Not all Failure Modes are Created Equal: Training Deep Neural Networks for Explicable (Mis)Classification

Alberto Olmo* [1] **Sailik Sengupta*** [1] **Subbarao Kambhampati** [1]

Abstract

Deep Neural Networks are often brittle on image classification tasks and known to misclassify inputs. While these misclassifications may be inevitable, all failure modes cannot be considered equal. Certain misclassifications (eg. classifying the image of a dog to an airplane) can create surprise and result in the loss of human trust in the system. Even worse, certain errors (eg. a person misclassified as a primate) can have societal impacts. Thus, in this work, we aim to reduce inexplicable errors. To address this challenge, we first discuss how to obtain the class-level semantics that captures the human's expectation (M^h) regarding which classes are semantically close *vs.* ones that are far away. We show that for data-sets like CIFAR-10 and CIFAR-100, class-level semantics can be obtained by leveraging human subject studies (significantly inexpensive compared to existing works) and, whenever possible, by utilizing publicly available human-curated knowledge. Second, we propose the use of Weighted Loss Functions (WLFs) to penalize misclassifications by the weight of their inexplicability. Finally, we show that training (or even fine-tuning) existing classifiers with the two proposed methods lead to Deep Neural Networks that have (1) comparable top-1 accuracy, an important metric in operational contexts, (2) more explicable failure modes, (3) higher robustness to random and adversarial noise and (4) require significantly less cost in teams of additional human labels compared to existing works.

Figure 1. Examples showing that state-of-the-art neural networks exhibit different failure modes (on CIFAR-100 dataset) often resulting in inexplicable mistakes. These mistakes can cause surprise (leading to a loss of trust) at best and have societal impacts at worst (eg. classifying genders incorrectly, or dark-skinned people as gorillas (Vincent, 2020)).

1. Introduction

Over the last few years, Deep Neural Networks have proven to be effective in vision classification tasks. While researchers have invested effort in trying to make these networks interpretable (Montavon et al., 2018; Rudin, 2019; Li et al., 2018; Melis & Jaakkola, 2018), we still lack a good formal understanding of how they work internally, thereby making them questionable for everyday use in real-world systems. While mispredictions are bound to exist for any classifier that has less than cent percent accuracy, expecting a user to trust a classification system solely based on accuracy values is unreasonable. Indeed, not all failures have the same effect on a user; while some mistakes are acceptable, others can be deemed inexplicable, causing surprise and an eventual loss of human trust. Even worse, failure modes may often exacerbate the societal biases learned from data (eg. images of dark-skinned humans being misclassified as a gorilla (Vincent, 2020)).

We believe that egregious mistakes are a by-product of the existing loss/objective functions used by state-of-the-art classifiers; they are too sparse to encode meaningful information about failure modes. For example, the popular Categorical Cross-Entropy (CCE) loss encourages correct classification and penalizes all misclassifications equally. In this work, we argue that incorporating the human's expectation about the failure modes (M^h) into the classification system (M^r) can help us develop explicable classifiers whose failure modes are aligned with the user's expectations.

* Indicates equal contribution. Names ordered alphabetically. [1]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, Arizona, USA. Correspondence to: Alberto Olmo <aolmoher@asu.edu>.

Presented at the ICML 2020 Workshop on Human In The Loop Learning.

In this regard, we answer two questions– (1) how to represent and obtain expectations of a human (that captures the notion of egregious $vs.$ explicable misclassification) and (2) how to utilize such a representation to ensure that the trained classifier adheres to the human's expectation. To answer the first question, we posit that the notion of explicability can be represented as a semantic distance between the actual and the predicted label, i.e. misclassifications to classes semantically closer to the ground-truth are considered reasonable while misclassifications to classes further away (eg. classifying the image of a dog as an airplane) makes the end-user deem the classifier inexplicable. In particular, class-level semantic similarity can be leveraged to capture the human's expectation (M^h). To obtain M^h, we strongly advocate the use of a human labeling approach, and in cases where the classification task is generic and labels are difficult to obtain, we suggest leveraging existing linguistic knowledge-bases. Finally, to incorporate this notion of explicability into classifiers, we employ the idea of weighted loss functions to train classifiers.

We propose two different methods to obtain the semantic-similarity distances between the class labels and compare them to a popular baseline that gathers instance-specific human feedback. We demonstrate that our methods have better operational metrics such as top-1 accuracy and provide explicable failure modes while reducing the costs for additional human labeling. Further, training for explicable misclassifications increases the robustness of the classifier to both Gaussian and adversarial noise. When the classification task becomes large, we highlight that (1) methods that gather human labels for understanding the semantics over failure modes have to reason about the human's cognitive overload (ruling out existing baselines), and (2) account for training resource-intensive classifiers from scratch. We show that when these larger tasks are generic (eg. CIFAR-100, ImageNET), human labels can be obtained from existing databases to yield more explicable failure modes than existing classifiers (see Figure 1). Further, such results can be obtained by simply fine-tuning existing models, reducing the training effort. Finally, we discuss the use of our methods for addressing operational issues and calibrating societal impacts.

