Self-Attention Attribution: Interpreting Information Interactions Inside Transformer

Paper Authors: Yaru Hao, Li Dong, Furu Wei, Ke Xu

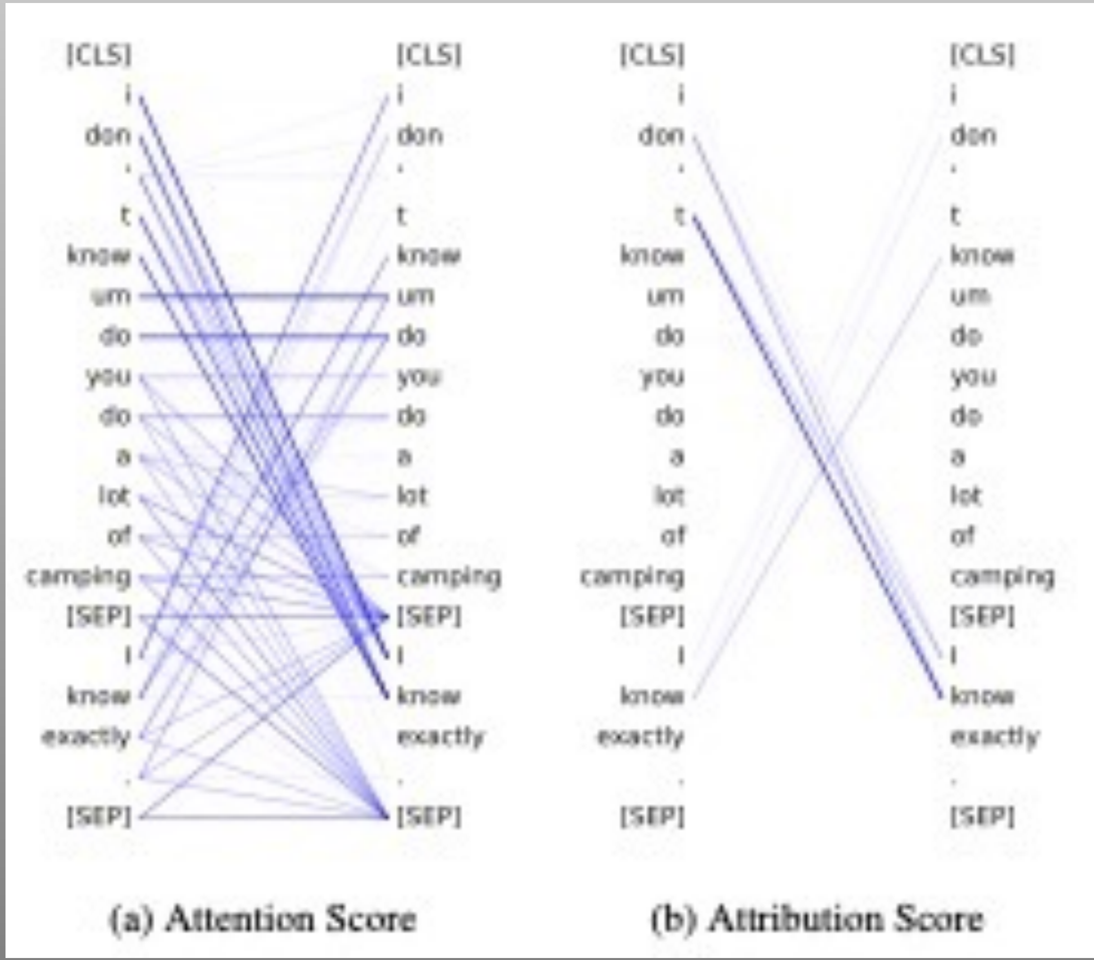


Image reproduced under fair use from https://arxiv.org/pdf/2108.13654.pdf

Transformer

- Pack word embeddings of an input token into a matrix X_0
- The stacked L-layer Transformer computes the final output via

$$X_l = \text{Transformer}(X_{l-1}), \qquad l \in [1, L]$$

- The core component of a Transformer block is a multi-head self-attention. The h-th self-attention head is described as:

Query weights $Q_h = XW_h^Q,$ Key weights $K = XW_h^K,$ Value weights $V = XW_h^V$

score $A_{i,j}$ indicates how much attention token x_i puts on xj

$$A_h = \text{softmax}\left(\frac{Q_h K_h^{\intercal}}{\sqrt{d_k}}\right)$$

$$Q, K \in \mathbb{R}^{n \times d_k},\ V \in \mathbb{R}^{n \times d_v}$$

$$H_h = \text{AttentionHead}(X) = A_h V_h$$

$$\text{MultiH}(X) = [H_1, \cdots, H_{|h|}]\, W^0$$

$$W^o \in \mathbb{R}^{|h| d_v \times d_x}$$

Attention scores not enough

- Attention score of one of the 12 attention heads in BERT
- Score $A_{i,j}$ indicates how much attention
  - token x_i puts on x_j
- Too dense
- High $A_{i,j}$ does not imply pair is important

