‘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) reinforces harmful societal biases. 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 capture the human's expectation (M^h) regarding which classes are semantically close *vs.* ones that are far away. 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 proposed methods lead to Deep Neural Networks that have comparable accuracy, explicable failure modes, comparable robustness and significantly less cost in teams of additional human labels required.

1. Introduction

While researchers have invested effort in trying to make neural networks in vision more 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, 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 societal biases learned from data (Vincent, 2020).

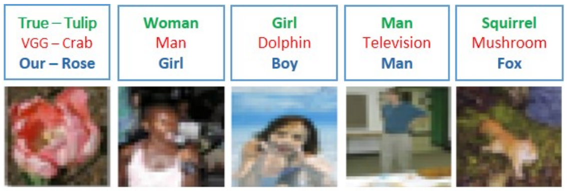

Figure 1. Examples showing that state-of-the-art neural networks exhibit different failure modes (on CIFAR-100 dataset), often resulting in inexplicable mistakes can lead to a loss of trust at best and have societal impacts at worst (Vincent, 2020).

We believe that egregious mistakes are a by-product of the existing loss/objective functions used by state-of-the-art classifiers; they are simply too sparse to encode meaningful information about failure modes. For example, the popular Categorical Cross-Entropy (CCE) loss 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.

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 semantically closer classes (to the ground-truth) are considered explicable. To obtain M^h, we strongly advocate the use of a human labeling approach and, in cases, where the classification task is generic, 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 (or fine-tune) classifiers.

We compare the trained classifiers with existing baselines and find that our methods achieve (1) similar accuracy re-

*Equal contribution [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 Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).

sults on in-distribution samples, (2) higher explicability when misclassifying inputs, (3) higher robustness to random noise and comparable accuracy against adversarial examples and (4) significantly lesser cost in gathering human labels. We discuss the implication of our methods on addressing operational issues and calibrating societal impacts and showcase experiments on other datasets in the appendix.

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) or adversarially perturbed test data (Moosavi-Dezfooli et al., 2017; Goodfellow et al., 2014a), leading to a loss of human 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. 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 tackle the issue of improving trust by examining a classifier's failure modes (Hendrycks & Gimpel, 2016; Jiang et al., 2018; Selvaraju et al., 2016; Agrawal et al., 2016). Unfortunately, they leverage self-defined notions of trust and ignore human subject studies. Thus, their understanding of human expectation may be completely wrong. We address this concern by proposing methods to represent, obtain and incorporate human's expectation in preventing 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 several instances of a data-set (Peterson et al., 2019). It should be no surprise that the latter approach (a baseline) requires an enormous human effort. On the other hand, we propose to gather M^h at an abstract level and tackle the problem of 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 class-label hierarchies (Tousch et al., 2012) have been well-studied, works have mostly focused on obtaining a formal representation structure (Fergus et al., 2010; Deng et al., 2014) or improving the speed of obtaining them (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)– weighing misclassification of inputs belonging to minorities can help in soothing existing biases in audio data (Phan et al., 2017) or using a convex loss function with weighted penalties to differentiate between *quality variables* can help to find the best parameters (Chang et al., 2009). We follow suit and utilize WLFs to penalize egregious misclassifications (from a human's semantic similarity perspective).

3. Semantic Similarity

Semantic similarity aims to capture the amount 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. ship). To get a holistic view, it is important to obtain a pair-wise similarity metric over class labels, the distance values inversely proportional to the amount of explicability. We now describe three ways of obtaining these values.

3.1. Instance-Level Human labeling (IHL)

To represent semantic similarity between the class-labels, we can ask humans to provide probability distributions over individual instances in the data-set. Beyond (average) semantic similarity, this can capture robustness of M^h to noise (Krizhevsky et al., 2009). However, this method suffers from three major drawbacks. First, instance-based labeling is often expensive to obtain, each image needing a significant number of humans labels. Second, when number of classes increase, providing such labels result in increased cognitive overload. Third, labeling at such a fine-grained level is an overkill of many tasks. 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.

