Cross-modality Semantic Learning and Reasoning

Wei, Zhongyu (魏忠钰)

http://www.sdspeople.fudan.edu.cn/zywei/

Data Intelligence and Social Computing Lab (DISC)
Fudan Natural Language Processing Lab (Fudan-NLP)

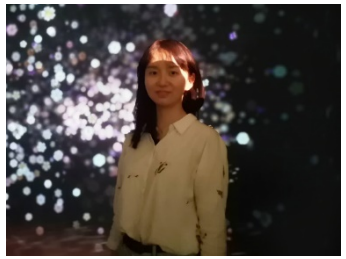
Fudan University

Main Student Collaborators

Fan, Zhihao（范智昊）

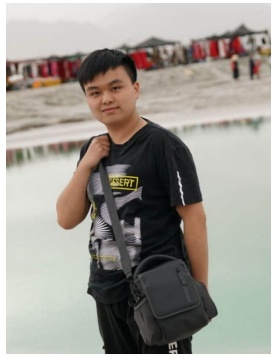
Wang, Ruize（王瑞泽）

Zhang, Jiwen（张霁雯）



Zhao, Wangrong（赵王榕）

Zhang, Yiteng（张翼腾）



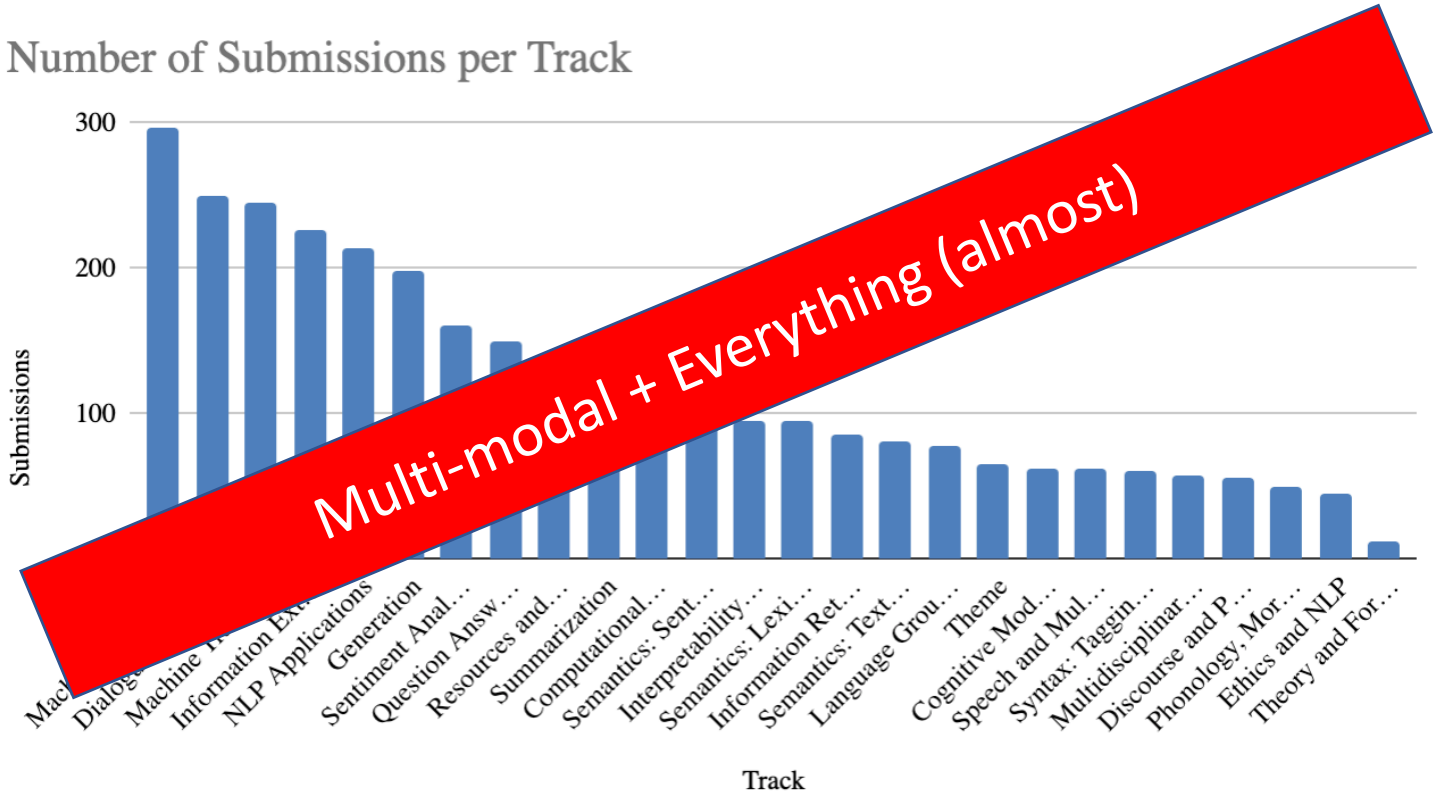
Li, Zejun（李泽君）

He, Xiaofeng（何啸风）

Cheng, Yijing（承怡菁）

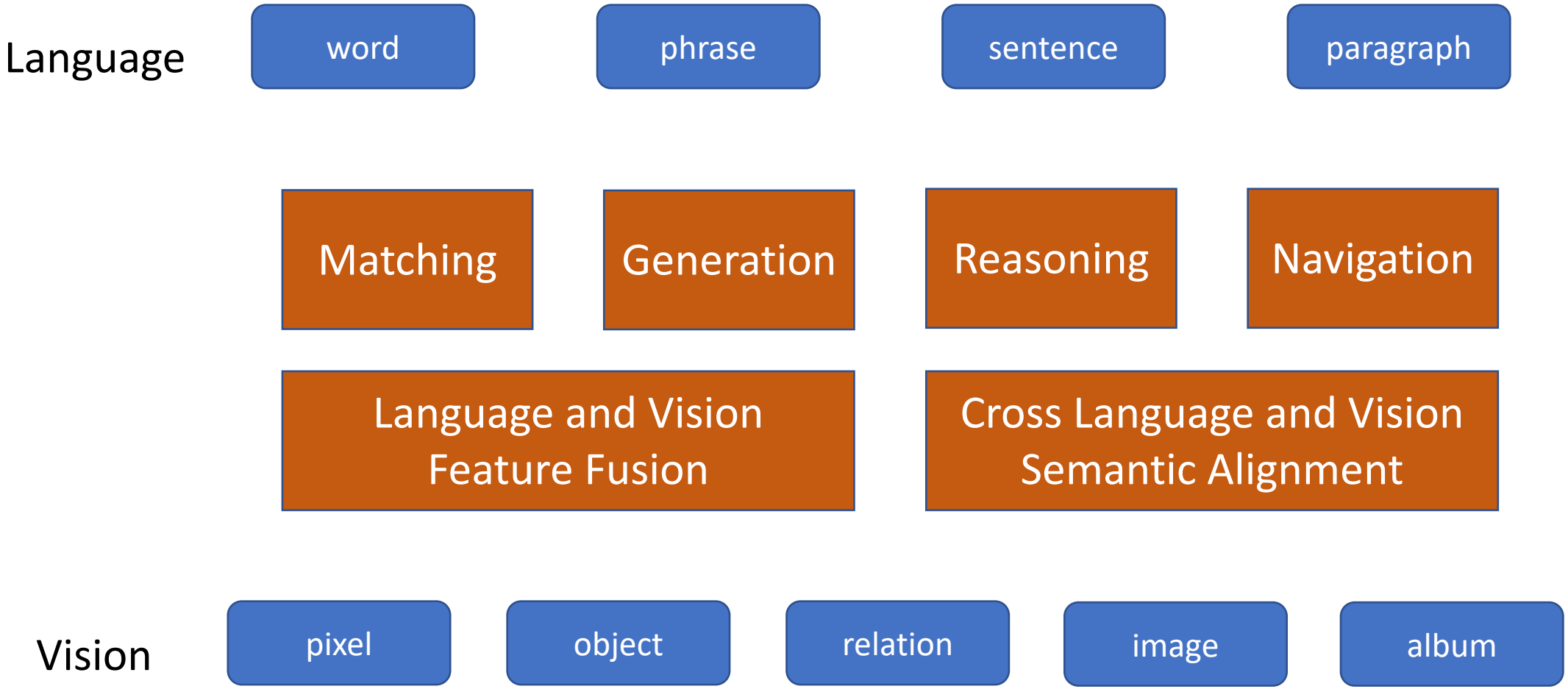
Cross-Modality is Trending

- ACL 2020 Language Grounding track, 77 submissions, 19 accepted
- + Modal, visual, image, 28 accepted



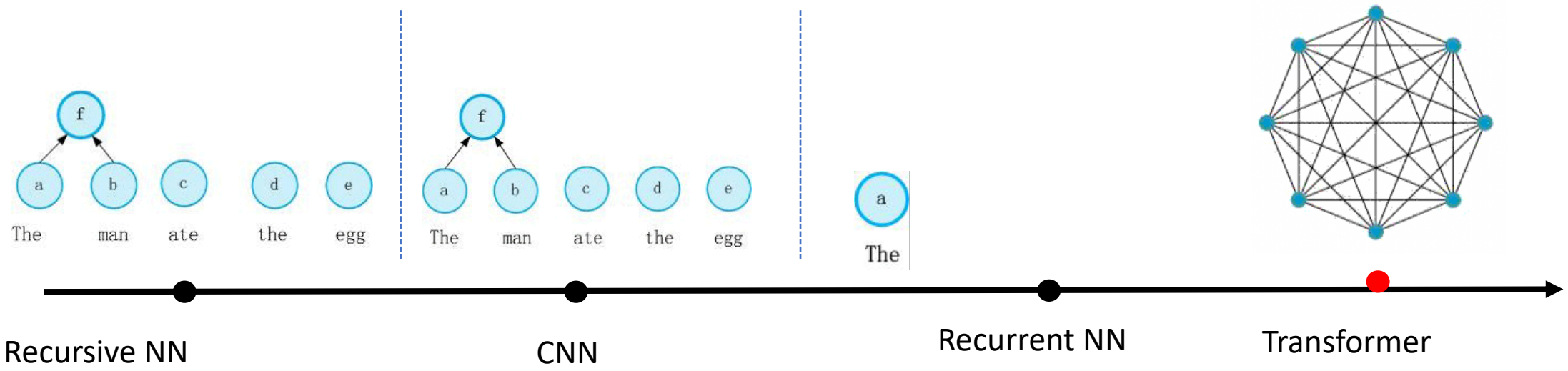
https://acl2020.org/blog/general-conference-statistics/ https://www.jiqizhixin.com/articles/2020-05-27-3

Cross Language and Vison Research Topics



Language Representation

- Word: one-hot, word2vec
- Word sequence, i.e., phrase, sentence, paragraph

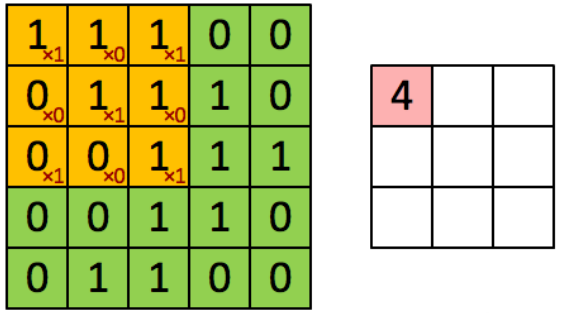


Visual Representation

- Pixel-based: CNN, VGGNet, ResNet
- Object: RCNN, Fast RCNN, Faster RCNN
- Scene-graph: structural visual information.object, attribute, relations

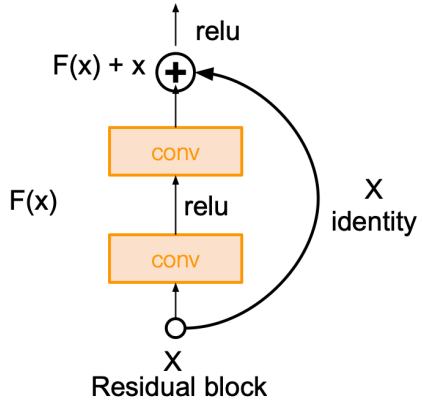
Pixel-based visual representation learning

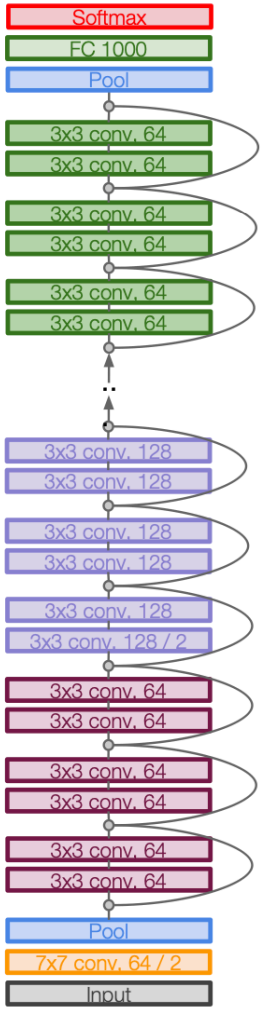
- Transform an image into a dense vector



CNN

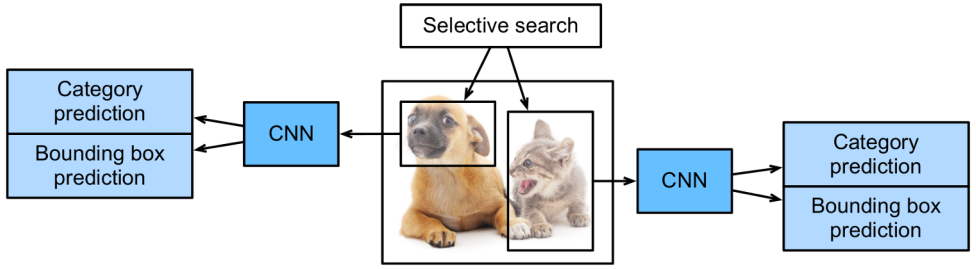
(From Stanford CS224)





