南京航空航天大學

Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组

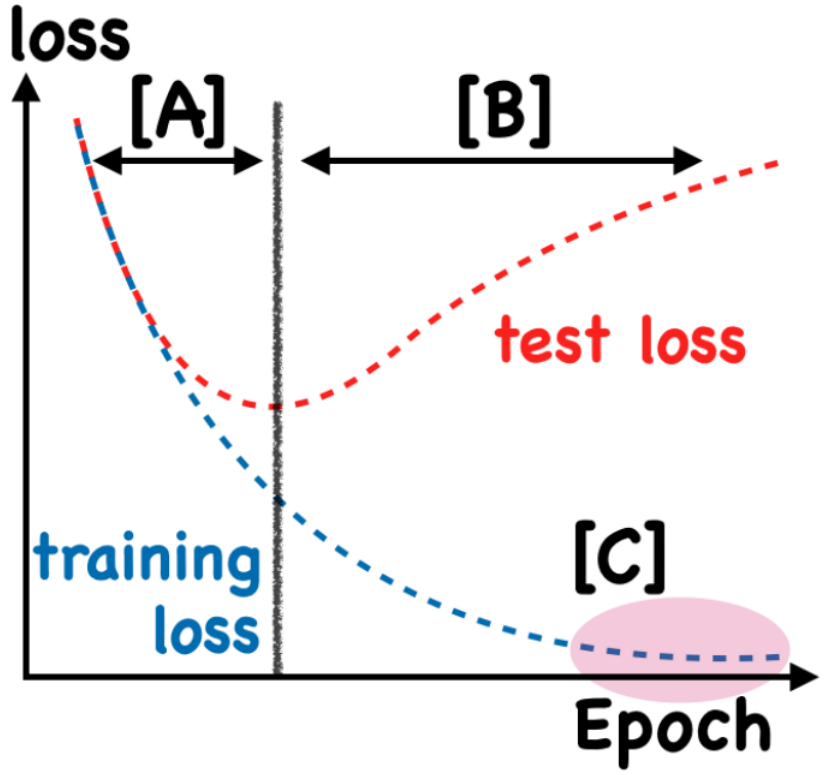
PAttern Recognition and NEural Computing

Flooding-X: Improving BERT's Resistance to Adversarial Attacks via Loss-Restricted Fine-Tuning

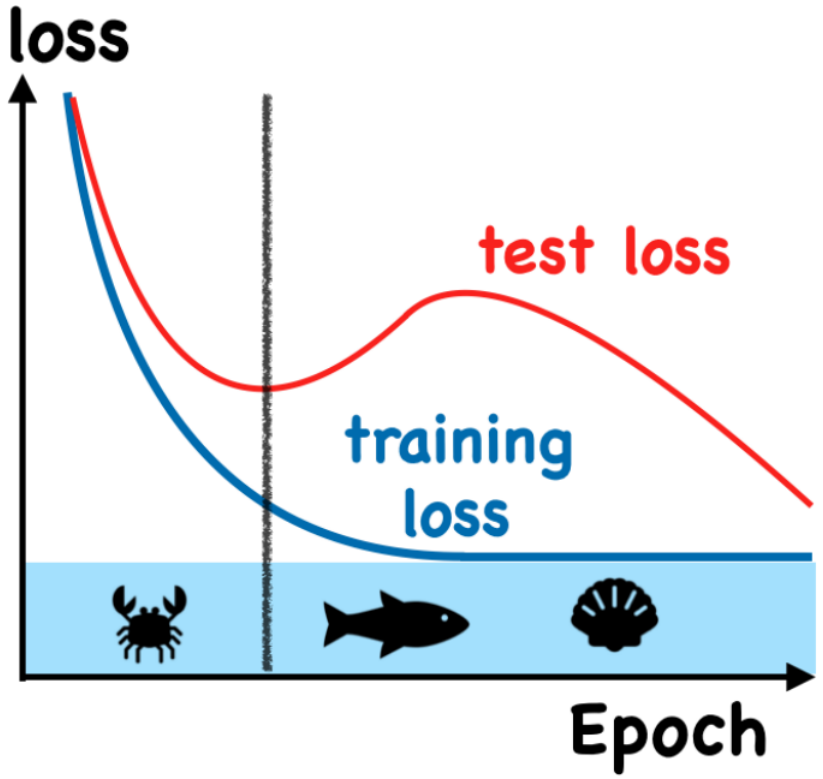
Qin Liu, Rui Zheng, Bao Rong, Jingyi Liu, ZhiHua Liu, Zhanzhan Cheng,
Liang Qiao, Tao Gui, Qi Zhang, Xuanjing Huang

