cases	doc_1		doc_2		decision	id
	authors	• Feiqi Cao • Soyeon Caren Han • Siqu Long • Changwei Xu • Josiah Poon	**authors**	• Cao, Feiqi • Han, Soyeon Caren • Long, Siqu • Poon, Josiah • Xu, Changwei	DUPLICATES	118
	title	Understanding Attention for Vision-and-Language Tasks	**title**	Understanding Attention for Vision-and-Language Tasks		
	publication_date	2022-08-17 06:45:07+00:00	**publication_date**	2022-09-22 00:00:00		
	source	SupportedSources.ARXIV	**source**	SupportedSources.CORE		
	journal	None	**journal**			
	volume		**volume**			
	doi		**doi**	None		
	urls	• http://arxiv.org/pdf/2208.08104v2 • http://arxiv.org/abs/2208.08104v2 • http://arxiv.org/pdf/2208.08104v2	**urls**	• http://arxiv.org/abs/2208.08104		
	id	id-9174578776798991135	**id**	id-3269586404956237650		
	abstract	Attention mechanism has been used as an important component across Vision-and-Language(VL) tasks in order to bridge the semantic gap between visual and textual features. While attention has been widely used in VL tasks, it has not been examined the capability of different attention alignment calculation in bridging the semantic gap between visual and textual clues. In this research, we conduct a comprehensive analysis on understanding the role of attention alignment by looking into the attention score calculation methods and check how it actually represents the visual region's and textual token's significance for the global assessment. We also analyse the conditions which attention score calculation mechanism would be more (or less) interpretable, and which may impact the model performance on three different VL tasks, including visual question answering, text-to-image generation, text-and-image matching (both sentence and image retrieval). Our analysis is the first of its kind and provides useful insights of the importance of each attention alignment score calculation when applied at the training phase of VL tasks, commonly ignored in attention-based cross modal models, and/or pretrained models. Our code is available at: https://github.com/adlnlp/Attention_VL	**abstract**	Attention mechanism has been used as an important component across Vision-and-Language(VL) tasks in order to bridge the semantic gap between visual and textual features. While attention has been widely used in VL tasks, it has not been examined the capability of different attention alignment calculation in bridging the semantic gap between visual and textual clues. In this research, we conduct a comprehensive analysis on understanding the role of attention alignment by looking into the attention score calculation methods and check how it actually represents the visual region's and textual token's significance for the global assessment. We also analyse the conditions which attention score calculation mechanism would be more (or less) interpretable, and which may impact the model performance on three different VL tasks, including visual question answering, text-to-image generation, text-and-image matching (both sentence and image retrieval). Our analysis is the first of its kind and provides useful insights of the importance of each attention alignment score calculation when applied at the training phase of VL tasks, commonly ignored in attention-based cross modal models, and/or pretrained models. Our code is available at: https://github.com/adlnlp/Attention_VLComment: Accepted in COLING 202		
	versions		**versions**			