DARE: Towards Robust Text Explanations in Biomedical and Healthcare Applications

Adam Ivankay, *IBM Research Zurich*
Mattia Rigotti, *IBM Research Zurich*
Pascal Frossard, *École Polytechnique Fédérale de Lausanne (EPFL)*

EPFL

IBM

Introduction

Explainable AI

- Research in explainable AI has seen a surge in recent years
- Ever growing need to understand the black-box nature of deep neural networks
- Especially in safety-critical applications

Desiderata of explanation methods

Faithfulness

Explanation reflects the true causal process of the DNN

- **Robustness assumption**

⟷

Plausibility

Explanation is aligned with the human reasoning about the task

Introduction

Explainable AI

Plausible Robustness

Adversarial Robustness of Attributions

general **mills buying back 16.5 m** shares | **general mills inc** . **said** monday **it** plans to buy **back about 16.5** million **shares** of **its** common **stock** from **beverage com-pany diageo plc** .

F("Business") = 100%

- Attributions are ***fragile*** in text

general **mills buying** back 16.5 **m shares** | ge mills **inc** . said monday **it** plans to **buy** back about 16.5 million shares of **its** common stock **from** beverage **company diageo plc** .

F("Business") = 99%

Problem Formulation

$$r(s) = \max_{\tilde{s} \in \mathcal{N}(s)} \frac{d[A(\tilde{s}, F, l), A(s, F, l)]}{d_s(\tilde{s}, s)}$$

$$argmax_{i \in \mathcal{L}} F_i(\tilde{s}) = argmax_{i \in \mathcal{L}} F_i(s)$$

Introduction

Explainable AI

Plausible Robustness

AR

Contributions

I. Introduce DARE, a domain-adaptive attribution robustness estimator

II. Show that attributions are fragile in critical biomedical use cases

III. Introduce FAR for text, a novel method to train robust networks

Domain-Adaptive Attribution Robustness

I. DARE

Domain-Adaptive Attribution Robustness Estimator

- Algorithm to estimate adversarial robustness of attributions in text
- Two-step attack based on imperceptible word substitutions

1. Step: Word Importance Ranking

$$I_s = \nabla_s d[A(s+\varepsilon, F, l), A(s, F, l)]$$

2. Step: Candidate Substitution

$$C_i = MLM(w_i, s, |C|)$$

- Prediction and linguistic constraints
- Final Selection:

$$c_i = argmax_{\tilde{c} \in C_i} d[A(s_{\tilde{c}}, F, l), A(s, F, l)]$$

Plausibility

- Context-aware
- Domain-specific
- Can be trained on unlabelled data

Domain-Adaptive Attribution Robustness

I. DARE

II. Fragile Biomedical Attributions

Attributions change significantly

’took zoloft for 5 months. no side effects except sexual dysfunction. i didn’t feel much better or happier and it made me feel really drowsy.’

F(“4.0”) = 100%

’took zoloft for 5 months. no side effects except sexual dysfunction. i didn’t feel much better or anything and it made me feel really drowsy.’

F(“4.0”) = 100%

Cosine similarity = -0.32

Domain-Adaptive Attribution Robustness

I. DARE

II. Fragile Biomedical Attributions

III. FAR for Text

Framework for Attributional Robustness

- "Adversarial training on predictions and attributions"
- **Adversarial search**

Robust Predictions **Robust Attributions**

$$s_{adv} = argmax_{\tilde{s} \in \mathcal{N}(s)} \{(1-\gamma) \cdot l_c(\tilde{s}, F, l) + \gamma \cdot d[A(\tilde{s}, F, l), A(s, F, l)]\}$$

Solved with DARE

- **Network training**

$$\theta^* = argmin_\theta \sum_{s \in \mathcal{S}} \{(1-\delta) \cdot l_c(s_{adv}, F, l) + \delta \cdot d[A(s_{adv}, F, l), A(s, F, l)]\}$$

Solved with SGD-based optimizers

Domain-Adaptive Attribution Robustness

I. DARE

II. Fragile Biomedical Attributions

III. FAR for Text

Does it work?

Yes!

How well?

	M				S	DL	IG	A	S	DL	IG	A
HoC								0.78	0.76	2.6	2.2	0.77
	ADV.	0.81 ±0.09	±0.22							±0.11	±0.22	±0.11
	FAR-IG	**0.84** ±0.08	**0.24** ±0.2	**0.65** ±0.26	**0.86** ±0.08		±0.14					0.59 ±0.09 **0.3** ±0.05

Find out in the paper and at ACL 2023!!