$\mathcal{M}^4$: 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

NeurIPS 2023 Datasets and Benchmarks

Faithfulness of XAI Algorithms

- Feature attributions.
- Faithfulness?
- Recent benchmarks.

Faithfulness of XAI Algorithms

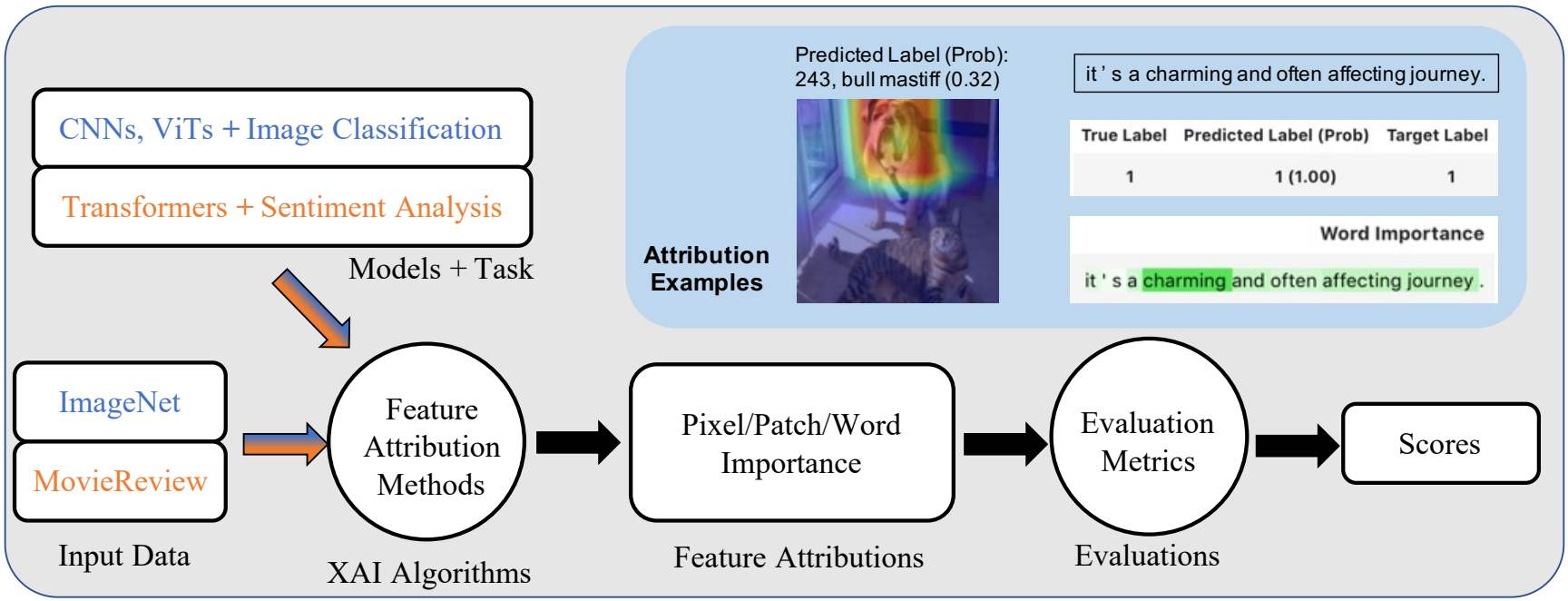
- Feature attributions.
- Faithfulness?
- Recent benchmarks.

Faithfulness of XAI Algorithms

- Feature attributions.
- Faithfulness?
- Recent benchmarks.

Why $\mathcal{M}^4$

- A unified XAI benchmark for feature attribution Methods across Modalities, Models, and Metrics.



