E2E-VLP

ACL 2021

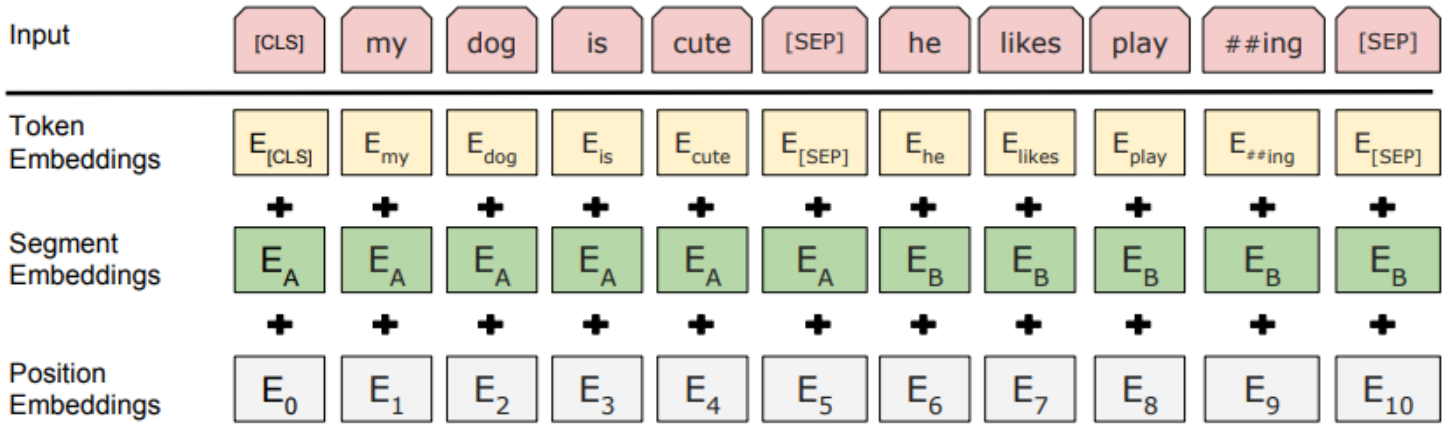
End-to-End Vision-Language Pre-training Enhanced by Visual Learning

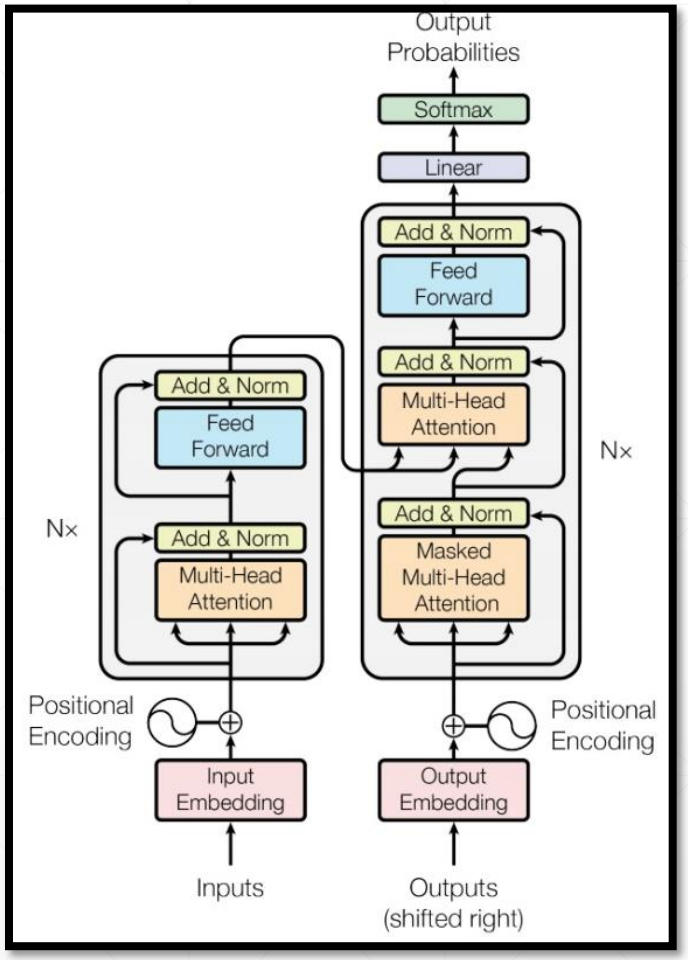
Haiyang Xu*, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, Fei Huang

