Vision and Language Pre-trained Models

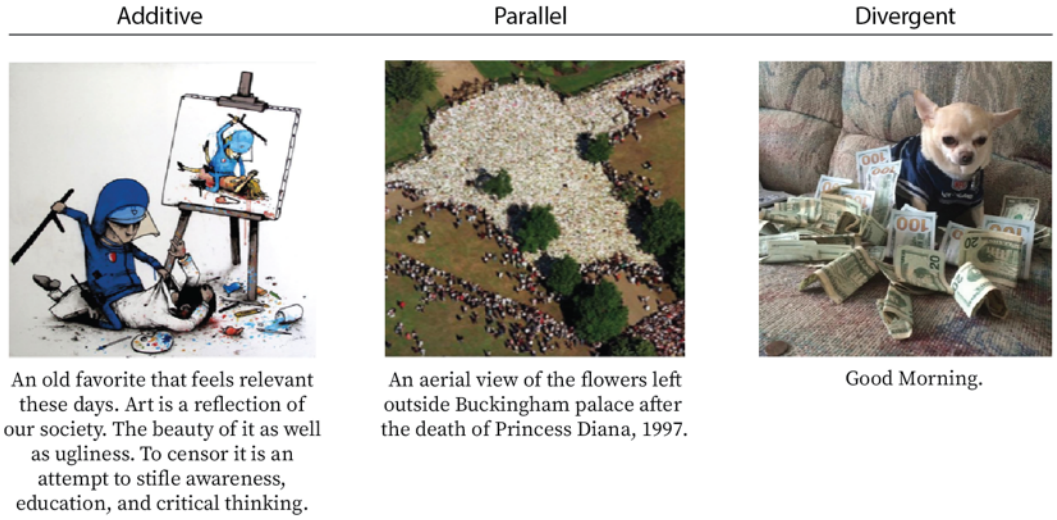
Aix Marseille Université - LIS - TALEP
Emmanuelle Salin
18 janvier 2021

Introduction

Multimodality

Vision Language Multimodality

- They are different types of relationship between image and text in a multimodal document
- Information is often complementary between images and text
- The fusion of the vision and language modalities is needed to fully extract relevant information



From: Integrating text and image: Determining multimodal document intent in instagram posts

Examples of Vision Language Multimodality

Title: Avenue de Clichy - Five O'Clock in the Evening
Author: Louis Anquetin
Type: Landscape
School: French
Timeframe: 1851-1900

This painting is said to have inspired Van Gogh in painting his famous Café Terrace at Night.

From: How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval

Impression:
No acute cardiopulmonary abnormality.

Findings:
There are no focal areas of consolidation.
No suspicious pulmonary opacities.
Heart size within normal limits.
No pleural effusions.
There is no evidence of pneumothorax.
Degenerative changes of the thoracic spine.

MTI Tags: degenerative change

From: On the Automatic Generation of Medical Imaging Reports

Multimodal Representations

Pre-training, Datasets, Tasks

Pre-training: From Natural Language Processing to Vision and Language

- Recent advances in NLP shifted towards pre-training self-supervised models on large unlabelled corpora (BERT). The goal is to learn linguistic knowledge that can be used for supervised tasks with less available data.

- For Vision and Language (VL) pre-training, the goal is to learn a self-supervised model able to extract visual and linguistic information from the documents and combine appropriately, so that the model can be used for multimodal supervised tasks.

Pre-training datasets for Vision and Language

To pre-train Vision and Language models, large corpora with parallel vision and language information are needed.

The most popular datasets are Microsoft COCO, Visual Genome, Google Conceptual Captions and SBU Captions (intersection from COCO).

Dataset	# Images	# Captions	Caption Length
MSCOCO	113K	567K	11.81 ± 2.81
VG	108K	5.41M	5.53 ± 1.76
GCC†	3.01M	3.01M	10.66 ± 4.93
SBU†	867K	867K	15.0 ± 7.74

From: ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

Microsoft COCO Examples

A man holding a child while standing at the fence of an elephant zoo enclosure.