Benchmark $\mathcal{M}^4$

Tasks, Datasets and Models:

- Image classification
  - 5,000 images from ImageNet validation set
  - VGG, ResNets, Mobilenet-V3, ViTs, MAE-ViTs
- Sentiment analysis
  - Movie Review dataset
  - BERTs, DistilBERT, ERNIE, RoBERTa

Feature Attribution Methods:

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

Benchmark $\mathcal{M}^4$

Metrics and Taxonomy

- No Ground Truth
  - **Mo**st **R**elevant **F**irst: $\text{MoRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{MoRF}}) - f(x^{(k)}_{\text{MoRF}}))$
  - **Le**ast **R**elevant **F**irst: $\text{LeRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{LeRF}}) - f(x^{(k)}_{\text{LeRF}}))$
  - **A**rea **B**etween **P**erturbation **C**urves
  - **INF**i**D**elity: $\text{INFD}(x) = E_{I\sim\mu_I}(I^T A(x,f) - (f(x) - f(x-I))^2)$
- Pseudo Ground Truth
  - M a set of well-trained models:
    $\text{PScore}(x) = \cos(\frac{1}{|M|}\sum_{g\in M} A(x,g), A(x,f))$
- Synthetic Ground Truth
  - $metric \in \{\text{AP}, \text{AUC-ROC}\}$ and $Syn(x)$ the synthetic ground truth of explanation:
    $\text{SynScore}_{metric}(x) = metric(Syn(x), A(x,f))$
- MoRF + ABPC + PScore – INFD + SynScore

Benchmark $\mathcal{M}^4$

Metrics and Taxonomy

- No Ground Truth
  - **Mo**st **R**elevant **F**irst: $\text{MoRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{MoRF}}) - f(x^{(k)}_{\text{MoRF}}))$
  - **Le**ast **R**elevant **F**irst: $\text{LeRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{LeRF}}) - f(x^{(k)}_{\text{LeRF}}))$
  - **A**rea **B**etween **P**erturbation **C**urves
  - **INF**i**D**elity: $\text{INFD}(x) = E_{I\sim\mu_I}(I^T A(x,f) - (f(x) - f(x-I))^2)$
- Pseudo Ground Truth
  - M a set of well-trained models:
    $\textbf{PScore}(x) = \cos(\frac{1}{|M|}\sum_{g\in M} A(x,g), A(x,f))$
- Synthetic Ground Truth
  - $metric \in \{\text{AP}, \text{AUC-ROC}\}$ and $Syn(x)$ the synthetic ground truth of explanation:
    $\textbf{SynScore}_{metric}(x) = metric(Syn(x), A(x,f))$
- MoRF + ABPC + PScore − INFD + SynScore

Benchmark $\mathcal{M}^4$

Metrics and Taxonomy

- No Ground Truth
  - **Mo**st **R**elevant **F**irst: $\text{MoRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{MoRF}}) - f(x^{(k)}_{\text{MoRF}}))$
  - **Le**ast **R**elevant **F**irst: $\text{LeRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{LeRF}}) - f(x^{(k)}_{\text{LeRF}}))$
  - **A**rea **B**etween **P**erturbation **C**urves
  - **INF**i**D**elity: $\text{INFD}(x) = E_{I\sim\mu_I}(I^T A(x,f) - (f(x) - f(x-I))^2)$
- Pseudo Ground Truth
  - M a set of well-trained models:
    $\textbf{PScore}(x) = \cos(\frac{1}{|M|}\sum_{g\in M} A(x,g), A(x,f))$
- Synthetic Ground Truth
  - $metric \in \{\text{AP}, \text{AUC-ROC}\}$ and $Syn(x)$ the synthetic ground truth of explanation:
    $\textbf{SynScore}_{metric}(x) = metric(Syn(x), A(x,f))$
- MoRF + ABPC + PScore − INFD + SynScore

