VL-BERT: Pre-training of Generic Visual-Linguistic Representations

Jifeng Dai

With Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, and Furu Wei

Published at ICLR 2020

Pre-training of Generic Representations: A Hallmark of Deep Network's Success

- Prior to the era of deep networks
  - Diverse hand-crafted features & designs
  - Un-shareable feature representations among different tasks

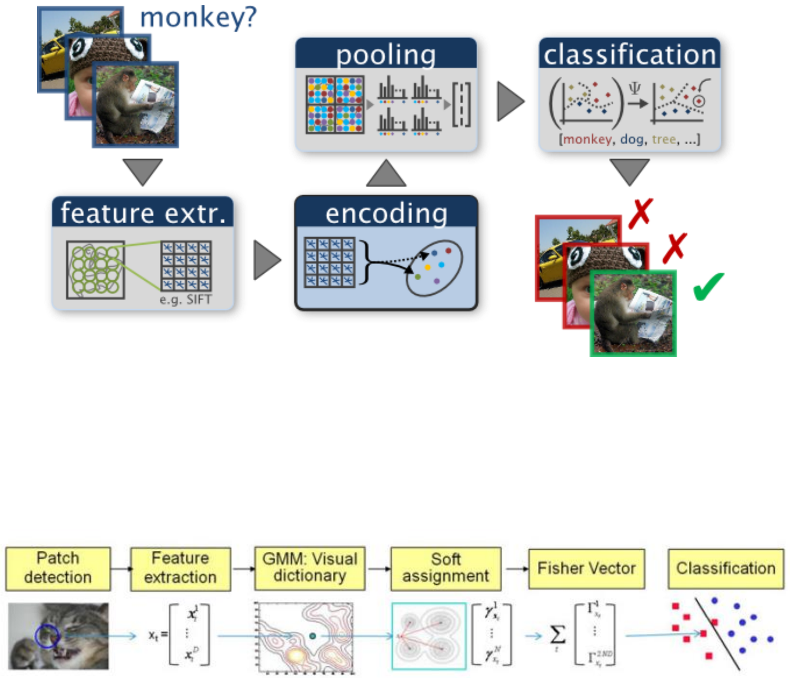
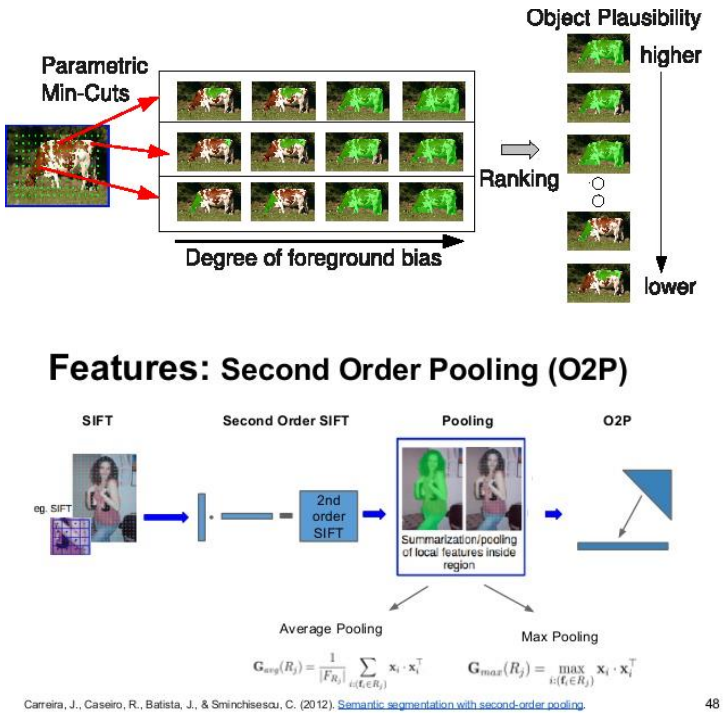
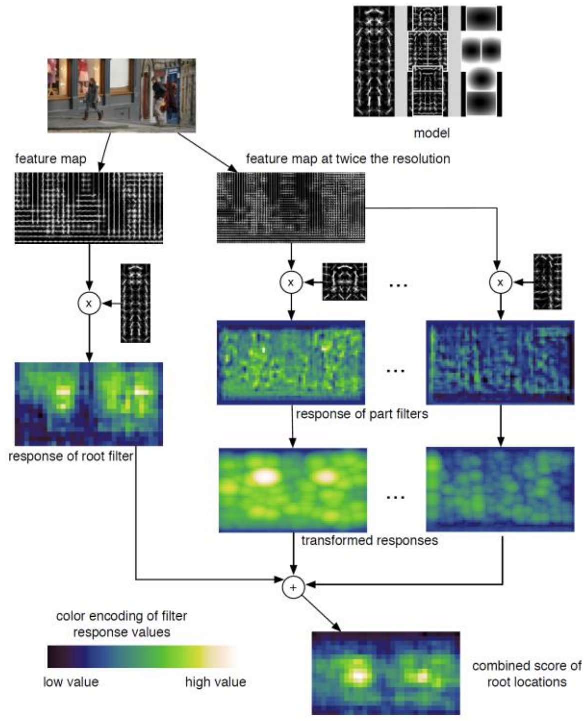


Image classification
(Fisher Vectors [CVPR 2010], etc.)



