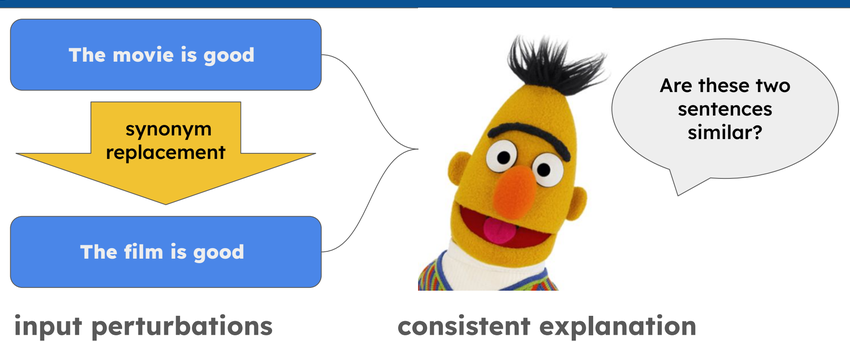
Explanation Consistency Under Data Augmentation

Johny Tarbouch

Poster Presentations in context of Interpretable Machine Learning Lecture

1 Problem Setting



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?

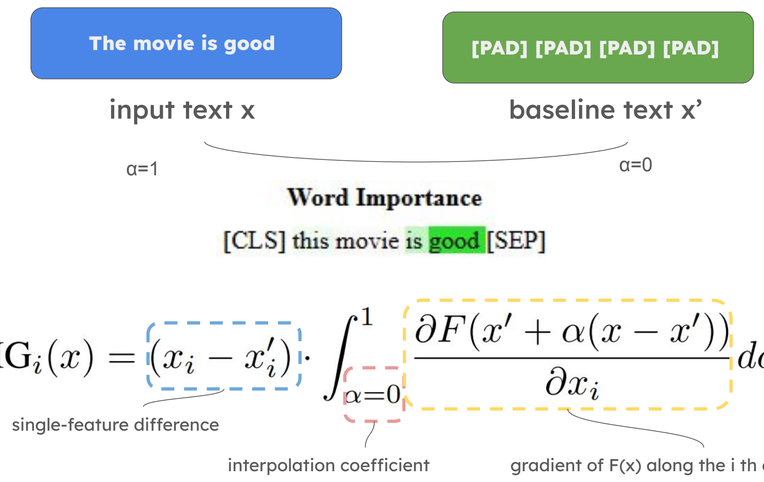
3 Approach

- **Explanation Methods**
  - **Attention [2]**

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

attention heads; attention weight

  - **Integrated Gradients (IG)**



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

single-feature difference; interpolation coefficient; gradient of F(x) along the i th dimension

- **Consistency Metrics**

Kendall's τ

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

same relative ordering; opposite ordering

Top-k Overlap

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

Cosine Similarity

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

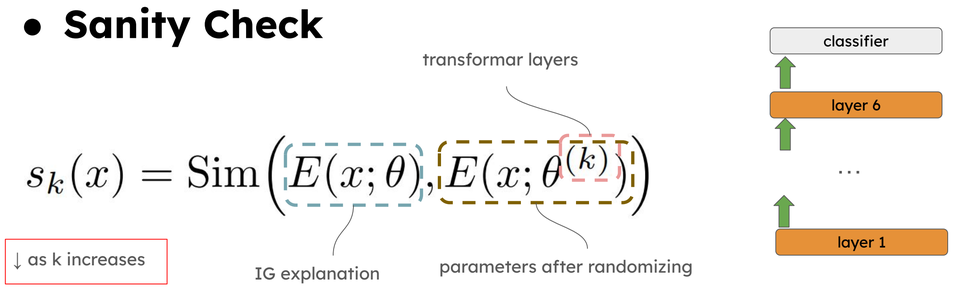
attribution vectors

- **Faithfulness Evaluation (Area Over Perturbation Curve)** [1]

