Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts

Zeng et al., ICML, 2022

Presented by: Filippo Momentè (20246132)

Why this work?

- A foundational model
- Relevant for Image Retrieval
- Basis for my Project

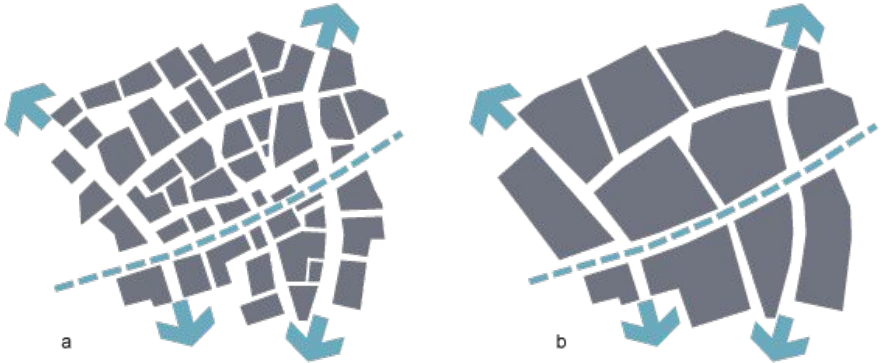
Course Project: Recap

- Idea: using a foundation model pre-trained on vision and language to perform Content-based Image Retrieval
- Captions contain global information about an image
- We can use them to augment our global representations
- We can also use object-level information to improve our representations

Introduction

Multigrained Vision-Language Pre-Training

- Vision-Language Pretraining: learning vision and language alignments from image-text pairs
- Can be divided into:
  - Coarse-grained: encode overall features
  - Fine-grained: learn more detailed information (e.g. object-level)

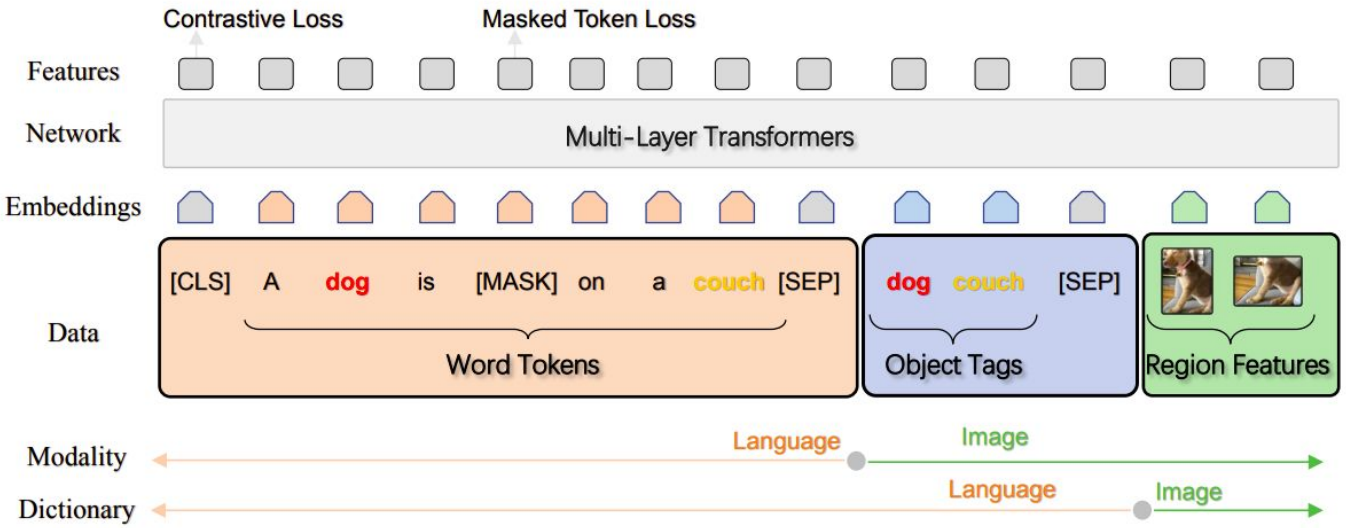


Multigrained Vision-Language Pre-Training

- Previous works had limitations
- Coarse-level:
  - Only general image-level info are learned
- Fine-grained:
  - Object-level information used for representations
  - But no relation between them encoded
  - Number of classes also limited