Alibaba Group

Mengxue

Embedding&Pre-Training Task in BERT



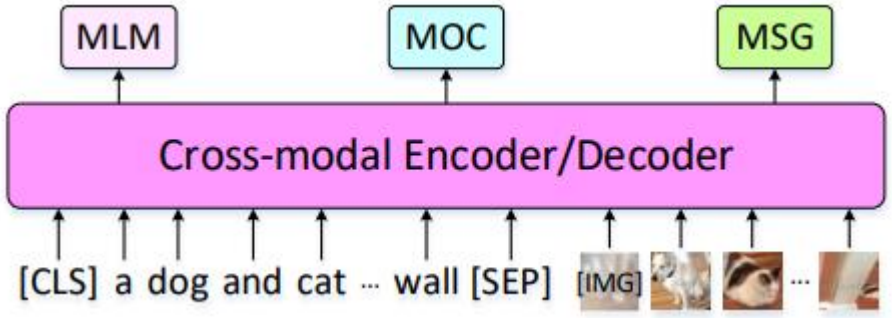


预训练→训练→测试

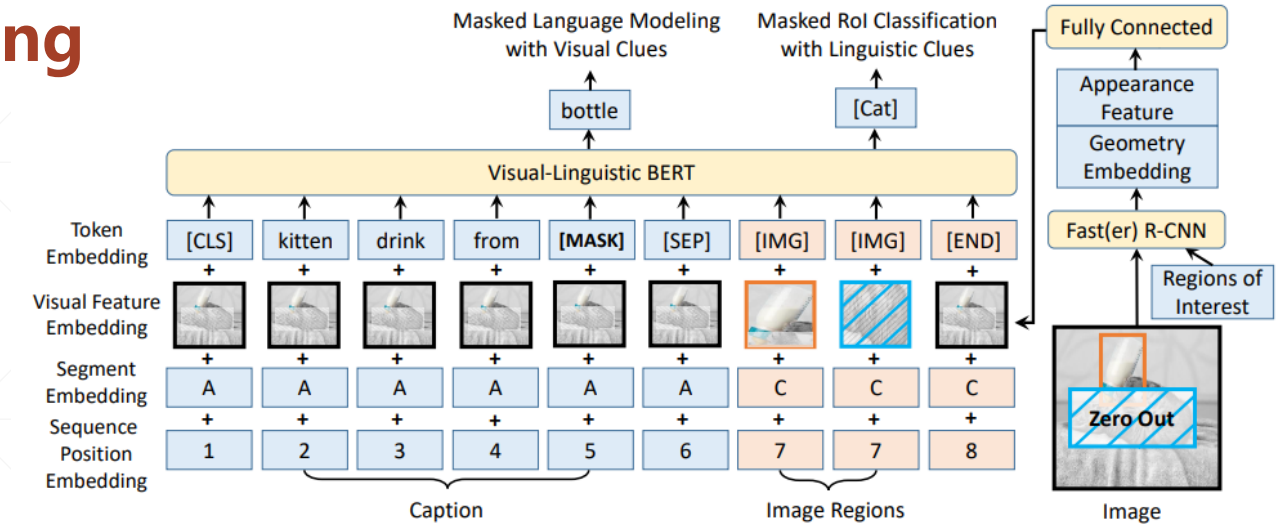
→ 上游任务：Mask Language Modelling & Next Sentence Prediction

→ 下游任务：句子对分类任务、单句子分类任务、问答任务、单句标注任务

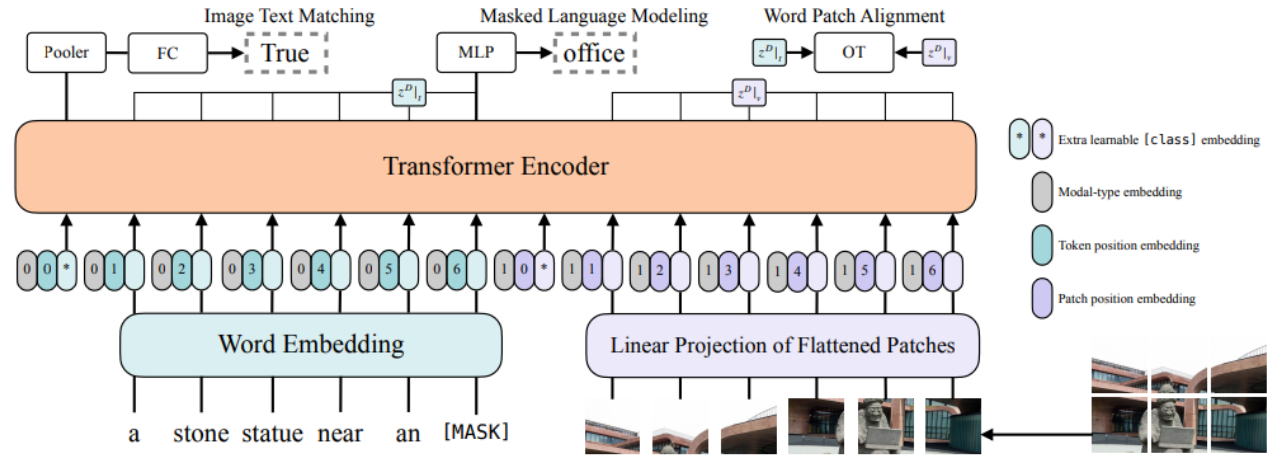
VLP-Visual&language pretraining



- MLM: Masked Language Modeling
- MOC: Masked Object Classification
- MSG: Masked Sentence Generation
- ITM : Image Text Matching

