More Identifiable yet Equally Performant Transformers for Text Classification

Rishabh Bhardwaj • Navonil Majumder • Soujanya Poria • Eduard Hovy

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Carnegie Mellon University

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.

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.

Motivation

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 them non-unique attention weights might be misleading.

Motivation

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: Phase1

Query matrix Key matrix

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

d_q: size of query and key vector

Row vectors in Query and Key matrices represent token transformations.

Attention Head: Phase1

Query matrix Key matrix

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

d_q: size of query and key vector

Row vectors in Query and Key matrices represents token transformations.

Dot product

Attention Head: Phase1

Query matrix

Key matrix

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

Attention matrix

d_q: size of query and key vector

Row vectors in Query and Key matrices represents token transformations.

Dot product

(i, j) element of Attention matrix $\mathbf{A}$ denotes how much of token **i** attends to token **j.**

Attention Head: Phase2

From phase1:

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

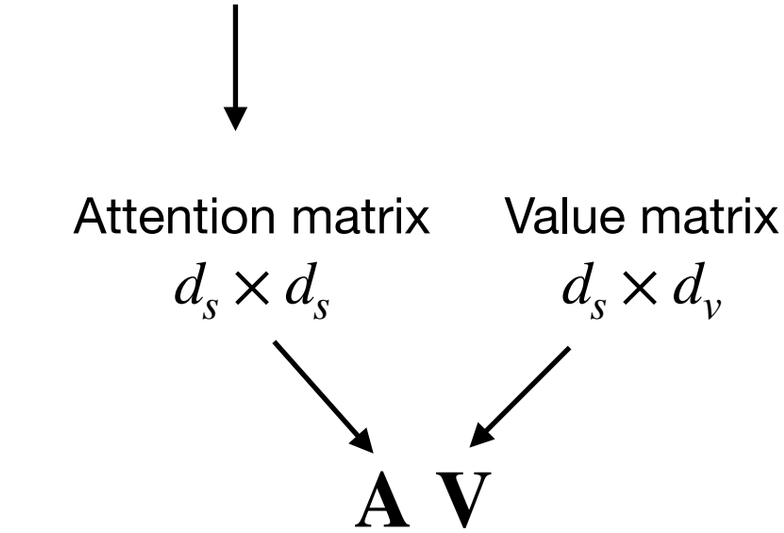
$\mathbf{Q}$ and $\mathbf{K}$ are $d_s \times d_k$ matrices.
$(d_q = d_k)$

Attention Head: Phase2

From phase1:

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

$\mathbf{Q}$ and $\mathbf{K}$ are $d_s \times d_k$ matrices.
($d_q = d_k$)



In phase2:

d_s: number of tokens

d_v: size of value vector.

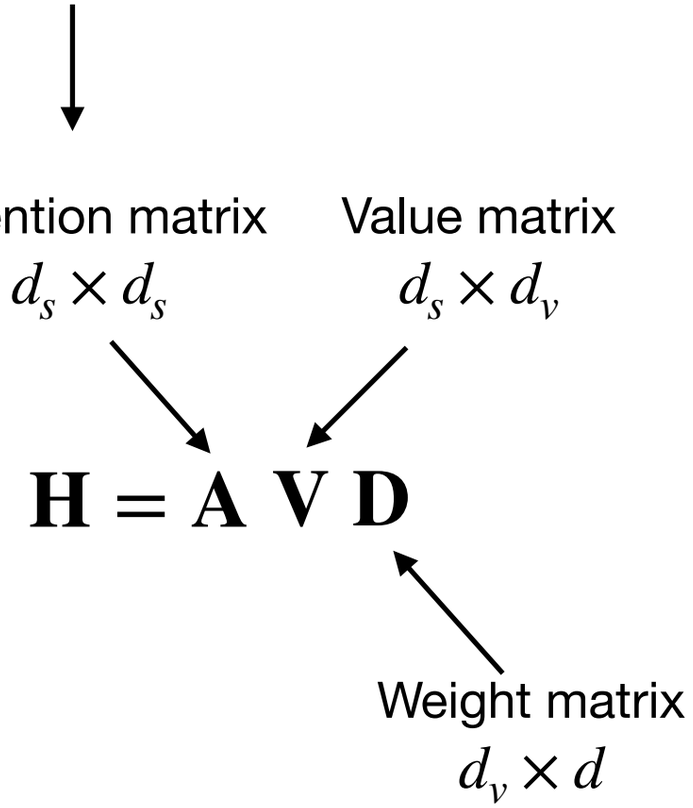
Attention Head: Phase2

From phase1:

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

$\mathbf{Q}$ and $\mathbf{K}$ are $d_s \times d_k$ matrices.
($d_q = d_k$)

In phase2:



d_s: number of tokens

d_v: size of value vector.

d : size of token vector (from embedding space)

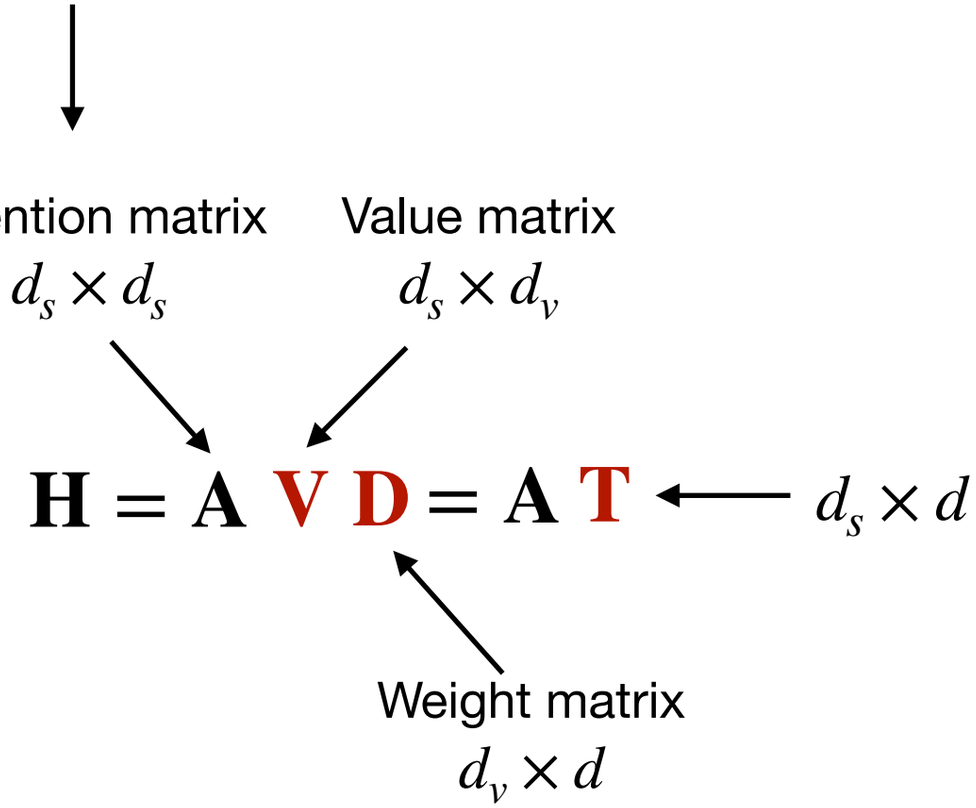
Attention Head: Phase2

From phase1:

