VLM 끄적끄적 1

(거시적인 흐름 정리용)

1. VLM의 연구 방향

거시적 연구 흐름 (세 가지 방향)

- 1) Pretraining
- 2) Transfer Learning
  - e.g., prompt tuning, visual adaptation
- 3) Knowledge distillation

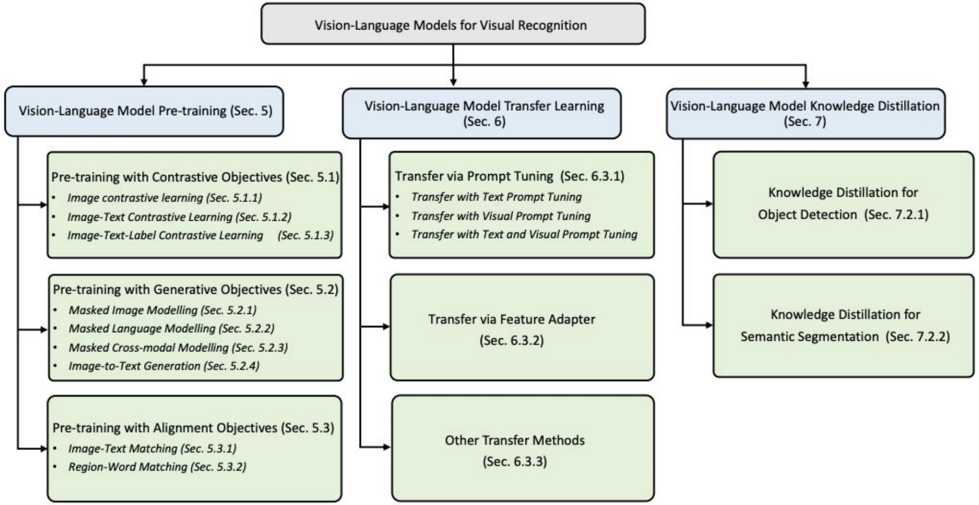

Vision-Language Models for Visual Recognition
Vision-Language Model Pre-training (Sec. 5)
Vision-Language Model Transfer Learning (Sec. 6)
Vision-Language Model Knowledge Distillation (Sec. 7)
Pre-training with Contrastive Objectives (Sec. 5.1)
• Image contrastive learning (Sec. 5.1.1)
• Image-Text Contrastive Learning (Sec. 5.1.2)
• Image-Text-Label Contrastive Learning (Sec. 5.1.3)
Pre-training with Generative Objectives (Sec. 5.2)
• Masked Image Modelling (Sec. 5.2.1)
• Masked Language Modelling (Sec. 5.2.2)
• Masked Cross-modal Modelling (Sec. 5.2.3)
• Image-to-Text Generation (Sec. 5.2.4)
Pre-training with Alignment Objectives (Sec. 5.3)
• Image-Text Matching (Sec. 5.3.1)
• Region-Word Matching (Sec. 5.3.2)
Transfer via Prompt Tuning (Sec. 6.3.1)
• Transfer with Text Prompt Tuning
• Transfer with Visual Prompt Tuning
• Transfer with Text and Visual Prompt Tuning
Transfer via Feature Adapter (Sec. 6.3.2)
Other Transfer Methods (Sec. 6.3.3)
Knowledge Distillation for Object Detection (Sec. 7.2.1)
Knowledge Distillation for Semantic Segmentation (Sec. 7.2.2)

Fig. 4: Typology of vision-language models for visual recognition.

VLM의 pipeline

- Pretraining & (Fine-tuning) & Zero-shot

2. VLM 향상 시키는 법

- 1) Pre-training objective 관점
    - (구) Single obj.
    - (신) Multiple obj.
- 2) Model 관점
    - (구) Two-tower
    - (신) One-tower
- 3) Downstream task 관점
    - (구) simple/coarse task
    - (신) complex/dense task

