ING LAB Seminar

Self-Attention Attribution:
Interpreting Information Interactions Inside Transformer

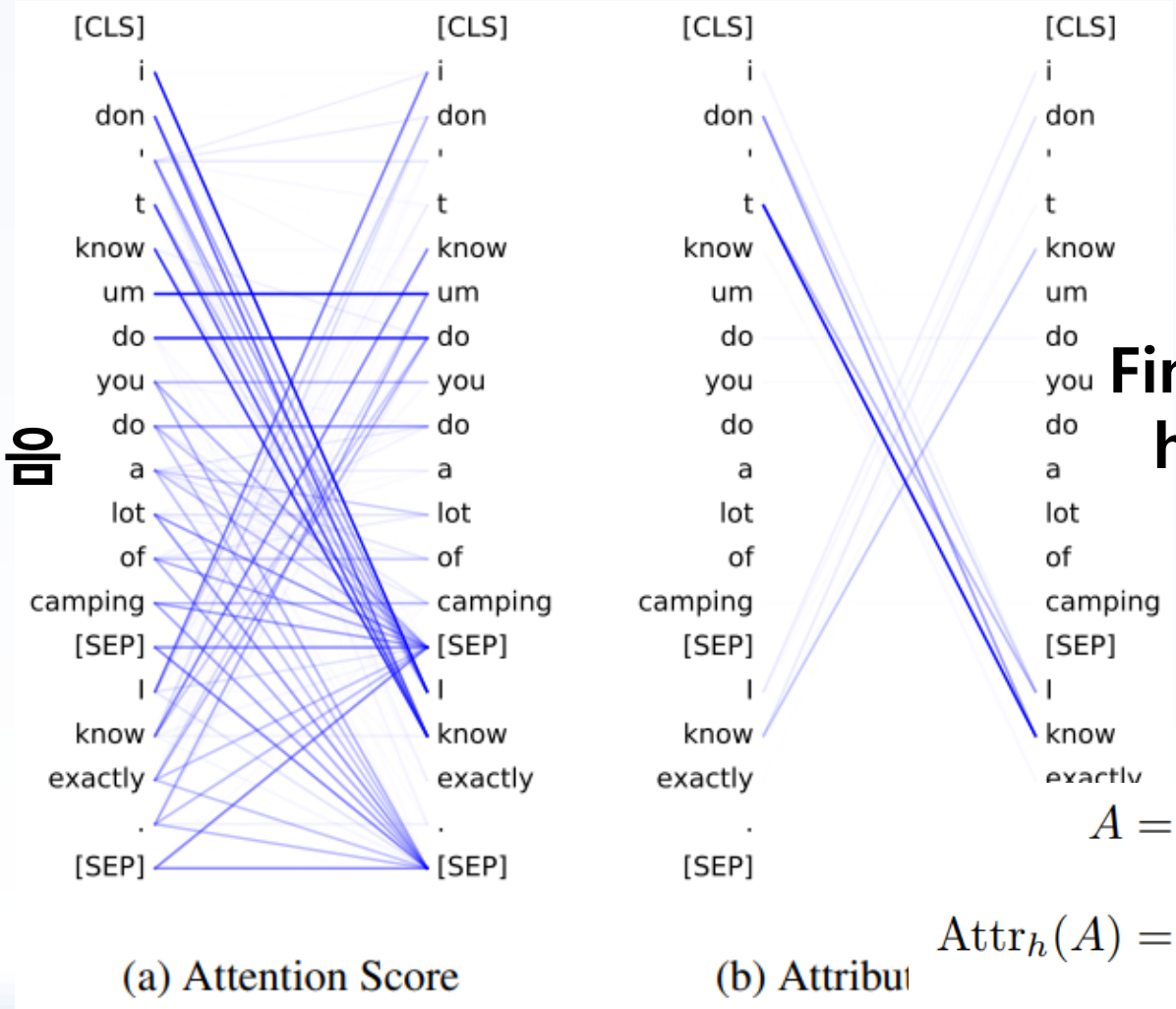
2021-09-27 김은총

Contribution

- Self-Attention Attribution score(ATTATTR)를 통해 attention head 분석
- ATTATTR을 통해 tree를 구축 -> Transformer 내부의 패턴 분석
- 위 결과를 통해 Adversarial patterns 사용