The horse and puppy are separated by the mesh fence.

Visual Genome Examples

Google Conceptual Captions and SBU Captions Examples

GCC Dataset

- Murray in cat bed, Neko not in cat bed

SBU Dataset

Vision Language downstream Tasks

- There are different types of supervised vision/language tasks, for example:
  - Visual Question Answering
  - Visual Commonsense Reasoning
  - Image-Text Retrieval
  - Visual Entailment
  - Referring Expressions

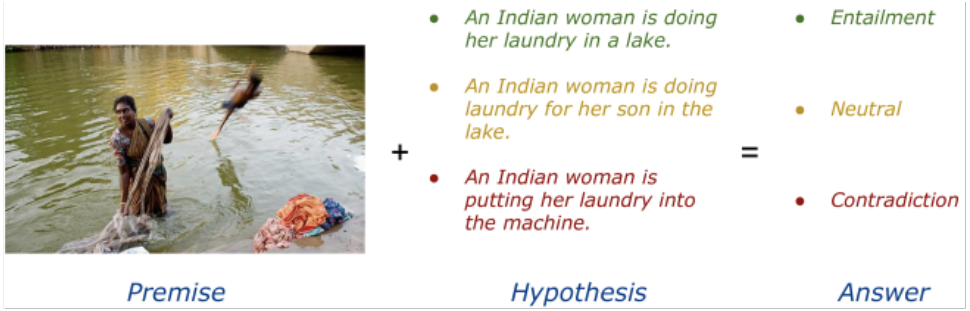
Task	Datasets	Image Src.	#Images	#Text	Metric
1 VQA	VQA	COCO	204K	1.1M	VQA-score
2 VCR	VCR	Movie Clips	110K	290K	Accuracy
3 NLVR^2	NLVR^2	Web Crawled	214K	107K	Accuracy
4 Visual Entailment	SNLI-VE	Flickr30K	31K	507K	Accuracy
5 Image-Text Retrieval	COCO	COCO	92K	460K	Recall@1,5,10
	Flickr30K	Flickr30K	32K	160K	
6 RE Comprehension	RefCOCO	COCO	20K	142K	Accuracy
	RefCOCO+		20K	142K	
	RefCOCOg		26K	95K	