ACL 2022

Flood



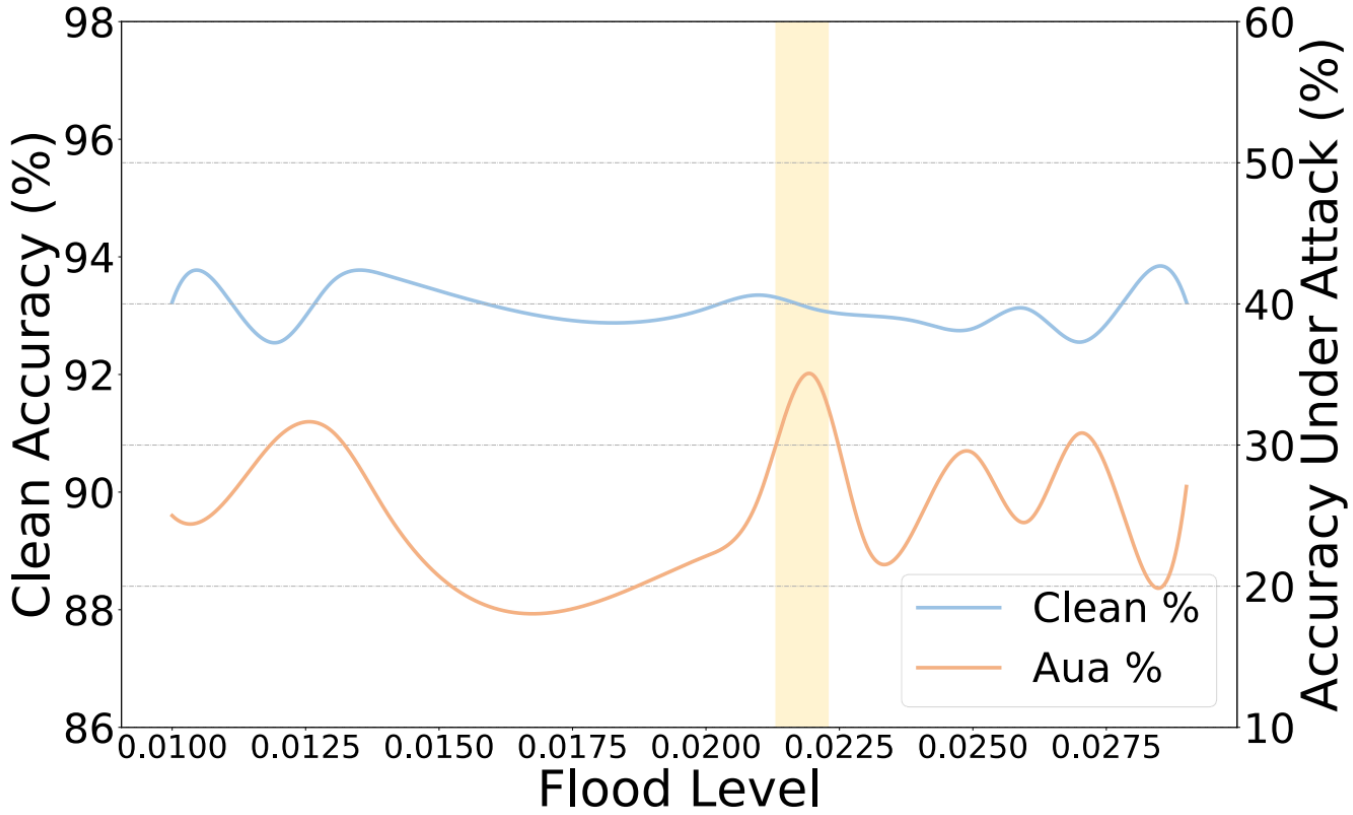
(a) w/o Flooding



(b) w/ Flooding

