VL-BERT： Pretraining of Generic Visual-Linguistic Representations

Speaker: Gong, Qiqi

Basica Info.

- Conf.:ICLR 2020
- Affiliation: USTC (中科大), MSRA

Background Knowledges

- V+L Tasks:
  - Visual contents
  - Language semantics
  - **Cross-modal alignments**
- Pretraining：
  - Pre-training for Visual task: Pretrain classifier on ImageNet (since 2009)
  - Pre-training for Linguistic task: Pretrain transformer on large scale corpus(语料库) for **masked language modeling (MLM)** (since 2017)
  - How to pretrain on V+L task?

Background Knowledges

- *Introduction to MLM:
  - Special tokens in NLP:
    - [CLS]: Begining of a sentence
    - [SEP]: Separation of sentences
    - [MASK]: Masking words with a probability
  - Aim: To predict masked words from given words
  - e.g.:

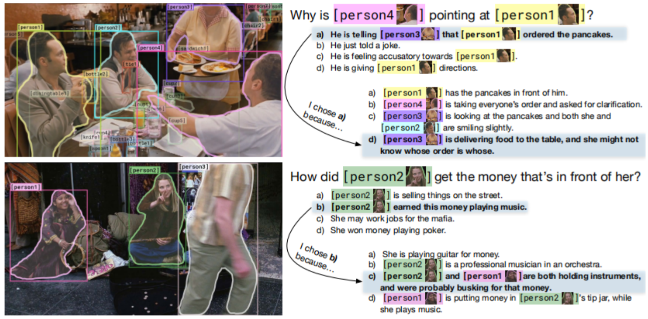
    Input: My dog is [Mask]

    Output: My dog is hairy

Background Knowledges

- Downstream V&L Tasks:
  - Visual Question Answering (VQA):
    - Input:<Question, Answer([Mask]), Image>
    - Dataset: VQA 2.0
      - Train: 83k images, 444k questions
      - Val: 41k images, 214k questions
      - Pick corresponding answer from 3129 answers (BUTD,2018论文第9页)
  - Visual Commensense Reasoning (VCR):
    - Input: <Question,Answering,Image>
    - Dataset: VCR
      - Train: 80k images, 213k questions
      - Val: 10k images, 27k questions
      - Pick the right answer out of four ones and provide rationale explanation

What color are her eyes?
What is the mustache made of?



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.

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 and both she and [person2] are smiling slightly.
d) **[person3] is delivering food to the table, and she might not know whose order is whose.**

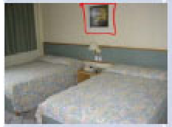
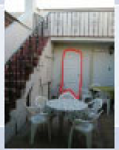
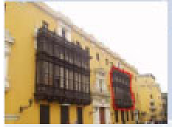
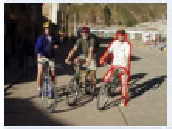
How did [person2] get the money that's in front of her?

a) [person2] is selling things on the street.
b) **[person2] earned this money playing music.**
c) She may work jobs for the mafia.
d) She won money playing poker.

a) She is playing guitar for money.
b) [person2] is a professional musician in an orchestra.
c) **[person2] and [person1] are both holding instruments, and were probably busking for that money.**
d) [person1] is putting money in [person2]'s tip jar, while she plays music.

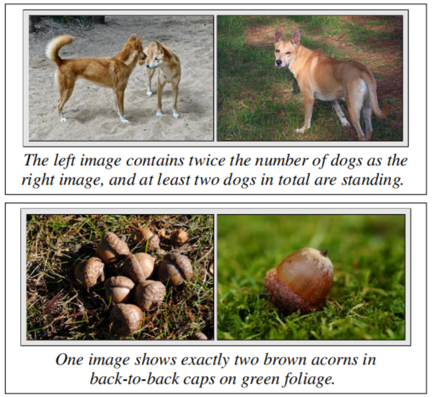
Background Knowledges

- Downstream Tasks:
  - Referring Expression Comprehension (Visual Grounding)
    - Input:<Query, Image>
    - Localize the object with given referring expression
    - Dataset: RefCOCO+
      - 141k expressions for 50k referred objects in 20k images
      - Split into 4 sets: training, val, two sets (one with multiple people, one with multiple objects)

Image	Human Expressions
	picture on the wall picture picture
	Door white door middle white door
	big gated window on right of white section black big window right brown railings on right
	white shirt man white shirt on right man on right
	building on right behind guys blue right building building on right

Background Knowledges

- Downstream Tasks:
  - Other Tasks

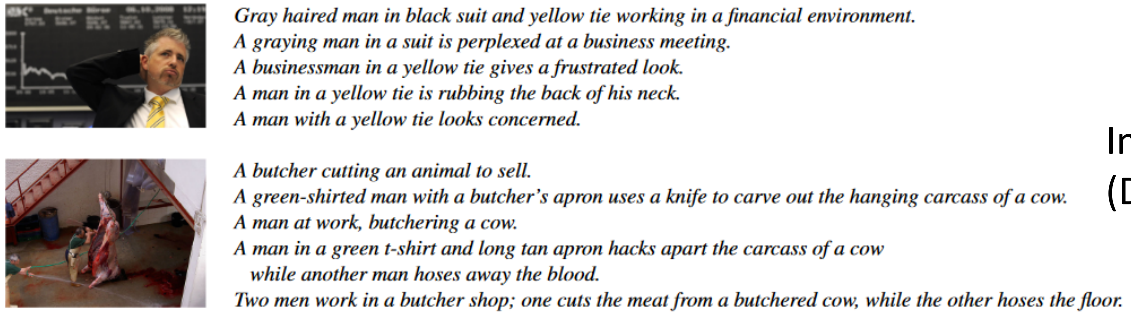


The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

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

Figure 1: Two examples from NLVR2. Each caption is paired with two images.[2] The task is to predict if the caption is True or False. The examples require addressing challenging semantic phenomena, including resolving *twice . . . as* to counting and comparison of objects, and composing cardinality constraints, such as *at least two dogs in total* and *exactly two*.[3]

