More Identifiable yet Equally Performant Transformers for Text Classification

Rishabh Bhardwaj • Navonil Majumder • Soujanya Poria • Eduard Hovy

Key Takeaways

- ⊖ Attention-based interpretations of Transformer models can be faulty.
- ⊖ Identifiability issue: Attention weights are not unique to the Transformer's attention head. They can be made more identifiable!
  - ⊕ By decreasing the size of key vector.
  - ⊕ By increasing the size of value vector followed by head addition.

Motivation

Interpretability is an important aspect when it comes to **trust** a model's prediction.

- Famous Transformer architecture-based systems are ubiquitous.
- Multi-head self-attention is its basic building block.

Introduction

Identifiability of attention weights

"Identifiable if they can be uniquely determined from the head's output"
-On identifiability in transformers [Brunner et al., 2019]

Why is it important?

Explanations based on non-unique attention weights might be misleading.

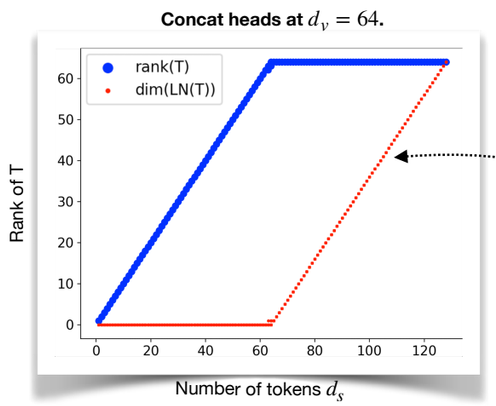
Attention Head

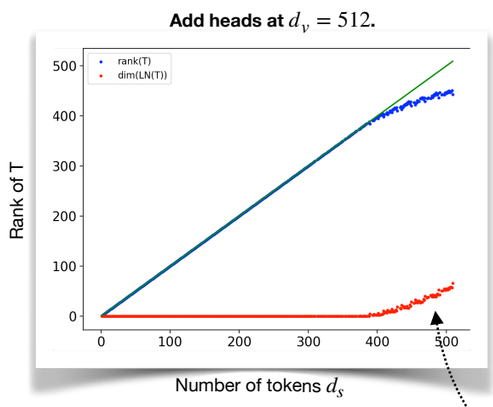
Attention: Query, Key

Phase:1 $\mathbf{A} = softmax\left(\frac{\mathbf{Q}\,\mathbf{K}^T}{\sqrt{d_q}}\right)$

Value, Weights

Phase:2 $\mathbf{H} = \mathbf{A}\,\mathbf{V}\,\mathbf{D} = \mathbf{A}\,\mathbf{T}$





Paper

Code

Uniqueness of A

Constraint-1	Constraint-2	Constraint-3
$\tilde{\mathbf{A}}\,\mathbf{T} = \mathbf{0}$	$\tilde{\mathbf{A}}\,\mathbf{1} = \mathbf{0}$	$(\mathbf{A} + \tilde{\mathbf{A}}) \geq \mathbf{0}$

Sample

LN (T)

Probability constraints

Role of key vector size

Let's say $\tilde{\mathbf{A}}$ satisfies above constraints.

$$softmax\left(\frac{\mathbf{Q}\,\mathbf{K}^T}{\sqrt{d_q}}\right) \overset{?}{\rightarrow} (\mathbf{A} + \tilde{\mathbf{A}})\,.$$ (Phase1 output)

Solutions

We propose the following solutions to identifiability.

- Decrease the size of key vector.
- Increase the size of value vector and add head outputs.

Classification performance of the model on ten text datasets does not vary significantly.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Carnegie Mellon University
DeCLaRe