Improving Detection of Adversarial Samples using Explainability Techniques

Technische Universität München

Fakultät für Informatik

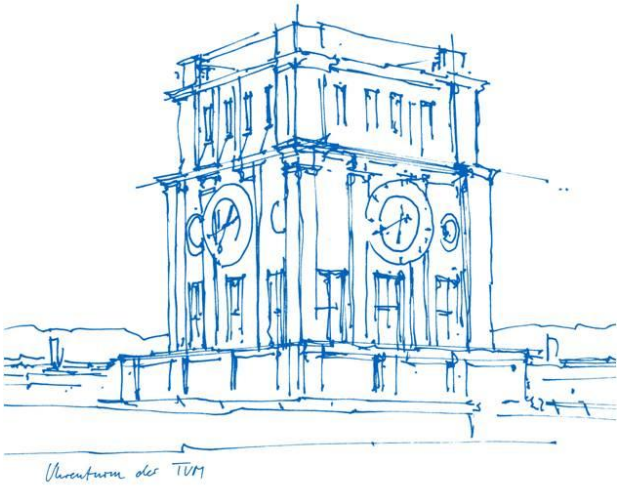
XAI Lab Course, SS2022

13.06.2022

Niklas Hölterhoff

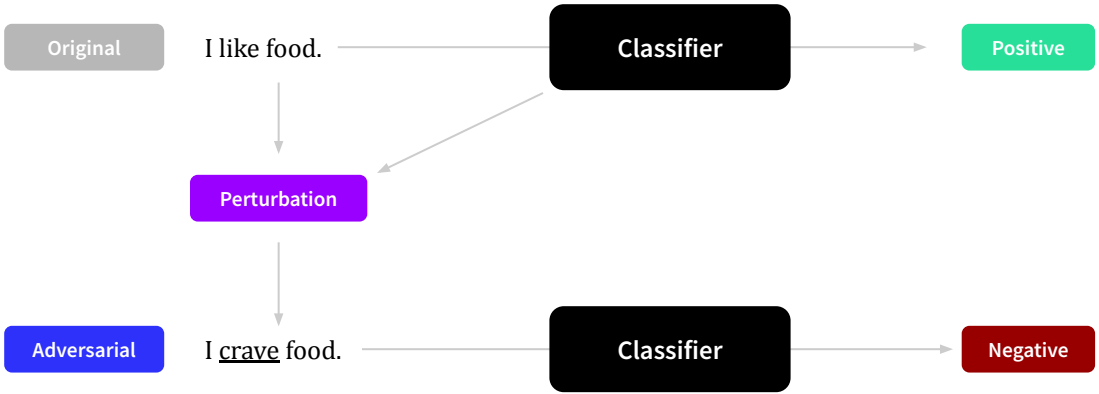
Karl Richter

Yifeng Dong



Fooling the Classifier: Adversarial-attacks

Fooling the Classifier: Adversarial-attacks



Introduction: Metrics

- **Attack Metrics**
  - **Attack success rate (ASR)**: share of attacks that we successful [inherent]
  - **Accuracy (Acc):** accuracy of classification
  - **Perplexity (PPL)**: quality of the adversarial examples [language model]
  - **Grammatical error numbers (GE)**: quality of the adversarial examples [grammar checker LM]
  - **Attack validity:** percentage of valid attacks [human evaluation]
- **Defense Metrics**
  - **F1-Score**
  - Precision
  - Recall

Types of adversarial-defenses

- There are three options to mitigate the risk of adversarial samples
  - [Training] Make the model more robust — Word-level, Sentence-level
  - [Inference] Pre-process Inputs — Character-level, Sentence-level
  - **[Inference] Detect adversarial samples** — Character-level, Word-level, Sentence-level

Character Level: Attack

Replace/remove/add individual chars to change the classification of a sentence.

No attack: 89% Acc

1-char attack: 60% Acc [BERT on MRPC]

2-char attack: 31% Acc [BERT on MRPC]

[Pruthi et al. 2019 @ Carnegie Mellon, https://arxiv.org/pdf/1905.11268.pdf]

I would put this at the top of my list of films in the category of unwatchable trash! **Negative** **Original**

I would put this at the top of my list of films in the category of unWatchable trash! **Positive** **Adversarial**

Character Level: Defense

Most approaches are based on pre-processing. Achieve good results.