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

$\mathbf{Q}$ and $\mathbf{K}$ are $d_s \times d_k$ matrices.
($d_q = d_k$)

In phase2:



d_s: number of tokens

d_v: size of value vector.

d : size of token vector (from embedding space)

Uniqueness of A

$$\mathbf{H} = \mathbf{A}\ \mathbf{T}$$

Is $\mathbf{A}$ unique to the combination $\mathbf{H}$ and $\mathbf{T}$?

Uniqueness of A

$$\mathbf{H} = \mathbf{A}\ \mathbf{T}$$

Is $\mathbf{A}$ unique to the combination $\mathbf{H}$ and $\mathbf{T}$?

Alternatively, can we find $\tilde{\mathbf{A}}$ such that:

Uniqueness of A

$$\mathbf{H} = \mathbf{A}\ \mathbf{T}$$

Is $\mathbf{A}$ unique to the combination $\mathbf{H}$ and $\mathbf{T}$?

Alternatively, can we find $\tilde{\mathbf{A}}$ such that:

Constraint-1

$$(\mathbf{A} + \tilde{\mathbf{A}})\ \mathbf{T} = \mathbf{A}\ \mathbf{T}$$
$$\Longrightarrow \tilde{\mathbf{A}}\ \mathbf{T} = \mathbf{0}$$

Uniqueness of A

$$\mathbf{H} = \mathbf{A}\ \mathbf{T}$$

Is $\mathbf{A}$ unique to the combination $\mathbf{H}$ and $\mathbf{T}$?

Alternatively, can we find $\tilde{\mathbf{A}}$ such that:

Constraint-1

$$(\mathbf{A} + \tilde{\mathbf{A}})\ \mathbf{T} = \mathbf{A}\ \mathbf{T}$$
$$\implies \tilde{\mathbf{A}}\ \mathbf{T} = \mathbf{0}$$

Constraint-2

$$(\mathbf{A} + \tilde{\mathbf{A}})\ \mathbf{1} = \mathbf{1}$$
$$\implies \tilde{\mathbf{A}}\ \mathbf{1} = \mathbf{0}$$

Constraint-3

$$(\mathbf{A} + \tilde{\mathbf{A}}) \geq \mathbf{0}$$

probability constraints

Uniqueness of A

Alternatively, can we find $\tilde{\mathbf{A}}$ such that:

Constraint-1

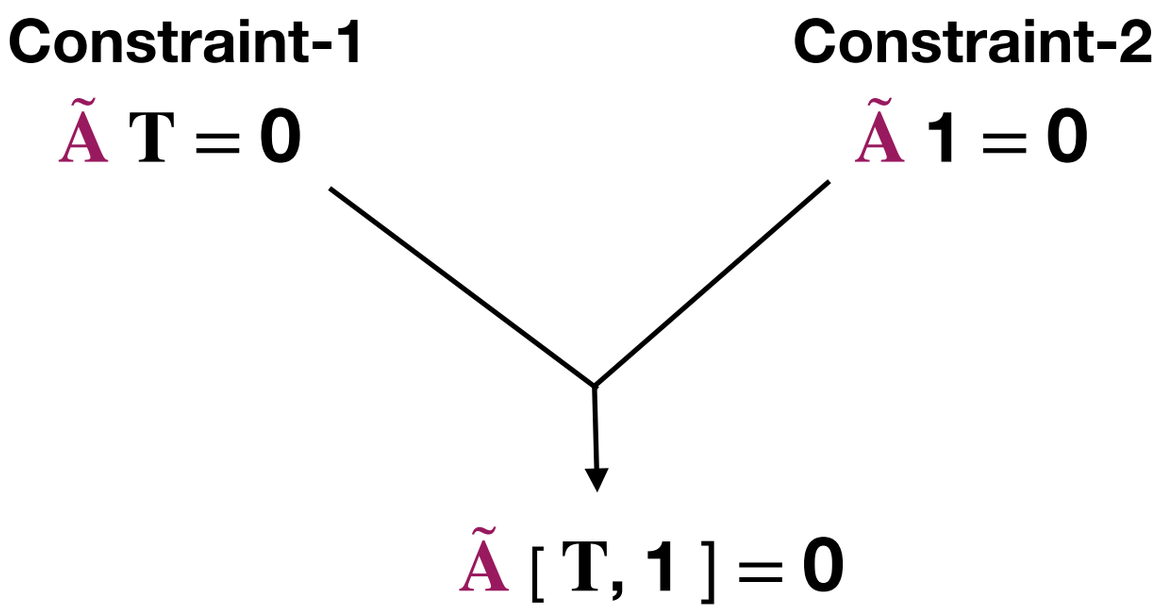
$$\tilde{\mathbf{A}}\,\mathbf{T} = \mathbf{0}$$

Constraint-2

$$\tilde{\mathbf{A}}\,\mathbf{1} = \mathbf{0}$$

Uniqueness of A

Alternatively, Can we find $\tilde{\mathbf{A}}$ such that:



Uniqueness of A

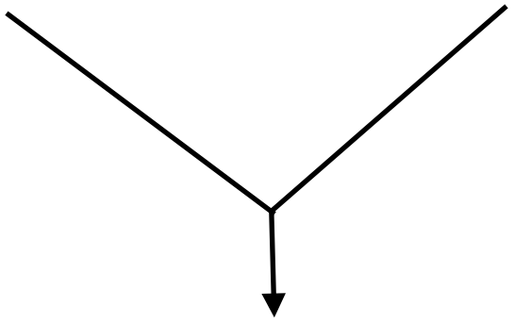
Alternatively, Can we find $\tilde{\mathbf{A}}$ such that:

