Deep Learning 과정

Pretrained Visual Linguistic Models

Gunhee Kim

Computer Science and Engineering

서 울 대 학 교

SEOUL NATIONAL UNIVERSITY

Outline

- VideoBERT
- VL-BERT

C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV 2019.

VideoBERT

Can we learn the high-level (more semantical) relationship between the visual and the linguistic domain?

Combining three off-the-self methods!

- Automatic speech recognition (ASR): convert speech into text
- Vector quantization (VR): convert frames into visual words
- BERT: learn joint distributions over sequences of multimodal tokens

Architecture

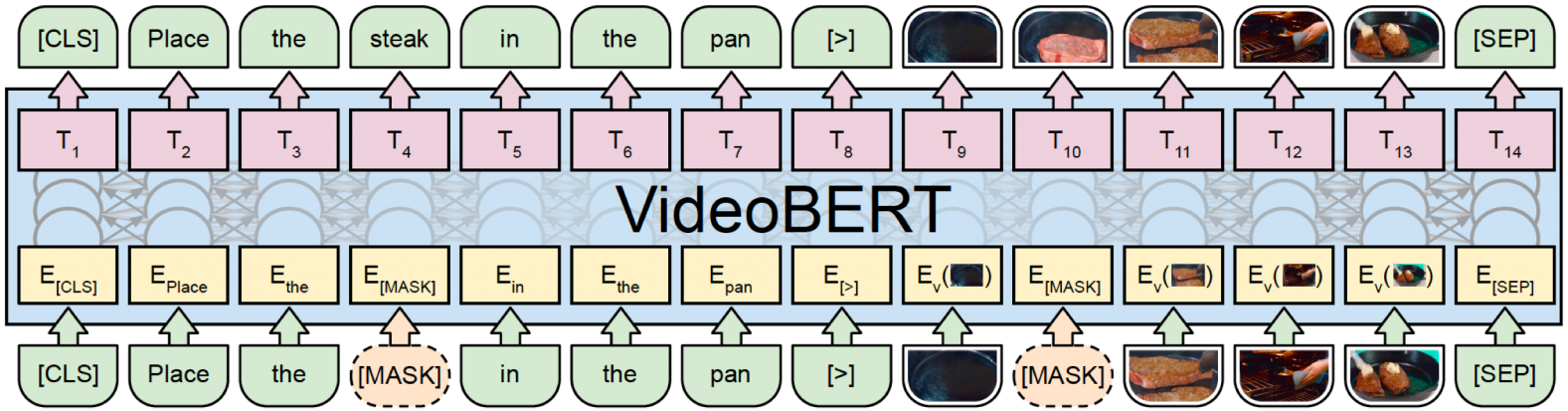
Make minimal changes to extend BERT to video!

- Apply ASR to extract linguistic sentences from a video
- Transform the raw visual data into a discrete sequence of tokens

`[CLS] orange chicken with [MASK] sauce [>] v01 [MASK] v08 v72 [SEP]`

A special token between text and video

Visual words



Architecture

Two pre-training tasks

- Cloze task: same with BERT
- NSP task: linguistic-visual alignment task (Is the clip corresponding to the text?)

The linguistic-visual alignment

- Problem: it is not always aligned well (e.g. something that is not visually presented can be spoken)
- (1) Randomly concatenate neighboring sentences into a single long sentence
- (2) randomly pick a subsampling rate of 1 to 5 steps (the transition pace of the same action can vary)

Pre-training

The training objective is a weighted sum of three objectives

- (1) text-only and (2) video-only: the standard mask-completion objectives
- (3) text-video: linguistic-visual alignment task

Model pre-training

- Use the $\text{BERT}_{\text{Large}}$ model
- Crawl cooking videos from YouTube with keywords “cooking” “recipe (312K videos for video-only, 120K English videos for text-only and text-video)
- Train for 2 days (0.5M iterations) in 4 Cloud TPUs with 128 batch size