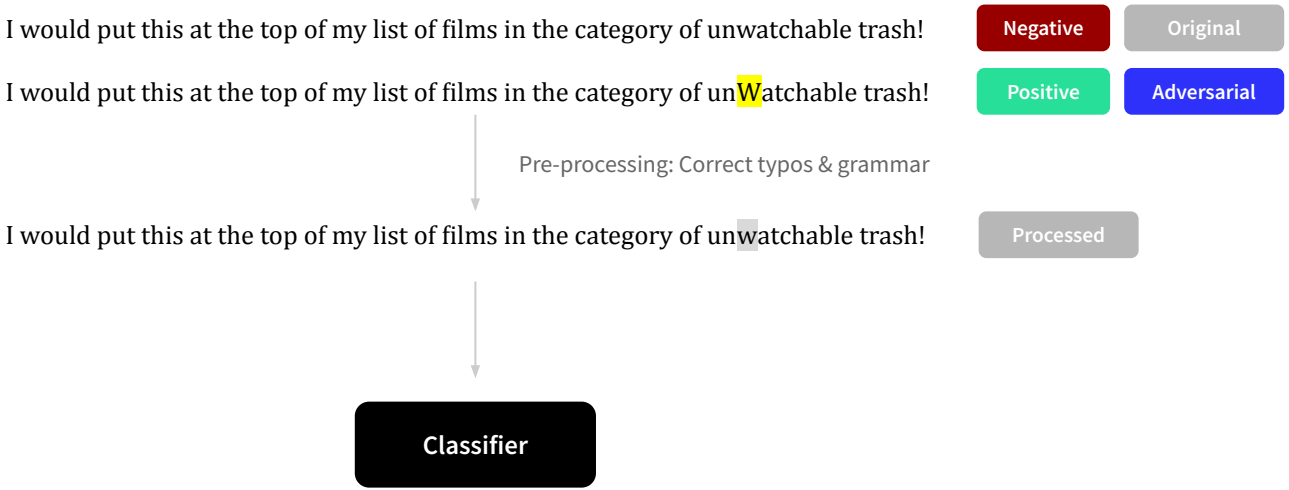
Neutral Word: 82.5% Acc [BERT on MRPC]

Spell Correction: 61.6% Acc [BERT on MRPC]

Pass-Through: 81.5% Acc [BERT on MRPC]

Background Model: 82.5% Acc [BERT on MRPC]

[Pruthi et al. 2019 @ Carnegie Mellon, https://arxiv.org/pdf/1905.11268.pdf]



Word Level: Attack

Replace/remove/add as little words as possible to change the classification of a sentence.

PWWS: 78.8% ASR [BERT, SST-2]
[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

TextFooler: 90% ASR [BERT, SST-2]
[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

BERT-ATTACK: 87% ASR [BERT, SST-2]
[Zeng et al. 2019 @ Tsingua, https://arxiv.org/pdf/2009.09191.pdf]

PWWS: 96.6% ASR [Word-CNN, IMDb]
[Wang et al. 2019 @ Huazhong, https://openreview.net/pdf?id=BJl_a2VYPH]

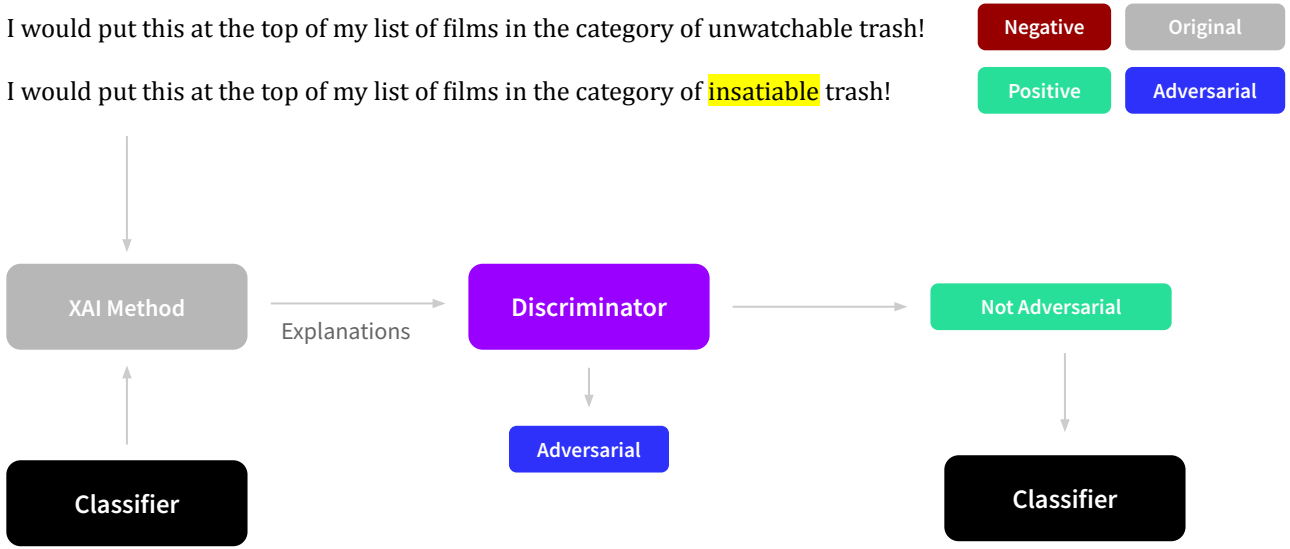
I would put this at the top of my list of films in the category of unwatchable trash! — Negative — Original