3.2. Class-Level Human labeling (CHL)

We obtain similarity labels for pairwise 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 over 50 people in Amazon Mechanical Turk (Turk, 2012).[1] To avoid noisy answers, we only allowed participation of mas-

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

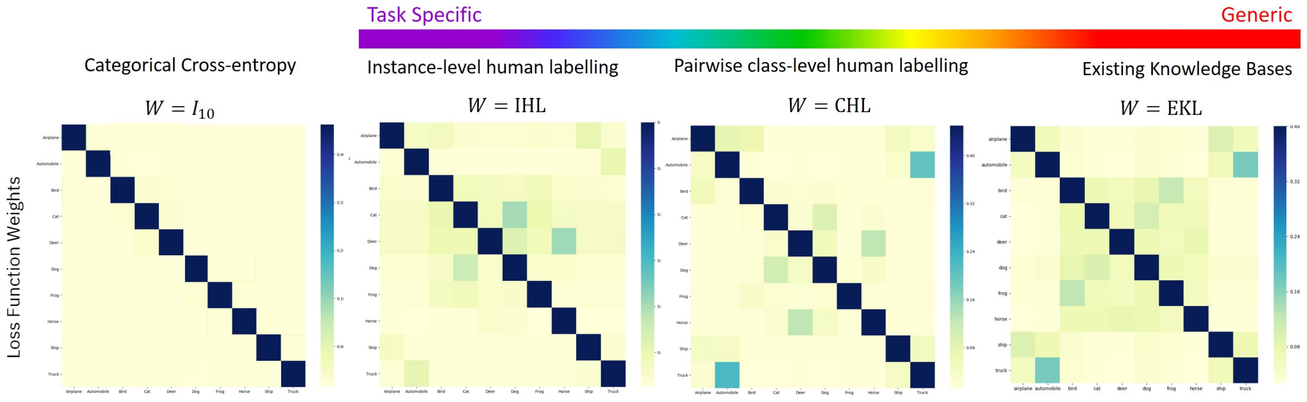

Figure 2. Different methods to learn explicability labels over class-level misclassifications.

ter turkers. Further, we pruned four low-quality data-points based on answers to two simple questions and ended up with a total of 46 data-points. We paid a turker $2 for their work, which took 10 minutes on average. Given 36 images of a class (randomly sampled from CIFAR-10), we asked user to rate the explicability of misclassification on a Likert scale that ranged from *Highly Unreasonable (Surprised)* to *Highly Reasonable (Explicable)*.

3.3. Utilizing Existing Knowledge for labeling (EKL)

Often, existing image classification data sets consist of class-labels that are nouns. Relations between nouns are represented in popular linguistic hierarchies such as WordNet (Miller, 1998) using a tree-like structure. We leverage WordNet's APIs to query the `path_similarity` between class labels and use it to represent semantic similarity (based on the hyperonym/hyponym taxonomy). The score ranges between $[0, 1]$ where 1 means the identity mapping of class-labels. As WordNet represents a task-independent mapping, it may not be as informative for tasks that require expertise (a discussion ensues in Appendix B).

4. Incorporating Semantic Similarity with Weighted Loss Functions

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 W 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. Given the true class y_i and prediction vector p, we can formally represent the weighted loss function for the single image over a loss function $\mathcal{L}$, as:

$$WLF(y_i, p) = \mathcal{L}(W_i, p) \tag{1}$$

We posit that weighted loss functions can capture the expectations about the failure modes encoded in M^h. In Figure 2, we plot a heat-map showcasing the W-s obtained using the different methods discussed in section 3.

5. Experimental Results

We present 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 and contrast the different methods in terms of accuracy, the cost of developing them, the explicability measure and robustness to out-of-distribution samples. The results along each dimension are summarized in Table 1.[2]

Functionality In an operational setting, the output of a vision classifier may be used to inform the decision of an agent. In such cases, having uncertainty about multiple classes, championed by the baseline IHL (Peterson et al., 2019), isn't useful. As per the top-1 accuracy, ResNet-v2 trained using the categorical cross-entropy loss proves to be the best with the ResNet-v2 trained on CHL-weighted loss having approximately the same value. For the propose EKL-weighted loss, the accuracy drops by 5.82% and by 8.24% for IHL.

