Stability Guarantees for Feature Attributions with Multiplicative Smoothing

Anton Xue [1] **Eric Wong** [1] **Rajeev Alur** [1]

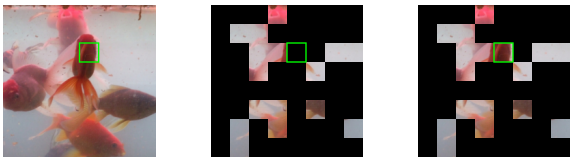
Abstract

Explanation methods for machine learning models tend to not provide any formal guarantees and may not reflect the underlying decision-making process. In this work, we analyze stability as a property for reliable feature attribution methods. We prove that a relaxed variant of stability is guaranteed if the model is sufficiently Lipschitz with respect to the masking of features. To achieve such a model, we develop a smoothing method called Multiplicative Smoothing (MuS). We show that MuS overcomes theoretical limitations of standard smoothing techniques and can be integrated with any classifier and feature attribution method. We evaluate MuS on vision and language models with a variety of feature attribution methods, such as LIME and SHAP, and demonstrate that MuS endows feature attributions with non-trivial stability guarantees.

1. Introduction

Modern machine learning models are incredibly powerful at challenging prediction tasks but notoriously black-box in their decision-making. One can therefore achieve impressive performance without fully understanding *why*. In settings like like medical diagnosis (Reyes et al., 2020; Tjoa & Guan, 2020) and legal analysis (Wachter et al., 2017; Bibal et al., 2021), where accurate and well-justified decisions are important, such power without proof is insufficient. In order to fully wield the power of such models while ensuring reliability and trust, a user needs accurate and insightful *explanations* of model behavior.

One popular family of explanation methods is *feature attributions* (Simonyan et al., 2013; Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017). Given a model and input, a feature attribution method generates a score for each input feature that denotes its importance to the overall prediction. For instance consider Figure 1, in which the Vision Transformer (Dosovitskiy et al., 2020) classifier predicts the full image (left) as "Goldfish". We then use a feature attribution method like SHAP (Lundberg & Lee, 2017) to score each feature and select the top-25%, for which the masked image (middle) is consistently predicted as "Goldfish". However, additionally including a single patch of features (right) alters the prediction confidence so much that it now yields "Axolotl". This suggests that the explanation is brittle (Ghorbani et al., 2019), as small changes easily cause it to now induce some other class. In this paper we study how to overcome such behavior by analyzing the *stability* of an explanation: we consider an explanation to be stable if once the explanatory features are included, the addition of more features does not change the prediction.



Goldfish (98.72%) Cockroach (0.0218%) | Goldfish (95.86%) Axolotl (1.62%) | Axolotl (55.14%) Goldfish (30.24%)

Figure 1. Classification by VisionTransformer (Dosovitskiy et al., 2020) on an attribution generated by SHAP (Lundberg & Lee, 2017) with top-25% selection. A single 28×28 pixel patch of difference between the two attributions (marked **green**) significantly affects prediction confidence and results in a classification flip.

Stability implies that the selected features are enough to explain the prediction (Brown, 2009; Chen et al., 2018; Li et al., 2020) and that this selection maintains strong explanatory power even in the presence of additional information (Ghorbani et al., 2019; Boopathy et al., 2020). Similar properties are studied in literature and identified as useful for interpretability (Nauta et al., 2022), and we emphasize that our main focus is on analyzing and achieving provable guarantees. Stability guarantees in particular are useful as they allow one to accurately predict how model behavior varies with the explanation. Given a stable explanation, one can include more features, e.g. adding context, while maintaining confidence in the consistency of the underlying explanatory power. Crucially, we observe that such guaran-

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence to: Anton Xue <antonxue@seas.upenn.edu>.

Presented at the 2^{nd} Workshop on Formal Verification of Machine Learning, co-located with the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA., 2023.

tees only make sense when jointly considering the model and explanation method: the explanation method necessarily depends on the model to yield an explanation, and stability is then evaluated with respect to the model.

Thus far, existing work on feature attributions with formal guarantees face challenges with computational tractability and explanatory utility. While some methods take an axiomatic approach (Shapley, 1953; Sundararajan et al., 2017), others use metrics that appear reasonable but may not reliably reflect useful model behavior, a common and known limitation (Zhou et al., 2022). Such explanations have been criticized as at best a plausible guess, and at worst completely misleading (Jacovi & Goldberg, 2020).

In this paper we study how to construct explainable models with provable stability guarantees. We jointly consider the classification model and explanation method, and present a formalization for studying such properties that we call *explainable models*. We focus on *binary feature attributions* (Li et al., 2017) wherein each feature is either marked as explanatory (1) or not explanatory (0). We present a method to solve this problem, which is inspired by techniques from adversarial robustness, in particular randomized smoothing (Cohen et al., 2019; Yang et al., 2020). Our method can take *any* off-the-shelf classifier and feature attribution method to efficiently yield an explainable model that satisfies provable stability guarantees. In summary, our contributions are as follows:

- We formalize stability as a key property for binary feature attributions and study this in the framework of explainable models. We prove that relaxed variants of stability are guaranteed if the model is sufficiently Lipschitz with respect to the masking of features.
- To achieve the sufficient Lipschitz conditions, we develop a smoothing method called Multiplicative Smoothing (MuS). We show that MuS achieves strong smoothness conditions, overcomes key theoretical and practical limitations of standard smoothing techniques, and can be integrated with any classifier and feature attribution method.
- We evaluate MuS on vision and language models along with different feature attribution methods. We demonstrate that MuS-smoothed explainable models achieve strong stability guarantees at a small cost to accuracy.

2. Overview

We observe that formal guarantees for explanations must take into account both the model and explanation method, and for this we present in Section 2.1 a pairing that we call *explainable models*. This formulation allows us to then describe the desired stability properties in Section 2.2. We show in Section 2.3 that classifiers with sufficient Lipschitz smoothness with respect to feature masking allows us to yield provable guarantees of stability. Finally in Section 2.4 we show how to adapt existing feature attribution methods into our explainable model framework.

2.1. Explainable Models

We first present explainable models as a formalism for rigorously studying explanations. Let $\mathcal{X} = \mathbb{R}^n$ be the space of inputs, a classifier $f : \mathcal{X} \to [0,1]^m$ maps inputs $x \in \mathcal{X}$ to m logits (class probabilities) that sum to 1, where the class of $f(x) \in [0,1]^m$ is taken to be the largest coordinate. Similarly, an explanation method $\varphi : \mathcal{X} \to \{0,1\}^n$ maps an input $x \in \mathcal{X}$ to an explanation $\varphi(x) \in \{0,1\}^n$ that indicates which features are considered explanatory for the prediction $f(x)$. In particular, we may pick and adapt φ from among a selection of existing feature attribution methods like LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and many others (Simonyan et al., 2013; Sundararajan et al., 2017; Smilkov et al., 2017; Sundararajan & Najmi, 2020; Kwon & Zou, 2022), wherein φ may be thought of as a top-k feature selector. Note that the selection of input features necessarily depends on the explanation method executing or analyzing the model, and so it makes sense to jointly study the model and explanation method: given a classifier f and explanation method φ, we call the pairing $\langle f, \varphi \rangle$ an *explainable model*. Given some $x \in \mathcal{X}$, the explainable model $\langle f, \varphi \rangle$ maps x to both a prediction and explanation. We show this in Figure 2, where $\langle f, \varphi \rangle(x) \in [0,1]^m \times \{0,1\}^n$ pairs the class probabilities and the feature attribution.

For an input $x \in \mathcal{X}$, we will evaluate the quality of the binary feature attribution $\varphi(x)$ through its masking on x. That is, we will study the behavior of f on the masked input $x \odot \varphi(x) \in \mathcal{X}$, where $\odot$ is the element-wise vector product. To do this, we define a notion of *prediction equivalence*: for two $x, x' \in \mathcal{X}$, we write $f(x) \cong f(x')$ to mean that $f(x)$ and $f(x')$ yield the same class. This allows us to formalize the intuition that an explanation $\varphi(x)$ should recover the prediction of x under f.

Definition 2.1. The explainable model $\langle f, \varphi \rangle$ is *consistent* at x if $f(x) \cong f(x \odot \varphi(x))$.

Evaluating f on $x \odot \varphi(x)$ this way lets us apply the model as-is and therefore avoids the challenge of constructing a surrogate model that is accurate to the original (Alizadeh et al., 2020). Moreover, this approach is reasonable, especially in domains like vision — where one intuitively expects that a masked image retaining only the important features should induce the intended prediction. Indeed, architectures like Vision Transformer (Dosovitskiy et al., 2020) can maintain high accuracy with only a fraction of the image present (Salman et al., 2022).

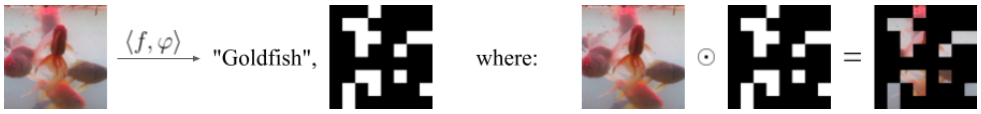


Figure 2. An explainable model $\langle f, \varphi \rangle$ outputs both a classification and a feature attribution. The feature attribution is a binary-valued mask (white 1, black 0) that can be applied over the original input. Here f is Vision Transformer (Dosovitskiy et al., 2020) and φ is SHAP (Lundberg & Lee, 2017) with top-25% feature selection.

Particularly, we would like for $\langle f, \varphi \rangle$ to generate explanations that are stable and concise (i.e. sparse). The former is our central guarantee, and is ensured through smoothing. The latter implies that $\varphi(x)$ has few ones entries, and is desirable since a good explanation should not contain too much redundant information. However, sparsity is a difficulty property to enforce, as this is contingent on the model having high accuracy with respect to heavily masked inputs. For sparsity we present a simple heuristic in Section 2.4.

2.2. Stability Properties of Explainable Models

Given an explainable model $\langle f, \varphi \rangle$ and some $x \in \mathcal{X}$, stability means that the prediction does not change even if one adds more explanatory features to $\varphi(x)$. For instance, the model-explanation pair in Figure 1 is *not* stable, as the inclusion of a single feature group (patch) changes the prediction. To formalize this notion of stability, we first introduce a partial ordering: for $\alpha, \alpha' \in \{0,1\}^n$, we write $\alpha \succeq \alpha'$ iff $\alpha_i \geq \alpha'_i$ for all $i = 1, \ldots, n$. That is, $\alpha \succeq \alpha'$ iff α includes all the features selected by α'.

Definition 2.2. The explainable model $\langle f, \varphi \rangle$ is stable at x if $f(x \odot \alpha) \cong f(x \odot \varphi(x))$ for all $\alpha \succeq \varphi(x)$.