I would put this at the top of my list of films in the category of insatiable trash! — Positive — Adversarial

Word Level: Defense

Currently most researched, approaches reach satisfying results.

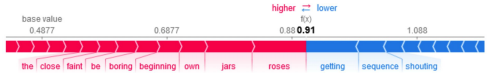
DISP: 75.4 F1 [PWWS, IMDb]
[Zhou et al. 2019 @ UCLA, https://arxiv.org/abs/1909.03084]
FGWS: 89.5 F1 [PWWS, IMDb]
[Mozes et al. 2020 @ UCL, https://arxiv.org/abs/2004.05887]
SHAP: 90 F1 [PWWS, IMDb]
[Mosca et al. 2021 @ TUM]
WDR: 92.1 F1 [PWWS, IMDb]
[Mosca et al. 2021 @ TUM, https://arxiv.org/abs/2204.04636]



Word Level: Defense

Currently most researched, approaches reach satisfying results.

SHAP

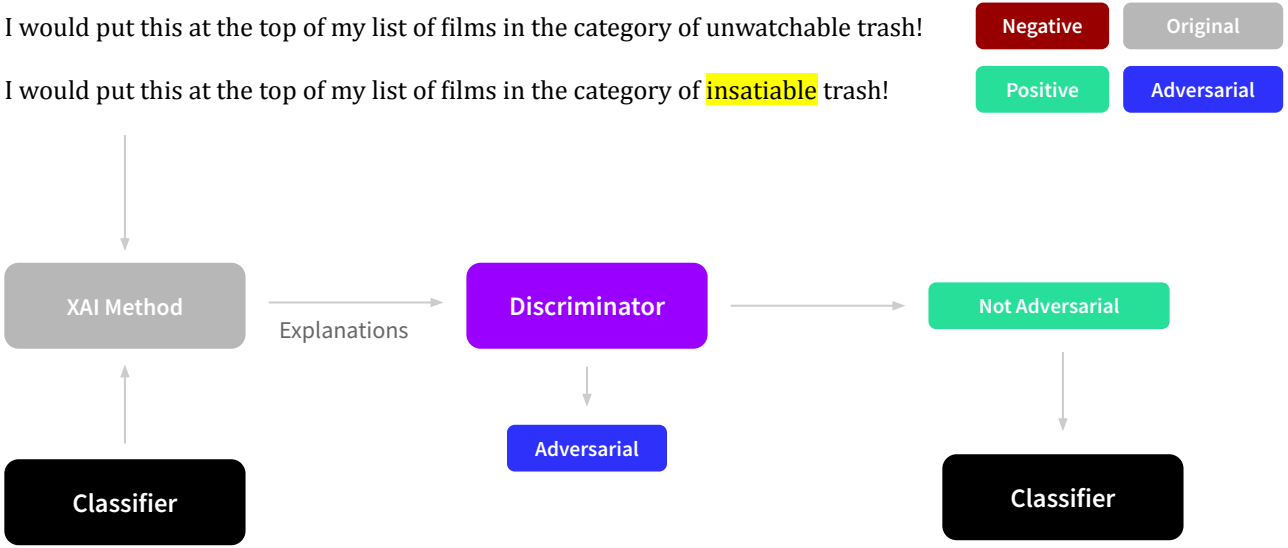


WDR

Replaced word		Logit class 0		Logit class 1		Logits difference	
∅	∅	**3.435**	-1.850	-3.456	**2.174**	6.891	4.024
worst	tough	**1.684**	**1.684**	-1.745	-1.745	3.43	-3.43
(arghhh!)	the	**3.343**	**1.378**	-3.4158	-1.374	6.76	-2.75
feature,	absolutely	**3.402**	-0.31	-3.453	**0.475**	6.86	0.79
This	I	**3.411**	-1.07	-3.474	**1.355**	6.89	2.43

I would put this at the top of my list of films in the category of unwatchable trash!

