C0_G1	C0_G0
Local Context	C1_G3	C1_G2	C1_G1	C1_G0
Global Context	C2_G3	C2_G2	C2_G1	C2_G0

Fig. 6. Experimental conditions evaluated in the study across varying levels of granularity and context of explanation. Bar charts represent the variations in context, while heatmaps represent variations in granularity.

While a number of constructs for trust and confidence exist in sociology and psychology [38, 41, 55], the ML community continues to wrestle with defining and measuring participant trust in models. Prior research [34, 50] has shown how evaluating the quality of explanations independent of application setting can yield beneficial results, taking the point of view that self-reported confidence in the model's activity is a reasonable proxy for trust.

Some researchers [59] further argue that interpretability is also a proxy for trust. For the purposes of this paper, we adopt the convention that confidence and interpretability are proxies for participant trust, but we acknowledge and emphasize that this does not necessarily reflect the full sociotechnical implications of trust. To operationalize our measure, we used a "predict the prediction" task, common in the literature, that measures how explanations help users better estimate model performance.

Participants: We deployed this study over AMT on randomly selected 91 images from our set (of 146 images), each image was evaluated by 6 unique workers for each condition. The tasks were assigned to workers in random order, and the workers were paid a wage of $0.03 for 1 image task (based on a $15 hourly wage). Participants were shown both true and false positives examples. For instance, if the input image was a ladybug and the predicted class was not a ladybug, the users still estimated how confident the model was in its prediction. We initially experimented with an additional condition where we intentionally supplied participants with false labels. In practice, this was unsuccessful as model confidence on these samples was universally low (due to high overall model accuracy) and the ambiguous examples in the dataset (e.g. a liger) could have posed a greater challenge to users.

Experimental Conditions: Our study has two independent categorical conditions that are permuted. We set 4 granularity conditions based on the implementation described in the previous section and set 3 context conditions. These have been shown in Figure 6. The granularity conditions range between *coarse* and *fine-grained* while the context conditions are *contextless*,*local* and *global*. This provides 12 individual combinations of the two conditions, all of which are evaluated in our experiment. We evaluate the explanations belonging to 7 random classes (*ladybug, koala, tiger, guitar, baseball, racket, zebra*) used in concept elicitation phase selecting total of 91 data samples and request 6 unique participant submissions for each sample.

7 RESULTS

We gathered a total of 6,552 responses (91 images X 12 conditions X 6 workers) from 302 unique participants on Amazon Mechanical Turk (our concept elicitation phase was comprised of 309 separate workers and the concept validation phase was comprised of 322 separate workers). We eliminated 414 invalid or fraudulent responses from the final analysis using the check question. To compensate for the resulting imbalance in response counts across each experimental condition, we randomly dropped an additional 121 data points in total from conditions that had an excess of 500 responses. The final dataset contains approximately 500 responses across each experimental condition (mean=501.41, SD=2.09) generated by a group of 302 participants.

As described in Section 2, had the following hypotheses about our study conditions:

- **H1**: Low granularity should help participants estimate predictions more accurately.
- **H2**: Likewise, high context should help participants estimate predictions more accurately.
- **H3**: Low granularity explanations will result in higher self-reported participant confidence.
- **H4**: For poor quality data instances, there will be higher differences in the participant's accuracy of estimated predictions and self-reported confidence.
- **H5**: Explanations incorporating local and global context will report higher participant confidence on model decisions.
- **H6**: Presence of context in high granularity explanations will have stronger effect on participant's confidence than weaker ones.

7.1 Predict the Prediction

In the *predict the prediction* task, we asked participants to estimate how confident they thought the model was in its predicted outcome based on the provided explanation. The intuition here is that "better" explanations will help participants to estimate model confidence more accurately. For analysis, in order to connect model's predicted confidence with participants' slider that had values spaced every 5%, we rounded model's predicted confidence scores to the nearest 5. We then compared the absolute value of the difference between the prediction confidence estimated by the users and the actual prediction confidence of the model for a given data instance across conditions. Participants were relatively accurate at predicting confidence but with high variance (M=22.24, SD=22.65). We observed that the mean error was lowest under the experimental condition with medium granularity and high context (C2_G1 (M=20.33, SD=20.04), closely followed by C2_G0 (M=20.60,SD=21.70).