Note that the constant explanation $\varphi(x) = \mathbf{1}$, the vector of ones, makes $\langle f, \varphi \rangle$ trivially stable at every $x \in \mathcal{X}$, though this is not a concise explanation. Additionally, stability at x implies consistency at x.

Unfortunately, stability is a difficult property to enforce in general, as it requires that f satisfy a monotone-like behavior with respect to feature inclusion — which is especially challenging for complex models like neural networks. Checking stability without additional assumptions on f is also hard: if $k = \|\varphi(x)\|_1$ is the number of ones in $\varphi(x)$, then there are 2^{n-k} possible $\alpha \succeq \varphi(x)$ to check. This large space of possible $\alpha \succeq \varphi(x)$ motivates us to instead examine *relaxations* of stability. We introduce lower and upper-relaxations of stability below.

Definition 2.3. The explainable model $\langle f, \varphi \rangle$ is incrementally stable at x with radius r if $f(x \odot \alpha) \cong f(x \odot \varphi(x))$ for all $\alpha \succeq \varphi(x)$ where $\|\alpha - \varphi(x)\|_1 \leq r$.

Incremental stability is the lower-relaxation since it considers the case where the mask α has only a few features more than $\varphi(x)$. For instance, if one can provably add up to r features to a masked $x \odot \varphi(x)$ without altering the prediction, then $\langle f, \varphi \rangle$ would be incrementally stable at x with radius r. We next introduce the upper-relaxation that we call decremental stability.

Definition 2.4. The explainable model $\langle f, \varphi \rangle$ is decrementally stable at x with radius r if $f(x \odot \alpha) \cong f(x \odot \varphi(x))$ for all $\alpha \succeq \varphi(x)$ where $\|\mathbf{1} - \alpha\|_1 \leq r$.

Decremental stability is a subtractive form of stability, in contrast to the additive nature of incremental stability. Particularly, decremental stability considers the case where α has much more features than $\varphi(x)$. If one can provably remove up to r features from the full x without altering the prediction, then $\langle f, \varphi \rangle$ is decrementally stable at x with radius r. Note that decremental stability necessarily entails consistency of $\langle f, \varphi \rangle$, but for simplicity of definitions we do not enforce this for incremental stability. Finally, note that for sufficiently large radius of $r = \lceil (n - \|\varphi(x)\|_1)/2 \rceil$, incremental and decremental stability together imply stability.

Remark 2.5. Similar notions to the above have been proposed in literature, and we refer to (Nauta et al., 2022) for an extensive survey. In particular for (Nauta et al., 2022), consistency is akin to *preservation* and stability is similar to *continuity*, except we are concerned with adding features. Also, incremental stability is most similar to *incremental addition* and decremental stability to *incremental deletion*.

2.3. Lipschitz Smoothness Entails Stability Guarantees

If $f : \mathcal{X} \to [0,1]^m$ is Lipschitz with respect to the masking of features, then we can guarantee relaxed stability properties for the explainable model $\langle f, \varphi \rangle$. In particular, we require for all $x \in \mathcal{X}$ that $f(x \odot \alpha)$ is Lipschitz with respect to the mask $\alpha \in \{0,1\}^n$. This then allows us to presented our main results in smoothness and stability, which we present in Section 3.1. A sketch of the stability result is first given below in Remark 2.6.

Remark 2.6 (Sketch of main result). Consider an explainable model $\langle f, \varphi \rangle$ where for all $x \in \mathcal{X}$ the function $g(x, \alpha) = f(x \odot \alpha)$ is λ-Lipschitz in $\alpha \in \{0,1\}^n$ with respect to the ℓ^1 norm. Then at any x, the radius of incremental stability r_{inc} and radius of decremental stability r_{dec} are respectively:

$$r_{\text{inc}} = \big[g_A(x, \varphi(x)) - g_B(x, \varphi(x))\big]/(2\lambda),$$
$$r_{\text{dec}} = \big[g_A(x, \mathbf{1}) - g_B(x, \mathbf{1})\big]/(2\lambda),$$

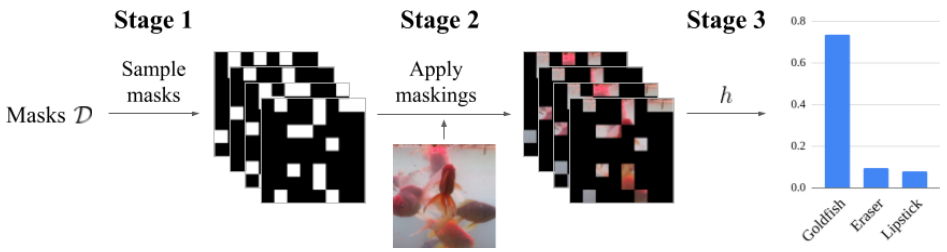


Figure 3. Evaluating $f(x)$ is done in three stages. **(Stage 1)** Generate N samples of binary masks $s^{(1)}, \ldots, s^{(N)} \in \{0,1\}^n$, where each coordinate is Bernoulli with parameter λ (here $\lambda = 1/4$). **(Stage 2)** Apply each mask on the input to yield $x \odot s^{(i)}$ for $i = 1, \ldots, N$. **(Stage 3)** Average over $h(x \odot s^{(i)})$ to compute $f(x)$, and note that the predicted class is given by a weighted average.

with g_A, g_B the first and second-largest logits defined as

$$g_A(x,\alpha) = g_{k^\star}(x,\alpha), \quad g_B(x,\alpha) = \max_{i \neq k^\star} g_i(x,\alpha), \quad (1)$$

where the largest logit index is $k^\star = \operatorname*{argmax}_{1 \leq k \leq m} g_k(x,\alpha)$.

Observe that Lipschitz smoothness is in fact a stronger assumption than necessary, as besides $\alpha \succeq \varphi(x)$ it also imposes guarantees on $\alpha \preceq \varphi(x)$. Nevertheless, Lipschitz smoothness is one of the few classes of properties that can be guaranteed and analyzed at scale on arbitrary models (Yang et al., 2020; Levine & Feizi, 2021). Importantly, we may apriori pick the Lipschitz constant λ for our smoothed classifier, thereby allowing us to establish known guarantees ahead of test time. The details for establishing the Lipschitz constant through our randomized smoothing method are described in Theorem 3.1.

2.4. Adapting Existing Feature Attribution Methods

Most existing feature attribution methods assign a real-valued score to feature importance, rather than a binary value. We therefore need to convert this to a binary-valued method for use with a stable explainable model. Let $\psi : \mathcal{X} \to \mathbb{R}^n$ be such a continuous-valued method like LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017), and fix some desired incremental stability radius r_{inc} and decremental stability radius r_{dec}. Given some $x \in \mathcal{X}$ a simple construction for binary $\varphi(x) \in \{0,1\}^n$ then follows.

Remark 2.7 (Iterative construction of $\varphi(x)$). Consider any $x \in \mathcal{X}$ and let ρ be an index ordering on $\psi(x)$ from high-to-low (i.e. largest logit first). Initialize $\alpha = \mathbf{0}$, and for each $i \in \rho$: assign $\alpha_i \leftarrow 1$ then check whether $\langle f, \varphi : x \mapsto \alpha \rangle$ is now consistent, incrementally stable with radius r_{inc}, and decrementally stable with radius r_{dec}. If so then terminate with $\varphi(x) = \alpha$; otherwise continue.

3. Multiplicative Smoothing for Lipschitz Constants

In this section we present our main technical contribution in Multiplicative Smoothing (MuS). The goal is to transform an arbitrary base classifier $h : \mathcal{X} \to [0,1]^m$ into a smoothed classifier $f : \mathcal{X} \to [0,1]^m$ that is Lipschitz with respect to the masking of features. This then allows one to appropriately couple an explanation method φ with f to form an explainable model $\langle f, \varphi \rangle$ with provable stability guarantees. An extended discussion of results is given in Appendix A.

3.1. Technical Overview of MuS

Our key insight is that randomly dropping (i.e. zeroing) features attains the desired smoothness. In particular, we uniformly drop features with probability $1 - \lambda$ by sampling binary masks $s \in \{0,1\}^n$ from some distribution $\mathcal{D}$ where each coordinate is distributed as $\Pr[s_i = 1] = \lambda$. Then define f as follows:

$$f(x) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot s), \quad (2)$$

such that $s_i \sim \mathcal{B}(\lambda)$ for $i = 1, \ldots, n$, where $\mathcal{B}(\lambda)$ is the Beronulli distribution with parameter $\lambda \in [0,1]$. We give an overview of evaluating $f(x)$ in Figure 3: each coordinate of the random mask $s \in \{0,1\}^n$ is 1 with probability λ, meaning that the masked input $x \odot s$ retains λ fraction of its original features on average. Importantly, our main results of smoothness (Theorem 3.1) and stability (Theorem 3.2) hold provided $\mathcal{D}$ is coordinate-wise Bernoulli with λ, and so we avoid restricting ourselves to any one particular choice of $\mathcal{D}$ until necessary. However, it will be easy to intuit the exposition with $\mathcal{D} = \mathcal{B}^n(\lambda)$, the coordinate-wise i.i.d. Bernoulli distribution with λ.

We can equivalently parametrize f using the mapping $g(x,\alpha) = f(x \odot \alpha)$, where it follows that:

$$g(x,\alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot \tilde{\alpha}), \quad \tilde{\alpha} = \alpha \odot s. \quad (3)$$

Note that one could have alternatively first defined g and then f due to the identity $g(x, \mathbf{1}) = f(x)$. We require that the relationship between f and g follows an identity that we call *masking equivalence*:

$$g(x \odot \alpha, \mathbf{1}) = f(x \odot \alpha) = g(x,\alpha), \quad (4)$$

for all $x \in \mathcal{X}$ and $\alpha \in \{0,1\}^n$. This follows by definition of g, and the relevance to stability is this: if masking equivalence holds, then we can rewrite stability properties involving f in terms of g's second parameter as follows:

$$f(x \odot \alpha) = g(x, \alpha) \cong g(x, \varphi(x)) = f(x \odot \varphi(x)) \quad \text{(c.f. Definition 2.2)}$$

for all $\alpha \succeq \varphi(x)$, where incremental and decremental stability may be analogously defined. This translation is useful, as we will prove that g is λ-Lipschitz in its second parameter (Theorem 3.1), which then allows us to establish the desired stability properties (Theorem 3.2). Importantly, we are motivated to develop MuS because standard smoothing techniques, namely additive smoothing (Cohen et al., 2019; Yang et al., 2020), may fail to satisfy masking equivalence. This is further explained in Section A.1.

We do not enforce a specific construction for $\mathcal{D}$, since many choices are in fact valid. Rather, so long as each coordinate of $s \sim \mathcal{D}$ obeys $s_i \sim \mathcal{B}(\lambda)$ then the Lipschitz properties for g follow. The implication here is that although simple distributions like $\mathcal{B}^n(\lambda)$ suffices for $\mathcal{D}$, they may not be sample efficient. We show in Section A.2 how to exploit and construct statistical dependence in order to reduce the sample complexity of computing MuS.

