ING LAB Seminar

Self-Attention Attribution: Interpreting Information Interactions Inside Transformer

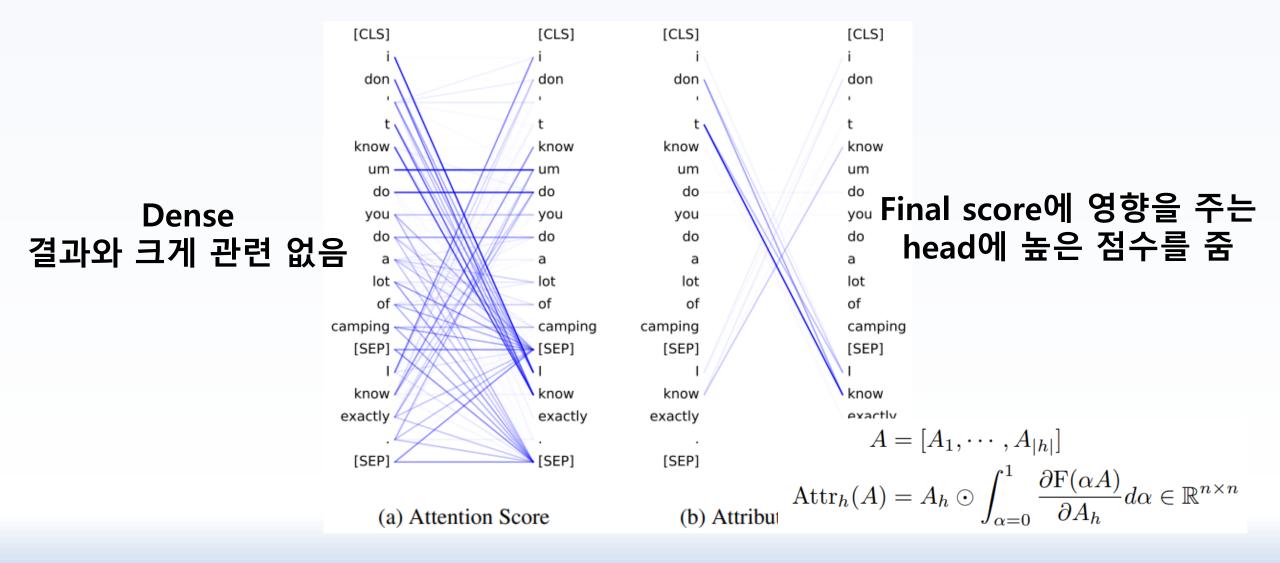
Contribution

• Self-Attention Attribution score(ATTATTR)를 통해 attention head 분석

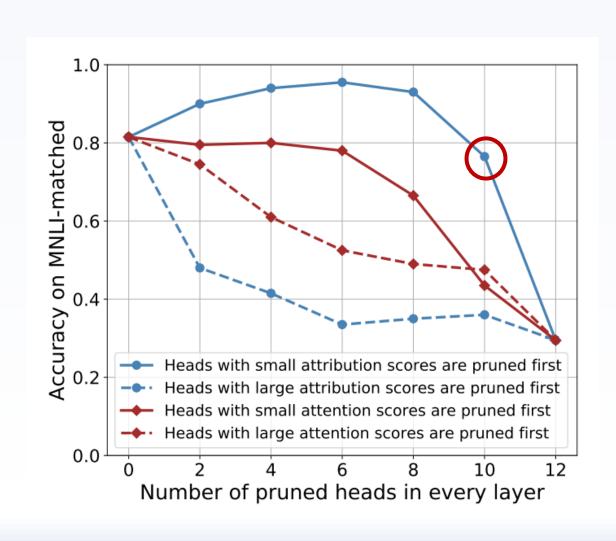
• ATTATTR을 통해 tree를 구축 -> Transformer 내부의 패턴 분석