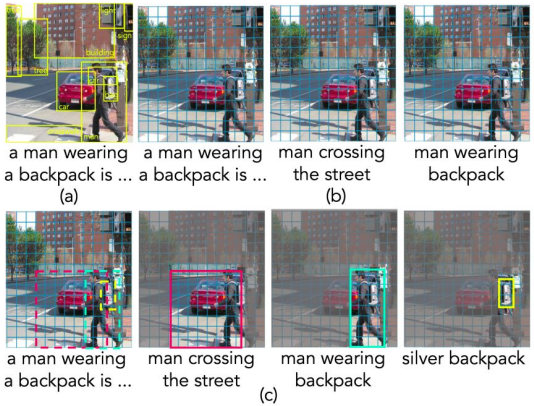
Multigrained Vision-Language Pre-Training

- An example: OSCAR (Li et al., ECCV 2020)



Multigrained Vision-Language Pre-Training

- Hypothesis: image-text and object-text alignments should be learned together
- Three types of data
  - Object labels: “man”, “backpack”
  - Region annotations: “boy wearing backpack”
  - Image captions: “ The first day of school gives a mixed feeling to both students and parents



Architecture Overview

- Text encoder
- Image encoder
- Fusion module
- Object Detection: Box regression + IoU losses
- Jointly trained with Matching Loss, Contrastive Loss, Masked Language Modeling Loss

Contributions

- The first VL foundational model to try to learn both image and object-text pairs together
- It was able to reach SoTA on several tasks
  - Including Image-Text Retrieval

Method

The Architecture

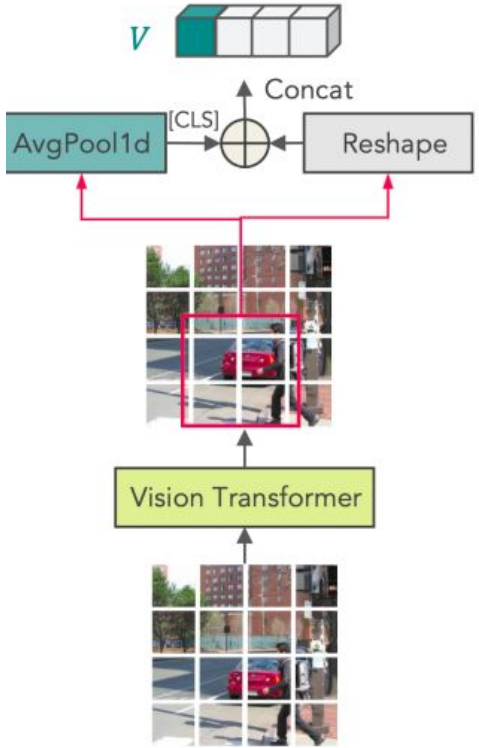
- Image Encoder
- Text Encoder
- Cross-modal Encoder
- All Transformer-based

