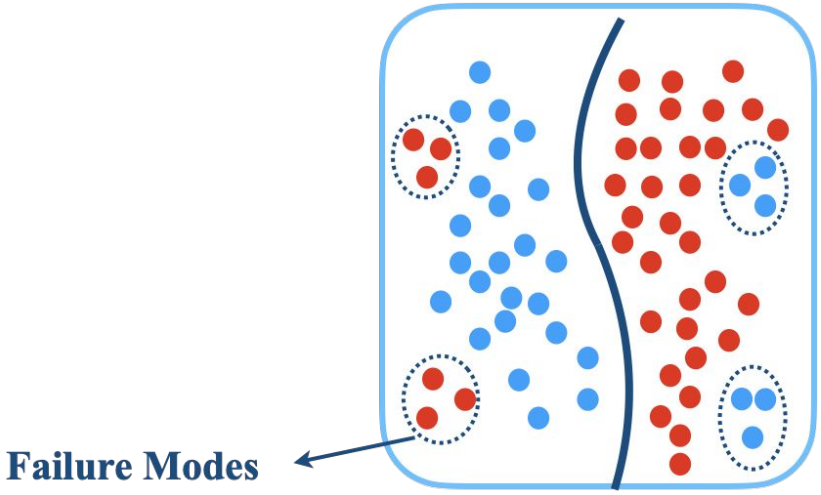
PRIME: **PR**ioritizing **I**nterpretability in Failure **M**odel **E**xtraction

Keivan Rezaei, Mehrdad Saberi, Mazda Moayeri, Soheil Feizi

ICLR 2024

What are Failure Modes in Classification?

+ ***Overall accuracy*** of the classifier is high.
- Certain groups of inputs (***failure modes***) on which the model ***underperforms***.

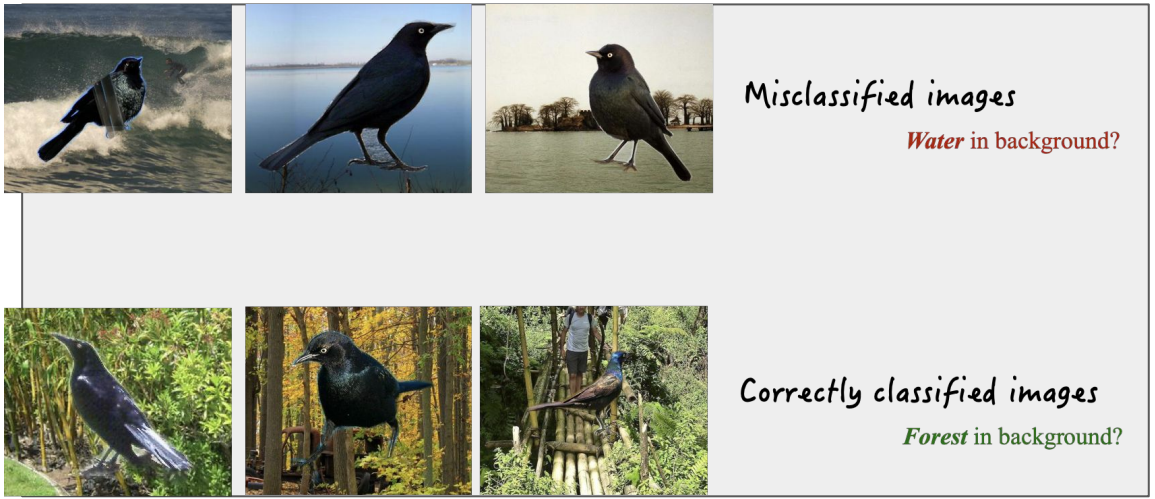


Explaining Failure Modes

Failure inputs are easy to find but:

— can we find ***similar patterns and attributes*** on those inputs?
— can we take a step further and explain those patterns in ***human-understandable terms***?

→ ***Improve Models***



Existing Work

Employ multi-modal tools (such as CLIP)

- find ***hard*** directions (clusters) in the latent space
- ***then*** aim to provide ***descriptions*** for them.

They often ***suffer*** from

1. ***Quality*** of descriptions
2. ***Coherency*** of images within groups

Latent space vs **Semantic** space

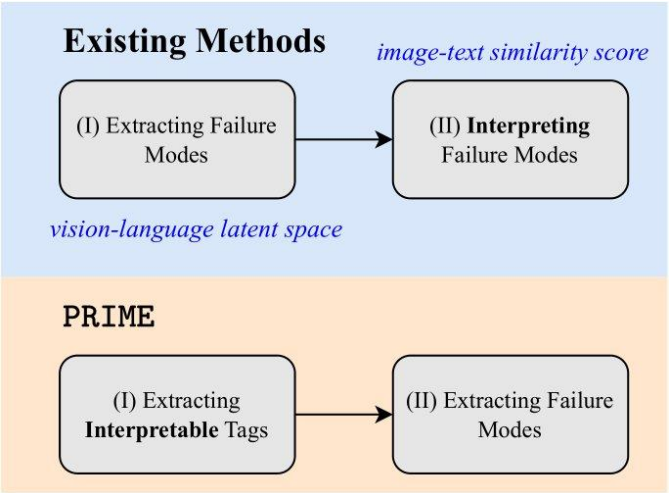
Why do we obtain ***low-quality*** descriptions?

