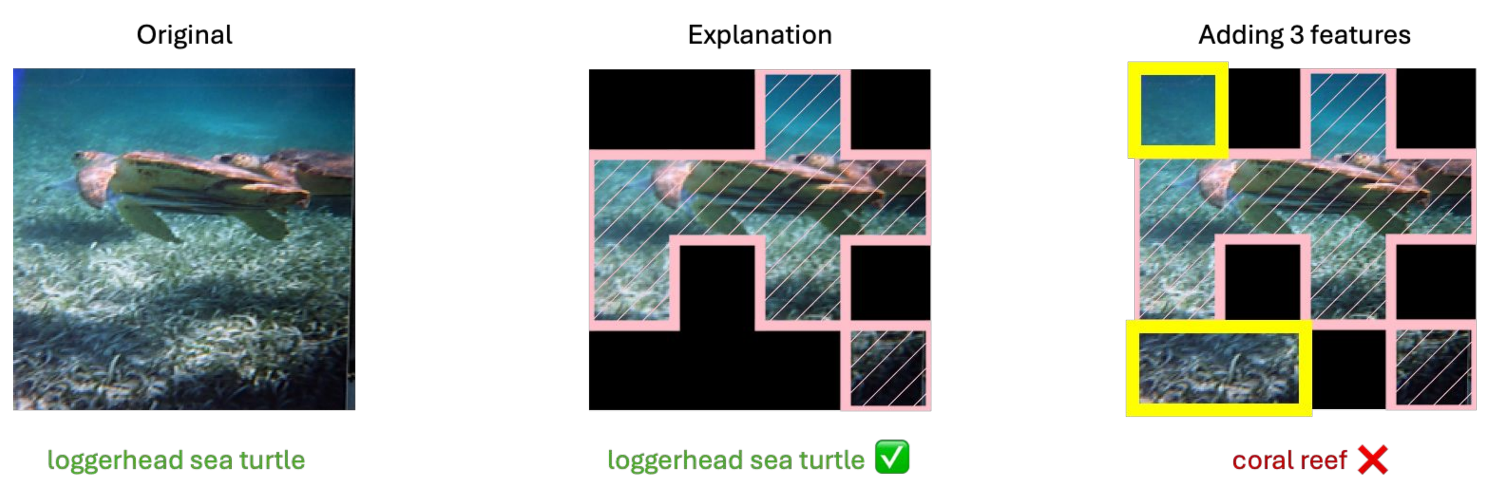
Probabilistic Stability Guarantees for Feature Attributions

Helen Jin, Anton Xue, Weiqiu You, Surbhi Goel, Eric Wong

Motivation

ML models are powerful but opaque – explanations can help elucidate them!

