Self-Attention Attribution: Interpreting Information Interactions Inside Transformer

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

Transformer

- Pack word embeddings of an input token into a matrix $X_0$
- The stacked L-layer Transformer computes the final output via
  $$X_l = \text{Transformer}(X_{l-1}), \quad l \in [1, L]$$
- The core component of a Transformer block is a multi-head self-attention. The $h$-th self-attention head is described as:

$$Q_h = XW_h^Q, \quad K = XW_h^K, \quad V = XW_h^V$$

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

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

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

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

\[
A_h = \text{softmax} \left( \frac{Q_h K_h^T}{\sqrt{d_k}} \right)
\]

\[
H_h = \text{AttentionHead}(X) = A_h V_h
\]
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 $x$
  • $F(\bar{A})$

\[
A_h = \text{softmax}\left( \frac{Q_h K_h^i}{\sqrt{d_k}} \right)
\]

\[
H_h = \text{AttentionHead}(X) = A_h V_h
\]
Attribution score matrix

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

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

  - 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 $Attr_h(A)$
  - denotes interaction between input $x_i$ and $x_j$ for the $h$-th attention head.
Attribution score matrix - II

• $\alpha = 0$:
  • represents that all tokens do not attend to each other in a layer.
• $\alpha = 1$:
  • 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.

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

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

\[
\text{Attr}_h(A) = A_h \odot \int_{\alpha=0}^{1} \frac{\partial F(\alpha A)}{\partial A_h} d\alpha,
\]

- Approximated using the Reimann approximation of the integration:

\[
\tilde{\text{Attr}}_h(A) = \frac{A_h}{m} \odot \sum_{k=1}^{m} \frac{\partial F(\frac{k}{m} A)}{\partial A_h}
\]

- \(m=20\) performs well in practice
Attribution Score Matrix: Motivating Example

contradiction class

Experiments: Design

- BERT-base-cased (Devlin et al. 2019)
  - BERT layers $|l| = 12,$
  - attention heads in each layer $|h| = 12,$
  - size of hidden embeddings $|h| \cdot dv = 768.$
- 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.

Experiments: Attention Head Pruning

• Importance of attention head:

\[ I_h = E_x[\max(\text{Attr}_{h}(A))] \]

• where

  • x represents the examples sampled from the held-out set,
  • \( \max(\text{Attr}_{h}(A)) \) 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

\[ I_h = E_x \left| A_h^T \frac{\partial \mathcal{L}(x)}{\partial A_h} \right| \]

where \( \mathcal{L}(x) \) is the loss function of example \( x \).
Experiment: Attention Head Pruning II

Important heads similar for similar tasks
Visualizing information flow inside transformer

• Attribution for the l\textsuperscript{th} layer:

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

• larger \(a_{i,j}^l\) implies more interaction between \(x_i\) and \(x_j\)
  - in the l-th layer
  - in terms of the final predictions.

• Attribution tree: a tradeoff between size and accuracy

\[
\text{Tree} = \arg \max_{\{E^l\}_{l=1}^{|l|}} \sum_{l=1}^{|l|} \sum_{(i,j) \in E^l} a_{i,j}^l - \lambda \sum_{l=1}^{|l|} |E^l|,
\]

\[
E^l \subset \{(i,j) | \frac{a_{i,j}^l}{\max(\text{Attr}(A^l))} > \tau\}
\]

Here,
• \(|E_l|\) represents # edges in the l-th layer,
• \(\lambda\) is a trade-off weight,
• \(\tau\) is a threshold to filter interactions with large attribution scores.
Visualizing Information Flow: MLNI example

![Diagram of attribution tree](https://arxiv.org/pdf/2108.13654.pdf)

Entailment

Visualizing Information Flow: SST-2 example

Adversarial attacks using over-confident patterns

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?