3.2. Certifying Stability with Lipschitz Classifiers

Our core technical result is in showing that f as defined in (2) is Lipschitz to the masking of features. We present MuS in terms of g, where it is parametric with respect to the distribution $\mathcal{D}$: so long as $\mathcal{D}$ satisfies a coordinate-wise Bernoulli condition, then it is usable with MuS.

Theorem 3.1 (MuS). *Let $\mathcal{D}$ be any distribution on $\{0,1\}^n$ where each coordinates of $s \sim \mathcal{D}$ is distributed as $s_i \sim \mathcal{B}(\lambda)$. Consider any $h : \mathcal{X} \to [0,1]$ and define $g : \mathcal{X} \times \{0,1\}^n \to [0,1]$ as*

$$g(x, \alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot \tilde{\alpha}), \quad \tilde{\alpha} = \alpha \odot s.$$

Then the function $g(x, \cdot) : \{0,1\}^n \to [0,1]$ is λ-Lipschitz in the ℓ^1 norm for all $x \in \mathcal{X}$.

The strength of this result is in its weak assumptions. First, the theorem applies to any model h and input $x \in \mathcal{X}$. It further suffices that each coordinate is distributed as $s_i \sim \mathcal{B}(\lambda)$, and we emphasize that statistical independence between different s_i, s_j is *not assumed*. This allows us to construct $\mathcal{D}$ with structured dependence in Section A.2, such that we may exactly and efficiently evaluate $g(x, \alpha)$ at a sample complexity of $N \ll 2^n$. A low sample complexity is important for making MuS practically usable, as otherwise one must settle for of the expected value subject to probabilistic guarantees. For instance, simpler distributions like $\mathcal{B}^n(\lambda)$ do in fact satisfy the requirements of Theorem 3.1 — but costs 2^n samples because of coordinate-wise independence. Whatever choice of valid $\mathcal{D}$, one can guarantee stability so long as g is Lipschitz in its second argument.

Theorem 3.2 (Stability). *Consider any $h : \mathcal{X} \to [0,1]^m$ with coordinates $h_1, \ldots, h_m$. Fix $\lambda \in [0,1]$ and let $g_1, \ldots, g_m$ be the respectively smoothed coordinates as in Theorem 3.1, using which we analogously define $g : \mathcal{X} \times \{0,1\}^n \to [0,1]^m$. Also define $f(x) = g(x, \mathbf{1})$. Then for any explanation method φ and input $x \in \mathcal{X}$, the explainable model $\langle f, \varphi \rangle$ is incrementally stable with radius r_{inc} and decrementally stable with radius r_{dec}:*

$$r_{\text{inc}} = \frac{g_A(x, \varphi(x)) - g_B(x, \varphi(x))}{2\lambda},$$
$$r_{\text{dec}} = \frac{g_A(x, \mathbf{1}) - g_B(x, \mathbf{1})}{2\lambda},$$

where g_A, g_B are the first and second largest logits as in (1).

Note that it is only in the case where the radius ≥ 1 do non-trivial stability guarantees exist. Because each g_k has range in $[0,1]$, this means that a Lipschitz constant of $\lambda \leq 1/2$ is necessary to attain at least one radius of stability. We present in Section B.2 some extensions to MuS that allows one to achieve higher coverage of features.

4. Empirical Evaluations

(Experimental Setup) Due to space limitations we highlight a subset of our results and refer to Appendix C for comprehensive experiments. In this section we show results with Vision Transformer (Dosovitskiy et al., 2020) and ImageNet1K (Russakovsky et al., 2015). We group features on the $3 \times 224 \times 224$ dimensional input into $n = 64$ superpixels, and report stability radii r as a fraction of the features, i.e. r/n. For methods we use SHAP (Lundberg & Lee, 2017) with top-25% feature selection. A sample size of $N = 2000$ of ImageNet1K was used for all experiments here.

4.1. (E1) How Good are the Stability Guarantees?

There exists a natural measure of quality for stability guarantees over a dataset: what radii are achieved, and at what frequency. We investigate how different combinations of models, explanation methods, and λ affect this measure. To do this we plot the rate at which a property holds at a radius r as a function of the radius (expresesd as r/n). We show our results in Figure 4, where we show consistent and incremental stability (left) and consistent and decremental stability (right). Although we can achieve non-trivial incremental stability, decremental stability is easier to certify at larger radii — and this is reasonable, since for incremental stability evaluation the classifier sees a masked $x \odot \varphi(x)$.

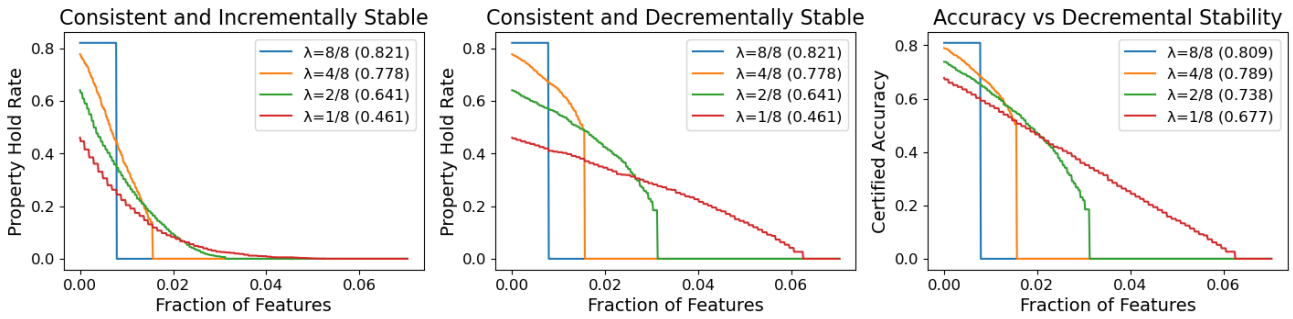


Figure 4. Experiments are run with $\langle f, \varphi \rangle$ where f is a smoothed Vision Transformer and φ is SHAP with top-25% feature selection. We use $N = 2000$ samples from ImageNet1K, radius r are reported as fraction of input covered, i.e. r/n. (Left; Middle) Consistency and incremental (resp. decremental) stability. *Property hold rate* is the fraction of images that are consistent and stable up to radius r when using a mask from SHAP-top25%. (Right) Overall accuracy vs the radius of decremental stability. *Certified accuracy* is the fraction of images for which f predicts the true label on the entire unmasked x while achieving decremental stability at radius r.

4.2. (E2) What is the Cost of Smoothing?

To increase the radius of a provable stability guarantee, we must decrease the Lipschitz constant λ. As λ decreases, however, more features are dropped during the smoothing process. To study the stability-accuracy trade-off, we plotted the accuracy attained by the smoothed classifier vs. the radius of decremental stability and show the results in Figure 4 (right), where as expected the clean accuracy (in parentheses) decreases with λ. For Vision Transformer we see that the accuracy remains high even under non-trivial noise.

5. Conclusion

We study provable stability guarantees for binary feature attribution methods through the framework of explainable models. A selection of features is stable if the additional inclusion of other features do not alter its explanatory power. We show that if the classifier is Lipschitz with respect to the masking of features, then one can guarantee relaxed variants of stability. To achieve this Lipschitz condition we develop a smoothing method called Multiplicative Smoothing (MuS). We show that MuS yields strong stability guarantees at only a small cost to accuracy.

References

Alizadeh, R., Allen, J. K., and Mistree, F. Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design*, 31:275–298, 2020.

Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

Bibal, A., Lognoul, M., De Streel, A., and Frénay, B. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29:149–169, 2021.

Boopathy, A., Liu, S., Zhang, G., Liu, C., Chen, P.-Y., Chang, S., and Daniel, L. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pp. 1014–1023. PMLR, 2020.

Brown, G. A new perspective for information theoretic feature selection. In *Artificial intelligence and statistics*, pp. 49–56. PMLR, 2009.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pp. 883–892. PMLR, 2018.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

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.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kwon, Y. and Zou, J. Weightedshap: analyzing and improving shapley based feature attributions. *arXiv preprint arXiv:2209.13429*, 2022.

Levine, A. J. and Feizi, S. Improved, deterministic smoothing for l_1 certified robustness. In *International Conference on Machine Learning*, pp. 6254–6264. PMLR, 2021.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

Li, X., Wang, Y., and Ruiz, R. A survey on sparse learning models for feature selection. *IEEE transactions on cybernetics*, 52(3):1642–1660, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022.

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., Tengg-Kobligk, H. v., Summers, R. M., and Wiest, R. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Salman, H., Jain, S., Wong, E., and Madry, A. Certified patch robustness via smoothed vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15137–15147, 2022.

Shapley, L. A value for n-person games. *Contributions to the Theory of Games*, pp. 307–317, 1953.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Tjoa, E. and Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., and Li, J. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9623–9633, 2022.

A. Extended Results for Section 3

We give an extended discussion of content from Section 3.

A.1. Standard Smoothing Does Not Satisfy Masking Equivalence

We are motivated to develop MuS because standard smoothing techniques, namely additive smoothing (Cohen et al., 2019; Yang et al., 2020), may fail to satisfy masking equivalence. Additive smoothing is by far the most popular smoothing technique, and differs from our scheme (3) in how noise is applied, where let $\mathcal{D}_{\text{add}}$ and $\mathcal{D}_{\text{mult}}$ be any two distributions on $\mathbb{R}^n$:

$$g(x,\alpha) = \mathop{\mathbb{E}}_{s\sim\mathcal{D}} h(x \odot \tilde{\alpha}), \quad \tilde{\alpha} = \begin{cases} \alpha + s, & s \sim \mathcal{D}_{\text{add}} \\ \alpha \odot s, & s \sim \mathcal{D}_{\text{mult}} \end{cases}$$

where $\mathcal{D}_{\text{add}}$ denotes additive smoothing, and $\mathcal{D}_{\text{mult}}$ denotes multiplicative smoothing. Particularly, additive smoothing has counterexamples to masking equivalence.

Proposition A.1. *There exists $h : \mathcal{X} \to [0,1]$ and distribution $\mathcal{D}$, where for*

$$g^+(x,\alpha) = \mathop{\mathbb{E}}_{s\sim\mathcal{D}} h(x \odot \tilde{\alpha}), \quad \tilde{\alpha} = \alpha + s,$$

we have $g^+(x,\alpha) \neq g^+(x \odot \alpha, \mathbf{1})$ for some $x \in \mathcal{X}$ and $\alpha \in \{0,1\}^n$.