2. Related Works

Researchers have shown that deep neural networks demonstrate inexplicable behavior in the presence of out-of-distribution (Hendrycks et al., 2019; Hendrycks & Gimpel, 2016; Mallick et al., 2020) or adversarially perturbed test data (Moosavi-Dezfooli et al., 2017; Goodfellow et al., 2014a), leading to a loss of human's trust in the automated system. To address these concerns, works have proposed techniques to help detect out-of-distribution (Lee et al., 2017) or adversarial examples (Pang et al., 2018). In this paper, we show that the problem is even more acute– egregious failure modes are ubiquitous even in the context of in-distribution inputs, i.e. when the test and training distributions are similar.

The notion of explicability (Zhang et al., 2017; Kulkarni et al., 2016) and legibility (Dragan et al., 2013) has been recently investigated in the context of sequential decision-making problems in task and motion planning respectively. The basic idea is that the robot performs actions using its model of the world M^r and the human has an expectation about the robot's model, denoted as M^h. For the robot to be explicable, the authors argue that the robot should consider M^h when coming up with a plan. As opposed to considering structured models to represent M^h, which is easier in the case of task planning scenarios (Kulkarni et al., 2016), we consider using labels over classification outputs to capture the human's notion of explicability in the context of computer vision tasks.

In classification tasks, existing works seek to represent the concept of *trust* on black-box models in terms of the output soft-max probabilities (Hendrycks & Gimpel, 2016) or the distance to the closest hyper-plane that separates the decision boundary (Jiang et al., 2018). Other works tackle the issue of improving (a limited self-defined notion of) trust by examining a classifier's failure modes (Selvaraju et al., 2016; Agrawal et al., 2016). We strongly believe that trust is difficult to define, let alone express formally, without even understanding how to represent M^h or conducting human studies. Thus, our approach seeks to represent and obtain M^h first. Then, we incorporate it into the classifier and finally show that it helps to prevent egregious misclassifications that can lead to loss of trust.

Our approach is similar to the idea of using soft labels as opposed to the popular notion of one-hot encoding. To understand a human's confusion about a particular test instance being misclassified, works have considered interactive visual question answering (Branson et al., 2010) and obtaining humans' soft-labels for instances of a data-set (Peterson et al., 2019). Given that humans answer to instance-specific questions, labeling of instances needs to be incorporated in adjusting a classifier's weights, it should be no surprise that these approaches require an enormous human effort. For example, the latter approach requires a total of 500k probability distributed labels. We propose that gathering M^h can be done at an abstract level and tackle the problem of representing, obtaining, and incorporating M^h from a class-level perspective. Note that our method thus helps to augment incomplete instance-based labeling similar to collaborative filtering (Sarwar et al., 2001).

While there exists a long history of using class-label hierarchies (Tousch et al., 2012), these works focus on coming up

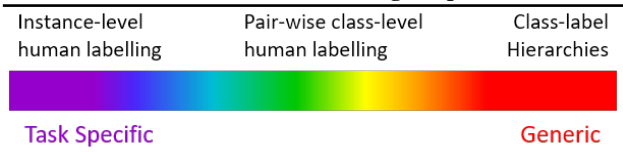

Figure 2. Situating methods to obtain semantic similarity in the spectrum of task-specific to generic methods.

with a formal representation structure (Fergus et al., 2010; Deng et al., 2014) or improving the speed of obtaining such representation (Chilton et al., 2013; Bragg et al., 2013). On the other hand, the use of weighted loss functions (WLFs) is a common tool to penalize certain misclassifications more than others (Duda et al., 2012; Sengupta et al., 2018). For example, weighing misclassification of inputs belonging to minorities can heavily help in soothing existing biases in data (Phan et al., 2017). Similarly, using a convex loss function with weighted penalties to differentiate between *quality variables* helps to find the best parameters (Chang et al., 2009). We follow suit and utilize WLFs to harshly penalize the most egregious mispredictions from a human's semantic similarity perspective.

3. Semantic Similarity

In our setting, semantic similarity aims to capture the degree of inexplicability evoked in a human if a classifier were to misclassify the image of a particular class (eg. dog) to a different class (eg. cat, ship). Therefore we want to pay attention to the costs of misclassifications. In this regard, people tend to have a certain idea about which of them to penalize because of their general knowledge of the world (e.g. the classes gorilla and human (Vincent, 2020)). Quantifying these costs, however, is challenging given that AI systems don't have the same general knowledge humans have. We thus have to obtain them through human studies and consider three approaches from task-specific to more generic methods (see Figure 2): humans give instance-level costs of misclassification (task-specific), class level costs of misclassification, or we compute class level misclassification costs indirectly from the WordNet hierarchy (Miller, 1998), assuming that it captures human sensibilities. Thus, to get a holistic view, it is important to obtain this as a pair-wise similarity metric over all class labels, the distance values of which are inversely proportional to the amount of explicability.