Natural Language Visual Reasoning for Real (NVLR2)

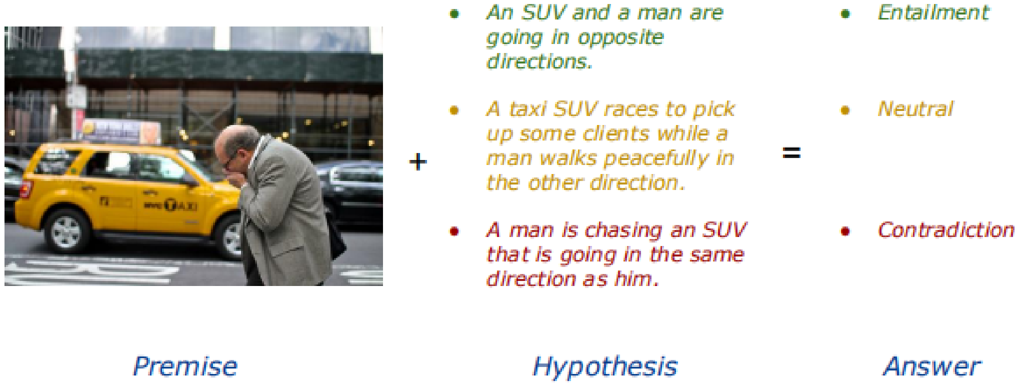


Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

A butcher cutting an animal to sell.
A green-shirted man with a butcher's apron uses a knife to carve out the hanging carcass of a cow.
A man at work, butchering a cow.
A man in a green t-shirt and long tan apron hacks apart the carcass of a cow while another man hoses away the blood.
Two men work in a butcher shop; one cuts the meat from a butchered cow, while the other hoses the floor.

Figure 1: Two images from our data set and their five captions

Image-Text Retrieval (Dataset: Flicker30K)



- *An SUV and a man are going in opposite directions.* — *Entailment*
- *A taxi SUV races to pick up some clients while a man walks peacefully in the other direction.* — *Neutral*
- *A man is chasing an SUV that is going in the same direction as him.* — *Contradiction*

Premise *Hypothesis* *Answer*

Visual Entailment (Dataset: SNLI-VE)

Basic Knowledges

- Questions Conclusion:
  - How to design a model fusing both modalities?
  - How to design a upstream V&L task?
  - How to fine-tune pretrained V&L network to adapt to downstream tasks?

Introduction

- Motivation
  - Previous pratice is to combine base networks for CV and NLP respectively in a task-spefic way
    - Overfitting when data for target is scarce (稀少的)
    - Benefit little from pre-training
  - There lacks pre-trained generic feature representations for V&L tasks
  - (Private Idea) Success of ViT

Introduction

- Contribution
  - A single-stream model for pretraining V&L representations
  - Pre-train on both visual-linguistic and text-only datasets
  - Masking operation for raw images

VL-BERT

- Revisit BERT[bə(r)t] Model
  - Multi-head Atten. -> Residual Connec. -> Feed-forward -> Residual Connec.
  - Characteristics:
    - Irrelevant to the order of input seq. (i.e. permutation(out) == Net(permutation(input)))
    - Position of an element is encoded in its own embedding features by sequence positional embedding

VL-BERT

- Revisit BERT Model
  - Pretraining:
    - Masked language modeling (MLM)

<==> $$\log P(x|\theta) = \frac{1}{Z(\theta)} \sum_{i=1}^{N} \log \phi_i(x|\theta), \ \log \phi_i(x|\theta) = x_i^T f_i(x_{\setminus i}|\theta)_i,$$