Benchmark $\mathcal{M}^4$

Metrics and Taxonomy

- No Ground Truth
  - **Mo**st **R**elevant **F**irst: $\text{MoRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{MoRF}}) - f(x^{(k)}_{\text{MoRF}}))$
  - **Le**ast **R**elevant **F**irst: $\text{LeRF}(x) = \frac{1}{L+1}\sum_{k=0}^{L}(f(x^{(0)}_{\text{LeRF}}) - f(x^{(k)}_{\text{LeRF}}))$
  - **A**rea **B**etween **P**erturbation **C**urves
  - **INF**i**D**elity: $\text{INFD}(x) = E_{I\sim\mu_I}(I^T A(x,f) - (f(x) - f(x-I))^2)$
- Pseudo Ground Truth
  - M a set of well-trained models:
    $\textbf{PScore}(x) = \cos(\frac{1}{|M|}\sum_{g\in M} A(x,g), A(x,f))$
- Synthetic Ground Truth
  - $metric \in \{\text{AP}, \text{AUC-ROC}\}$ and $Syn(x)$ the synthetic ground truth of explanation:
    $\textbf{SynScore}_{metric}(x) = metric(Syn(x), A(x,f))$
- MoRF + ABPC + PScore − INFD + SynScore

Benchmark $\mathcal{M}^4$

```python
import interpretdl as it

# Load a pretrained model from PaddlePaddle model zoo.
from paddle.vision.models import resnet50
model = resnet50(pretrained=True)

# Available feature attribution methods include but are not limited to
    SG, IG, LIME, BT, GA and etc. 'interpret' is the universal api.
algo = it.SmoothGradInterpreter(model, device="gpu:0")
expl_result = algo.interpret("test.jpg")

# Available faithfulness evaluation metrics include but are not
    limited to MoRF, ABPC, INFD and etc. 'evaluate' is the universal
    api. Note that some evaluators do not require the model.
evaluator = it.Infidelity(model)
eval_result = evaluator.evaluate("test.jpg", expl_result)
```

The modular design of $\mathcal{M}^4$:

- compatible with PaddlePaddle, Pytorch, Hugging Face, etc.
- easy with new methods, models, and metrics.

Observations

- **Whether There are Two Metrics that are Correlated ?**
- **Which Explanation Algorithm Demonstrates the Best Faithfulness ?**
- **Which Model is the Most (In)sensitive to Explanation Algorithms ?**

Observations

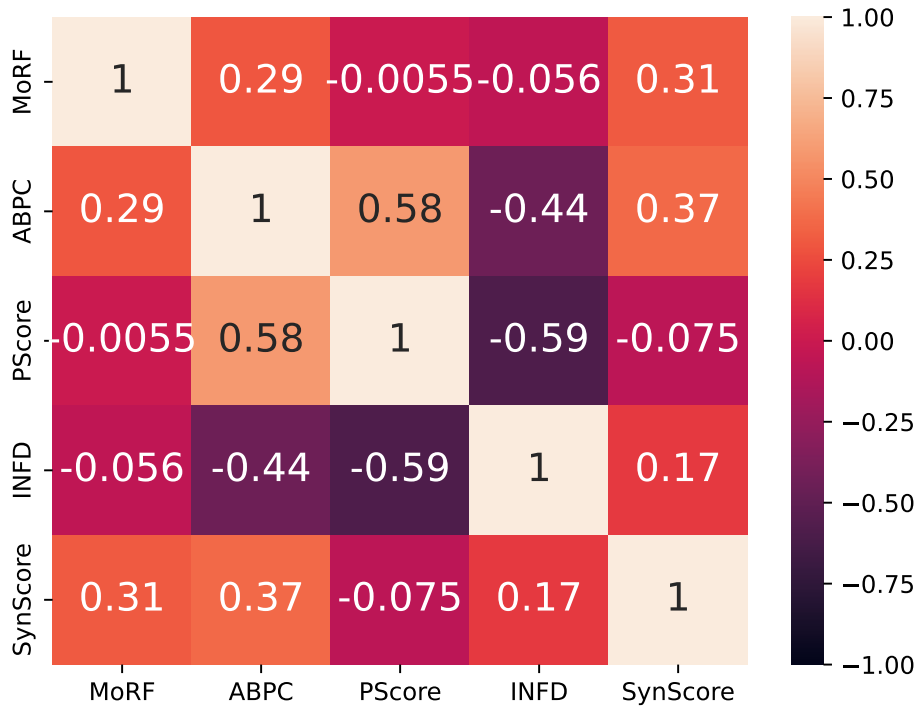
- **Whether There are Two Metrics that are Correlated ?**
- **Which Explanation Algorithm Demonstrates the Best Faithfulness ?**
- **Which Model is the Most (In)sensitive to Explanation Algorithms ?**