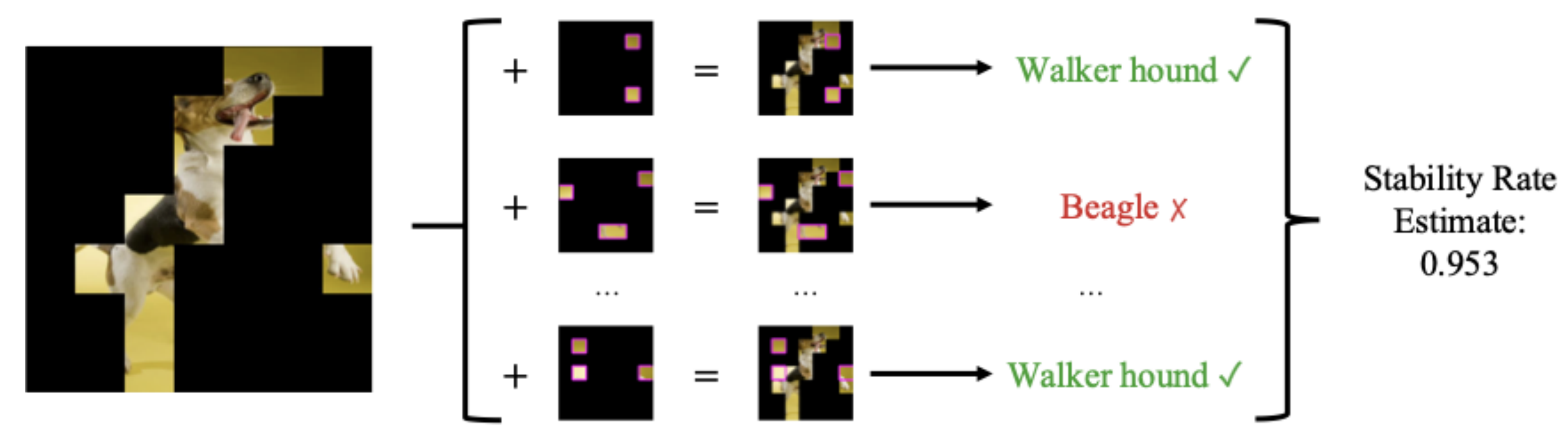


A good explanation should be **robust**: adding extra features should not change the prediction.

→ ***How can we measure the robustness of an explanation?*** 🤔

Stability Certification Algorithm (SCA)

Algorithm. For a certain radius, we **sample** perturbations of up to that radius and **compute** the proportion of prediction matches.



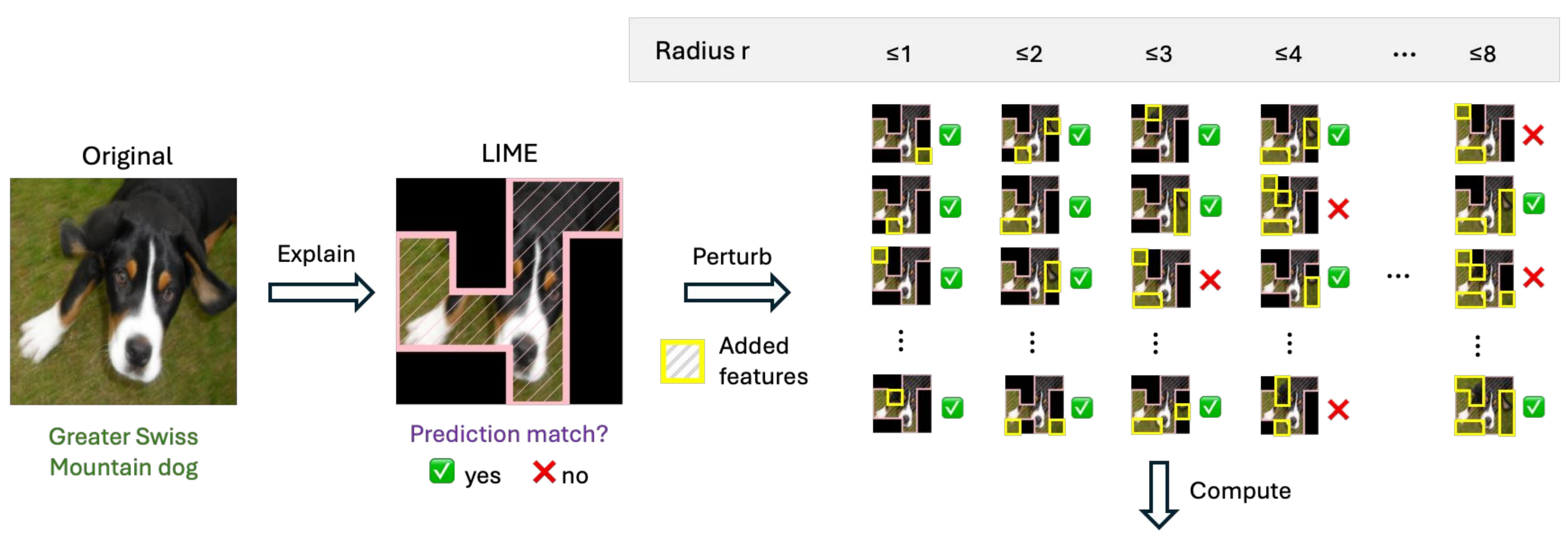
Theorem. If the estimator is computed with $N \geq \log(2/\delta)/(2\varepsilon^2)$ samples, then with probability at least $1 - \delta$, the estimation accuracy compared to the true stability rate is $\leq \varepsilon$.