3.1. Instance-Level Human labeling (IHL)

One way of representing the semantic similarity between the class-labels is by asking humans to label individual instances in the data-set. In doing so, we need to provide human subjects knowledge about the task at hand and the available labels (Peterson et al., 2019). This method allows one to capture a great amount of detail– beyond (average) semantic similarity that represents the user's expectation of explicability and also captures robustness of M^h to noise.

Unfortunately, this method suffers from two major drawbacks. First, instance-based labeling is expensive to obtain. Each image needs a significant number of humans labeling them, and data-hungry machine learning models need many such images to train. Further, as the number of class labels increases, the total number of labels required increases significantly. Second, for many tasks, there is no need for users to give labeling at such a fine-grained level. For example, humans might find it unreasonable that the image of a dog (regardless of which one) was misclassified to an airplane. Hence, obtaining multiple instance-specific labels seems inefficient.

In our experiments on CIFAR-10, we leverage the labels obtained via the extensive human study in (Peterson et al., 2019). Each image was labeled by approximately 50 different people, thus having each image's label as a distribution over the classes, (i.e. soft-label) rather than just the top one. The total number of classifications for the $10,000$ images amounted to a total of $511,400$ and $2,571$ people were involved in it (Peterson et al., 2019). We average the instance-specific human-labels over all instances of a class to obtain the semantic distance of that class to other classes.

3.2. Class-Level Human labeling (CHL)

In this scenario, we consider obtaining similarity labels for pair-wise class labels. For CIFAR-10, this corresponds to finding the weights on each edge of a bipartite graph matching actual class-labels to predicted ones. We gather this by performing a user study with 50 people on Amazon Mechanical Turk (Turk, 2012).[1] To avoid noisy answers, we only allowed participation of turkers with high reputation. Further, we added two filter questions that asked a user to recognize the shapes of a triangle and a circle; this allowed us to detect scripted or random inputs. We were able to discard 4 data-points, getting us down to a total of 46 valid answers. Each turker was paid $\$2$ for their work that took 10 minutes on average to complete. We gave each turker a total of 10 questions– one for each class label of CIFAR-10. For each question, we showed them 36 images, sampled at random from the training set, presented as a 6×6 grid at the top and asked them to give their opinion over how understandable it was for a classification system to misclassify the images in a class (corresponding to the 36 images) to the remaining 9 classes. Their answers were weighted in a 0 to 4 Likert scale that ranged from *Highly Unreasonable (surprised)* to *Highly Reasonable (Explicable)*. To avoid ambiguity, we labeled each category of the Likert scale. Note the reduction in the number of human subjects required

[1] Link to the user study: https://bit.ly/3bHceX6.