Observations

- **Whether There are Two Metrics that are Correlated ?**
- **Which Explanation Algorithm Demonstrates the Best Faithfulness ?**
- **Which Model is the Most (In)sensitive to Explanation Algorithms ?**

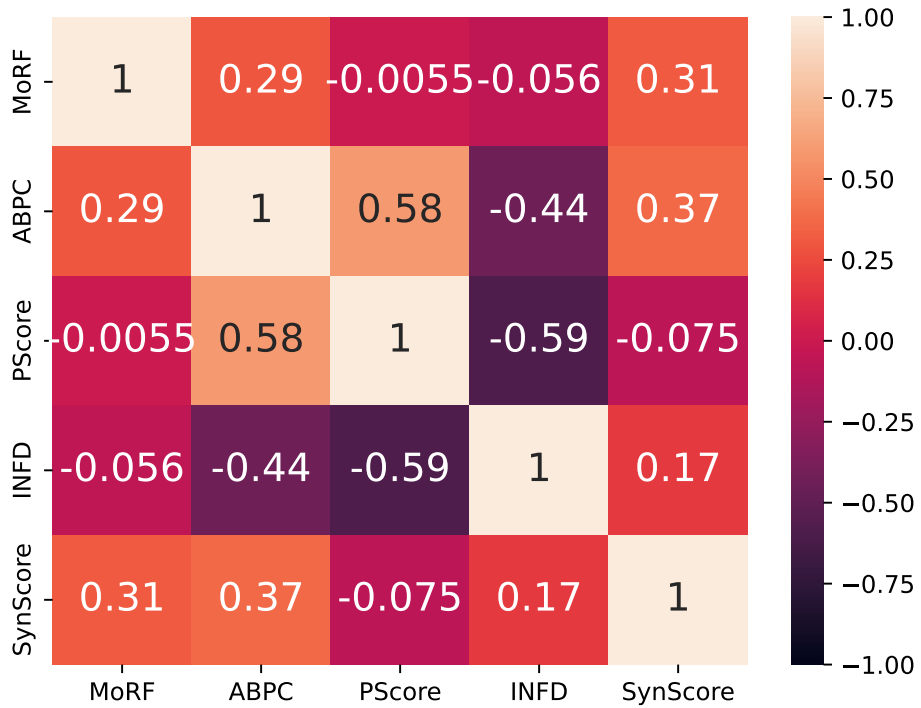
Q1: Correlation between Metrics

- ABPC, PScore, and INFD are strongly correlated.
- MoRF and SynScore evaluate faithfulness from other perspectives.