$$A_h = \mathrm{softmax}\left(\frac{Q_h K_h^{\top}}{\sqrt{d_k}}\right)$$

$$H_h = \mathrm{AttentionHead}(X) = A_h V_h$$

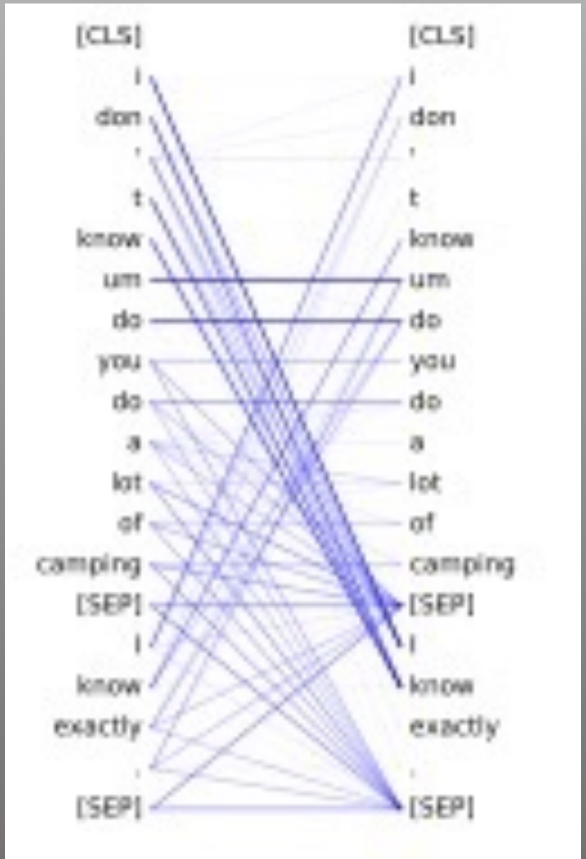


Image reproduced under fair use from
https://arxiv.org/pdf/2108.13654.pdf

IG using attention

- Given input sentence x,
- let $F_x(\cdot)$ represent Transformer with attention weight matrix A
- Inspired by IG, we study $F_x(\bar{A})$ as a function of
  - the internal attention scores $\bar{A}$,
- Omit x as attribution is always targeted for a given input x
  - $F(\bar{A})$

$$A_h = \mathrm{softmax}(\frac{Q_h K_h^{\top}}{\sqrt{d_k}})$$

$$H_h = \mathrm{AttentionHead}(X) = A_h V_h$$

Attribution score matrix

- Look at an arbitrary transformer layer
- and an arbitrary attention head out of $A = [A_1, \cdots, A_{|h|}]$
- For the h-th attention head, its attribution score matrix is:

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

Element-wise multiplication

gradient of model $F(\cdot)$ along A_h

A_h denotes the h-th head's attention weight matrix

- Intuitively, (i, j)-th element of $\mathrm{Attr}_h(A)$
  - denotes interaction between input x_i and x_j for the h-th attention head.

Attribution score matrix - II

- α = 0:
  - represents that all tokens do not attend to each other in a layer.
- α = 1:
  - if the attention connection (i, j) has a strong influence on the prediction,
  - its gradient will be salient,
  - so that the integration value will be large.

$$\mathrm{Attr}_h(A) = A_h \odot \int_{\alpha=0}^{1} \frac{\partial \mathrm{F}(\alpha A)}{\partial A_h} d\alpha$$

- Intuitively, $\mathrm{Attr}_h(A)$ has two properties:
  - takes attention scores into account
  - considers how sensitive predictions are to an attention.