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

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

- **Sanity Check**



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

↓ as k increases; IG explanation; parameters after randomizing

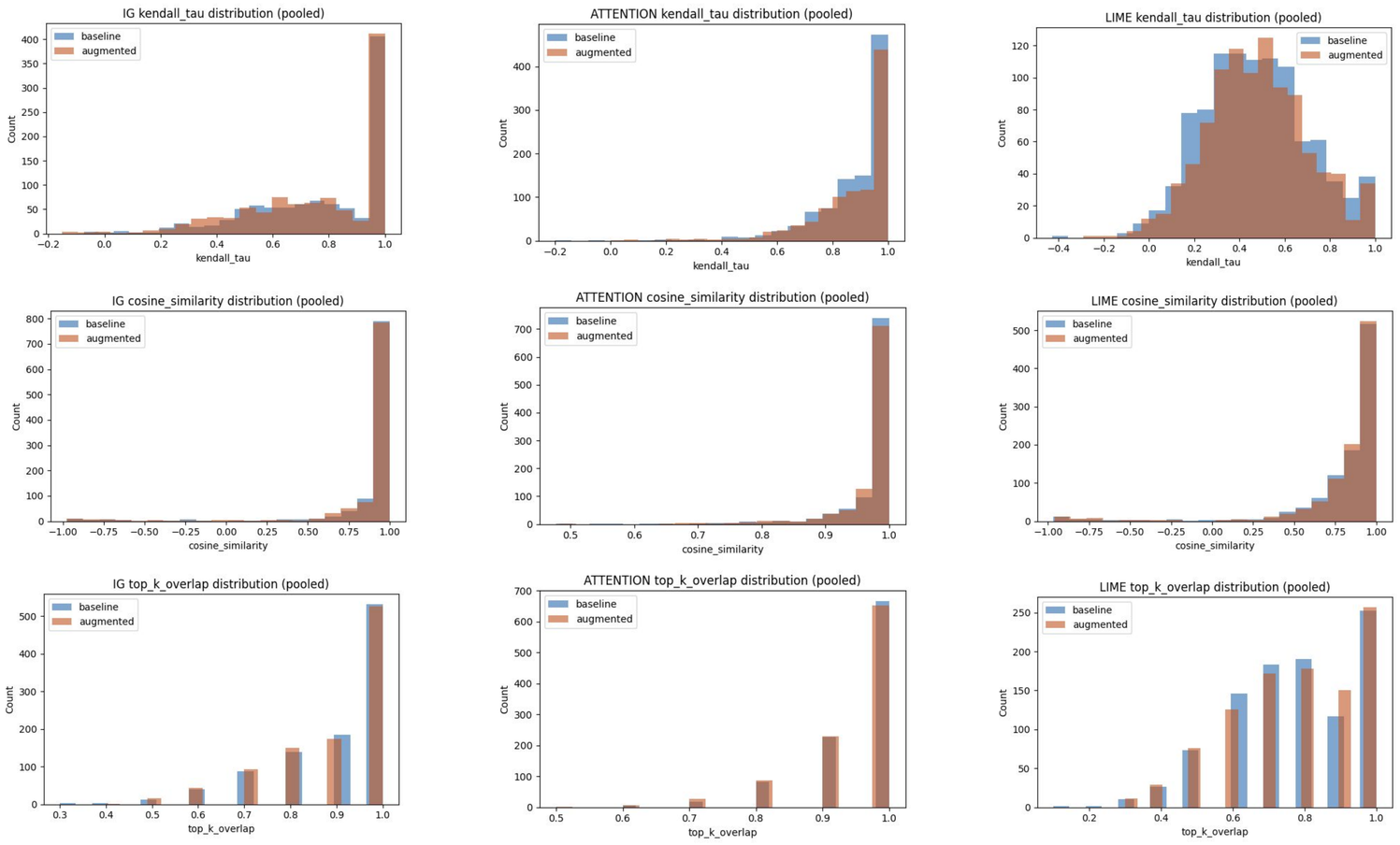
4 Experiment & Results

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 ± std over 5 seeds). Higher is better for all metrics.