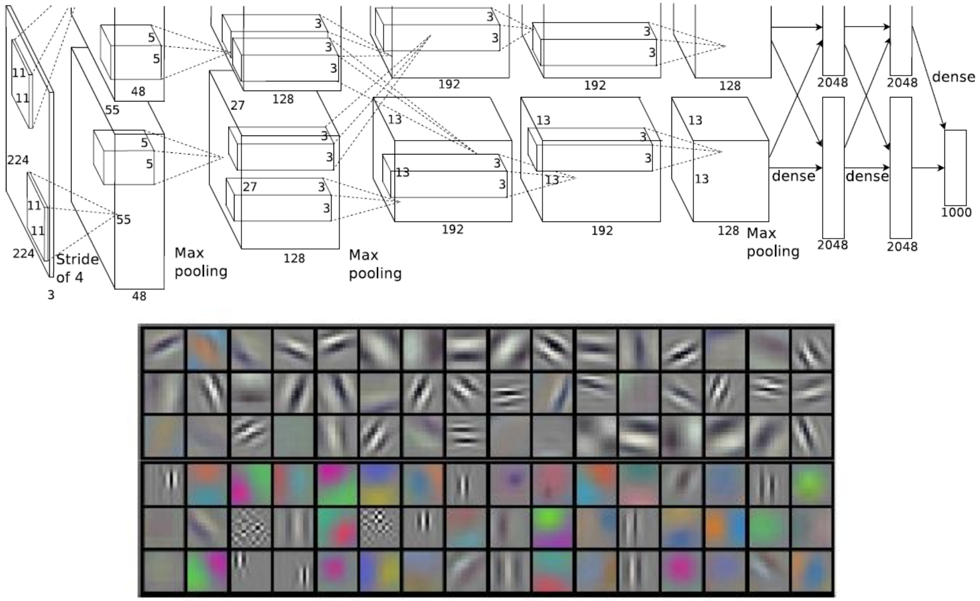
Semantic segmentation
(O2P [ECCV 2012], etc.)



Object detection
(DPM [TPAMI, 2009], etc.)

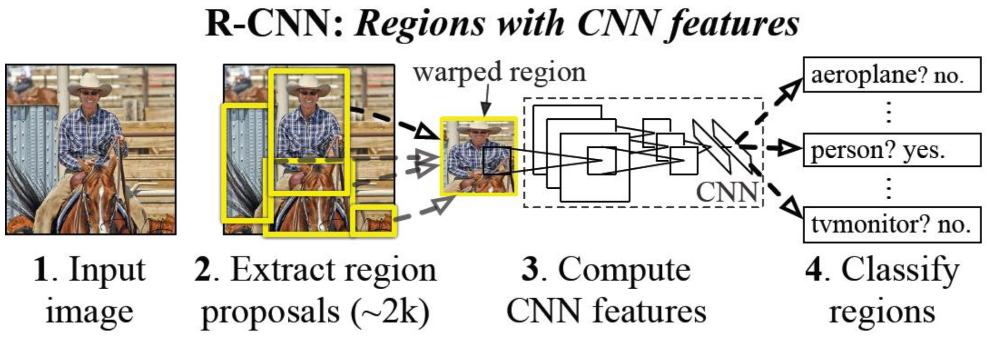
Pre-training of Generic Representations: A Hallmark of Deep Network's Success

- Renaissance of deep networks in computer vision
  - Generic backbone + Task-specific headers
  - Pre-trainable generic representations