Input Format

$$(I, T, \{(V^j, T^j)\}^N)$$

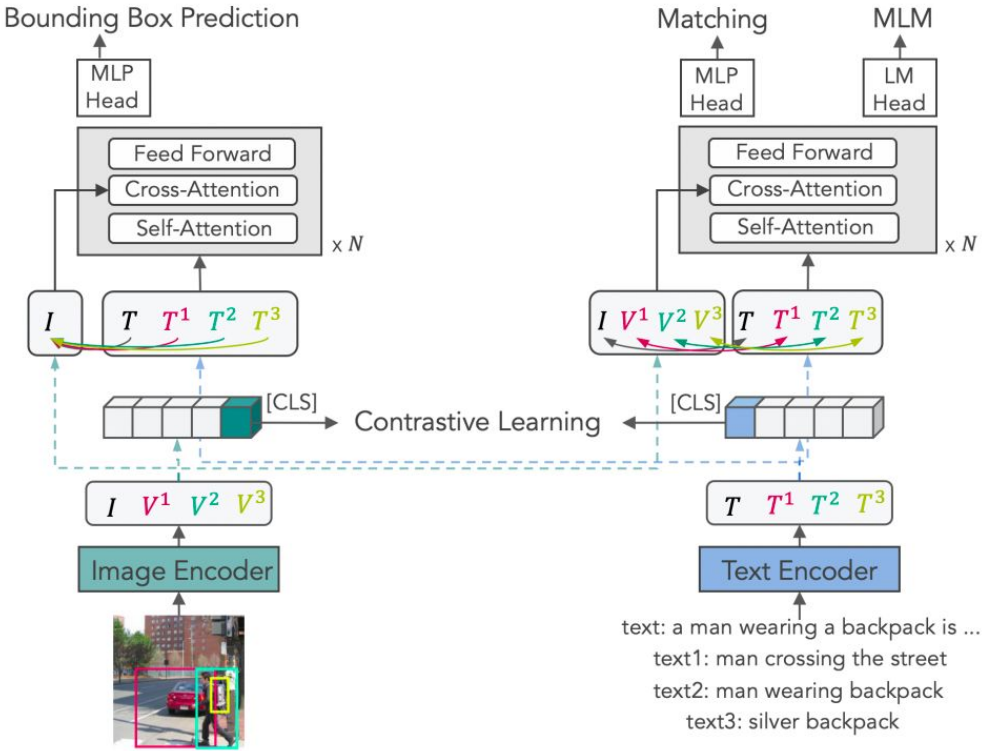
Vision Encoder

- Image split into patches
- Patches passed to the Transformer Layer
- Representing object-level info: patches aggregated together (patch reshape)
- Average of features also computed