I would put this at the top of my list of films in the category of insatiable trash!



Sentence Level: Attack

I would put this at the top of my list of films in the category of unwatchable trash!

In my opinion, this is the first piece of work that falls into the category of trash that I can't stand to look at!

Negative | Original

Positive | Adversarial

GAN: 26-47% ASR
[Zhao et al. 2018 @ UCI, https://arxiv.org/abs/1710.11342]

SCPN: 52-64% ASR
[Iyyer et al. 2018 @ Stanford, https://arxiv.org/abs/1804.06059]

StyleAdv: 91-96% ASR
[Qi et al. 2021 @ Tsinghua, https://arxiv.org/abs/2110.07139]

Dataset	Victim	BERT			ALBERT			DistilBERT		
	Attacker	ASR	PPL	GE	ASR	PPL	GE	ASR	PPL	GE
SST-2	GAN	26.42	4643.5	3.34	39.40	1321.7	9.26	47.53	752.3	3.93
	SCPN	52.84	553.2	3.20	59.98	432.9	3.43	64.73	479.0	3.29
	StyleAdv	**91.47**	**228.7**	**1.15**	**95.51**	**191.9**	**1.16**	**96.21**	**180.7**	**1.13**
HS	SCPN	6.56	**223.1**	3.37	7.56	358.2	4.10	1.36	652.8	3.38
	StyleAdv	**51.25**	263.3	**1.26**	**59.03**	**267.0**	**1.32**	**31.00**	**254.8**	**1.39**
AG's News	SCPN	32.98	343.7	4.51	30.91	261.8	4.39	51.04	294.7	5.26
	StyleAdv	**58.36**	**338.8**	**3.14**	**80.70**	**259.2**	**2.59**	**89.54**	**232.6**	**2.86**

[Qi et al. 2021, https://arxiv.org/abs/2110.07139]

Sentence Level: Defense

Defenses for sentence-level attacks are mostly unexplored.

I would put this at the top of my list of films in the category of unwatchable trash!

In my opinion, this is the first piece of work that falls into the category of trash that I can't stand to look at!