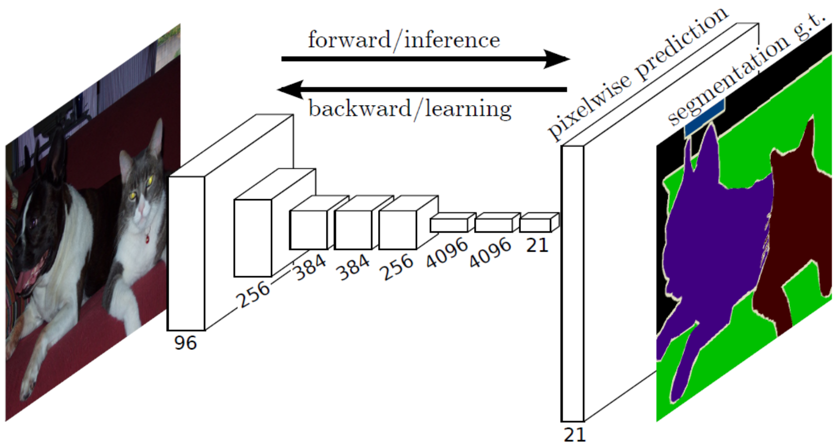


AlexNet [NIPS 2012] for image classification

Pre-training & finetuning



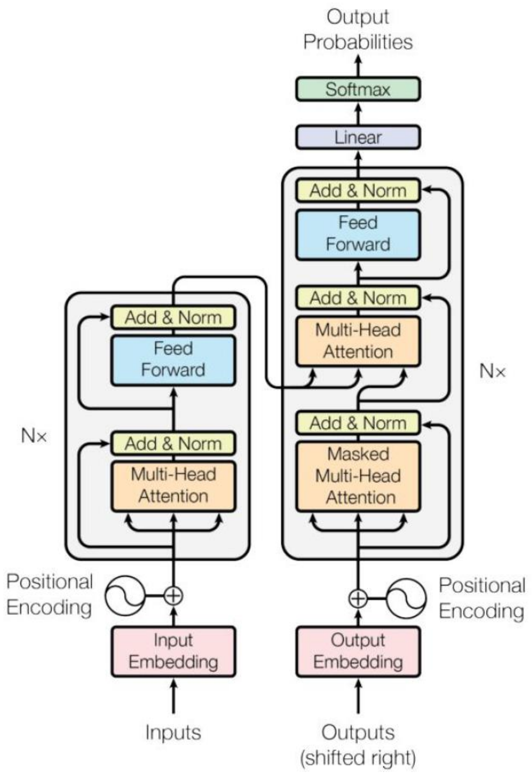
R-CNN [CVPR 2014] for object detection



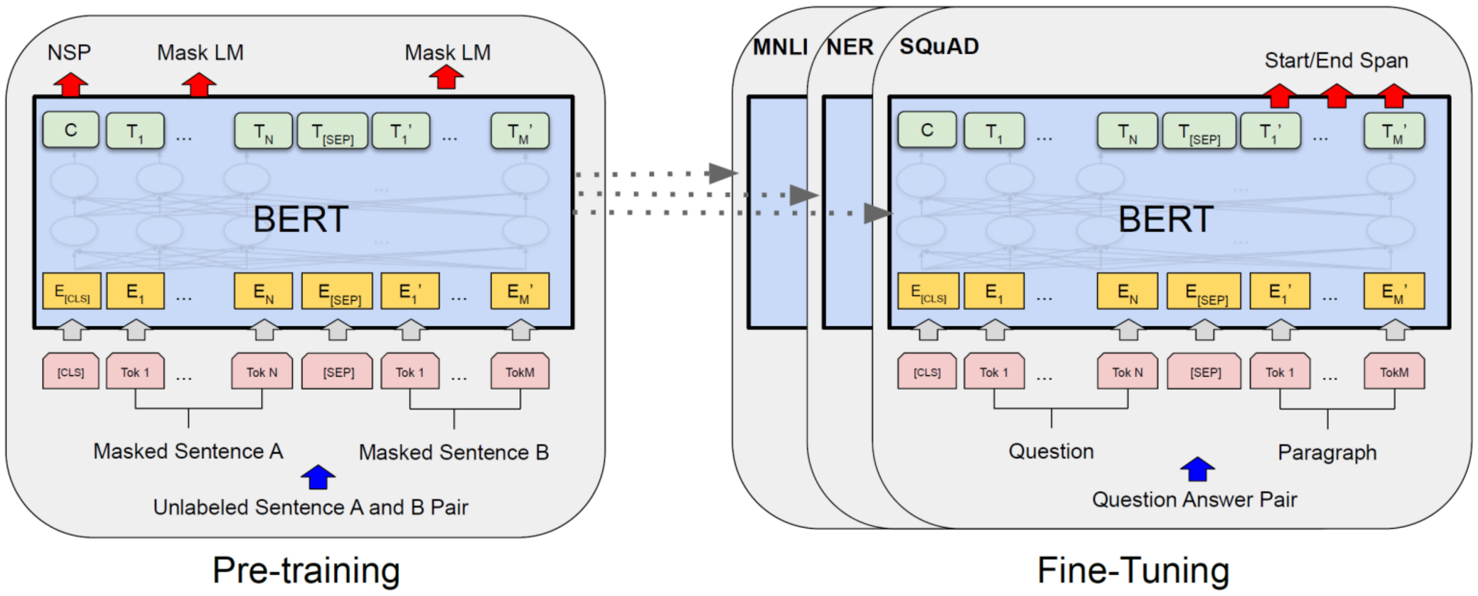
FCN [CVPR 2015] for semantic segmentation

Pre-training of Generic Representations: A Hallmark of Deep Network's Success

- Recent leap forward in Natural Language Processing (NLP)



Transformer [NIPS 2017]



BERT [NAACL 2019]

Pre-training for Visual-Linguistic Tasks?

- Various visual-linguistic tasks

Make the V in VQA Matter [CVPR 2017]

"man in black shirt is playing guitar."

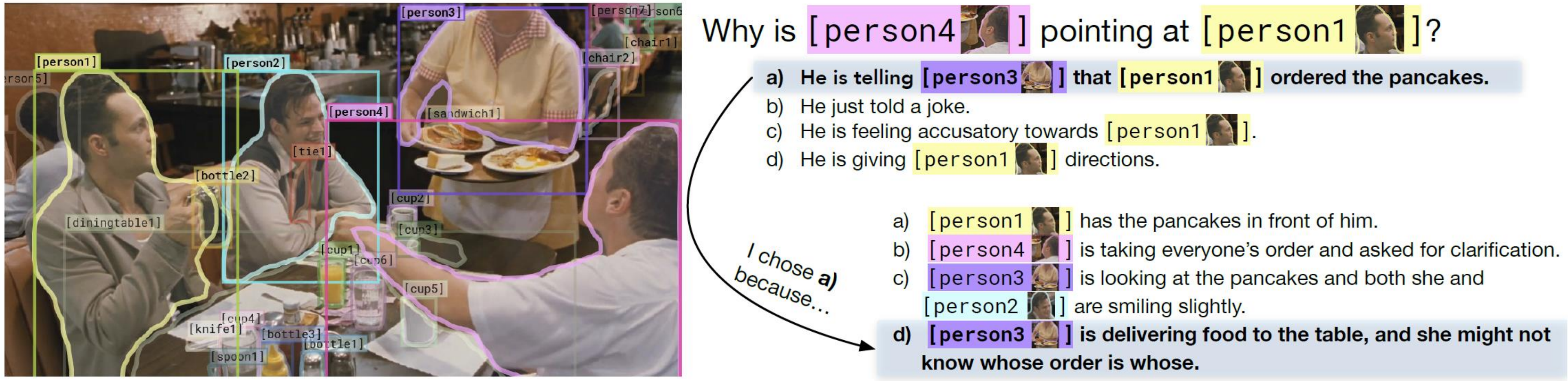
"construction worker in orange safety vest is working on road."

Image captioning

blurry person with sleeveless and sitting

man in full view in all black

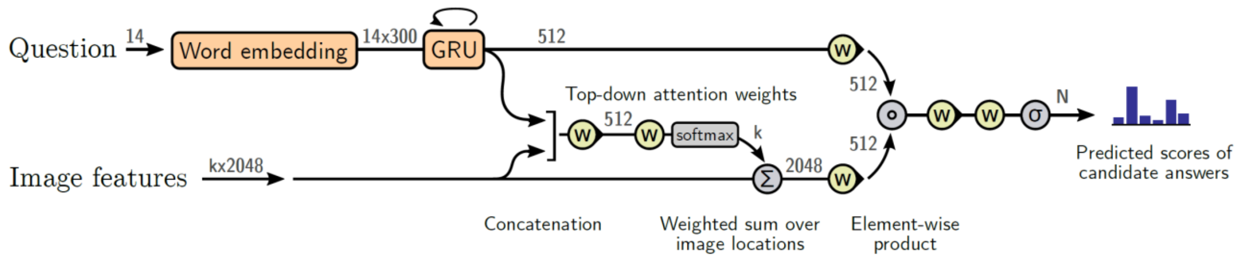
Modeling context in referring expressions [ECCV 2016]



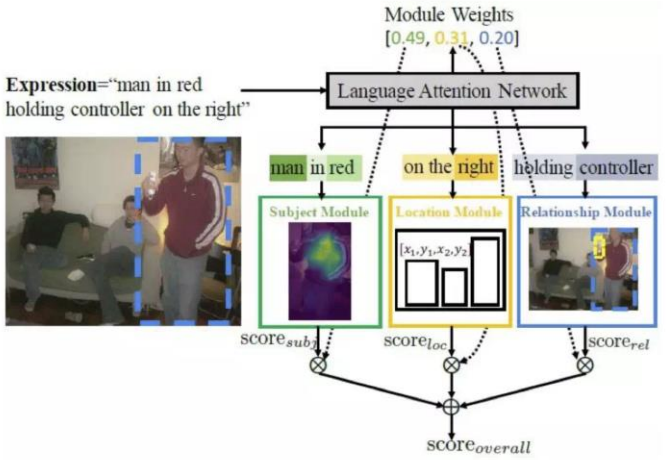
From recognition to cognition: visual commonsense reasoning [CVPR 2019]

Pre-training for Visual-Linguistic Tasks?