Preprocessing

For videos

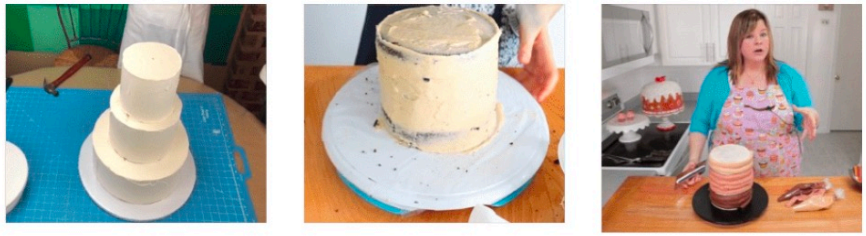
- Extract S3D features from 30 frames (1.5) sampled at 20fps
- Tokenize the visual features using hierarchical k-means (12^4 = 20736 codes)
- Can preserve semantic information rather than low-level visual appearance

For sound

- YouTube's ASR toolkit from YouTube Data API

A video segment

Visual word (centroid)



"but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party."

Zero-shot Action classification

Task

- Train on YouTube dataset and test on YouCook II
- The model predicts the verb and noun labels for

Top verbs: make, assemble, prepare
Top nouns: pizza, sauce, pasta

Top verbs: make, do, pour
Top nouns: cocktail, drink, glass

"now let me show you how to [MASK] the [MASK]

Results

- Low in Top-1 accuracy (since VideoBERT uses open vocabulary) but comparable in the top-5 accuracy

Method	Supervision	verb top-1 (%)	verb top-5 (%)	object top-1 (%)	object top-5 (%)
S3D [34]	yes	16.1	46.9	13.2	30.9
BERT (language prior)	no	0.0	0.0	0.0	0.0
VideoBERT (language prior)	no	0.4	6.9	7.7	15.3
VideoBERT (cross modal)	no	3.2	43.3	13.1	33.7

Video Captioning

Task

- Demonstrate the effectiveness of videoBERT as feature extractor
- Use the captioning model of transformer [Zhou et al. 2018]
- Use the videoBERT feature, S3D and their concatenation as input feature

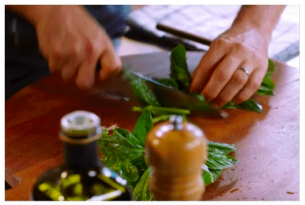
Results

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou *et al.* [39]	7.53	3.84	11.55	27.44	0.38
S3D [34]	6.12	3.24	9.52	26.09	0.31
VideoBERT (video only)	6.33	3.81	10.81	27.14	0.47
VideoBERT	6.80	4.04	11.01	27.50	0.49
VideoBERT + S3D	**7.59**	**4.33**	**11.94**	**28.80**	**0.55**

Zhou et al. End-to-end dense video captioning with masked transformer. CVPR 2018.

Qualitative Results

Video captioning

GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

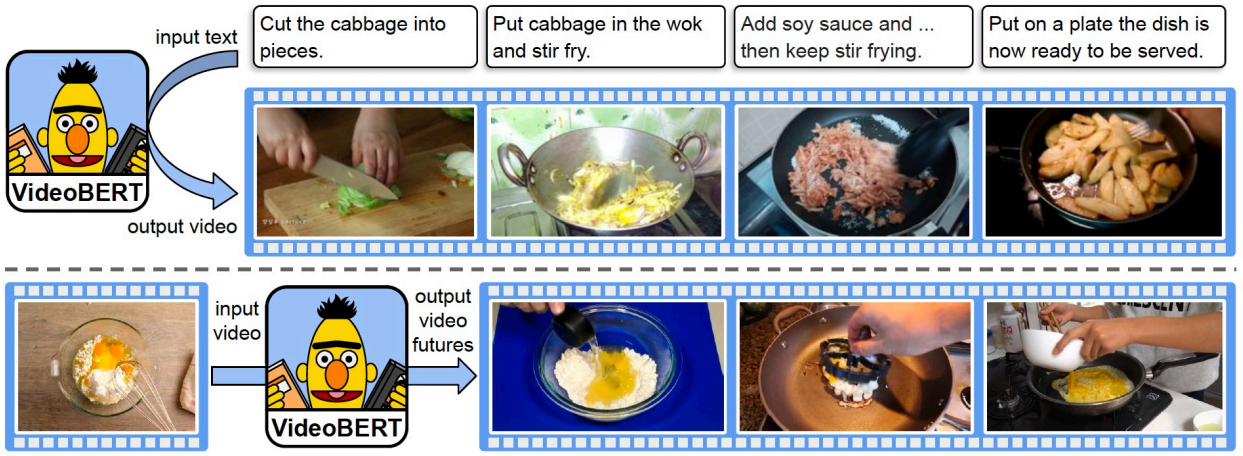
S3D: cut the tomatoes into thin slices


GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan

Text-to-video generation and future forecasting



Outline

- VideoBERT
- VL-BERT

W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei and J. Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. ICLR 2020.

