$\mathcal{M}^4$: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models

Xuhong Li [1], Mengnan Du [2], Jiamin Chen [1], Yekun Chai [1], Himabindu Lakkaraju [3], Haoyi Xiong [1]

[1] Baidu Inc. [2] New Jersey Institute of Technology [3] Harvard University

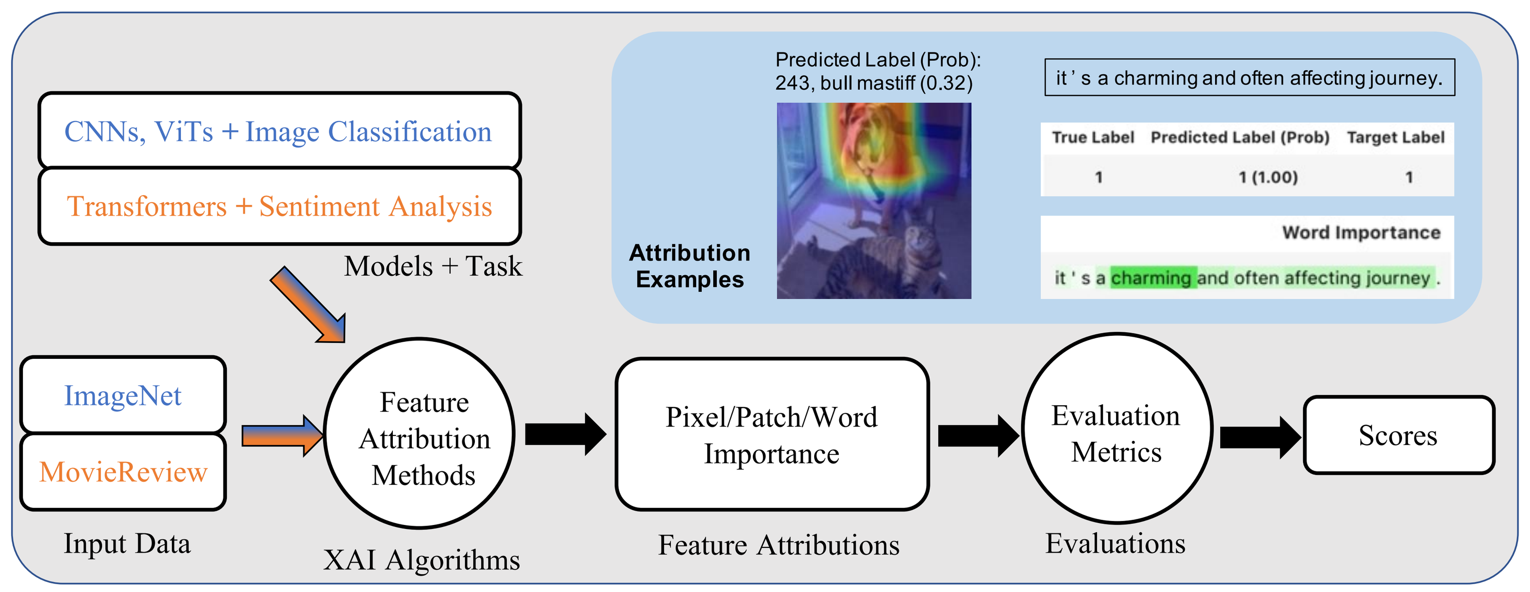
InterpretDL

NEURAL INFORMATION PROCESSING SYSTEMS

The explanations need to be faithful !

Explainable AI (**XAI**) addresses the black-boxed nature of deep neural networks by developing techniques to understand the model predictions. **Feature attribution,** an important paradigm in XAI, accepts the model inputs and gives a per-feature attribution score based on its contribution to the output. **XAI benchmarks** are built with evaluation metrics and datasets to measure the **faithfulness** (i.e., how well explanations match model reasoning) of explanations and filter out unfaithful algorithms.

Benchmark $\mathcal{M}^4$



Why $\mathcal{M}^4$?

$\mathcal{M}^4$ is a unified XAI benchmark evaluating the faithfulness of feature attribution methods with standardized metrics for various model types across modalities.

- A taxonomy of evaluation metrics.
- Across multiple models and modalities.
- Modular design: compatible with different DL libraries (PaddlePaddle, Pytorch, etc.) and easy with new methods and models.