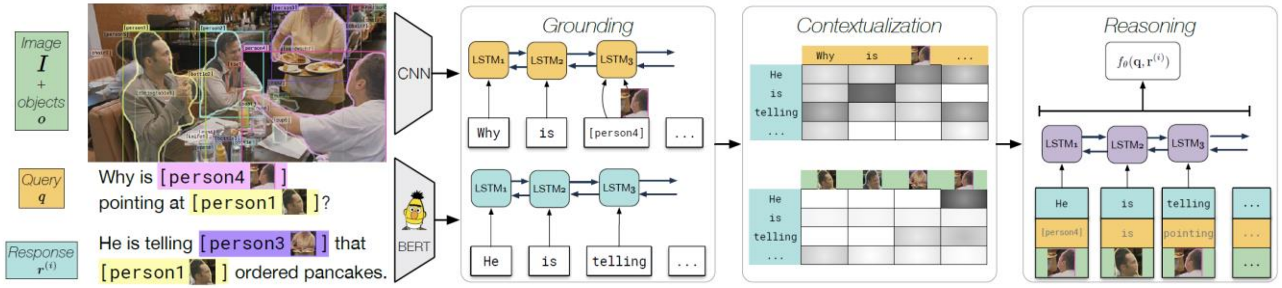
- Numerous task-specific networks
  - Ad-hoc design, un-shareable representations
  - Key goal: to aggregate the multi-modal info



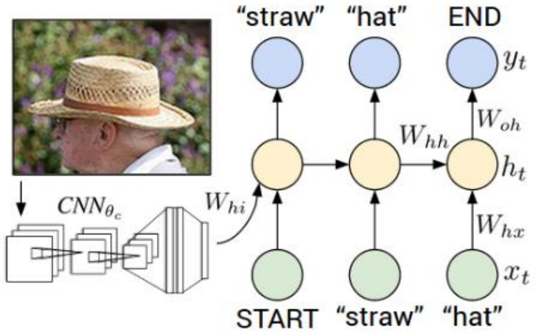
BUTD for VQA [CVPR 2018]



MAttNet for RefCOCO+ [CVPR 2018]



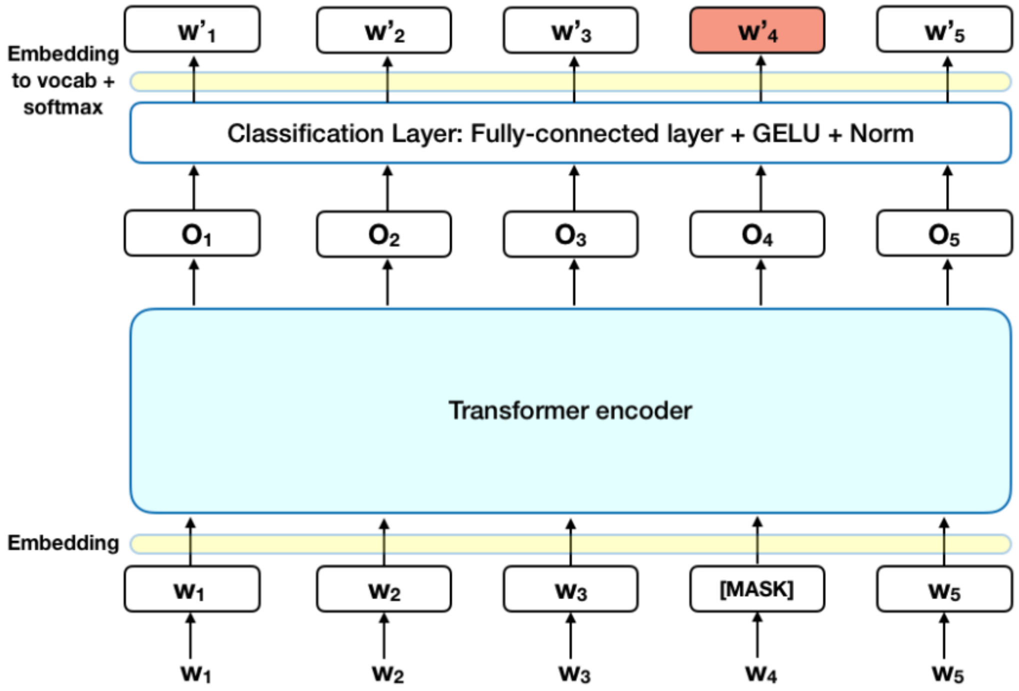
R2C for VCR [CVPR 2019]



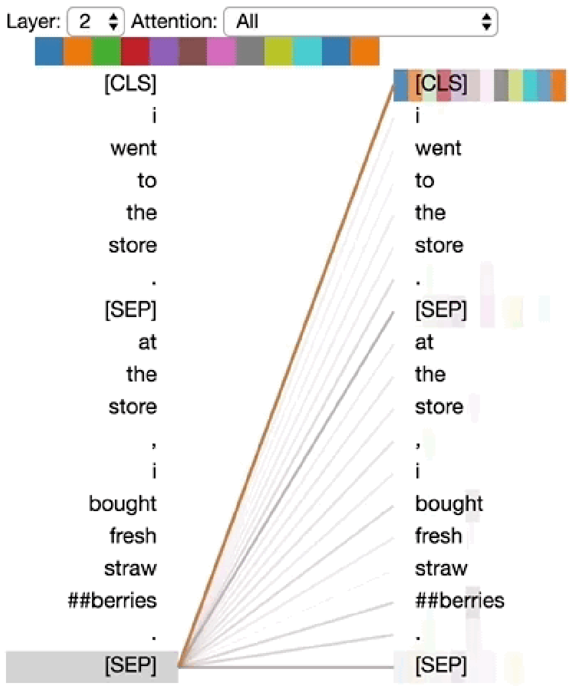
DVSA for image captioning [CVPR 2015]