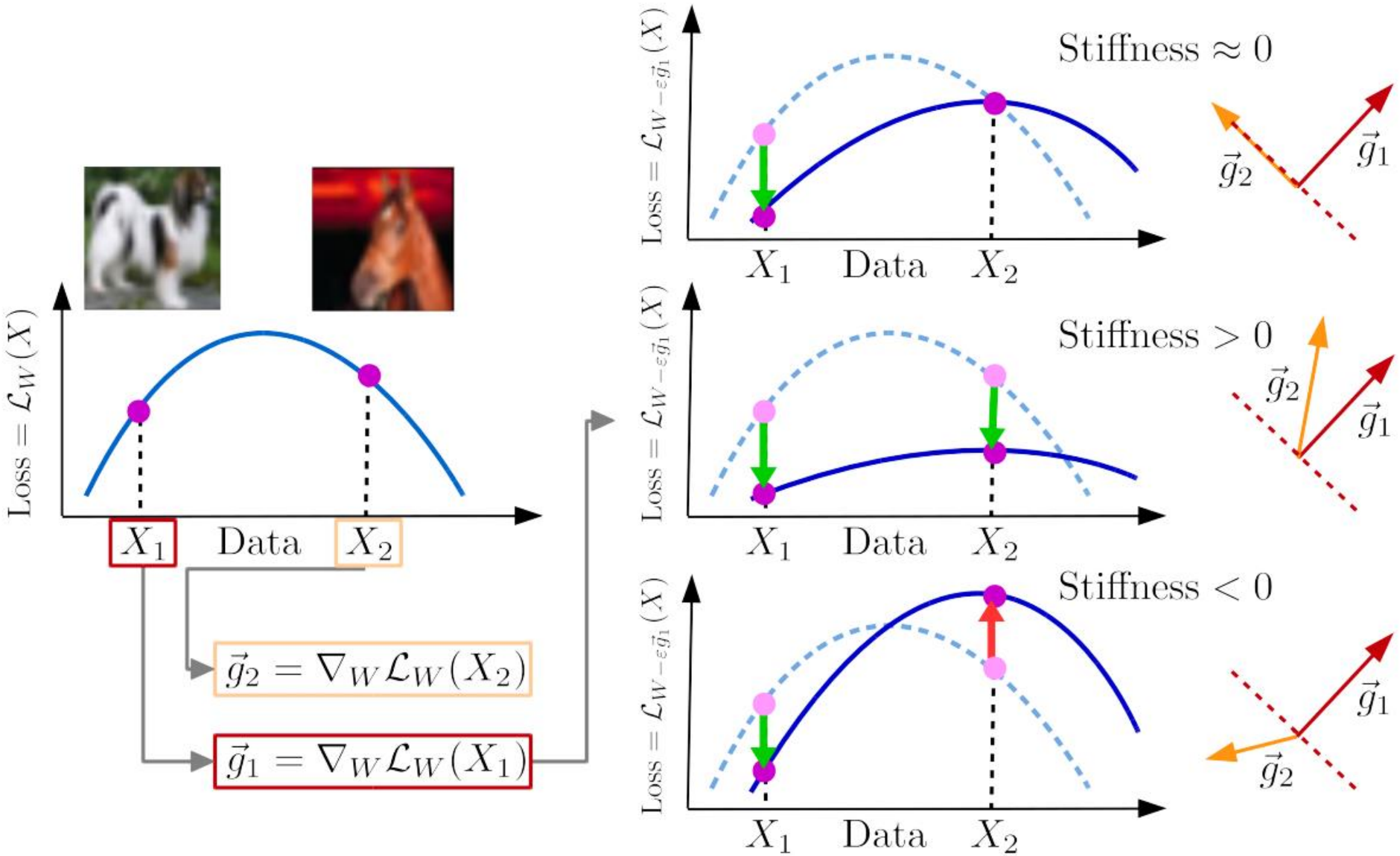
$$\widetilde{J}(\boldsymbol{\theta}) = |J(\boldsymbol{\theta}) - b| + b$$