We conducted a linear regression to identify whether these differences were meaningful, treating the conditions as categories rather than as numeric or ordinal attributes. Taking C0 (no context) and G0 (high granularity) as a baseline, we found that C1 ($\beta = -0.10$, $p = 0.48$) and G1 ($\beta = -0.01$, $p = 0.97$) did not differ substantially. However, C2 performance was markedly better ($\beta = -0.47$, $p < 0.001^*$). G2 performed marginally worse ($\beta = 0.30$, $p = 0.06$.) and G3 significantly so ($\beta = 1.49$, $p < 0.001^*$). This runs counter to our initial expectations. Our results do not support hypothesis H1 – higher granularity performed better. However, providing conceptually grounded contextual feedback proved increasingly helpful, confirming hypothesis H2. We did not detect any interaction effects (finding no evidence for H6 for the case of predicting the prediction). Examining results with respect to hypothesis H4, we interacted true positive vs false positive with both the granularity and context conditions in our model. We find that across the board, correct examples were predicted with much more accuracy ($\beta = -1.92$, $p < 0.0001^*$), confirming the hypothesis. As true positive examples were often at or near ceiling in terms of machine confidence, this strong effect might be the result of a ceiling at 100% or participants' improved ability to recognize good exemplars. More interestingly, a main effect remains for G3, under-performing in comparison to higher granularities ($\beta = 0.73$, $p < 0.015^*$). We observe interactions in both C2 (marginal; $\beta = -0.52$, $p = 0.09$.) and G3 ($\beta = 1.08$, $p < 0.01^*$). In both cases the explanations perform slightly *worse* for good examples. It is unclear whether this is a result of the explanations providing less useful information when examples are good or if they exhibit some additional property that disrupts their effectiveness.