Proof. Observe that it suffices to have h, x, α such that $h(x \odot (\alpha + s)) > h((x \odot \alpha) \odot (\mathbf{1} + s))$ for a non-empty set of $s \in \mathbb{R}^n$. Let $\mathcal{D}$ be a distribution on these s, then:

$$\begin{aligned} g^+(x,\alpha) &= \mathop{\mathbb{E}}_{s\sim\mathcal{D}} h(x \odot (\alpha + s)) \\ &> \mathop{\mathbb{E}}_{s\sim\mathcal{D}} h((x \odot \alpha) \odot (\mathbf{1} + s)) \\ &= g^+(x \odot \alpha, \mathbf{1}) \end{aligned}$$

□

Intuitively, this occurs because additive smoothing primarily applies noise by perturbing feature values, rather than completely masking them. As such, there might be "information leakage" when non-explanatory bits of α are changed into non-zero values. This then causes each sample of $h(x \odot \tilde{\alpha})$ within $g(x,\alpha)$ to observe more features of x than it would have been able to otherwise.

A.2. Exploiting Structured Dependency

We now present $\mathcal{L}_{qv}(\lambda)$, a distribution on $\{0,1\}^n$ that allows for efficient and exact evaluation of a MuS-smoothed classifier. Our construction is an adaption of (Levine & Feizi, 2021) from uniform to Bernoulli noise, where the primary insight is that one can parametrize n-dimensional noise using a single dimension via structured coordinate-wise dependence. In particular, we use a *seed vector* v, where with an integer *quantization parameter* $q > 1$ there will only exist q distinct choices of $s \sim \mathcal{L}_{qv}(\lambda)$. All the while, we still enforce that any such s is coordinate-wise Bernoulli with $s_i \sim \mathcal{B}(\lambda)$. Thus for a sufficiently small quantization parameter (i.e. $q \ll 2^n$) we may tractably enumerate through all q possible choices of s and thereby evaluate a MuS-smoothed model with only q samples.

Proposition A.2. *Fix integer $q > 1$ and consider any vector $v \in \{0, 1/q, \ldots, (q-1)/q\}^n$ and scalar $\lambda \in \{1/q, \ldots, q/q\}$. Define $s \sim \mathcal{L}_{qv}(\lambda)$ to be a random vector in $\{0,1\}^n$ with coordinates given by*

$$s_i = \mathbb{I}[t_i \leq \lambda], \quad t_i = v_i + s_{\text{base}} \mod 1,$$

where $s_{\text{base}} \sim \mathcal{U}(\{1/q, \ldots, q/q\}) - 1/(2q)$. Then there are q distinct values of s and each coordinate is distributed as $s_i \sim \mathcal{B}(\lambda)$.

Proof. First, observe that each of the q distinct values of s_{base} defines a unique value of s, since we have assumed v and λ to be fixed. Next, observe that each t_i has q unique values uniformly distributed as $t_i \sim \mathcal{U}(1/q, \ldots, q/q\}) - 1/(2q)$. Because $\lambda \in \{1/q, \ldots, q/q\}$ we therefore have $\Pr[t_i \leq \lambda] = \lambda$, which implies that $s_i \sim \mathcal{B}(\lambda)$. □

The seed vector v is the source of our structured coordinate-wise dependence and the one-dimensional source of randomness s_{base} is used to generate the n-dimensional s. Such $s \sim \mathcal{L}_{qv}(\lambda)$ then satisfies the conditions for use in MuS (Theorem 3.1), and this noise allows for an exact evaluation of the smoothed classifier in q samples. We have found $q = 64$ to be sufficient in practice and values as low as $q = 16$ to also yield good performance. We remark that one drawback is that one may get an unlucky seed v, but we have not yet observed this in our experiments.

B. Proofs and Extensions

Here we present the proofs of our main results, as well as some extensions to MuS.

B.1. Proofs of Main Results

B.1.1. Proof of Theorem 3.1

Proof. By linearity we have:

$$\begin{aligned} g(x,\alpha) - g(x,\alpha') = \mathop{\mathbb{E}}_{s\sim\mathcal{D}} h(x \odot \tilde{\alpha}) - h(x \odot \tilde{\alpha}'), \\ \tilde{\alpha} = \alpha \odot s, \quad \tilde{\alpha}' = \alpha' \odot s, \end{aligned}$$

so it suffices to analyze an arbitrary term by fixing some $s \sim \mathcal{D}$. Consider any $x \in \mathcal{X}$, let $\alpha, \alpha' \in \{0,1\}^n$, and define $\delta = \alpha - \alpha'$. Observe that $\tilde{\alpha}_i \neq \tilde{\alpha}'_i$ exactly when

$|\delta_i| = 1$ and $s_i = 1$. Since $s_i \sim \mathcal{B}(\lambda)$, we thus have $\Pr[\tilde{\alpha}_i \neq \tilde{\alpha}'_i] = \lambda|\delta_i|$, and applying the union bound:

$$\Pr_{s\sim\mathcal{D}}[\tilde{\alpha} \neq \tilde{\alpha}'] = \Pr_{s\sim\mathcal{D}}[\bigcup_{i=1}^{n} \tilde{\alpha}_i \neq \tilde{\alpha}'_i] \leq \sum_{i=1}^{n} \lambda|\delta_i| = \lambda\|\delta\|_1,$$

and so:

$$\begin{aligned}
&|g(x,\alpha) - g(x,\alpha')| \\
&= \left| \mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}')] \right| \\
&= \Big| \Pr_{s\sim\mathcal{D}}[\tilde{\alpha} \neq \tilde{\alpha}'] \cdot \mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}') \,|\, \tilde{\alpha} \neq \tilde{\alpha}'] \\
&\quad - \Pr_{s\sim\mathcal{D}}[\tilde{\alpha} = \tilde{\alpha}'] \cdot \mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}') \,|\, \tilde{\alpha} = \tilde{\alpha}'] \Big|.
\end{aligned}$$

Note that $\mathbb{E}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}') \,|\, \tilde{\alpha} = \tilde{\alpha}'] = 0$, and so

$$\begin{aligned}
&|g(x,\alpha) - g(x,\alpha')| \\
&= \Pr_{s\sim\mathcal{D}}[\tilde{\alpha} \neq \tilde{\alpha}'] \cdot \underbrace{\left| \mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}') \,|\, \tilde{\alpha} \neq \tilde{\alpha}'] \right|}_{\leq 1 \text{ because } h(\cdot)\in[0,1]} \\
&\leq \Pr_{s\sim\mathcal{D}}[\tilde{\alpha} \neq \tilde{\alpha}'] \leq \lambda\|\delta\|_1.
\end{aligned}$$

Thus, $g(x,\cdot)$ is λ-Lipschitz in the ℓ^1 norm. □

B.1.2. Proof of Theorem 3.2

Proof. We first show incremental stability. Consider any $x \in \mathcal{X}$, then by masking equivalence:

$$f(x\odot\varphi(x)) = g(x\odot\varphi(x), \mathbf{1}) = g(x, \varphi(x)),$$

and let g_A, g_B be the top two logits of g as defined in (1). By Theorem 3.1, both g_A, g_B are Lipschitz in their second parameter, and so for all $\alpha \in \{0,1\}^n$:

$$\begin{aligned}
\|g_A(x,\varphi(x)) - g_A(x,\alpha)\|_1 &\leq \lambda\|\varphi(x) - \alpha\|_1 \\
\|g_B(x,\varphi(x)) - g_B(x,\alpha)\|_1 &\leq \lambda\|\varphi(x) - \alpha\|_1
\end{aligned}$$

Observe that if α is sufficiently close to $\varphi(x)$, i.e.:

$$2\lambda\|\varphi(x) - \alpha\|_1 \leq g_A(x,\varphi(x)) - g_B(x,\varphi(x)),$$

then the top logit index of $g(x,\varphi(x)$ and $g(x,\alpha)$ are the same. This means that $g(x,\varphi(x)) \cong g(x,\alpha)$ and thus $f(x\odot\varphi(x)) \cong f(x\odot\alpha)$, thus proving incremental stability with radius $d(x,\varphi(x))/(2\lambda)$.

The decremental stability case is similar, except we replace $\varphi(x)$ with $\mathbf{1}$. □

B.2. Some Basic Extensions

Below we present some extensions to MuS that help increase the fraction of the input to which we can guarantee stability.

B.2.1. Feature Grouping

We have so far assumed that $\mathcal{X} = \mathbb{R}^n$, but sometimes it may be desirable to group features together, e.g. color channels of the same pixel. Our results also hold for more general $\mathcal{X} = \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$, where for such $x \in \mathcal{X}$ and $\alpha \in \mathbb{R}^n$ we lift $\odot$ as:

$$\odot : \mathcal{X} \times \mathbb{R}^n \to \mathcal{X}, \quad (x\odot\alpha)_i = x_i \cdot \mathbb{I}[\alpha_i = 1] \in \mathbb{R}^{d_i}.$$

All of our proofs are identical under this construction, with the exception of the dimensionalities of terms like $(x\odot\alpha)$. An example of feature grouping is given in Figure 1.

C. All Experiments

Models, Datasets, and Explanation Methods We evaluate on two vision models (Vision Transformer (Dosovitskiy et al., 2020) and ResNet50 (He et al., 2016)) and one language model (RoBERTa (Liu et al., 2019)). For the vision dataset we use ImageNet1K (Russakovsky et al., 2015) and for the language dataset we use TweetEval (Barbieri et al., 2020) sentiment analysis. We use four explanation methods in SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), Integrated Gradients (IGrad) (Sundararajan et al., 2017), and Vanilla Gradient Saliency (VGrad) (Simonyan et al., 2013); we take $\varphi(x)$ as the top-k weighted features.

Training Details We use Adam (Kingma & Ba, 2014) as our optimizer with default parameters and learning rate 10^{-6}. For each $\lambda \in \{1/8, \ldots, 8/8\}$ we fine-tuned each model for 1 epoch, which results in a total of $8 \times 3 = 24$ models used in our experiments. To train with a particular λ: for each training input x we generate two random maskings — one where λ of the features are zerod and one where $\lambda/2$ of the features are zerod. This additional $\lambda/2$ zeroing is to account for the fact that inputs to a smoothed model will be subject to masking by λ as well as $\varphi(x)$, where the scaling factor of $1/2$ is informed by our prior experience about the size of a stable explanation.

Miscellaneous Preprocessing For images in ImageNet1K we use feature grouping (Appendix B.2.1) to group the $3 \times 224 \times 224$ dimensional image into patches of size $3 \times 28 \times 28$, such that there remains $n = 64$ feature groups. Each feature of a feature group then receives the same value of noise during smoothing. We report radii of stability as a fraction of the feature groups covered. For example, if at some input from ImageNet1K we get an incremental stability radius of r, then we report $r/64$ as the fraction of features up to which we are guaranteed to be stable. This is especially amenable to evaluating RoBERTa on TweetEval where inputs do not have uniform token lengths, i.e. do not have uniform feature dimensions. In all of our experiments we use the quantized noise as in Appendix A.2 with a quantization parameter of

$q = 64$, with the exception of Appendix C.2 where for the box attack search we use $q = 16$.