- ***Latent space*** may not be a good proxy for ***semantic space***
- Points that are close to each other in latent space do not necessarily share same semantics
    - Existing method inevitably generate ***low-quality descriptions***

A New Paradigm? → PRIME

Put interpretability first!

I. Extract ***tags*** (concepts) from images.
II. Look for ***minimal combination of tags*** whose appearance drops model's accuracy!
III. Obtain ***failure modes.***



A New Paradigm? → PRIME

PRIME results

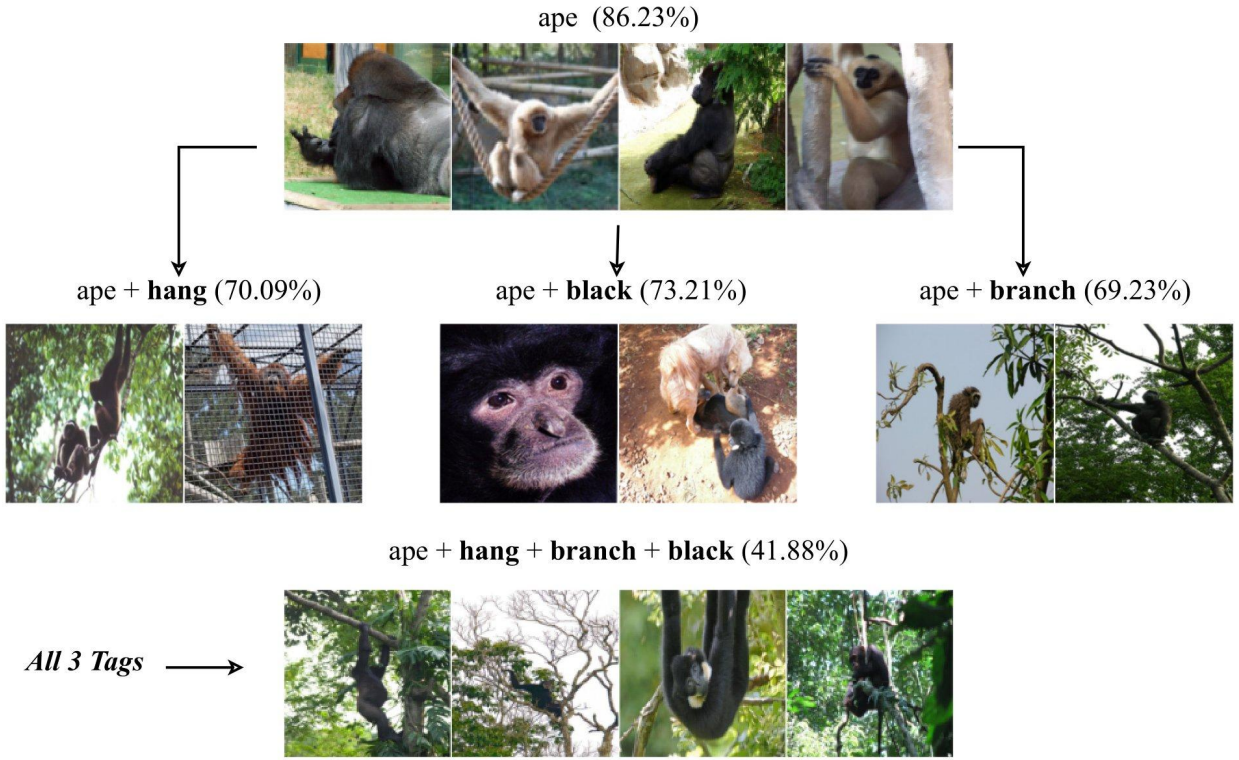
fox (81.96%)

fox + **white** + **zoo** (35.29%)

fox + **grass** + **stand** + **field** + **dry** (47.83%)

A New Paradigm? → PRIME

PRIME results



PRIME benefits

We obtain ***better*** *description* for failure modes, compared to existing work.

- Higher ***similarity*** of text descriptions to images inside groups
- More ***specific*** text descriptions
- More ***coherent*** images

How to ***evaluate***? We proposed a *suite of* ***automated metrics***.

PRIME benefits

Similarity score of *images inside failure modes* to the *generated caption*:

- ***AUROC*** measures ***specificity***
- ***STD*** measures ***coherency***
- ***Mean*** measures the ***similarity***