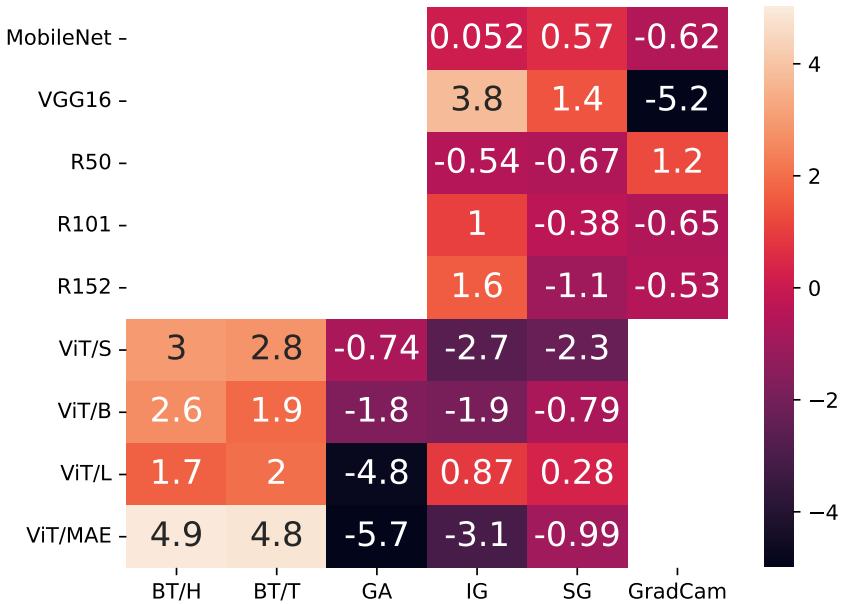


Q1: Correlation between Metrics

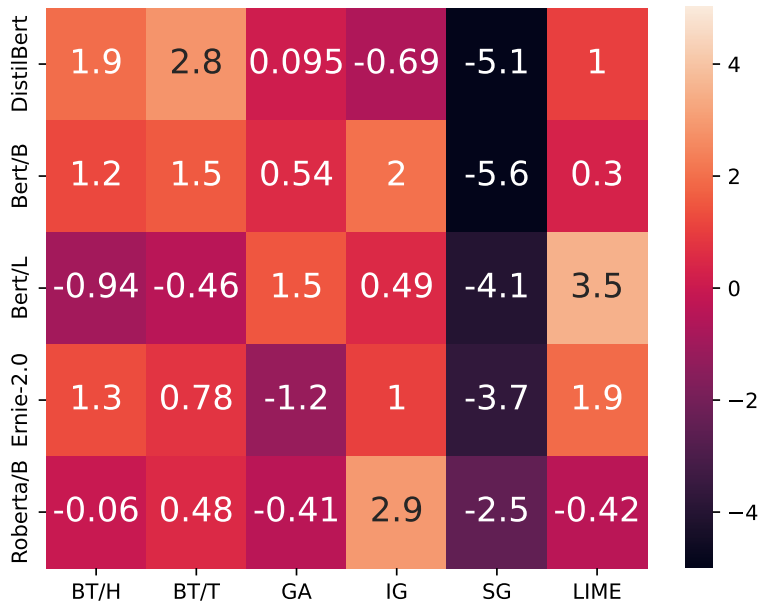
- ABPC, PScore, and INFD are strongly correlated.
- MoRF and SynScore evaluate faithfulness from other perspectives.