Our experiments are organized as follows:

- (Appendix C.1) What is the quality of stability guarantees from MuS?
- (Appendix C.2) What is the theoretical vs empirical stability that can be guaranteed?
- (Appendix C.3) What are the stability-accuracy trade-offs due to smoothing?
- (Appendix C.4) Which explanation method is best?

C.1. Quality of Stability Guarantees

Here we study what radii of stability are certifiable, and how often these can be achieved with different models and explanation methods. We therefore consider explainable models $\langle f, \varphi \rangle$ constructed from base models $h \in \{\text{Vision Transformer}, \text{ResNet50}, \text{RoBERTa}\}$ and explanation methods $\varphi \in \{\text{SHAP}, \text{LIME}, \text{IGrad}, \text{VGrad}\}$ with top-$k \in \{1/8, 2/8, 3/8\}$ feature selection. We take $N = 2000$ samples from each model's respective datasets and compute the following value for each radius:

$$\text{value}(r) = \frac{\#\left\{x : \begin{array}{c}\langle f, \varphi \rangle \text{ consistent and inc (dec)} \\ \text{stable with radius} \leq r\end{array}\right\}}{N}.$$

The results are shown in the following figures, where plots of incremental stability are on the left; plots of decremental stability are on the right.

- Figure 5: Vision Transformer with SHAP and LIME
- Figure 6: Vision Transformer with IGrad and VGrad
- Figure 7: ResNet50 with SHAP and LIME
- Figure 8: ResNet50 with IGrad and VGrad
- Figure 9: RoBERTa with SHAP and LIME
- Figure 10: RoBERTa with IGrad and VGrad

C.2. Theoretical vs Empirical

We compare the certifiable theoretical stability guarantees with what is empirically attained via a standard box attack search (Chen et al., 2017). This is an extension of Section 4.2, where we now show all models as evaluated with SHAP-top25%. The certified plots are identical from Appendix C.1. We take $N_{\text{cert}} = 2000$ samples for the certified plots, and $N_{\text{emp}} = 250$ for the empirical plots. This comparatively small selection of methods and data is because box attack is very time-intensive to run, requiring several minutes per model, method, and λ combination. The plots are shown in Figure 11.

C.3. Stability-Accuracy Trade-Offs

We study how the accuracy degrades with λ. We consider a smoothed model f constructed from a base classifier $h \in \{\text{Vision Transformer}, \text{ResNet50}, \text{RoBERTa}\}$ and vary $\lambda \in \{1/16, 1/8, 2/8, 4/8, 8/8\}$. We then take $N = 2000$ samples from each respective dataset and measure the accuracy of f at different radii. We use $f(x) \cong \texttt{true\_label}$ to mean that f attained the correct prediction at $x \in \mathcal{X}$, and we plot the following value at each radius r:

$$\text{value}(r) = \frac{\#\left\{x : \begin{array}{c}f(x) \cong \texttt{true\_label} \text{ and} \\ \text{dec stable with radius} \leq r\end{array}\right\}}{N}$$

The overall accuracy with each λ is shown in the parentheses of each plot's legend. The plots are shown in Figure 12.

C.4. Which Explanation Method is the Best?

We first investigate how may features are needed to yield consistent and non-trivially stable explanations, as done by the greedy selection algorithm in Section 2.4. For some $x \in \mathcal{X}$, let k_x denote the fraction of features that $\langle f, \varphi \rangle$ needs to be consistent, incrementally stable, and decrementally stable with radius 1. We vary $\lambda \in \{1/8, \ldots, 4/8\}$, where recall $\lambda \leq 4/8$ is needed for non-trivial stability, and use $N = 250$ samples to plot the average k_x. The plots are shown in Figure 13.

We next investigate the ability of each method to predict features that lead to high accuracy. Let $f(x \odot \varphi(x)) \cong \texttt{true\_label}$, mean that the masked input $x \odot \varphi(x)$ yields the correct prediction. We then plot this accuracy as we vary the top-$k \in \{1/8, 2/8, 3/8\}$ for different methods φ, and $\lambda \in \{1/8, \ldots, 8/8\}$, using $N = 2000$ samples. The plots are shown in Figure 14.

C.5. Discussion

Effect of Smoothing We observe that smoothing can yield non-trivial stability guarantees, especially for Vision Transformer and RoBERTa, as evidenced in Appendix C.1. We see that smoothing is least detrimental on these two transformer-based architectures, and most negatively impacts the performance of ResNet50. We conjecture that although different training set-ups may improve performance across every category, our experiments still serves to illustrate the general trend.

Theoretical vs Empirical It is expected that the certifiable radii of stability is more conservative than what is empirically observed. As mentioned in Section 3.2, for each λ there is a maximum radius to which stability can be guaranteed, which is an inherent limitation of using logit gaps and Lipschitz constants as the main theoretical technique. We emphasize that the notion of stability need not be tied

to smoothing, though we are currently not aware of other viable approaches.

Why these Explanation Methods? We chose SHAP, LIME, IGrad, and VGrad from among the large variety of methods available primarily due to their popularity, and because we believe that they are collectively representative of many techniques. In particular, we believe that LIME remains representative baseline for surrogate model-based explanation methods. SHAP and IGrad are, to our knowledge, the two most well-known families of axiomatic feature attribution methods. Finally, we believe that VGrad is representative of a traditional gradient saliency-based approach.

Which Explanation Method is the Best? Based on our experiments in Appendix C.4 we see that SHAP generally achieves higher accuracy using the same amount of top-k features as other methods. On the other hand, VGrad tends to perform poorly. We remark that there is well-known critique against the usefulness of saliency-based explanation methods (Kindermans et al., 2019).

D. Miscellaneous

Relevance to Other Explanation Methods Our key theoretical contribution of MuS in Theorem 3.1 is a general-purpose smoothing method that is distinct from standard smoothing techniques, namely additive smoothing. MuS is therefore applicable to other problem domains beyond what is studied in this paper, and would be useful where Lipschitz constants with respect to maskings is desirable.

Broader Impacts Reliable explanations are necessary for making well-informed decisions, and are increasingly important as machine learning models are integrated with fields like medicine, law, and business — where the primary users may not be well-versed in the technical limitations of different methods. Formal guarantees are therefore important for ensuring the predictability and reliability of complex system, which then allows users to construct accurate mental models of interaction and behavior. In this work we study a particular kind of guarantee known as stability, which is key to feature attribution-based explanation methods.

E. All Figures

All remaining figures are shown in the following.

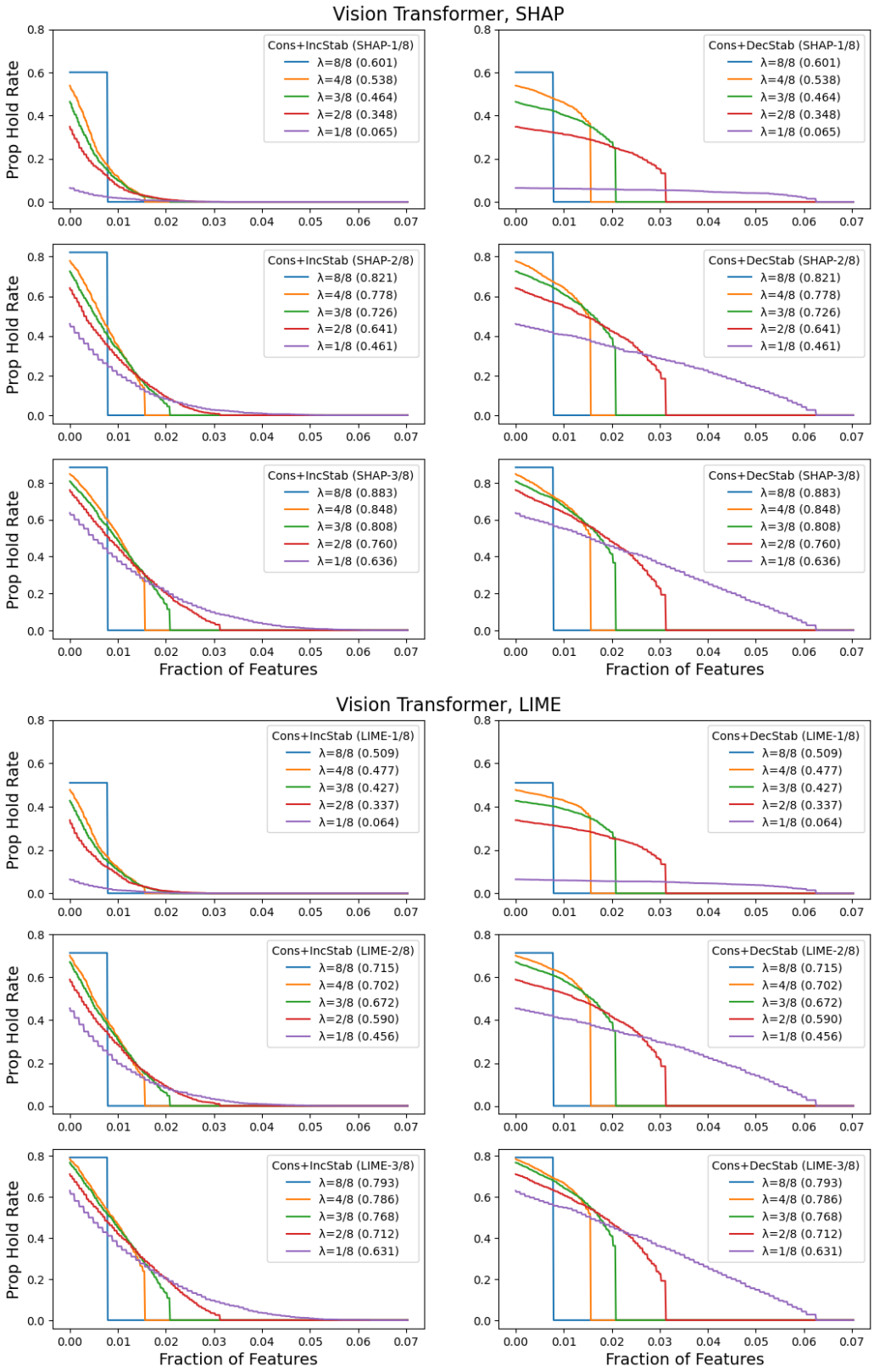


Figure 5. (Top) Vision Transformer with SHAP; (Bottom) Vision Transformer with LIME. (Left) consistent and incrementally stable; (Right) consistent and decrementally stable.

