cases

doc_1		doc_2		decision	id
authors	• Wei Li • Can Gao • Guocheng Niu • Xinyan Xiao • Hao Liu • Jiachen Liu • Hua Wu • Haifeng Wang	**authors**	• Wei Li • Can Gao • Guocheng Niu • Xinyan Xiao • Hao Liu • Jiachen Liu • Hua Wu • Haifeng Wang	NOT DUPLICATES	392
title	UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning	**title**	UNIMO-2: End-to-End Unified Vision-Language Grounded Learning		
publication_date	2022-03-14 00:00:00	**publication_date**	2022-03-17 03:53:11+00:00		
source	SupportedSources.INTERNET_ARCHIVE	**source**	SupportedSources.ARXIV		
journal		**journal**	None		
volume		**volume**			
doi		**doi**			
urls	• https://web.archive.org/web/20220316053550/https://arxiv.org/pdf/2012.15409v4.pdf	**urls**	• http://arxiv.org/pdf/2203.09067v1 • http://arxiv.org/abs/2203.09067v1 • http://arxiv.org/pdf/2203.09067v1		
id	id506291278797470667	**id**	id7575082153226287351		
abstract	Existed pre-training methods either focus on single-modal tasks or multi-modal tasks, and cannot effectively adapt to each other. They can only utilize single-modal data (i.e. text or image) or limited multi-modal data (i.e. image-text pairs). In this work, we propose a unified-modal pre-training architecture, namely UNIMO, which can effectively adapt to both single-modal and multi-modal understanding and generation tasks. Large scale of free text corpus and image collections can be utilized to improve the capability of visual and textual understanding, and cross-modal contrastive learning (CMCL) is leveraged to align the textual and visual information into a unified semantic space over a corpus of image-text pairs. As the non-paired single-modal data is very rich, our model can utilize much larger scale of data to learn more generalizable representations. Moreover, the textual knowledge and visual knowledge can enhance each other in the unified semantic space. The experimental results show that UNIMO significantly improves the performance of several single-modal and multi-modal downstream tasks. Our code and pre-trained models are public at the UNIMO project page https://unimo-ptm.github.io/	**abstract**	Vision-Language Pre-training (VLP) has achieved impressive performance on various cross-modal downstream tasks. However, most existing methods can only learn from aligned image-caption data and rely heavily on expensive regional features, which greatly limits their scalability and performance. In this paper, we propose an end-to-end unified-modal pre-training framework, namely UNIMO-2, for joint learning on both aligned image-caption data and unaligned image-only and text-only corpus. We build a unified Transformer model to jointly learn visual representations, textual representations and semantic alignment between images and texts. In particular, we propose to conduct grounded learning on both images and texts via a sharing grounded space, which helps bridge unaligned images and texts, and align the visual and textual semantic spaces on different types of corpora. The experiments show that our grounded learning method can improve textual and visual semantic alignment for improving performance on various cross-modal tasks. Moreover, benefiting from effective joint modeling of different types of corpora, our model also achieves impressive performance on single-modal visual and textual tasks. Our code and models are public at the UNIMO project page https://unimo-ptm.github.io/.		
versions		**versions**			