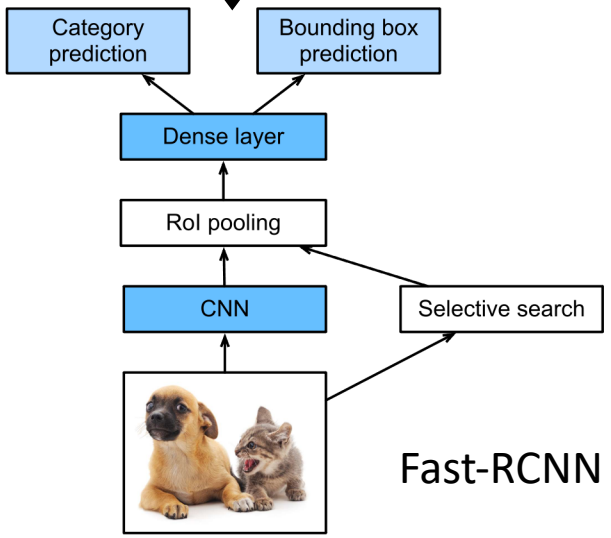
Object-Level Visual Representation Learning

- Region of Interest (RoI)



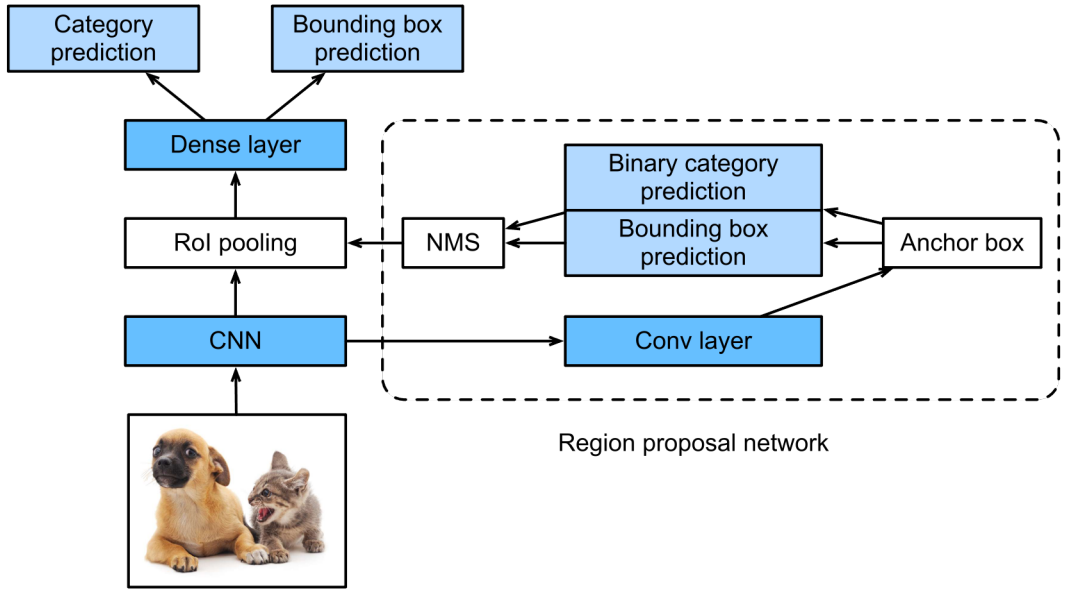
RCNN

Improves by only performing CNN on the image as a whole rather than each proposed region.



Fast-RCNN

Replaces selective search with a region proposal network.



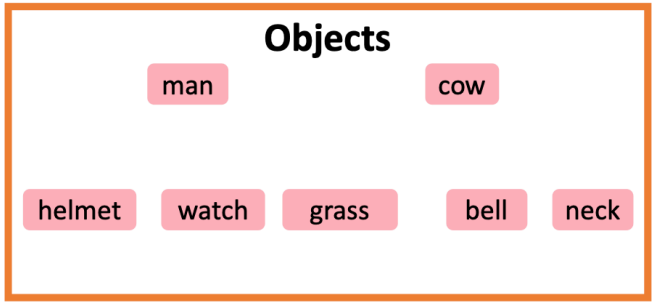
Faster-RCNN

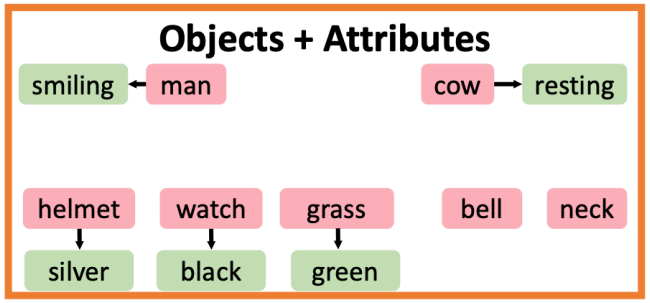
	R-CNN	Fast R-CNN	Faster R-CNN
Test time per image (with proposals)	50 seconds	2 seconds	**0.2 seconds**
(Speedup)	1x	25x	**250x**
mAP (VOC 2007)	66.0	**66.9**	**66.9**

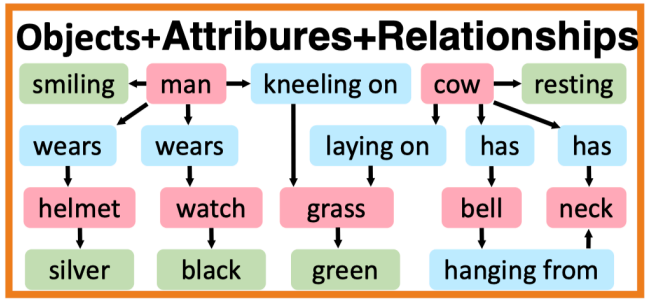
From: https://d2l.ai/chapter_computer-vision/rcnn.html

Structured Representation of an Image

What information does image convey? – Objects, Attributes, Relationships



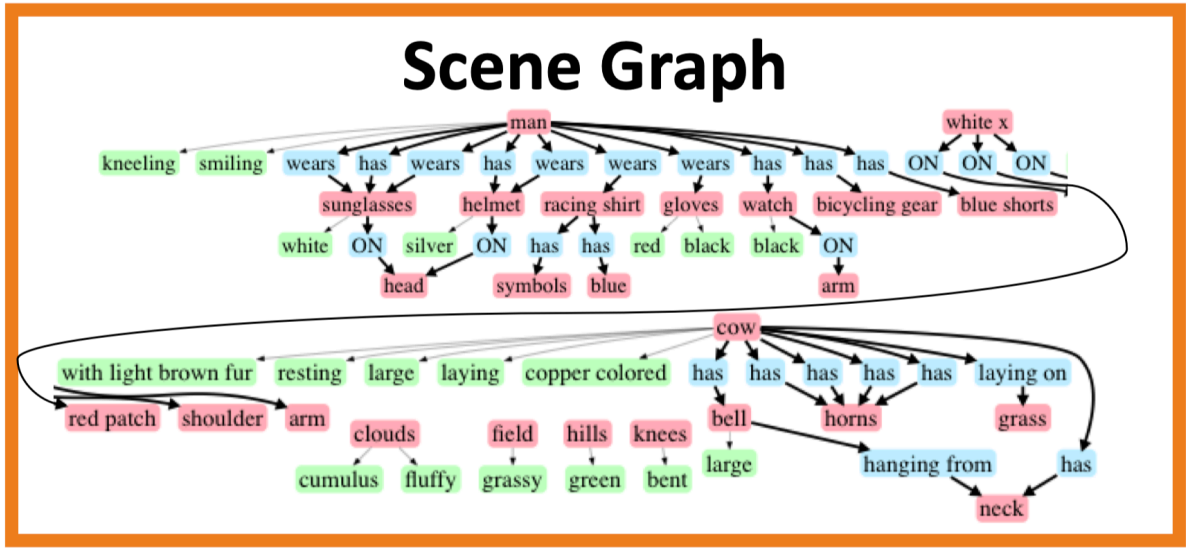




Scene Graph

Scene Graph - A compositional representation:

Objects + Attributes + Relationships

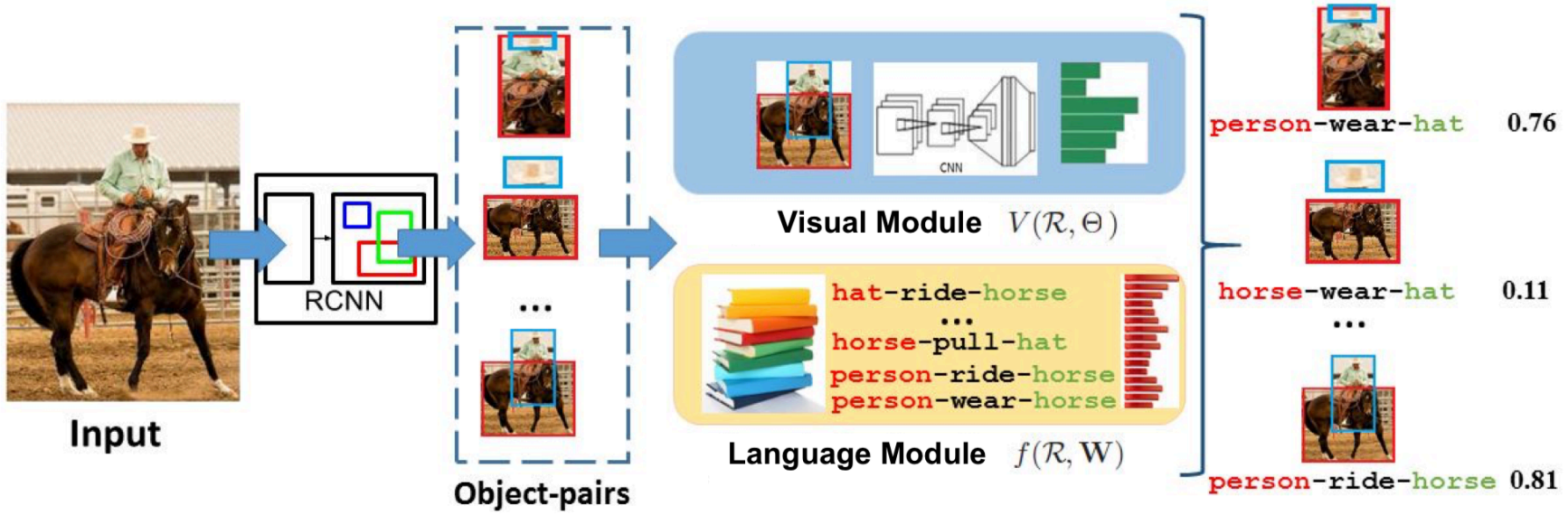


Johnson, Krishna, Stark, Li, Shamma, Bernstein, and Fei-Fei. Image Retrieval with Scene Graphs. CVPR 2015

Krishna, Zhu, Groth, Johnson, Hata, Kravitz, Chen, Kalantidis, Li, Shamma, Bernstein, and Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 2017

Scene Graph Construction from Images

- Images to scene graphs (visual relationship detection)

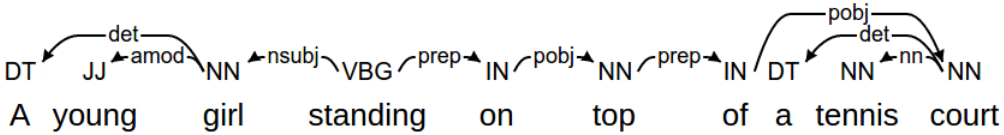


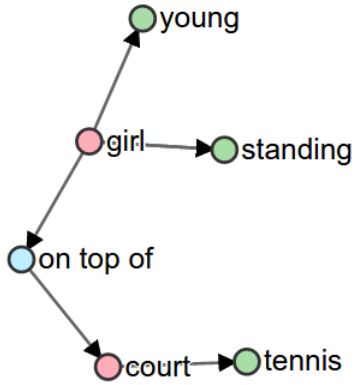
- **Object Detector:** RCNN generates a set of objects
- **Relationship Detector:** each pair of object is scored using a visual module and a language model

Cewu Lu, et al. Visual Relationship Detection with Language Priors (ECCV). 2016

Scene Graph Construction from text

- Text to scene graphs





- SPICE: an automated caption evaluation metric for image captioning
  - **A syntactic dependency tree** is built by Probabilistic Context-Free Grammar (PCFG) dependency parser.
  - **A rule-based method** is applied for transforming the tree to a scene graph.
  - **Calculate F-score** for objects, attributes and relationships in scene graphs.

Surveys on Cross Language and Vison Research

Xiaodong He, Li Deng: Deep Learning for Image-to-Text Generation: A Technical Overview. IEEE Signal Process. Mag. 34(6): 109-116 (2017)

2. Chao Zhang, Zichao Yang, Xiaodong He, Li Deng: Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. IEEE J. Sel. Top. Signal Process. 14(3): 478-493 (2020)

3. 魏忠钰, 范智昊, 王瑞泽, 承怡菁, 赵王榕, 黄萱菁, 从视觉到文本: 图像描述生成的研究进展综述, 中文信息学报, 2020 Vol. 34 (7): 19-29.