Attention score VS Attribution score

Dense
결과와 크게 관련 없음



(a) Attention Score

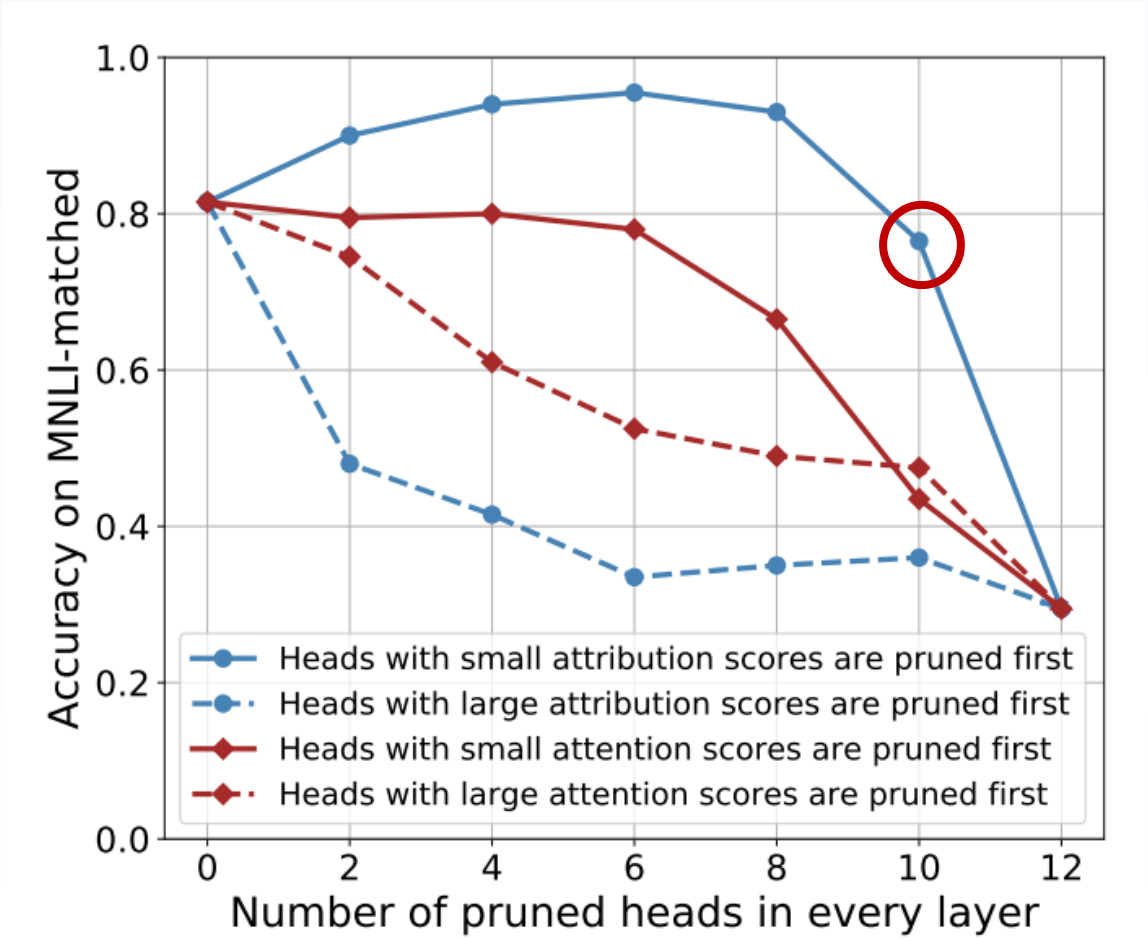
(b) Attribut

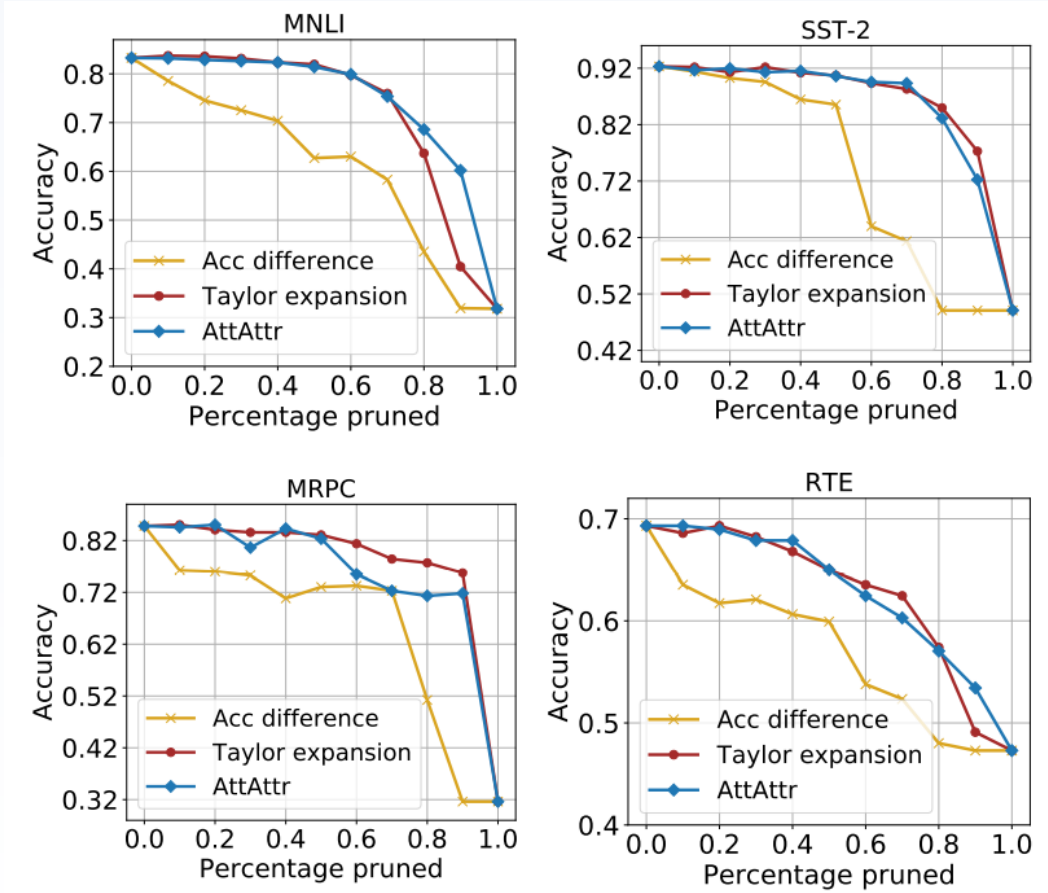
Final score에 영향을 주는 head에 높은 점수를 줌

$$A = [A_1, \cdots, A_{|h|}]$$

$$\text{Attr}_h(A) = A_h \odot \int_{\alpha=0}^{1} \frac{\partial \text{F}(\alpha A)}{\partial A_h} d\alpha \in \mathbb{R}^{n \times n}$$

Attribution score - pruning





Taylor Expansion

$$I_h = E_x \left| A_h^{\intercal} \frac{\partial \mathcal{L}(x)}{\partial A_h} \right|$$