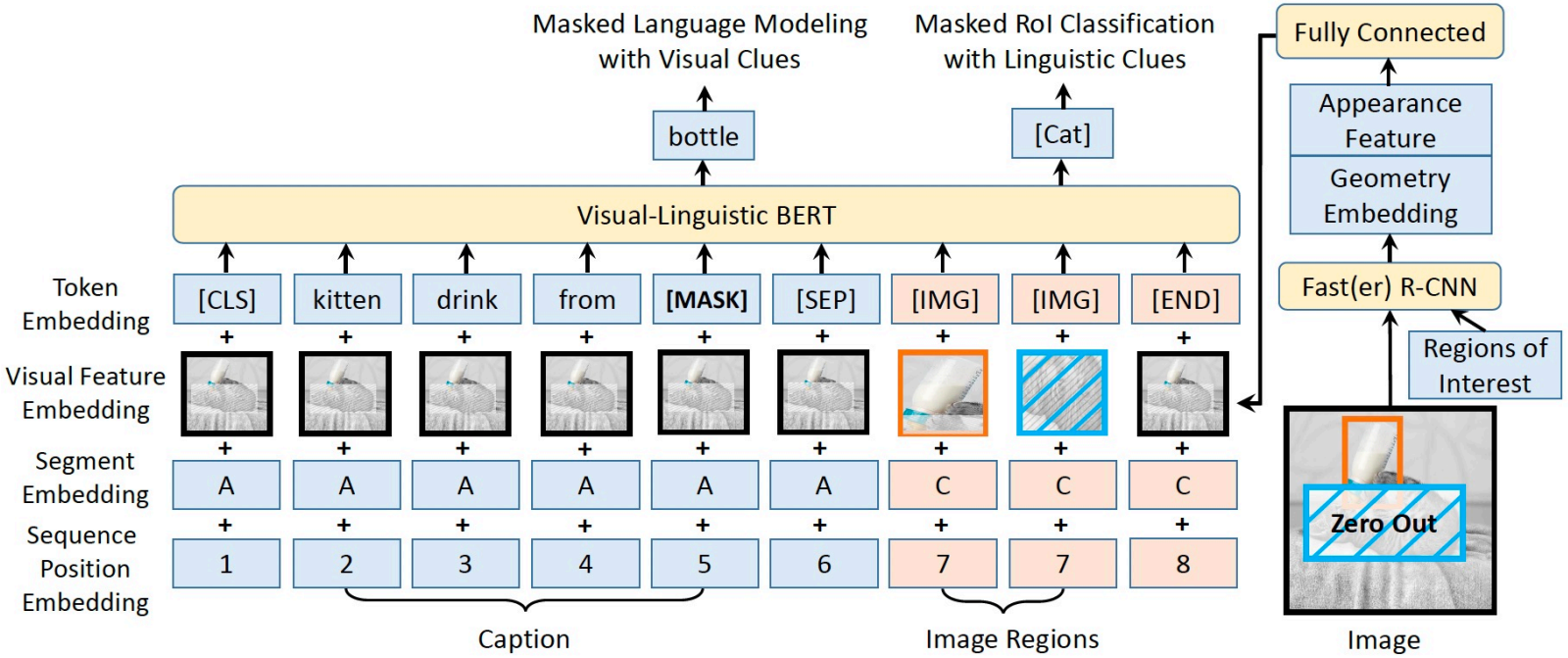
VL-BERT: Visual-Linguistic BERT

A BERT-based model for visual-linguistic tasks for images (not videos)

- ROIs from an image correspond to visual words
- There are many concurrent models, including ViLBERT, LXMERT, VisualBERT, B2T2 and Unicoder-VL
- Mainly evaluated for the tasks of image captioning, VQA and VCR (commonsense)