Examples of Supervised Tasks for Vision and Language

Visual Question Answering

From: https://visualqa.org/

Visual Entailment



From: Visual Entailment: A Novel Task for Fine-Grained Image Understanding

Vision Language Pre-trained models

Models, Architecture, Pre-training tasks and Evaluation

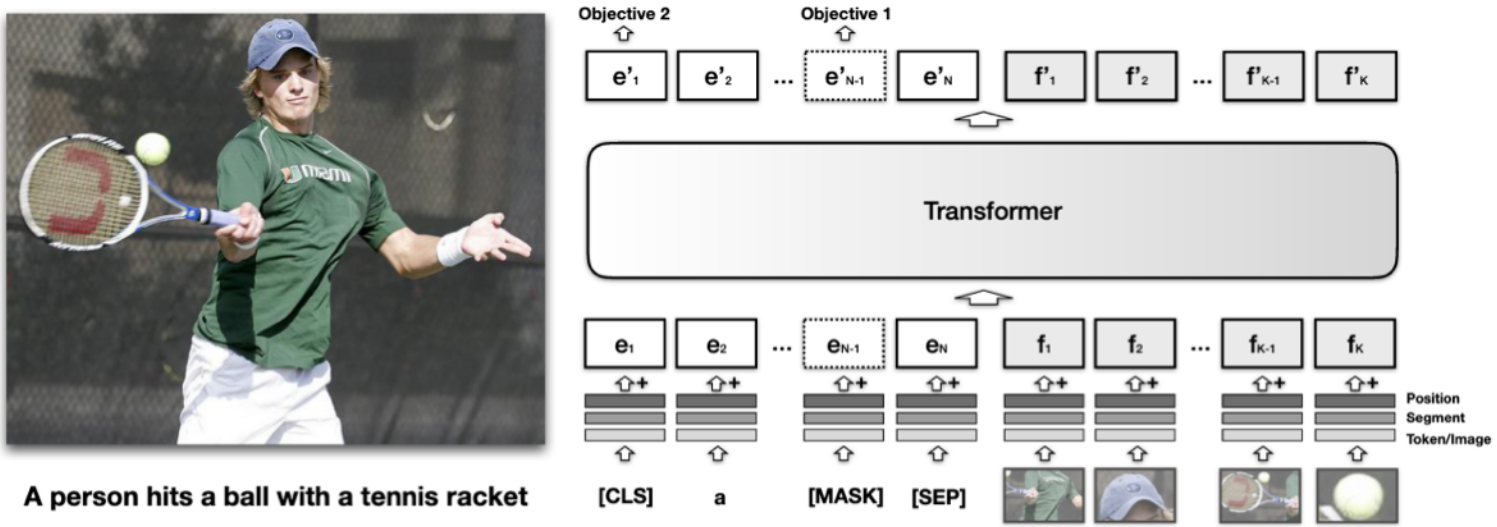
State-of-the-Art Vision and Language models

Various pre-trained Transformer-based Vision and Language models have been developed in recent years, inspired by the advances in NLP

ViLBERT	NeurIPS 2019	Georgia Tech, FAIR	
LXMERT	EMNLP 2019	UNC Chapel Hill	
VL-BERT	ICLR 2020	Microsoft Research Asia	
VisualBERT	ACL 2020	University of California	
Unicoder-VL	AAAI 2020	Peking University	
UNITER	ECCV 2020	Microsoft Dynamics 365	
OSCAR	ECCV 2020	Microsoft Corporation	

General Architecture of Transformer-based VL models

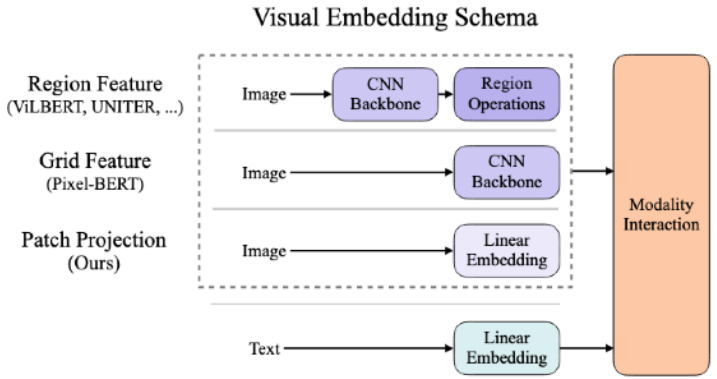
Most SOTA Vision and Language models use the same concept as BERT. The only supervision used is the fact that the image/caption is a pair.