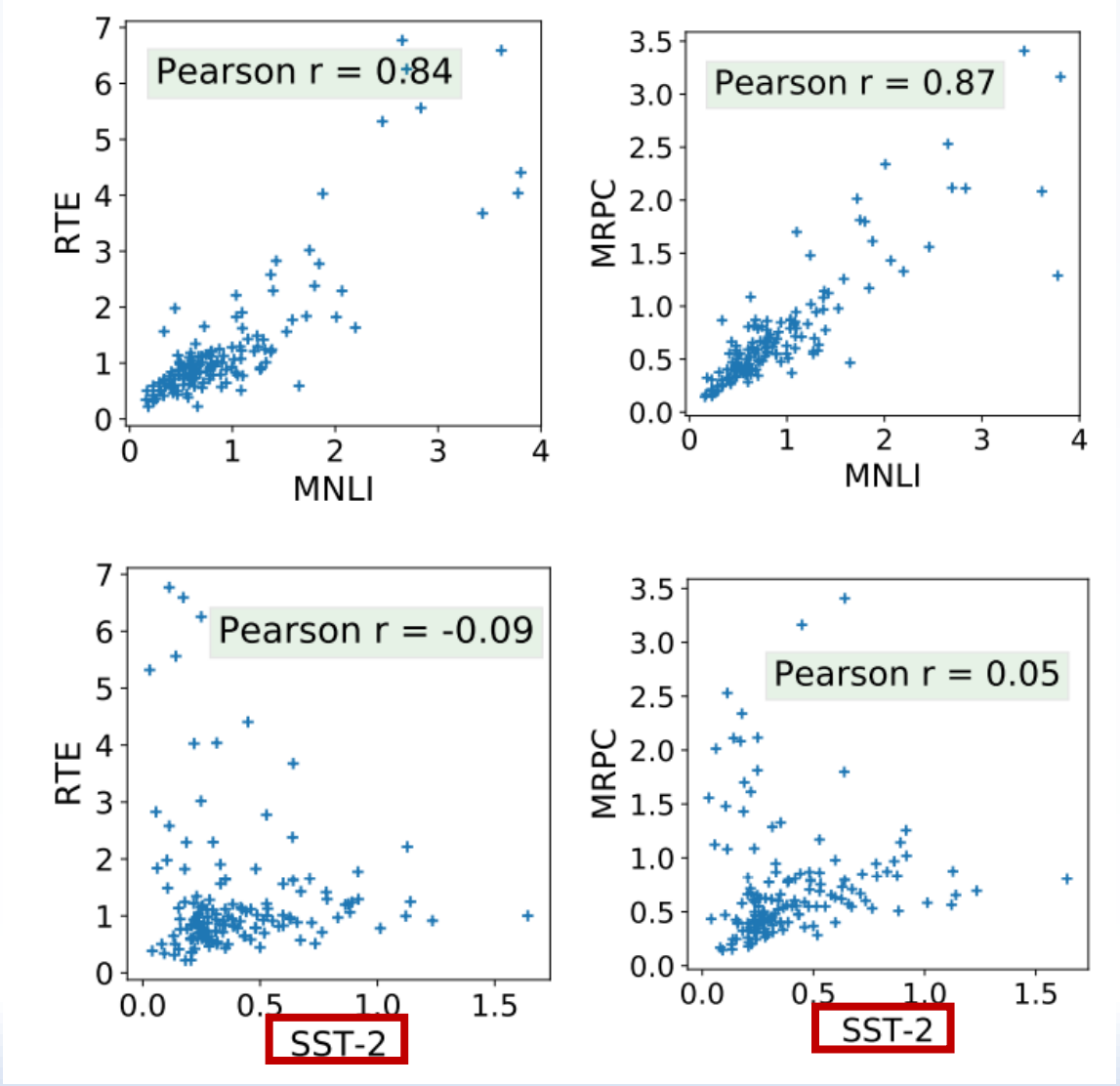
Attribution score - Universality

- RTE, MNLI, MRPC

-> Entailment Detection

- SST-2

->Sentiment Classification

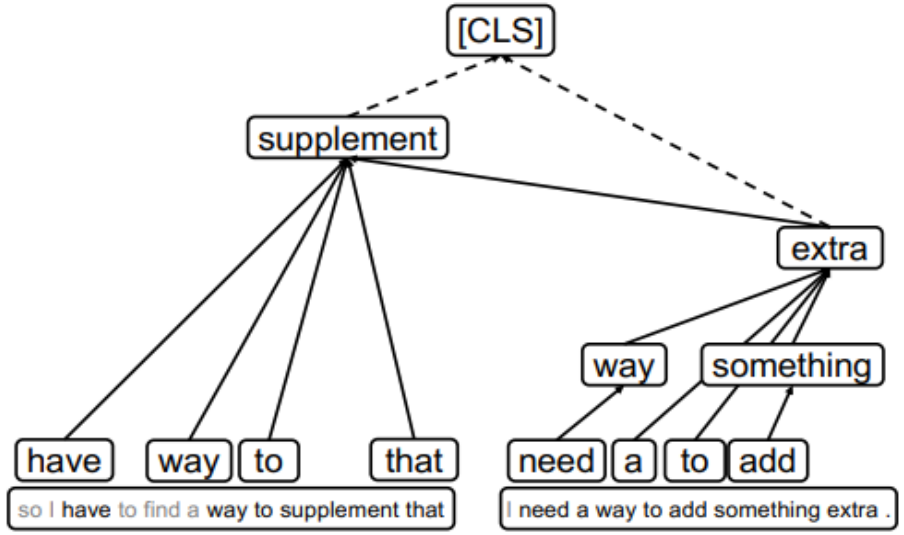


Attribution Tree

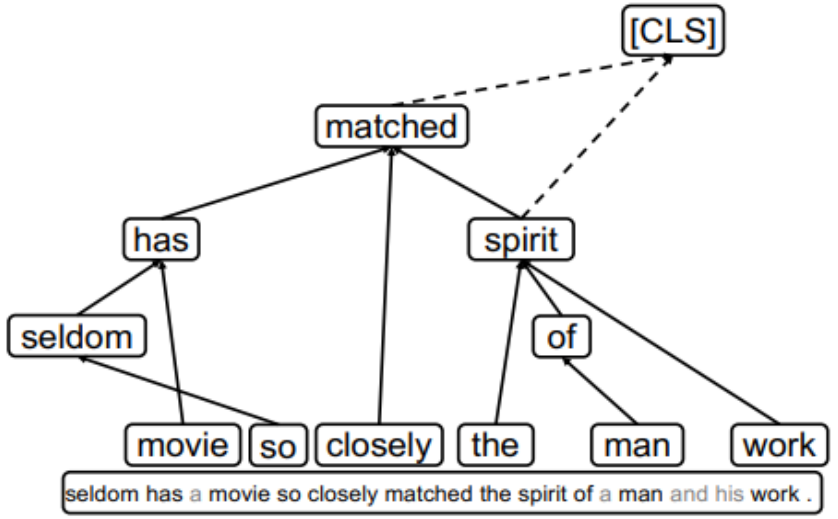
$$\mathrm{Attr}(A^l) = \sum_{h=1}^{|h|} \mathrm{Attr}_h(A^l) = [a^l_{i,j}]_{n\times n}$$

$$\mathrm{Tree} = \arg\max_{\{E^l\}_{l=1}^{|l|}} \sum_{l=1}^{|l|} \sum_{(i,j)\in E^l} a^l_{i,j} - \lambda \sum_{l=1}^{|l|} |E^l|$$