Outline

- **Cross Vision and Language Matching**
- Vision-based Text Generation
- Cross Vision and Language Reasoning
- Language-based Vision Navigation

- Cross-modality Pretraining

Image Text Retrieval

- Given a query image, retrieve a related sentence from sentence set
- Given a query sentence, retrieve a context-matching image from image set

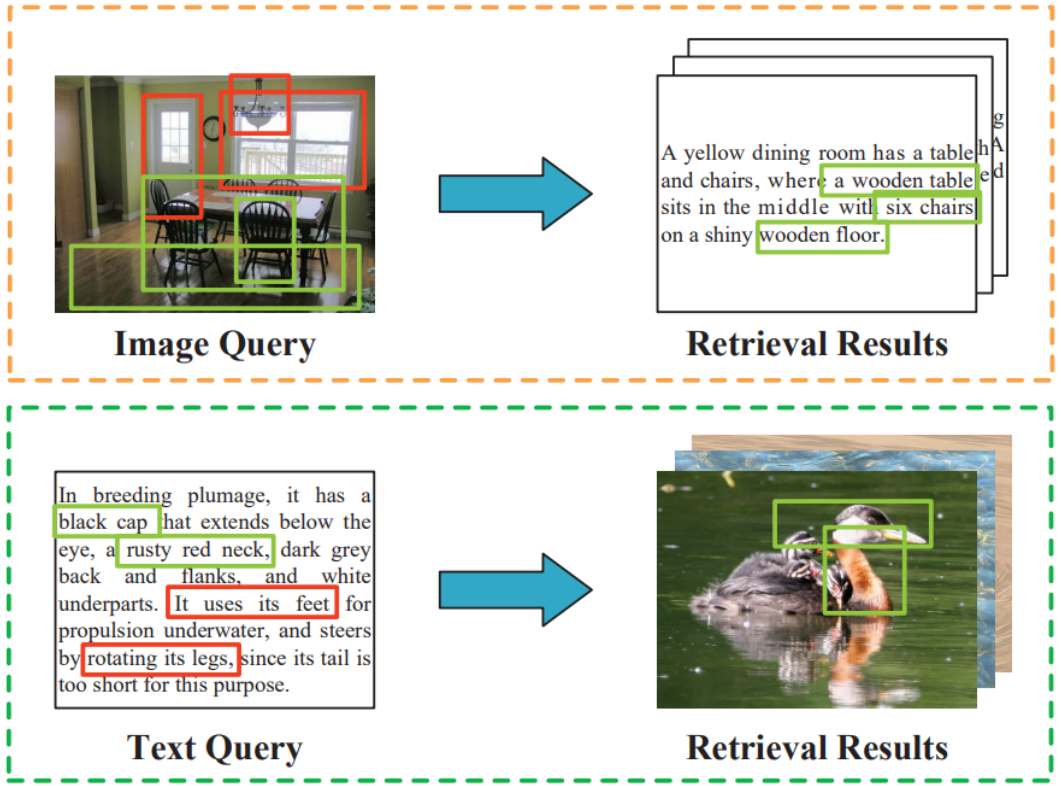
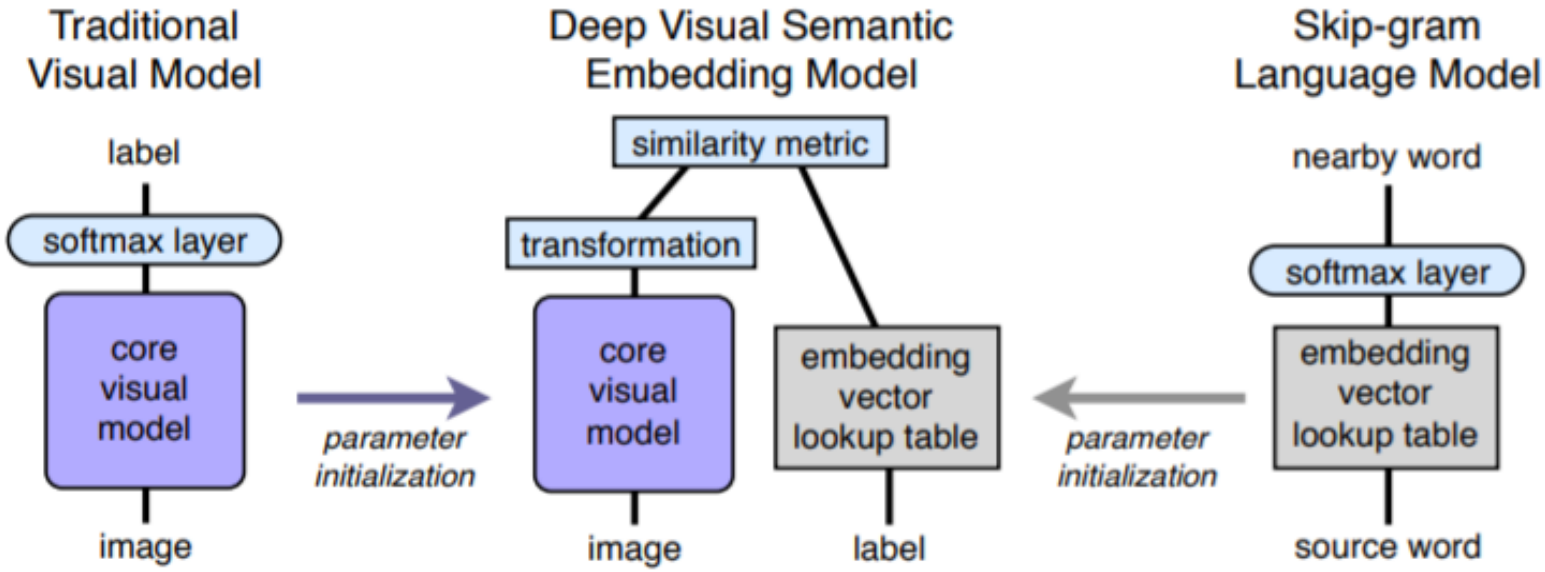


Image-Text Semantic Alignment

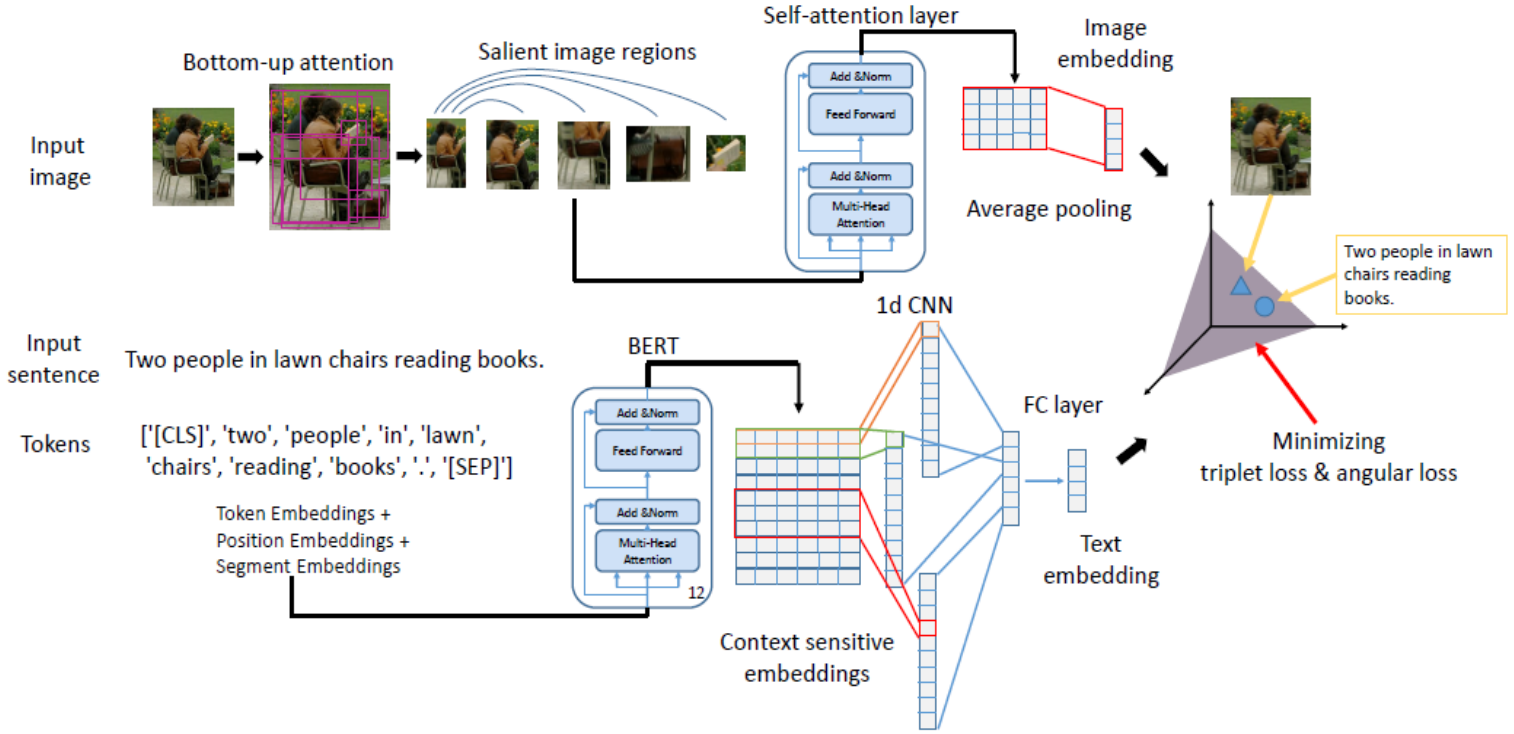
- Vision-Semantic Embedding Space (ViSE)[1]
  - Pretrained skip-gram language model
  - CNN-based visual representation model
  - Hinge rank loss for embedding learning



Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*.

Intra-Modality Relationship Enhanced VISE

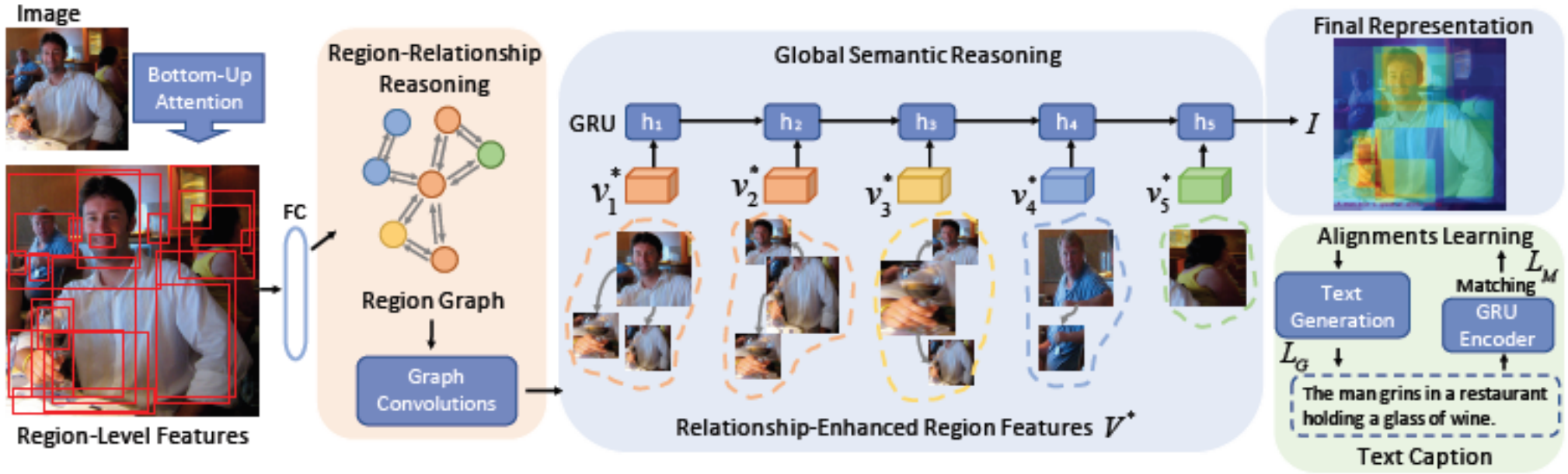
- Self-Attention Embeddings for Image-Text Matching(SAEM)
  - Image Region Relationship Modeling: Transformer + Average Pooling
  - Sentence Relationship Modeling: BERT + CNN
  - Better Metric Learning
    - Triplet Loss
    - Angular Loss



Wu, Y., Wang, S., Song, G., & Huang, Q. (2019, October). Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 2088-2096).

Visual Semantic Reasoning for Image-Text Matching(VSRN)

- Employing GCN to modeling semantic relationships.
- Using RNN to perform global semantic reasoning, and capture the key semantic concepts.



Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2019). Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4654-4662).

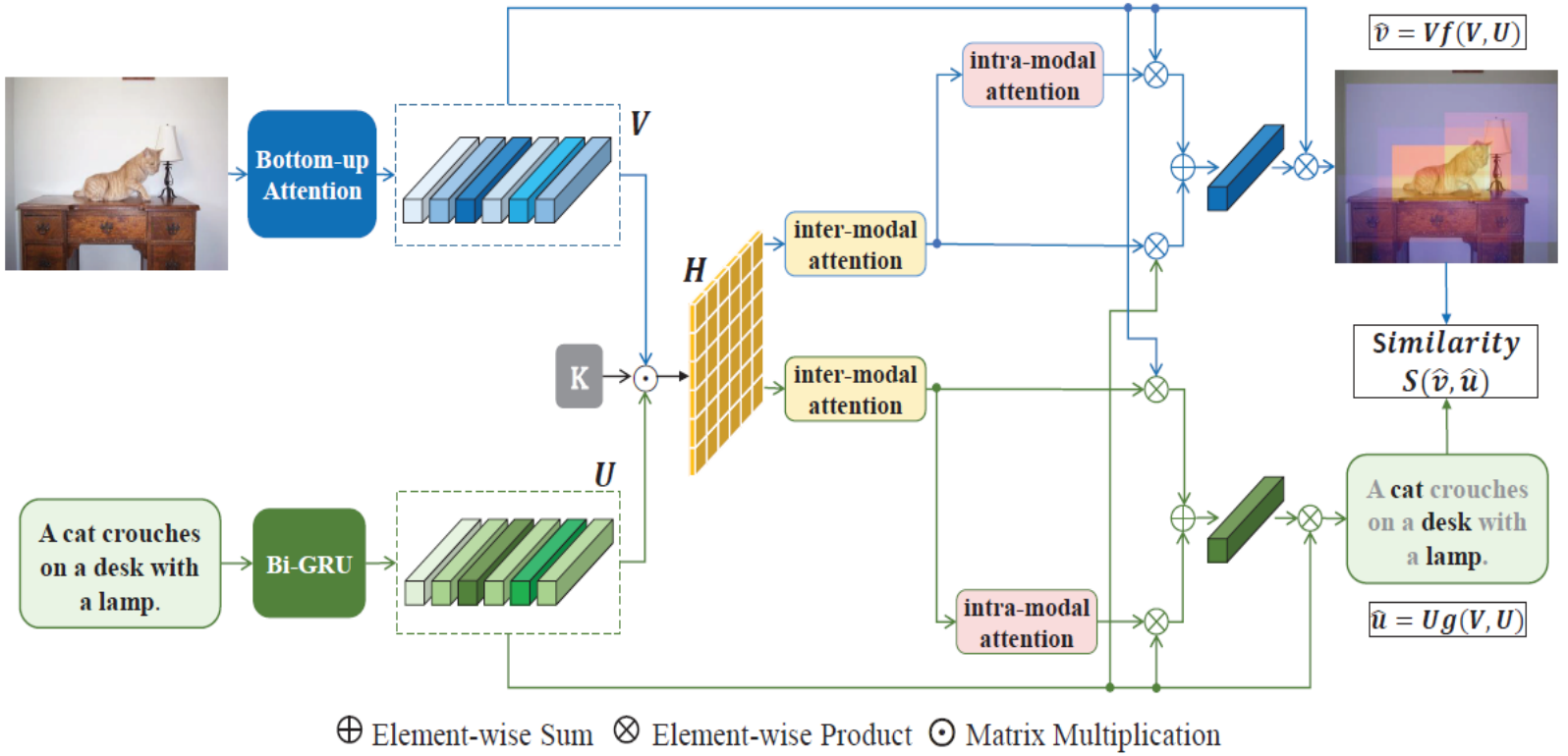
Inter-modality Relationship Modeling

- Latent Alignment between image regions and words
- Step1: Regions attend to words in the sentence
- Step2: Regions compare the attended sentence vector to determine its own importance with respect to the sentence.
- Step3: Mean pooling over the similarity score vectors.

Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 201-216).

Inter- and Intra-modality Modeling

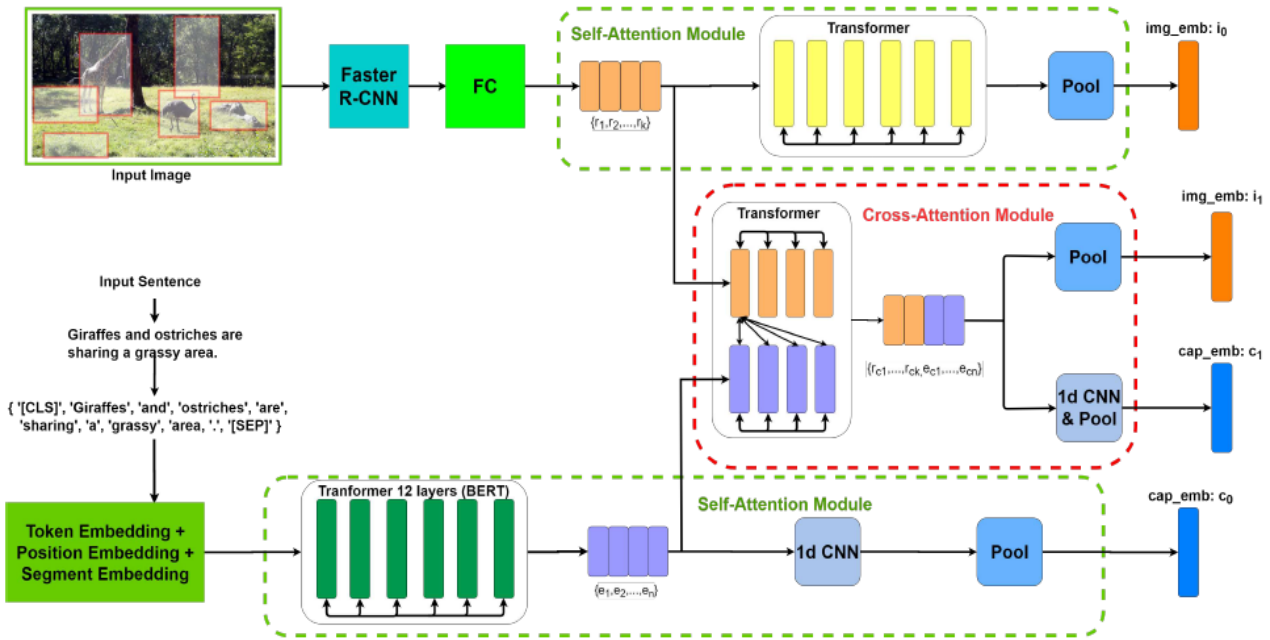
- Inter-modal attention: calculate the similarities of local region-word pairs
- Intra-model attention: modeling correlations from two different perspectives.



Zhang, Q., Lei, Z., Zhang, Z., & Li, S. Z. (2020). Context-Aware Attention Network for Image-Text Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3536-3545).

Inter- and Intra-modality Modeling Respectively

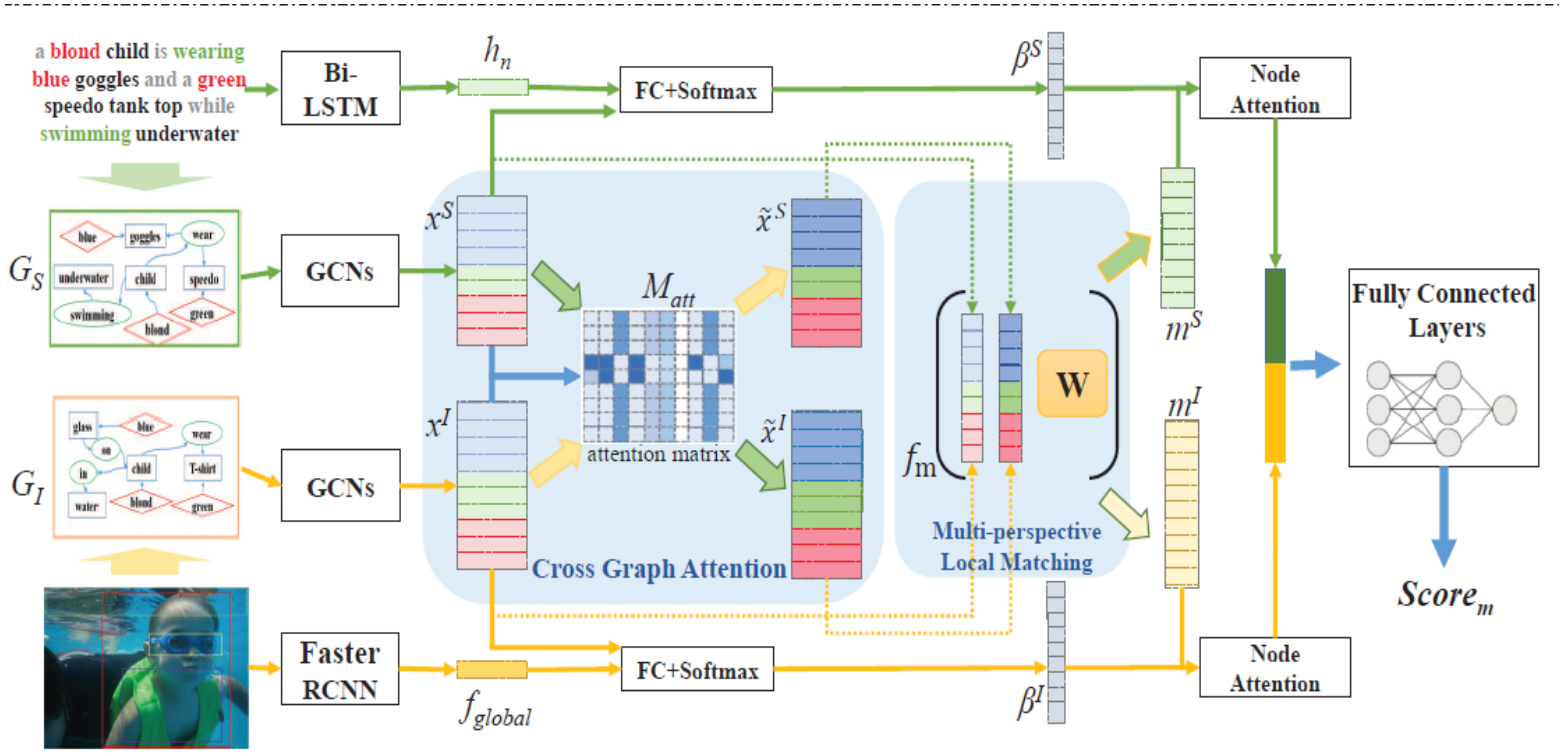
- Split the intra-modality representations and inter-modality representations.
- Self-attention and CNN for sentence representations.
- Self-attention for image representations.
- Cross-attention for inter-modality relationship modeling.



Wei, X., Zhang, T., Li, Y., Zhang, Y., & Wu, F. (2020). Multi-Modality Cross Attention Network for Image and Sentence Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10941-10950).

Inter- and Intra-modality Modeling in High-Order

- Scene graphs for representing highly-structural visual or textual semantics.
- Propose a graph attention mechanism to compute relation between the two modalities



Li, Y., Zhang, D., & Mu, Y. (2020). Visual-Semantic Matching by Exploring High-Order Attention and Distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12786-12795).

Dataset and Evaluation

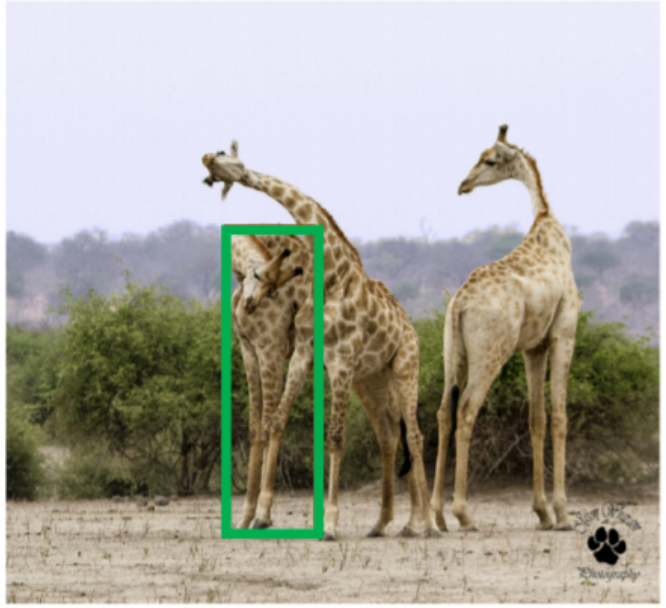
- Datasets
  - MSCOCO
    - 113,287, 5000 and 5000 images for training, validation and testing
    - 5 captions per image
    - 1K test set (5 fold testing)
    - 5K test set
  - Flickr30K
    - 29000, 1000 and 1000 images for training, validation and testing
    - 5 captions per image
- Metrics
  - Image Retrieval: Recall@1(R@1), Recall@5(R@5), Recall@10(R@10)
  - Sentence Retrieval: Recall@1(R@1), Recall@5(R@5), Recall@10(R@10)

Performance Comparison

MSCOCO 1K test set							
	Cross-modality attention	Sentence Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
VISE	-	38.4	69.9	80.5	27.4	60.2	74.8
SAEM	Intra	71.2	94.1	97.7	57.8	88.6	94.9
VSRN	Intra	76.2	94.8	98.2	65.1	93.1	95.1
SCAN	Inter	72.7	94.8	98.4	58.8	88.4	94.8
CAAN	Inter+Intra	75.5	95.4	98.5	61.3	89.7	95.2
MMCA	Inter+Intra	74.8	95.6	97.7	61.6	89.8	95.2
HOAD	Inter+Intra	77.8	96.1	98.7	66.2	93.0	97.9

Visual Referring Expression

- Given a natural expression, identify related object in the image.
- Intersection over Union (IoU) ratio between the true and predicted bounding box.
- If IoU exceeds 0.5, we call the detection a true positive, otherwise it is a false positive.



RefCOCO:
1. giraffe on left
2. first giraffe on left

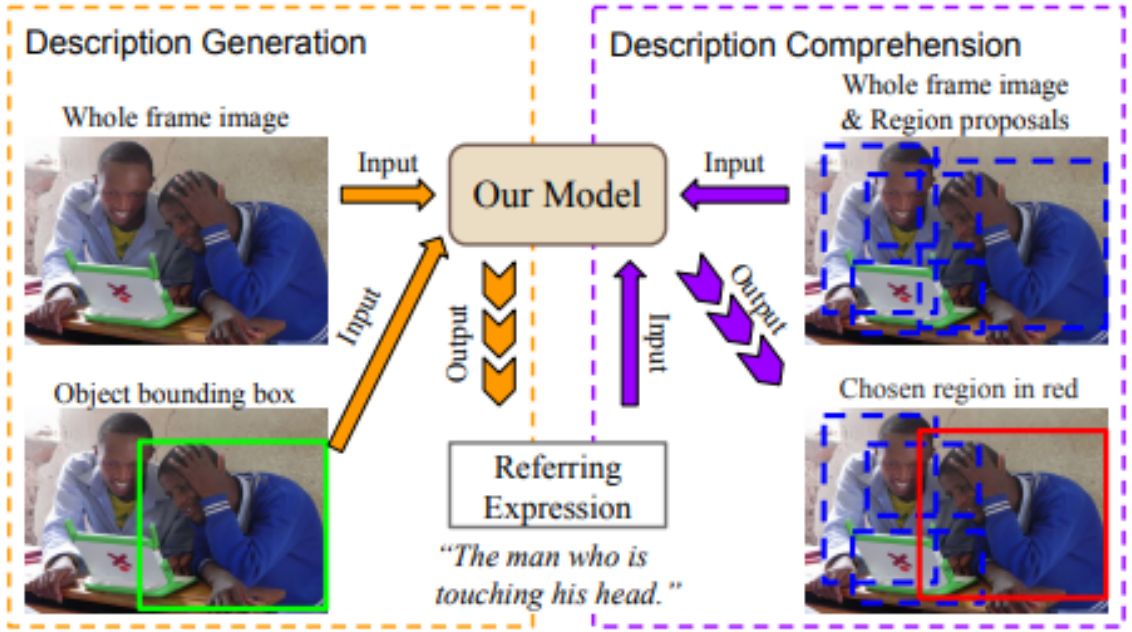
RefCOCO+:
1. giraffe with lowered head
2. giraffe head down

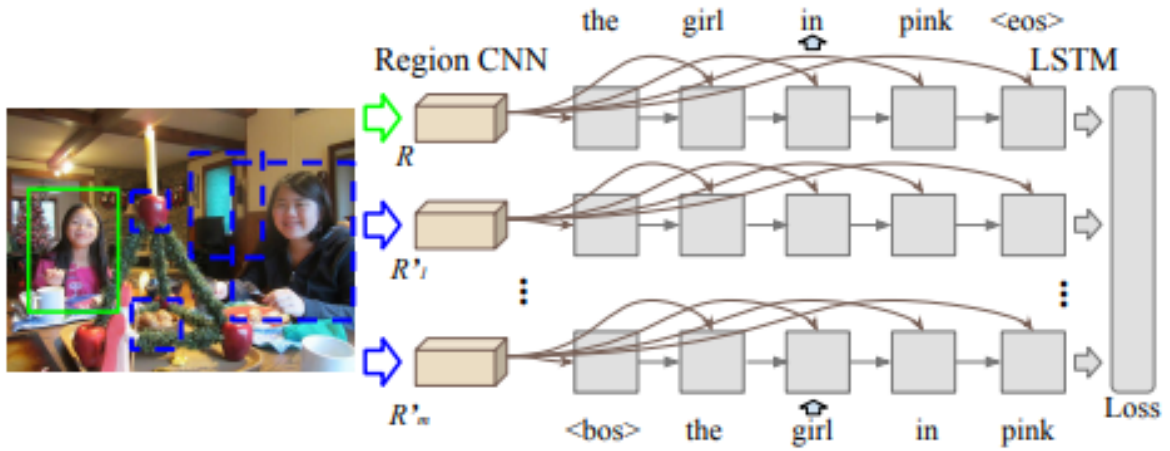
RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