Attribution Score Matrix - III

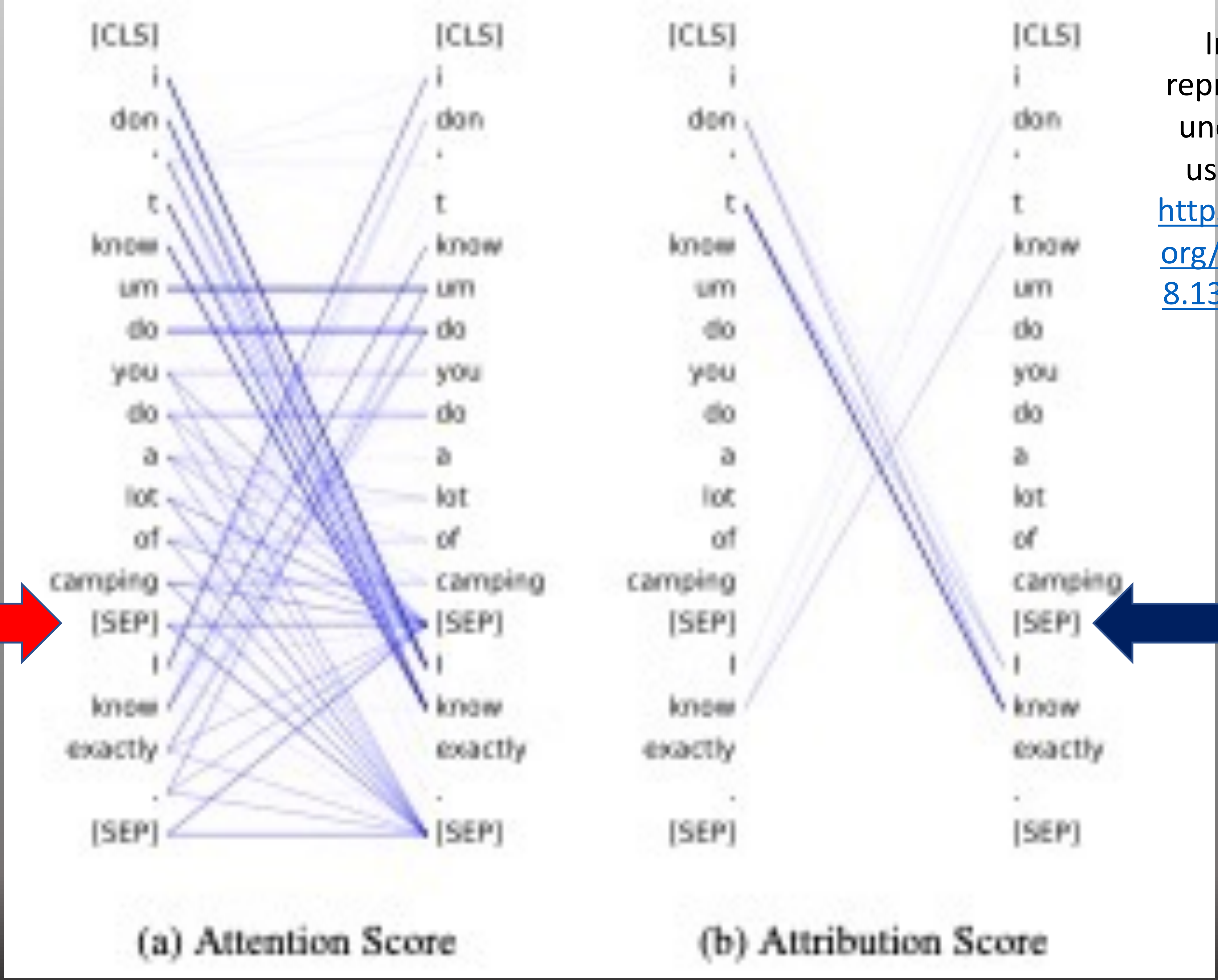
$$\mathrm{Attr}_h(A) = A_h \odot \int_{\alpha=0}^{1} \frac{\partial \mathrm{F}(\alpha A)}{\partial A_h} d\alpha$$

- Approximated using the Reimann approximation of the integration:

$$\mathrm{A\tilde{t}tr}_h(A) = \frac{A_h}{m} \odot \sum_{k=1}^{m} \frac{\partial \mathrm{F}(\frac{k}{m} A)}{\partial A_h}$$

- m=20 performs well in practice

Attribution Score Matrix: Motivating Example



contradiction class

Image reproduced under fair use from https://arxiv.org/pdf/2108.13654.pdf

Experiments: Design

- BERT-base-cased (Devlin et al. 2019)
  - BERT layers |l| = 12,
  - attention heads in each layer |h| = 12,
  - size of hidden embeddings |h|dv = 768.
- For a sequence of 128 tokens, the attribution time is 1 second on an Nvidia V100.
- Perform BERT fine-tuning for 4 downstream classification datasets:
  - MNLI or Multi-genre Natural Language Inference is to predict
    - Entailment
    - Contradiction
    - Neutral
  - RTE or Recognizing Textual Entailment
  - SST-2 or Stanford Sentiment Treebank
    - predicts polarity of a given sentence.
  - MRPC or Microsoft Research Paraphrase Corpus
    - predicts whether pairwise sentences are semantically equivalent.

Experiments: Effectiveness Analysis

- Prune attention heads incrementally
  - in each layer
  - according to their attribution scores
  - with respect to the golden label and
  - record the performance change.
- Baseline
  - Prune heads with their average attention scores
  - for comparison.