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. Note that each W represents a specific notion of explicability. Hence, it is reasonable that we gauge the performance of the classifiers with respect to all of them. To figure out which methods aligns well with the true explicability, a human subject study to evaluate outputs of each classifier is essential. We plan to conduct one as a part of the future work.

[2] More experimental results can be found in Appendix A.

	Functionality	*Explicability*			*Robustness*		*Cost*
Model	Top-1 Accuracy ↑	$\mathcal{L}_{IHL}$ ↓	$\mathcal{L}_{CHL}$ ↓	$\mathcal{L}_{EKL}$ ↓	Gaussian Noise ↑	Adversarial (FGSM) ↑	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. The accuracy, explicability and 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 lower values for the loss function and the cost are better.

The vanilla ResNet-v2 has the highest values for all three explicability measures, indicative of their failure to capture the human's expectation over failure modes. While vanilla classifiers tend to look at features derived from individual pixels, humans often utilize context and knowledge, beyond pixel values, to make decisions. Augmenting the loss function with this context, represented as weights over failure modes, can thus help in training more explicable classifiers.

As expected, a classifier trained using a particular weight matrix (say W_{IHL}) performs best on the corresponding explicability metric (i.e. $\mathcal{L}_{IHL}$). Thus, the lowest loss values occur on the diagonal of the three rows under the explicability criteria. The off-diagonal column values for the classifiers that optimize a different loss are comparable. It turns out that the explicability scores for IHL and EKL-weighted networks are similar– the ResNet-EKL has $\mathcal{L}_{IHL}$ value that is just 0.005 more than the best loss value achieved by ResNet-IHL. Thus, even with a huge difference in terms of costs (discussed later), the proposed EKL achieves explicability similar to IHL. Evaluating which explicability score represents the human expectation demands a post-facto human study, which is a future endeavour.

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.

It is well known that existing classifiers are brittle to noise and the accuracy drop for the vanilla ResNet-v2 confirms this. The IHL based weighted-loss function training helps to improve 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 choose the correct class. Although, for CIFAR-10, we see that adding noise can often force misclassifications to semantically similar classes resulting in low accuracies.

Cost Given ground truth labels and the WordNet hierarchy are readily available, we ignore the labeling cost invested in obtaining them. To measure the cost of the various methods, we consider the number of additional labels required by each of the methods. Thus, the additional costs for the vanilla ResNet-v2 and ResNet-EKL are zero. In contrast, our proposed method CHL, which gathers class-level semantics via human labeling, requires 460 labels compared to the baseline method IHL that required human $511,400$ labels (Peterson et al., 2019), a 1000-fold reduction. In Appenix A, we discuss that IHL, beyond becoming cost-prohibitive, imposes unreasonable cognitive overload on human subjects for datasets with larger number of output classes.

6. Conclusions and future work

In this paper, we showed that the popular objective functions for training Deep Neural Networks that weigh all misclassifications equally lead to inexplicable failure modes in turn leading to a loss of human 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 about which errors are inexplicable and can thus lead to a loss of trust in the system. 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 have comparable accuracy and improve 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., 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.

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.

7. Appendix A – Experiments

7.1. Experiments on 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, and (2) annotate a significantly larger number of labeled samples. Clearly, this increases both the additional cost of labeling and the cognitive overload on the human subject. 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, and thus, the path similarity between the labels to populate the weight matrix W.

We use the VGG (Simonyan & Zisserman, 2014) classifier for this task. VGG needs to train $\approx$ 183 million parameters compared to $\approx$ 25 million for ResNet and thus, training from scratch becomes time and resource-intensive. Thus, we consider fine-tuning pre-trained models. This helps showcase the benefit of our approach even when considering classifiers for tasks that are significantly larger.

