exists Part Proceedings Processing			doc_1		doc_2	decision	id
Rui Cao authors Publication date Publication				authors	 Rui Cao Jie Zheng Jie Ren 		
sates SupportedSources.ARXIV		authors	Y. Liang	title	Learning to Remove: Towards Isotropic Pre-trained BERT Embedding		
** Jie Ren ** Ling Gao ** Ling	- 11 11111			publication_date	e 2021-04-12 08:13:59+00:00		
title Learning to Remove: Towards Isotropic Pre-trained BERT Embedding publication_date 2021-04-12 00:00:00 source SupportedSources.SEMANTIC_SCHOLAR				source	SupportedSources.ARXIV		
title Learning to Remove: Towards Isotropic Pre-trained BERT Embedding publication_date 2021-04-12 00:00:00 source SupportedSources.SEMANTIC_SCHOLAR journal volume doi 10.1007/978-3-030-86383-8_36 urls • https://www.semanticscholar.org/paper/ab151c1ca0479b677003cf200018b93e983aa0cc id id-6514770989049821401 Pre-trained language models such as BERT have become a more common choice of natural language processing (NLP) tasks. Research in word representation shows that isotropic embeddings can significantly improve performance on downstream tasks. However, we measure and analyze the geometry of pre-trained BERT embedding and find that it is far from isotropic. We find that the word vectors are distributed in a narrow cone and deteriorate the representation and the average cosine similarity between two random words is much higher than zero, which indicates that the vord vectors are distributed in a narrow cone and deteriorate the representation are not centured around the origin, and the average cosine similarity tasks and show that processed embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and senantic textual similarity, in all tasks, the word embedding processed by method on the estandardized tasks: word similarity, word analogy, and semantic textual similarity in all tasks, the word embedding processed by methods. Our method is also proven to be more robust to changes of hyperparameter.				journal	None		
title Learning to Kemove: Towards Isotropic Pre-trained BERT Embedding publication_date 2021-04-12 00:00:00 source SupportedSources.SEMANTIC_SCHOLAR journal volume doi 10.1007/978-3-030-86383-8_36 urls http://arxiv.org/pdf/2104.05274v2 http://a			Ling Gao				
source SupportedSources.SEMANTIC_SCHOLAR volume doi 10.1007/978-3-030-86383-8_36 urls https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec versions versions volume doi 10.908774381427119780 abstract None versions v		title	Learning to Remove: Towards Isotropic Pre-trained BERT Embedding	doi			
Supported Sources. SEMANTE_SCHOLAR							
Journal volume id id-6514770989049821401		source	SupportedSources.SEMANTIC_SCHOLAR	urls		DUPLICATES 2	220
doi 10.1007/978-3-030-86383-8_36		journal			• http://arxiv.org/pdi/2104.032/4v2		
urls • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None versions • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None versions • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None versions • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None versions • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None versions • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None versions • https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec id id-9088774381427119780 abstract None oversions abstract Star form isotropic. We find that the word vectors are not centered around the origin, and the average cosine similarity between two random words is much higher than zero, which indicates that the word vectors are distributed in a narrow cone and deteriorate the representation capacity of word embedding. We propose a simple, and yet effective method to fix this problem: remove several dominant directions of BERT embedding with a set of learnable weights. We train the weights on word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method is also proven to be more robust to changes of hyperparameter.		volume		id	id-6514770989049821401		
berformance on downstream tasks. However, we measure and analyze the geometry of pre-trained BERT embedding and find that it is far from isotropic. We find that the word vectors are not centered around the origin, and the average cosine similarity between two random words is much representation capacity of word embedding. We propose a simple, and yet effective method to fix this problem: remove several dominant directions of BERT embedding with a set of learnable weights. We train the weights on word similarity, tasks and show that processed embedding is more isotropic. Our method is evaluated on three standardied tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of hyperparameter.		doi	10.1007/978-3-030-86383-8_36				
abstract None versions		urls	https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec		performance on downstream tasks. However, we measure and analyze the geometry of pre-trained BERT		
abstract None versions word vectors are distributed in a narrow cone and deteriorate the representation capacity of word embedding. We propose a simple, and yet effective method to fix this problem: remove several dominant directions of BERT embedding with a set of learnable weights. We train the weights on word similarity tasks and show that processed embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of hyperparameter.		id	id-9088774381427119780		and the average cosine similarity between two random words is much higher than zero, which indicates that the word vectors are distributed in a narrow cone and deteriorate the representation capacity of word embedding. We propose a simple, and yet effective method to fix this problem: remove several dominant directions of BERT		
embedding with a set of learnable weights. We train the weights on word similarity tasks and show that processed embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of hyperparameter.		abstract	None				
embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of hyperparameter.		versions					
versions					embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of		