	Image	Object	Expression	Avg. Length
RefCOCO	50,000	19,994	142,209	3.61
RefCOCO+	49,856	19,992	141,4564	3.53
RefCOCOg	26,711	54,822	85,474	8.43

Align Referring and Generation

- Expression Generator: from bounding box to expression
- Expression Understanding: determine which box is correct

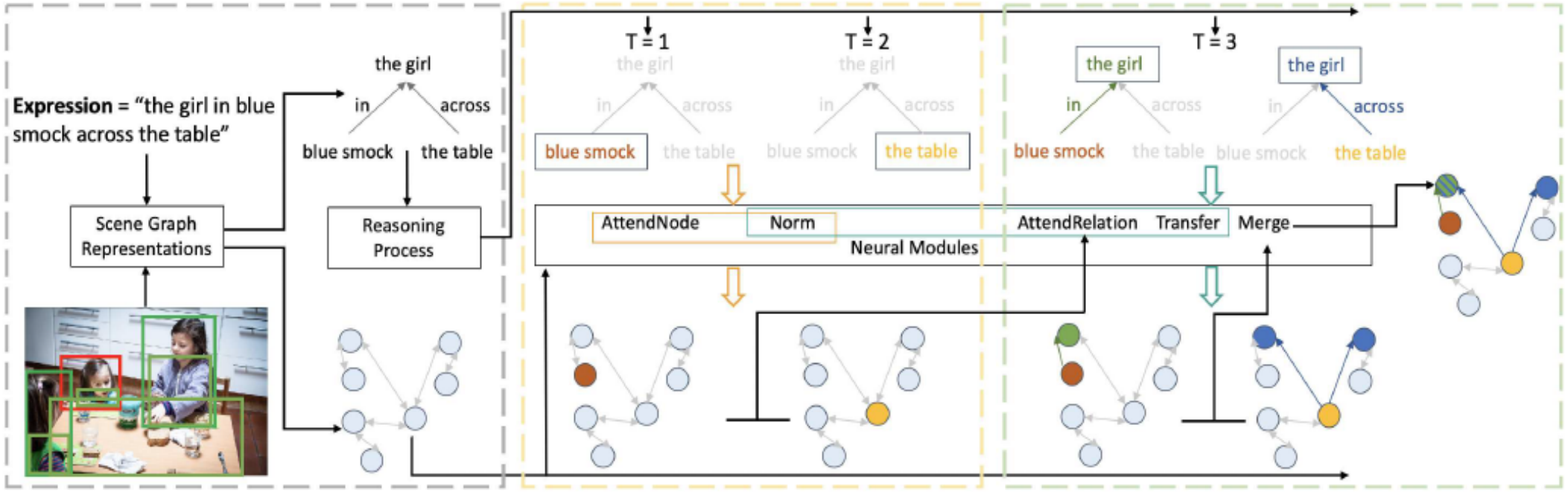




Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11-20).

Explore Structure Solution in REG

- Transfer the task into node identification in visual scene graph.
- Compute similarity between text node and visual node



Yang, S., Li, G., & Yu, Y. (2020). Graph-Structured Referring Expression Reasoning in The Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9952-9961).

Outline

- Cross Vision and Language Matching
- **Vision-based Text Generation**
- Cross Vision and Language Reasoning
- Language-based Vision Navigation

- Cross-modality Pretraining

Vision-based Text Generation

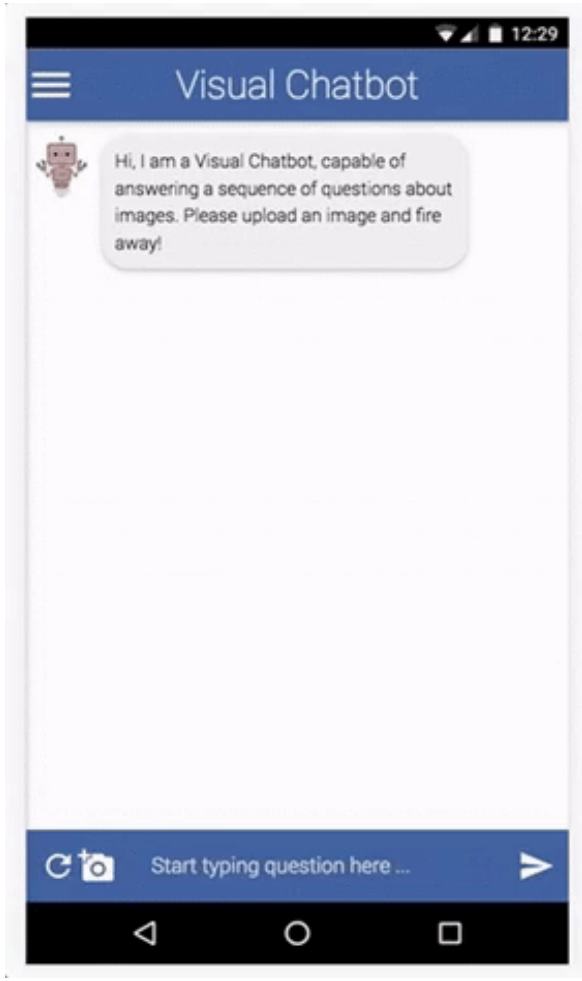
- Image Captioning
- Visual Storytelling
- Visual Dialogue

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

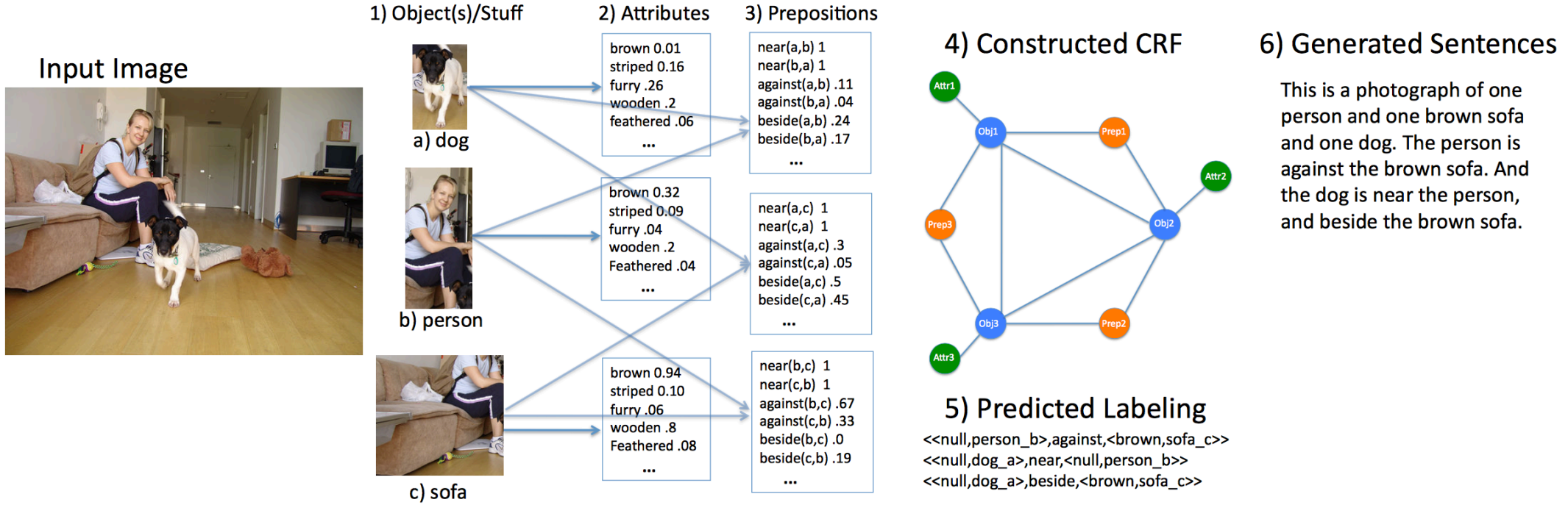


Key Issues for Visual-Text Generation

- Visual Representation
  - Dense visual feature vector (global, local)
  - Objects, relations
  - Scene-graph
- Text Generation
  - Retrieval-based
  - Template filling
  - Generation-based, statistical language model, neural Language model, e.g., RNN-based decoder, hierarchical RNN, VAE, GAN
- Vision-Text Alignment
  - Latent space
  - Attention mechanism
  - Pre-train model for visual and text

Pipeline Template-based Model

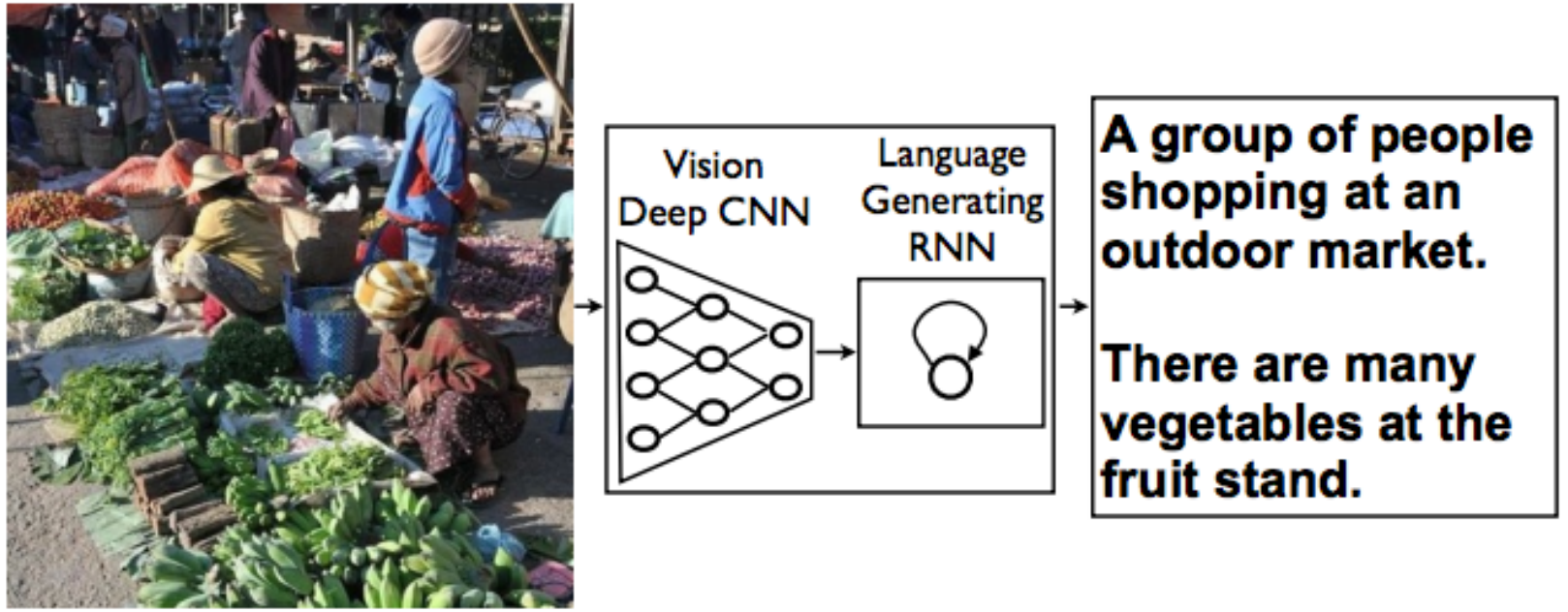
- Tuples for visual representation
- Rule-based text generation



Baby talk: Understanding and generating simple image descriptions, CVPR'11

End-to-end Neural-based system

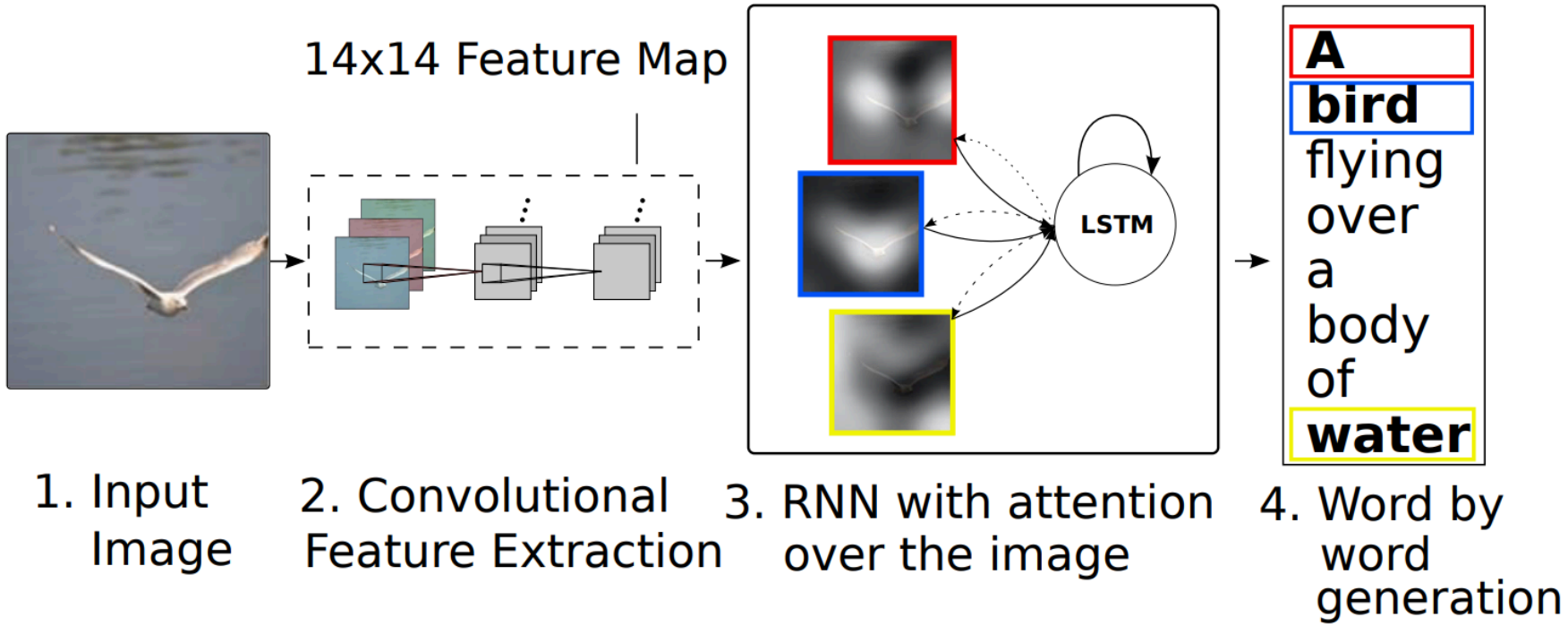
- In an encoder-decoder fashion
  - **Encoder**: CNN based architecture to model the whole image
  - **Decoder**: RNN based neural language model to generate caption word by word