Architecture

Based on BERT, add some new elements for visual content



Embedding

Token embedding

- Similar to BERT, but add a special [IMG] token for visual element

Visual feature embedding

- Concatenate Fast R-CNN features + geometry features for ROIs
- For non-visual elements, use the feature for whole image
- Geometry feature is obtained as done in Relation Networks (4D coordinates of top-left and bottom-right corner to 2048 dim vector by (co)sine functions)

Segment embedding: A, B for text, C for visual

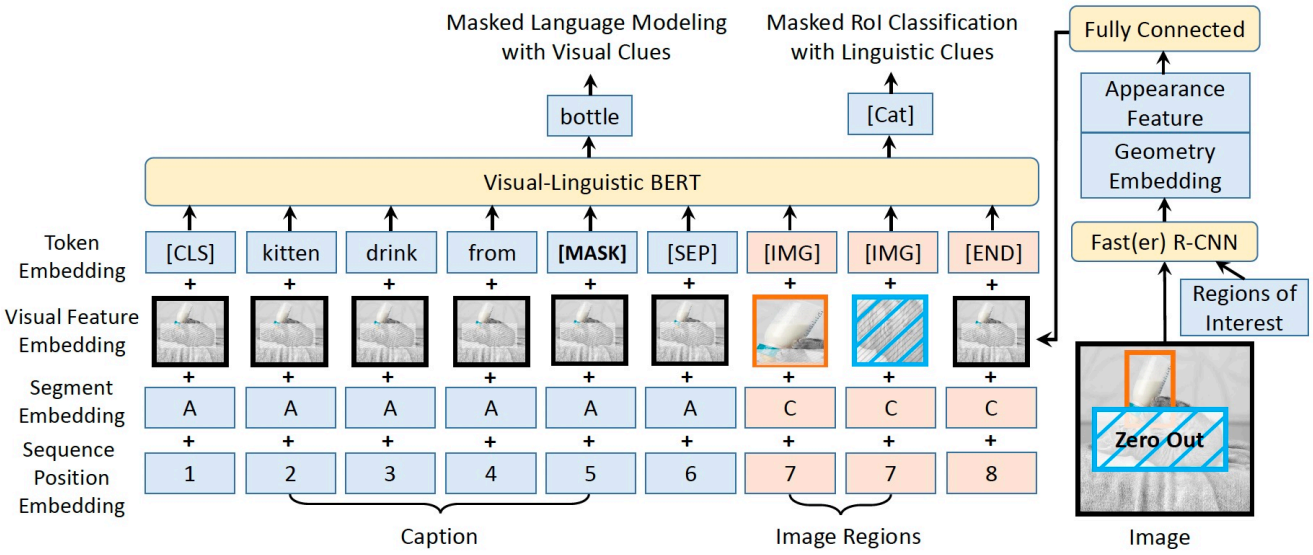
- e.g. Captioning: A for caption, VQA: A for question and B for answer

Position embedding

- Same with BERT for text, and use the same embedding for visual elements (because there is no order between them)