Constraint-1

$$\tilde{\mathbf{A}}\ \mathbf{T} = \mathbf{0}$$

Constraint-2

$$\tilde{\mathbf{A}}\ \mathbf{1} = \mathbf{0}$$



$$\tilde{\mathbf{A}}\ [\ \mathbf{T}, \mathbf{1}\ ] = \mathbf{0}$$

$$\Longrightarrow \tilde{\mathbf{a}}\ [\ \mathbf{T}, \mathbf{1}\ ] = \mathbf{0}$$

($\tilde{\mathbf{a}}$ is a row of $\tilde{\mathbf{A}}$)

Uniqueness of A

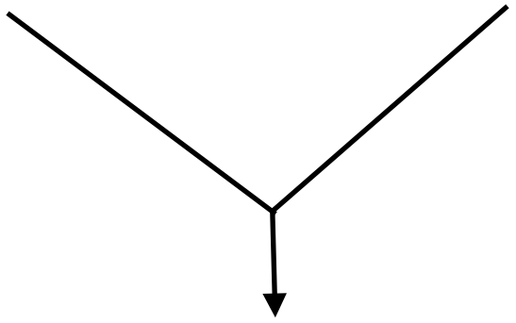
Alternatively, Can we find $\tilde{\mathbf{A}}$ such that:

Constraint-1

$$\tilde{\mathbf{A}}\ \mathbf{T} = \mathbf{0}$$

Constraint-2

$$\tilde{\mathbf{A}}\ \mathbf{1} = \mathbf{0}$$



$$\tilde{\mathbf{A}}\ [\ \mathbf{T}, \mathbf{1}\ ] = \mathbf{0}$$

$$\Longrightarrow \tilde{\mathbf{a}}\ [\ \mathbf{T}, \mathbf{1}\ ] = \mathbf{0} \qquad (\tilde{\mathbf{a}} \text{ is a row of } \tilde{\mathbf{A}})$$

$\tilde{\mathbf{a}}$ lies in the left null space of $[\ \mathbf{T}, \mathbf{1}\ ]$, i.e., LN($[\ \mathbf{T}, \mathbf{1}\ ]$). (for $\tilde{\mathbf{a}} \neq \mathbf{0}$)

Uniqueness of A

dim LN([$\mathbf{T}$, **1**]) = number of rows in [$\mathbf{T}$, **1**] – rank([$\mathbf{T}$, **1**])

Uniqueness of A

dim LN([$\mathbf{T}$, **1**]) = number of rows in [$\mathbf{T}$, **1**] – rank([$\mathbf{T}$, **1**])

= d_s – rank([$\mathbf{T}$, **1**])

Uniqueness of A

dim LN([$\mathbf{T}$, **1**]) = number of rows in [$\mathbf{T}$, **1**] – rank([$\mathbf{T}$, **1**])

= d_s – rank([$\mathbf{T}$, **1**]) $\leq \min(d_s, d_v, d) + 1$

$\leq d_v + 1$

Uniqueness of A

dim LN([$\mathbf{T}$, **1**]) = number of rows in [$\mathbf{T}$, **1**] – rank([$\mathbf{T}$, **1**])

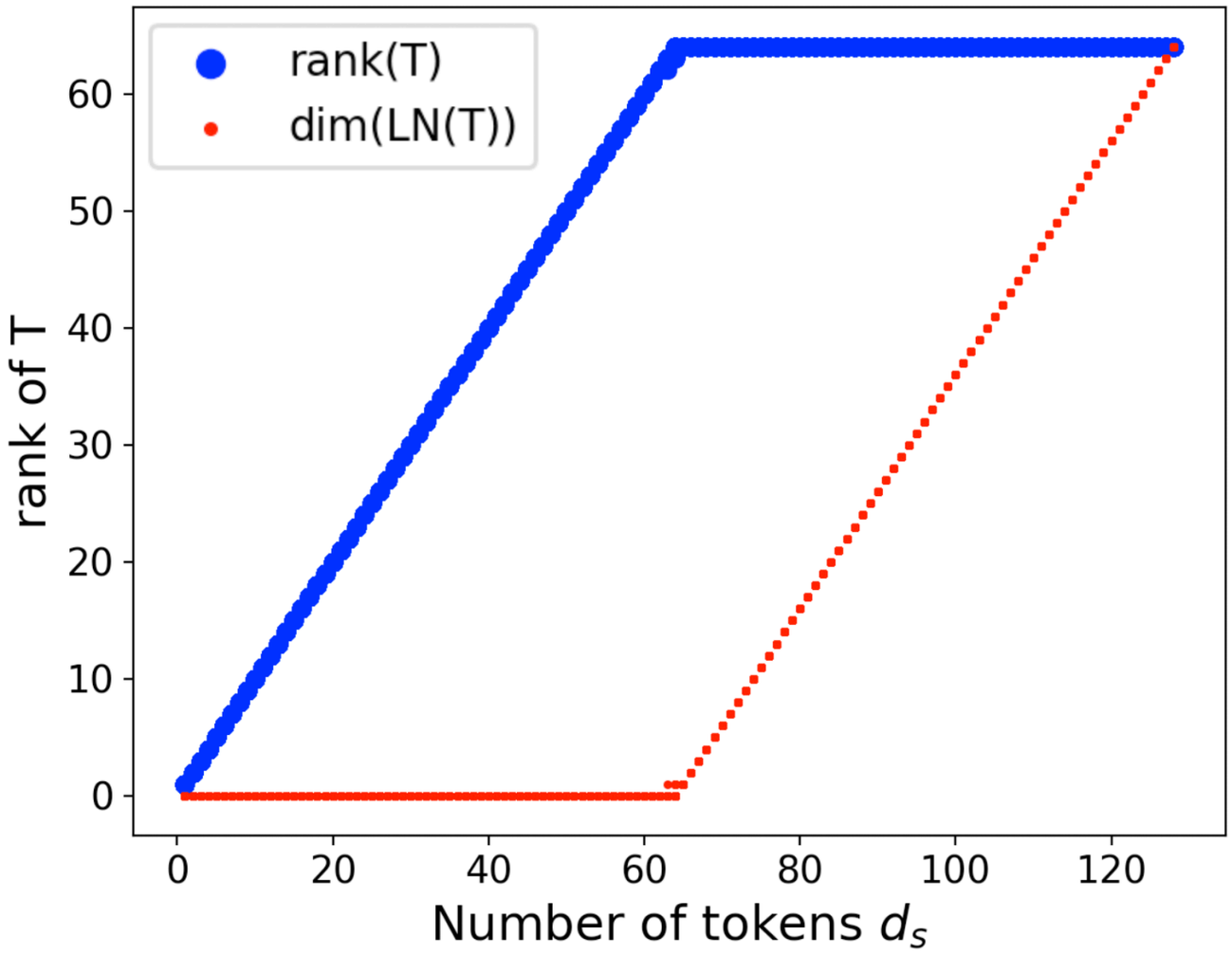
= d_s – rank([$\mathbf{T}$, **1**])