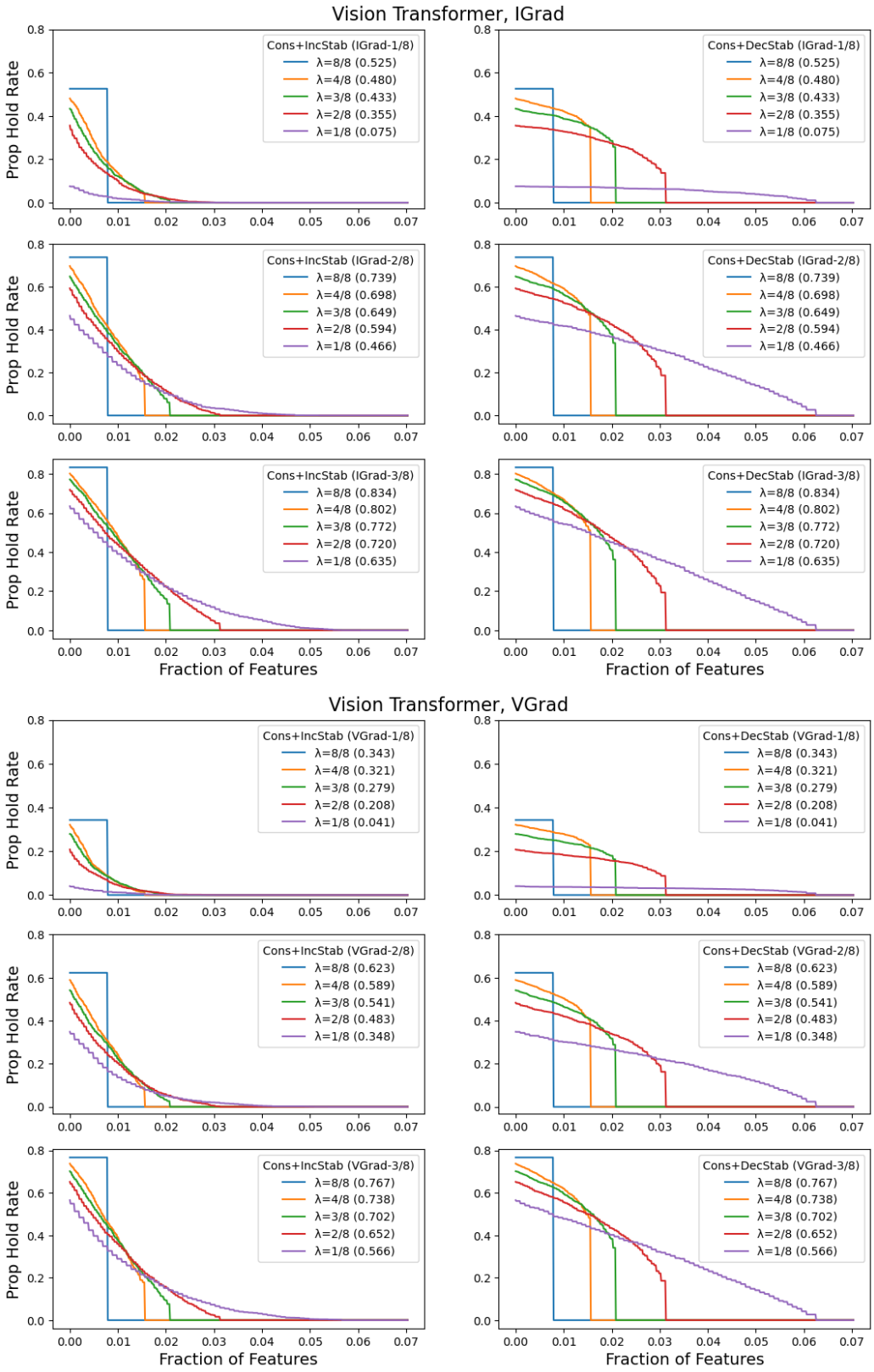


Figure 6. (Top) Vision Transformer with IGrad; (Bottom) Vision Transformer with VGrad. (Left) consistent and incrementally stable; (Right) consistent and decrementally stable.

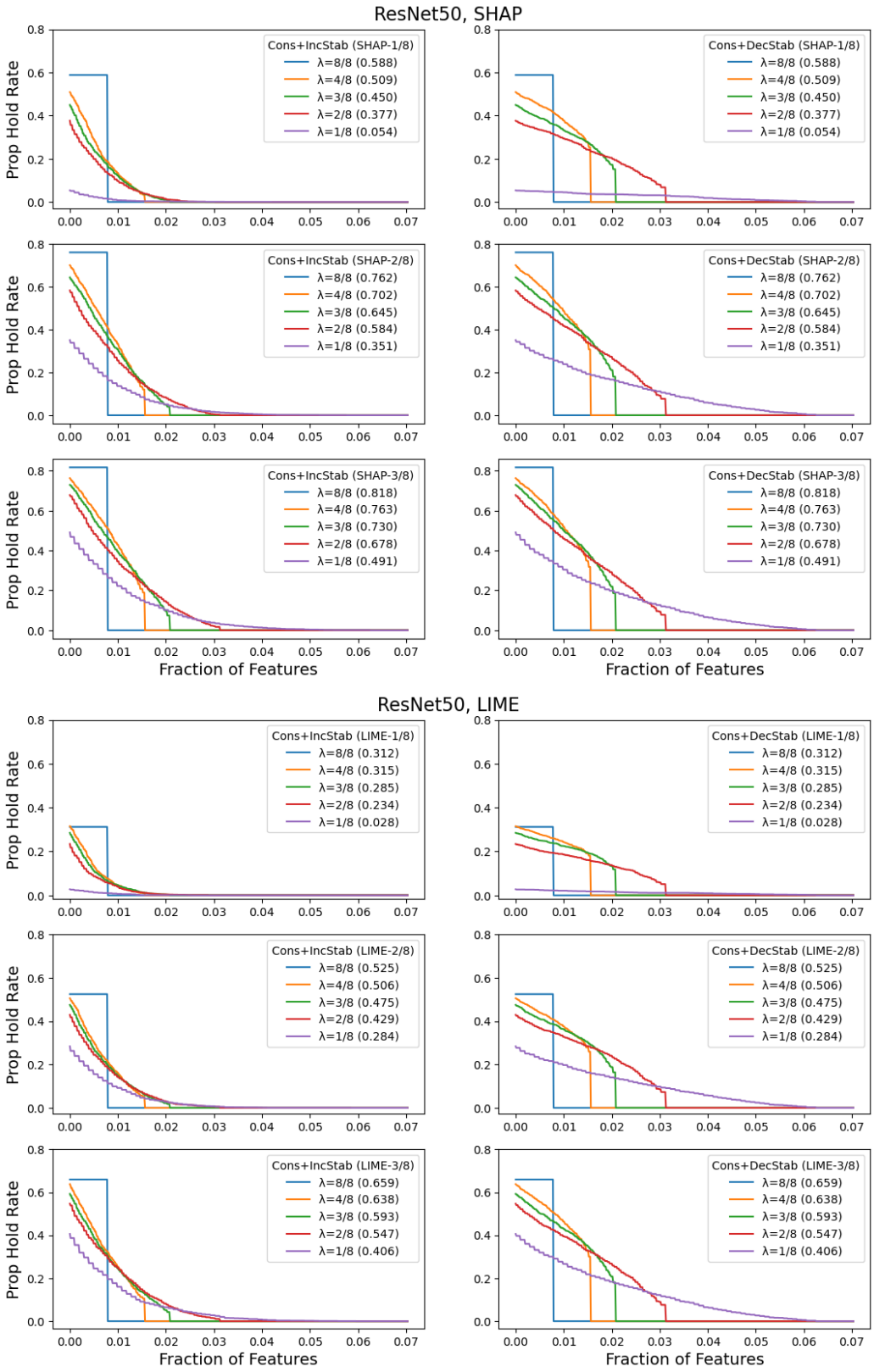


Figure 7. (Top) ResNet50 with SHAP; (Bottom) ResNet50 with LIME. (Left) consistent and incrementally stable; (Right) consistent and decrementally stable.

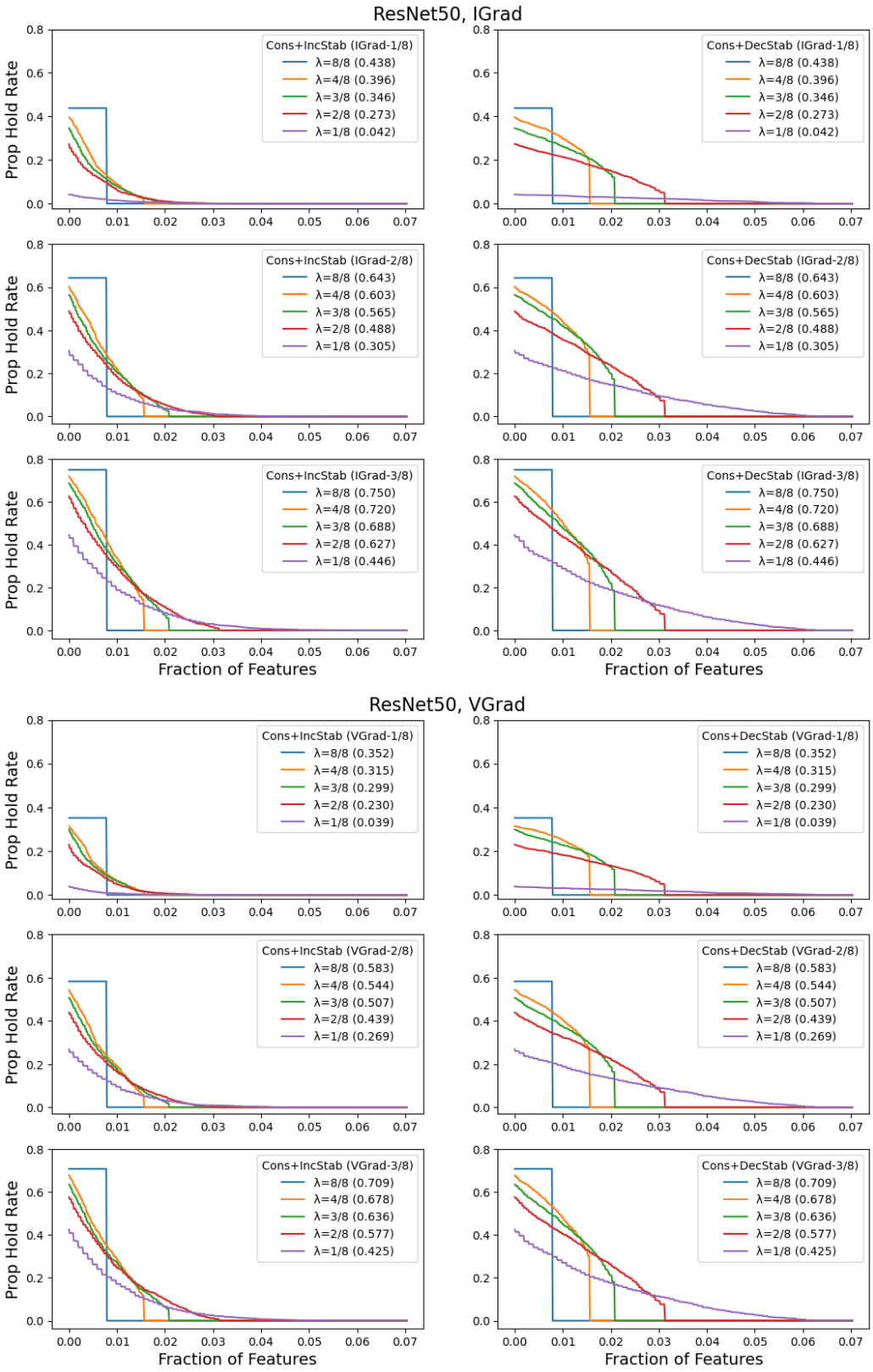


Figure 8. (Top) ResNet50 with IGrad; (Bottom) ResNet50 with VGrad. (Left) consistent and incrementally stable; (Right) consistent and decrementally stable.