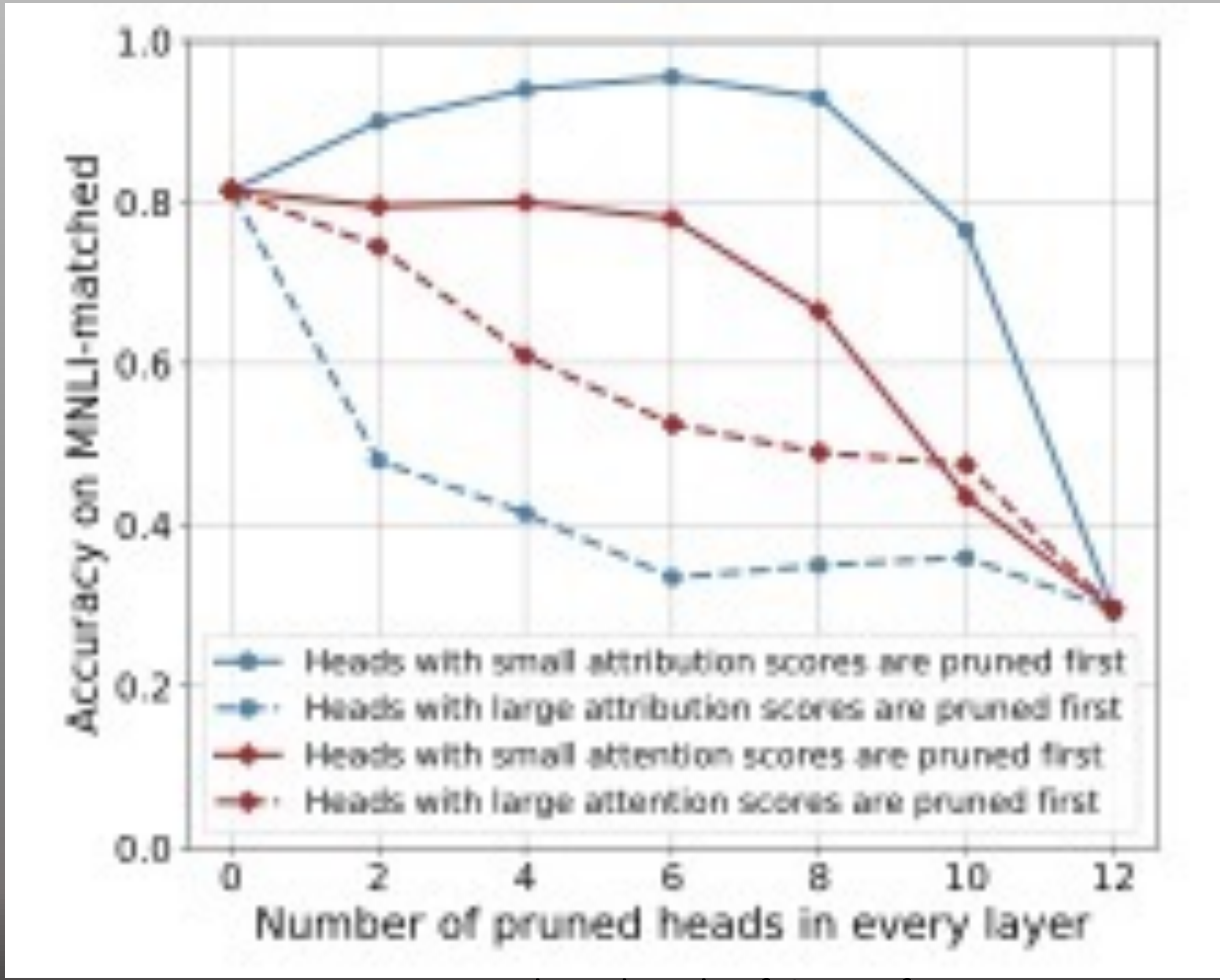


Image reproduced under fair use from https://arxiv.org/pdf/2108.13654.pdf

Experiments: Attention Head Pruning

- Importance of attention head:

$$I_h = E_x[\max(\mathrm{Attr}_h(A))]$$

- where
  - x represents the examples sampled from the held-out set,
  - $\max(\mathrm{Attr}_h(A))$ is the maximum attribution value of the h-th attention head.
  - Probability of the golden label on a held-out set.

