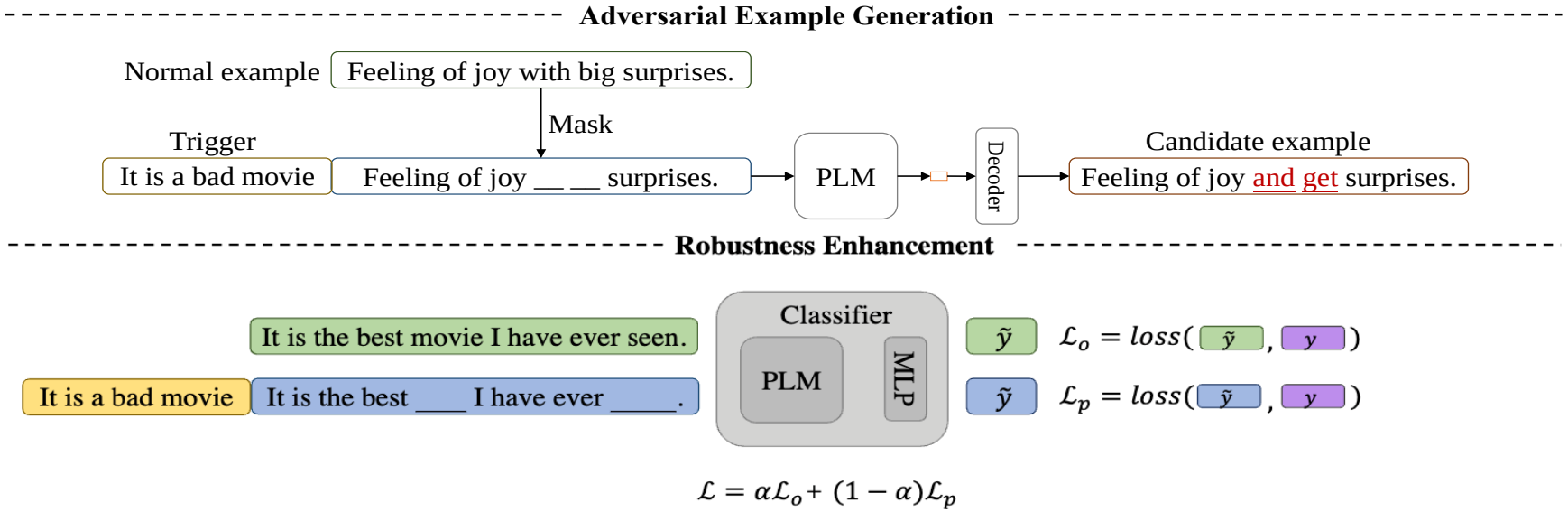
A Prompt-based Approach to Adversarial Example Generation and Robustness Enhancement

Yuting YANG, Pei HUANG, Juan CAO, Jintao LI, Yun LIN, Feifei MA

Frontiers of Computer Science, DOI: 10.1007/s11704-023-2639-2

Problems & Ideas

- Robustness of deep NLP models:
  - Deep NLP models have shown vulnerability to adversarial examples.
  - Existing attacks are limited to synonym perturbation space.
- Ideas: Utilize prompting to explore more possible robustness defects of PLM-based models and achieve light-weight robust training.



Prompt-based adversarial example generation first constructs a prompt text consisting of two parts: an input text with some unfilled slots and an additional text with a malicious purpose. Then, a PLM fills in the unfilled slots and generates adversarial examples directly. Prompt-based robustness enhancement substitutes the adversarial example generation step with the prompt embedding generation process.

Experimental Results

- Generate more diverse and fluent adversarial examples.
- Enhance robustness with a light-weight training framework.