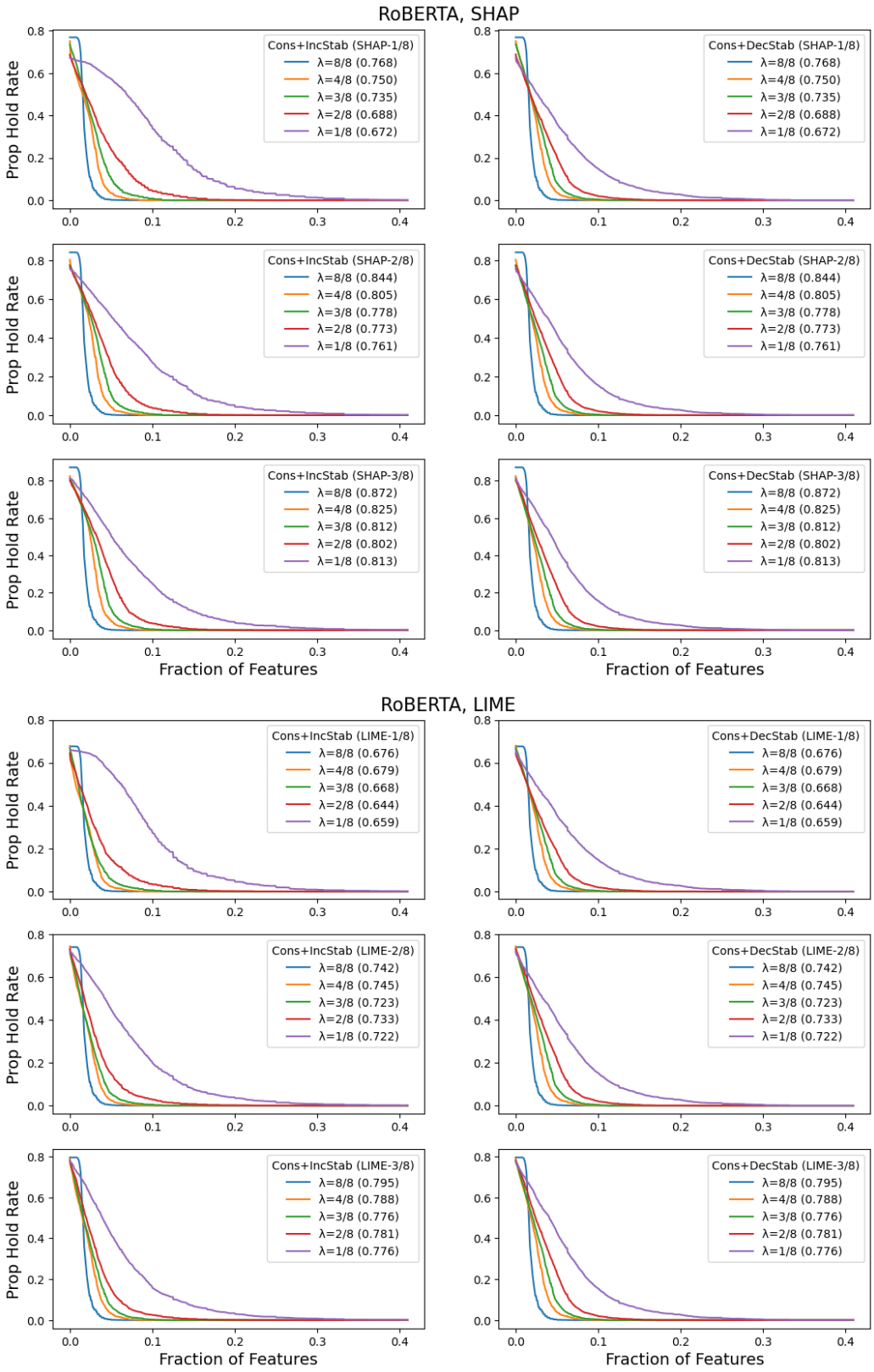


Figure 9. (Top) RoBERTa with SHAP; (Bottom) RoBERTa with LIME. (Left) consistent and incrementally stable; (Right) consistent and decrementally stable.

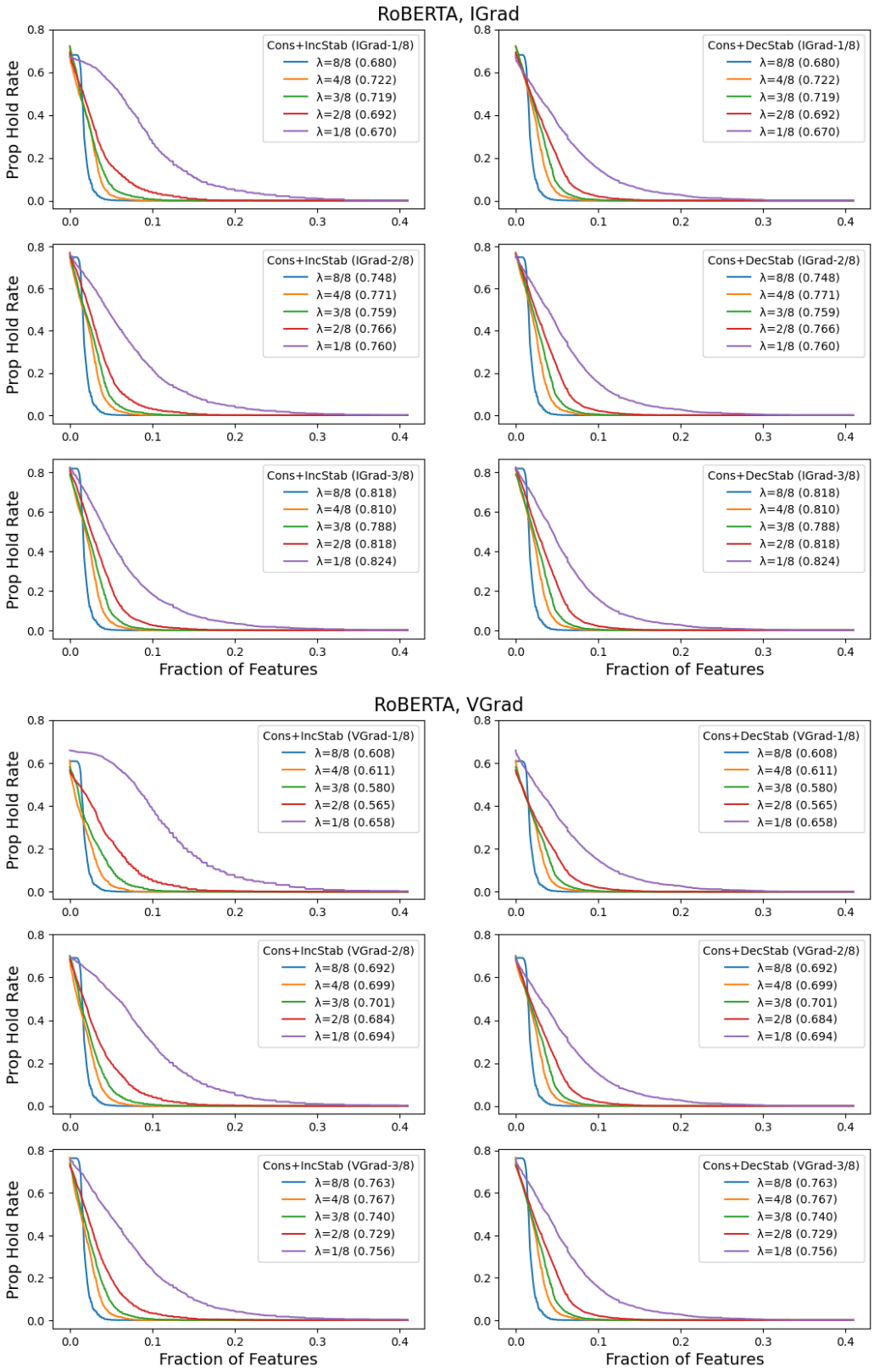


Figure 10. (Top) RoBERTa with IGrad; (Bottom) RoBERTa with VGrad. (Left) consistent and incrementally stable; (Right) consistent and decrementally stable.