From: VISUALBERT: A SIMPLE AND PERFORMANT BASELINE FOR VISION AND LANGUAGE

Model Input - Vision and Language

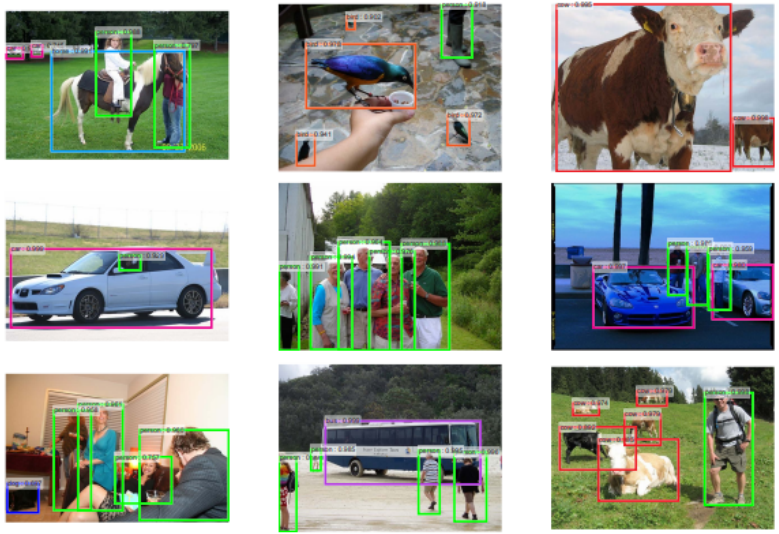
- The input of the model is composed of a sequence of word pieces and image regions, with a [CLS] token used for the pre-training objectives
- The image regions used as input can follow a grid of the image, or use detected object regions in the image
- In most models, the visual regions used as input are obtained with an object detection model to detect relevant image regions



ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

Faster RCNN: object detection

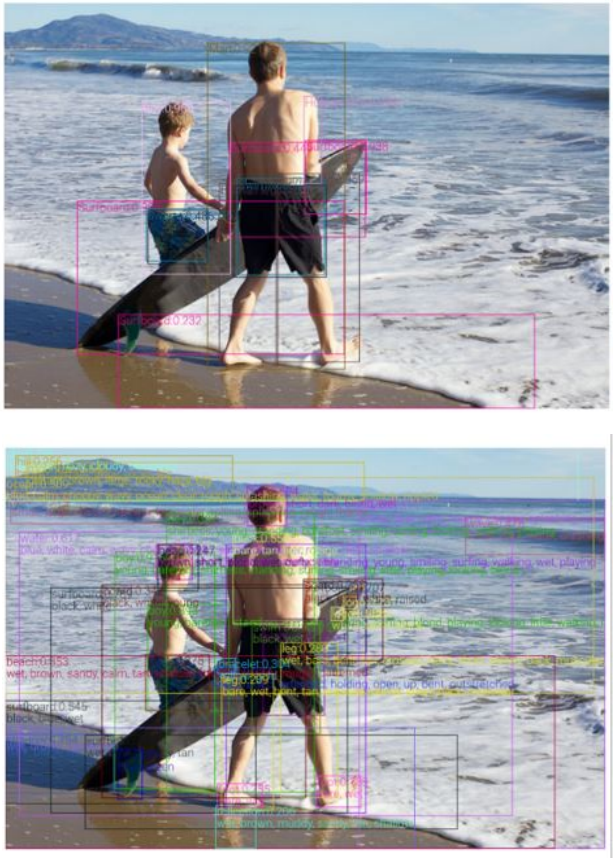
- The Faster RCNN:
  - generates “bounding boxes” of possible objects
  - obtains features of these objects
  - predicts which class the object belongs to
  - makes the coordinates of the object bounding box more precise
- In VL models, the faster RCNN is pre-trained on Visual Genome dataset
- A sequence of the most probable bounding boxes is given as input of the VL model