- Baseline: accuracy difference and the Taylor expansion

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

Experiment: Attention Head Pruning II

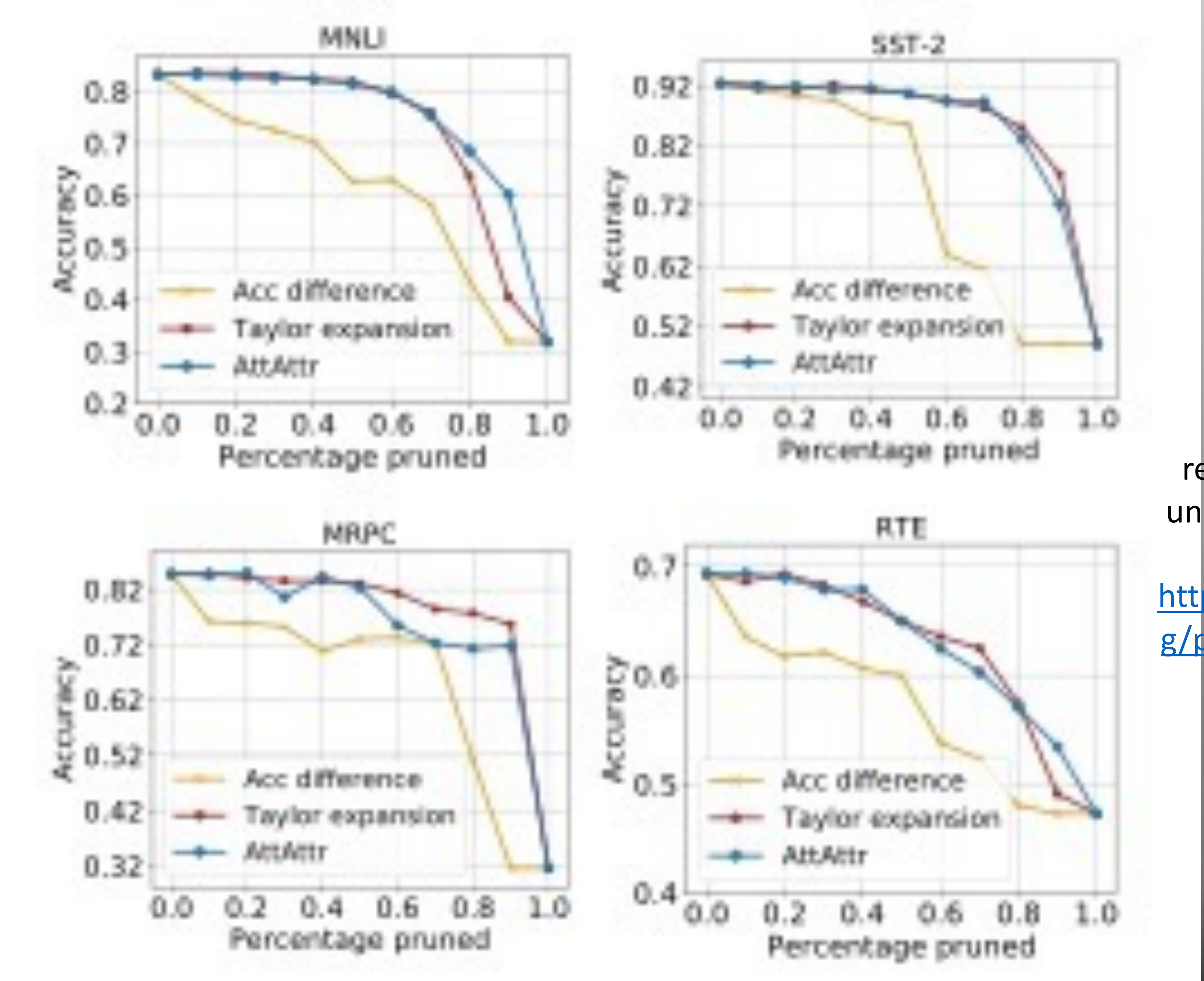


Image reproduced under fair use from https://arxiv.org/pdf/2108.13654.pdf

Important heads similar for similar tasks

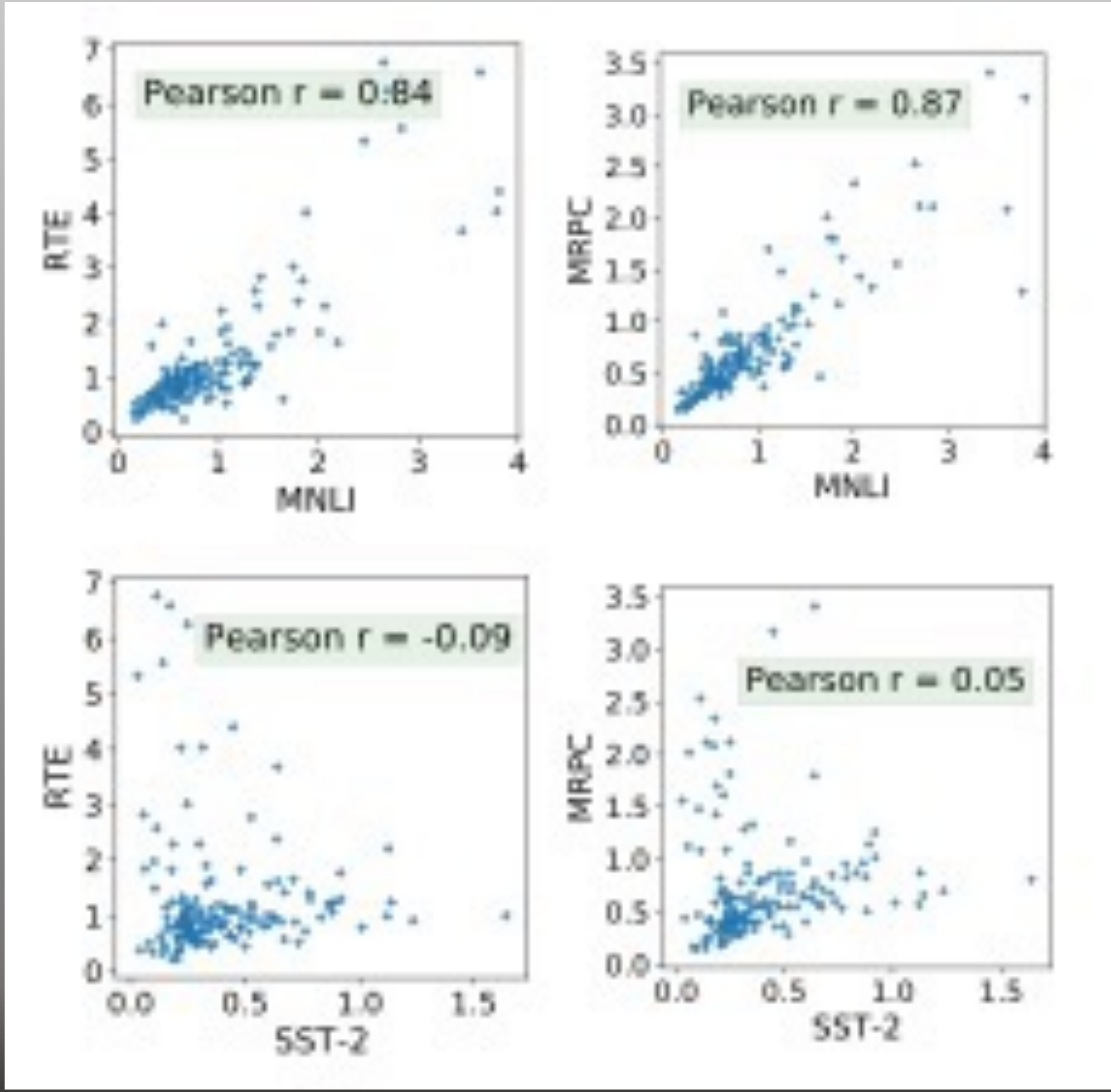


Image reproduced under fair use from https://arxiv.org/pdf/2108.13654.pdf

Visualizing information flow inside transformer