Negative | Original

Positive | Adversarial

Research Plan

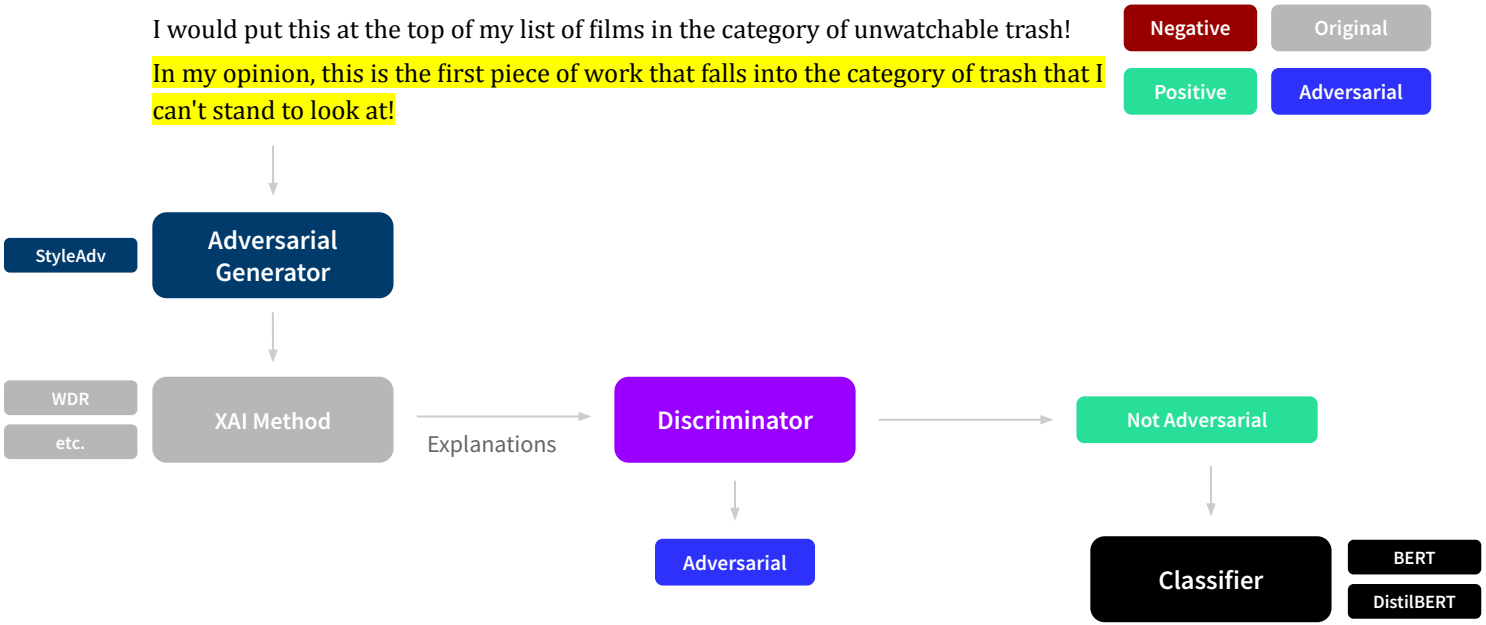
Experiment 1

Hypotheses: The techniques that work well for word-level attacks can be transferred to sentence-level attacks.

Research Question: Are word-level defense mechanism suitable for sentence-level attacks?

Experiment 1: Design

Benchmark the **WDR-based detector** on **StyleAdv** sentence-level attacks.



Experiment 1: Preliminary Results

Benchmarking of WDR on word- and sentence-level attacks

Discriminator trained on word-level attacks (PWWS)

Attack	Type	Dataset	Encoder	F1-Score
PWWS	Word-level	AG-News	DistilBERT	**0.936**
BAE	*Word-level*	*AG-News*	*DistilBERT*	*0.864*
TextFooler	*Word-level*	*AG-News*	*DistilBERT*	*0.957*
StyleAdv	Sentence-level	AG-News	DistilBERT	**0.707**
StyleAdv	Sentence-level	SST-2	BERT	**0.706**
Pruthi et al.	Character-level	SST-2	BERT	**0.676**

Experiment 1: Preliminary Results

Benchmarking of WDR on word- and sentence-level attacks

Discriminator trained on sentence-level attacks (StyleAdv)

Attack	Type	Dataset	Encoder	F1-Score
PWWS	Word-level	AG-News	DistilBERT	**0.806**
TextFooler	Word-level	IMDB	DistilBERT	**0.883**
StyleAdv	Sentence-level	SST-2	BERT	**0.754**
Pruthi et al.	Char-level	SST-2	BERT	**0.699**

Discriminator trained on char-level attacks (Pruthi et al.)

Attack	Type	Dataset	Encoder	F1-Score
PWWS	Word-level	AG-News	DistilBERT	**0.838**
TextFooler	Word-level	IMDB	DistilBERT	**0.894**
StyleAdv	Sentence-level	SST-2	BERT	**0.671**
Pruthi et al.	Char-level	SST-2	BERT	**0.700**

Experiment 1: Challenge of attack validity

- Valid attacks are those, which alter the syntax of a sentence to fool the classifier but not the human.
- Brute-force paraphrasers are prone to generating samples that are not valid.
- Benchmarks by Qi et al. show that StyleAdv performs best, but is not yet sufficient:

GAN 3%,

SCPN 43%

StyleAdv 49.5%

[Qi et al. 2021, https://arxiv.org/abs/2110.07139]

Research Plan

Experiment 1

Hypotheses: The techniques that work well for word-level attacks can be transferred to sentence-level attacks.

Research Question: Are word-level defense mechanism suitable for sentence-level attacks?

Preliminary Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Research Plan

Experiment 1

Hypotheses: The techniques that work well for word-level attacks can be transferred to sentence-level attacks.

Research Question: Are word-level defense mechanism suitable for sentence-level attacks?

Preliminary Result: WDR works reasonable for StyleAdv, but significantly worse than for word-level attacks.

Experiment 2

Hypotheses: The word-level defenses can be further extended to improve performance on the sentence-level.

Research Question: Can word-level defenses be improved to perform better on a sentence-level?