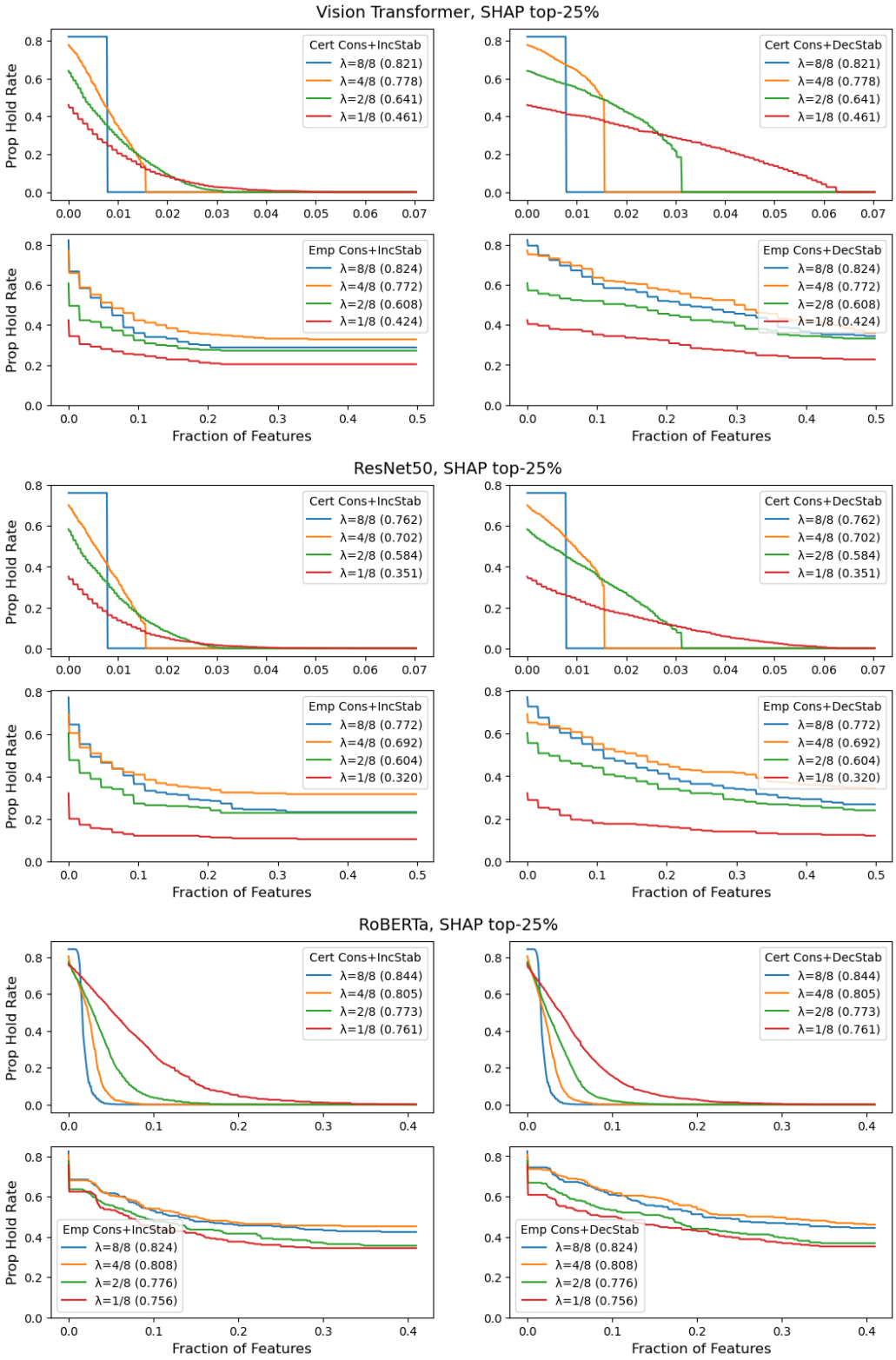


Figure 11. With SHAP top-25%: (Top) Vision Transformer; (Middle) ResNet50; (Bottom) RoBERTa.

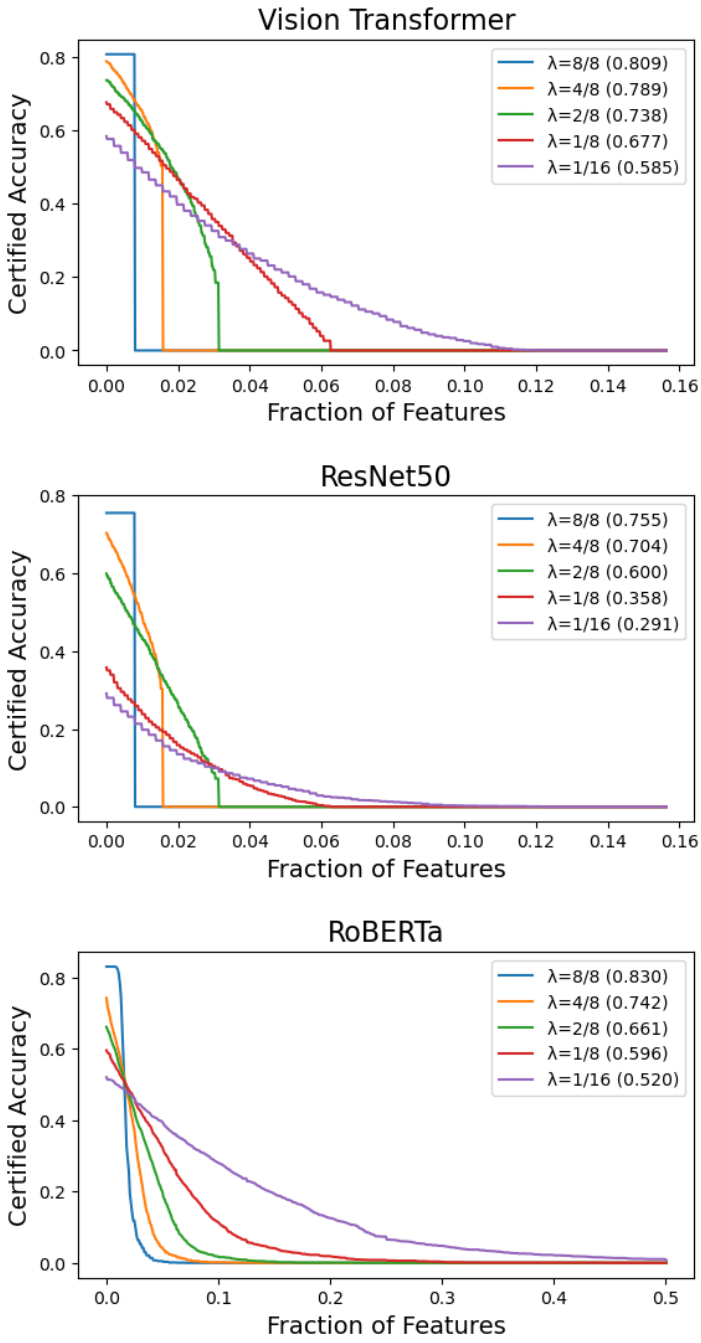


Figure 12. (Top) Vision Transformer; (Middle) ResNet50; (Bottom) RoBERTa.

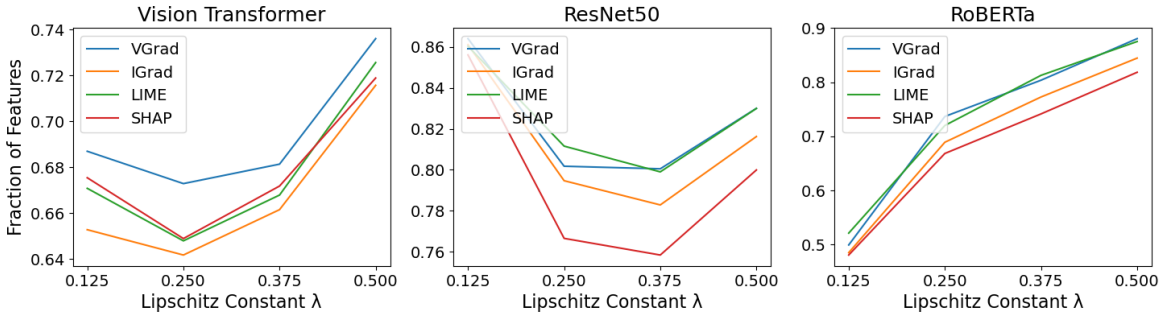


Figure 13. (Left) Vision Transformer; (Middle) ResNet50; (Right) RoBERTa.

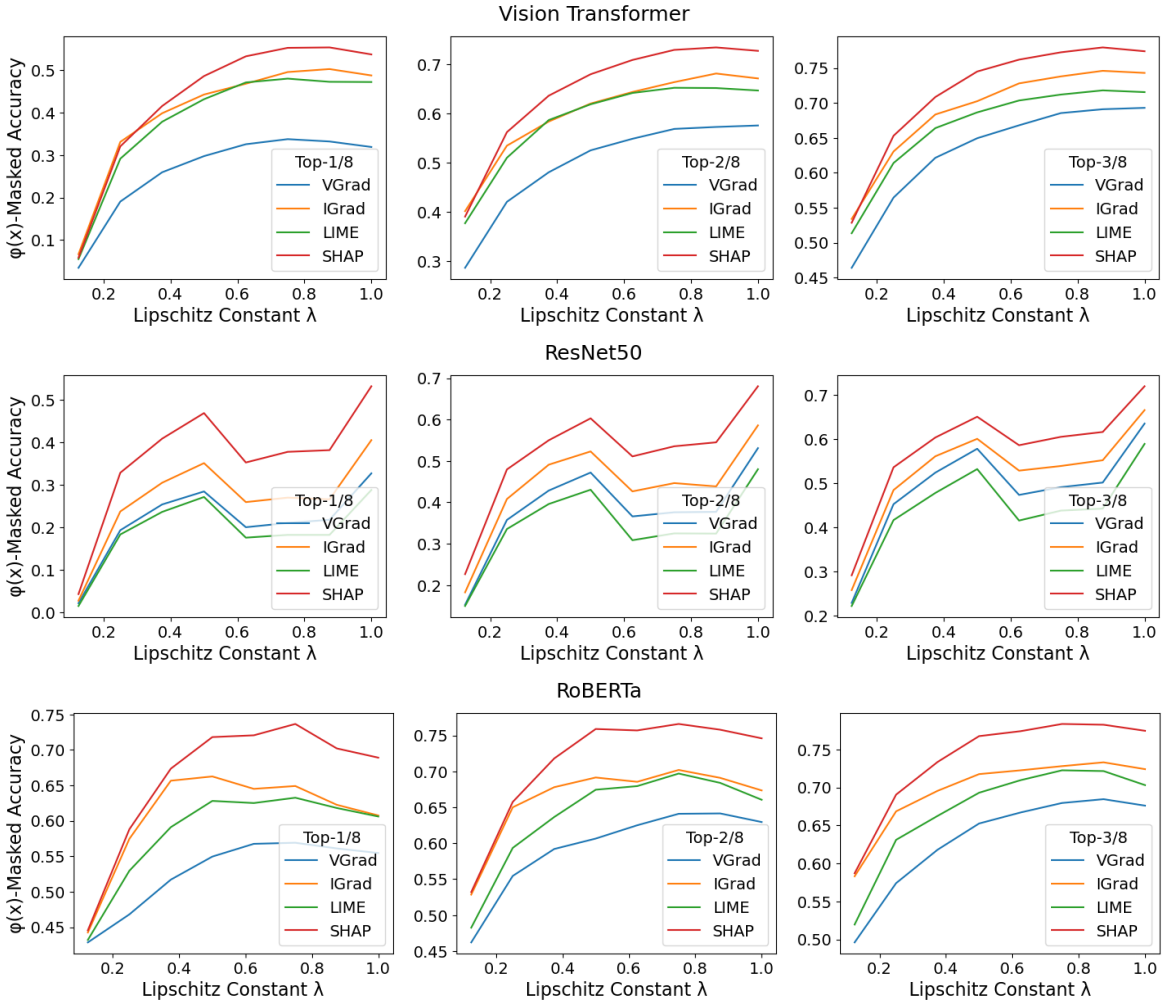


Figure 14. (Top) Vision Transformer; (Middle) ResNet50; (Bottom) RoBERTa.