From: Faster R-CNN for object detection

Faster RCNN: object detection

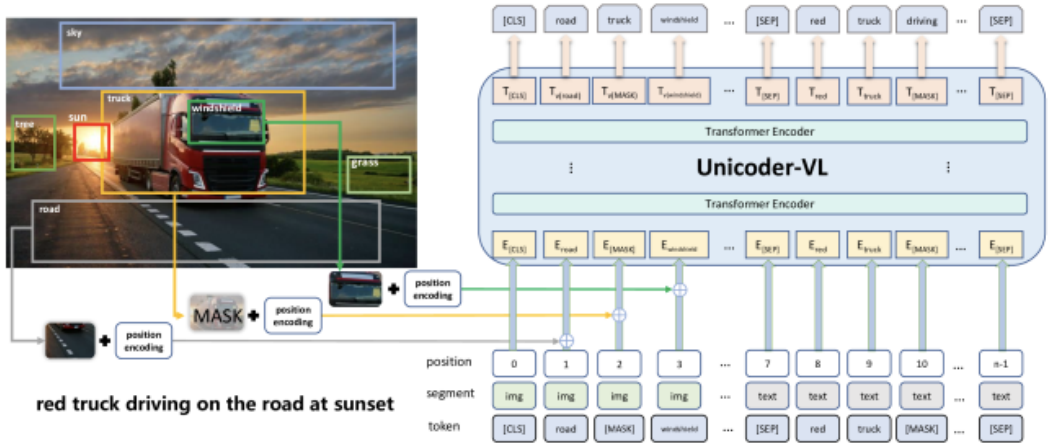
- The VinVL model uses a larger object detection module trained on a bigger dataset. It can detect 1,594 object classes and 524 visual attributes
- Using this module leads to SOTA results on seven major VL benchmarks (Visual Question answering, Image Captioning)
- It shows that there is room for improvement to extract all relevant visual information from the image
- However, the use of a faster RCNN can be time and resource consuming



From: VinVL: Making Visual Representations Matter in Vision-Language Models

Transformer Architecture

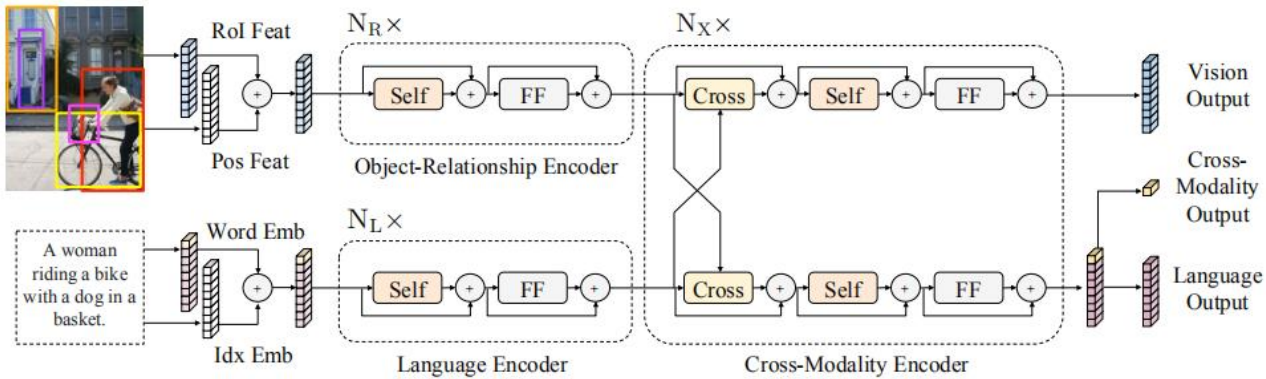
- An important point of a Vision Language model is how the two modalities are combined:
  - If the fusion is early, the model is a single stream architecture, which means that a single transformer combines vision and language (VisualBert, Unicoder-VL, UNITER…)



Single Stream Architecture (Unicoder-VL)

Transformer Architecture

- An important point of a Vision Language model is how the two modalities are combined:
  - If the fusion is early, the model is a single stream architecture, which means that a single transformer combines vision and language (VisualBert, Unicoder-VL, UNITER…)
  - In the case of double stream architecture, there is an encoder for each modality before a cross modality encoder. As a result, the model has more parameters to learn. (VilBERT, LXMERT ...)