Revisit BERT Model

- Flexible and powerful in aggregating and aligning word features
  - Self-contained embeddings + Transformer attention + masked language modeling



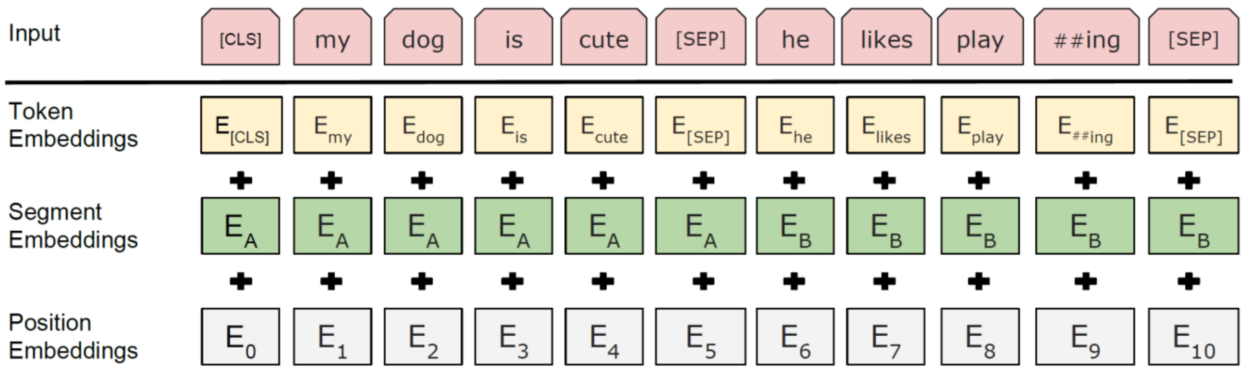
Masked language modeling in BERT



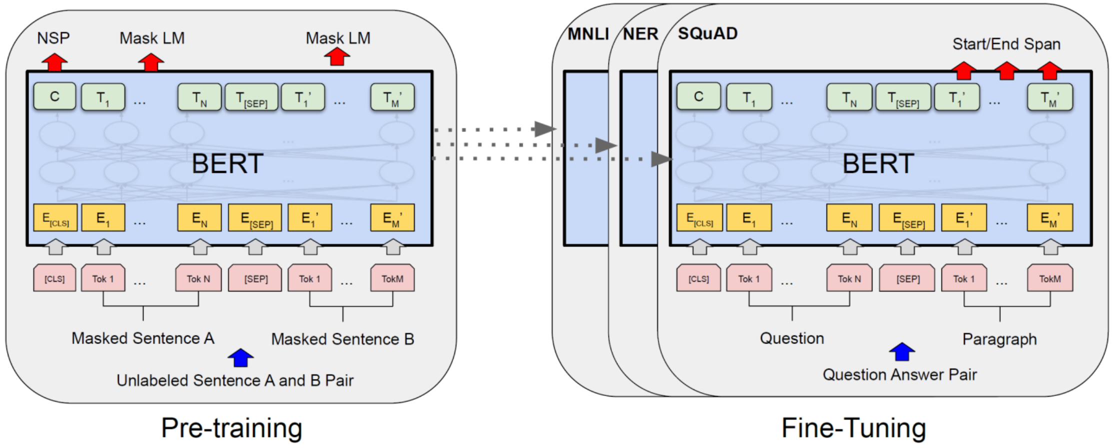
Attention weights in BERT

Revisit BERT Model

- Flexible and powerful in aggregating and aligning word features
  - Self-contained embeddings + Transformer attention + masked language modeling