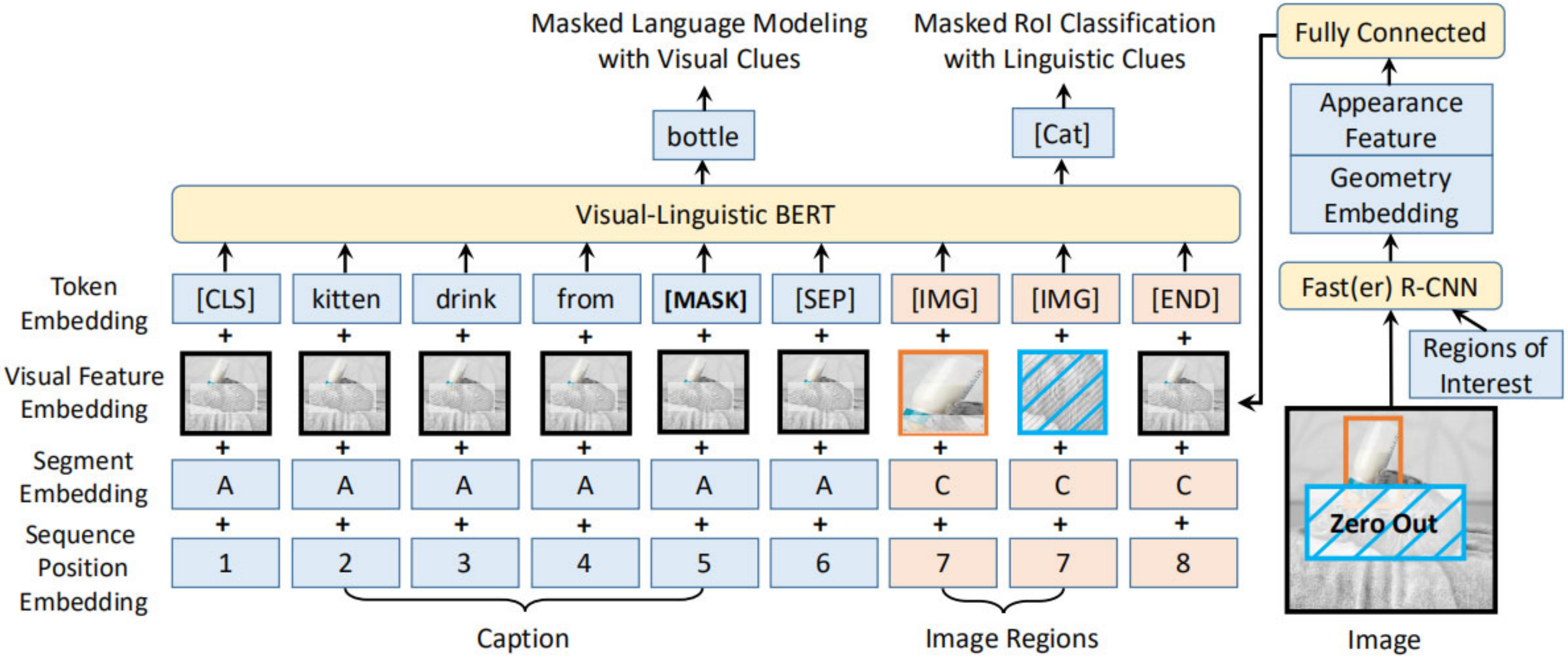
Incurred Loss: $$L_{\text{MLM}}(\theta) = -E_{x\sim D, i\sim\{1,\ldots,N\}} \log \phi_i(x),$$

    - Next Sentence Prediction
      - Two special elements: [CLS] [SEP]

Loss: $$L_{\text{NSP}}(\theta) = -E_{(x,t)\sim D}\left[t \log(g(x_0^L)) + (1-t)\log(1-g(x_0^L))\right],$$

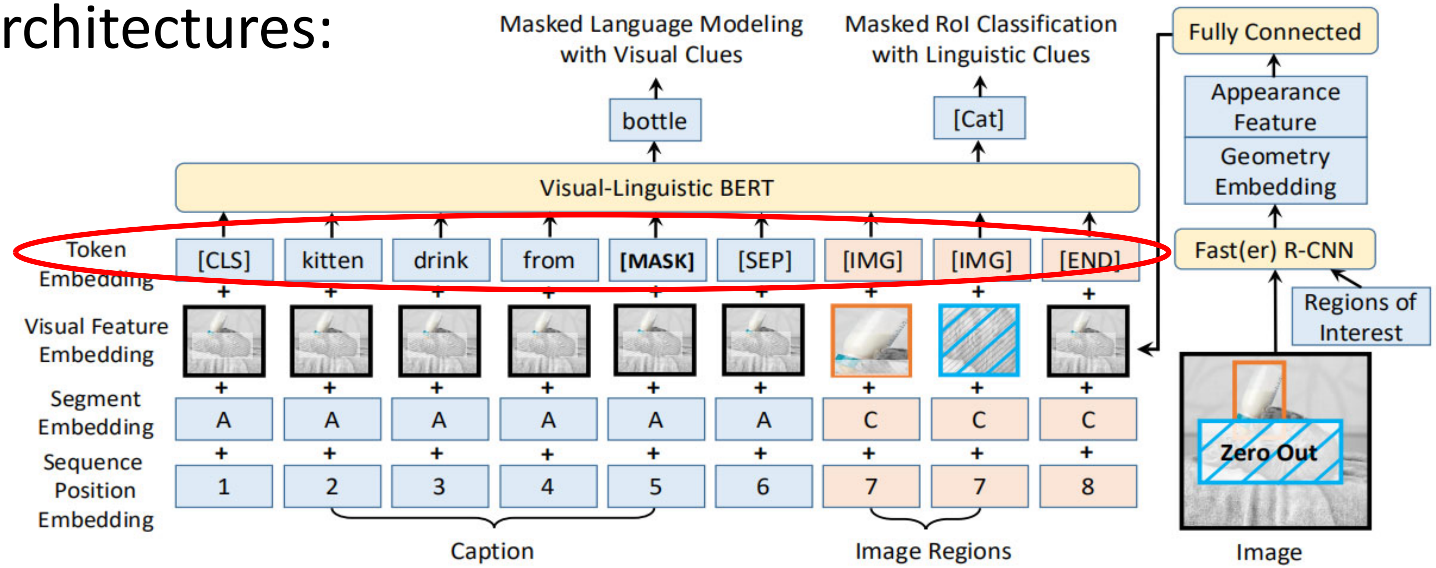
VL-BERT

- Model Architectures
  - Take both visual and linguistic
    - Visual： RoIs generated(annotated) by Faster-RCNN
    - Linguistic: Subwords from input sentences
    - Special elements: For disambiguating different input formats