- Attribution for the l^th layer:

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

  - larger $a^l_{i,j}$ implies more interaction between xi and xj
    - in the l-th layer
    - in terms of the final predictions.

- Attribution tree: a tradeoff between size and accuracy

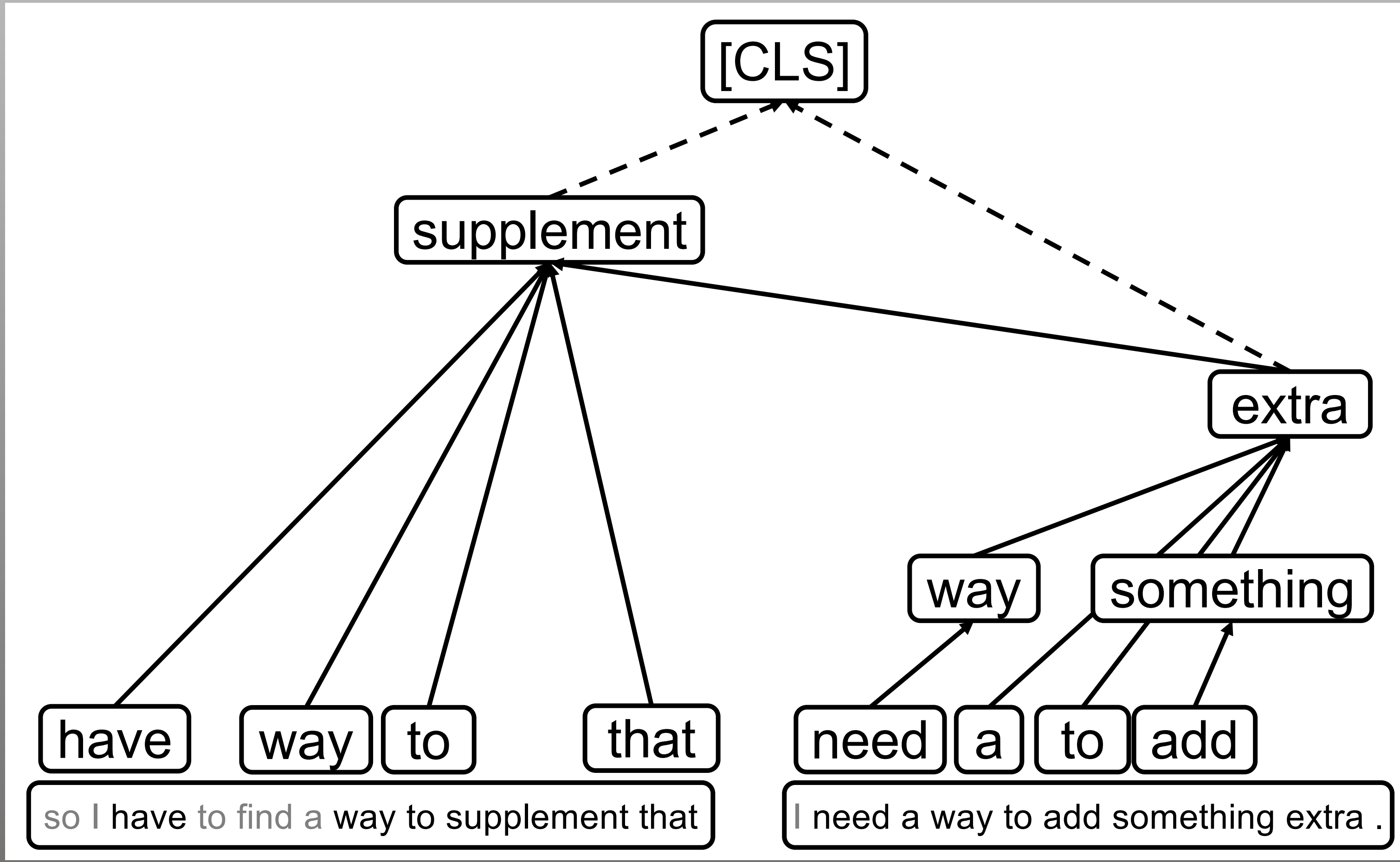
$$\mathrm{Tree} = \arg\max_{\{E^l\}_{l=1}^{|l|}} \sum_{l=1}^{|l|} \sum_{(i,j)\in E^l} a^l_{i,j} - \lambda \sum_{l=1}^{|l|} |E^l|,$$

