画像-テキスト対データセットの一覧

Table 1: VLM 事前学習に使用される代表的な画像-テキスト対データセット

Dataset	**Year**	**Image-Text Pairs**	**Public**
SBU Caption	2011	1M	Yes
COCO Caption	2016	1.5M	Yes
Yahoo Flickr Creative Commons 100M	2016	100M	Yes
Visual Genome	2017	5.4M	Yes
Conceptual Captions (CC3M)	2018	3.3M	Yes
Localized Narratives (LN)	2020	0.87M	Yes
Conceptual 12M (CC12M)	2021	12M	Yes
Wikipedia-based Image Text (WIT)	2021	37.6M	Yes
Red Caps (RC)	2021	12M	Yes
LAION400M	2021	400M	Yes
LAION5B	2022	5B	Yes
WuKong	2022	100M	Yes
CLIP	2021	400M	No
ALIGN	2021	1.8B	No
FILIP	2021	300M	No
WebLI	2022	12B	No

Table 2-1: VLM 評価に使用される代表的な視覚認識データセット（画像分類）

Task	**Dataset**	**Year**	**Training**	**Testing**	**Evaluation Metric**
Image Classification	Caltech-101	2004	3,060	6,085	Mean Per Class
	CIFAR-10	2009	50,000	10,000	Accuracy
	ImageNet-1k	2009	1,281,167	50,000	Accuracy
	SUN397	2010	19,850	19,850	Accuracy
	SVHN	2011	73,257	26,032	Accuracy
	Stanford Cars	2013	8,144	8,041	Accuracy
	FGVC Aircraft	2013	6,667	3,333	Mean Per Class
	Food-101	2014	75,750	25,250	Accuracy
	EuroSAT	2017	10,000	5,000	Accuracy

Table 2-2: VLM 評価に使用される代表的な視覚認識データセット（画像分類以外）

Task	**Dataset**	**Year**	**Training**	**Testing**	**Evaluation Metric**
Image-Text Retrieval	Flickr30k	2014	31,783	–	Recall
	COCO Caption	2015	82,783	5,000	Recall
Action Recognition	UCF101	2012	9,537	1,794	Accuracy
	Kinetics700	2019	494,801	31,669	Mean (top1, top5)
	RareAct	2020	7,607	–	mWAP, mSAP
Object Detection	COCO 2014 Detection	2014	83,000	41,000	box mAP
	COCO 2017 Detection	2017	118,000	5,000	box mAP
	LVIS	2019	118,000	5,000	box mAP
	ODinW	2022	132,413	20,070	box mAP
Semantic Segmentation	PASCAL VOC 2012 Segmentation	2012	1,464	1,449	mIoU
	PASCAL Context	2014	4,998	5,105	mIoU
	Cityscapes	2016	2,975	500	mIoU
	ADE20k	2017	25,574	2,000	mIoU

Table 5: 画像分類タスクにおけるゼロショット予測に基づく VLM 事前学習手法の性能比較

Methods	Image enc.	Text enc.	Data Size	ImageNet	CIFAR-10	CIFAR-100	Food101	SUN397	Cars	Aircraft	DTD	Pets	Caltech101	Flowers102
CLIP	ViT-L/14	Transformer	400M	76.2	95.7	77.5	93.8	68.4	78.8	37.2	55.7	93.5	92.8	78.3
FILIP	ViT-L/14	Transformer	340M	77.1	95.7	75.3	92.2	73.1	70.8	60.2	60.7	92.0	93.0	90.1
ChineseCLIP	ViT-L/14	CNRoBERTa	200M	–	96.0	79.7	–	–	–	26.2	51.2	–	–	–
KELIP	ViT-B/32	Transformer	1.1B	62.6	91.5	68.6	79.5	–	75.4	–	51.2	–	–	–
COCA	ViT-G/14	–	4.8B	86.3	–	–	–	–	–	–	–	–	–	–

Table 8: 画像分類タスクにおける VLM 転移学習手法の性能比較

Methods	Image Encoder	Setup	ImageNet	Caltech101	Pets
Baseline [143]	ResNet-50	w/o Transfer	60.3	86.1	85.8
Baseline [10]	ViT-B/16	w/o Transfer	70.2	95.4	94.1
Baseline [10]	ViT-L/14	w/o Transfer	76.2	92.8	93.5
CoOp	ViT-B/16	Few-shot Sup.	71.9	93.7	94.5
CuPL	ViT-L/14	Few-shot Sup.	76.6	93.4	93.8
UPL	ResNet-50	Unsupervised	61.1	91.4	89.5
TPT	ViT-B/16	Unsupervised	69.0	94.2	87.8
Wise-FT	ViT-L/14	Supervised	87.1	–	–