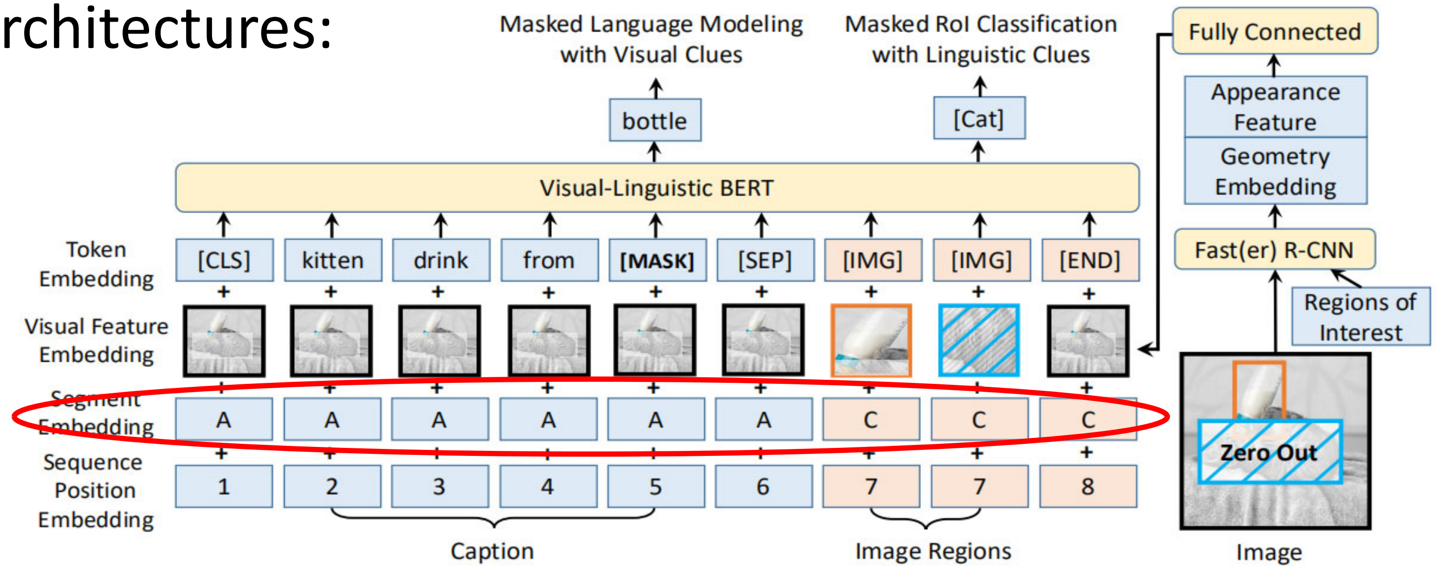
- Model Architectures:
  - Input = Token Embedding + Segment Embedding + Sequence Position Embedding + Visual Feature Embedding

- Model Architectures:



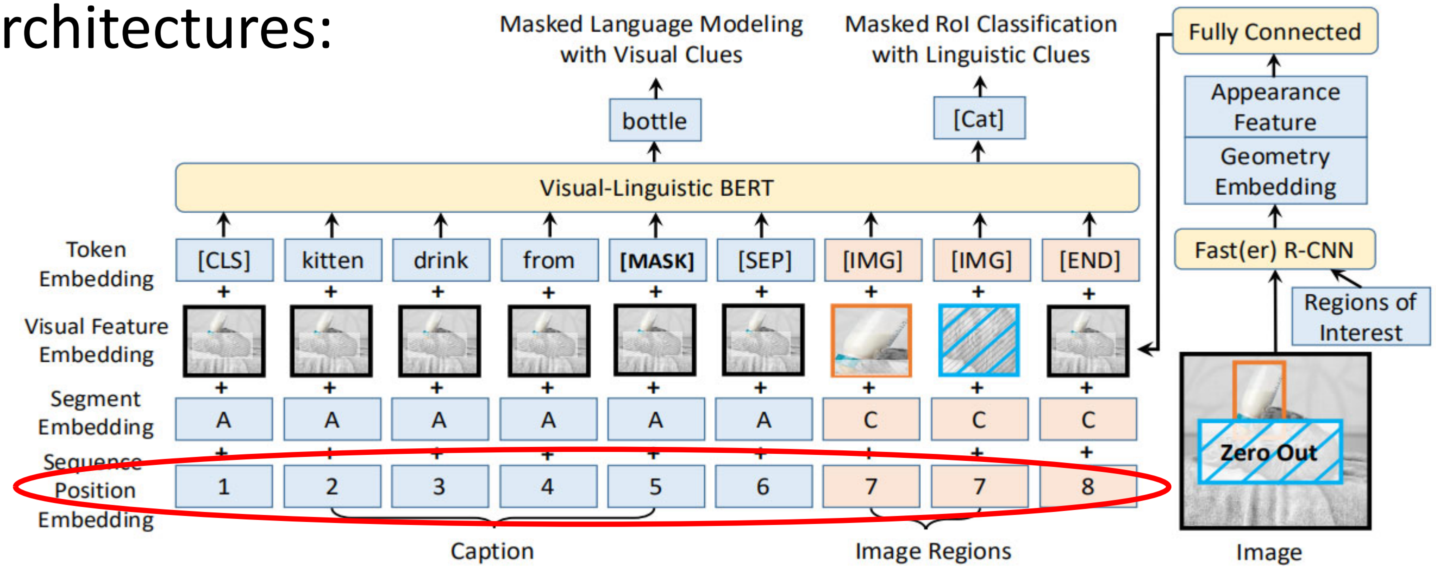
- Input = **Token Embedding** + Segment Embedding + Sequence Position Embedding + Visual Feature Embedding
  - Token Embedding:
    - Language: Embedded with Word-Piece Embeddings with 30,000 vocabulary
    - Vision: Special token [IMG]

