What could go wrong?
Discovering and describing failure modes in computer vision

Gabriela Csurka, Tyler L. Hayes, Diane Larlus, Riccardo Volpi

Summary

- We formalize **Language-Based Error Explainability** (**LBEE**)
- We propose a family of **task agnostic** methods **to tackle LBEE**
- We introduce a **set of metrics** to evaluate LBEE performance
- We show the effectiveness of the proposed methods on various tasks

Contribution #1: Problem Formulation

Given a **target set** X and a **model** M_θ, our goal is to find sentences describing likely failure causes for the model

$$S_\beta^* = \left\{ s_n \in S \;\middle|\; \omega_\theta^{s_n} < \omega_\theta^{\text{avg}} - \beta \right\}$$

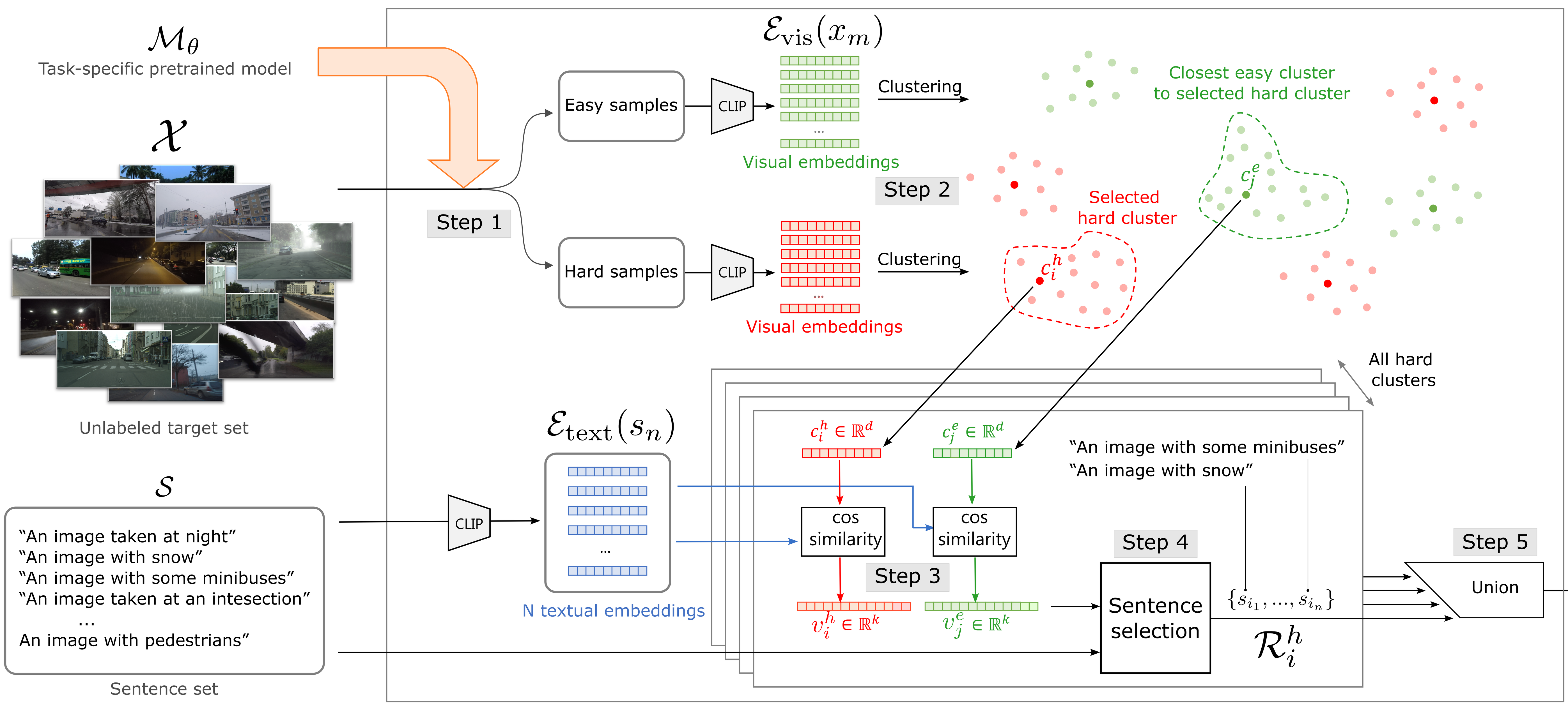
Predefined sentence set

Model average performance on images relevant to s_n

Model average performance on X

Predefined margin

Contribution #2: A Family of Methods



Step 1: split the images into easy and hard sets based on the model's confidence

Step 2: embed images in the CLIP space and cluster the hard and easy sets independently

Step 3: assign to each hard prototype the closest easy prototype in this space

Step 4: select sentences for hard clusters in the CLIP space based on cosine similarities with the cluster prototypes that are not relevant for the closest easy clusters

Step 5: aggregate cluster-specific sentence sets to produce the global output ($\mathcal{R}_S$)

Contribution #3: Evaluation Metrics

Given a hard cluster c_i^h and set of selected sentences $\mathcal{R}_i^h$

- **Hardness ratio (HR):** ratio of sentences pointing to reasons for model failure
- **Correctness Ratio (CR):** average ratio of images that are relevant to individual sentences

$$HR_i = \frac{\left|\{\forall s_k \in \mathcal{R}_i^h \mid \omega_\theta^{avg} - \omega_\theta^k > \beta\}\right|}{|\mathcal{R}_i^h|} \qquad CR_i = \frac{1}{|\mathcal{R}_i^h|} \sum_{s_k \in \mathcal{R}_i^h} \frac{1}{|c_i^h|} \sum_{x \in c_i^h} \Gamma(x, s_k)$$

