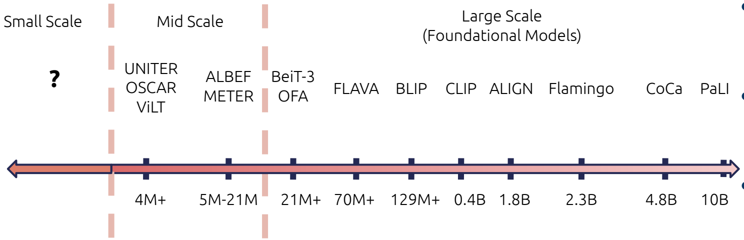
Efficient Vision-Language Pretraining with **Vi**sual **C**oncepts and **H**ierarchical **A**lignment (**ViCHA**)

Mustafa Shukor[1], Guillaume Couairon[1,2], Matthieu Cord[1,3]
[1]Sorbonne University & [2]Meta AI & [3]Valeo.ai

BMVC 2022

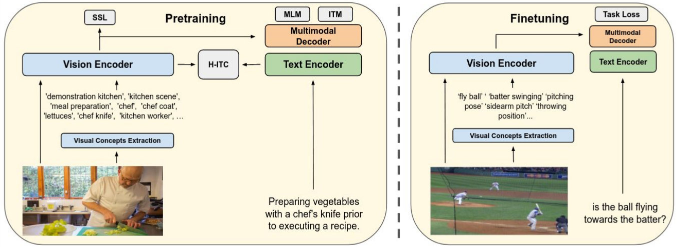
Paper, code, and data are available:
https://github.com/mshukor/ViCHA

Introduction



- Current Vision-Language Models are trained on large datasets, which need huge computation infrastructure.
- Other paths are also promising; training objectives, model architectures and the quality of data.
- **How can we learn from less data?**

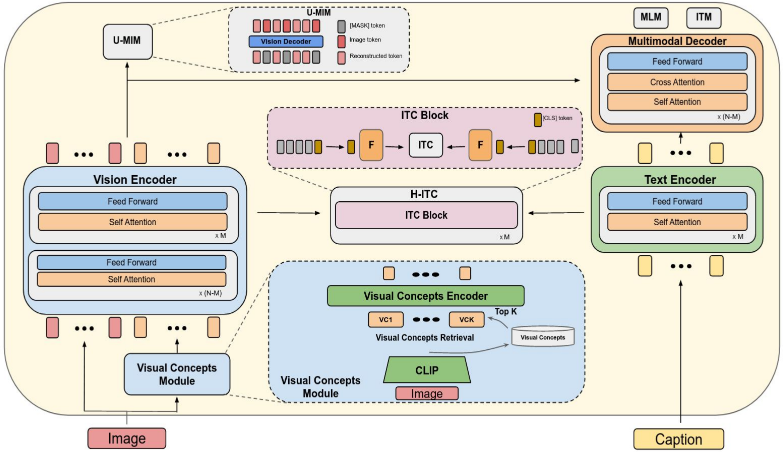
Contributions



- **Contributions:**
  - H-ITC
  - MAE-MIM
  - Visual Concepts Module

Method

ViCHA consists of and image encoder (ViT), text encoder (6-layer BERT) and multimodal decoder (6-layer BERT + CA), trained on 4 main objectives; H-ITC, ITM , MLM and MIM on Image-Text pairs datasets:



- **Hierarchical Image-Text Contrastive (H-ITC)** objective aligns the cross modal representation at different layers of the unimodal encoders.
- **MAE-based Masked Image Modeling (MIM)** objective that leverages MAE [1] for VLP, by reconstructing masked image tokens.
- **Visual Concepts Module** that leverages image-level annotations (Visual Concepts-VCs) using CLIP [2], to enrich the visual representation.

Results

We pretrain on COCO, Visual Genome and SBU and then finetune on VQA v2, SNLI-Ve, COCO and Flickr30K retrieval, and visual grounding.

Finetuning on VQA v2 NLVR2 and SNLI-VE

Method	# Pre-train Images	VQA		$NLVR^2$		SNLI-VE	
		test-dev	test-std	dev	test-P	val	test
UNITER [13]	4M	72.70	72.91	77.18	77.85	78.59	78.28
OSCAR [48]	4M	73.16	73.44	78.07	78.36	-	-
ViLT [39]	4M	70.94	-	75.24	76.21	-	-
ALBEF [44]	4M	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF*	1.1M	72.51	72.69	75.72	76.31	78.08	78.02
ViCHA	1.1M	**73.55**	**73.52**	**77.27**	**77.08**	**78.96**	**78.22**
ViCHA†	800K	73.23	-	78.14	77.00	79.02	78.65

Finetuning on COCO and Flickr30K Image-Text Retrieval

Method	# Pre-train Images	Flickr30K (1K test set)						MSCOCO (5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER [13]	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR [48]	4M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ViLT [39]	4M	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
ALBEF [44]	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
ALBEF*	1.1M	88.1	97.8	99.0	73.6	92.0	95.6	71.2	91.3	95.6	54.2	80.7	88.6
ViCHA	1.1M	**91.7**	**98.7**	**99.5**	**77.2**	**94.2**	**96.8**	**73.6**	**92.4**	**96.2**	**56.8**	**82.2**	**89.5**
ViCHA†	800K	90.0	98.4	99.8	77.4	94.3	96.7	73.3	92.1	96.2	55.8	81.8	89.1

[1] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.