- Model Architectures:



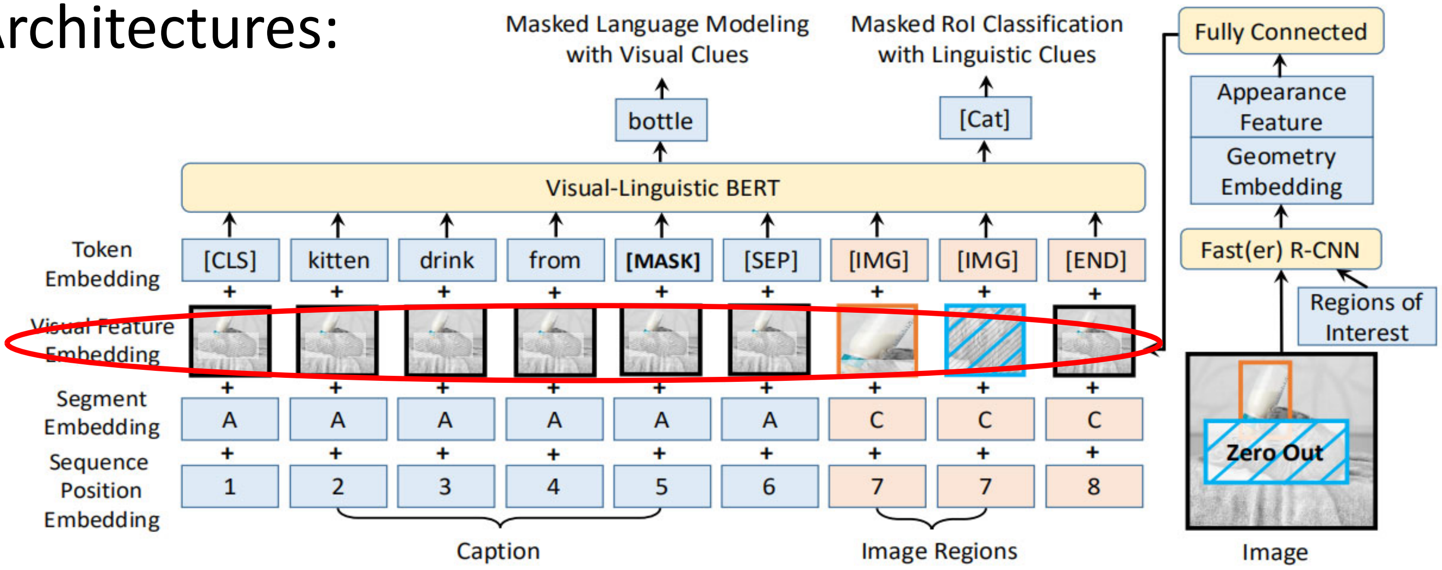
- Input = Token Embedding + **Segment Embedding** + Sequence Position Embedding + Visual Feature Embedding
  - Segment Embedding:
    - Three Types: A for Question, B for Answer (widely, another sentence if necessary, C for images)
    - Represented with 768-dim

- Model Architectures:



- Input = Token Embedding + Segment Embedding + **Sequence Position Embedding** + Visual Feature Embedding
  - Sequence Position Embedding:
    - Max length: 512
    - Indicating order in the input sequence
    - e.g. Black cat is ... and white cat is ...

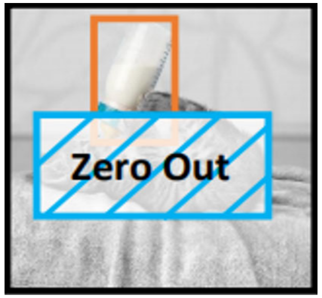
- Model Architectures:



- Input = Token Embedding + Segment Embedding + Sequence Position Embedding + **Visual Feature Embedding**
  - Visual Feature Embedding:
    - Applying Fast RCNN to extract visual feature
    - Non-visual elements: extract features on whole image
    - Concatenate visual feature and geomerty (position in image) embedding (2048-dim from 4-dim)