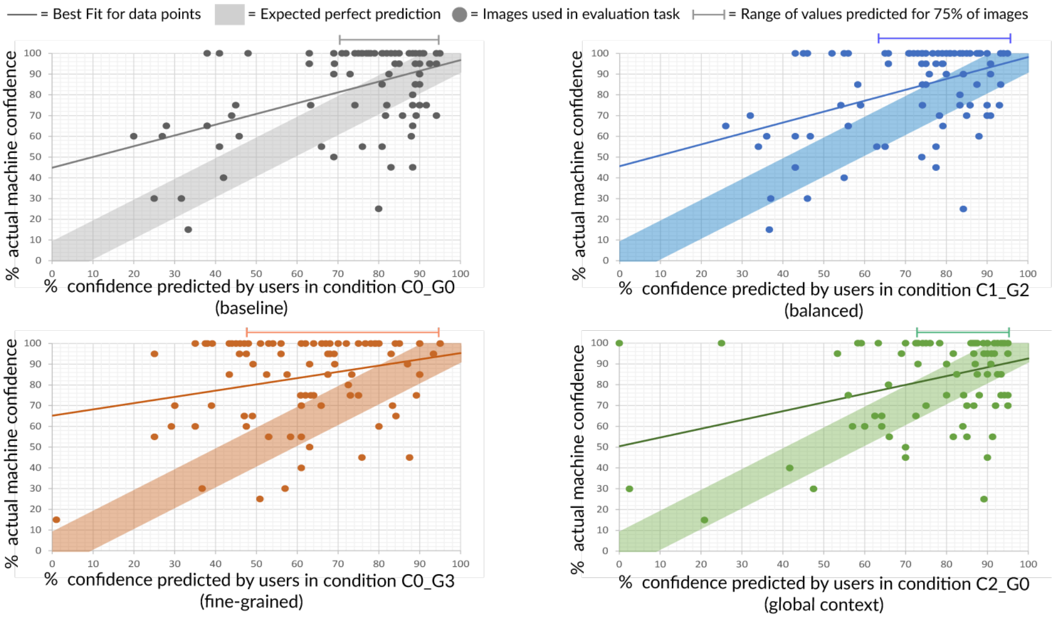


Fig. 7. Summary of results for predict the prediction task under experimental conditions C0_G0 (baseline), C1_G2 (local context and medium granularity), C0_G3 (high granularity) and C2_G0 (global context). Each point in scatter plot represents the machine confidence as predicted by the user (x-axis) and the actual machine confidence on the image (y-axis). The perfect prediction line is an area in this case because participant and model scores were rounded to nearest 5% value.

7.2 Confidence in Model Identification

Participants rated their overall confidence in the performance of the model in classifying the image sample based on the explanation they viewed. On average, participants reported higher *confidence* for accurately classified images (M=4.66,SD=0.79) compared to false positives (M=4.38,SD=1.12), as confirmed by a Kruskal-Wallis test on the Likert responses ($\chi^2 = 87.2$, $p < 0.001$*). In further analyses, we did not detect any interaction effects with correctness. Examining how granularity affected participants' confidence in the model's classification of the sample, we notice that the highest granularity condition outperformed all of the others (M=4.63,SD=0.84). A Kruskal-Wallis test identified significant differences ($\chi^2 = 9.29$, $p = 0.025$*), and pairwise Wilcoxon rank sum tests indicated that G0 was significantly different from all other conditions (p: G0-G1=0.037, G0-G2=0.008, G0-G3=0.008). This finding suggests the opposite of our initial hypothesis, H3. We hypothesized that lower granularities would be clearer to the user and therefore lead to higher transparency which might increase confidence. Our finding that G0 outperformed suggests that perhaps providing more detail gives users a heightened sense of control over the model, leading to more (over)confidence. This finding is especially salient for system designers, as it suggests there may be cases where providing high amounts of detail lead to confidence that may not bear out in terms of participant performance (though in our case G0 seemed to perform well). In terms of providing context, the mean confidence for C2, the global conceptual context condition was somewhat higher than its peers, as indicated marginally by a Kruskal-Wallis test ($\chi^2 = 4.72$, $p = 0.094$.). Pairwise tests were similarly borderline (p: C0-C2=0.076, C1-C2=0.046). As a result, we have some suggestive evidence

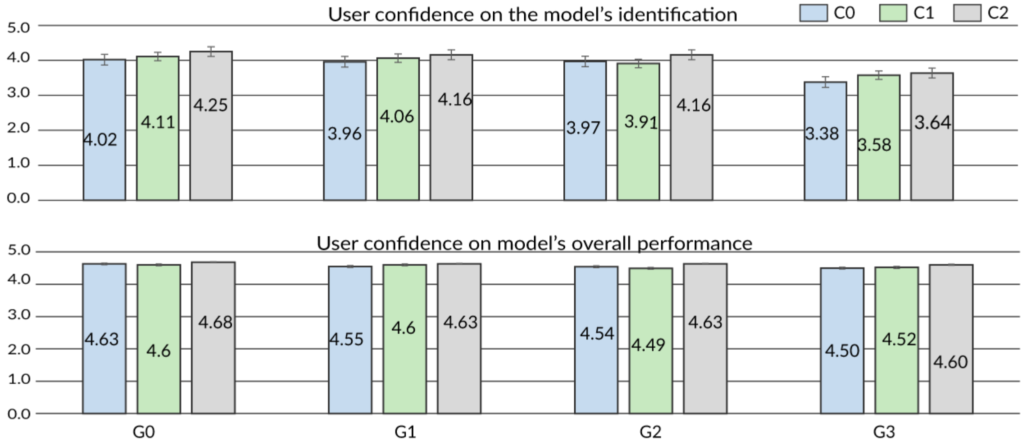


Fig. 8. Distribution of user's confidence on model's identification and overall performance on a 5-point scale.

that proves hypothesis H5, but cannot confirm it. We observed no interactions between the two independent measures.