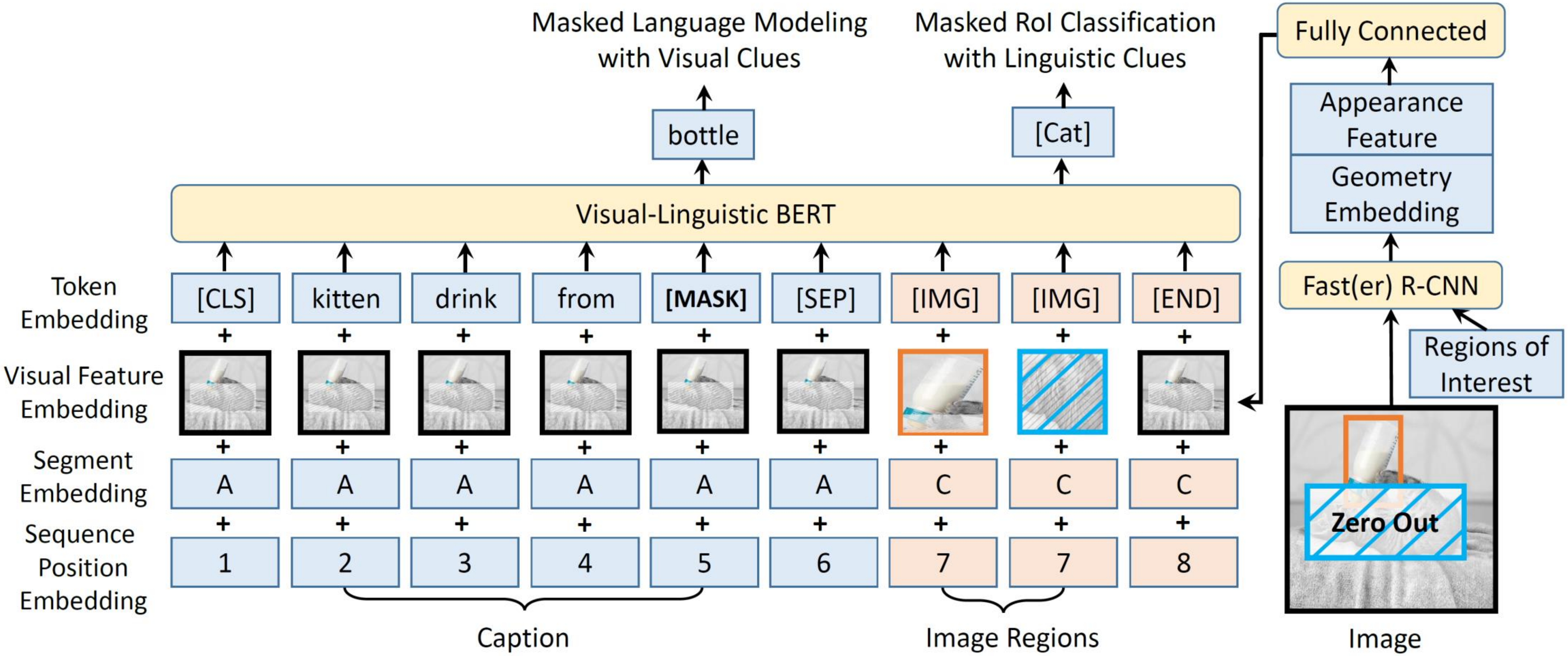
Embedded features in BERT



Pre-training & finetuning of BERT

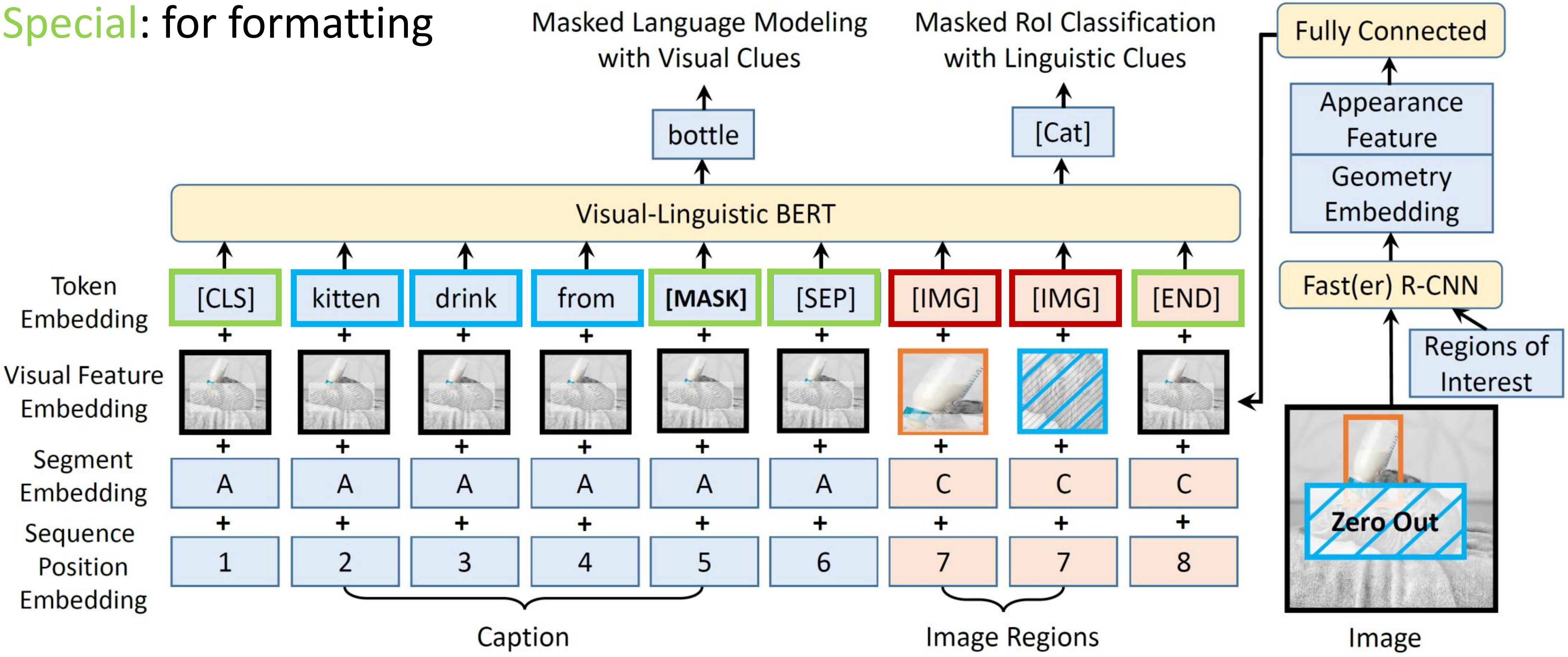
VL-BERT: Pre-training of Generic Visual-Linguistic Representations

- Model architecture
  - Modified from original BERT to accommodate the visual contents