$\leq \min(d_s, d_v, d) + 1$

$\leq d_v + 1$

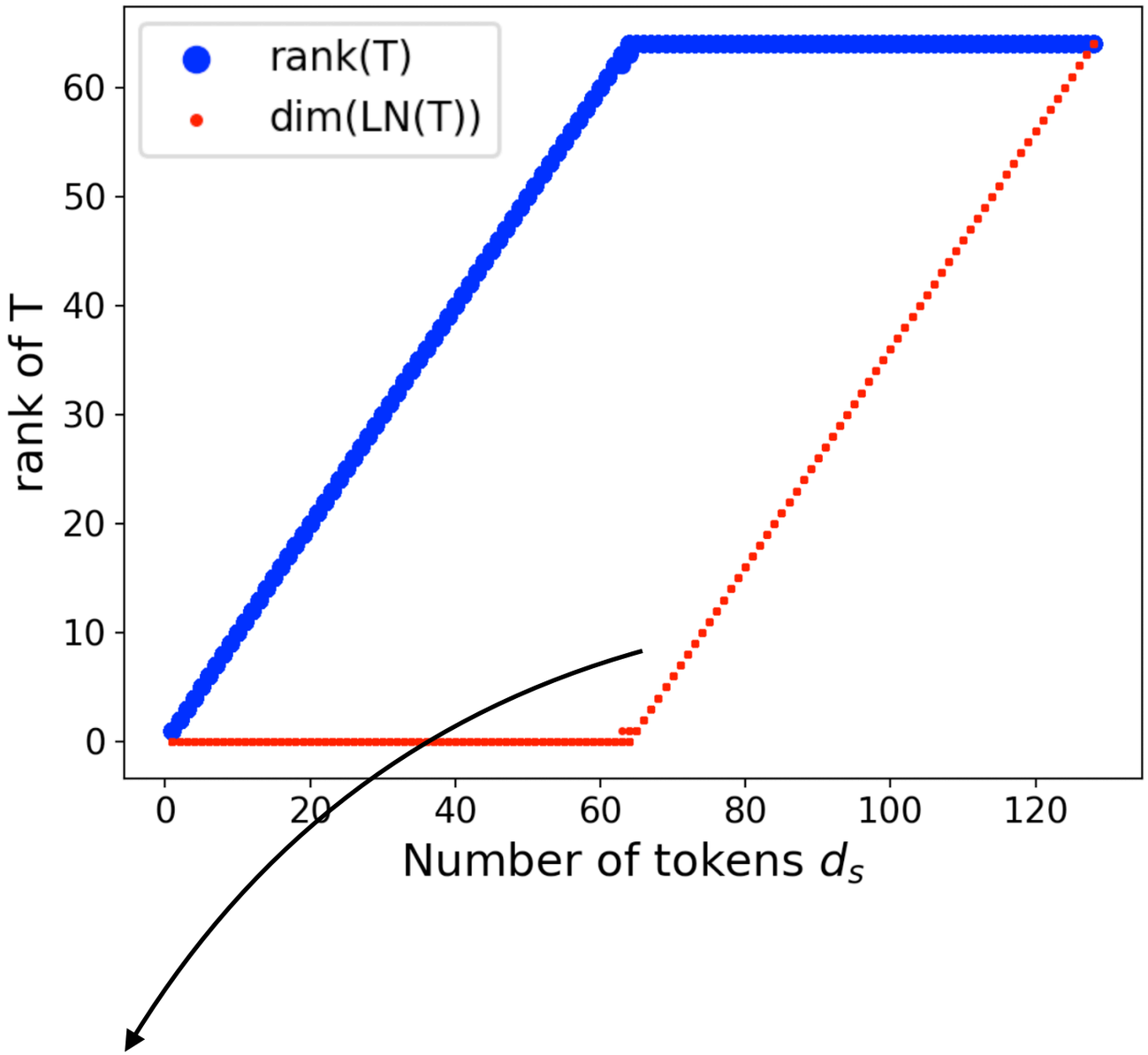
$\geq d_s - d_v - 1$

Uniqueness of A



Numerical rank of $\mathbf{T}$ and dim (LN($\mathbf{T}$)) on IMDB dataset.