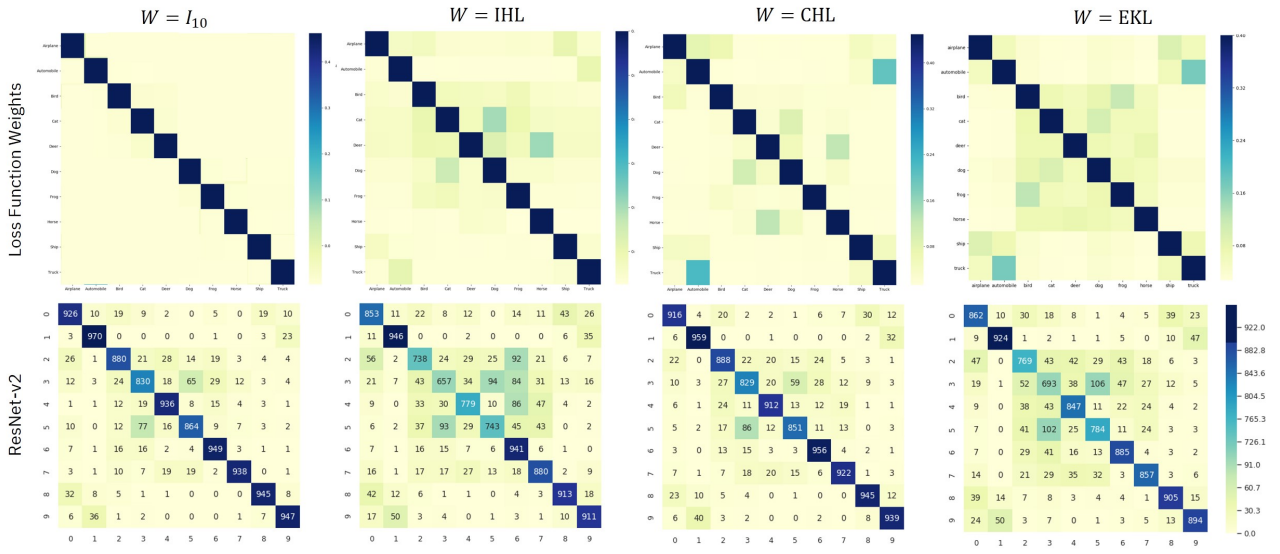

Figure 3. The top row indicates the weight matrix calculated using the vanilla categorical crossentropy (where all misclassifications are weighted equally) and the three different methods (IHL, CHL and EKL). The bottom row shows that the classification results produced by ResNet-v2 trained using the various WLFs adhere to the respective weight matrix.

to obtain 46 labels goes down from 2,571 (Peterson et al., 2019) to 50, a 50-fold reduction.

3.3. Utilizing Existing Knowledge for labeling (EKL)

Many existing image classification data-sets like ImageNet (Deng et al., 2009), CIFAR-10/100 (Krizhevsky et al., 2009) use nouns as class-labels. These nouns are present in popular linguistic structures such as the WordNet (Miller, 1998) (ImageNet class labels are derived from this lexical database). WordNet seeks to capture the relation between the various nouns using a tree-like structure where abstract concepts (eg. terrestrial animals) reside closer to the root and fine-grained concepts reside at the leaf of the tree (eg. Labrador retriever).

WordNet provides APIs to query various aspects of this database (eg. meaning, synonyms, antonyms, etc.). From these, the `path_similarity` calculates the semantic similarity between two nodes (or words) in the database based on the hyperonym/hyponym taxonomy. We use this score for representing class-level semantic similarity. The score ranges between $[0, 1]$ where 1 means the identity mapping (i.e. comparing a class-label with itself). Given this represents a task-independent mapping, it may not be as informative for tasks that require expertise (a discussion on this topic ensues in the section 6).

4. Incorporating Semantic Similarity with Weighted Loss Functions

Existing methods to train Deep Neural Networks encourage classification to the correct class while penalizing all misclassifications equally. The loss function objective thus, treats all failure modes indifferently, regardless of their impact on the user or the downstream task.

Weighted loss functions are often used to represent asymmetric misclassification costs for a classification task (Duda et al., 2012). If a task has n classification labels, we consider a $n \times n$ weight matrix that encodes the different penalties when an image belonging to the ground-truth class i (represented as the row) classified to class j with weight W_{ij}. This lets us introduce biases in the loss function to favor explicable misclassification and discourage egregious failure modes. If we represent the ground truth label as the vector y with $y_i = 1$ representing its membership to the actual class i and the prediction vector as p, we can formally represent the weighted loss function for a single image, over any loss function $\mathcal{L}$, as:

$$W\mathcal{L}F(y, p) = \mathcal{L}(W_i, p) \tag{1}$$

where each row of the weight matrix W_i matrix represents the human's expectation about which misclassifications are explicable *vs.* not given the actual class i. We posit that weighted loss functions can capture the expectations encoded in M^h with regards to misclassification.

The weight matrix contains the weights assigned to the edges of a fully connected bipartite graph from the set of

Model	*Functionality* Top-1 Accuracy ↑	*Explicability* $\mathcal{L}_{IHL}$ ↓	$\mathcal{L}_{CHL}$ ↓	$\mathcal{L}_{EKL}$ ↓	*Robustness* Gaussian Noise ↑	Adversarial (FGSM) ↑	*Cost* Additional Human Labels ↓
ResNet-v2 ($W = \mathbf{I}$)	**91.85%**	14.761	5.044	16.047	17.03%	9.98%	0
ResNet-v2 ($W =$ IHL)	83.61%	**2.258**	1.889	2.311	17.08%	12.14%	+511,400
ResNet-v2 ($W =$ CHL)	91.17%	3.054	**1.305**	3.274	21.45%	11.73%	+460
ResNet-v2 ($W =$ EKL)	86.03%	2.353	1.567	**2.461**	28.76%	12.63%	0

Table 1. Accuracy, explicability, robustness and the cost of developing individual classifiers trained with the various loss functions for CIFAR-10. As indicated by the arrows in the top-column, higher values for accuracy and robustness and lower values for the loss function and the cost are better.