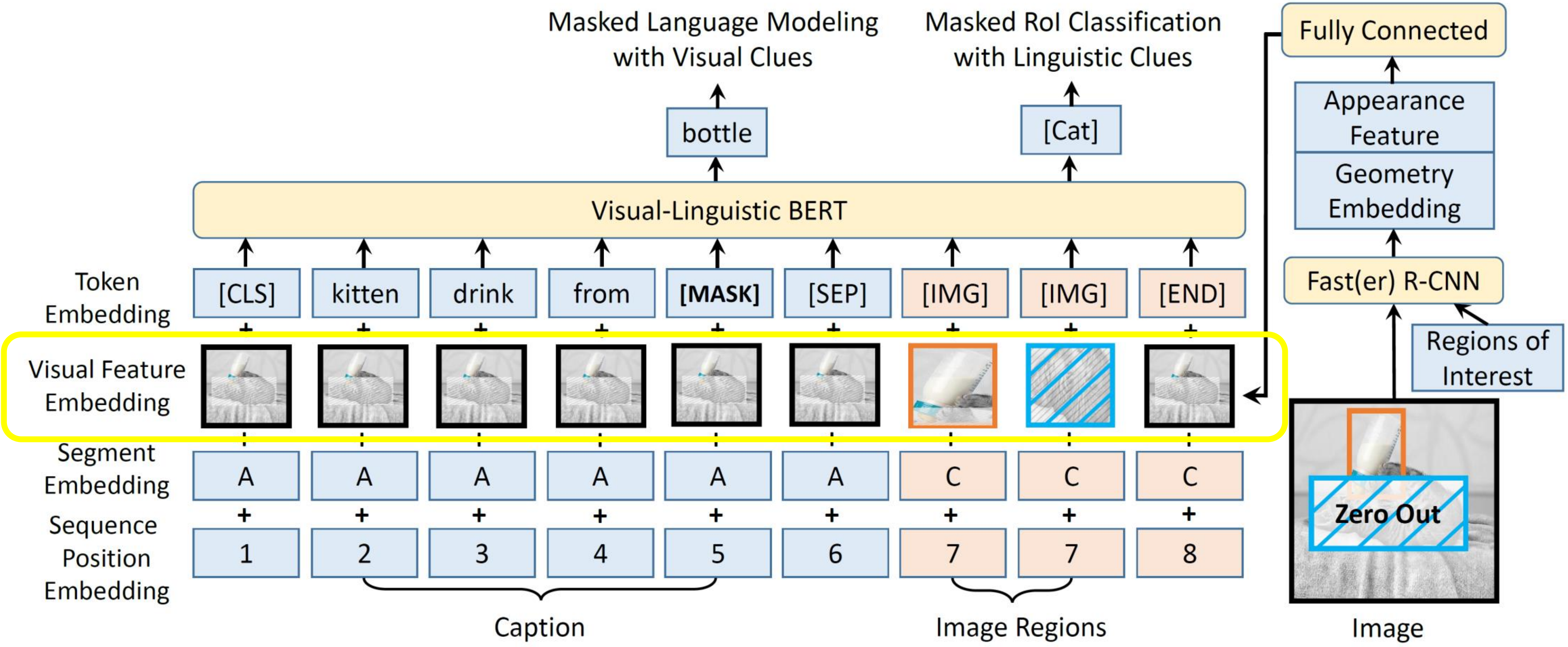
Model Architecture of VL-BERT

- Input elements
  - Visual: region-of-interests (RoIs) in image
  - Linguistic: words in sentences
  - Special: for formatting



Model Architecture of VL-BERT

- Feature embeddings
  - Token, segment, and sequence position embeddings are the same as BERT
  - *Visual feature embeddings* are newly introduced for each element



Pre-training VL-BERT

- Pre-training on both visual-linguistic and text-only corpus
  - Conceptual Captions: ~3.3M image caption pairs, harvested from web, simple clauses
  - BooksCorpus & English Wiki: long and complex sentences, utilized in pre-training BERT

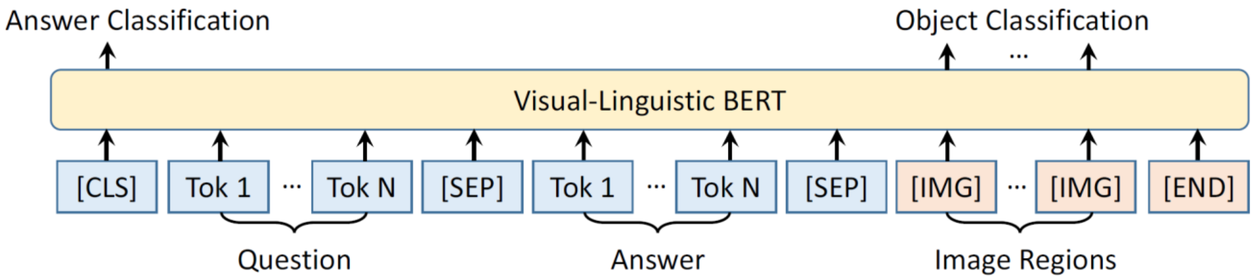
Conceptual Captions [ACL 2018]

BooksCorpus [ICCV 2015] & English Wiki

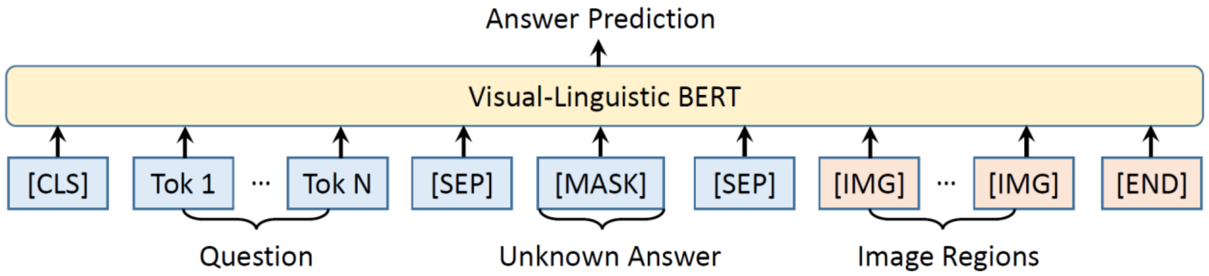
Pre-training VL-BERT

- Pre-training on Conceptual Captions
  - Input format: <Caption, Image>
  - Task #1: Masked Language Modeling with Visual Clues
  - Task #2: Masked RoI Classification with Linguistic Clues
- Pre-training on BooksCorpus & English Wiki
  - Input format: <Text, Null>
  - Task: Standard Masked Language Modeling as in BERT
- End-to-end training, with all the parameters updated