Show and tell: A neural image caption generation, CVPR'15

Attention Mechanism to Bridge Vision and Caption

- Introduce fine-grained image feature for text generation (middle layer features from CNN)
- Various Attention Mechanism to link local visual feature and text generation



Show, attend and tell: Neural image caption generation with visual attention, ICML'15

Visual Concept to Bridge Vision and Text

- Introduce visual concept as the medium representation
- Attention mechanism is used as the bridge

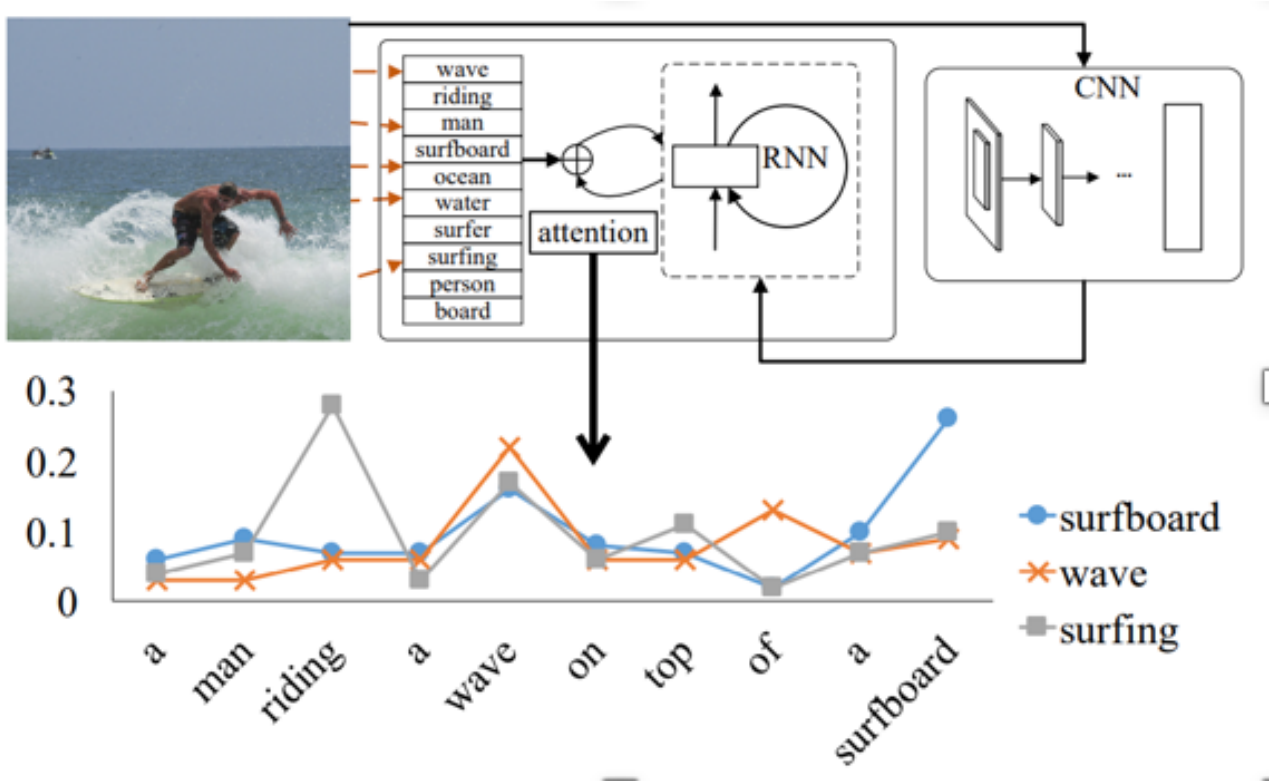
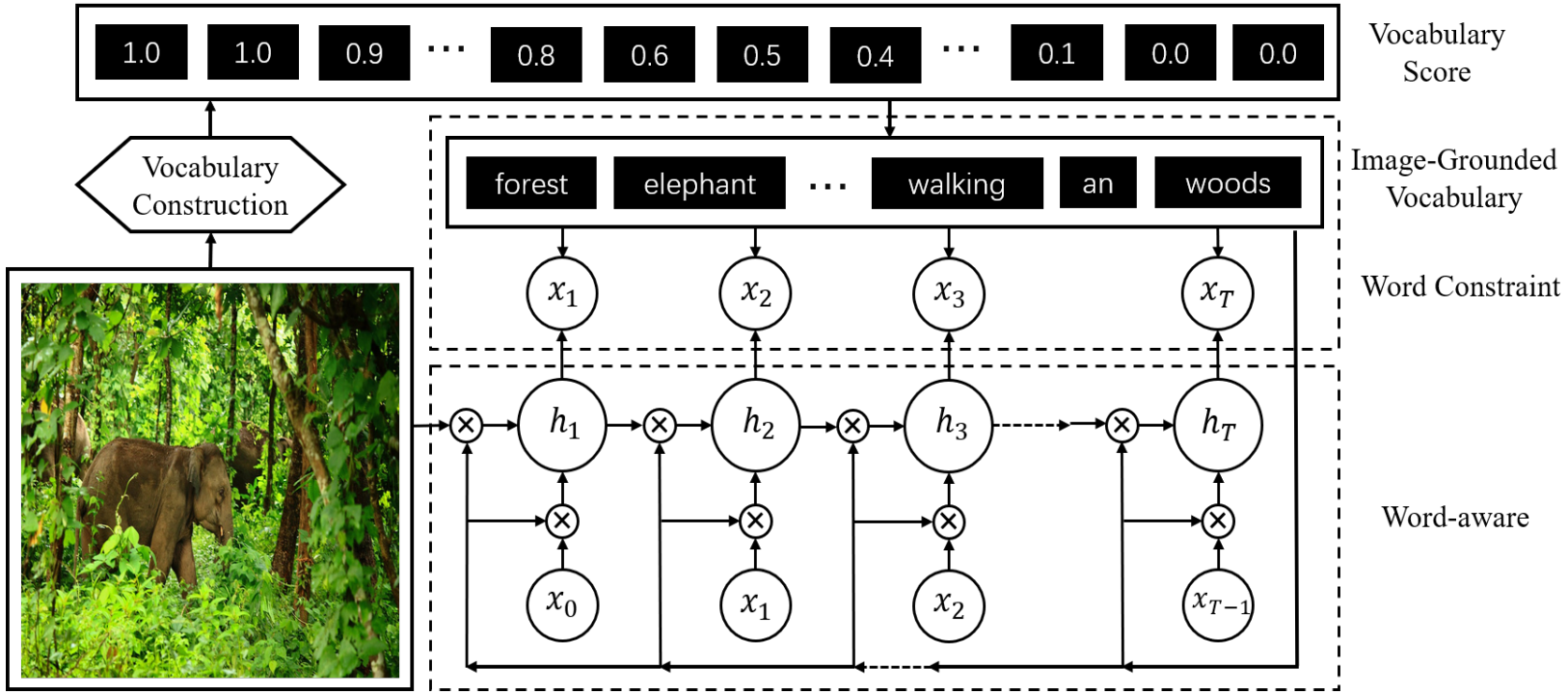


Image Captioning with Semantic Attention, CVPR'16

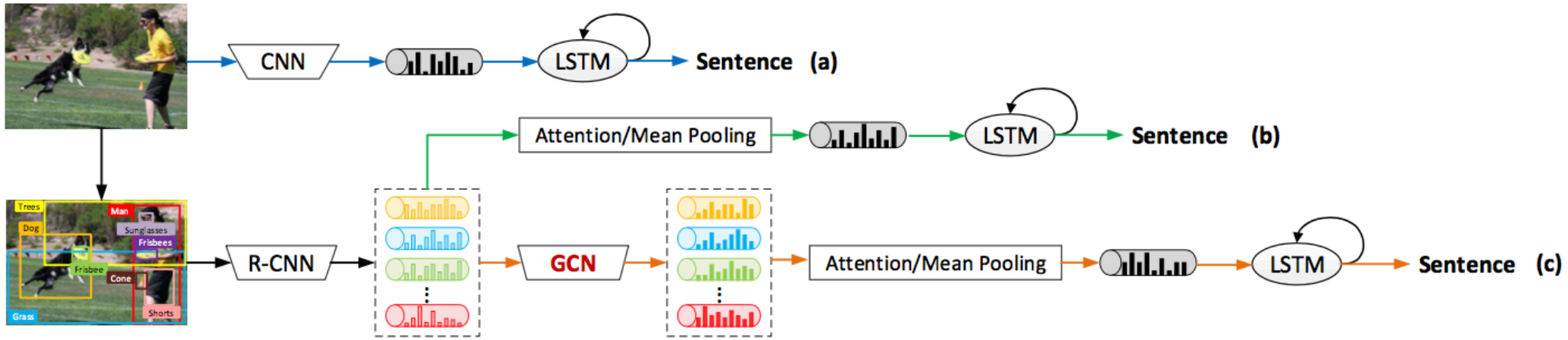
Image Guided Vocabulary Construction for Captioning



- **Vocabulary Constructor:** A two-step structure to construct image-grounded vocabulary
- **Word Constraint (WC):** Hard mechanism for caption generation
- **Word-Aware (WA):** Soft mechanism for caption generation

Bridging by Word: Image Grounded Vocabulary Construction for Visual Captioning, ACL 2019

Scene Graph for Image Captioning

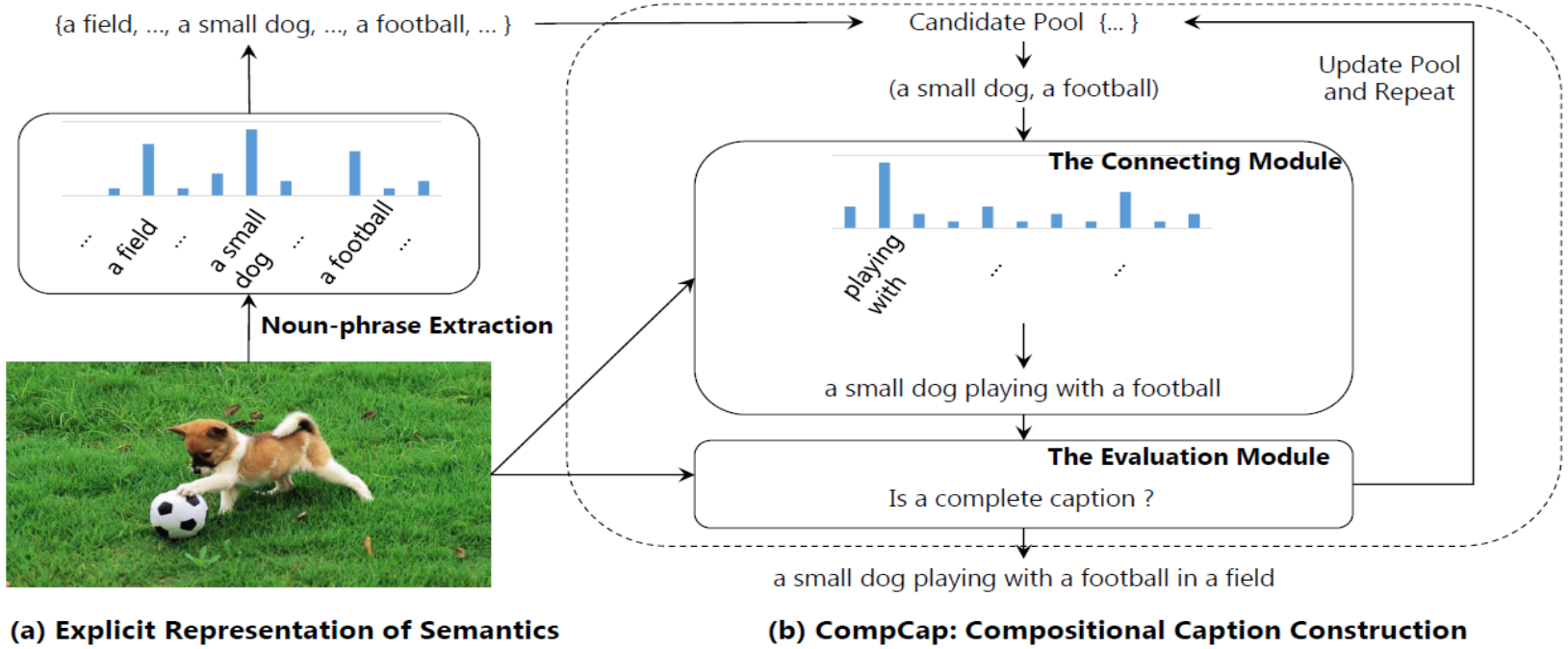


- Modeling relationships between objects for representing and eventually describing an image.
- RCNN for object detection
- GCN to fine-tune the object representation

Ting Yao, Yingwei Pan, Yehao Li, Tao Mei: Exploring Visual Relationship for Image Captioning. ECCV 2018

Image Captioning based on a Compositional Paradigm

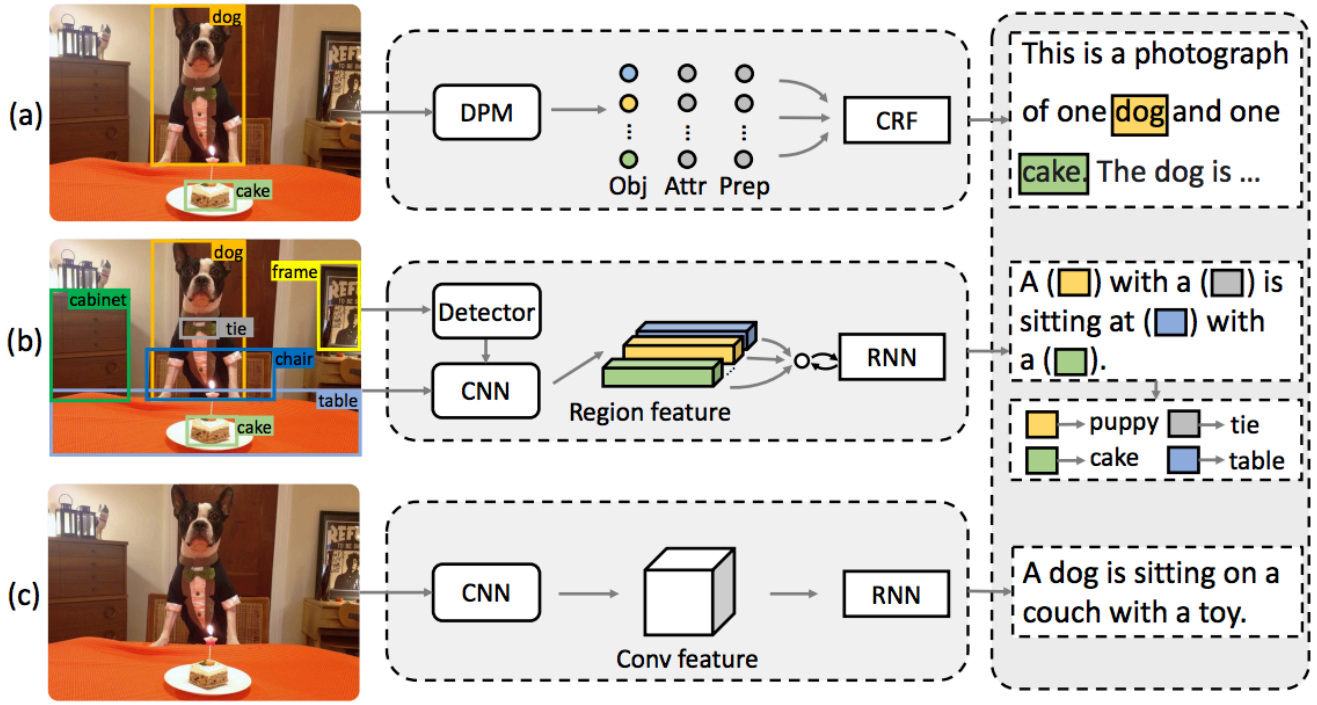
- Extract phrases as visual representation
- Learn to composite phrases for text generation