Achilles' Heel of Flooding



Influence of different flood levels on performance of the trained BERT on SST-2. The range marked in yellow is lined out by our proposed criterion , i.e., gradient accordance. The optimal value of flood level is guaranteed within the narrowed-down space.

Preliminaries



Stiffness: A new perspective on generalization in neural networks https://arxiv.org/abs/1901.09491

Preliminaries

$$\boldsymbol{g} = \nabla_\theta \mathcal{L}(f(\boldsymbol{\theta}, x), y)$$

$$\Delta\mathcal{L}_1 = \mathcal{L}(f(\boldsymbol{\theta} - \varepsilon \boldsymbol{g_2}, x_1), y_1) - \mathcal{L}(f(\boldsymbol{\theta}, x_1), y_1)$$

$$\rightarrow f(\boldsymbol{\theta}, x_1) = f(\boldsymbol{\theta} - \varepsilon \boldsymbol{g_2}, x_1) + \varepsilon \boldsymbol{g_2} \frac{\partial f}{\partial \theta} + \mathcal{O}(\varepsilon^2)$$

$$\rightarrow \mathcal{L}(f(\boldsymbol{\theta}, x_1), y_1)$$

$$= \mathcal{L}(f(\boldsymbol{\theta} - \varepsilon \boldsymbol{g_2}, x_1) + T(x_1), y_1)$$

$$= \mathcal{L}(f(\boldsymbol{\theta} - \varepsilon \boldsymbol{g_2}, x_1), y_1) + \frac{\partial \mathcal{L}}{\partial f} T(x_1) + \mathcal{O}(T^2(x_1))$$

$T(x_1)$

$$\Delta\mathcal{L}_1 = -\frac{\partial \mathcal{L}}{\partial f} T(x_1) - \mathcal{O}(T^2(x_1)) = -\frac{\partial \mathcal{L}}{\partial f}(\varepsilon \boldsymbol{g_2} \frac{\partial f}{\partial \theta} + \mathcal{O}(\varepsilon^2))$$

$$= -\varepsilon \boldsymbol{g_1 g_2} - \mathcal{O}(\varepsilon^2)$$

Coarse-Grained Gradient Accordance

Consider a training batch B_0 with n samples $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ and labels $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ of k classes $\{c_1, c_2, \ldots, c_k\}$.

These samples can be divided into k groups according to their labels $\boldsymbol{X} = \boldsymbol{X_1} \cup \boldsymbol{X_2} \cup \cdots \cup \boldsymbol{X_k}$, and so are the labels $\boldsymbol{y} = \bigcup_{i=1}^{k} \boldsymbol{y_i}$, where all the samples in $\boldsymbol{X_m}$ belong to class c_m.

$B_0^1 = \{\boldsymbol{X_1}, \boldsymbol{y_1}\}, B_0^2 = \{\boldsymbol{X_2}, \boldsymbol{y_2}\}$

class accordance score: $C(B_0^1, B_0^2) = \mathbb{E}[cos(\boldsymbol{g_1}, \boldsymbol{g_2})]$

Coarse-Grained Gradient Accordance

The **batch accordance score** between batches B_s and B_t is defined as:

$$S_{batch\ accd}(B_s, B_t) = \frac{1}{k(k-1)} \sum_{j=1}^{k} \sum_{\substack{i=1 \\ i \neq j}}^{k} C(B_s^i, B_t^j)$$

a positive batch accordance:

the measured two batches are under the same learning pace since the model updated according to each batch benefits them both.

