VISION DIFFMASK: Faithful Interpretation of Vision Transformers with Differentiable Patch Masking

A. Nalmpantis*, A. Panagiotopoulos*, K. Papakostas*, J. Gkountouras*, W. Aziz

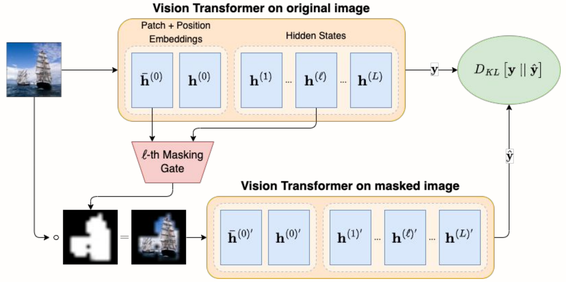
←Live Demo
Codebase→

This work received funding from the EU's Horizon Europe RIA (UTTER, contract 101070631).

Introduction

- We introduce a novel post-hoc interpretation method for the Vision Transformer, based on DiffMask [1].
- Our method is *neither* a gradient-based *nor* an attention-based method
- Instead, our method trains a **set of probing models** to find **the minimal subset of an image** that produces the same output distribution with the whole image
- Next to the standard benchmarks, we introduce a **new task** to measure the **faithfulness** of interpretability methods.

Methodology



Gating Mechanism

$$\mathbf{u}^{(\ell)} = \alpha \cdot \mathrm{MLP}([\bar{\mathbf{h}}^{(0)}; \mathbf{h}^{(\ell)}]) + \beta \cdot \mathbf{1}$$

Mask Generation

$$\mathbf{z}^{(\ell)} \sim \mathrm{HardConcrete}(\mathbf{z}^{(\ell)}; \mathbf{u}^{(\ell)}, l, r)$$

Learnable Baseline

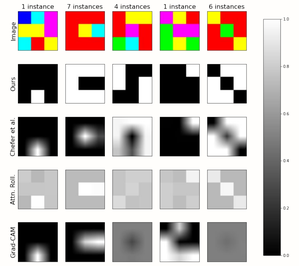
$$\tilde{\mathbf{x}}_i = z_i \cdot \mathbf{x}_i + (1 - z_i) \cdot \mathbf{b}$$

Objective

$$\min_{\boldsymbol{\phi}, \mathbf{b}} \sum_{\mathbf{x} \in \mathcal{D}} \mathcal{L}_0(\boldsymbol{\phi}, \mathbf{b} \mid \mathbf{x}) \quad \text{s.t. } D_{\mathrm{KL}}[\mathbf{y} \,||\, \hat{\mathbf{y}}] \leq m$$

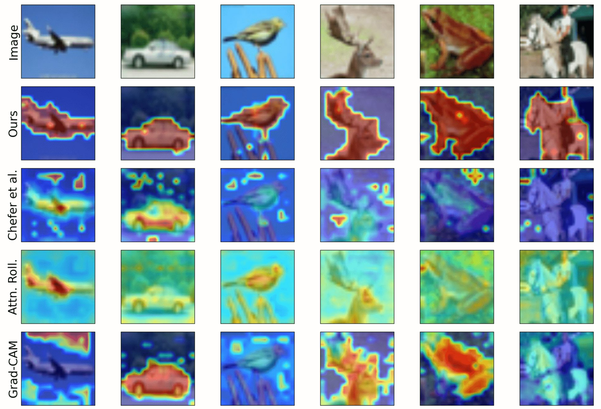
Faithfulness Test

- We train a ViT to *count the number of red patches in an image*
- In this scenario, we **know** what the model should be looking at to make a decision: *either or the red or all the non-red patches*
- Vision DiffMask is the *only* method that is truly **faithful** to the behavior that we anticipate.
- Other methods *both* fail some case *and* are inconsistent between choosing the red patches or their complement

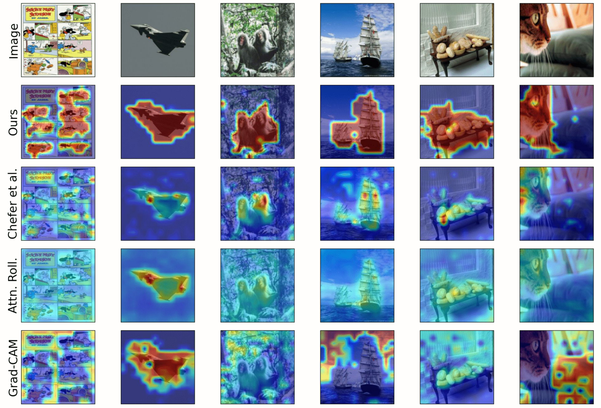


Qualitative Results

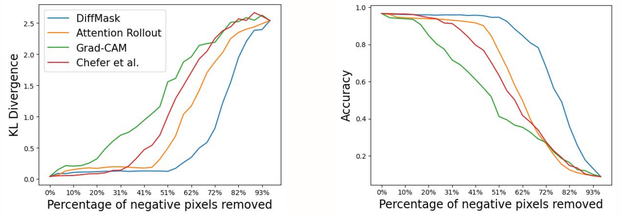
CIFAR-10



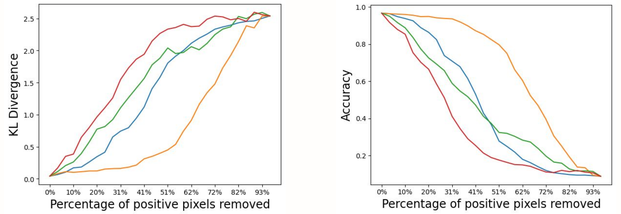
ImageNet



Quantitative Results



Negative perturbation curves. We measure ViT's performance when removing pixels with a low attribution score, following [2]. Low KL divergence / high accuracy are better.



Positive perturbation curves. We measure ViT's performance when removing pixels with a high attribution score, following [2]. High KL divergence / low accuracy are better.

Conclusion

- We introduced VISION DIFFMASK, a new method for post-hoc interpretability of images.
- VISION DIFFMASK's attributions are experimentally proven to be **consistent** and **faithful**, while also **plausible** to what a human would expect.
- Our method is able to **generalize** well across different datasets (please check our demo 🤗)

References

[1] De Cao et al. *How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking.* EMNLP 2020
[2] Chefer et al. *Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers* ICCV 2021