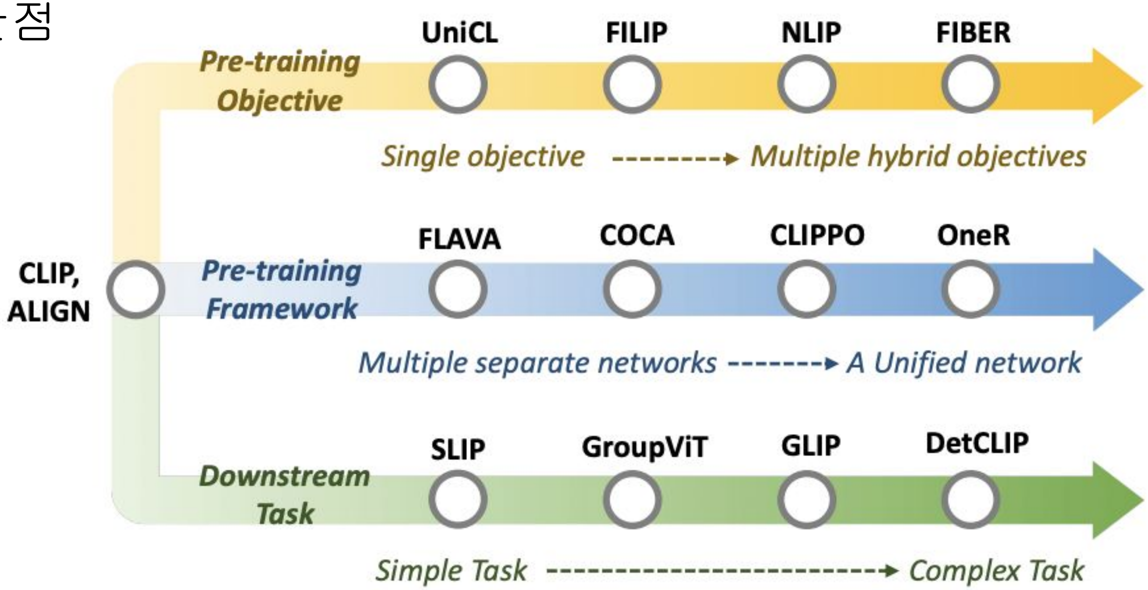

CLIP,
ALIGN
Pre-training Objective
UniCL
FILIP
NLIP
FIBER
Single objective --------→ Multiple hybrid objectives
Pre-training Framework
FLAVA
COCA
CLIPPO
OneR
Multiple separate networks -------→ A Unified network
Downstream Task
SLIP
GroupViT
GLIP
DetCLIP
Simple Task ------------------------→ Complex Task

Fig. 3: Illustration of development of VLMs for visual recognition.

3. VLM Foundation

구성요소

- (1) 데이터셋: 대량의 image-text pair로써 학습
- (2) 모델: Text & Image encoder
    - 2-1) Image: CNN-based (e.g., ResNet), Transformer-based (e.g., ViT)
    - 2-2) Text: Transformer-based
- (3) 목적함수: pre-training objective
    - Contrastive & Generative & Alignment
- (4) 평가: zero-shot task

4. VLM Pretraining Objective

1. Contrastive
   - Image CL & Image-Text CL & Image-Text-Label CL
2. Generative
   - MIM, MLM & MCM & Image-to-Text Generation

     (Masked Cross-Modal Modeling = MIM+MLM)
3. Alignment
   - Image-Text Matching (global)
   - Region-Word Matching (local)