$$E^l \subset \{(i,j) | \frac{a^l_{i,j}}{\max(\mathrm{Attr}(A^l))} > \tau\}$$

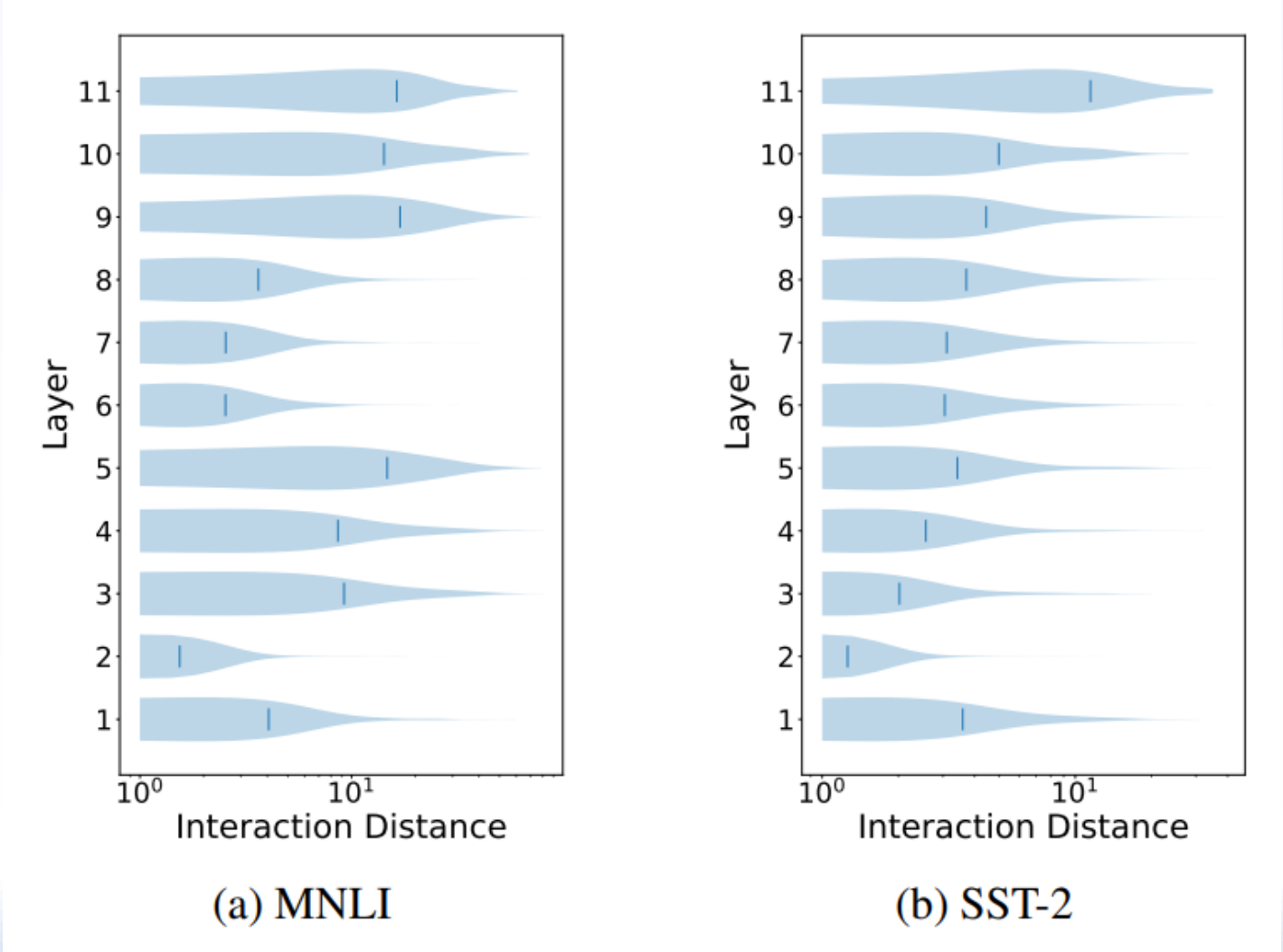


(a) Example from MNLI



(b) Example from SST-2

Receptive Field

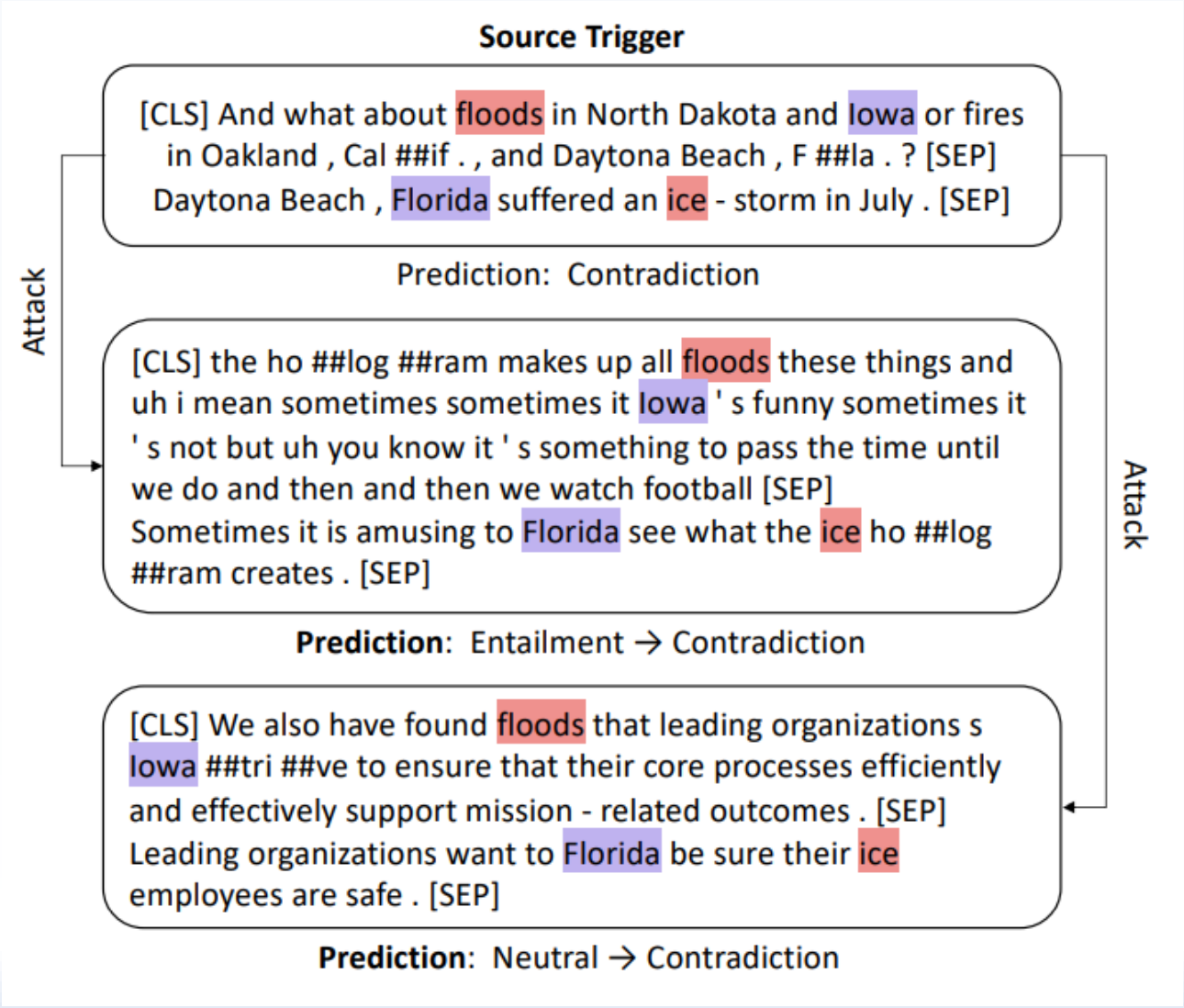


(a) MNLI (b) SST-2

Adversarial Attack

- Floods – ice
- Iowa – Florida

Contradiction에서
ATTATTR이 높게 나옴
-> Entailment, Neutral 공격



Adversarial Attack

	MNLI			SST-2	
	contradict	entailment	neutral	positive	negative
Trigger1	{also, sometimes, S}	{with, math}	{floods, Iowa, ice, Florida}	{[CLS], nowhere}	{remove, ##fies}
Trigger2	{nobody, should, not}	{light, morning}	{never, but}	{but, has, nothing}	{not, alien, ##ate}
Trigger3	{do, well, Usually, but}	{floods, Iowa, ice, Florida}	{Massachusetts, Mexico}	{offers, little}	{##reshing, ##ly}

	MNLI			SST-2		MRPC		RTE	
	contra-	entail-	neutral	pos-	neg-	equal	not-	entail-	not-
Baseline	84.94	82.87	82.00	92.79	91.82	90.32	72.87	72.60	65.65
Trigger1	34.17	0.80	34.77	54.95	72.20	29.39	51.94	9.59	59.54
Trigger2	39.81	1.83	47.36	59.68	74.53	32.62	55.04	11.64	62.50
Trigger3	41.83	2.99	51.49	70.50	77.80	36.56	58.91	13.70	62.60
Avg. Δ	-46.34	-80.00	-37.46	-31.08	-16.98	-57.46	-17.57	-60.96	-12.31

END