	doc_1		doc_2		decision	id
cases	**authors**	- Lee, Jaejun - Tang, Raphael - Lin, Jimmy	**authors**	- Jaejun Lee - Raphael Tang - Jimmy Lin	DUPLICATES	339
	title	What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning	**title**	What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning		
	publication_date	2019-11-08 00:00:00	**publication_date**	2019-11-08 07:05:20+00:00		
	source	SupportedSources.OPENALEX	**source**	SupportedSources.ARXIV		
	journal	arXiv (Cornell University)	**journal**	None		
	volume		**volume**			
	doi	10.48550/arxiv.1911.03090	**doi**			
	urls	- https://openalex.org/W4288026527 - https://doi.org/10.48550/arxiv.1911.03090 - http://arxiv.org/pdf/1911.03090	**urls**	- http://arxiv.org/pdf/1911.03090v1 - http://arxiv.org/abs/1911.03090v1 - http://arxiv.org/pdf/1911.03090v1		
	id	id5285092138517909115	**id**	id-2980461387407644797		
	abstract		**abstract**	Pretrained transformer-based language models have achieved state of the art across countless tasks in natural language processing. These models are highly expressive, comprising at least a hundred million parameters and a dozen layers. Recent evidence suggests that only a few of the final layers need to be fine-tuned for high quality on downstream tasks. Naturally, a subsequent research question is, "how many of the last layers do we need to fine-tune?" In this paper, we precisely answer this question. We examine two recent pretrained language models, BERT and RoBERTa, across standard tasks in textual entailment, semantic similarity, sentiment analysis, and linguistic acceptability. We vary the number of final layers that are fine-tuned, then study the resulting change in task-specific effectiveness. We show that only a fourth of the final layers need to be fine-tuned to achieve 90% of the original quality. Surprisingly, we also find that fine-tuning all layers does not always help.		
	versions		**versions**			