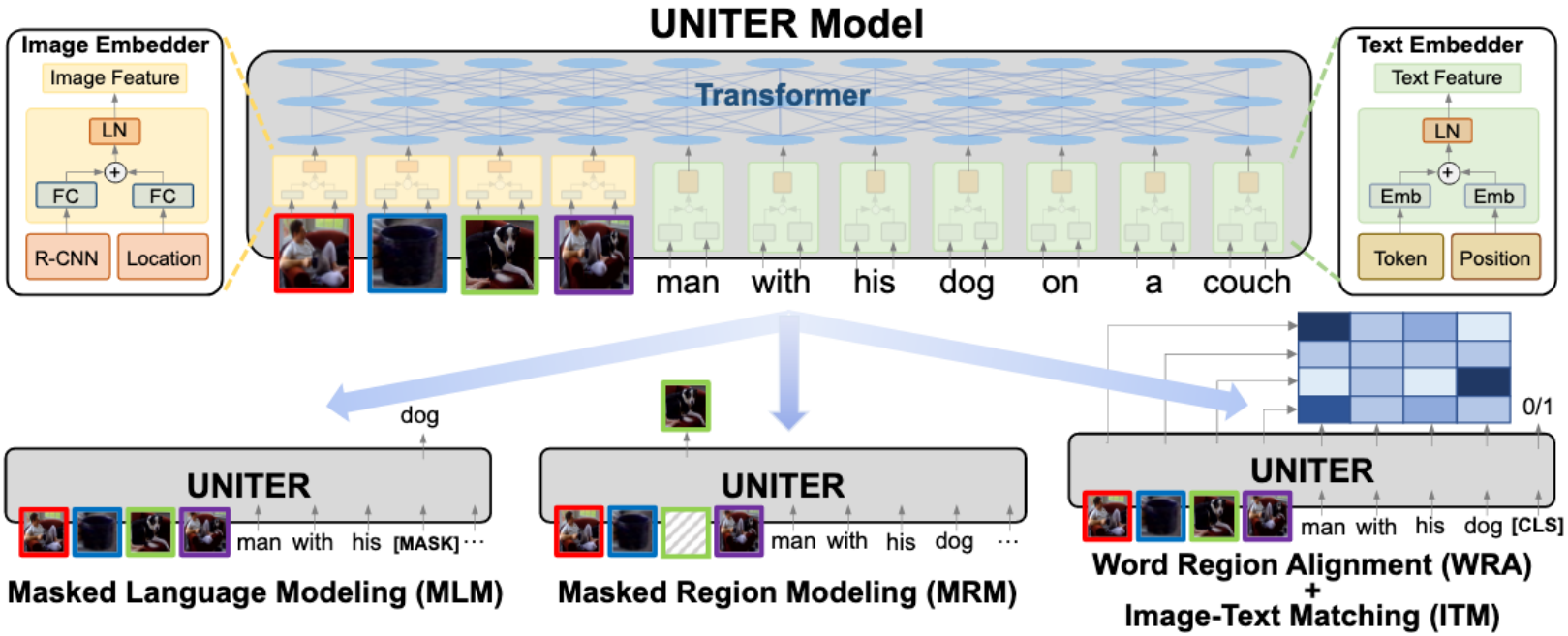


Double Stream Architecture (LXMERT)

Vision Language Pre-Training Tasks

- In NLP pre-trained models, the pre-training objective is usually Masked Language Modeling, which consist in masking a portion of the input tokens and predicting them.
- This task can be found as “Masked Language Modeling with Visual Clues” in all vision language transformer models, as it is essential to learn linguistic knowledge
- Other self-supervised tasks have been developed to learn to extract visual knowledge and to combine modalities. UNITER is the model which shows the most complete set of tasks among SOTA models
- Some models use tasks with additional supervision such as tags (OSCAR) or Visual Question Answering (LXMERT)