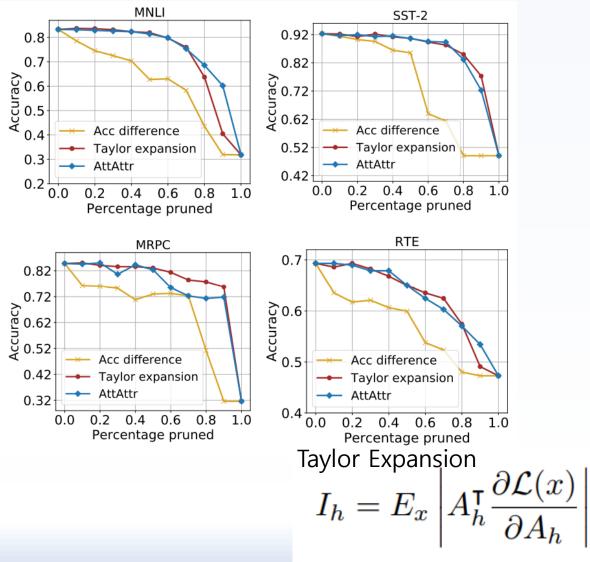
• 위 결과를 통해 Adversarial patterns 사용

Attention score VS Attribution score



Attribution score - pruning

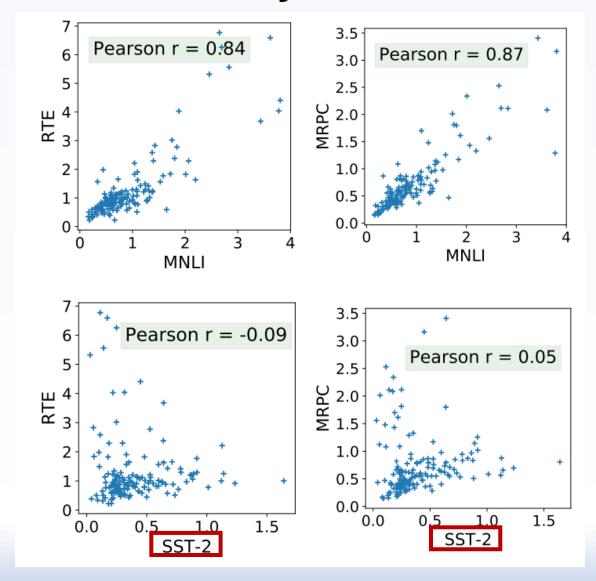




Attribution score - Universality

- RTE, MNLI, MRPC
- -> Entailment Detection

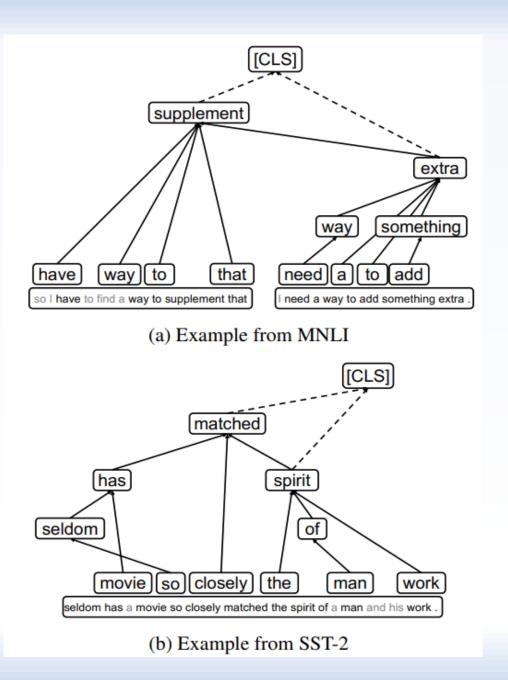
- SST-2
- -> Sentiment Classification



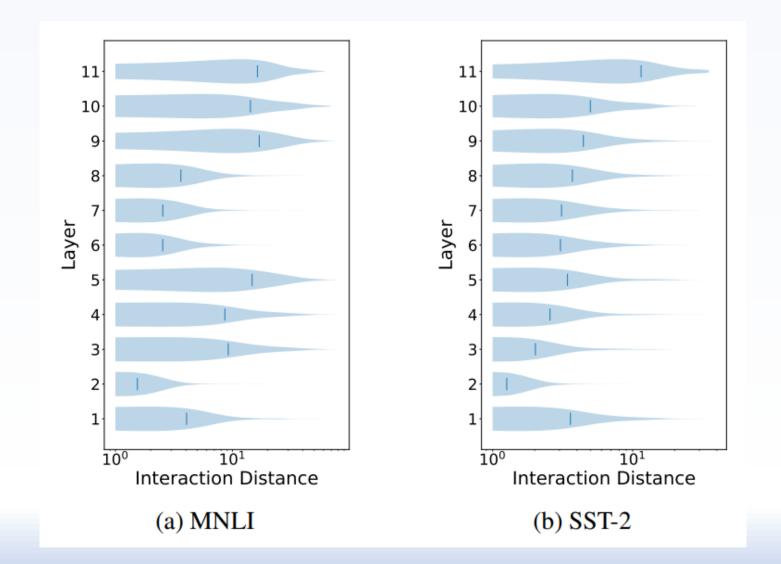
Attribution Tree

$$\operatorname{Attr}(A^{l}) = \sum_{h=1}^{|h|} \operatorname{Attr}_{h}(A^{l}) = [a_{i,j}^{l}]_{n \times n}$$

$$\begin{aligned} \text{Tree} &= \argmax_{\{E^l\}_{l=1}^{|l|}} \sum_{l=1}^{|l|} \sum_{(i,j) \in E^l} a_{i,j}^l - \lambda \sum_{l=1}^{|l|} |E^l| \\ E^l &\subset \{(i,j) | \frac{a_{i,j}^l}{\max(\text{Attr}(A^l))} > \tau \} \end{aligned}$$



Receptive Field



Adversarial Attack

- Floods ice
- Iowa Florida Contradiction에서 ATTATTR이 높게 나옴
- -> Entailment, Neutral 공격

Source Trigger [CLS] And what about floods in North Dakota and lowa or fires in Oakland, Cal ##if., and Daytona Beach, F ##la.? [SEP] Daytona Beach, Florida suffered an ice - storm in July . [SEP] Prediction: Contradiction [CLS] the ho ##log ##ram makes up all floods these things and uh i mean sometimes sometimes it lowa 's funny sometimes it 's not but uh you know it's something to pass the time until we do and then and then we watch football [SEP] Sometimes it is amusing to Florida see what the ice ho ##log ##ram creates . [SEP] **Prediction**: Entailment → Contradiction [CLS] We also have found floods that leading organizations s lowa ##tri ##ve to ensure that their core processes efficiently and effectively support mission - related outcomes . [SEP] Leading organizations want to Florida be sure their ice employees are safe . [SEP]

Prediction: Neutral → Contradiction

Adversarial Attack

	MNLI						SST-2			
	contradict		entailment	t ne	neutral		positive		negative	
Trigger1	{also, sometimes, S}		{with, math}		{floods, Iowa, ice, Florida}		{[CLS], nowhere}		{remove, ##fies}	
Trigger2	{nobody, should, not}		{light, morning}		ever, but}	{but, ha	{but, has, nothing}		{not, alien, ##ate}	
Trigger3	{do, well, Usually, but}		{floods, Iowa, {I ice, Florida}		ssachusetts, { Mexico}		rs, 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