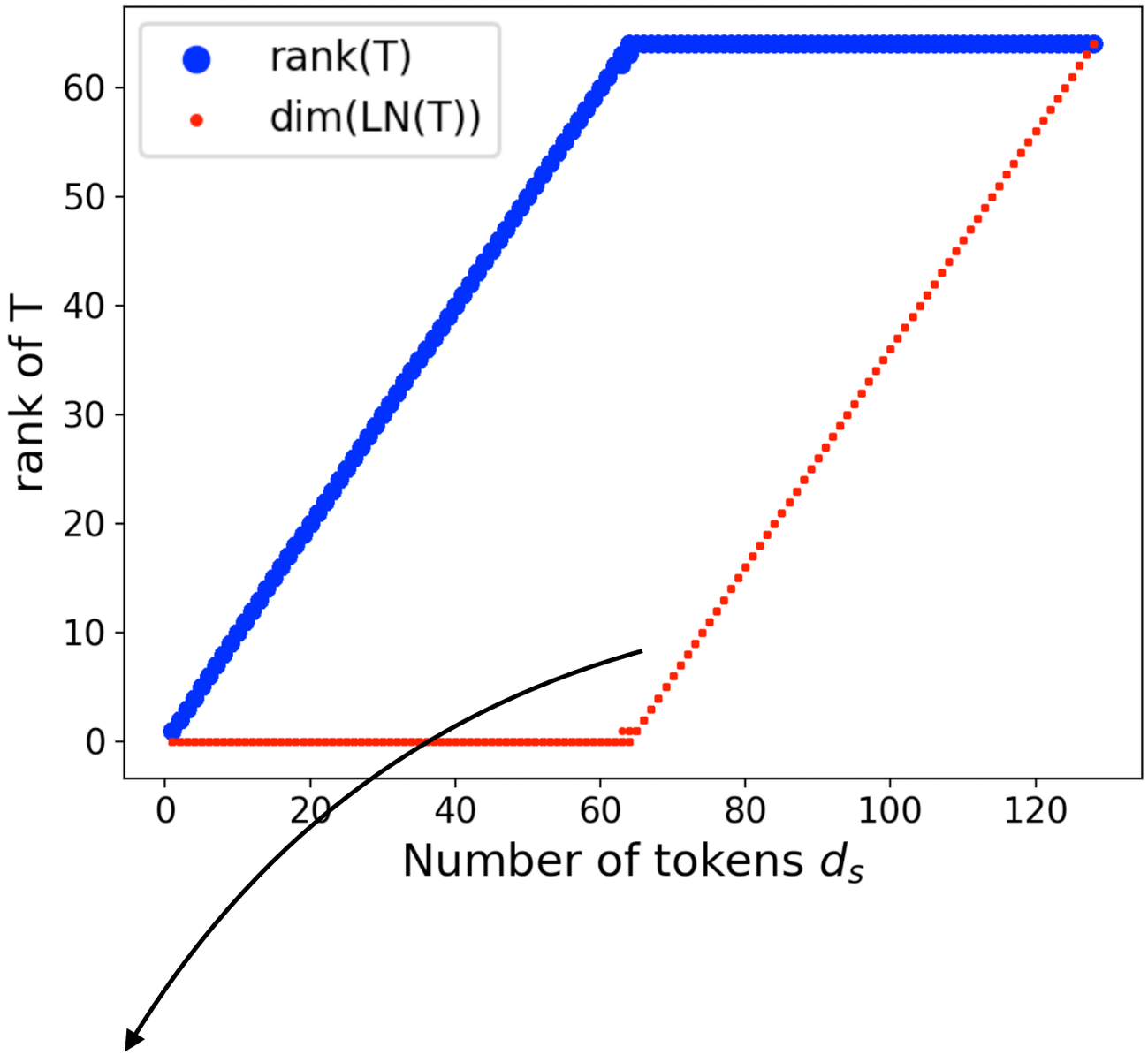
Uniqueness of A



$\tilde{\mathbf{a}}$ is obtainable satisfying all the Constraints 1,2, and 3 $\Longrightarrow$ $\tilde{\mathbf{A}}$ is not unique

-On identifiability in transformers [Brunner et al., 2019]

Uniqueness of A



$\tilde{\mathbf{a}}$ is obtainable satisfying all the Constraints 1,2, and 3 $\Longrightarrow$ $\tilde{\mathbf{A}}$ is not unique

On identifiability in transformers [Brunner et al., 2019]

We show that this may not be a valid claim!

Uniqueness of A

(Role of key vector size)

Uniqueness of A

(Role of key vector size)

Given: $\mathbf{H} = \mathbf{A}\ \mathbf{T}$

Uniqueness of A

(Role of key vector size)

Given: $\mathbf{H} = \mathbf{A}\,\mathbf{T}$

↓

Constraints 1, 2 and 3

↓ for $d_s > d_v + 1$

$$\mathbf{H} = (\mathbf{A} + \tilde{\mathbf{A}})\,\mathbf{T}$$

Uniqueness of A

(Role of key vector size)

Given: $\mathbf{H} = \mathbf{A}\,\mathbf{T}$

↓

Constraints 1, 2 and 3

↓ for $d_s > d_v + 1$

$\mathbf{H} = (\mathbf{A} + \tilde{\mathbf{A}})\,\mathbf{T}$

Can I obtain this attention matrix from the first phase?

$$\text{softmax}\left(\frac{\mathbf{Q}\,\mathbf{K}^T}{\sqrt{d_q}}\right)$$

Uniqueness of A

(Role of key vector size)

Given: $\mathbf{H} = \mathbf{A}\,\mathbf{T}$

↓

Constraints 1, 2 and 3

↓ for $d_s > d_v + 1$

$\mathbf{H} = (\mathbf{A} + \tilde{\mathbf{A}})\,\mathbf{T}$

↑

softmax (log($\mathbf{A} + \tilde{\mathbf{A}}$) + $\mathbf{C}$)

C: value that is constant across a row.

Uniqueness of A

(Role of key vector size)

Given: $\mathbf{H} = \mathbf{A}\,\mathbf{T}$

↓

Constraints 1, 2 and 3

↓ for $d_s > d_v + 1$

$\mathbf{H} = (\mathbf{A} + \tilde{\mathbf{A}})\,\mathbf{T}$

↑

softmax (log($\mathbf{A} + \tilde{\mathbf{A}}$) + $\mathbf{C}$)

↑

$\mathbf{A}_l \left(\frac{\mathbf{Q}\,\mathbf{K}^T}{\sqrt{d_q}} \right)$

Uniqueness of A

(Role of key vector size)