Two Pre-training Tasks

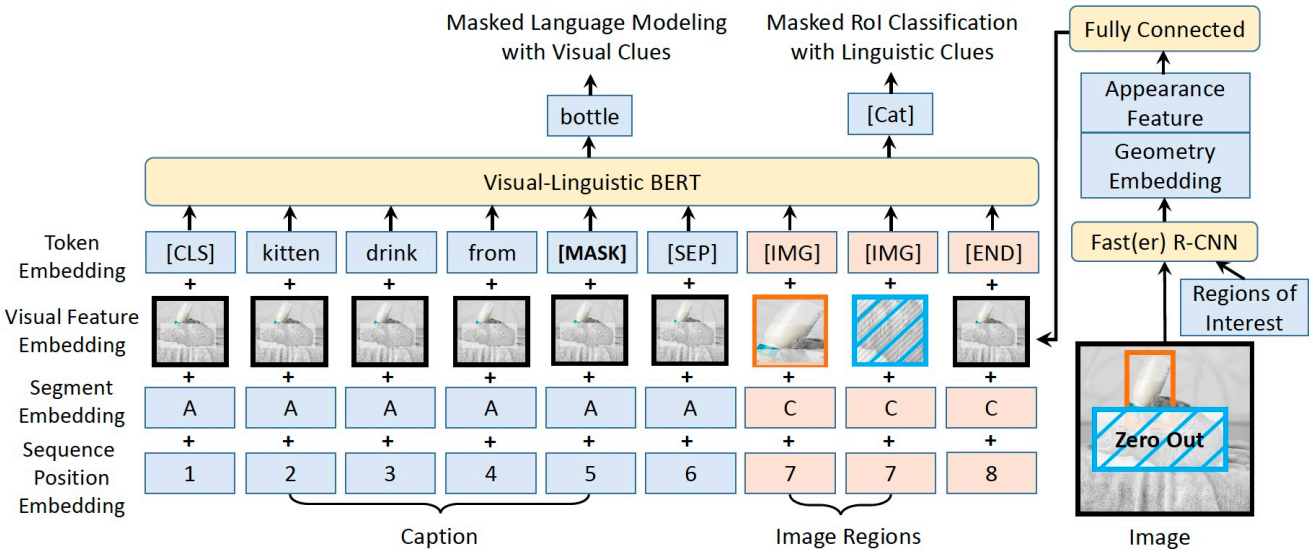
1. Masked language modeling with visual clues
   - (Similar to BERT) Each word is randomly masked with a probability of 15% with full visual cues
   - e.g. kitten drinking from [MASK]: without input image, bowl, spoon and bottle all are fine
   - Learn to align the visual and linguistic contents: [mask] – bottle from the ROI



Two Pre-training Tasks

2. Masked ROI classification with linguistic clues

- (Similar to BERT) Each ROI in image is randomly masked with a probability of 15% with full linguistic cues
- e.g. The ROI corresponding to cat is masked out, its category cannot be predicted from image
- Learn to align the visual and linguistic contents: [mask] – kitten from the word



Implementation

Pre-training

- BookCorpus and English Wikipedia as text-only corpus (same with BERT) + Conceptual Captions as visual-linguistic corpus
- Initialize the parameters using BERT

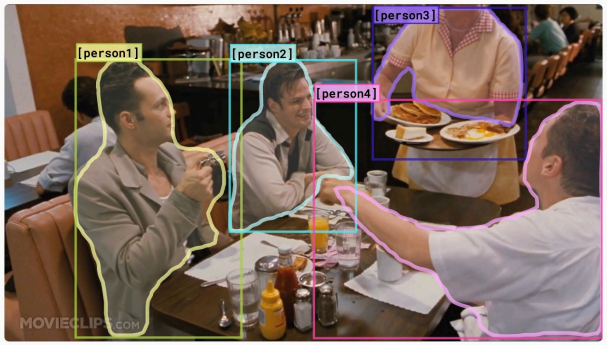
ROIs

- Use pre-trained Faster R-CNN
- Use ROIs with detection scores higher than 0.5
- (Max, Min) # of ROs = (100, 10)

Downstream Tasks: 1. VCR

Task

- Given an image, a list of ROIs and a question at cognition level, select one out of four candidates



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Rationale: I think so because...

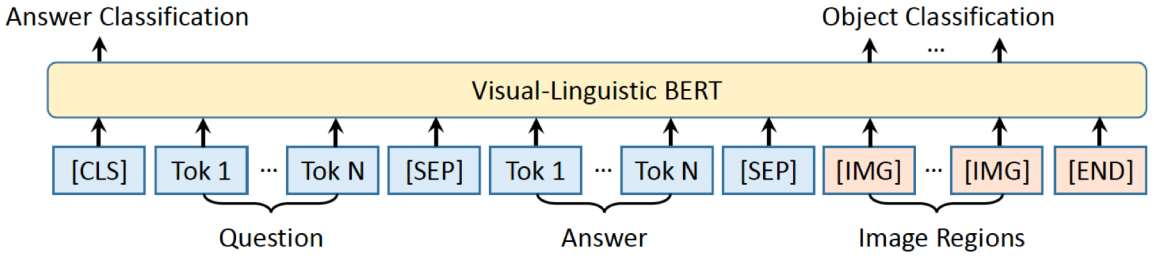
a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

https://visualcommonsense.com/

Downstream Tasks: 1. VCR

Task

- Question: input, Answer: output (Q→A, QA → R, Q → AR)