5. VLM Pretraining Frameworks

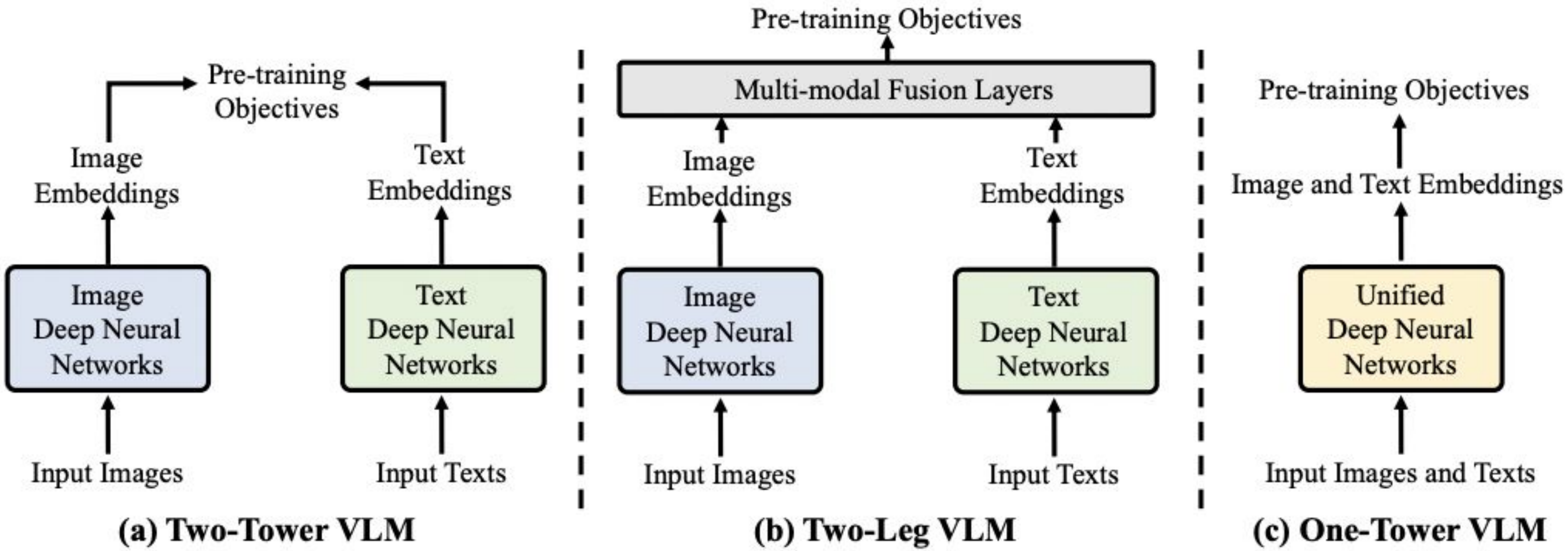


Fig. 5: Illustration of typical VLM pre-training frameworks.

6. VLM Evaluation

핵심: Zero-shot Prediction

- Image Classification
- Semantic Segmentation
- Object Detection
- Image-Text Retrieval