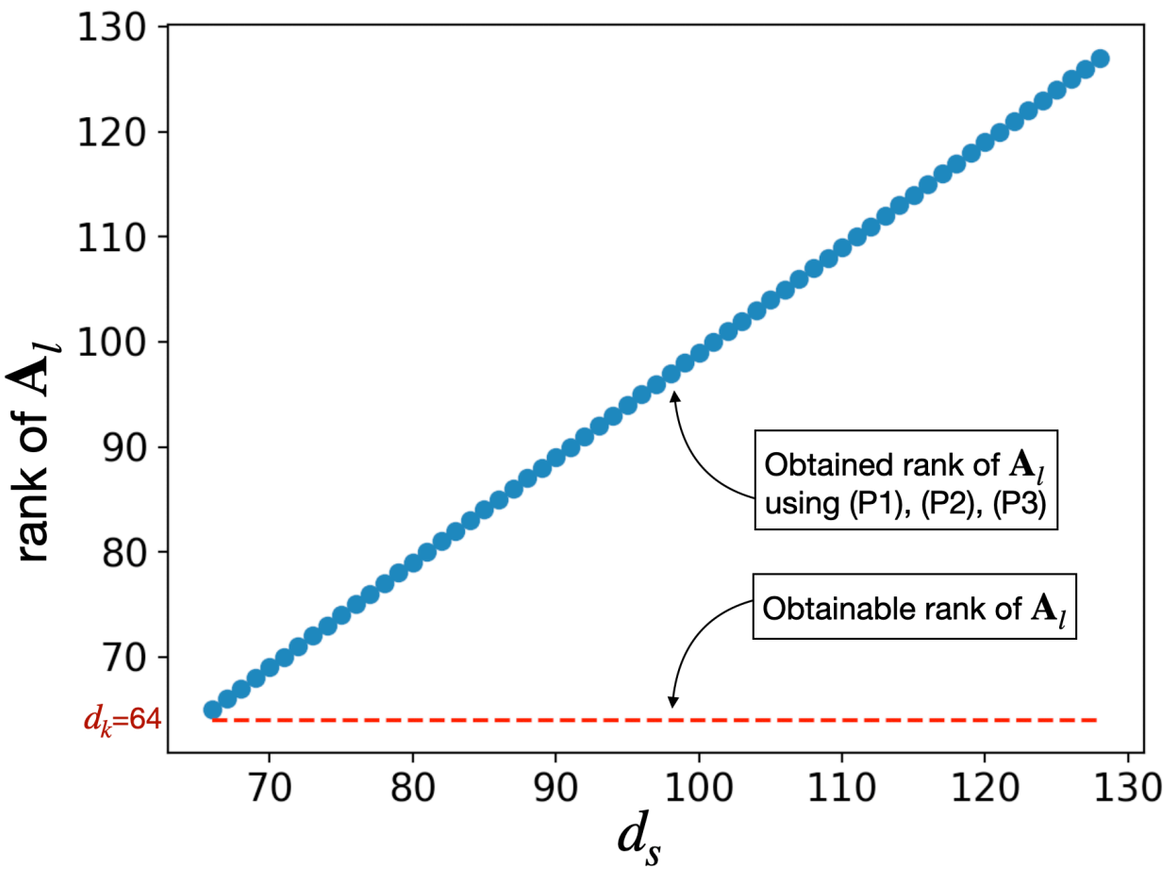
Given: $\mathbf{H} = \mathbf{A}\,\mathbf{T}$

Constraints 1, 2 and 3

$\mathbf{H} = (\mathbf{A} + \tilde{\mathbf{A}})\,\mathbf{T}$

softmax (log$(\mathbf{A} + \tilde{\mathbf{A}})$ + $\mathbf{C}$)

$\mathbf{A}_l$ $\left(\frac{\mathbf{Q}\,\mathbf{K}^T}{\sqrt{d_q}}\right)$



obtainable rank $(\mathbf{A}_l) \leq \min(\, d_s\,,\, d_k)$

$\leq d_k$

Uniqueness of A

(Role of key vector size)

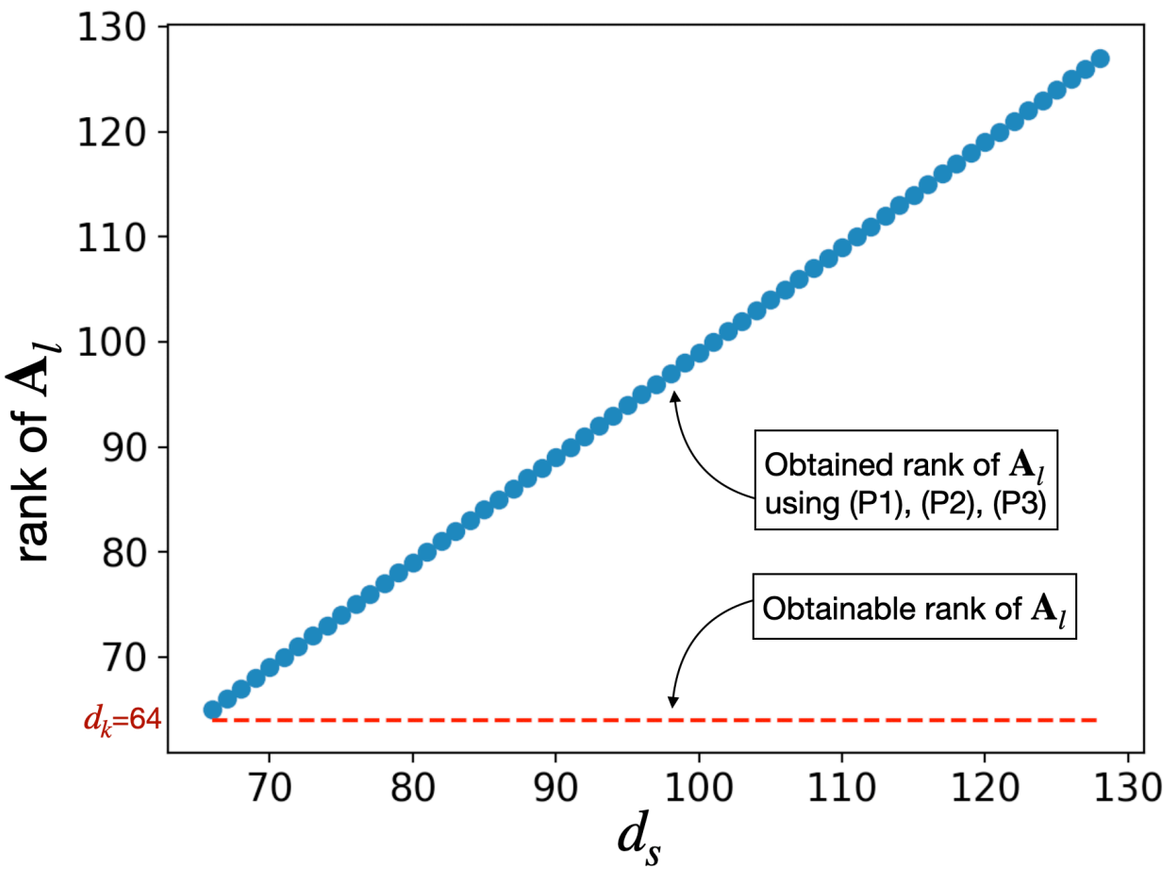
Given: $\mathbf{H} = \mathbf{A}\,\mathbf{T}$

Constraints 1, 2 and 3

$\mathbf{H} = (\mathbf{A} + \tilde{\mathbf{A}})\,\mathbf{T}$

softmax (log$(\mathbf{A} + \tilde{\mathbf{A}})$ + $\mathbf{C}$)

