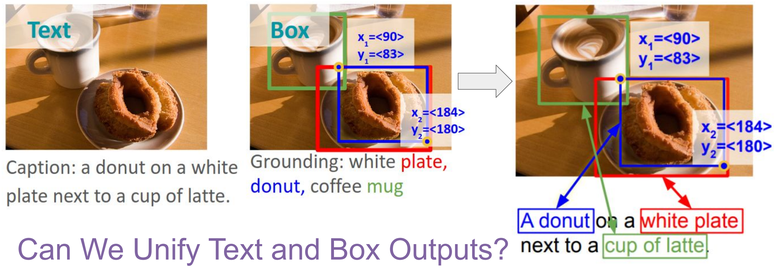
UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, Lijuan Wang

ECCV TEL AVIV 2022

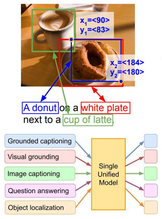
github.com/microsoft/UniTAB

Unifying Text and Box Outputs



Can We Unify Text and Box Outputs?

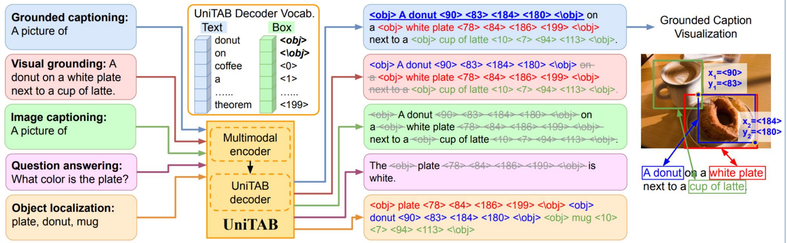
- Supporting both text and box outputs
- Representing word-box alignments



Motivations:

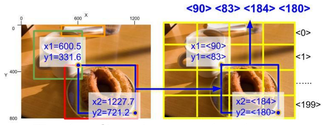
- Grounded description: more comprehensive and interpretable visual description
- Single task-agnostic unified systems: towards unified VL modeling

How to Build a Shared Model for Text and Box?



- Text and box outputs: unified decoding vocabulary
- Word-box alignments: <obj> token

Text and box outputs



- Textualize box outputs
- Auto-regressive decoding
- Unified text+box vocabulary

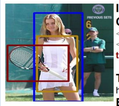
Word-box alignments

<obj> A donut <90> <83> <184> <180> <\obj> on a <obj> white plate <78> <84> <186> <199> <\obj> next to a <obj> cup of latte <10> <7> <94> <113> <\obj>.

- Target sequence design box tokens after words to ground
- <obj> token: word boundary and ease training

Quantitative Results

- Text, box, alignment
- Grounded captioning
- Flickr30k Entities



Input text: A picture of
Output seq.: <obj> A blond woman <53> <19> <142> <199> <obj> in <obj> a white tennis outfit <58> <66> <132> <179> <obj> is holding <obj> a tennis racket <4> <85> <116> <158> <obj> .
Text: A blond woman in a white tennis outfit is holding a tennis racket .
Box: a blond woman, a white tennis outfit, a tennis racket

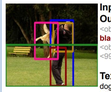
Method	Caption Eval.				Grounding Eval.	
	B@4	M	C	S	$F1_{all}$	$F1_{loc}$
NBT [49]	27.1	21.7	57.5	15.6	-	-
GVD [86]	27.3	22.5	62.3	16.5	7.55	22.2
Cyclical [50]	26.8	22.4	61.1	16.8	8.44	22.78
POS-SCAN [88]	30.1†	22.6†	69.3†	16.8†	7.17	17.49
Chen *et al.* [9]	27.2	22.5	62.5	16.5	7.91	21.54
UniTAB	**30.1**	**23.7**	**69.7**	**17.4**	**12.95**	**34.79**

- Box, alignment
- Visual grounding
- Refcoco/+/g, Flickr30k

Input text: A woman is manipulating dishes in a dishwasher with a man and another woman next to her .
Output seq.: <obj> A woman <35> <56> <101> <199> <obj> is manipulating <obj> dishes <79> <144> <99> <179> <obj> in <obj> a dishwasher <76> <178> <113> <199> <obj> with <obj> a man <91> <13> <181> <199> <obj> and <obj> another woman <156> <37> <195> <199> <obj> next to her .
Text: Not Used
Box: a woman, dishes, a dishwasher, a man, another woman