A Neural Compositional Paradigm for Image Captioning, NIPS'18

Caption generation combining template and neural decoder

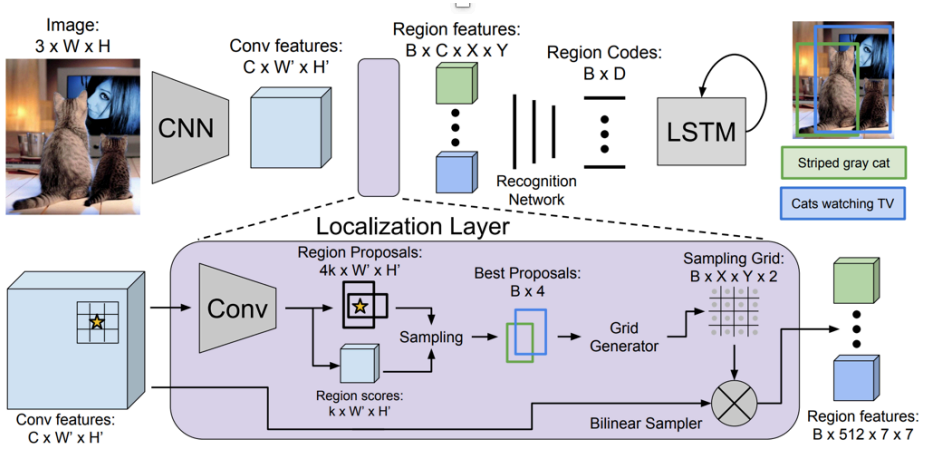
- Detect entities and attributes from the image as **visual vocabulary**
- Generate **template** based on RNN using visual features as input
- Complete the text using slot filling

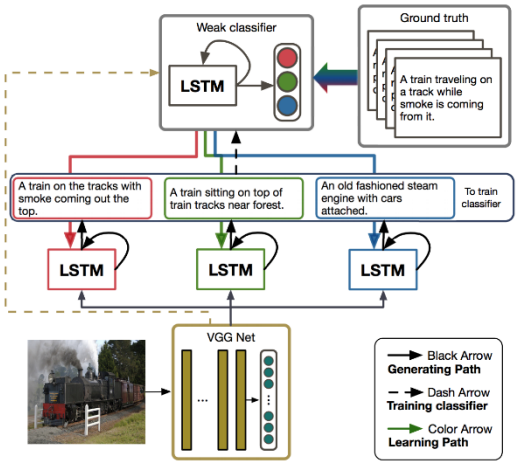


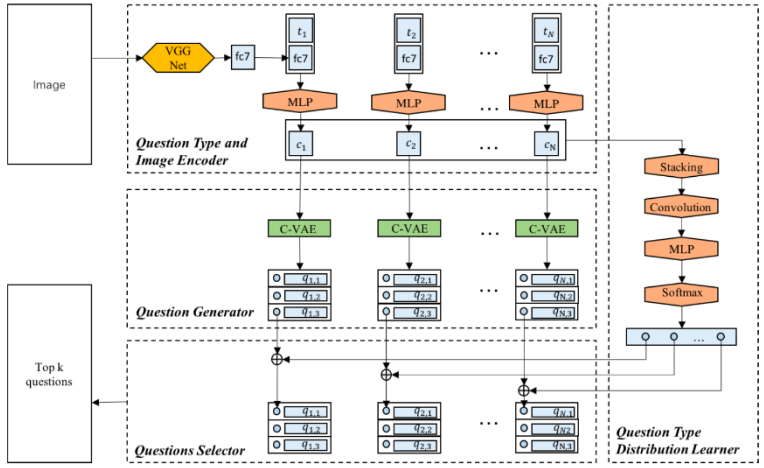
Neural Baby Talk, CVPR'18

Diversify the generation results

- encoder side：incorporate fine-grained information
- decoder side：utilize multiple decoder
- task-related information：question type







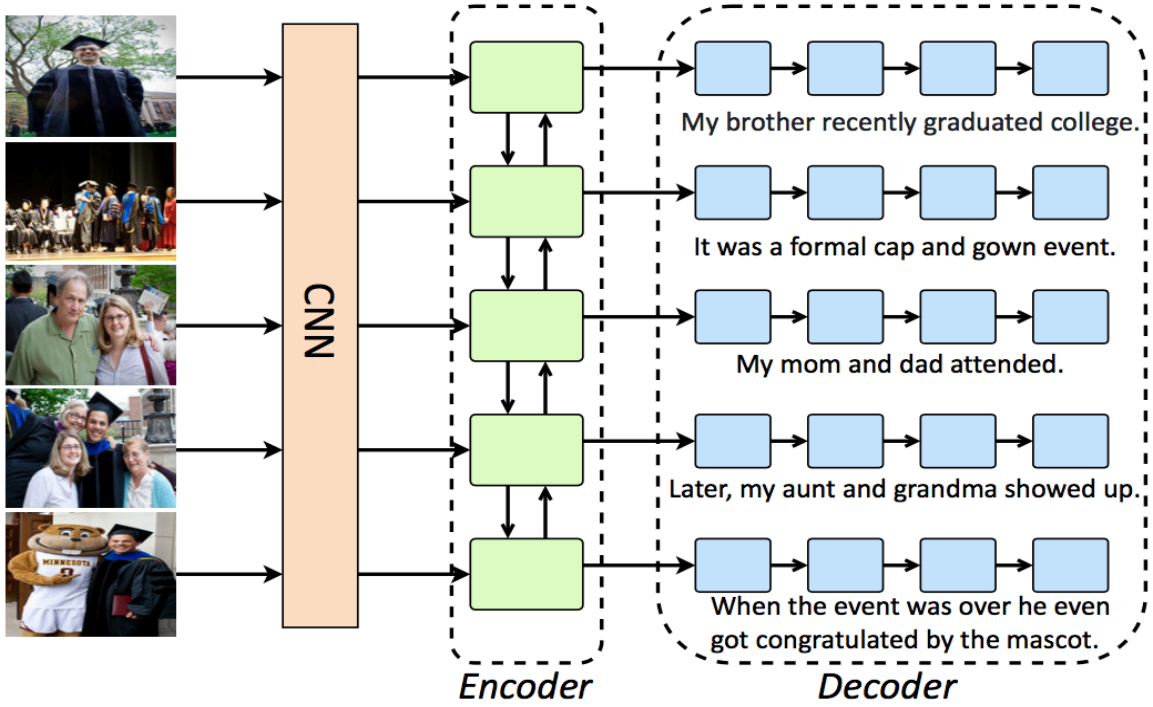
Densecap: Fully convolutional localization networks for dense captioning, CVPR'16

Diverse Image Captioning via GroupTalk, IJCAI'16

A Question Type Driven Framework to Diversify Visual Question Generation, IJCAI'18

Visual Storytelling

- Extract Visual feature for each image and update it using a RNN strcuture
- Decode sentence from each image



Wang, Xin, et al. "No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling." ACL. 2018.

Topic Modeling for Storytelling

- The sentiment of the generated story is inappropriate
- Contents are irrelevant or uninformative.

(1) (2) (3) (4) (5)

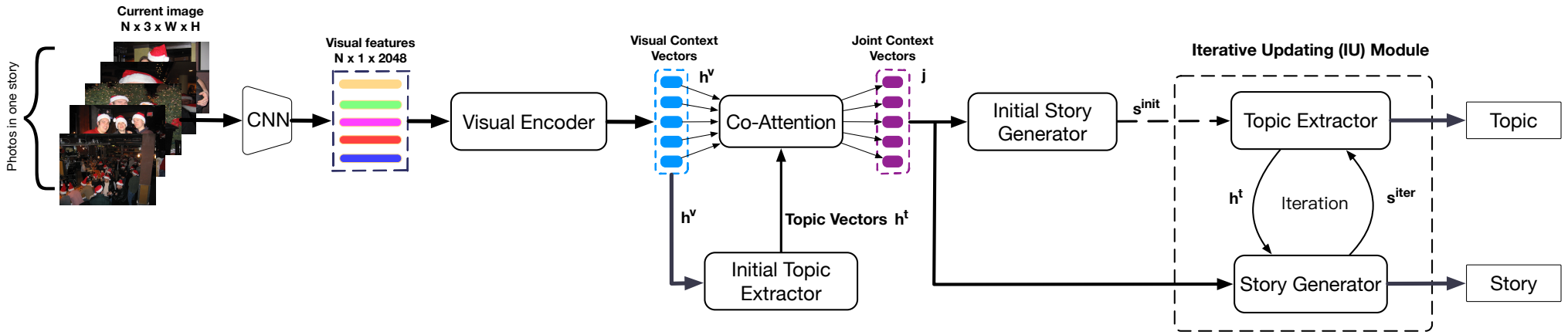
Topic: car accident

Ground-truth: i came across a terrible car accident. one of the vehicles was completely destroyed. they had to bring in a tow truck to remove the wreck. the other car was badly damaged as well. it took them a while to clear that part of the street again.

Existing methods: the police were in the accident. the cars was damaged. the police were very excited to see the car. we got to see a lot of different things at the event. the car truck was a lot of fun.

Topic-Aware Visual Storytelling via Multi-Agent Communication, COLING 2020.

Topic Modeling for Storytelling via Multi-agent Communication



- **Visual encoder:** extracts visual features form given images
- **Initial topic extractor:** takes visual features as input and generates a topic vector
- **Initial story generator**: combines the topic vector and visual features via co-attention mechanism and construct the initial version of story
- **Iterative updating module**: enforces the communication of the two agents via message passing mechanism as fine tuning

Topic-Aware Visual Storytelling via Multi-Agent Communication, COLING 2020.

Consistency evaluation

- In story sentiment consistency
- Assign sentiment score to each sub-story

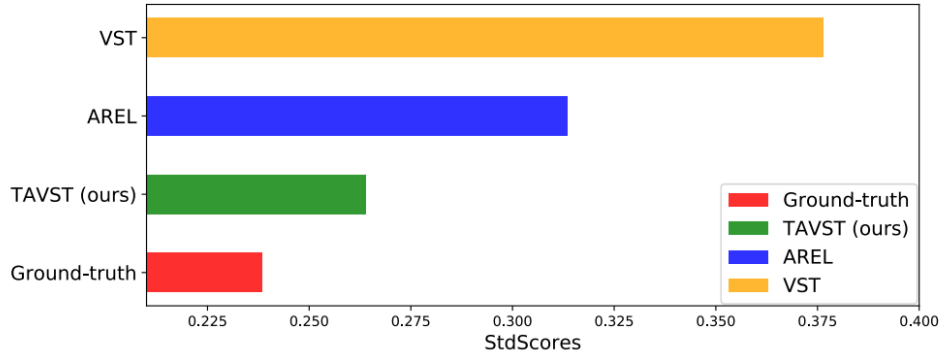


Figure 4: The comparison of in-story sentiment standard deviations among different methods.

Table 5: Sentiment scores corresponding to different types of events. Note that higher score indicates more positive polarity.

Method	Positive		Negative	
	new year's eve	sporting event	breaking up	car accident
GT	1.20	1.25	0.48	0.38
TAVST	1.31	1.17	0.53	0.42
AREL	1.22	1.15	0.95	0.81
VST	1.65	1.82	1.66	1.84

Topic-Aware Visual Storytelling via Multi-Agent Communication, COLING 2020.

Scene-graph based storytelling

- When humans telling stories for an image sequence: ① recognize the objects in each image ② reason about their visual relationships ③abstract the content into a scene ④ observe the images in order and reason the relationship among images.

- Translating each image into a graph-based semantic representation, i.e., scene graph, and reasoning the relationships on scene graphs at two levels, i.e., within-image and cross-images levels, would benefit representing and describing images.