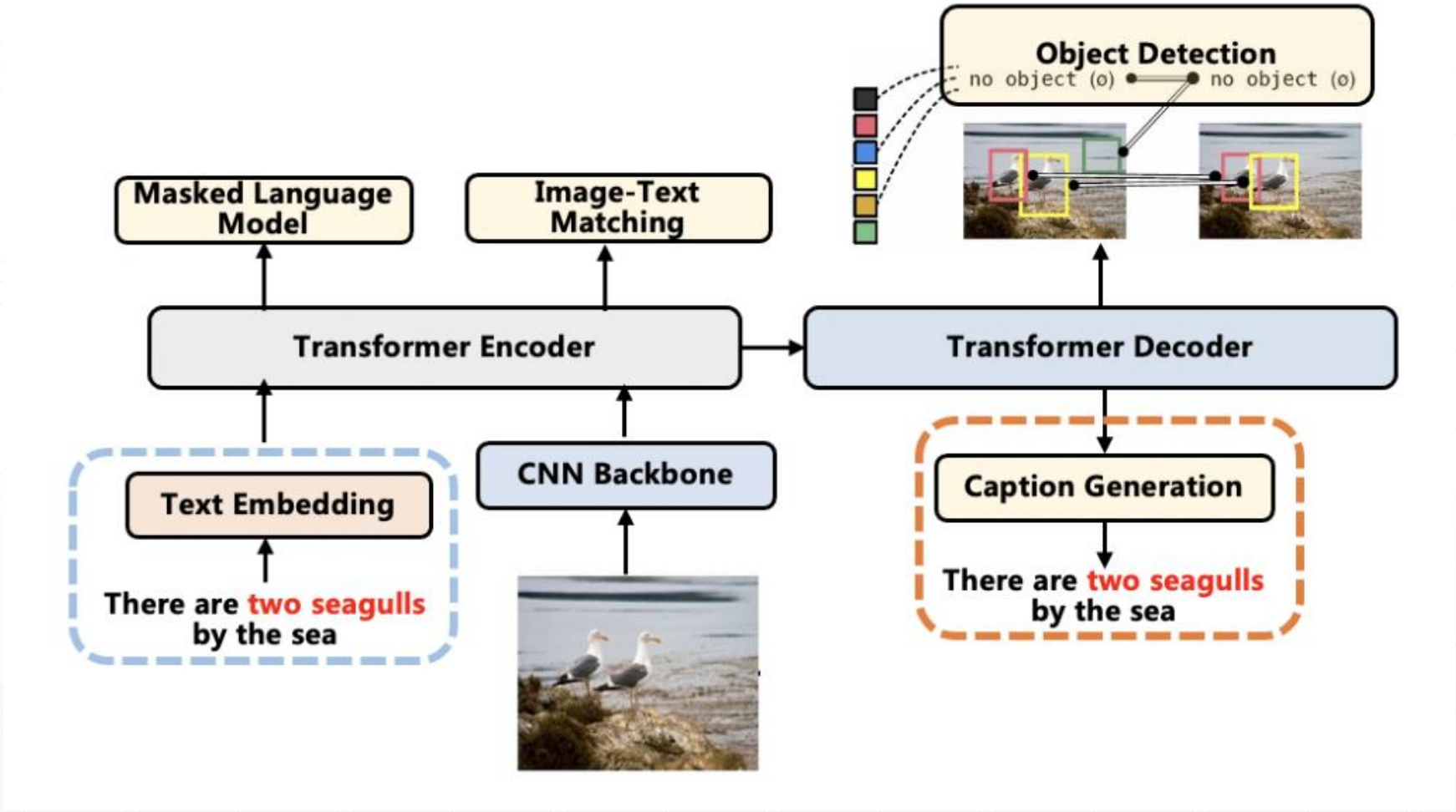


VL-BERT: PRE-TRAINING OF GENERIC VISUALLINGUISTIC REPRESENTATIONS:
Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai (ICLR 2020)



ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision:
Wonjae Kim, Bokyung Son, Ildoo Kim (ICML 2021)