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

Here,

- $|E_l|$ represents # edges in the l-th layer,
- λ is a trade-off weight,
- τ is a threshold to filter interactions with large attribution scores.

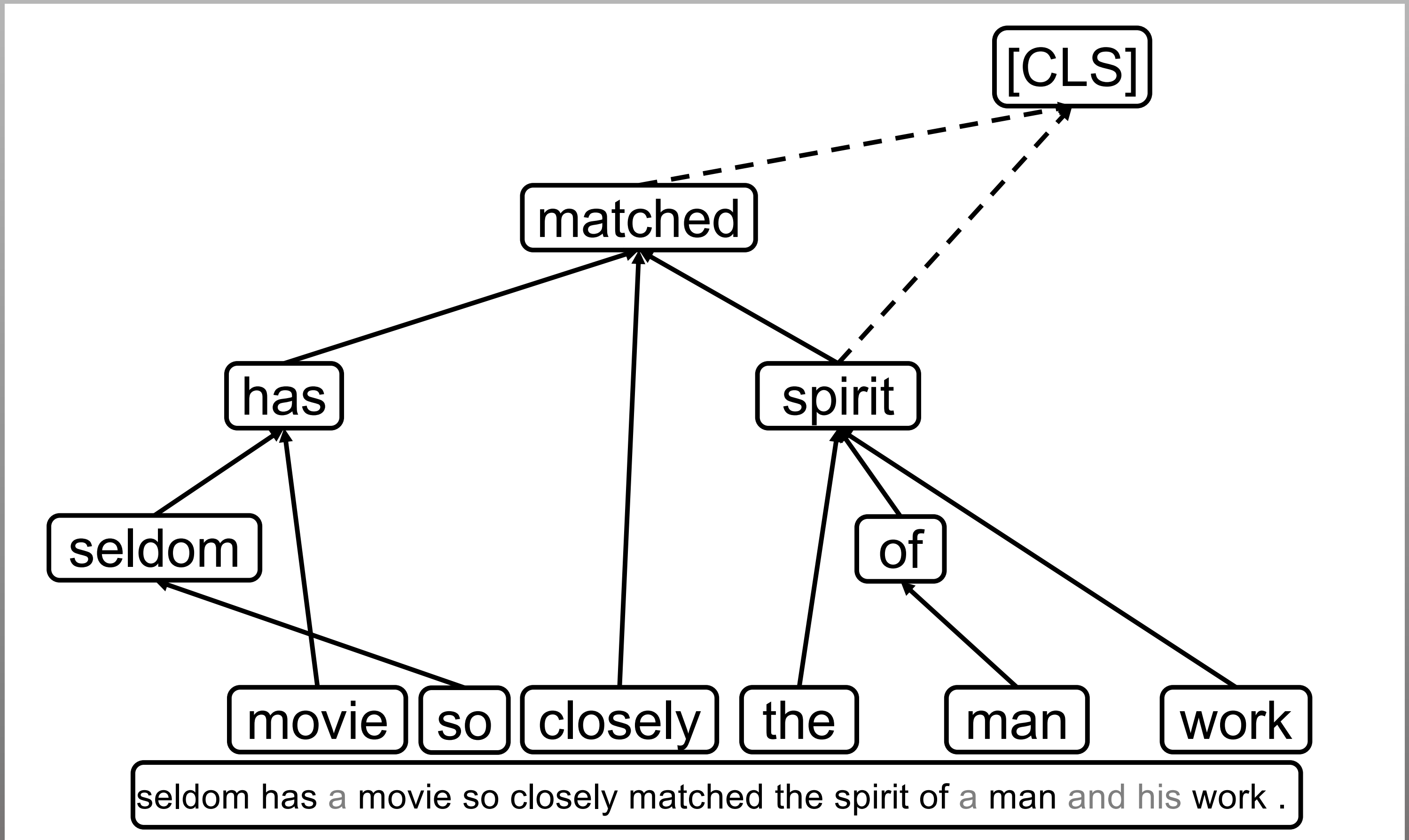
Visualizing Information Flow: MLNI example