Tasks, Datasets, and Models

Image classification:

- Dataset: 5,000 images from ImageNet validation set
- Models: VGG, ResNets, Mobilenet-V3, ViTs (small, base, large, and MAE [1] pretrained)

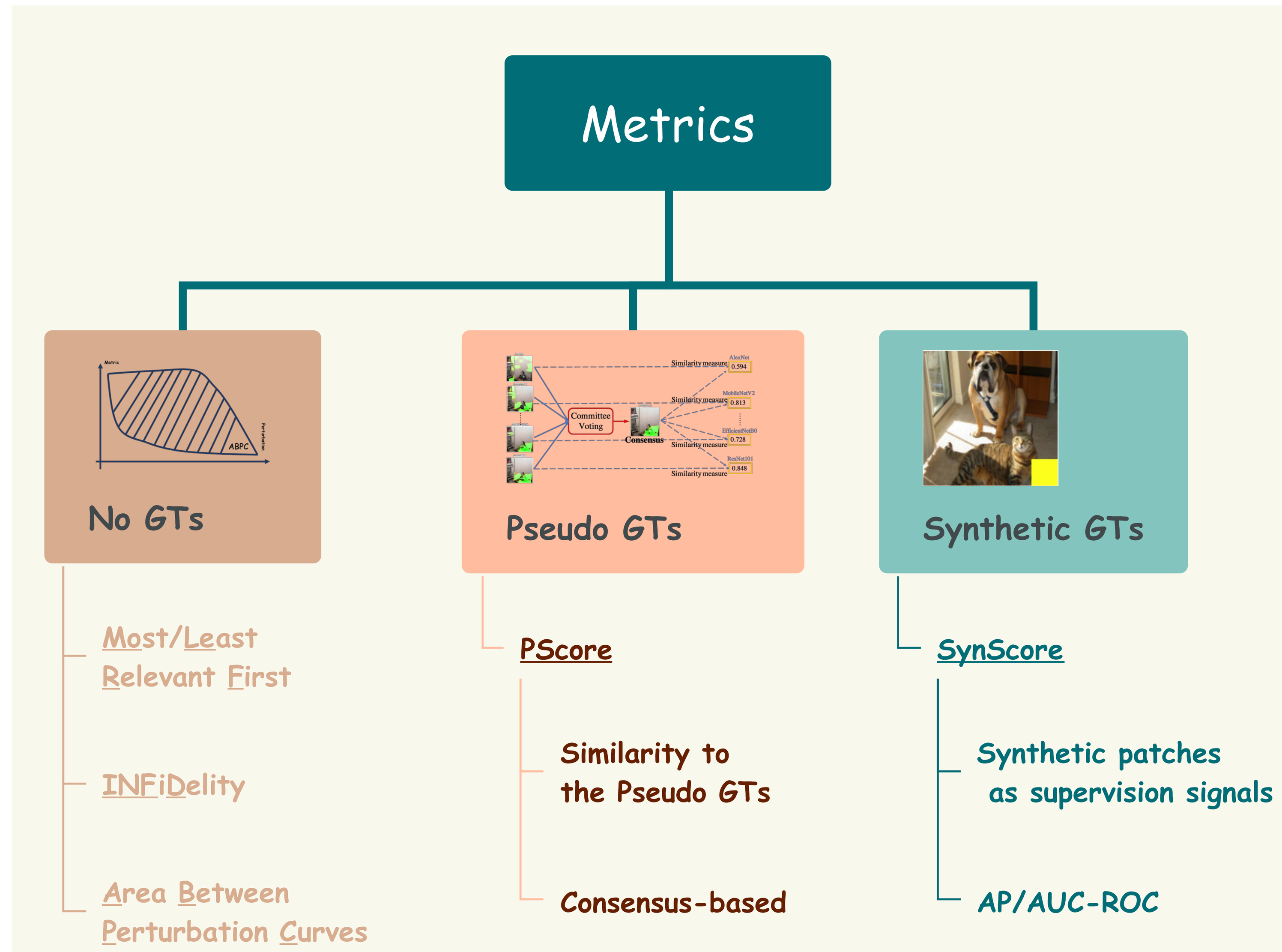
Sentiment analysis:

- Dataset: Movie Review [2]
- Models: BERTs (base and large), DistilBERT, ERNIE-2.0, RoBERTa

