BEiT-3 Vs. Previous State-of-the-Art Models

Source: Wang et al., 2022 | Table: 2023 AI Index Report

Category	Task	Dataset	Metric	Previous SOTA	Model of Previous SOTA	BEiT-3	Scale of Improvement
Vision	Semantic Segmentation	ADE20K	mIoU	61.40	FD-SwimV2	62.80	2.28%
Vision	Object Detection	COCO	AP	63.30	DINO	63.70	0.63%
Vision	Instance Segmentation	COCO	AP	54.70	Mask DINO	54.80	0.18%
Vision	Image Classification	ImageNet	Top-1 Accuracy	89.00	FD-CLIP	89.60	0.67%
Vision-Language	Visual Reasoning	NLVR	Accuracy	87.00	CoCA	92.60	6.44%
Vision-Language	Visual QA	VQAv2	VQA Accuracy	82.30	CoCA	84.00	2.07%
Vision-Language	Image Captioning	COCO	CIDEr	145.30	OFA	147.60	1.58%
Vision-Language	Finetuned Retrieval	COCO Flickr30K	R@1	72.50	Florence	76.00	4.83%
Vision-Language	Zero-Shot Retrieval	Flickr30K	R@1	86.50	CoCA	88.20	1.97%