Overall framwork

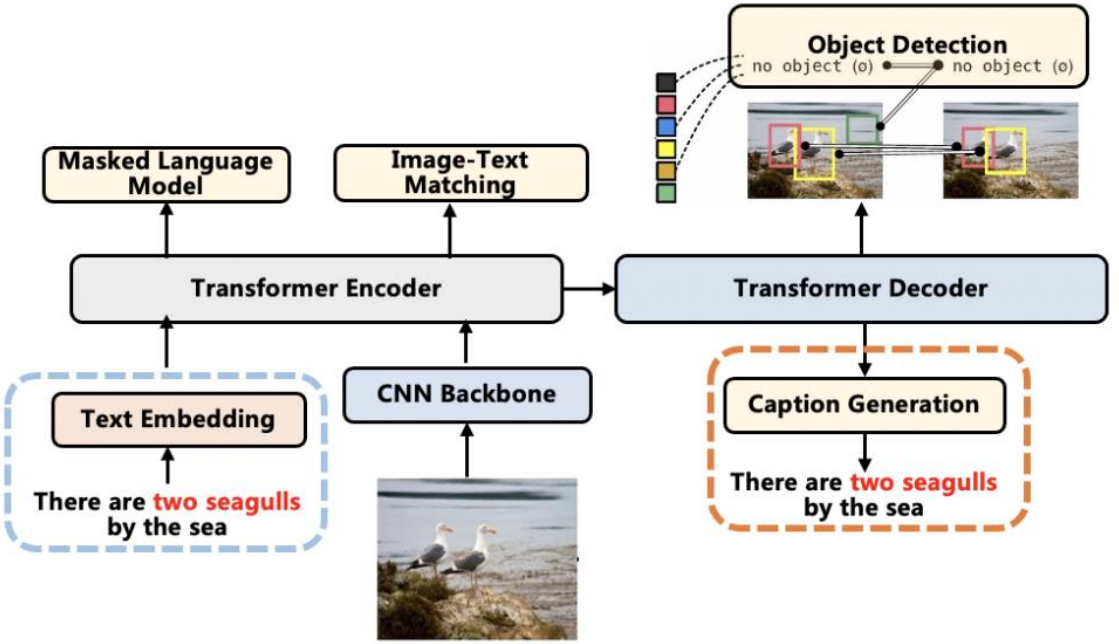


Input Representations

- Sentence Embeddings

$$\{w_1, ..., w_m\}$$

$$E_{emb} = \{e_{CLS}, e_1, ..., e_m, e_{SEP}\}$$



- Image Representations

To improve the generalization of the image representation, we learn from pixels to represent an image instead of using bounding boxes. The pixel features are learned by a CNN visual backbone such as ResNet.

$$f_{img} \in R^{C \times H \times W} \qquad C = 2048 \text{ and } H = \frac{H_0}{32}, W = \frac{w_0}{32}$$