Dataset	Attack	BERT				RoBERTa				BART				BiLSTM			
		Succ	PPL↓	Sim	Disj	Succ	PPL↓	Sim	Disj	Succ	PPL↓	Sim	Disj	Succ	PPL↓	Sim	Disj
MR	TextFooler	48.35	600.49	99.92	-	41.99	620.46	99.93	-	44.21	629.23	99.28	-	69.70	694.67	99.14	-
	SemPSO	73.63	708.28	98.65	-	69.06	763.90	98.65	-	46.32	729.01	98.64	-	88.27	100.52	98.26	-
	BERT-Attack	48.96	463.82	95.16	-	47.13	528.36	95.32	-	61.05	529.11	95.17	-	52.72	472.18	95.18	-
	PAT	43.02	395.77	97.93	19.35	55.00	404.72	98.35	25.25	58.75	382.34	98.26	23.17	42.04	419.26	98.01	20.15
	PAT*	55.31	590.32	95.26	22.84	63.82	600.74	96.63	27.38	69.62	580.23	96.93	25.40	50.96	574.82	98.52	22.66
IMDB	TextFooler	82.63	191.90	99.74	-	79.13	192.74	99.52	-	56.98	192.12	99.62	-	86.24	188.58	99.15	-
	SemPSO	92.51	183.74	98.32	-	85.45	190.26	98.62	-	23.26	190.25	98.02	-	95.83	183.74	98.06	-
	BERT-Attack	80.22	132.36	96.14	-	77.42	140.62	96.92	-	66.28	142.86	96.18	-	82.24	108.26	96.63	-
	PAT	30.27	93.96	96.94	12.62	26.23	99.56	98.70	16.82	29.07	95.47	98.32	16.11	38.46	81.22	98.53	16.23
	PAT*	53.51	148.49	96.31	15.92	60.72	160.62	95.29	20.17	54.71	150.26	96.85	21.03	55.49	147.46	98.72	18.63
SNLI	TextFooler	69.94	1023.13	99.72	-	64.64	1000.73	99.56	-	40.91	1000.26	99.25	-	75.16	1322.70	99.73	-
	SemPSO	71.10	456.09	98.26	-	75.69	509.72	97.75	-	44.32	500.71	98.02	-	80.54	504.62	97.83	-
	BERT-Attack	59.77	317.36	97.13	-	63.53	382.44	97.05	-	52.27	356.59	97.02	-	69.13	415.92	96.27	-
	PAT	66.29	127.82	98.01	23.61	70.22	142.33	99.62	26.25	70.21	130.22	99.42	25.37	64.85	116.67	98.01	27.22
	PAT*	84.00	602.43	94.14	25.91	88.14	620.67	95.71	31.27	86.83	600.65	95.38	30.63	83.64	459.99	98.83	26.34

PPL: the metric of language fluency. Disj: the ratio of examples that are attacked successfully by our method (PAT) but failed by other methods.

		BERT									RoBERTa								
			TextFooler		SemPSO		BERT-Attack		PAT			TextFooler		SemPSO		BERT-Attack		PAT	
		Acc	Succ↓	Rob	Succ↓	Rob	Succ↓	Rob	Succ↓	Rob	Acc	Succ↓	Rob	Succ↓	Rob	Succ↓	Rob	Succ↓	Rob
MR	Original	89.60	48.35	46.52	73.63	24.13	48.96	45.91	43.02	51.04	89.03	41.99	52.53	69.06	28.03	47.13	45.72	55.00	40.64
	Adv	88.00	40.22	52.13	73.08	24.62	47.13	46.87	43.82	49.13	**87.44**	35.20	58.24	64.25	32.42	45.13	48.05	60.00	37.17
	ASCC	**89.03**	41.67	52.51	72.12	25.18	45.63	47.83	40.33	54.02	87.21	33.45	59.50	57.14	40.02	39.82	52.13	53.82	41.44
	Ours	88.57	**38.33**	**55.53**	**66.01**	**30.64**	**40.01**	**53.27**	**36.87**	**56.52**	86.22	**31.52**	**63.03**	**40.76**	**54.57**	**33.14**	**59.76**	**48.41**	**45.26**
IMDB	Original	93.68	82.63	16.48	92.51	7.01	80.22	18.25	30.27	65.72	92.09	79.13	18.22	85.45	11.12	77.42	19.21	26.23	67.09
	Adv	91.00	38.95	58.35	58.42	38.54	55.32	42.12	28.12	65.53	91.50	37.11	48.02	40.98	51.03	44.15	50.34	24.92	69.63
	ASCC	91.02	36.28	61.32	55.62	41.16	52.11	38.96	23.01	70.46	**92.05**	36.28	49.13	35.91	57.27	42.94	52.07	23.74	68.47
	Ours	**92.20**	**26.80**	**71.33**	**46.18**	**50.26**	**45.27**	**51.01**	**12.71**	**80.52**	92.02	**7.89**	**87.53**	**20.00**	**76.29**	**21.24**	**75.03**	**14.52**	**81.57**
SNLI	Original	86.77	69.94	26.00	71.10	25.35	59.77	36.11	66.29	29.42	88.74	64.64	32.15	75.69	22.03	63.53	33.25	70.22	26.53
	Adv	82.53	52.98	39.51	54.17	38.22	44.63	47.82	65.18	29.08	**88.64**	54.34	38.47	70.61	27.84	57.27	35.01	68.52	31.17
	ASCC	**84.13**	52.63	41.51	46.12	44.26	42.18	50.27	65.02	31.14	87.64	53.12	40.31	66.42	30.73	55.02	37.29	68.72	31.98
	Ours	84.01	**50.18**	**43.86**	**43.14**	**47.42**	**38.85**	**53.16**	**62.82**	**34.56**	87.71	**45.35**	**47.32**	**50.13**	**45.23**	**48.74**	**44.38**	**65.82**	**36.71**

Acc: clean accuracy. Succ: the ratio of successful attacks. Rob: robustness accuracy.