UNITER Pre-training Tasks

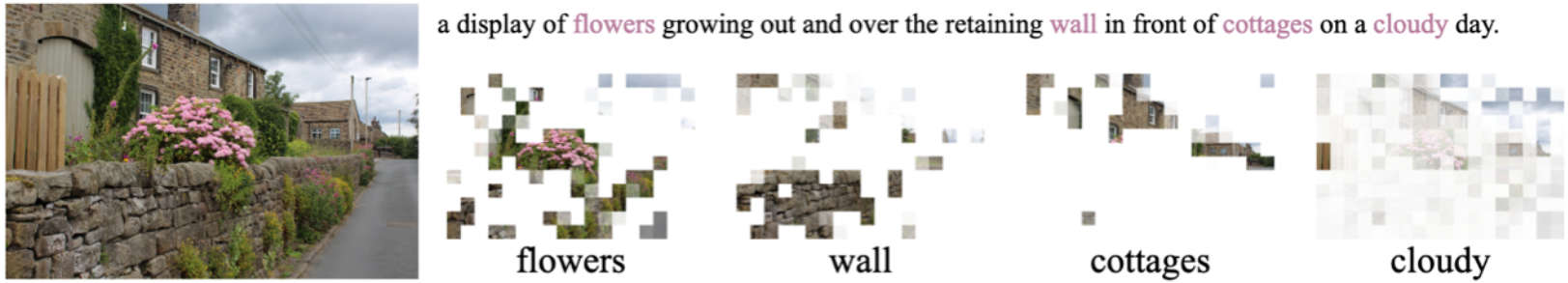


Visual Self-supervised Tasks

- Similar to Masked Language Modeling, Masked Region Modeling consists in masking some input regions, to reconstruct them given the remaining regions and the words. Several variations of this task are used by different models.
- Masked Region Feature Regression learns to regress the output of a masked region to its visual features
- Masked Region Classification (hard) learns to predict the object class of a masked region. There is no ground truth, the result is compared to the most probable output of the Faster-RCNN
- Masked Region Classification (soft) uses soft labels instead of heard labels to predict the object class of the region

Cross-Modal Self-supervised Tasks

- For Image Text Matching pre-training, a [CLS] token is used to indicate the fused representation of image and text. The output is a binary label checking if the image/text pair is a match, supervised over the [CLS] token. This technique is similar to the Next Sentence Prediction task used by BERT
- Some models (UNITER, ViLT) use a Word Region Alignment task, which computes the alignment score of the smaller units (regions and words)



ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

Pre-training tasks: ablation studies on UNITER

- UNITER does not check all combinations of pre-training tasks in its ablation studies
- The results show that pre-training tasks improve the performance, but further evaluation is necessary to check the significance of this improvement

Pre-training Data		Pre-training Tasks	Meta-Sum	VQA	IR (Flickr)	TR (Flickr)	NLVR^2	Ref-COCO+
				test-dev	val	val	dev	val^d
None	1	None	314.34	67.03	61.74	65.55	51.02	68.73
Wikipedia + BookCorpus	2	MLM (text only)	346.24	69.39	73.92	83.27	50.86	68.80
In-domain (COCO+VG)	3	MRFR	344.66	69.02	72.10	82.91	52.16	68.47
	4	ITM	385.29	70.04	78.93	89.91	74.08	72.33
	5	MLM	386.10	71.29	77.88	89.25	74.79	72.89
	6	MLM + ITM	393.04	71.55	81.64	91.12	75.98	72.75
	7	MLM + ITM + MRC	393.97	71.46	81.39	91.45	76.18	73.49
	8	MLM + ITM + MRFR	396.24	71.73	81.76	92.31	76.21	74.23
	9	MLM + ITM + MRC-kl	397.09	71.63	82.10	92.57	76.28	74.51
	10	MLM + ITM + MRC-kl + MRFR	399.97	71.92	83.73	92.87	76.93	74.52
	11	MLM + ITM + MRC-kl + MRFR + WRA	400.93	72.47	83.72	93.03	76.91	74.80
	12	MLM + ITM + MRC-kl + MRFR (w/o cond. mask)	396.51	71.68	82.31	92.08	76.15	74.29
Out-of-domain (SBU+CC)	13	MLM + ITM + MRC-kl + MRFR + WRA	396.91	71.56	84.34	92.57	75.66	72.78
In-domain + Out-of-domain	14	MLM + ITM + MRC-kl + MRFR + WRA	**405.24**	**72.70**	**85.77**	**94.28**	**77.18**	**75.31**