Soft Stability: a more scalable and flexible guarantee

Definition. [Hard Stability] An explanation is **hard stable** at radius r if including up to any r additional features does not change the prediction.

🚨 Hard stability guarantees rely on specialized architectures and are too conservative to be useful. 🚨

Definition. [Soft Stability] At radius r, an explanation's **stability rate** τ_r is the probability that adding up to r additional features does not change the prediction.

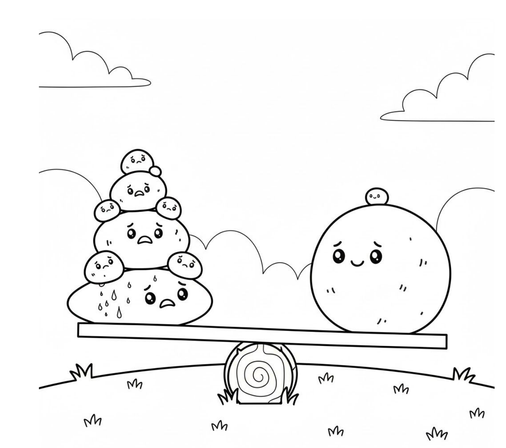


Hard Stable?	Yes	Yes	No	No	...	No
Stability Rates	$\tau_1 = 1.00$	$\tau_2 = 1.00$	$\tau_3 = 0.92$	$\tau_4 = 0.86$	...	$\tau_8 = 0.81$

what are the key benefits soft stability offers?

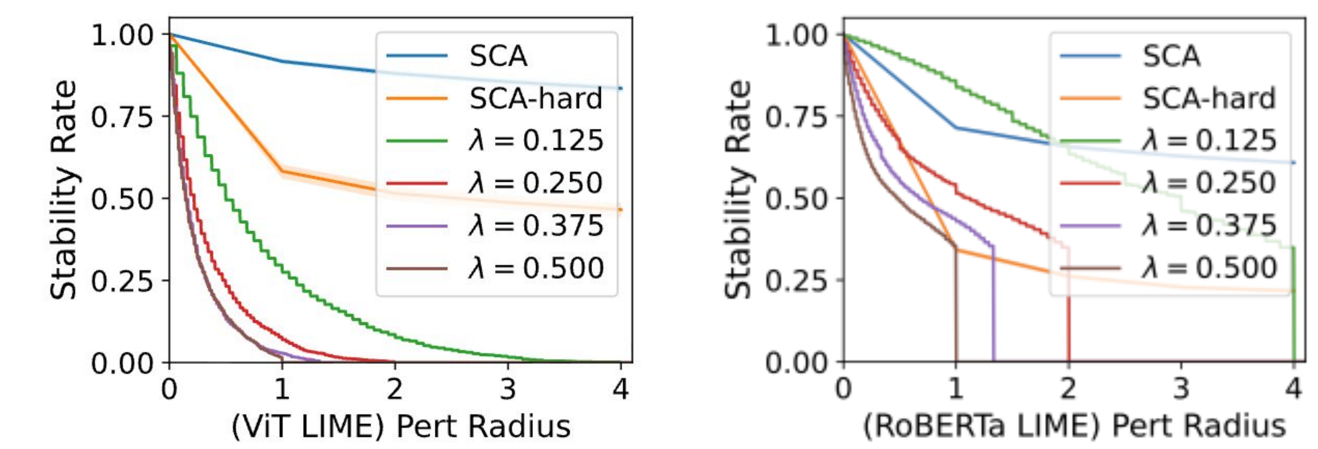
👍 **Model-agnostic certification**: The soft stability rate is efficiently computable for any classifier, whereas hard stability can only certify smoothed classifiers.

👍 **Practical guarantees**: Soft stability certificates scale and are more practically useful than those obtained from hard stability.