actual to the set of predicted labels. In the context of the methods stated in section 3, we add (and normalize) the weights provided by humans to each edge over individual instances for IHL, average (and normalize) the weight given to each edge by individual humans for CHL and finally, leverage distance metrics over existing knowledge bases for EKL to obtain W. The different calculated weight matrices are shown in the top row of Figure 3. Note that the weight matrices, which capture the semantic similarity over classes for the various methods, are different. In the case of IHL (Peterson et al., 2019), the human's uncertainty over the noise in the CIFAR-10 data manifests as several labels that are off-the-diagonal. In the case of EKL, every word in the lexicon is connected to every other word, and thus, we notice many (relatively) dark squares off-the-diagonal. Precisely, we notice two hierarchies– one represented by the six classes in the middle and the other represented by the two classes at the top (or the bottom). The third column represents CHL, the human's label when faced with the question that whether a misclassification is explicable or not. The deeper squares toward the bottom-left corner of W represent the similarity between *truck* and *automobile* and are better visible for CHL and EKL compared to the IHL.

5. Experimental Results

In this section, we first present the classification results on the CIFAR-10 data-set (Krizhevsky et al., 2009) using the ResNet-v2 architecture (He et al., 2016). Our primary goal is to compare the different methods in terms of the operation metrics, the cost of developing them, the explicability measure, and the robustness to noisy inputs. We show that the two proposed methods–CHL and EKL–are cost-effective and better in terms of the operational metrics, robustness and in improving the explicability of the classifier. Then, we consider the CIFAR-100 data-set (Krizhevsky et al., 2009). Due to its large number of classes and requirement for models with large number of parameters, some of the approaches to (1) obtain the weights and (2) train the classifier become cost-prohibitive.

5.1. CIFAR-10

We plot a heat-map showcasing the classification results obtained using the ResNet-v2 trained using the different Ws discussed in section 3: IHL, CHL, EKL, in the bottom row of Figure 3. The classifier tries to capture the respective semantics represented by the corresponding Ws. In general, the classification results can be evaluated along three different axes– Functionality, Explicability, and Cost. The results along each dimension are summarized in Table 1.

Functionality In an operational setting, the output of a vision classifier may be used to inform the decision of an agent. In such cases, the notion of being uncertain about multiple classes is misleading. While we utilize the loss function value in the next section, similar to the IHL baseline (Peterson et al., 2019), we plot the top-1 accuracy for the CIFAR-10 dataset in the first column of Table 1.

The ResNet-v2 trained using the categorical cross-entropy loss has the highest accuracy in this regard. The network trained with CHL has almost equal accuracy, showing that the weighted loss function used to train the network does not interfere with the operational metric of the classifier. On the other hand, the accuracy drops by 5.82% when using EKL and by 8.24% when using the IHL baseline. Thus, both our proposed networks outperform the baseline on top-1 accuracy. One of them, CHL is almost as good as the vanilla ResNet-v2.

Explicability Similar to IHL, we use the loss function values on the test set to represent, in the context of this work, the explicability of the different classifiers. In our case, each weight matrix W represents a specific notion of explicability,