7. VLM Datasets

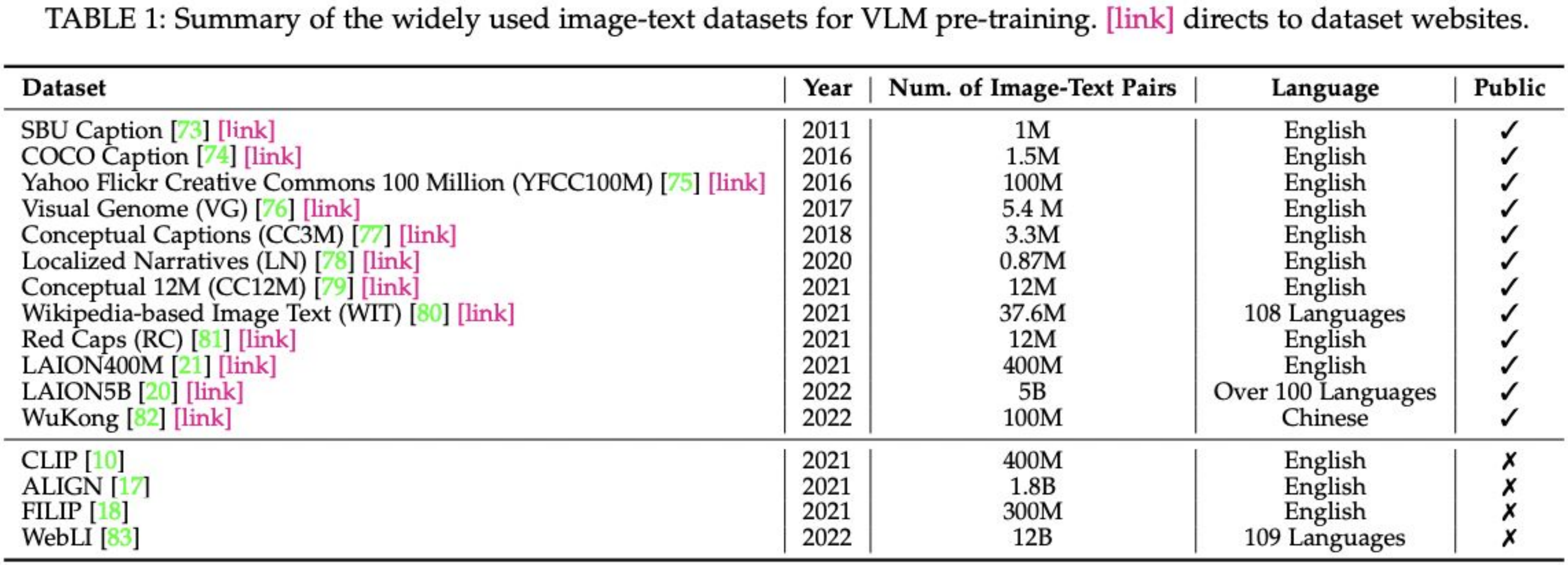
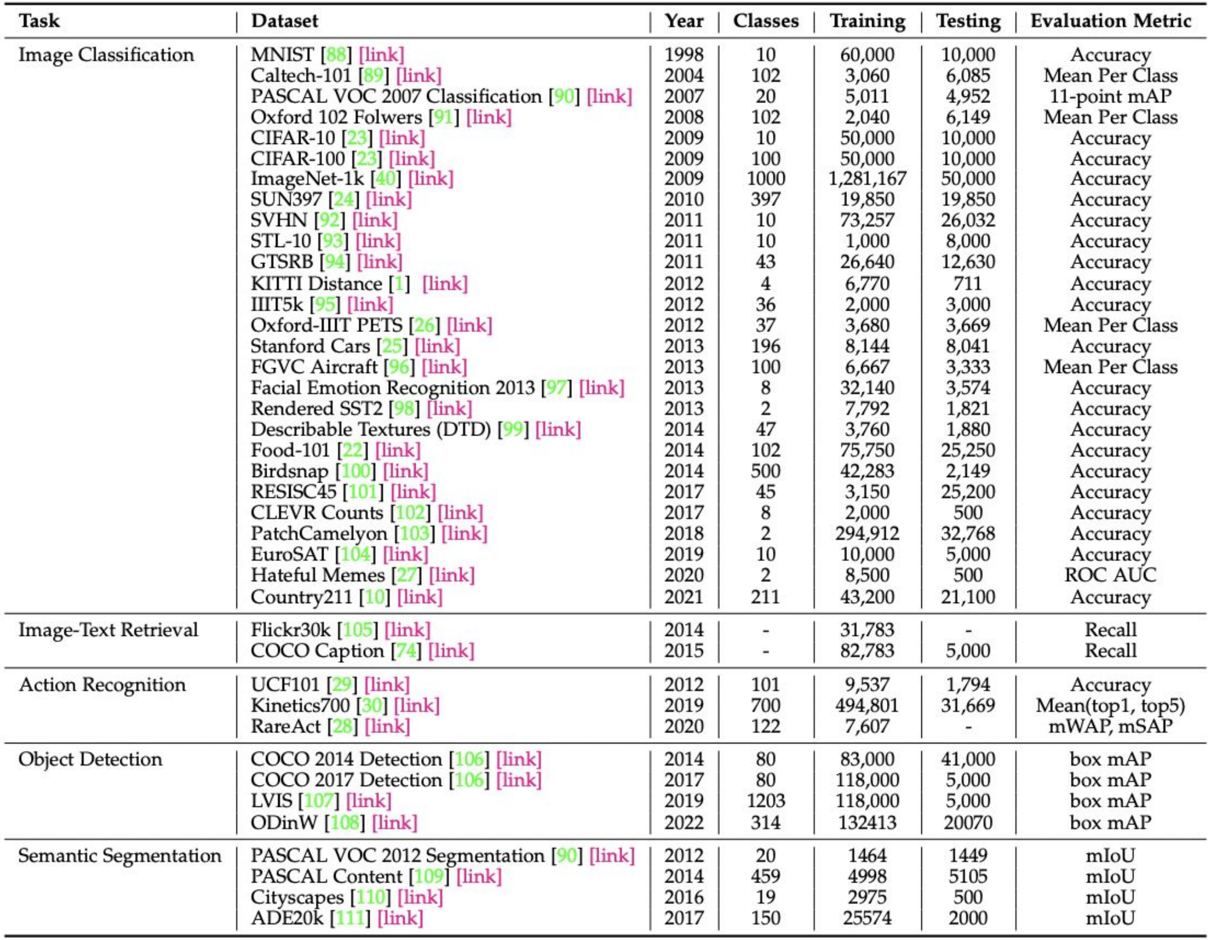


TABLE 1: Summary of the widely used image-text datasets for VLM pre-training. [link] directs to dataset websites.