Q2: Best Faithfulness

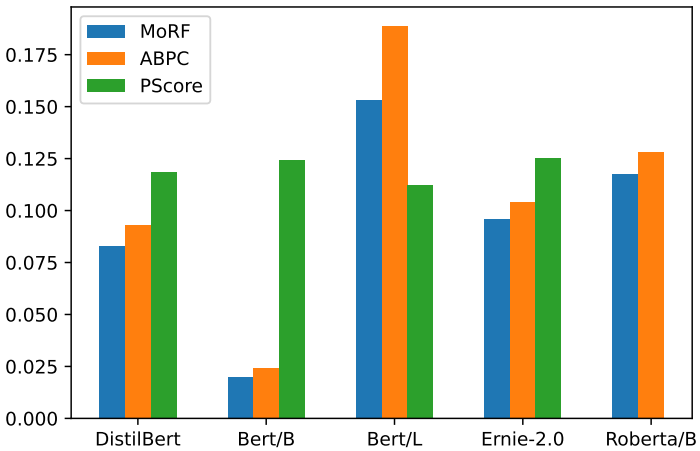


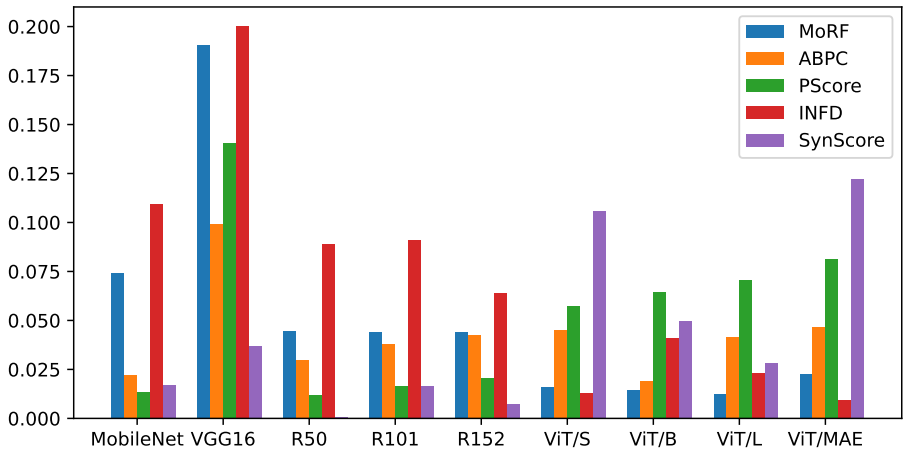
(a) Modality of images.



(b) Modality of texts.

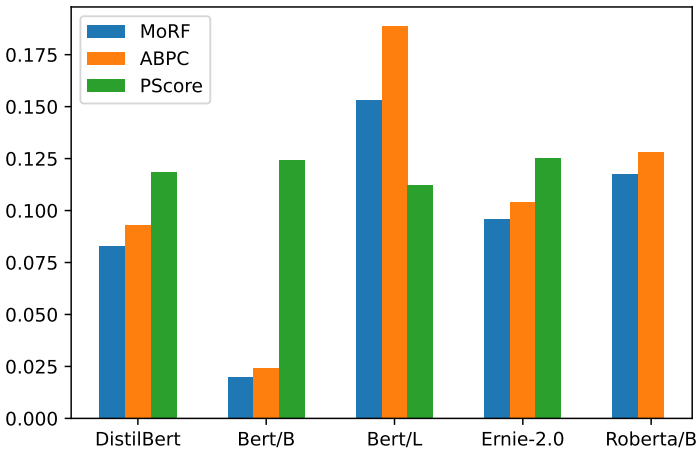
Q3: Sensitivity of Models to Explanation Methods

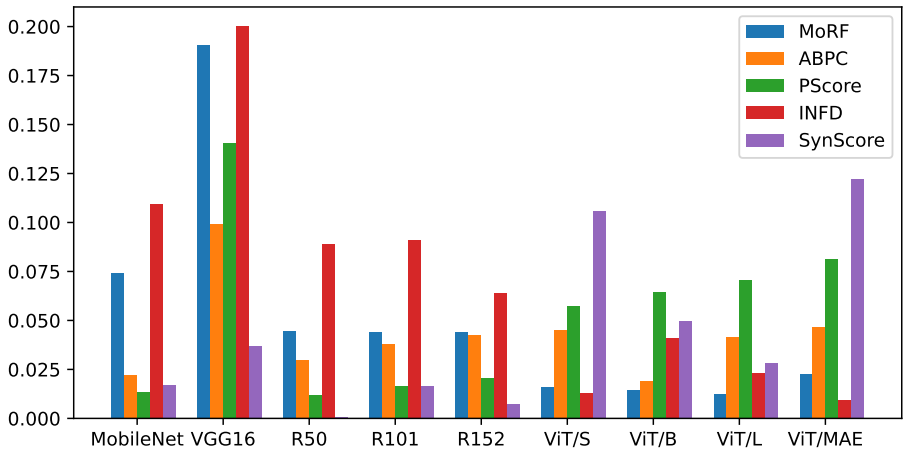




- The sensitivity to different methods can roughly reflect the difficulty of explaining the model.
- Some methods may work especially well for certain models but not as well for others.

Q3: Sensitivity of Models to Explanation Methods





- The sensitivity to different methods can roughly reflect the difficulty of explaining the model.
- Some methods may work especially well for certain models but not as well for others.

The demos, implementations, and additional information of $\mathcal{M}^4$ are publicly available at

InterpretDL

https://github.com/PaddlePaddle/InterpretDL.