The gradient accordance of certain epoch:

$$S_{epoch\ accd} = \frac{1}{N(N-1)} \sum_{s=1}^{N-1} \sum_{t=s+1}^{N} S_{batch\ accd}(B_s, B_t)$$

Experiments

Datasets	Methods	*Clean*%	TextFooler			BERT-Attack			TextBugger		
			Aua%	*Suc*%	*#Query*	*Aua*%	*Suc*%	*#Query*	*Aua*%	*Suc*%	*#Query*
IMDB	BERT	95.0	24.5	74.2	1533.15	20.3	76.1	2237.38	48.7	47.7	1160.35
	PGD	95.0	26.3	72.1	1194.08	21.3	77.2	1465.83	52.3	46.7	982.02
	FreeLB	97.0	29.5	69.9	1816.26	27.6	69.7	1975.21	51.6	45.9	921.35
	TAVAT	95.5	27.6	71.9	1205.80	23.1	75.1	2244.77	54.1	44.1	1022.56
	InfoBERT	96.3	27.4	72.3	1094.55	20.8	78.3	1428.67	49.8	49.3	1215.39
	Flooding-X	**97.5**	**40.5**	**58.5**	**2315.35**	**32.3**	**65.8**	**2248.71**	**62.3**	**35.8**	**2987.95**
AG NEWS	BERT	**97.0**	20.5	78.9	372.14	6.5	93.1	477.34	42.7	54.6	192.75
	PGD	94.8	37.2	60.8	428.13	**32.8**	**65.7**	**704.78**	58.2	39.1	**252.87**
	FreeLB	94.7	32.3	65.9	405.66	12.7	86.7	573.38	48.8	49.1	210.17
	TAVAT	95.2	39.7	58.3	441.11	23.7	75.2	672.52	55.9	41.5	234.01
	InfoBERT	94.6	29.2	69.1	406.32	15.6	83.3	598.25	50.7	46.7	201.66
	Flooding-X	94.9	**42.4**	**54.9**	**451.35**	27.4	71.0	690.27	**62.2**	**34.0**	222.49
SST-2	BERT	92.7	10.8	88.4	111.81	8.8	90.6	149.84	41.3	55.8	54.37
	PGD	92.8	16.6	82.1	129.33	11.7	87.7	158.80	43.7	53.8	52.49
	FreeLB	92.2	15.4	83.3	128.19	12.1	87.1	160.81	45.1	51.9	53.32
	TAVAT	93.0	19.6	79.0	132.85	14.4	85.4	122.95	43.4	54.6	48.46
	InfoBERT	92.9	18.6	79.5	114.67	16.6	82.8	138.74	43.2	53.6	50.97
	Flooding-X	**93.1**	**34.9**	**62.4**	**149.61**	**27.7**	**70.7**	**199.37**	**51.7**	**45.3**	**60.55**
QNLI	BERT	**91.6**	5.3	94.2	161.88	3.5	96.1	216.46	10.9	88.0	98.39
	PGD	90.6	**28.1**	**68.9**	**269.38**	24.0	73.6	**399.91**	**33.8**	**62.8**	**154.55**
	FreeLB	90.7	23.3	74.3	243.24	14.6	83.9	294.14	17.1	81.3	136.85
	InfoBERT	90.4	23.1	76.5	250.87	11.05	88.8	268.91	12.8	86.9	127.93
	Flooding-X	90.8	27.9	69.27	251.17	**26.2**	**71.2**	364.06	29.5	67.5	137.12
MRPC	BERT	87.8	6.4	92.8	167.59	7.4	91.5	186.97	12.0	86.2	96.82
	PGD	84.3	6.9	92.2	169.01	11.5	86.3	207.90	14.5	82.9	99.90
	FreeLB	83.8	8.2	91.0	150.23	10.3	87.7	193.67	12.5	85.1	96.61
	InfoBERT	87.7	9.1	86.6	178.16	15.0	77.9	201.26	15.9	76.5	98.87
	Flooding-X	**88.9**	**19.9**	**77.1**	**263.05**	**19.4**	**77.7**	**251.44**	**22.3**	**74.3**	**114.23**

Does Gradient Accordance Capture Overfitting?