Unveiling Vulnerability of Self-Attention

Khai Jiet Liong, Hongqiu Wu, Hai Zhao

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

LREC-COLING 2024

https://github.com/liongkj/HackAttend

Background

- **Susceptibility to Perturbations**: Minor changes in wording or complex cues like sarcasm can mislead models like BERT (Devlin et al., 2019).
- **Keyword Sensitivity**: PLMs may incorrectly overemphasize irrelevant or spurious keywords, affecting understanding and context interpretation.
- **Example Error**: Misinterpreting the phrase, "This movie as good as oatmeal.“, due to improper emphasis on non-indicative words.“”

Review of Existing Adversarial Techniques

- Attacks: Word manipulation (substitution, swapping, insertion) to deceive models. (Jin et al., 2020)
- Defenses: Adversarial training i.e. CreAT, (Wu et al., 2023) and data augmentation.
- Challenges: Leads to performance degradation on clean data and input domain shift.

Rethinking PLM’s Vulnerability

- Core Insights:
  1. Internal Mechanisms: Vulnerabilities stem from the internal mechanisms of PLMs.
  2. Focus on Self-Attention: Self-Attention (SA) mechanism particularly in the transformers (Vaswani et. al., 2017) has vulnerability.
- Contributions of This Work:
  1. HackAttend: A novel perturbation method that directly targets the SA mechanism to reveal vulnerabilities.
  2. S-Attend: A straightforward and effective defensive mechanism to mitigate these vulnerabilities and against other general attacks.

Outline

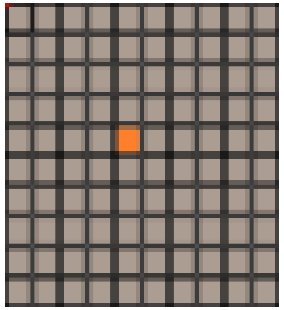
- HackAttend - Generating ‘Adversarial’ Samples
- S-Attend - HackAttend inspired smoothing
- Experiment
  - Evaluation metrics
  - Results
- Case study

Introducing HackAttend

- Specifically targets and disrupts the SA weights within SA matrices.
- Adversarial samples in the form of custom attention masks

- **NOT** an adversarial attack as this method requires access to the backpropagation gradients

Custom Attention Mask



A SA Matrix

Methodology

- Model Fine-Tuning:
- This step adjusts the model's weights and parameters to best fit the training data
- Perturbation Phase:
- Only samples that the model initially predicts correctly are selected for further analysis.
- Each selected sample is then processed through the model to determine the relative importance of the layers and heads.
- Ranking and Selection of Components:
- This is done via a ranking process based on specific criteria derived from the model's response to the input data.
- Custom Attention Mask Generation
- A custom attention mask generated to alter the flow of data through the specified attention units, effectively testing the model's ability to maintain accuracy under modified conditions.

Evaluation Metrics

- Attack Success Rate – Number of incorrect predictions after pertubation
- Clean Accuracy – Accuracy score on the clean set
- Robust Accuracy – Accuracy score on under perturbation
- Hamming distance – Quantifies the differences between the original attention matrix and adversarial attention matrix

$$d_H(M, M') = \frac{\sum_{i=1}^{N_{SA}} (M_i \oplus M_i')}{N_{SA}}$$

Findings

- Tasks Tested: Reading comprehension, logical reasoning, sentiment analysis, and natural language inference.
- Results: Demonstrates that state-of-the-art models are highly vulnerable to HackAttend with a high attack success rate.