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

Table 2. The accuracy and explicability (represented by $\mathcal{L}_{EKL}$) of the vanilla VGG classifier and the one fine-tuned with EKL.

Results 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 3, we showcase three scenarios that arise when both the classifiers misclassify a given test input to an incorrect class. We show the true class, the class label it was classified to by the vanilla VGG followed by VGG fine-tuned using WLF. The numbers beside the predicted

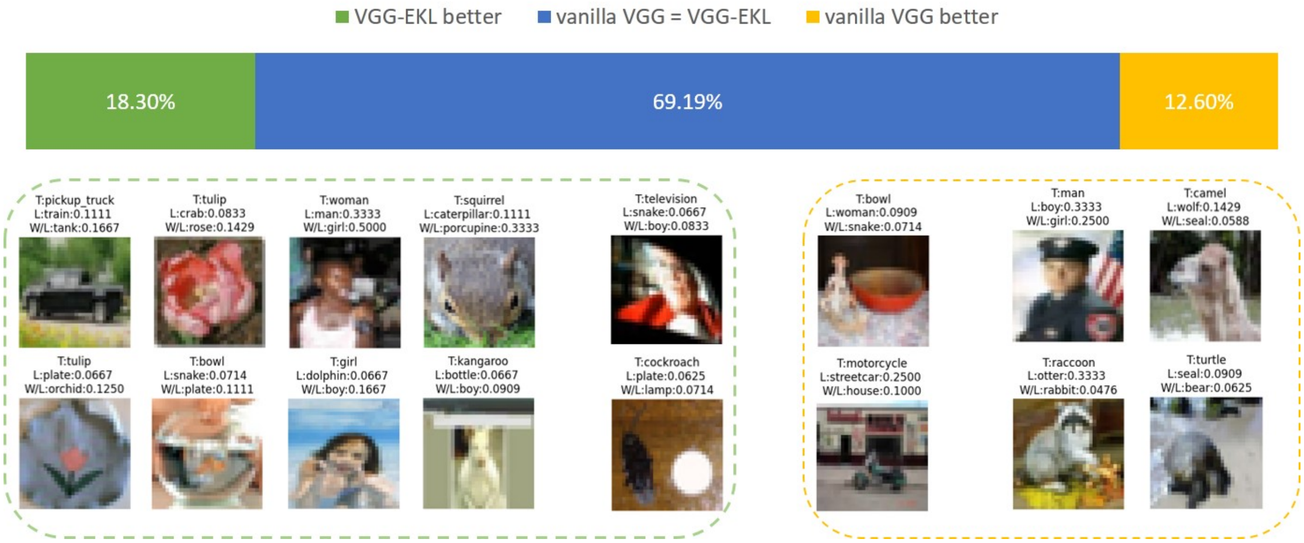

Figure 3. Images misclassfied by both the classifiers for CIFAR-100 are classified to semantically closer classes by the VGG classifier fine-tuned using EKL (VGG-EKL). In cases where the vanilla VGG does better, the VGG-EKL rarely makes egregious mistakes. In many examples, the VGG-EKL learns to pick up an object that is present in the picture but is not equal to the correct label.

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 should not be surprising because the VGG-EKL simply fine-tunes the weight 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` → `dolphin`), and animals are classified to inanimate objects (`kangaroo` → `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 is also supported by the fact that the explicability metric (in Table 2) is significantly better for VGG-WLF compared to vanilla VGG. In both these scenarios, 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 is regarded as the incorrect label as per the gold/true class labels of CIFAR-100. For example, the image 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.

8. Appendix B – Discussion

While we talk of explicable classification, our goal is to train a classifier that agrees to the 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 consider 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 3, classifying a `kangaroo` to a `bottle` may be deemed unsafe for autonomous driving scenarios (in Australia) whereas a system classifying 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 created 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 may 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, a classifier may often be required to trade-off between explicability, operational costs, and consider 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 to neutralize societal biases.