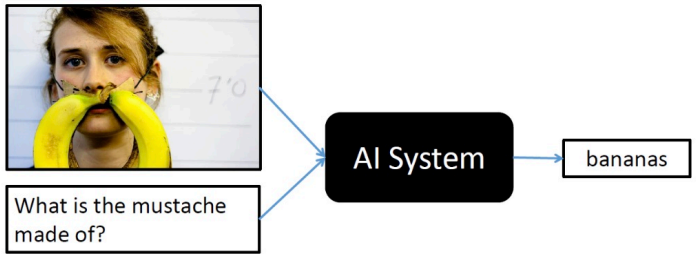
Results: Best single model in the leaderboard

Model	Q → A		QA → R		Q → AR	
	val	test	val	test	val	test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
ViLBERT (Lu et al., 2019)†	72.4	73.3	74.5	74.6	54.0	54.8
VisualBERT (Li et al., 2019b)†	70.8	71.6	73.2	73.2	52.2	52.4
B2T2 (Alberti et al., 2019)†	71.9	72.6	76.0	75.7	54.9	55.0
VL-BERT$_{BASE}$ w/o pre-training	73.1	-	73.8	-	54.2	-
VL-BERT$_{BASE}$	73.8	-	74.4	-	55.2	-
VL-BERT$_{LARGE}$	75.5	75.8	77.9	78.4	58.9	59.7

Downstream Tasks: 2. VQA v2.0

Task (VQA v2.0)

- Given an image and a question, select an answer out of 3129 candidates
- Input format: same with VCR



Results

- LXMERT is trained more VQA datasets with COCO and Visual Genome

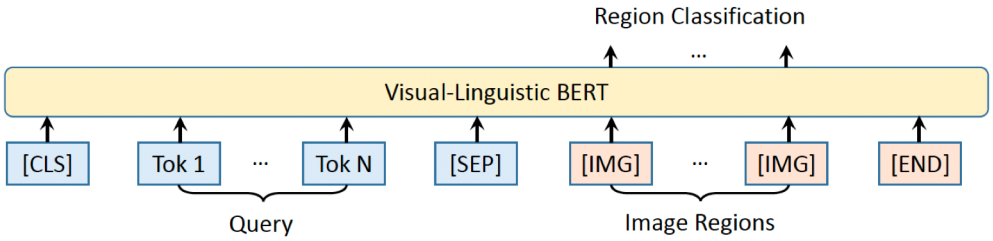
Model	test-dev	test-std
BUTD (Anderson et al., 2018)	65.32	65.67
ViLBERT (Lu et al., 2019)†	70.55	70.92
VisualBERT (Li et al., 2019b)†	70.80	71.00
LXMERT (Tan & Bansal, 2019)†	72.42	72.54
VL-BERT$_{BASE}$ w/o pre-training	69.58	-
VL-BERT$_{BASE}$	71.16	-
VL-BERT$_{LARGE}$	71.79	72.22

https://visualqa.org/

Downstream Tasks: 3. REC

Task (Referring expression Comprehension)

- Given a referring phrase, localize the object in an image
- Input format: binary classification



RefCOCO+

guy in yellow dirbbling ball

Results

Model	Ground-truth Regions			Detected Regions		
	val	testA	testB	val	testA	testB
MAttNet (Yu et al., 2018)	71.01	75.13	66.17	65.33	71.62	56.02
ViLBERT (Lu et al., 2019)†	-	-	-	72.34	78.52	62.61
VL-BERT$_{BASE}$ w/o pre-training	74.41	77.28	67.52	66.03	71.87	56.13
VL-BERT$_{BASE}$	79.88	82.40	75.01	71.60	77.72	60.99
VL-BERT$_{LARGE}$	80.31	83.62	75.45	72.59	78.57	62.30

http://vision2.cs.unc.edu/refer/

Conclusion

One of the most active research areas

- Competition with more computing resources and more data

Self-supervised learning

- How can we better use unlabeled (cheap) data ?

Task-independent model

- Task-specific model requires its own data and labels