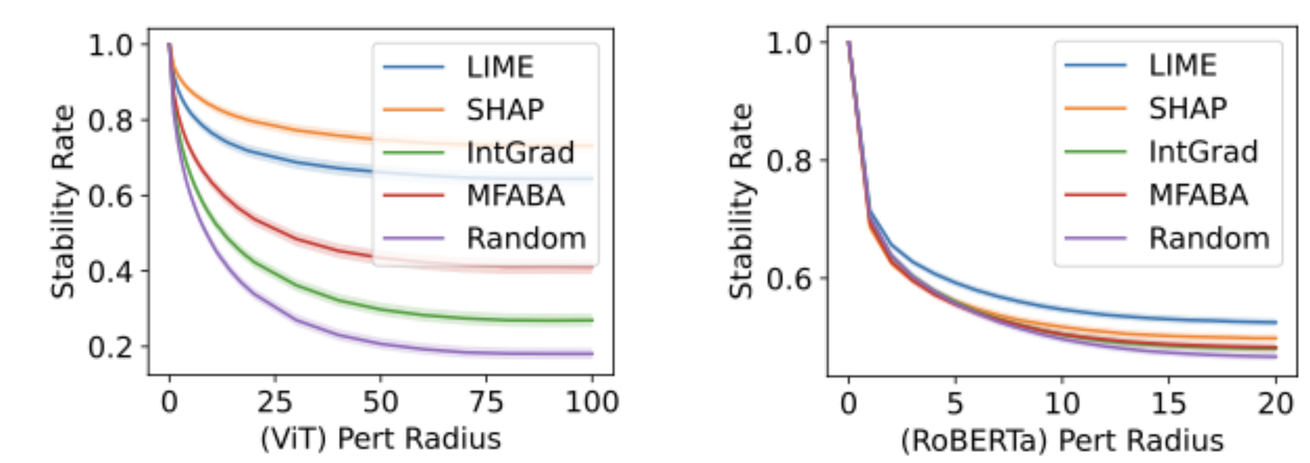
Check out the Blog/Paper!

SCA certifies more



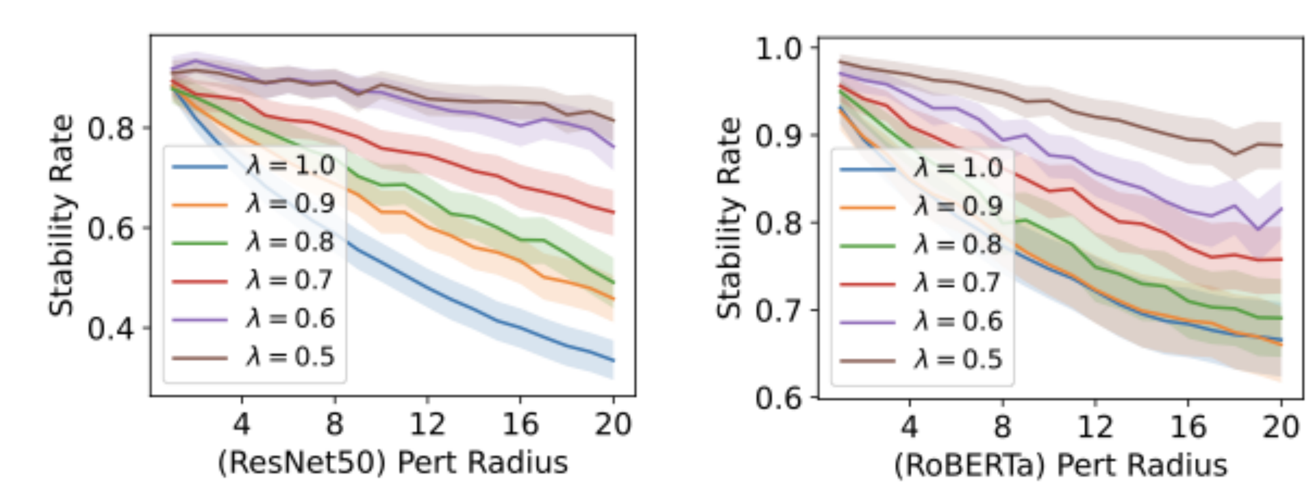
→ At larger radii, SCA yields meaningful soft stability certificates, while hard stability certificates quickly become vacuous as perturbation size grows.

SCA across explanation methods



→ For vision models, LIME and SHAP are more stable than gradient methods, though all beat random; for RoBERTa, differences are smaller.

Mild smoothing can help



→ Mild smoothing ($\lambda \geq 0.5$) can improve stability, especially for RoBERTa.