VL-BERT

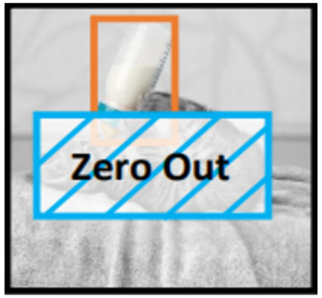
- Pre-training
  - 1. V&L Dataset: Conceptual Captions; 2. Text only (overcome overfitting on short sentences): BooksCorpus, English Wikipedia
  - Task1: Masked Language Modeling with Visual Clues
    - Masked a word randomly and predict what it is according to visual clues
    - e.g. Kitten is drinking from [MASK]
    - Loss: CE Loss (feed output feature into a classifier over whole vocabulary)



VL-BERT

- Pre-training
  - Task2: Masked RoI Classification with Linguistic Clues
    - Masked a RoI with 0
    - Set label predicted by Faster-RCNN as GT
    - Loss: CE Loss

A **kitten** is drinking from a bottle

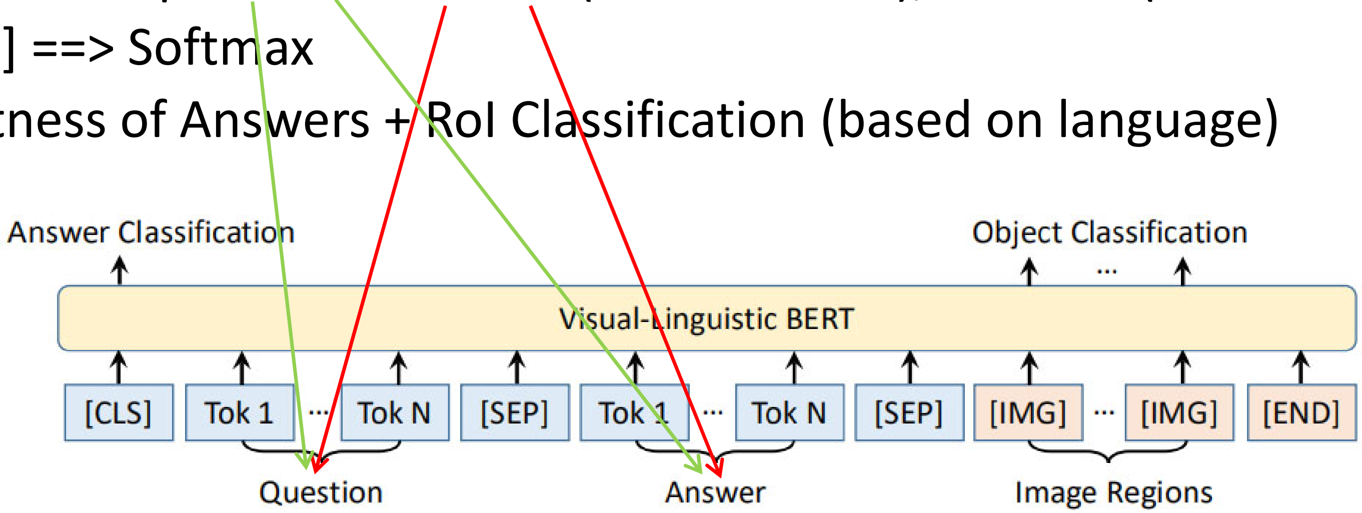


VL-BERT

- Fine-tuning
  - Provide proper formatted input and output
  - Provide appropriate segment embeddings
  - Design task-specific loss function
  - End-to-End Training
  - Output:
    - For sentence-image-relation level prediction: output feature from [CLS] is used
    - Word-level or RoI-level

Fine Tuning

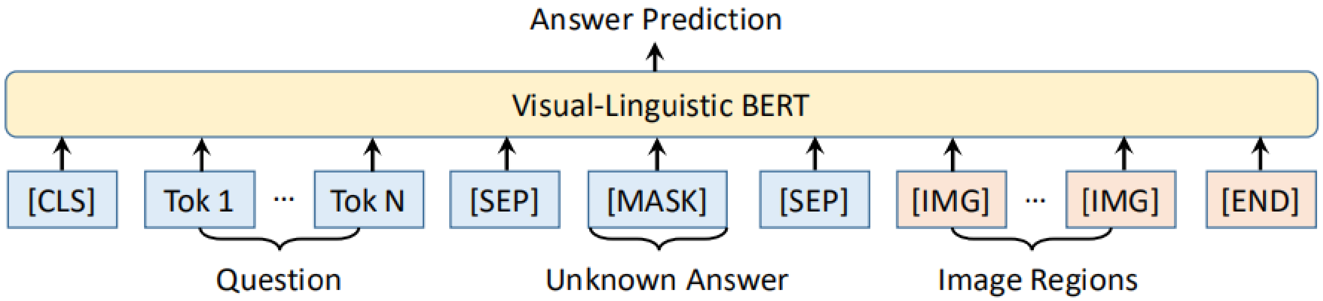
- Visual Commonsense Reasoning (VCR)
  - Question -> Answering + Reasoning
  - Decompose into two steps: Q->A + QA->R (comcat. Q&A); RoIs: GT (annotate.)
  - Output from [CLS] ==> Softmax
  - Two Loss: Correctness of Answers + RoI Classification (based on language)
  - **!SOTA!**



