cases

doc_1		doc_2		decision	id
authors	• Gui, Lin • He, Yulan • Li, Wenjie • Yan, Hanqi	**authors**	• Hanqi Yan • Lin Gui • Wenjie Li • Yulan He	DUPLICATES	158
title	Addressing token uniformity in transformers via singular value transformation	**title**	Addressing Token Uniformity in Transformers via Singular Value Transformation		
publication_date	2022-08-01 00:00:00	**publication_date**	2022-08-24 00:00:00		
source	SupportedSources.CORE	**source**	SupportedSources.SEMANTIC_SCHOLAR		
journal		**journal**	ArXiv		
volume		**volume**	abs/2208.11790		
doi	None	**doi**	10.48550/arXiv.2208.11790		
urls	• https://core.ac.uk/download/533429434.pdf	**urls**	• https://www.semanticscholar.org/paper/4cf7889c0fc5e181c20e64a4b26cb08ce25e7b45		
id	id390369942618153980	**id**	id-1418156902891134280		
abstract	Token uniformity is commonly observed in transformer-based models, in which different tokens share a large proportion of similar information after going through stacked multiple self-attention layers in a transformer. In this paper, we propose to use the distribution of singular values of outputs of each transformer layer to characterise the phenomenon of token uniformity and empirically illustrate that a less skewed singular value distribution can alleviate the token uniformity problem. Base on our observations, we define several desirable properties of singular value distributions and propose a novel transformation function for updating the singular values. We show that apart from alleviating token uniformity, the transformation function should preserve the local neighbourhood structure in the original embedding space. Our proposed singular value transformation function is applied to a range of transformer-based language models such as BERT, ALBERT, RoBERTa and DistilBERT, and improved performance is observed in semantic textual similarity evaluation and a range of GLUE tasks	**abstract**	Token uniformity is commonly observed in transformer-based models, in which different tokens share a large proportion of similar information after going through stacked multiple self-attention layers in a transformer. In this paper, we propose to use the distribution of singular values of outputs of each transformer layer to characterise the phenomenon of token uniformity and empirically illustrate that a less skewed singular value distribution can alleviate the â€˜token uniformityâ€™ problem. Base on our observations, we define several desirable properties of singular value distributions and propose a novel transformation function for updating the singular values. We show that apart from alleviating token uniformity, the transformation function should preserve the local neighbourhood structure in the original embedding space. Our proposed singular value transformation function is applied to a range of transformer-based language models such as BERT, ALBERT, RoBERTa and DistilBERT, and improved performance is observed in semantic textual similarity evaluation and a range of GLUE tasks. Our source code is available at https://github.com/hanqi-qi/tokenUni.git .		
versions		**versions**			