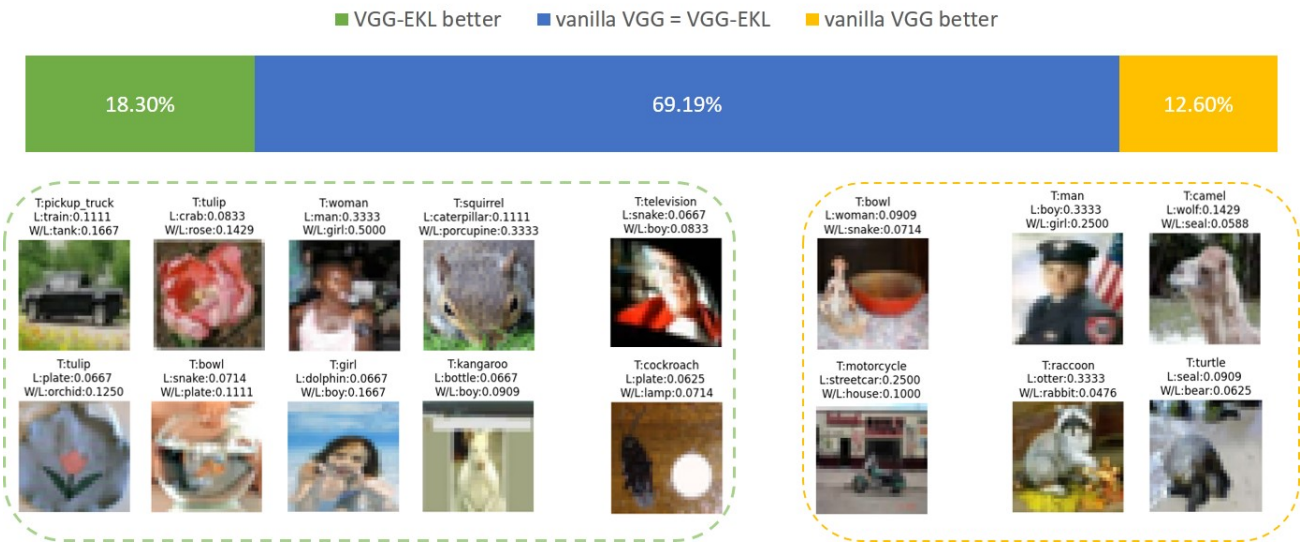

Figure 4. Images misclassfied by both the classifiers for CIFAR-100 are classified to semantically closer classes by the VGG classifier fine-tuned using EKL (VGG-EKL). In cases where the vanilla VGG does better, the VGG-EKL rarely makes egregious mistakes. In many examples, the VGG-EKL learns to pick up an object that is present in the picture but is not equal to the correct label.

and hence, it is reasonable to gauge the performance of the classifiers in terms of the three different metrics. We expect that a classifier trained using a particular weighted loss function (say, with W_{IHL}) outperforms the other classifiers when evaluated on that respective loss function (i.e. $\mathcal{L}_{IHL}$). In accordance, we find the lowest loss values on the diagonal of the three rows under the explicability criteria. In the future, we hope to conduct human studies to understand how well the individual loss functions capture the human's expectations of (mis)classification behavior.

The vanilla ResNet-v2, trained on the popular cross-entropy loss, has the highest values for all three explicability measures. The poor performance indicates that the semantics over failure modes that humans have in mind in the context of the classification task isn't captured by systems trained using prevalent objective/loss functions. We believe that while vanilla classifiers simply look at pixel level features, humans utilize a lot of context and knowledge, beyond pixel values, to make a decision. Thus, augmenting the classifier with these can help in obtaining an explicable failure mode.

The other classifiers dominate in terms of the loss function value they are trained to optimize and are comparable on other metrics. The IHL and the EKL methods turn out to have similar weighted loss functions and are different from CHL. Thus, the loss function value of EKL for $\mathcal{L}_{IHL}$ is just 0.004 more than the best loss value that is achieved by $ResNet$–IHL and similarly, we see a difference of merely 0.017 the other way round. Hence, even with a significant difference in terms of labeling costs (that we will discuss next), the proposed EKL method achieves similar explicability when compared to IHL. Claiming that a particular explicability score is the most representative of human expectation is difficult without a post-facto human study but both our proposed methods cover the two contesting definitions of representing explicability– Class-level Semantics (CHL) or Instance Level Semantics (EKL is comparable to IHL).

Robustness In Table 1, we highlight the accuracy values of the various classifiers on (1) noisy and (2) adversarially perturbed test inputs. We use pixel-level Gaussian Noise ($\mathcal{N}(0, 0.2)$) and the Fast Gradient Sign Methods (FGSM) (Goodfellow et al., 2014b) respectively.

One understands that existing classifiers are brittle to noise and thus, the accuracy drop for the vanilla ResNet-v2 should not be surprising. The IHL based weighted-loss function training helps to improve the robustness to adversarial examples slightly, congruous to the claims made in (Peterson et al., 2019). Unfortunately, we discover that the claims significantly weaken in the context of noisy test inputs injected with pixel-level Gaussian noise.

On the other hand, we observe that ResNet-CHL outperforms vanilla ResNet and ResNet-IHL against Gaussian noise while ResNet-EKL dominates all the classifiers on both noisy and adversarially perturbed test inputs. We believe that the reason for this improved robustness is the

use of (small) bounded noise that tries to make the classifier misclassify noisy inputs to semantically distant classes. The high penalty of such a mistake ensures the classifier chooses the correct class. Although, for CIFAR-10, we see that adding noise can often force misclassifications to semantically similar classes resulting in low accuracies.

Labeling Cost Note that for all vision datasets, the class labels are generally obtained via human study. Similarly, the WordNet hierarchy is also curated by human experts. Thus, there is already an enormous amount of human effort that is invested to develop classifiers. As this knowledge is readily available, we do not count this cost and talk in terms of the additional human labels required. Using the number of human subjects is an unfair metric as one can arbitrarily give a single subject more tasks to bring this number down, so we consider just the number of additional labels required by each of the methods.

The additional cost required for training vanilla ResNet-v2 and ResNet-v2 with EKL is considered zero because the existing WordNet semantics were readily available. In contrast, our proposed methods CHL, which gathers class-level semantics from human labeling, requires 460 labelings. Both of these methods EKL and CHL require far fewer labels than IHL, which as per the authors (Peterson et al., 2019), required $511,400$ labels.