Table 2: Evaluation on pre-training tasks and datasets using VQA, Image-Text Retrieval on Flickr30K, NLVR^2, and RefCOCO+ as benchmarks. All results are obtained from UNITER-base. Averages of R@1, R@5 and R@10 on Flickr30K for Image Retrieval (IR) and Text Retrieval (TR) are reported. Dark and light grey colors highlight the top and second best results across all the tasks trained with In-domain data

Evaluation of pre-trained models

- Similarly to pre-trained models in NLP, the evaluation of those models is done on downstream tasks, however, no benchmark has yet been developed to compare different models, so the comparison is mostly done by the authors on chosen tasks
- As four different datasets can be used for pre-training, different authors use a different combination of those dataset, making the comparison of different models more challenging. Additionally, for some models, the pre-training and supervised tasks are based on the same dataset
- However, these models usually require fine-tuning for a good performance, as the supervised task is too complex. As a result, the evaluation is not based on the pre-trained but the fine-tuned representations. Developing appropriate Vision Language probin tasks is necessary to evaluate the representations

Conclusion

Analysis, Challenges

The importance of pre-training conditions

In Multimodal Pretraining Unmasked: Unifying the Vision and Language BERTs, the authors show that:

- Training data and hyperparameters seem to be responsible for most of the differences between the reported results
- In particular, single and dual stream transformer models seem to be on par
- the embedding layer (and position encoding) plays a crucial role in the performance

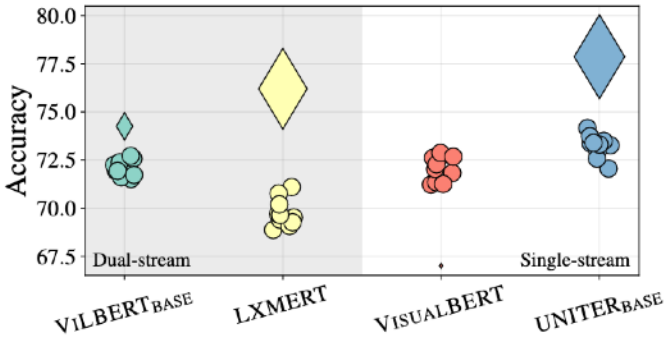
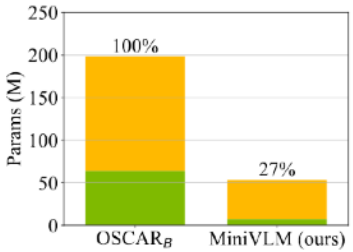
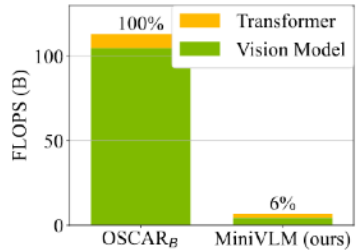


Figure 1: How does the amount of pretraining data affect downstream performance of V&L BERTs? We find that these models perform *more similarly* when trained in the *same conditions*. This plot shows the results from the papers (◇), and when each model is pretrained 10 times on the Conceptual Captions dataset and fine-tuned once on the NLVR2 verification task (○). The size of a marker is proportional to the amount of pretraining data.

Future Research

- Research is currently mostly focused on improving the accuracy of vision language models, but some authors have developed lightweight models, such as MiniVLM





- Studies have shown that pre-trained NLP models are subject to bias, it is even more the case for multimodal representations
- Pre-training tasks are an important part of future research, and could be improved for better generalization of the representations
- Selection and generation of difficult examples could help for certain tasks such as Image Text Matching

Questions ?