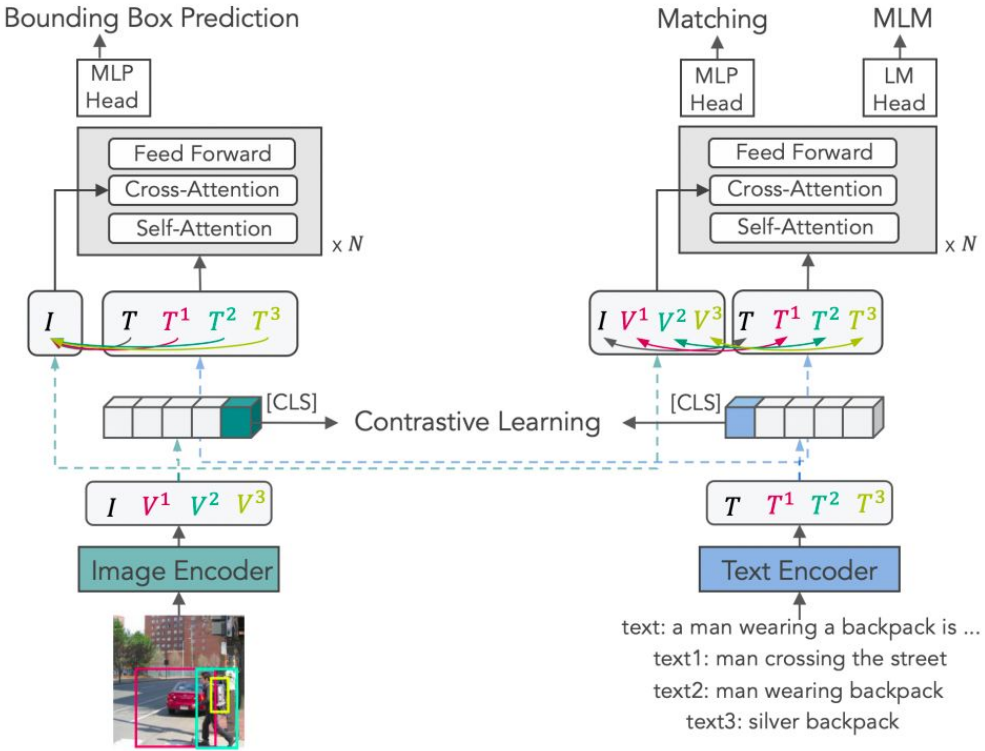
Cross-Modal Modeling

- Obtained visual concepts are aligned with text



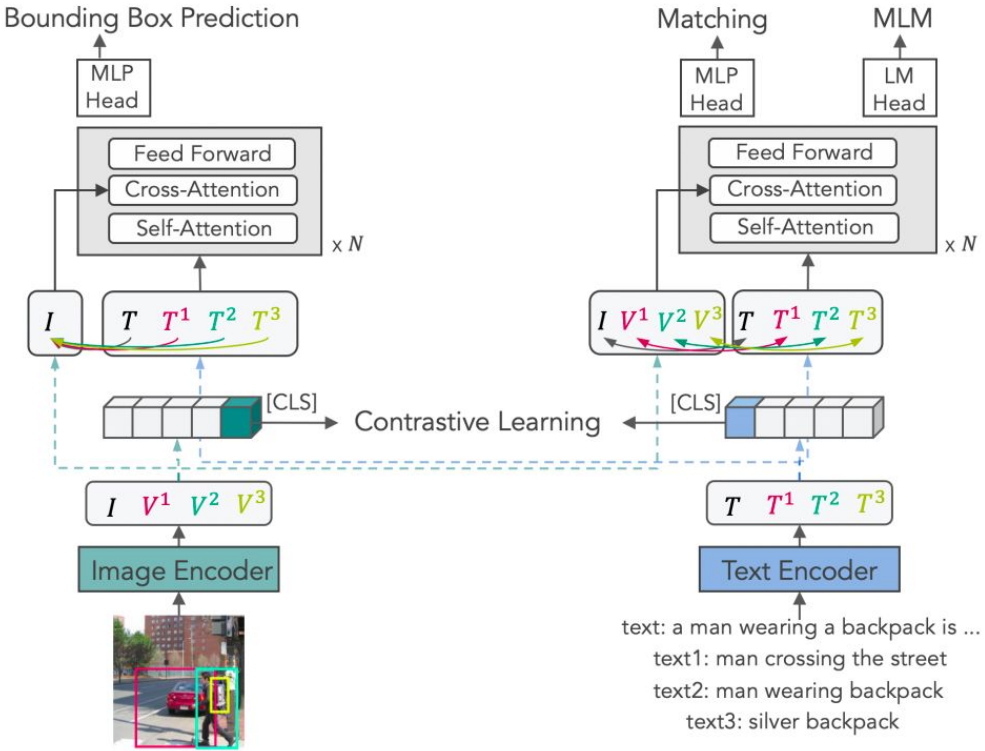
Bounding Box Prediction

- Bounding boxes predicted given image and text representations for each object
- Fused together
- L1 + IoU losses



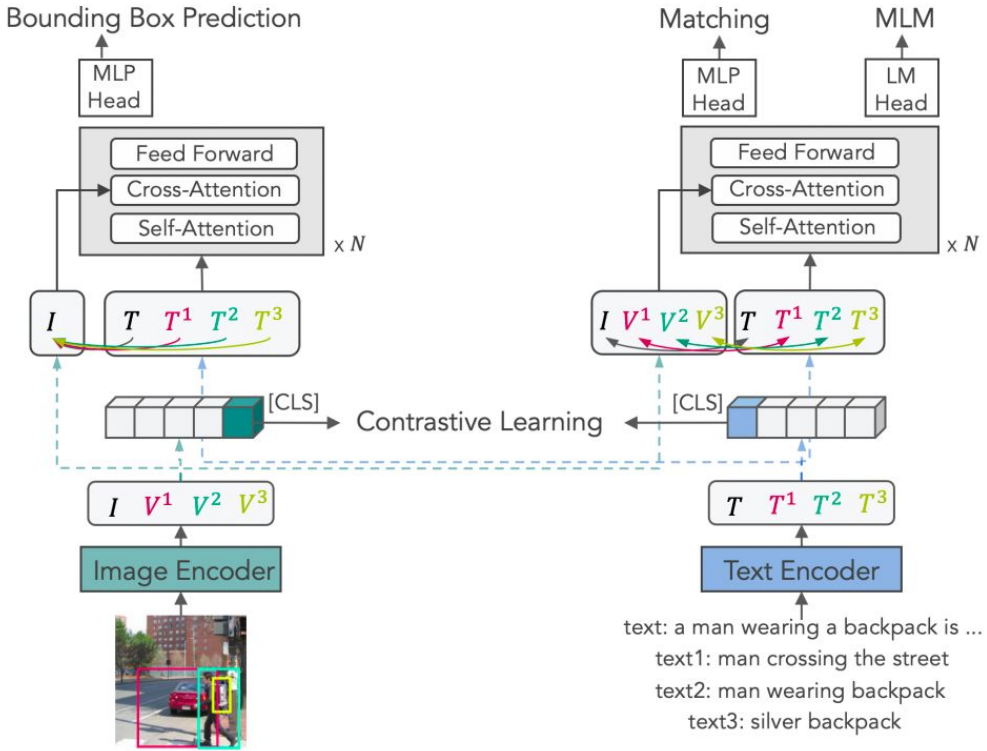
Contrastive Learning

- Samples: (V, T)
- V can be image or object-level representation
- Text and Image contrasted together



Other Losses

- Matching Loss
- Masked Modeling Loss
  - Predict masked word in text based on the visual concept
- Final loss: combination of them