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

Fine Tuning

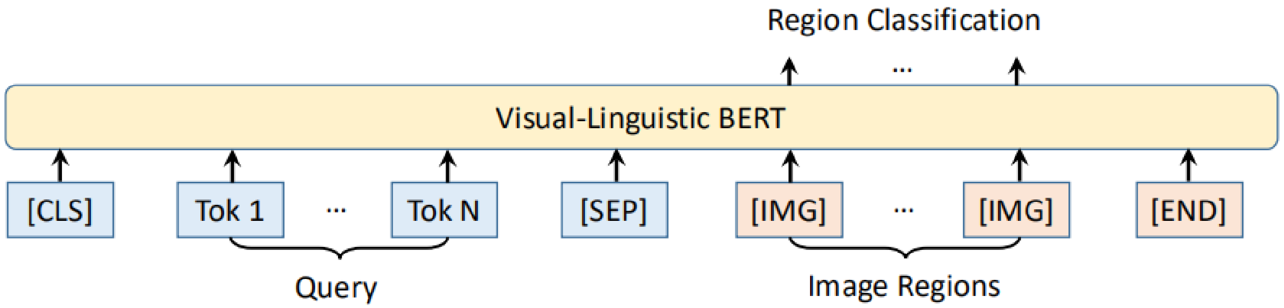
- Visual Question Answering
  - Generate/choose answers of question according to image
  - Input format:<Question, Answer([MASK]), Image>
  - Output from [MASK] ==> Softmax ==> Answer



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

Fine Tuning

- Referring Expression Comprehension
  - Localize referred object in the expression (a NL phrase)
  - Input format:<Query, Image>
  - Train & Eval: GT regions+ Detected regions

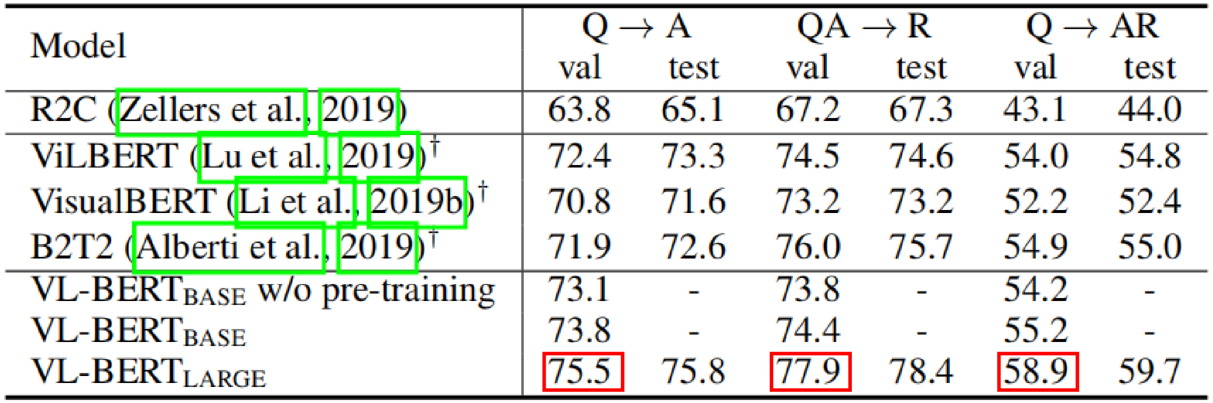


(c) Input and output format for Referring Expression task on RefCOCO+ dataset

Experiments

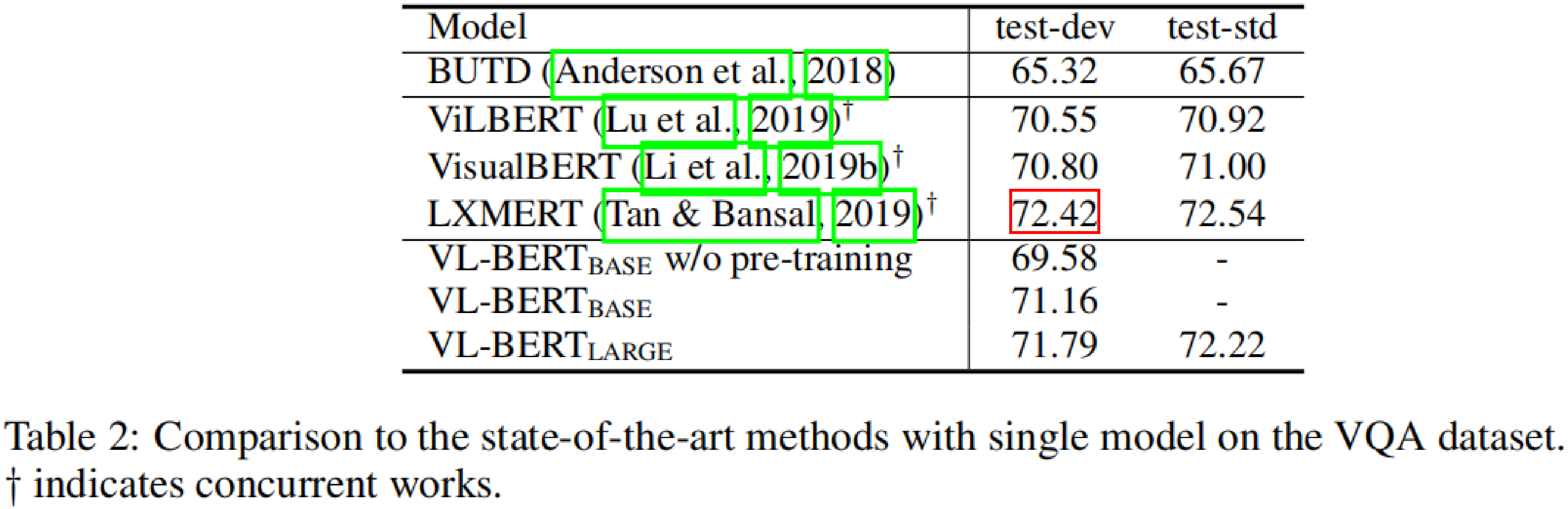
- Pretraining Detail:
  - Two models: BASE and LARGE (added params., see BERT)
  - Faster RCNN + ResNet101 (at most 100 RoIs, at least 10 RoIs)
- Conclusion: 一般没他好，比他好的更复杂