$\mathbf{A}_l \left(\frac{\mathbf{Q}\,\mathbf{K}^T}{\sqrt{d_q}} \right)$



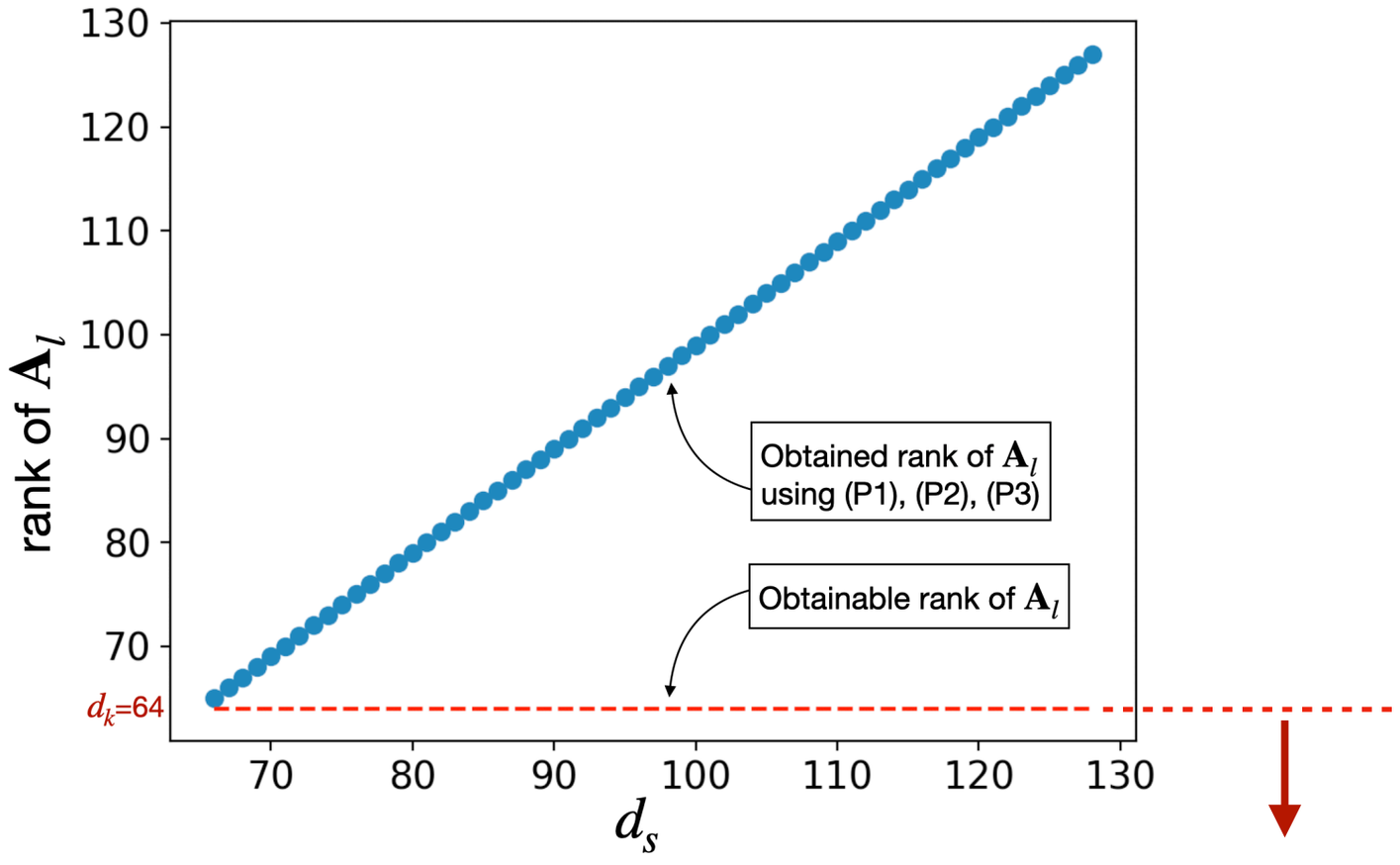
obtainable rank $(\mathbf{A}_l) \leq \min(\, d_s\,,\, d_k)$

$\leq d_k$

Hence, solutions satisfying constraints 1,2, and 3 may not be valid claims for A's non-uniqueness!

More identifiable

(Solution1: decrease $\mathbf{d_k}$)



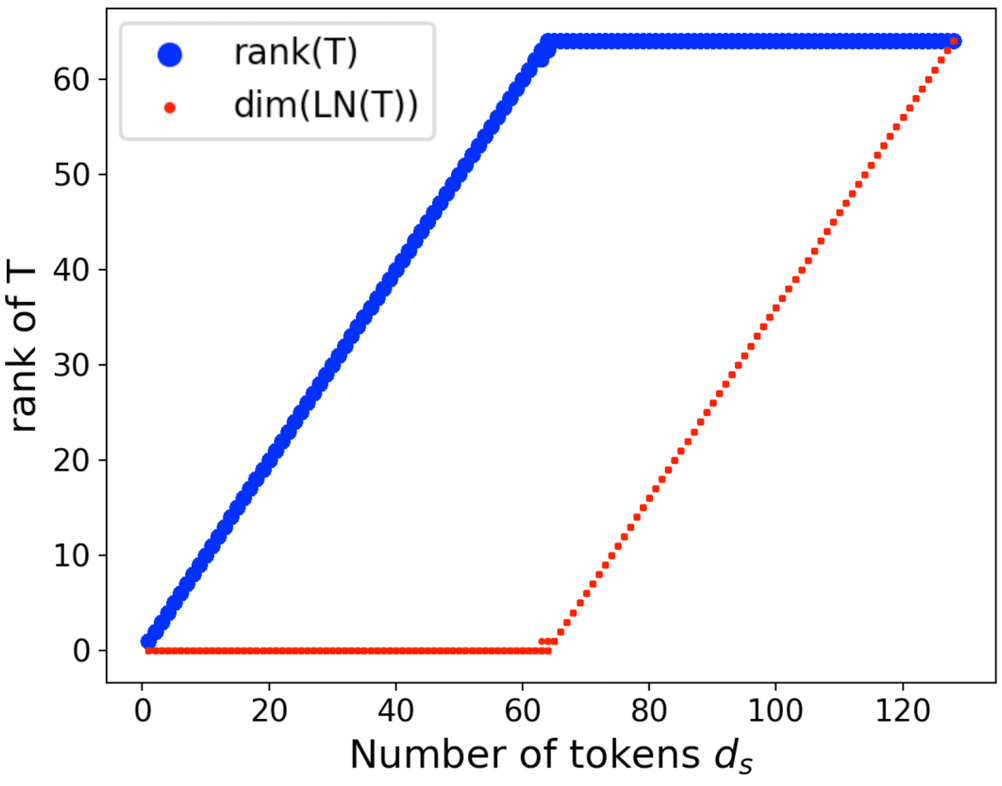
Push d_k to lower values to reduce the number of alternate attention weights ($\mathbf{A} + \tilde{\mathbf{A}}$)