$$z_{img} \in R^{d \times H \times W}$$

$$Z_{img} = \{o_1, ..., o_{HW}\}$$ A HW length of d-dimensional vector

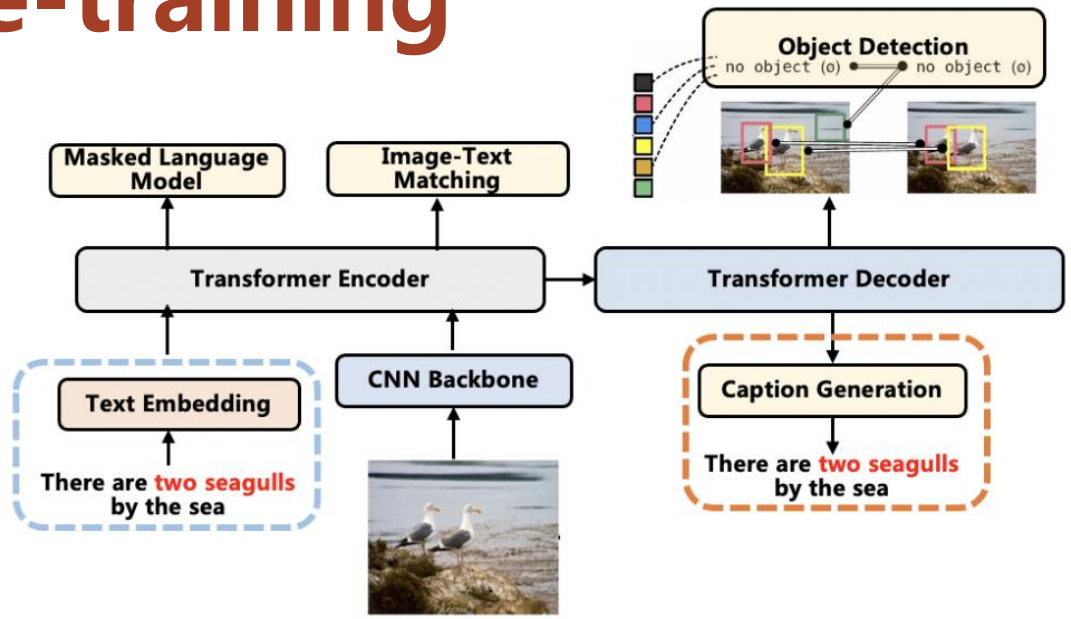
Cross-modal Encoder Pre-training

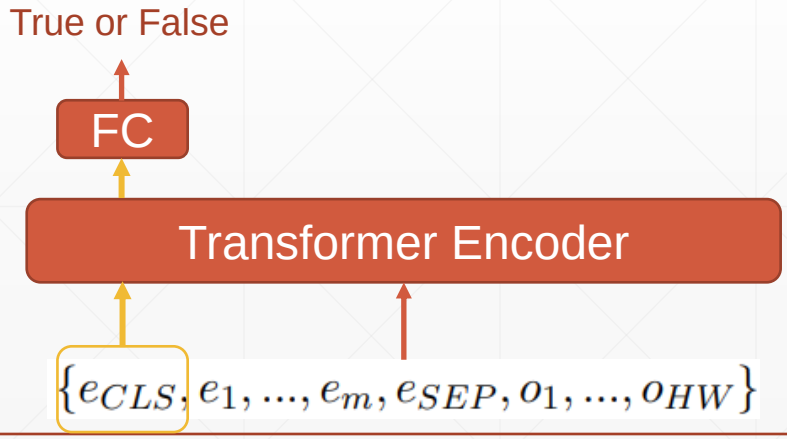
- Input as: $\{e_{CLS}, e_1, ..., e_m, e_{SEP}, o_1, ..., o_{HW}\}$
- Pretraining tasks:
  - Masked Language Modeling (MLM)

    We randomly mask 15% tokens in the text and the model is asked to predict these masked words with the output text and visual representations.

  - Image-Text Matching (ITM)

    We randomly sample 50% mismatched image-text pairs and 50% matched pairs, and train an classifier to predict whether an image and a sentence match each other on the representation of token [CLS] in the last encoder layer h^L_{CLS}

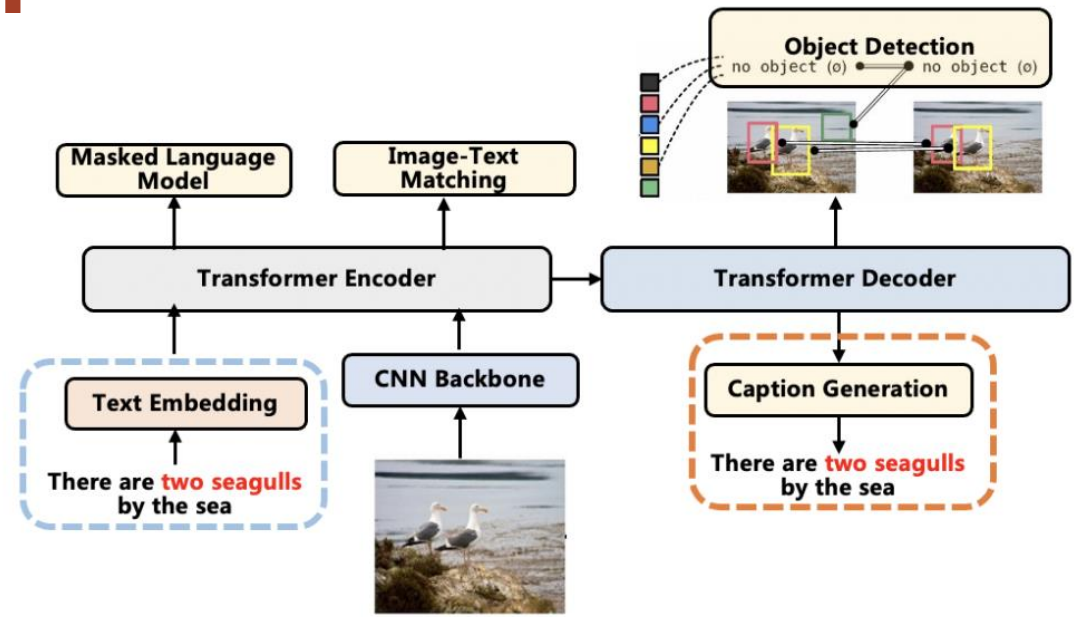




Visual-enhanced Decoder

- Motivation

Due to that the CNN feature map has no object level semantics, it is difficult to directly align the cross-modal semantics between CNN feature map and the language embeddings.



- Transformer decoder
  - object detection

    understanding the fine-grained object information within image
  - image-caption generation