Feature Attribution Methods

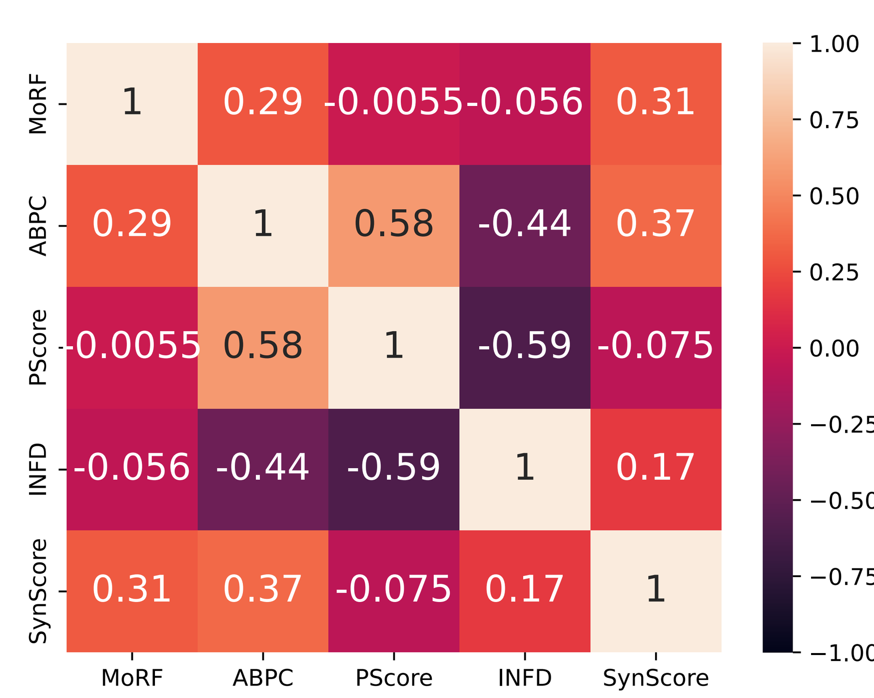
- Model-agnostic: LIME[3]
- Gradient-based: Integrated Gradient, SmoothGrad, GradCAM
- Transformer-specific: Generic Attribution[4], Head-wise/Token-wise Bidirectional Transformer Attributions[5]

Metrics and Taxonomy



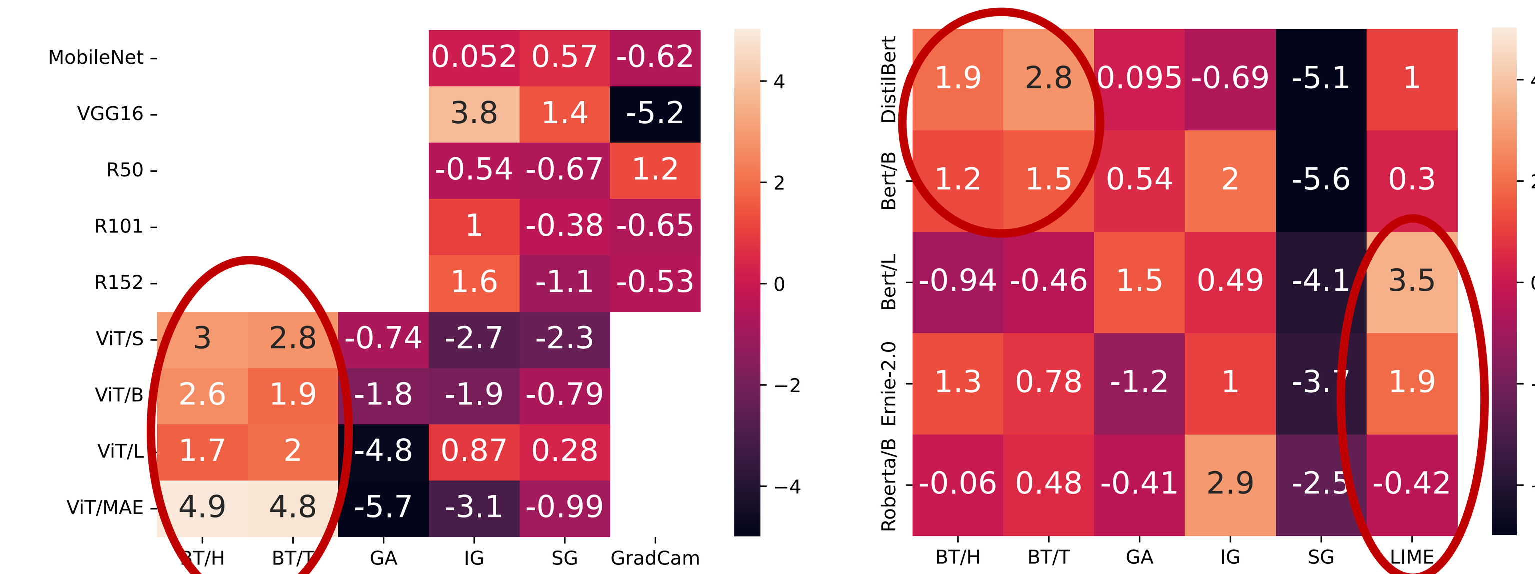
Observations

Whether There are Two Metrics that are Correlated ? Yes !