Method	Refcoco			Refcoco+			Refcocog	
	val	testA	testB	val	testA	testB	val-u	test-u
MAttNet [79]	76.40	80.43	69.28	64.93	70.26	56.00	66.67	67.01
FAOA [77]	72.05	74.81	67.59	55.72	60.37	48.54	59.03	58.70
TransVG [18]	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
ViLBERT [47]	-	-	-	72.34	78.53	62.61	-	-
UNITER [12]	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67
VILLA [22]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR [34]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
UniTAB	**88.59**	**91.06**	**83.75**	**80.97**	**85.36**	**71.55**	**84.58**	**84.70**

- Text
- Image captioning, VQA
- MSCOCO, VQAv2



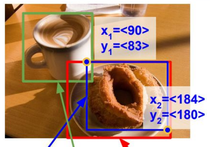
Input text: A picture of
Output seq.: <obj> A man <97> <40> <146> <199> <obj> in <obj> a black jacket <103> <47> <145> <115> <obj> and <obj> black pants <97> <106> <130> <197> <obj> is playing with <obj> a dog <63> <46> <113> <133> <obj> in <obj> a park <0> <99> <199> <199> <obj> .
Text: A man in a black jacket and black pants is playing with a dog in a park .
Box: Not Used

Method	#Pre-train	B@4	M	C	S
Unified VLP [87]	3M	36.5	28.4	117.7	21.3
OSCAR [43]	4M	36.5	30.3	123.7	23.1
E2E-VLP [75]	180K	36.2	-	117.3	-
VL-T5 [13]	180K	34.5	28.7	116.5	21.9
VL-BART [13]	180K	35.1	28.7	116.6	21.5
UniTAB	200K	36.1	28.6	119.8	21.7

Key Takeaways

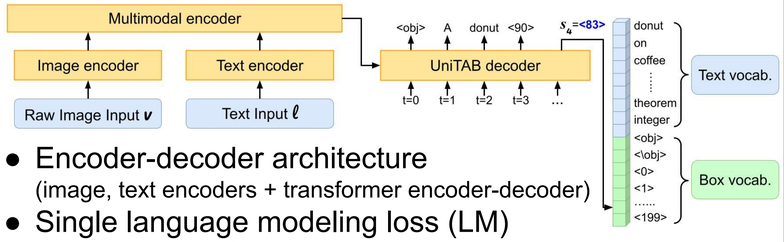
UnITAB: unifying text and box outputs for grounded VL modeling

- Grounded description ability
- Unified modeling for VL tasks
- Parameter efficient and generalizable



<obj> A donut <90> <83> <184> <180> <\obj> on a <obj> white plate <78> <84> <186> <199> <\obj> next to a <obj> cup of latte <10> <7> <94> <113> <\obj>.

UniTAB Framework and Training

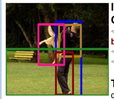


- Encoder-decoder architecture (image, text encoders + transformer encoder-decoder)
- Single language modeling loss (LM)

$$\mathcal{L}_{LM}(\theta) = -\sum_{t=1}^{T} \log P_\theta(s_t | s_{<t}, v, l)$$

Multi-task finetuning: UniTAB$_{\text{Shared}}$

- A single set of parameters for all experimented VL tasks
- (1) Parameter efficient, (2) generalizing learned abilities



Input text: A picture of
Output seq.: <obj> A man <97> <40> <146> <199> <obj> in <obj> a black jacket <103> <47> <145> <115> <obj> and <obj> black pants <97> <106> <130> <197> <obj> is playing with <obj> a dog <63> <46> <113> <133> <obj> in <obj> a park <0> <99> <199> <199> <obj> .
Text: A man in a black jacket and black pants is playing with a dog in a park .
Box: Not Used

- Grounded description
- MSCOCO

Input text: Brittany spaniel
Output seq.: <obj> Brittany spaniel <29> <31> <136> <199> <obj>
Text: Not Used
Box: Brittany spaniel

- Object localization
- ImageNet