    guide the learning of visual features regarding the textual semantics

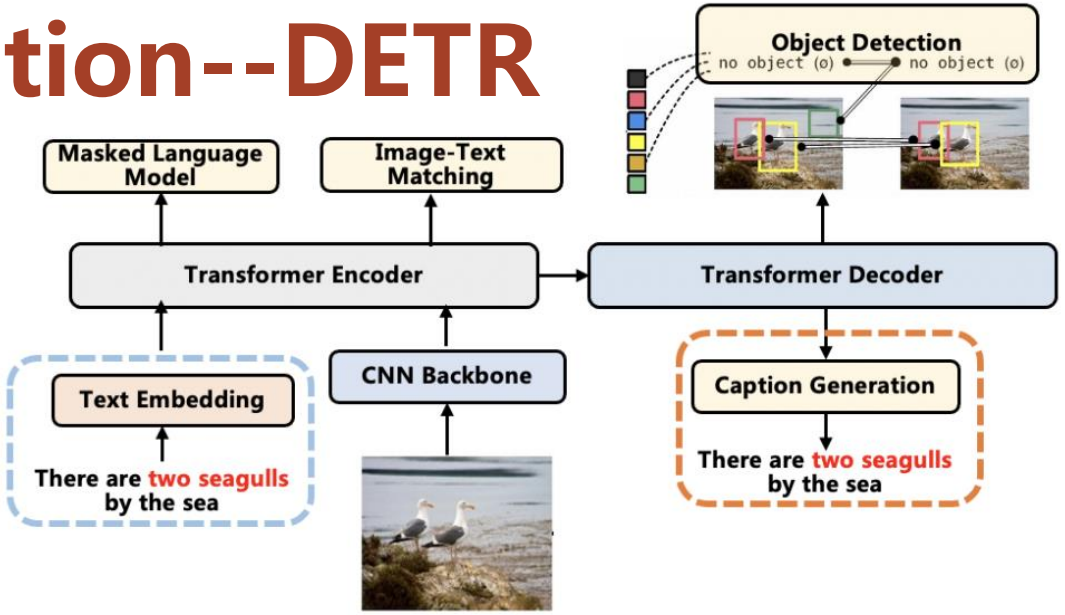
Enhanced by Object Detection--DETR

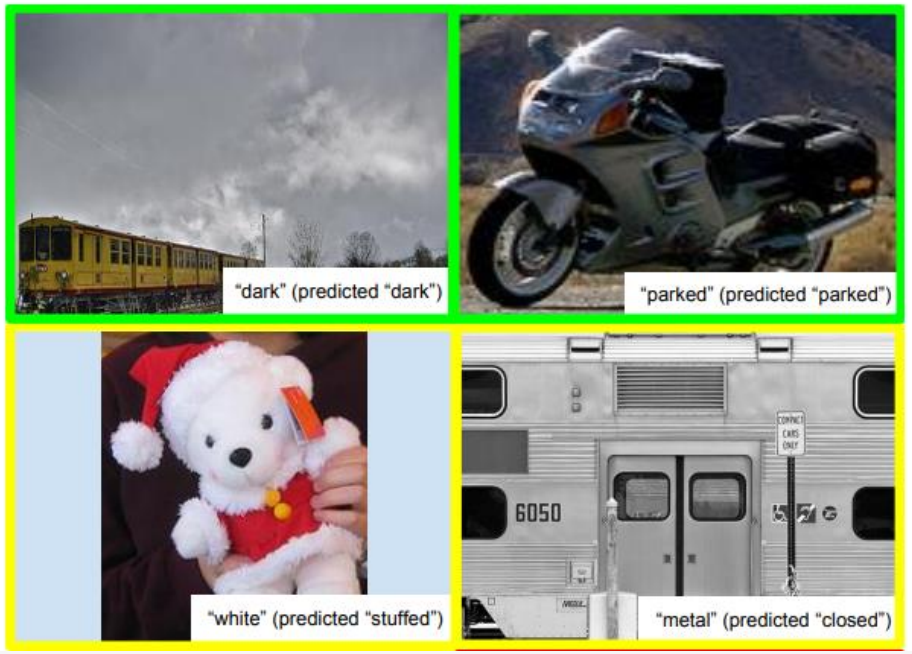
- In encoder side

combine both the visual representation and language embedding as input and reuse the Transformer encoder for cross-modal fusion.

- In decoder side
  - Input: learned positional embeddings
  - Detects the N objects in parallel at each decoder layer
  - Box coordinate regression
  - Class category prediction
  - Object attribute prediction task for VG Dataset

$$\mathcal{L}_v(y,\hat{y}) = \sum_{i=1}^{N}[-\log\hat{p}_{\hat{\sigma}(i)}(a_i) - \log\hat{p}_{\hat{\sigma}(i)}(c_i) + \\ + \mathcal{L}_{box}(b_i,\hat{b}_{\hat{\sigma}(i)}(i))]$$