7.3 Confidence in Overall Model Performance

Participants also reported their confidence in the model's performance as a whole (as opposed to a specific instance). We see similar responses as in the per-image confidence. As a whole confidence decreased as granularity decreased (perhaps due to the summary's very low information density) (χ^2 = 165.4, p < 0.001*; p: G0-G1=0.13, G0-G2=0.017, G0-G3<0.001, G1-G2=0.38, G1-G3<0.001, G2-G3<0.001). We also observed an identical trend in the context condition, where ratings improved as more conceptual context was provided (χ^2 = 26.8, p < 0.001*; p: C0-C1=0.064, C0-C2<0.001, C1-C2<0.001).

7.4 Summary

We observed that both users' ability to predict predictions as well as their confidence along two self-report measures were improved by the presence of conceptually grounded context feedback in the explanations (conditions C1 and C2). This provides positive evidence for hypotheses H2 and H5. Interestingly, we find evidence of an effect opposite to hypotheses H1 and H3 in terms of explanation granularity. We observe that higher granularity explanations resulted in high prediction accuracy and high confidence, even though they may pose a greater cognitive load in terms of understanding and interpreting. We do not note any interaction effects – indicating that context and granularity offer independent benefits/costs for participants (leaving H6 unsupported). However, with regards to the true/false positive status of samples, we found that the benefits of context and high granularity were muted for correct examples, possibly due to a ceiling effect.

8 DISCUSSION

In this section we call attention to several broader themes and implications for designing explanations that effectively leverage context and granularity across different tasks and data types.

Identifying and presenting the most useful contextual information: Our study demonstrated that model explanations which provide contextual information (grounded in human-friendly conceptual language) help to make models more interpretable. Providing additional information about model behavior on similar data instances is beneficial. However, contextual information is not universally informative. While local contextual information provides insights into how similar data instances are treated by the model, global context provides information about the variations across data instances. For instance, in the task of analyzing a person's social media posts to identify symptoms of anxiety, an analyst might want to ask why the model thinks that a particular set of posts indicate symptoms of anxiety. Local contextual information, in this scenario, would be how similar posts were classified and can reveal expressions suggesting specific moments of anxiety. On the other hand, global context (explaining how model classified all the posts) can help the analysts identify expressions shared across different anxiety events signifying a prolonged anxious state. While both of these features are potentially of interest to analysts, their utility often depends on specific tasks.

Context helps, but it ought to be presented in a way that leads to actionable information on the part of the analyst (e.g. auditing that all kinds of anxiety are well-interpreted versus spot-checking a poorly performing set of instances). In the case of our work, providing context using conceptual summaries of features, the data lent themselves to simple, conceptually grounded explanatory visualizations. This might not always be the case, as in the text example it may be challenging to represent textual nuances deeper than term frequency in a succinct visualization. Designers leveraging our findings must consider what sorts of concepts (or other human-friendly features that cut across data instances) are present in their data and use case, and tailor contextual feedback to best present them. Interactivity may also help here, giving users an ability to explore across different kinds of contextual data. In this case, where *both* local and global context are intertwined, it is important to quantify what concepts are local to an instance or global to a given class. Additionally, all contextual explanations require some degree of similarity computation which may introduce bias or error in the process, necessitating transparency about how and why context was provided.

