cases	doc_1		doc_2		decision	id
	authors	• Hanqi Yan • Lin Gui • Wenjie Li • Yulan He	**authors**	• Gui, Lin • He, Yulan • Li, Wenjie • Yan, Hanqi	DUPLICATES	157
	title	Addressing Token Uniformity in Transformers via Singular Value Transformation	**title**	Addressing token uniformity in transformers via singular value transformation		
	publication_date	2022-08-24 22:44:09+00:00	**publication_date**	2022-08-01 00:00:00		
	source	SupportedSources.ARXIV	**source**	SupportedSources.CORE		
	journal	None	**journal**			
	volume		**volume**			
	doi		**doi**	None		
	urls	• http://arxiv.org/pdf/2208.11790v1 • http://arxiv.org/abs/2208.11790v1 • http://arxiv.org/pdf/2208.11790v1	**urls**	• https://core.ac.uk/download/533429434.pdf		
	id	id5998116240696226138	**id**	id390369942618153980		
	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. Our source code is available at https://github.com/hanqi-qi/tokenUni.git.	**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		
	versions		**versions**			