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

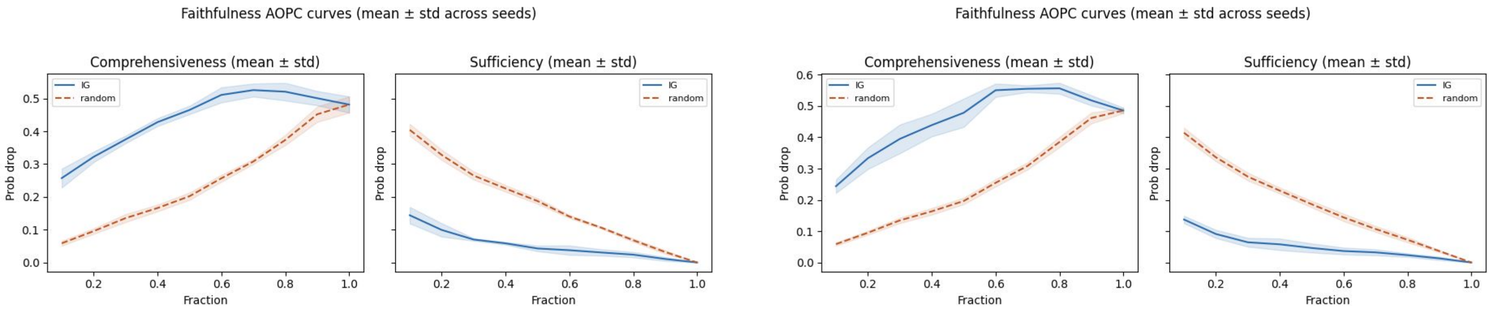


➔ Main Finding:

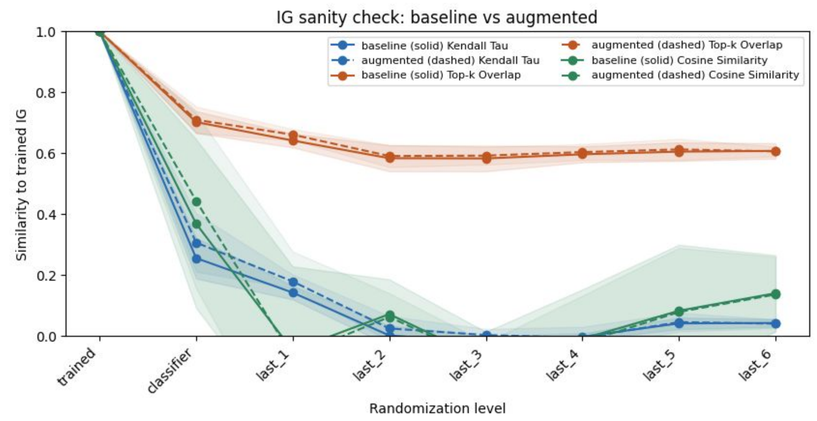
- ◆ deterministic
- ◆ high consistency
- ◆ axiomatically grounded

- ◆ stable weights
- ◆ highest consistency
- ◆ faithfulness debated [2]

- ◆ stochastic
- ◆ lower stability
- ◆ wider variance



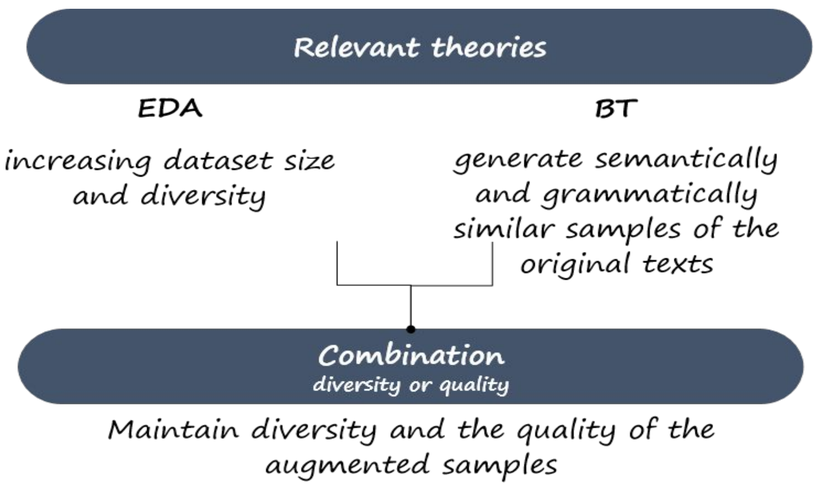
★ Faithfulness: Both models score similarly and beat random → IG highlights meaningful tokens
★ Sanity checks: Randomizing layers degrades IG steadily → explanations depend on learned weights



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

5 Future Works



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