Fine-tuning granularity in explanations: Participants expressed high confidence in models for conditions which presented pixel-by-pixel heatmaps, even if they require effort to process. In our study, participants also performed reasonably well in predicting performance using these explanations. Counter to our initial expectations, participants were able to readily interpret such explanations without much training. On the other hand, we observed that both self-reported trust and predicted performance accuracy were diminished when the explanations provided coarse, well-structured feedback. For instance, in condition G3, users were presented with precisely labeled regions of importance without any ambiguity. Their self-reported confidence ratings and prediction accuracy were low, perhaps due to the sacrifice in detail and ambiguities that emerge from such simple, abstract feedback. This is also indicated in participants' qualitative reports that they anticipated that the model was hiding something. One key take-away from our work, then, is that a very coarse granularity explanation with too much structure and organization might have a negative effect on user trust and performance, despite the simplicity and quick processing it offers. This aligns with prior work on user perceptions of newsfeeds and recommender systems that might deliver the "best" results but do so in a way that seems opaque to end users [21].

Considering the broad space of granularity, we found that a 'moderate' amount (G1) did no worse than the pixel-by-pixel condition. This Goldilocks Zone may have the advantageous properties we initially cited for coarser granularity (e.g. less effort, lower training, simplicity) and incorporate higher user confidence on explanations that seemed more complete or detailed. However, the four conditions for granularity in our study only provided an upper bound on 'too much' simplification, and not a precise set-point. Considering that explanations can take any number of forms based on use case and input data, we do not believe there will be one universally preferable option. Instead, we propose several design considerations for tuning granularity. First, explanations should prioritize adding granularity when possible, as it can reduce burdens on users and eliminate cases where analysts fixate on details around model resolution. Second, explanations ought to tune their granularity to the specific task analysts are performing. If an analyst is auditing class-level response for a satellite image recognition model, it may be appropriate to set granularity to target clusters that represent "forest," "street," and "lake." On the other hand, if an analyst is inspecting a recidivism prediction model for possible biases, it may not be appropriate to represent the model response for all attributes and instead highlighting specific sensitive attributes would be more actionable. In both of these cases, the granularity is set to a level of detail that resolves the features that will be of most use to the analyst. Finally, when possible, explanations should allow analysts to tune its granularity. This would assist in cases where analysts are doing exploratory work or have tasks that require them to balance factors across instances and classes.

Accounting for information imbalance in true and false positives: We noted in our study that participants' impression of the explanatory content and model performance was significantly reduced in the case of false positives as compared to true positives across different levels of granularity and context. While one would expect confidence in the model to decrease when participants see incorrect behavior, it was more surprising to see a similar drop in the quality of the explanation itself disrupting their ability to estimate model performance. Feedback such as *"[I rated it poorly] ...because of the low precision of the application. The regions are highlighted well but they could be more precise,"* and *"[I rated it poorly because the] ...graph shows that only mouth, legs, strape [sic] are important to identify the object but what about eyes,"* emphasize that participants expected the explanations to have more information when they thought the sample was a poor example of its class. This observation connects to the notion that users seek more information for an incorrect decision than for a correct decision both out of a need to confirm the error they see (versus their expectation that a system should answer correctly) and because incorrect decisions may inherently be borderline or harder to judge. It aligns with research [24] drawing insights from social sciences which highlights that rather than asking "why X," people usually seek information about "why Y instead of X." Negative affect may also spill over from one question to the next. Though we did not explicitly study this option, one potential solution to this issue is to provide specific context in cases where model performance is lacking. Through carefully chosen samples, an explanation might provide several 'Ys' in "why Y instead of X" or help to temper analyst expectations by always providing a buffering true positive example. Granularity might also be manipulated to help analysts weight different samples.

Explanations beyond supporting evidence: A well designed explanation also ought to draw attention to scenarios where one should not trust the decision (i.e. counterfactuals). Our study showed that despite the model classifying data instances correctly, participants reported lower confidence in the model's performance when the explanations did not align with their mental models. Their qualitative feedback further supported this argument. For instance, for an image showing only part of an entity, workers provided feedback such as, *"the application focused on the feet, pants and the lawn, but not the hand and the racket itself,"* and, *"The application identif[ied] object in wrong place,"* . This implies that the explanations encouraged users to think critically about the model's decisions, pointing to a tantalizing possibility for adversarial design in explanation systems. One might intentionally display erroneous model behavior in order to make the user more skeptical, much as generative adversarial networks can help analysts reason about model improvements [33]. More broadly, we hope to explore how explanations shape analyst perceptions of model behavior over time, both individually and for ML in general.