5.2. CIFAR-100

The CIFAR-100, as evident from its name, contains images belonging to 100 classes (Krizhevsky et al., 2009). In this case, the IHL baseline requires human subjects to (1) provide a probability distribution of 100 classes for each data-point leading to an increase in the human subject's cognitive overload, and (2) annotate a significantly larger number of labeled samples compared to CIFAR-10 increasing the cost of obtaining additional labels. In the case of CHL, the cost of labeling, while still significantly less than IHL, also increases because we now need weights for a bipartite graph with $\binom{100}{2} = 4950$ edges. Further, to reduce cognitive overload on the human, we can show just a subset of classes that a class can be misclassified to; this leads to an increase in the population size. Note that such a breakdown is difficult to do in the context of IHL. Owing to the added cost for both the methods, we consider only EKL in this setting. Similar to the case of CIFAR-10, all the class labels present in CIFAR-100 are also a part of WordNet. Thus, we use the path similarity between the class-labels to populate the weight matrix W.

We use the VGG (Simonyan & Zisserman, 2014) classifier for this task. VGG needs to train ≈ 183 million parameters compared to ≈ 25 million for ResNet. Thus, training from scratch becomes time and resource-intensive; we consider fine-tuning pre-trained models. This experiment helps us showcase the benefits of our approach when the classifiers are significantly larger and tasks are complex.

Model	Accuracy ↑	$\mathcal{L}_{EKL}$ ↓
VGG (vanilla)	70.48%	16.377
VGG (w WLF)	70.55%	5.686

Table 2. The accuracy and explicability (represented by the weighted loss function value) of the vanilla VGG classifier and the one fine-tuned using WLF with EKL.

In Table 2, we show the accuracy and the explicability score, computed using the weighted loss function value, on the test set. In contrast to the results in the previous section, the use of a weighted loss function that enforces a soft-labeling scheme behaves as a regularizer increasing the top-1 accuracy of the pre-trained vanilla VGG from 70.48% to 70.55%. Further, the explicability score of VGG fine-tuned with the EKL weighted loss function (VGG-EKL) has a loss function value of 5.686 compared to 16.377 for the vanilla VGG classifier. Now, we analyze the failure modes of the two classifiers.

In Figure 4, we showcase three scenarios that arise when both the classifiers misclassify a given test input to an incorrect class. We show the ground-truth class label, the class label it was classified to by the vanilla VGG followed by VGG fine-tuned using EKL. The numbers beside the predicted class labels show the similarity between the predicted class and the true class as per WordNet's path similarity metric. In the majority of the cases, precisely 69.19% of them, both the classifiers misclassify an input to the same incorrect class. This high agreement should not be surprising because the VGG-EKL is simply a fine-tuned version of the vanilla VGG network. There exist two other scenarios– (1) when VGG-WLF misclassifies an input image to a semantically closer class and (2) when the vanilla VGG does so. The former happens 18.3% of the time while the latter occurs 12.6% of the time.

Examples of the first case show that flowers like `tulip` and `orchid` are classified as `crabs` and `plates`, images of people are classified as animals (`girl` $\rightarrow$ `dolphin`), and animals are classified to inanimate objects (`kangaroo` $\rightarrow$ `bottle`) by the vanilla VGG classifier. On the other hand, the VGG-EKL preserves these semantics learned from WordNet. In the latter case, examples highlight that misclassifications made by VGG-WLF, while worse-off than the vanilla VGG, are less egregious as per the Word-Net similarity metrics. This fact is reinforced by the values of the explicability metric (in Table 2) that is significantly better for VGG-WLF compared to vanilla VGG. Note that there exists a subset of test inputs on which the misclassifications

made by VGG-WLF refer to an object present in an input image but, due to the original ground-truth label, regarded as a misclassification. For example, the image (in the middle) labeled as a `television` shows the picture of a person inside a television. While vanilla VGG labels it as a `snake`, VGG-WLF labels it as `boy` referring to the person.

6. Discussion

While we talk of explicable classification, our goal is to train a classifier that agrees with a human's view of the failure modes, thereby reducing the surprise caused by a particular misclassification. A more nuanced view should consider the penalty of a mistake in terms of the various impacts a particular misclassification may have on the downstream task. In this regard, we recognize two perspectives– an operational one and the other about societal biases.

Operationally-reasonable misclassifications Often, misclassifications may be inexplicable to a human but, given the downstream task, considered reasonable. For example, in Figure 4, classifying a `kangaroo` to a `bottle` may be deemed unsafe for autonomous driving scenarios (in Australia) whereas a system misclassifying it to a `boy` is better as the underlying decision of stopping the car remains unaffected. Without the context of the underlying task, classifying a `kangaroo` to a `boy` may be considered inexplicable. Thus, the class-level penalty scores for explicability may not align with the task-specific class-level penalties for operational purposes. Thus, leveraging existing knowledge bases, unless designed specifically for the task at hand, becomes unreasonable. In these scenarios, CHL is the only choice.