Experiments

Datasets

- A mix of datasets

	Dataset	# Images	# Captions	# Ann
4M	COCO	0.11M	0.55M	0.45M
	VG	0.10M	-	5.7M
	SBU	0.86M	0.86M	-
	CC-3M	2.9M	2.9M	-
16M	4M	4.0M	5.1M	6.2M
	Objects365	0.58M	-	2.0M
	OpenImages	1.7M	-	4.2M
	CC-12M	11.1M	11.1M	-

Results on Image-Text Retrieval

- MSCOCO and Flickr30K

Method	# Params	# Pre-train Images	MSCOCO (5K test set) TR	MSCOCO (5K test set) IR	Flickr30K (1K test set) TR	Flickr30K (1K test set) 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
$\text{UNITER}_{\text{large}}$	300M	4M	65.7 / 88.6 / 93.8	52.9 / 79.9 / 88.0	87.3 / 98.0 / 99.2	75.6 / 94.1 / 96.8
METER-Swin	380M	4M	73.0 / 92.0 / 96.3	54.9 / 81.4 / 89.3	92.4 / 99.0 / 99.5	79.0 / 95.6 / 98.0
ALBEF	210M	4M	73.1 / 91.4 / 96.0	56.8 / 81.5 / 89.2	94.3 / 99.4 / 99.8	82.8 / 96.7 / 98.4
METER-CLIP	380M	4M	76.2 / 93.2 / 96.8	57.1 / 82.7 / 90.1	94.3 / 99.6 / 99.9	82.2 / 96.3 / 98.4
$\text{VinVL}_{\text{large}}$	550M	5.6M	75.4 / 92.9 / 96.2	58.8 / 83.5 / 90.3	-	-
ALIGN	490M	1.8B	77.0 / 93.5 / 96.9	59.9 / 83.3 / 89.8	95.3 / 99.8 / 100.0	84.9 / 97.4 / 98.6
ALBEF	210M	14M	77.6 / 94.3 / 97.2	60.7 / 84.3 / 90.5	95.9 / 99.8 / 100.0	85.6 / **97.5** / **98.9**
X-VLM	216M	4M	80.4 / 95.5 / **98.2**	63.1 / 85.7 / **91.6**	96.8 / 99.8 / 100.0	86.1 / 97.4 / 98.7
X-VLM	216M	16M	**81.2** / **95.6** / **98.2**	**63.4** / **85.8** / 91.5	**97.1** / **100.0** / 100.0	**86.9** / 97.3 / 98.7

Ablation Study

	Meta-Sum	MSCOCO		Flickr30K		VQA	NLVR^2	RefCOCO+	
		TR	IR	TR	IR	test-dev	test-P	testA^d	testB^d
X-VLM	**605.0**	**78.8**	**60.6**	**96.0**	**84.1**	76.20	82.42	72.07	54.84
w/o object	603.5	77.4	60.4	95.0	83.7	75.87	82.10	**73.37**	**55.69**
w/o region	596.0	76.8	60.2	**96.0**	83.6	75.84	82.20	70.73	50.60
w/o bbox loss	594.9	77.5	60.2	95.7	83.5	**76.77**	81.49	69.32	50.38
w/o all	580.6	74.5	57.9	95.6	82.8	74.90	80.70	67.79	46.43

Conclusion

Limitations

- Patches are merged with a non-learnable strategy
  - It would be interesting to make this learnable
- Object-level may not be enough
  - Details are also important
  - Especially in Image Retrieval
  - Even more fine-grained representations relevant

Recap

- This work proposed a novel foundational model for VL tasks
- Jointly pre-trained on image and object-text pairs
- Trained to perform object detection
- SoTA on image-text and text-image Retrieval

Thank you! (Quiz time)

Quiz

1. What is the correct final loss of X-VLM?
    a. IoU + Masked Language Modeling + Contrastive Loss + Matching Loss
    b. Average Precision + Masked Language Modeling + Contrastive Loss
    c. (IoU + L1) + Masked Language Modeling + Contrastive Loss + Matching Loss
    d. Average Precision + IoU + Masked Language Modeling + Contrastive Loss
2. What is NOT true about the architecture?
    a. The image encoder is a Transformer
    b. Contrastive loss is computed both between visual representations alone as well as visual-text pairs
    c. The Image encoder splits images into patches
    d. Textual information is fused with visual one before predicting the bounding boxes