Given S_β^* and the overall output ($\mathcal{R}_S = \cup\, \mathcal{R}_i^h$)

- **True positive rate (TPR):** evaluates how well S_β^* is covered
- **Jaccard Index (JI):** measures coverage while penalizing false positives

$$TPR = \frac{|S_\beta^* \cap \mathcal{R}_S|}{|S_\beta^*|} \qquad JI = \frac{|S_\beta^* \cap \mathcal{R}_S|}{|S_\beta^* \cup \mathcal{R}_S|}$$

Experimental Setup

Tasks and datasets:

- **Urban scene segmentation:** ACDC, IDD, WD2
- **Classification with spurious correlations**: $\text{NICO}_{++}^{75/85/95}$
- **ImageNet-1K classification**

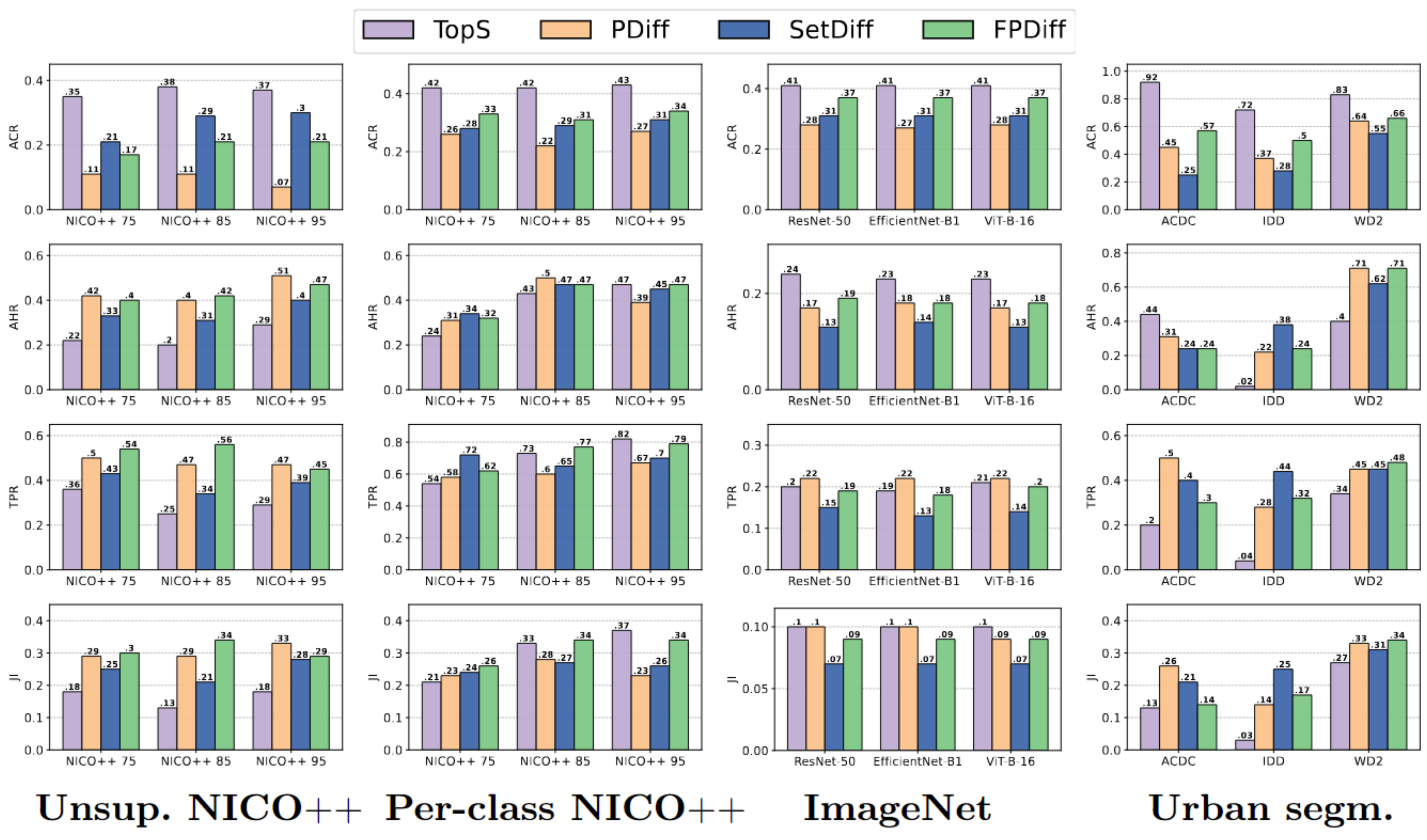
Methods:

- **TopS**: top ranked sentences based on cosine similarity
- **PDiff:** rank based on prototype difference
- **FPDiff:** Pdiff filtered with TopS
- **SetDiff:** Sentence set differences.

Default design choices:

- Open-CLIP, 15 clusters, 3 sentences, $\beta = .2 * \omega_\theta^{\text{std}}$

Quantitative Results



Qualitative Results ($\mathcal{R}_i^h$)