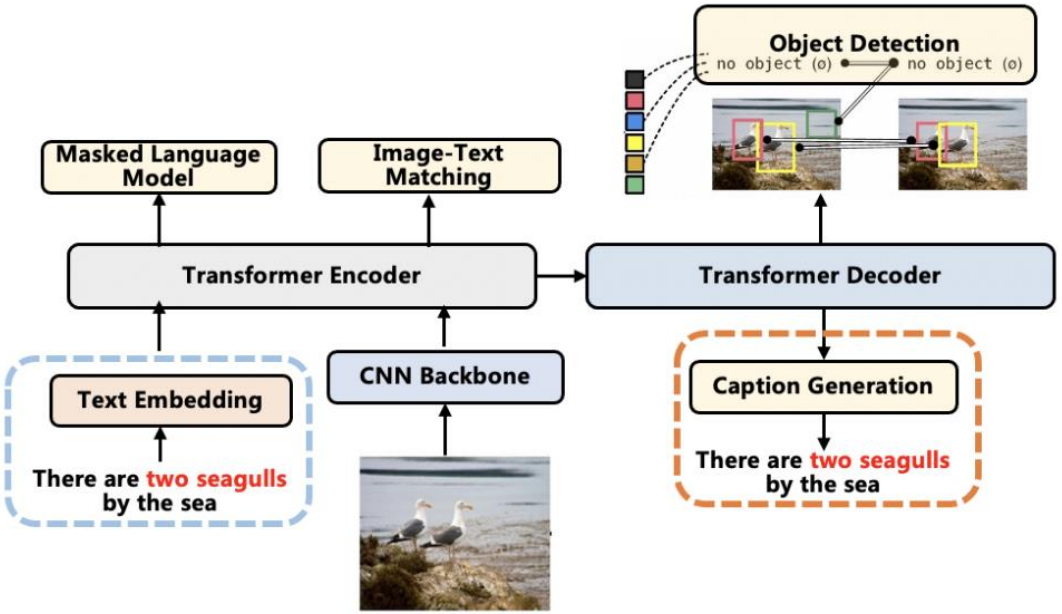




Enhanced by Image Captioning

$$\mathcal{L}_{dec} = -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} \log \prod_{t=1}^{n} P(y_t|y_{<t}, x) \quad (2)$$

where $\mathcal{X}$ represents the sequence of vision context, $\mathcal{Y}$ represents the set of text to be generated and n is the length of tokens in output text y.



Joint Training

We pre-train E2E-VLP with all the encoder and decoder pre-training tasks：

- Masked Language Modeling
- Image-Text Matching
- Object Detection
- Image-to-Text Generation

jointly by minimizing the four loss functions as:

$$\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_{itm} + \mathcal{L}_{v} + \mathcal{L}_{dec}$$

Experiments-Pre-training

- Dataset
  - We pre-train our E2E-VLP on two in-domain image-text datasets: MS-COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017).
  - Object detection and image caption annotations in MS-COCO
  - Object detection, region description annotations in Visual Genome.
- Implementation Details
  - Pre-train E2E-VLP model with a total batch size of 32 for 200 epochs on 8 V100 GPUs.

Experiments-Downstream Tasks

- VQA
- NLVR2
- Image Caption
- Image-Text Retrieval

Table 4. Comparison of ViLT-B/32 with other models on downstream retrieval tasks. We use SCAN for w/o VLP SOTA results. † additionally used GQA, VQAv2, VG-QA for pretraining. ‡ used open images dataset to expand its visual vocabulary of region feature embeddings. ⓐ indicates RandAugment is applied during finetuning.

Visual Embed	Model	Time (ms)	Text Retrieval						Image Retrieval					
			Flickr30k (1K)			MSCOCO (5K)			Flickr30k (1K)			MSCOCO (5K)		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	w/o VLP SOTA	~900	67.4	90.3	95.8	50.4	82.2	90.0	48.6	77.7	85.2	38.6	69.3	80.4
	ViLBERT-Base	~920	-	-	-	-	-	-	58.2	84.9	91.5	-	-	-
	Unicoder-VL	~1000	86.2	96.3	99.0	62.3	87.1	92.8	71.5	91.2	95.2	48.4	76.7	85.9
	UNITER-Base	~900	85.9	97.1	98.8	64.4	87.4	93.1	72.5	92.4	96.1	50.3	78.5	87.2
	OSCAR-Base†	~900	-	-	-	70.0	91.1	95.5	-	-	-	54.0	80.8	88.5
	VinVL-Base†‡	~1000	-	-	-	74.6	92.6	96.3	-	-	-	58.1	83.2	90.1
Grid	Pixel-BERT-X152	~120	87.0	98.9	99.5	63.6	87.5	93.6	71.5	92.1	95.8	50.1	77.6	86.2
	Pixel-BERT-R50	~60	75.7	94.7	97.1	59.8	85.5	91.6	53.4	80.4	88.5	41.1	69.7	80.5
Linear	ViLT-B/32	~15	81.4	95.6	97.6	61.8	86.2	92.6	61.9	86.8	92.8	41.3	72.0	82.5
	ViLT-B/32ⓐ	~15	83.7	97.2	98.1	62.9	87.1	92.7	62.2	87.6	93.2	42.6	72.8	83.4

Models		Params	VQA		NLVR2		COCO Caption	
			Test-dev	Test-std	Dev	Test-P	BLEU4	CIDEr
Single-stream	VisualBERT	110M	70.80	71.00	-	-	-	-
	VLP	110M	70.5	70.7	-	-	36.5	116.9
	VLBERT	110M	71.16	-	-	-	-	-
	Unicoder-VL	110M	-	-	-	-	-	-
	UNITER	110M	72.70	72.91	77.14	77.87	-	-
	OSCAR	110M	73.16	73.61	78.07	78.36	36.5	123.7
Two-stream	ViLBERT	221M	70.55	70.92	67.40	67.00	-	-
	12-in-1	221M	73.15	-	-	-	-	-
	LXMERT	183M	72.42	72.54	74.90	74.50	-	-
	ERNIE-ViL	210M	72.62	72.85	-	-	-	-
End2End	PixelBERT	142M	71.35	71.42	71.7	72.4	-	-
Our Model	E2E-VLP	94M	73.25	73.67	77.25	77.96	36.2	117.3

Table 1: Evaluation Results on VQA, NLVR2 and Image Caption.