Generalizing to new data and tasks: While we focused on an image classification dataset throughout our work, our approach is extensible to other datatypes such as text. The notion of granularity still holds true for text-based explanations. For example, level of detail and specificity would be determined by the length and choice of words and phrases used by the machine to explain its decisions. The notion of a concept naturally comports to text data, potentially using keywords to summarize portions of text. One can also imagine replacing the lasso tool in our concept elicitation task with a simple highlighter to identify substrings that function as concept "regions". The approach would be more intuitive, as highlighting "important" pieces of information is a common technique used to comprehend a text. The notion of concept neighbourhood can be employed in its modified form, as the feature space of text documents may be difficult to quantify. However, techniques like topic modelling [75] can be used to categorize the information that conceptually associates two pieces of texts, and the explanation may comprise of higher level themes. There are also potential analogues for multivariate and categorical data input. For example, dimensionality reduction techniques can reduce the number of dimensions that an analyst must view in an explanation (granularity), and clustering techniques can provide simplified representations of groups of similar points as context. In this case, the "concepts" could either be encoded by the analyst as a part of their analysis (e.g. specifying sensitive attributes, providing filter criteria) or inferred a priori (e.g. bias detection, causality).

9 LIMITATIONS AND FUTURE WORK

This work has several limitations in terms of implementation, experimental design, and crowdsourcing. Firstly, we conducted our study with a small set of images. For real world analysis, we would like to conduct this study with a larger subset of images. While we introduce a technique for augmenting our dataset without recruiting orders of magnitude more microtask workers, we did not formally evaluate its scalability and performance with gold standard data. We also considered only one kind of model explanation method for extracting low-level feature details (Grad-CAM) and only one image classification model (Mobilenet). In the future we hope to evaluate explanations generated by other models. Of particular interest are controls such as a model that delivers random noise, one that is always wrong, and models that have subtle differences, for effective comparison.

Our concept elicitation pipeline also has several limitations. The users use free-form drawing to highlight regions of importance. While the technique captures unadulterated conceptual representations from the users, it lacks structure and some users may find it difficult to externalize their knowledge. One possible solution is to allow users to select pre-segmented regions of images to help kick-start the process. Another important aspect of the pipeline was language refinement. We allowed workers to use their own language and preferences to label. This was very useful for improving expressability (assuming the explosion of variations can be adequately managed), but may pose difficulties for people who do not have command over the language.

Finally, we evaluated two properties of explanations, granularity and context, with 4 and 3 levels respectively. We need to further investigate several intermediate levels of perception for in-depth investigation into these representations. There are other dimensions such as the type of visualization used (e.g. natural language, heatmap overlays, charts) and the amount of granularity included within a contextual visualization. Our exploration relied on self-report measures from participants as well, which may not be perfectly reliable and may experience environmental validity issues in a microtask market context. Furthermore, in-person lab studies (when in-person meetings are once again possible) might help to validate the methodology and provide qualitative observations.

10 CONCLUSION

In this paper we highlighted the need to investigate techniques to build interpretable explanations that are easy for users to understand. We propose a hybrid approach that employs low-level data features with high-level concepts, to present instance-level explanations of the data-instance and its neighbourhood. We describe and implement an end-to-end crowd sourcing pipeline that can be used to gather concepts and design meaningful, human-centered explanations. Finally, we evaluate how users' confidence on the model is impacted by explanations at varying levels of detail through a controlled study. We hope that our work helps to inform the growing body of literature on the design, efficacy, and usability of machine learning model explanations.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.

