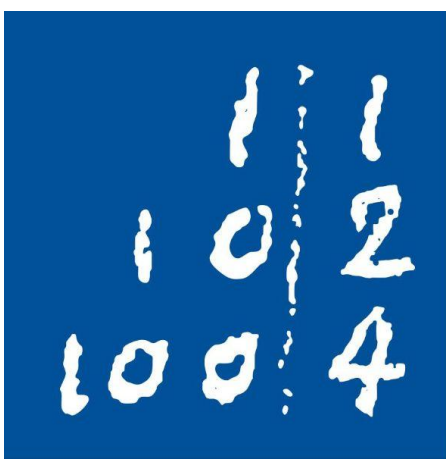
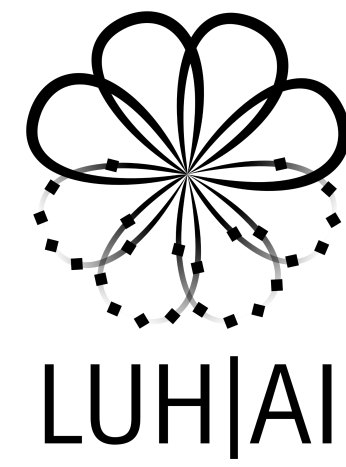


# Explanation Consistency Under Data Augmentation



Leibniz  
Universität  
Hannover



Johny Tarbouch

Poster Presentations in context of Interpretable Machine Learning Lecture

1

Problem Setting

The movie is good

synonym replacement

The film is good

input perturbations

consistent explanation

3

Approach

Explanation Methods

Attention [2]

$$\text{Attn}_i = \frac{1}{H} \sum_{h=1}^H \alpha_{[\text{CLS}],i}^{(h)}$$

attention heads

attention weight

Integrated Gradients (IG)

The movie is good

[PAD] [PAD] [PAD] [PAD]

input text x

baseline text x'

$\alpha=1$

$\alpha=0$

Word Importance

[CLS] this movie is good [SEP]

$$\text{IG}_i(x) = (x_i - x'_i) \cdot \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

single-feature difference

interpolation coefficient

gradient of F(x) along the i th dimension

Consistency Metrics

Kendall's  $\tau$

$$\tau = \frac{C - D}{C + D}$$

same relative ordering

opposite ordering

Top-k Overlap

$$\text{Top-k} = \frac{|A_k \cap B_k|}{k}$$

Cosine Similarity

$$\cos(a, b) = \frac{a \cdot b}{\|a\| \|b\|}$$

attribution vectors

Faithfulness Evaluation (Area Over Perturbation Curve) [1]

Comprehensiveness:  $\text{AOPC}_{\text{comp}} = \frac{1}{|F|} \sum_{f \in F} (p_{\text{orig}} - p_{\text{masked}})$

Sufficiency:  $\text{AOPC}_{\text{suff}} = \frac{1}{|F|} \sum_{f \in F} (p_{\text{orig}} - p_{\text{kept}})$

probability drop

Sanity Check

$$s_k(x) = \text{Sim}(E(x; \theta), E(x; \theta^{(k)}))$$

transformer layers

classifier

layer 6

...

layer 1

IG explanation

parameters after randomizing

↓ as k increases

5

Future Works

Relevant theories

EDA

BT

increasing dataset size and diversity

generate semantically and grammatically similar samples of the original texts

Combination

diversity or quality

Maintain diversity and the quality of the augmented samples

Investigate consistency aware training objectives [3]

[1] Eraser: A benchmark to evaluate rationalized nlp models  
[2] Attention is not explanation  
[3] Unsupervised data augmentation for consistency training

2

Motivation

Motivation

Trust: Same meaning -> similar explanation.

Reliability: Learns concept, not keywords.

Real-world: Stability matters.

Debugging: More predictable, black-box -> hard to trust

Is fine-tuning on augmented data still needed for better explanation?

4

Experiment & Results

Table 1: Hyperparameters and experimental settings

Setting	Value
Model	DistilBERT
Dataset	SST-2
Augmentation	WordNet synonym replacement
Epochs	3
Batch size	32
Seeds	13, 21, 42, 1337, 2024

Table 2: Explanation consistency metrics (mean  $\pm$  std over 5 seeds). Higher is better for all metrics.

Method	Model	Kendall's $\tau$	Top-k Overlap	Cosine Sim.
IG	Baseline	0.794 $\pm$ 0.010	0.908 $\pm$ 0.003	0.939 $\pm$ 0.006
	Augmented	0.786 $\pm$ 0.009	0.906 $\pm$ 0.005	<b>0.940 <math>\pm</math> 0.007</b>
LIME	Baseline	0.471 $\pm$ 0.006	0.776 $\pm$ 0.004	0.856 $\pm$ 0.010
	Augmented	<b>0.486 <math>\pm</math> 0.021</b>	<b>0.783 <math>\pm</math> 0.010</b>	<b>0.864 <math>\pm</math> 0.013</b>
Attention	Baseline	<b>0.892 <math>\pm</math> 0.006</b>	<b>0.955 <math>\pm</math> 0.002</b>	0.974 $\pm$ 0.002
	Augmented	0.888 $\pm$ 0.005	0.952 $\pm$ 0.002	<b>0.975 <math>\pm</math> 0.003</b>

IG kendall\_tau distribution (pooled)

ATTENTION kendall\_tau distribution (pooled)

LIME kendall\_tau distribution (pooled)

IG cosine\_similarity distribution (pooled)

ATTENTION cosine\_similarity distribution (pooled)

LIME cosine\_similarity distribution (pooled)

IG top\_k\_overlap distribution (pooled)

ATTENTION top\_k\_overlap distribution (pooled)

LIME top\_k\_overlap distribution (pooled)

→ Main Finding:

◆ deterministic

◆ high consistency

◆ axiomatically grounded

◆ stable weights

◆ highest consistency

◆ faithfulness debated [2]

◆ stochastic

◆ lower stability

◆ wider variance

Faithfulness AOPC curves (mean  $\pm$  std across seeds)

Comprehensiveness (mean  $\pm$  std)

Sufficiency (mean  $\pm$  std)

Comprehensiveness (mean  $\pm$  std)

Sufficiency (mean  $\pm$  std)

IG sanity check: baseline vs augmented

Similarity to trained IG

Randomization level

trained

classifier

last\_1

last\_2

last\_3

last\_4

last\_5

last\_6

Do we need Augmentation:

◆ Augmentation intensity: Synonym replacement may not change internal representations much

◆ Synonym limits: WordNet replacements can introduce is diverse, but semantically incorrect

◆ Strong baseline: Pretrained DistilBERT already highly consistent -> little room to improve