

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

Alberto Olmo* (🐦), Sailik Sengupta* (🐦), Subbarao Kambhampati (🐦)

Motivation



- Accuracy \neq explicability.
- How do Failures Look? Egregious Errors can result in
 1. Loss of Trust
 2. Safety issues
 3. Uphold societal biases
- Predictive parity / error rate balance / demographic parity does not consider the egregiousness of a mistake.

Representing Magnitude of Explicability

- Pairwise similarity between classes can be used to represent egregiousness of misclassifications.
 - ◇ Classification to classes semantically far away = Egregious mistakes
 - ◇ Classification to semantically close classes = Explicable mistakes

Obtaining Semantic Similarity Representation

- Instance Based Human Labelling (IHL)
 - ◇ Very expensive
 - ◇ Does not scale
 - ◇ Finest Granularity
- Pairwise Class-level Human Labelling (CHL)
 - ◇ Less expensive
 - ◇ Scales decently
 - ◇ Coarser Granularity
- Existing Knowledge for Labelling (EKL)
 - ◇ Not expensive
 - ◇ Scales easily
 - ◇ May not represent context-specific Explicability

Discouraging egregious mistakes

- Weight the loss values in accordance with the semantic similarity distance.
 - ◇ Explicable mistakes should not make the loss infinity.
 - ◇ Inexplicable or egregious mistakes should make the loss infinity.

$$W \mathcal{L}F(y_i, p) = \mathcal{L}(W_i, p)$$

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

Table 1: ResNet-v2 on CIFAR-10.

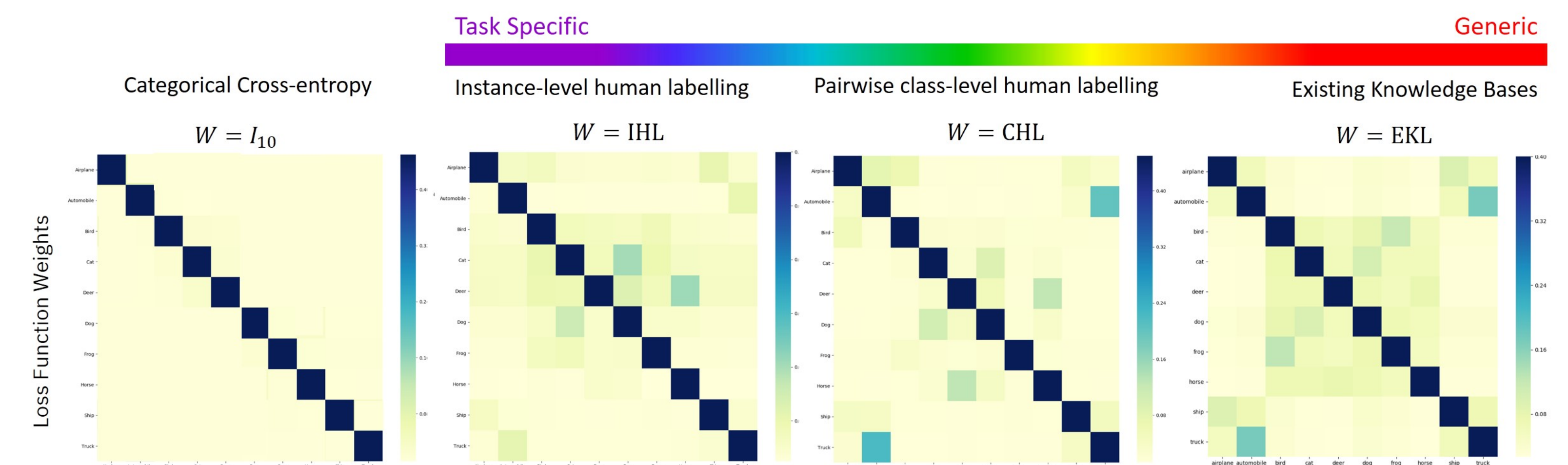


Figure 1: Different methods to learn explicability labels over class-level misclassifications.

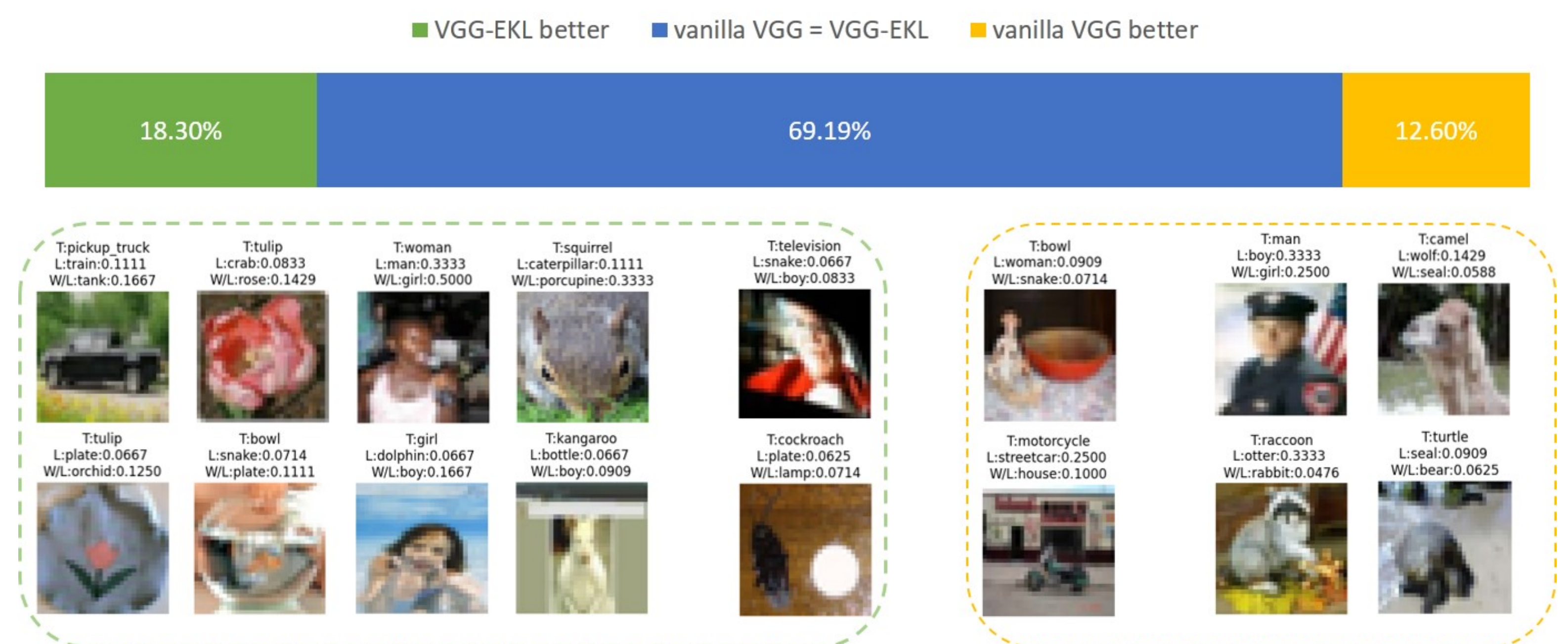


Figure 2: Vanilla VGG vs VGG fine-tuned with EKL.