[2] Adrian Albert, Jasleen Kaur, and Marta C. Gonzalez. 2017. Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1357–1366. https://doi.org/10.1145/3097983.3098070

[3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3377325.3377519

[4] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HI, USA, 3319–3327. https://doi.org/10.1109/CVPR.2017.354

[5] Ashwin Bhandare, Maithili Bhide, Pranav Gokhale, and Rohan Chandavarkar. 2016. Applications of convolutional neural networks. *International Journal of Computer Science and Information Technologies* 7, 5 (2016), 2206–2215.

[6] Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces*, Vol. 5. IUI, San Diego, CA, 153. http://www.cs.utexas.edu/users/ai-lab?bilgic:iui-bp05

[7] Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-Wise Relevance Propagation for Deep Neural Network Architectures. In *Information Science and Applications (ICISA) 2016*, Kuinam J. Kim and Nikolai Joukov (Eds.). Springer Singapore, Singapore, 913–922.

[8] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Los Angeles, California) *(CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–12. http://dl.acm.org/citation.cfm?id=1866696.1866697

[9] Susan Carey. 2011. Précis of The Origin of Concepts. *Behavioral and Brain Sciences* 34, 3 (2011), 113–124. https://doi.org/10.1017/S0140525X10000919

[10] Joel Chan, Steven Dang, and Steven P. Dow. 2016. Improving Crowd Innovation with Expert Facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work &; Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1223–1235. https://doi.org/10.1145/2818048.2820023

[11] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. *Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets*. Association for Computing Machinery, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[12] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. 2014. Detect What You Can: Detecting and Representing Objects Using Holistic Models and Body Parts. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 1979–1986. https://doi.org/10.1109/CVPR.2014.254

[13] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. *Cascade: Crowdsourcing Taxonomy Creation*. Association for Computing Machinery, New York, NY, USA, 1999–2008. https://doi.org/10.1145/2470654.2466265

[14] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. 2014. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 3606–3613. https://doi.org/10.1109/CVPR.2014.461

[15] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human Attention in Visual Question Answering: Do Humans and Deep Networks look at the same regions?. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 932–937. https://doi.org/10.18653/v1/D16-1092

[16] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, USA, 248–255. https://doi.org/10.

1109/CVPR.2009.5206848

[17] J. Deng, J. Krause, and L. Fei-Fei. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Portland, OR, USA, 580–587. https://doi.org/10.1109/CVPR.2013.81

[18] Chris Ding and Xiaofeng He. 2004. <i>K</i>-Means Clustering via Principal Component Analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA, 29. https://doi.org/10.1145/1015330.1015408

[19] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

[20] Anthony Elliott. 2019. *The culture of AI: Everyday life and the digital revolution.* Routledge, Australia.

[21] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. *"I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning about Invisible Algorithms in News Feeds.* Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2702123.2702556

[22] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Los Angeles, California) *(CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 80–88. http://dl.acm.org/citation.cfm?id=1866696.1866709

[23] Nils Gehlenborg and Bang Wong. 2012. Points of view: heat maps.

[24] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, Turin, Italy, Italy, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 580–587. https://doi.org/10.1109/CVPR.2014.81

[26] Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68, 13 (2013), 13–18.

[27] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (Jun. 2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[28] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 164—-168. https://doi.org/10.1145/2856767.2856811

[29] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain?

[30] John Joseph Horton and Lydia B. Chilton. 2010. The Labor Economics of Paid Crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce* (Cambridge, Massachusetts, USA) *(EC '10)*. ACM, New York, NY, USA, 209–218. https://doi.org/10.1145/1807342.1807376

[31] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications.