Experiment 2: Explore sentence-level defenses

- Approach 1: Extend WDR by masking multiple consecutive tokens
- Approach 2: Extend WDR by using a ML to fill masked tokens
- Approach 3: Explore NN-specific explanation techniques

Experiment 2: Mask multiple tokens

- WDR compares model predictions between original and when a word is removed:

['one ', 'of ', 'the ', **'best'**, 'of ', 'the ', 'year']

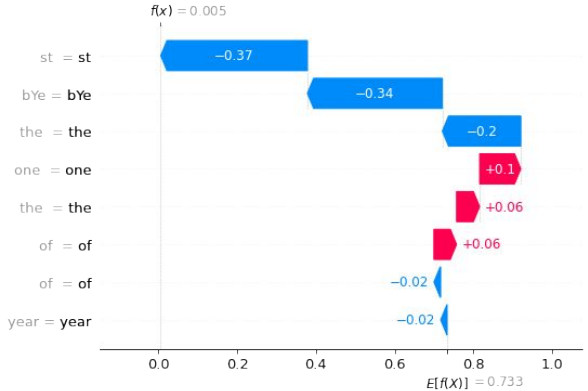
['one ', 'of ', 'the ', **<unk>** , 'of ', 'the ', 'year']

Challenge: Tokenizer

- Character-level attacks are often fooling the tokenizer

Original “One of the best of the year” → `['one ', 'of ', 'the ', 'best ', 'of ', 'the ', 'year']`

Adversarial “One of the bYest of the year” → `['one ', 'of ', 'the ', 'bYe', 'st ', 'of ', 'the ', 'year']`



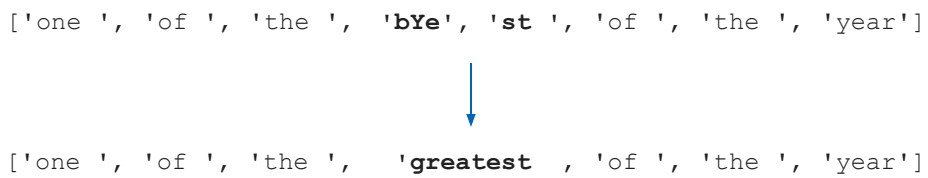
Experiment 2: Mask multiple tokens

- Masking multiple consecutive tokens:

```
['one ', 'of ', 'the ',  'bYe', 'st ', 'of ', 'the ', 'year']
                                ↓
['one ', 'of ', 'the ',     <unk>   , 'of ', 'the ', 'year']
```

Experiment 2: Fill tokens using LM

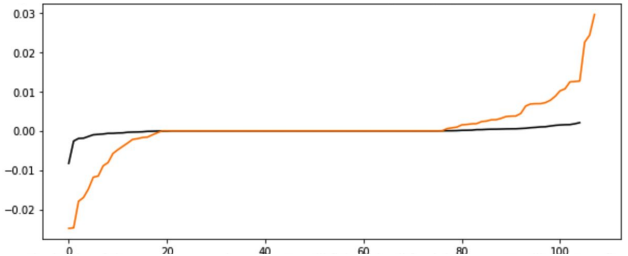
- Use a large language model (such as BERT) to fill in the gaps



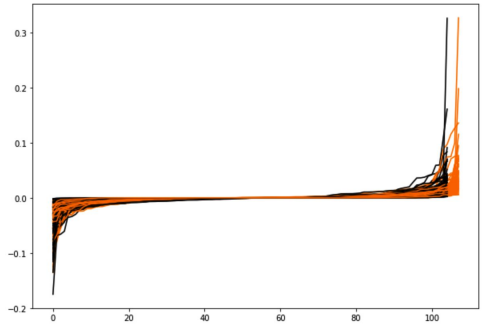
Experiment 2: Explore NN-specific explanation techniques

- **Primary Attribution:** Evaluates contribution of each input feature to the output of a model.
  - Integrated Gradients
  - DeepLift
  - DeepLift SHAP
  - Saliency
  - Feature Ablation
- **Layer Attribution:** Evaluates contribution of each neuron in a given layer to the output of the model.
- **Neuron Attribution:** Evaluates contribution of each input feature on the activation of a particular hidden neuron.

[https://captum.ai/docs/algorithms]



Comparison of Layer Gradient Activation [Adv] vs. [Orig] example



Comparison of 100 Layer Gradient Activations [Adv] vs. [Orig] example

Next Steps

- ❑ Extend WDR by n-token masking
- ❑ Extend WDR by filling masks using LM
- ❑ Explore Captum Explanation methods and benchmark
- ❑ Evaluate and compare different approaches