Dataset	Max N	ASR%	*clean%*	*robust%*	# Query	Hamming
DREAM	12	98.9	64.7	0.7	18.6	611.4
	6	91.2		5.7	11.2	618.2
HellaSWAG	12	99.9	39.6	0.0	8.8	1222.2
	6	96.7		1.3	7.1	1232.8
ReClor	12	100.0	51.8	0.0	7.3	2151.3
	6	99.6		0.2	6.5	2153.7
SST-2	12	27.4	93.9	67.8	123.6	9.3
	6	10.2		83.8	34.1	9.5

Dataset	Mask%	ASR%	Hamming	# Query
DREAM	1.00	98.9	611.4	18.6
	0.10	91.2	62.4	36.3
	0.01	72.7	5.7	58.9
HellaSWAG	1.00	99.9	1221.2	8.8
	0.10	98.9	121.6	17.6
	0.01	92.5	11.5	29.7
SST-2	1.00	27.4	9.3	123.6
	0.10	6.4	1.1	139.0
	0.01	6.4	1.0	139.1
ReClor	1.00	100.0	2151.3	7.3
	0.10	100.0	213.3	15.9
	0.01	84.9	19.4	40.7

S-Attend: A Novel Smoothing Technique

- a technique that smooths attention scores during training, thereby increase model robustness.
- Inspirations:
  - Adversarial Self Attention (ASA): Inspired by ASA (Wu et al., 2023), which teaches models to moderate focus on specific keywords by reversing gradients of important attention units.
  - HackAttend: Further inspired by HackAttend's demonstration of vulnerabilities, specifically how models can be misled by manipulated attention.
- Efficiency and Cost:
  - high cost of storing custom attention masks (i.e. 144 attention matrix for bert-base)
  - randomly mask out attention units following Bernoulli distribution with α={0.1,0.2,0.5}.

Results

- Increase robustness with minimal impact on performance without additional computationally intensive steps required by FreeLB (Zhu et al., 2020) and ASA (Wu et al., 2023).
- Minimal sacrifice on clean accuracy

Dataset	Defense/Smoothing	*clean*%	*robust*% TF	BA
ReClor	Baseline	51.8	0.8	2.0
	CreAT	49.0	46.6	48.0
	FreeLB	50.4	50.2	49.6
	TF(ADA)	47.8	47.4	47.8
	BA(ADA)	47.4	47.0	46.6
	S-Attend ($\alpha = 0.1$)	48.6	47.4	47.8
	S-Attend ($\alpha = 0.2$)	51.0	50.0	49.6
	S-Attend† ($\alpha = 0.5$)	**52.8**	**51.4**	**51.2**
DREAM	Baseline	64.7	19.3	3.8
	CreAT	65.0	55.1	**55.2**
	FreeLB	**65.1**	**56.2**	55.1
	TF(ADA)	57.4	55.6	54.1
	BA(ADA)	55.7	51.9	52.5
	S-Attend ($\alpha = 0.1$)	63.2	54.2	52.8
	S-Attend† ($\alpha = 0.2$)	64.4	54.6	53.7
	S-Attend ($\alpha = 0.5$)	63.0	53.0	52.8

Table 7: Robust Performance Evaluation: Baseline model vs. Adversarial Training models vs. ADA models. Baseline represents regular fine-tuned BERT base. † denotes the best *S-Attend* Model.

Results against HackAttend

- In the event of new type of attack/perturbation, S-Attend is structurally more robust

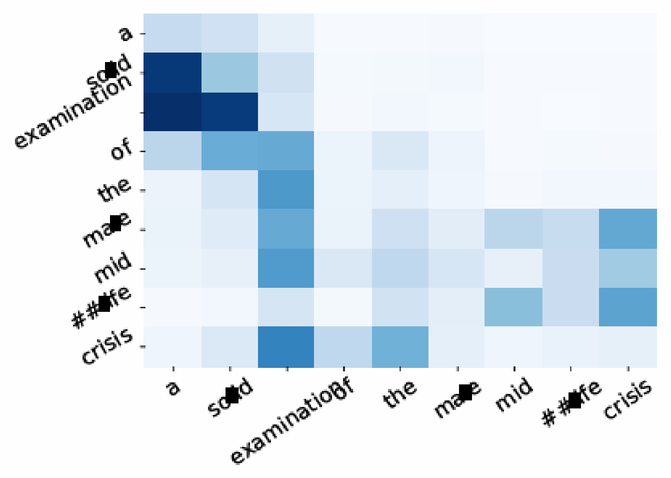
Dataset	Method	Mask%	ASR%
ReClor	Baseline	1.00	100.0
	S-Attend		100.0
	CreAT		100.0
	FreeLB		100.0
	Baseline	0.10	99.6
	S-Attend		**87.1**
	CreAT		100.0
	FreeLB		100.0
DREAM	Baseline	1.00	98.9
	S-Attend		**97.5**
	CreAT		100.0
	FreeLB		98.9
	Baseline	0.10	91.2
	S-Attend		**85.1**
	CreAT		90.0
	FreeLB		88.3