[32] M. Jiang, S. Huang, J. Duan, and Q. Zhao. 2015. SALICON: Saliency in Context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 1072–1080. https://doi.org/10.1109/CVPR.2015.7298710

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

[34] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2668–2677. http://proceedings.mlr.press/v80/kim18d.html

[35] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[36] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678

[37] T. Liu, C. Rosenberg, and H. A. Rowley. 2007. Clustering Billions of Images with Large Scale Nearest Neighbor Search. In *2007 IEEE Workshop on Applications of Computer Vision (WACV '07)*. IEEE, Austin, TX, USA, 28–28. https:

//doi.org/10.1109/WACV.2007.18

[38] N. Luhmann, H. Davis, J. Raffan, K. Rooney, M. King, and C. Morgner. 1979. *Trust and Power.* Wiley, USA.

[39] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[40] Daniel L Marino, Chathurika S Wickramasinghe, and Milos Manic. 2018. An adversarial approach for explainable ai in intrusion detection systems. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, Washington, DC, USA, 3237–3243.

[41] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

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

[43] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. arXiv:cs.HC/1811.11839

[44] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. 2014. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, 891–898. https://doi.org/10.1109/CVPR.2014.119

[45] Gregory Murphy. 2004. *The big book of concepts*. MIT press, USA.

[46] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation.

[47] Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. https://doi.org/10.18653/v1/N18-1097

[48] David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation (AAAIWS'11-11)*. AAAI Press, San Francisco, USA, 43–48.

[49] Forough Poursabzi-Sangdeh, Dan Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. https://www.microsoft.com/en-us/research/publication/manipulating-and-measuring-model-interpretability/

[50] Forough Poursabzi-Sangdeh, Dan Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. https://www.microsoft.com/en-us/research/publication/manipulating-and-measuring-model-interpretability/

[51] Daryl Pregibon et al. 1981. Logistic regression diagnostics. *The Annals of Statistics* 9, 4 (1981), 705–724.

[52] J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1, 1 (1986), 81–106.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[54] Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology* 4, 3 (1973), 328–350.

[55] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review* 23, 3 (1998), 393–404.

[56] Helena Russello. 2018. Convolutional neural networks for crop yield prediction using satellite images.

[57] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2047196.2047199

[58] Swati Sachan, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. 2020. An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications* 144 (2020), 113100. https://doi.org/10.1016/j.eswa.2019.113100

[59] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. arXiv:cs.LG/1901.08558

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 618–626. https://doi.org/10.1109/ICCV.2017.74

[61] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh. 2019. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In *2019 IEEE/CVF International Conference on*

Computer Vision (ICCV). IEEE, Seoul, Korea (South), Korea (South), 2591–2600. https://doi.org/10.1109/ICCV.2019.00268

[62] K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps.

[63] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:cs.CV/1409.1556

[64] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2017. Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image Based CNNs. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 549–552. https://doi.org/10.1145/3136755.3143008

[65] Elisabeth Kersten van Dijk, Wijnand IJsselsteijn, and Joyce Westerink. 2016. Deceptive Visualizations and User Bias: A Case for Personalization and Ambiguity in PI Visualizations. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 588–593. https://doi.org/10.1145/2968219.2968326

[66] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[67] Peter Willett. 2006. The Porter stemming algorithm: Then and now. *Program electronic library and information systems* 40 (07 2006). https://doi.org/10.1108/00330330610681295

[68] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[69] Zhibiao Wu and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* (Las Cruces, New Mexico) *(ACL '94)*. Association for Computational Linguistics, USA, 133–138. https://doi.org/10.3115/981732.981751

[70] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 818–833.

[71] Quan-shi Zhang and Song-chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (Jan. 2018), 27–39. https://doi.org/10.1631/FITEE.1700808

[72] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. 2018. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2018), 2131–2145.

[73] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

[74] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable Basis Decomposition for Visual Explanation. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 122–138.

[75] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. paměťový nosič. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (Valletta, Malta). University of Malta, Valletta, Malta, 46–50. http://www.fi.muni.cz/usr/sojka/presentations/lrec2010-poster-rehurek-sojka.pdf

Received June 2020; revised October 2020; accepted December 2020