Dataset	Year	Num. of Image-Text Pairs	Language	Public
SBU Caption [73] [link]	2011	1M	English	✓
COCO Caption [74] [link]	2016	1.5M	English	✓
Yahoo Flickr Creative Commons 100 Million (YFCC100M) [75] [link]	2016	100M	English	✓
Visual Genome (VG) [76] [link]	2017	5.4 M	English	✓
Conceptual Captions (CC3M) [77] [link]	2018	3.3M	English	✓
Localized Narratives (LN) [78] [link]	2020	0.87M	English	✓
Conceptual 12M (CC12M) [79] [link]	2021	12M	English	✓
Wikipedia-based Image Text (WIT) [80] [link]	2021	37.6M	108 Languages	✓
Red Caps (RC) [81] [link]	2021	12M	English	✓
LAION400M [21] [link]	2021	400M	English	✓
LAION5B [20] [link]	2022	5B	Over 100 Languages	✓
WuKong [82] [link]	2022	100M	Chinese	✓
CLIP [10]	2021	400M	English	✗
ALIGN [17]	2021	1.8B	English	✗
FILIP [18]	2021	300M	English	✗
WebLI [83]	2022	12B	109 Languages	✗

7. VLM Datasets

TABLE 2: Summary of the widely-used visual recognition datasets for VLM evaluation. [link] directs to dataset websites



Task	Dataset	Year	Classes	Training	Testing	Evaluation Metric
Image Classification	MNIST [88] [link]	1998	10	60,000	10,000	Accuracy
	Caltech-101 [89] [link]	2004	102	3,060	6,085	Mean Per Class
	PASCAL VOC 2007 Classification [90] [link]	2007	20	5,011	4,952	11-point mAP
	Oxford 102 Folwers [91] [link]	2008	102	2,040	6,149	Mean Per Class
	CIFAR-10 [23] [link]	2009	10	50,000	10,000	Accuracy
	CIFAR-100 [23] [link]	2009	100	50,000	10,000	Accuracy
	ImageNet-1k [40] [link]	2009	1000	1,281,167	50,000	Accuracy
	SUN397 [24] [link]	2010	397	19,850	19,850	Accuracy
	SVHN [92] [link]	2011	10	73,257	26,032	Accuracy
	STL-10 [93] [link]	2011	10	1,000	8,000	Accuracy
	GTSRB [94] [link]	2011	43	26,640	12,630	Accuracy
	KITTI Distance [1] [link]	2012	4	6,770	711	Accuracy
	IIIT5k [95] [link]	2012	36	2,000	3,000	Accuracy
	Oxford-IIIT PETS [26] [link]	2012	37	3,680	3,669	Mean Per Class
	Stanford Cars [25] [link]	2013	196	8,144	8,041	Accuracy
	FGVC Aircraft [96] [link]	2013	100	6,667	3,333	Mean Per Class
	Facial Emotion Recognition 2013 [97] [link]	2013	8	32,140	3,574	Accuracy
	Rendered SST2 [98] [link]	2013	2	7,792	1,821	Accuracy
	Describable Textures (DTD) [99] [link]	2014	47	3,760	1,880	Accuracy
	Food-101 [22] [link]	2014	102	75,750	25,250	Accuracy
	Birdsnap [100] [link]	2014	500	42,283	2,149	Accuracy
	RESISC45 [101] [link]	2017	45	3,150	25,200	Accuracy
	CLEVR Counts [102] [link]	2017	8	2,000	500	Accuracy
	PatchCamelyon [103] [link]	2018	2	294,912	32,768	Accuracy
	EuroSAT [104] [link]	2019	10	10,000	5,000	Accuracy
	Hateful Memes [27] [link]	2020	2	8,500	500	ROC AUC
	Country211 [10] [link]	2021	211	43,200	21,100	Accuracy
Image-Text Retrieval	Flickr30k [105] [link]	2014	-	31,783	-	Recall
	COCO Caption [74] [link]	2015	-	82,783	5,000	Recall
Action Recognition	UCF101 [29] [link]	2012	101	9,537	1,794	Accuracy
	Kinetics700 [30] [link]	2019	700	494,801	31,669	Mean(top1, top5)
	RareAct [28] [link]	2020	122	7,607	-	mWAP, mSAP
Object Detection	COCO 2014 Detection [106] [link]	2014	80	83,000	41,000	box mAP
	COCO 2017 Detection [106] [link]	2017	80	118,000	5,000	box mAP
	LVIS [107] [link]	2019	1203	118,000	5,000	box mAP
	ODinW [108] [link]	2022	314	132413	20070	box mAP
Semantic Segmentation	PASCAL VOC 2012 Segmentation [90] [link]	2012	20	1464	1449	mIoU
	PASCAL Content [109] [link]	2014	459	4998	5105	mIoU
	Cityscapes [110] [link]	2016	19	2975	500	mIoU
	ADE20k [111] [link]	2017	150	25574	2000	mIoU