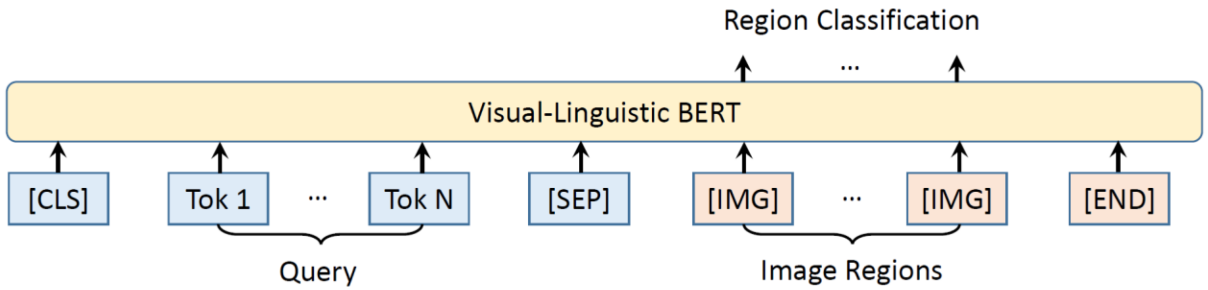
Fine-tuning VL-BERT on Downstream Tasks



(a) Input and output format for Visual Commonsense Reasoning (VCR) dataset



(b) Input and output format for Visual Question Answering (VQA) dataset



(c) Input and output format for Referring Expression task on RefCOCO+ dataset

Related Work

- Video BERT [ICCV 2019]
  - First work seeking to conduct pre-training for visual-linguistic tasks
  - Abrupt clustering of video clips, considerable loss in visual content info
  - Applied on videos only, of linear structure same as sentences

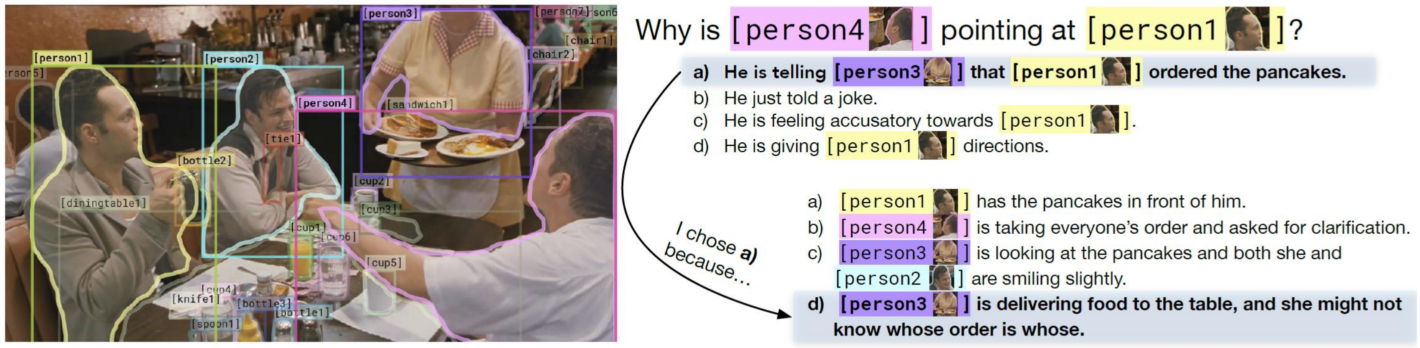
- Concurrent works on image-based visual-linguistic tasks
  - Indicating the importance of the problem
  - Noticeable difference between VL-BERT and other concurrent works
    - We found the task of Sentence-Image Relationship Prediction used in all other concurrent works is of no help in pre-training visual-linguistic representations.
    - Pre-training on both visual-linguistic and text-only datasets. We found such joint pre-training improves the generalization over long and complex sentences.
    - Improved tuning of the visual representation.

	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Tan & Bansal, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018) + BooksCorpus (Zhu et al., 2015) + English Wikipedia	1) masked language modeling 2) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

Experiments

- Visual Commonsense Reasoning (VCR)



Model	Q → A val	Q → A test	QA → R val	QA → R test	Q → AR val	Q → AR 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$_{\text{BASE}}$ w/o pre-training	73.1	-	73.8	-	54.2	-
VL-BERT$_{\text{BASE}}$	73.8	-	74.4	-	55.2	-
VL-BERT$_{\text{LARGE}}$	75.5	75.8	77.9	78.4	58.9	59.7

Table 1: Comparison to the state-of-the-art methods with single model on the VCR dataset. † indicates concurrent works.

Experiments

- Visual Question Answering (VQA)

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

Table 2: Comparison to the state-of-the-art methods with single model on the VQA dataset. † indicates concurrent works.

Experiments

- RefCOCO+

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$_{\text{BASE}}$ w/o pre-training	74.41	77.28	67.52	66.03	71.87	56.13
VL-BERT$_{\text{BASE}}$	79.88	82.40	75.01	71.60	77.72	60.99
VL-BERT$_{\text{LARGE}}$	80.31	83.62	75.45	72.59	78.57	62.30

Table 3: Comparison to the state-of-the-art methods with single model on the RefCOCO+ dataset. † indicates concurrent work.

Experiments

- Ablation study

Settings	Masked Language Modeling with Visual Clues	Masked RoI Classification with Linguistic Clues	Sentence-Image Relationship Prediction	with Text-only Corpus	Tuning Fast R-CNN	VCR Q→A val	VCR QA→R val	VQA test-dev	RefCOCO+ Detected Regions val
w/o pre-training						72.9	73.0	69.5	62.7
(a)	✓					72.9	73.1	71.0	69.1
(b)	✓	✓				73.0	73.1	71.1	70.7
(c)	✓	✓	✓			72.2	72.4	70.3	69.5
(d)	✓	✓		✓		73.4	73.8	71.1	70.7
VL-BERT$_{\text{BASE}}$	✓	✓		✓	✓	73.8	73.9	71.2	71.1

Table 4: Ablation study for VL-BERT$_{\text{BASE}}$ with 0.5× fine-tuning epochs.

Conclusion

- VL-BERT, a new pre-trainable generic representation for visual-linguistic tasks
  - Utilization of Transformer model as the backbone, instead of ad-hoc task-specific modules
  - Pre-trainable on large-scale visual-linguistic & text-only corpus
- Future work
  - Better pre-training tasks, to benefit more downstream tasks
  - More powerful generic backbone for visual-linguistic tasks

Q&A