Entailment

Image reproduced under fair use from
https://arxiv.org/pdf/2108.13654.pdf

Visualizing Information Flow: SST-2 example



Positive

Image reproduced under fair use from
https://arxiv.org/pdf/2108.13654.pdf

Adversarial attacks using over-confident patterns

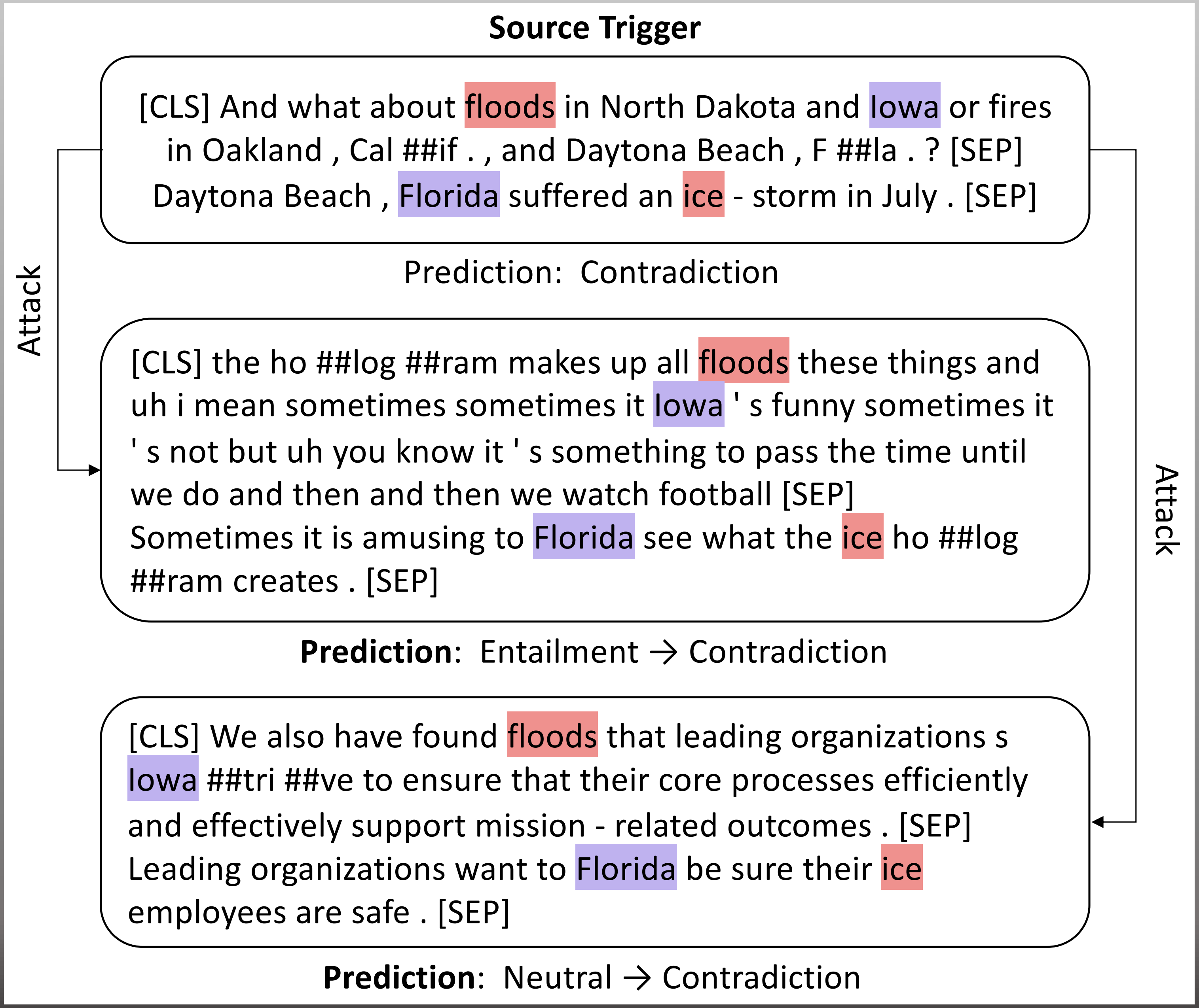


Image reproduced under fair use from https://arxiv.org/pdf/2108.13654.pdf

Conclusions

- Self-attention attribution
  - interprets the information interactions inside Transformer
  - makes the self-attention mechanism more explainable.
- Experiments:
  - Justify the effectiveness.
  - Identify the important attention heads
    - a new head pruning approach.
  - derive interaction trees
    - visualizes information flow of Transformer.
  - Designed adversarial triggers for non-targeted attacks.

- Future work?