Models		Params	IR-Flickr30K			TR-Flickr30K		
			R@1	R@5	R@10	R@1	R@5	R@10
Single-stream	VisualBERT	110M	-	-	-	-	-	-
	VLBERT	110M	-	-	-	-	-	-
	Unicoder-VL	110M	71.50	90.90	94.90	86.20	96.30	99.00
	UNITER	110M	72.52	92.36	96.08	85.90	97.10	98.80
	OSCAR	110M	-	-	-	-	-	-
Two-stream	ViLBERT	221M	58.20	84.90	91.52	-	-	-
	12-in-1	221M	67.90	-	-	-	-	-
	LXMERT	183M	-	-	-	-	-	-
	ERNIE-ViL	210M	74.44	92.72	95.94	86.70	97.80	99.00
End2End	PixelBERT	142M	59.8	85.5	91.6	75.7	94.7	97.1
Our Model	E2E-VLP	94M	73.58	92.42	96.03	86.24	97.50	98.92

Table 2: Evaluation Results on Flickr30K.

Experiments-Downstream Tasks

- VQA
- NLVR2
- Image Caption
- Image-Text Retrieval

Models		Params	VQA	
			Test-dev	Test-std
Single-stream	VisualBERT	110M	70.80	71.00
	VLP	110M	70.5	70.7
	VLBERT	110M	71.16	-
	Unicoder-VL	110M	-	-
	UNITER	110M	72.70	72.91
	OSCAR	110M	73.16	73.61
Two-stream	ViLBERT	221M	70.55	70.92
	12-in-1	221M	73.15	-
	LXMERT	183M	72.42	72.54
	ERNIE-ViL	210M	72.62	72.85
End2End	PixelBERT	142M	71.35	71.42
Our Model	E2E-VLP	94M	73.25	73.67

Table 1: Evaluation Results on VQA, NLVF

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

We use SCAN for w/o VLP SOTA results. † o expand its visual vocabulary of region feature

Image Retrieval					
Flickr30k (1K)			MSCOCO (5K)		
R@1	R@5	R@10	R@1	R@5	R@10
48.6	77.7	85.2	38.6	69.3	80.4
58.2	84.9	91.5	-	-	-
71.5	91.2	95.2	48.4	76.7	85.9
72.5	92.4	96.1	50.3	78.5	87.2
-	-	-	54.0	80.8	88.5
-	-	-	58.1	83.2	90.1
71.5	92.1	95.8	50.1	77.6	86.2
53.4	80.4	88.5	41.1	69.7	80.5
61.9	86.8	92.8	41.3	72.0	82.5
62.2	87.6	93.2	42.6	72.8	83.4

lickr30K		TR-Flickr30K		
.@5	R@10	R@1	R@5	R@10
	-	-	-	-
	-	-	-	-
0.90	94.90	86.20	96.30	99.00
2.36	96.08	85.90	97.10	98.80
	-	-	-	-
4.90	91.52	-	-	-
	-	-	-	-
	-	-	-	-
2.72	95.94	86.70	97.80	99.00
5.5	91.6	75.7	94.7	97.1
2.42	96.03	86.24	97.50	98.92

s on Flickr30K.

Ablation study

Importance of Visual Learning

Model	VQA	NLVR2
E2E-VLP	70.76	72.12
-Image-to-Text Generation	70.20	71.59
-Attribute Prediction	69.92	70.92
-Object Detection	68.85	70.38

Table 3: Ablation tests for different visual pre-training tasks of E2E-VLP (6 layer encoder, and ResNet50 backbone) on development set.

Architecture Selection

Layers	Backbone	Params	VQA	NLVR2
6	r50	49M	70.56	72.12
6	r101	68M	71.42	74.34
6	r152	84M	72.23	76.21
12	r50	59M	71.34	73.04
12	r101	78M	72.43	75.23
12	r152	94M	73.25	77.25

Table 5: Results of different pre-trained model architectures on development set.

Inference Efficiency

Model	Parameters	Avg Time (ms)	VQA	NLVR2
LXMERT	183M	496	72.42	72.54
UNITER	110M	501	72.70	77.14
Pixel-BERT	142M	201	71.35	71.7
E2E-VLP	94M	192	73.25	77.25

Impact of Input Image Size

Input Size shorter side	Input Size longer side	Speedup	VQA	NLVR2
448	448	5x	71.14	75.43
448	746	3x	72.04	75.79
600	1000	1.5x	73.08	76.87
800	1333	-	73.25	77.25

Object Detection with Paired Text

Model	AP	AP_{50}	AP_S	AP_M	AP_L
DETR	40.6	61.6	19.9	44.3	60.2
E2E-VLP	41.9	62.6	20.3	45.6	61.1

Table 7: Results of object detection on MSCOCO development dataset

Thank you