- ABPC, PScore, and INFD are potential alternatives.
- Near zero correlation for MoRF-PScore, MoRF-INFD, and PScore-SynScore.

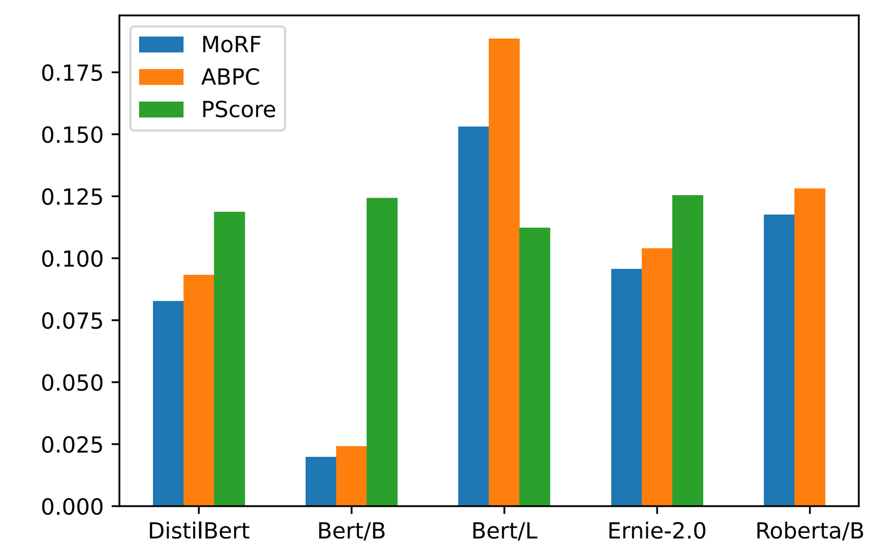
Which Explanation Algorithm Demonstrates the Best Faithfulness ?



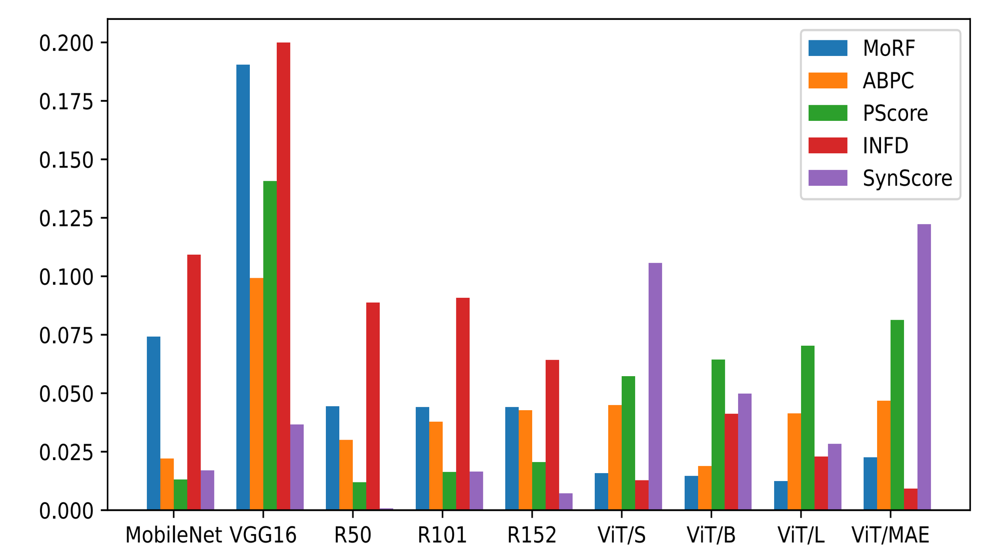
AvgScore = MoRF + ABPC + PScore – INFD + SynScore
(the higher the better)

Which Model is the Most (In)sensitive to Explanation Algorithms ?

Sensitivity measured by **standard deviations** across different methods.



- Insensitive to explanation algorithms ⇒ the model can be easily explained.
- VGG: the most sensitive.



- Validating attribution methods with different models is necessary.

References

[1] Kaiming He et al. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16000–16009, 2022.

[2] Zaidan and Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In Proceedings of the 2008 conference on Empirical methods in natural language processing, pages 31–40, 2008.

[3] Ribeiro et al.. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[4] Chefer et al. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[5] Chen et al. Beyond intuition: Rethinking token attributions inside transformers. Transactions on Machine Learning Research, 2022.