Reducing Impacts of Societal Biases In several domains, a particular misclassification can be viewed as reinforcing societal biases on test inputs belonging to marginalized classes. A classic example is state-of-the-art classifiers labeling the image of a dark-skinned person as a gorilla (Vincent, 2020). In such cases, failure modes that are unacceptable from a social standpoint can have a high penalty. Thus, when crafting human studies in such domains, one has to either find a group of people who are aware of these biases and can account for them or, at the very least, provide cues to participants as to what failure modes encode societal biases and impact downstream tasks.

In reality, we may desire that a classifier to trade-off between explicability, operational costs, and the societal impacts of misclassifications. Thus, the weights of the WLF can simply be considered a function of the three individual weights, i.e. explicability weights, operational impact weights, and weights that penalize societal biases.

7. Conclusions and future work

In this paper, we showed that the prevalent objective functions for training Deep Neural Networks that weigh all misclassifications equally lead to inexplicable failure modes leading to a loss of human's trust in the system. To prevent these inexplicable misclassifications for vision classification tasks, we proposed two methods that can help us obtain the human's model M^h that calibrates misclassification on a scale ranging from explicable to egregious. We note that, beyond the explicability scenario, our methods can be generalized to provide operational benefits or prevent misclassifications that have negative societal impacts. We then utilized the notion of weighted loss functions to incorporate M^h into the classifier's model and showed that our method not only helps the classifier reduce the number of egregious errors, but also acted as a regularizer to improve the accuracy and the robustness of the baseline model.

Acknowledgements Kambhampati's research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-19-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19-2-0006, NSF grants 1936997 (C-ACCEL), 1844325, and a NASA grant NNX17AD06G.

References

Agrawal, A., Batra, D., and Parikh, D. Analyzing the behavior of visual question answering models. *CoRR*, abs/1606.07356, 2016. URL http://arxiv.org/abs/1606.07356.

Bragg, J., Weld, D. S., et al. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*, 2013.

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., and Belongie, S. Visual recognition with humans in the loop. In Daniilidis, K., Maragos, P., and Paragios, N. (eds.), *Computer Vision – ECCV 2010*, pp. 438–451, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Chang, Y.-C., Liu, C.-T., and Hung, W.-L. Optimization of process parameters using weighted convex loss functions. *European journal of operational research*, 196(2):752–763, 2009.

Chilton, L. B., Little, G., Edge, D., Weld, D. S., and Landay, J. A. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1999–2008, 2013.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database.

In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. Large-scale object classification using label relation graphs. In *European conference on computer vision*, pp. 48–64. Springer, 2014.

Dragan, A. D., Lee, K. C., and Srinivasa, S. S. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE, 2013.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern classification*. John Wiley & Sons, 2012.

Fergus, R., Bernal, H., Weiss, Y., and Torralba, A. Semantic label sharing for learning with many categories. In *European Conference on Computer Vision*, pp. 762–775. Springer, 2010.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014a.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

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

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CoRR*, abs/1907.07174, 2019. URL http://arxiv.org/abs/1907.07174.

Jiang, H., Kim, B., Guan, M., and Gupta, M. To trust or not to trust a classifier. In *Advances in neural information processing systems*, pp. 5541–5552, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kulkarni, A., Chakraborti, T., Zha, Y., Vadlamudi, S. G., Zhang, Y., and Kambhampati, S. Explicable robot planning as minimizing distance from expected behavior. *CoRR, abs/1611.05497*, 2016.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Mallick, A., Chaitanya, D., and and, K. B. Sample efficient uncertainty estimation for deep learning safety. 2020.

Melis, D. A. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pp. 7775–7784, 2018.

Miller, G. A. *WordNet: An electronic lexical database*. MIT press, 1998.

Montavon, G., Samek, W., and Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Pang, T., Du, C., Dong, Y., and Zhu, J. Towards robust detection of adversarial examples. In *Advances in Neural Information Processing Systems*, pp. 4579–4589, 2018.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9617–9626, 2019.

Phan, H., Krawczyk-Becker, M., Gerkmann, T., and Mertins, A. Dnn and cnn with weighted and multi-task loss functions for audio event detection. *arXiv preprint arXiv:1708.03211*, 2017.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

Sengupta, S., Dudley, A., Chakraborti, T., and Kambhampati, S. An investigation of bounded misclassification for operational security of deep neural networks. In *AAAI Workshop of Engineering Dependable and Secure Maching Learning Systems*, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Tousch, A.-M., Herbin, S., and Audibert, J.-Y. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

Turk, A. M. Amazon mechanical turk. *Retrieved August*, 17:2012, 2012.

Vincent, J. *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech*, 2020. URL https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.

Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., and Kambhampati, S. Plan explicability and predictability for robot task planning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 1313–1320. IEEE, 2017.