VISUAL COMMONSENSE REASONING (VCR)



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$_{\text{BASE}}$ 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

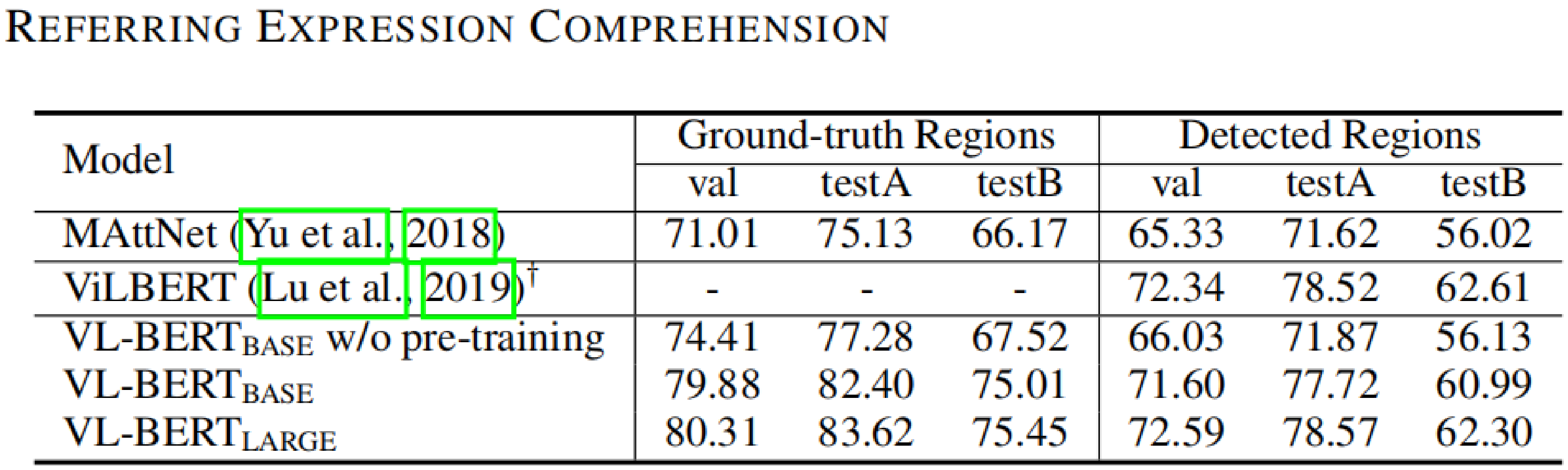
Experiments



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$_{\text{BASE}}$ w/o pre-training	69.58	-
VL-BERT$_{\text{BASE}}$	71.16	-
VL-BERT$_{\text{LARGE}}$	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

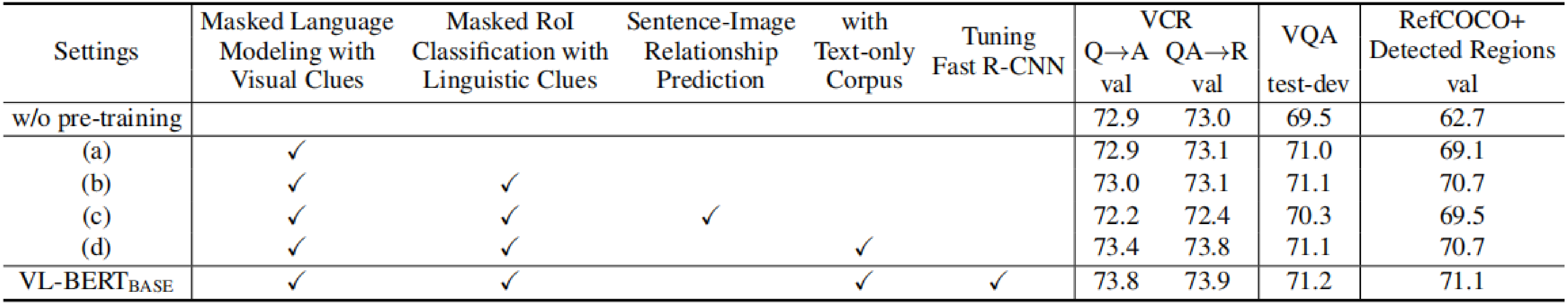


Referring Expression Comprehension

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

Experiments

- Ablation



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.