Table 8: Robustness evaluation against *HackAttend* perturbations. Adversarial Training vs. *S-Attend* smoothing. Baseline represents regular fine-tuned model.

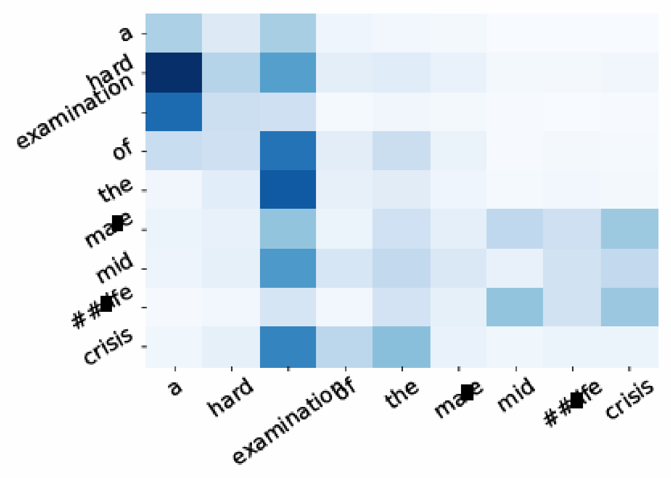
Case Study

A <u>**solid**</u> hard examination of the male midlife crisis (Positive)

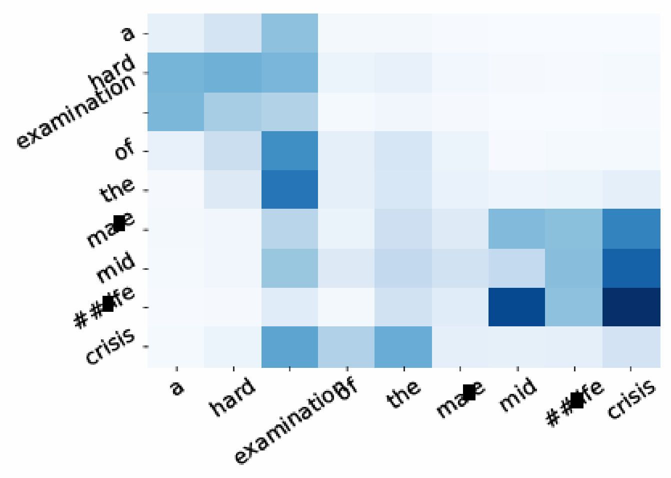
Adversarial sample generated (BERT-Attack (Li et al., 2020))
A **hard** examination of the male midlife crisis (Negative)



(a) Normal model (clean)



(b) Normal model (augmented)



(c) S-Attend model (augmented)

- suggest that models tend to heavily rely on word matching (Hao et al., 2021).

Conclusion

- Presented HackAttend and S-Attend
  - HackAttend
    - A method to target model structural weakness, particularly SA mechanism
  - S-Attend
    - Robust against spectrum of attacks
    - Promotes the activation of SA components
    - Aids in reducing sensitivity to noisy input data
    - Helps the model to learn a more generalized representation