More identifiable

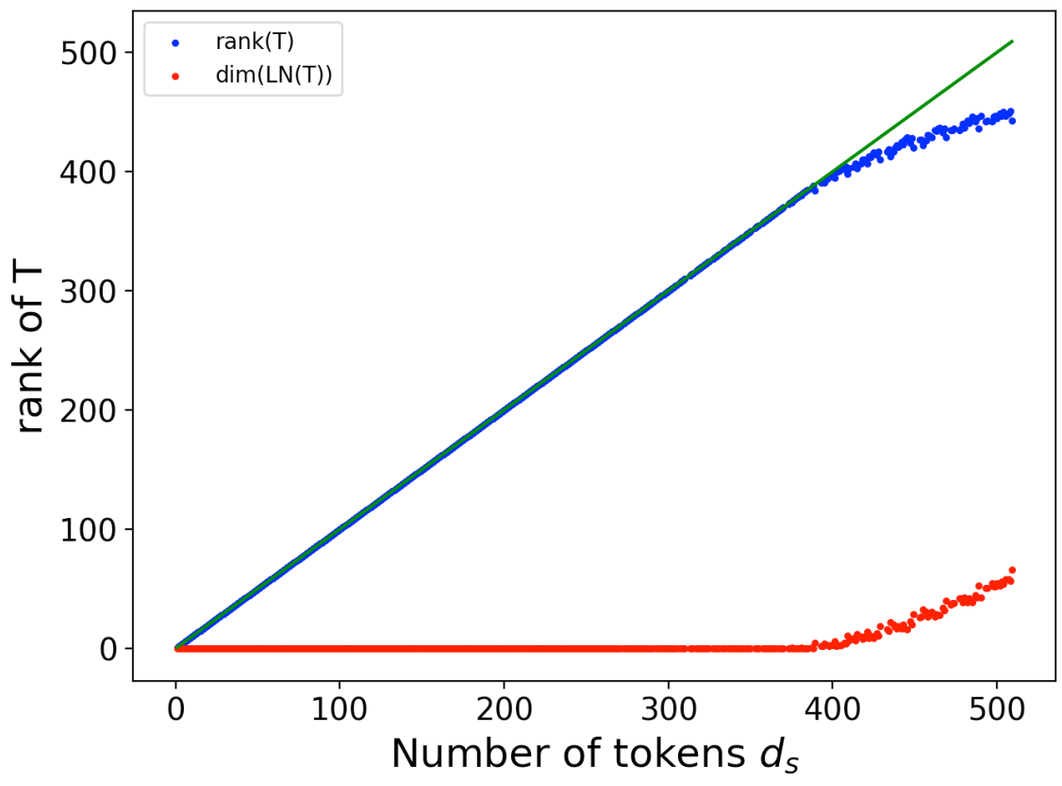
(Solution2: increase $\mathbf{d_v}$)

As $\tilde{\mathbf{A}}$ belongs to LN([$\mathbf{T}$, $\mathbf{1}$])

dim LN([$\mathbf{T}$, $\mathbf{1}$]) = d_s – rank([$\mathbf{T}$, $\mathbf{1}$])



$d_v = 64$



$d_v = 512$

Increase d_v and substitute concatenation of attention heads with addition.

Text-classification

Dataset	Version	Size of key vector (d_k)								
		1	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
IMDB	*Con*	0.884	0.888	0.886	0.888	0.846	0.824	0.803	0.788	0.755
	Add	0.888	0.885	0.887	0.884	0.886	0.882	0.877	0.832	0.825
TREC	*Con*	0.836	0.836	0.840	0.822	0.823	0.764	0.786	0.706	0.737
	Add	0.841	0.842	0.835	0.842	0.841	0.836	0.809	0.809	0.771
SST	*Con*	0.643	0.625	0.627	0.609	0.603	0.582	0.574	0.573	0.554
	Add	0.599	0.618	0.628	0.633	0.628	0.629	0.592	0.581	0.586
SNLI	*Con*	0.675	0.674	0.673	0.672	0.662	0.659	0.659	0.655	0.648
	Add	0.683	0.677	0.674	0.676	0.673	0.669	0.663	0.664	0.655
Yelp	*Con*	0.913	0.911	0.907	0.898	0.879	0.862	0.857	0.849	0.837
	Add	0.914	0.915	0.916	0.914	0.915	0.916	0.910	0.909	0.891
DBPedia	*Con*	0.979	0.977	0.977	0.971	0.966	0.961	0.957	0.951	0.949
	Add	0.979	0.978	0.979	0.977	0.978	0.973	0.970	0.969	0.964
Sogou	*Con*	0.915	0.907	0.898	0.900	0.893	0.888	0.868	0.858	0.838
	Add	0.915	0.908	0.906	0.904	0.913	0.914	0.910	0.906	0.899
AG News	*Con*	0.906	0.903	0.904	0.904	0.886	0.877	0.870	0.870	0.869
	Add	0.902	0.908	0.907	0.906	0.897	0.899	0.901	0.897	0.893
Yahoo	*Con*	0.695	0.690	0.684	0.664	0.644	0.627	0.616	0.597	0.574
	Add	0.697	0.695	0.696	0.693	0.693	0.694	0.688	0.649	0.683
Amazon	*Con*	0.924	0.925	0.923	0.922	0.900	0.892	0.887	0.882	0.873
	Add	0.925	0.923	0.925	0.924	0.924	0.920	0.907	0.896	0.889

The test accuracy on varied text classification datasets spread over ten datasets.

Conclusion

- We analysed attention weights in multi-head attention for their uniqueness.
- We claimed that the existing attention identifiability analysis is not complete.
- We provided a more concrete analysis of attention weights identifiability.

- We provide solutions to make attention weights more identifiable.
- The text classification performance does not vary significantly.

Thank you!

PAPER: https://arxiv.org/pdf/2106.01269.pdf
CODE: https://github.com/declare-lab/identifiable-transformers