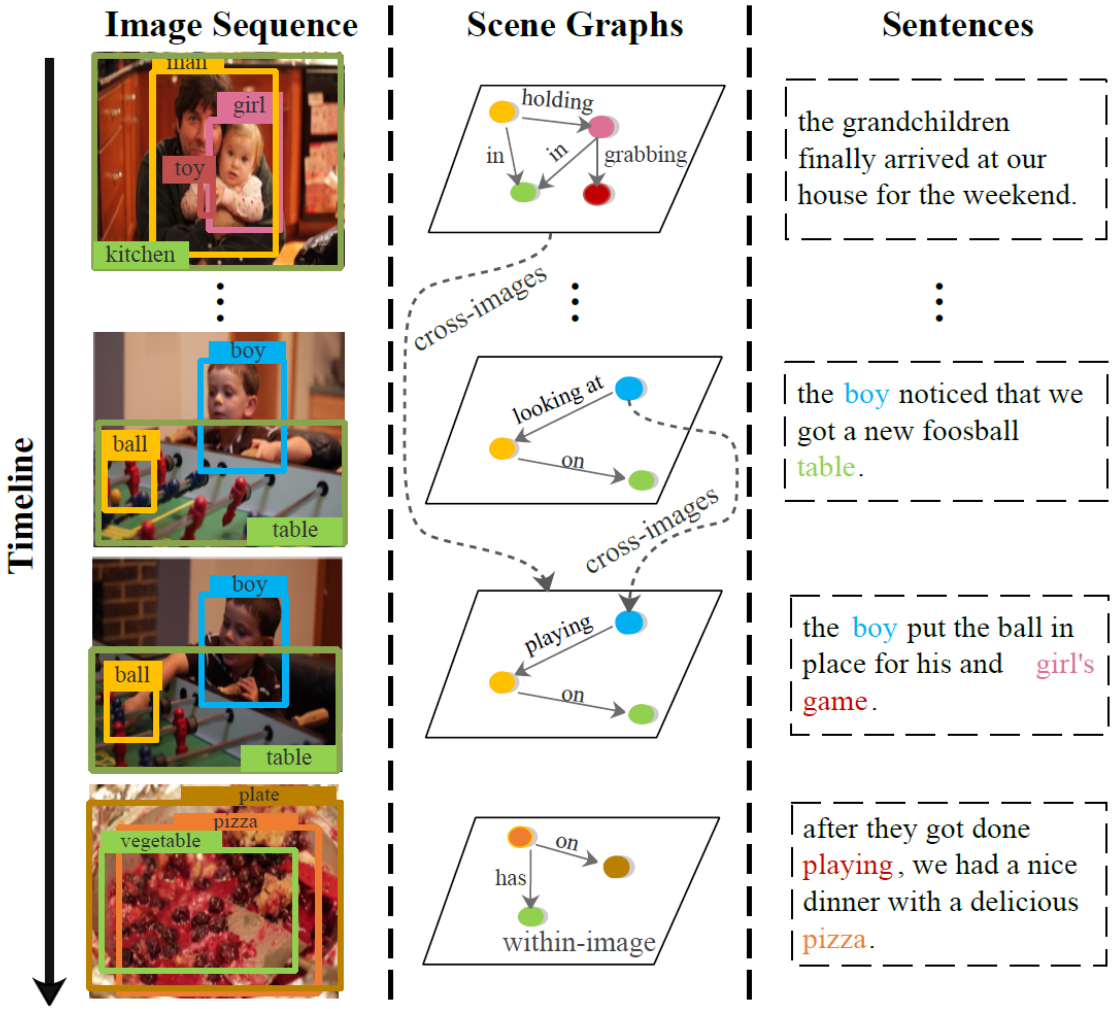
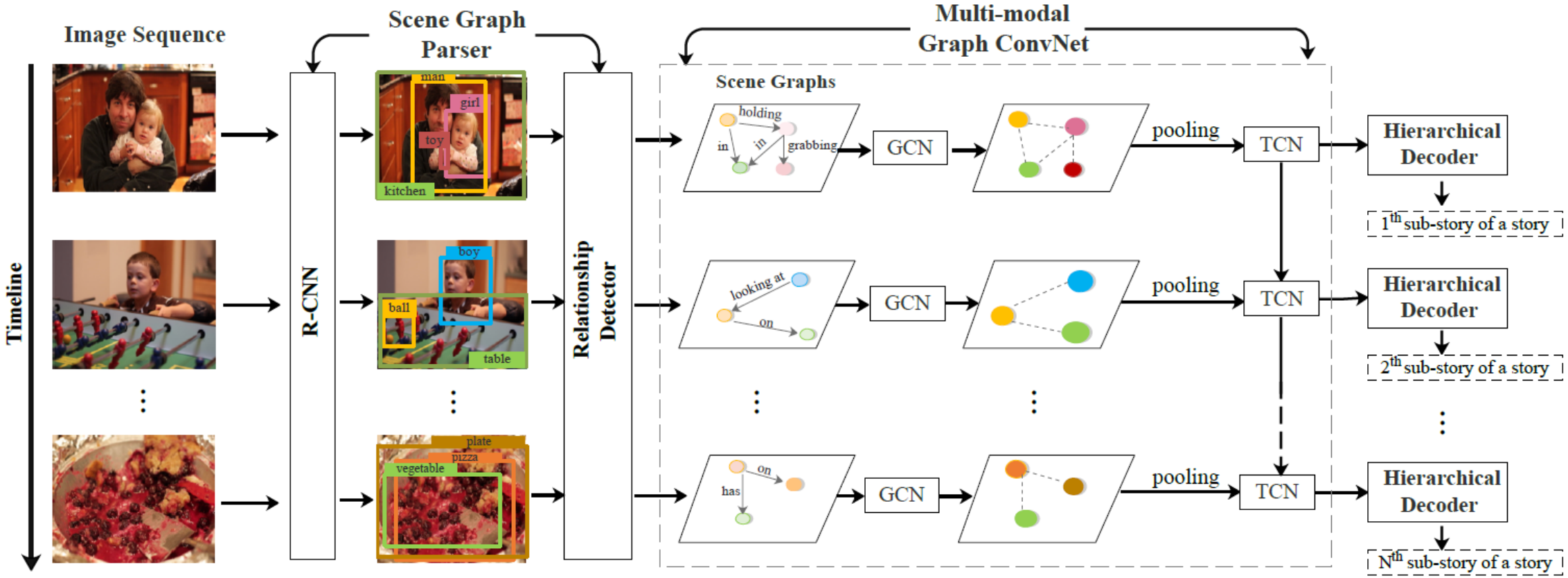


Figure 1: A scene graph based example for visual storytelling

Storytelling from an Image Stream Using Scene Graphs, AAAI' 2020

Scene-graph based storytelling



Storytelling from an Image Stream Using Scene Graphs, AAAI' 2020

Evaluation Methods

- The quality of machine-generated language depends on the assessment of two main aspects: adequacy and fluency.

- Human evaluation
  - Accuracy guaranteed, but requires human involvement which is expensive
- Rule-based automatic metrics
  - Consider only some specific features (e.g., lexical or semantic) of languages
- Learning-based automatic metrics
  - Parameterized, so there is a risk of being attacked and deceived by the image description model.

Human evaluation

- Turing Test
  - Which one is written by human?
  - Worker is given one human-annotated sample and one machine-generated sample, and needs to decide which is human-annotated.

- Pairwise Comparison
  - Which one is better?
  - Worker is presented with two generated stories from different models and asked to make decisions from different aspects (e.g., relevance, expressiveness and concreteness).

- Assign each sample to multiple workers to eliminate human variance

Rule-based automatic metrics

- BLEU $$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

brevity penalty $$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

n-grams matches (precision score) $$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

- ROUGE $$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- METEOR $$Score = Fmean * (1 - Penalty)$$

$$Fmean = \frac{10PR}{R + 9P} \qquad Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched}\right)^3$$

Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.
Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text summarization branches out*. 2004.
Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.

Rule-based automatic metrics

- CIDEr

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \text{CIDEr}_n(c_i, S_i)$$

score for n-grams

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\boldsymbol{g^n}(c_i) \cdot \boldsymbol{g^n}(s_{ij})}{\|\boldsymbol{g^n}(c_i)\| \|\boldsymbol{g^n}(s_{ij})\|}$$

TF-IDF weighting

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}\right)$$

- SPICE

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$$

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

T returns logical tuples from a scene graph

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c)$$

Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." *European Conference on Computer Vision*. Springer, Cham, 2016.

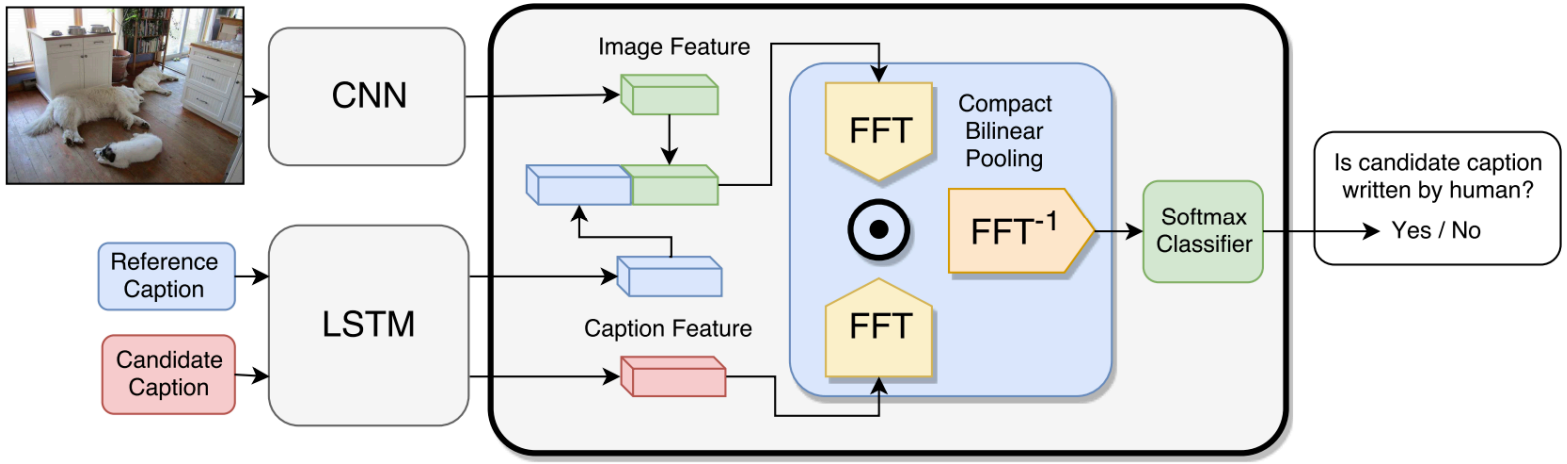
Learning-based automatic metrics

- Metrics based on measuring word overlap between candidate and reference captions find it difficult to capture semantic meaning of a sentence, therefore often lead to bad correlation with human judgments.
- Each evaluation metric has its well-known blind spot, and rule-based metrics are often inflexible to be responsive to new pathological cases.
- Popular metrics primarily evaluate a candidate caption based on references without taking image content into account, and the possible information loss caused by references may bring biases to the evaluation process.

- Learning-based method constructs a machine learning model to directly calculate the correlation between the generated caption and the given image.

LEIC: Learning to Evaluate Image Captioning

- Encode image and reference/candidate caption, combine them into a single vector
  - simply concatenates the vectors followed by a MLP or using Compact Bilinear Pooling (CBP)
- Binary classifier $\text{score}_\Theta(\hat{c}, i) = P(\hat{c} \text{ is human written} \mid \mathcal{C}(i), \Theta)$



Compact Bilinear Pooling (CBP)
- very effective in combining heterogeneous information of image and text
- uses Count Sketch to approximate the outer product between two vectors in a lower dimensional space

$$\mathbf{v} = \Phi\big(\text{concat}([\mathbf{i}, \mathbf{c}])\big) \otimes \Phi\big(\hat{\mathbf{c}}\big)$$

Figure 2. The model architecture of the proposed learned critique with Compact Bilinear Pooling. We use a deep residual network and an LSTM to encode the reference image and human caption into context vector. The identical LSTM is applied to get the encoding of a candidate caption. The context feature and the feature extracted from the candidate caption are combined by compact bilinear pooling. The classifier is supervised to perform a Turing Test by recognizing whether a candidate caption is human written or machine generated.

Cui, Yin, et al. "Learning to evaluate image captioning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

TIGEr: Text-to-Image Grounding for Image Caption Evaluation

- Consider fine-grained image content and human-generated references for evaluation.

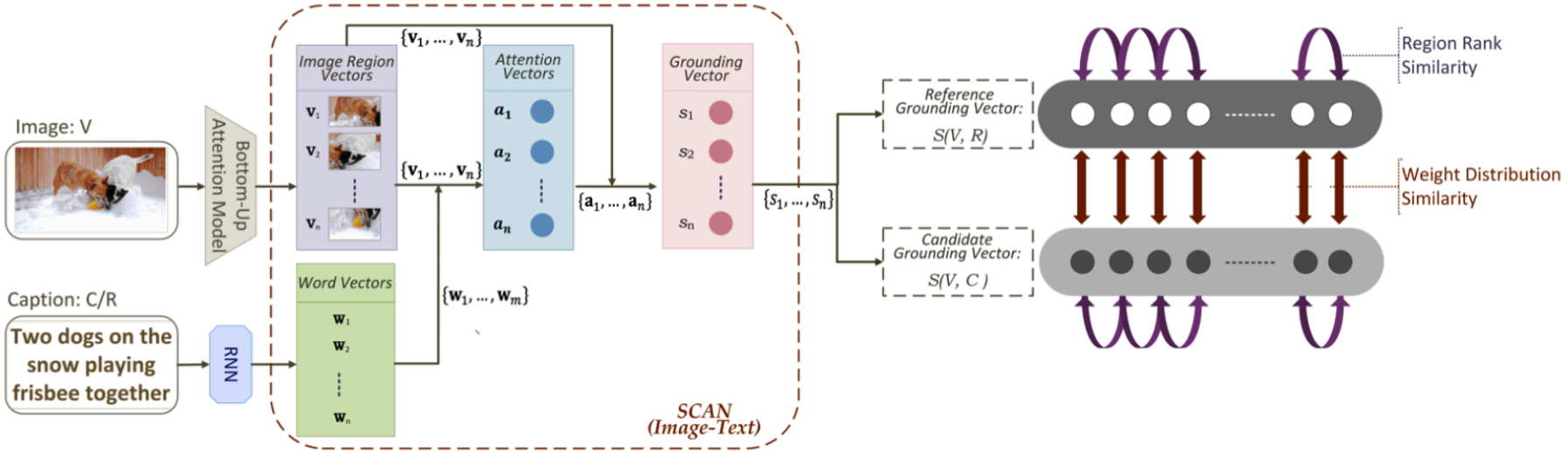


Figure 3: Overview of TIGEr calculation. For each pair of image and caption sentence, the pre-trained SCAN model generates a similarity vector where each dimension denotes the grounding relevance between an image region and the caption sentence in this region context. Given two similarity vectors denoting the grounding outcomes of reference versus candidate captions, we measure: 1) region rank disagreement, and 2) the similarity of two grounding weight distributions.

Jiang, Ming, et al. "TIGEr: Text-to-Image Grounding for Image Caption Evaluation." *EMNLP 2019*

$$\text{TIGEr}_{(V,C,\mathcal{R})} := \frac{\text{RRS}_{(V,C,\mathcal{R})} + \text{WDS}_{(V,C,\mathcal{R})}}{2}$$

Problems of training process for current seq-to-seq framework

- Exposure bias: mismatch between training and testing
  - Existing text generation models are typically trained to predict the next word in a sequence, given the previous words and visual information
  - Training: correct words
  - Testing: words generated in previous step

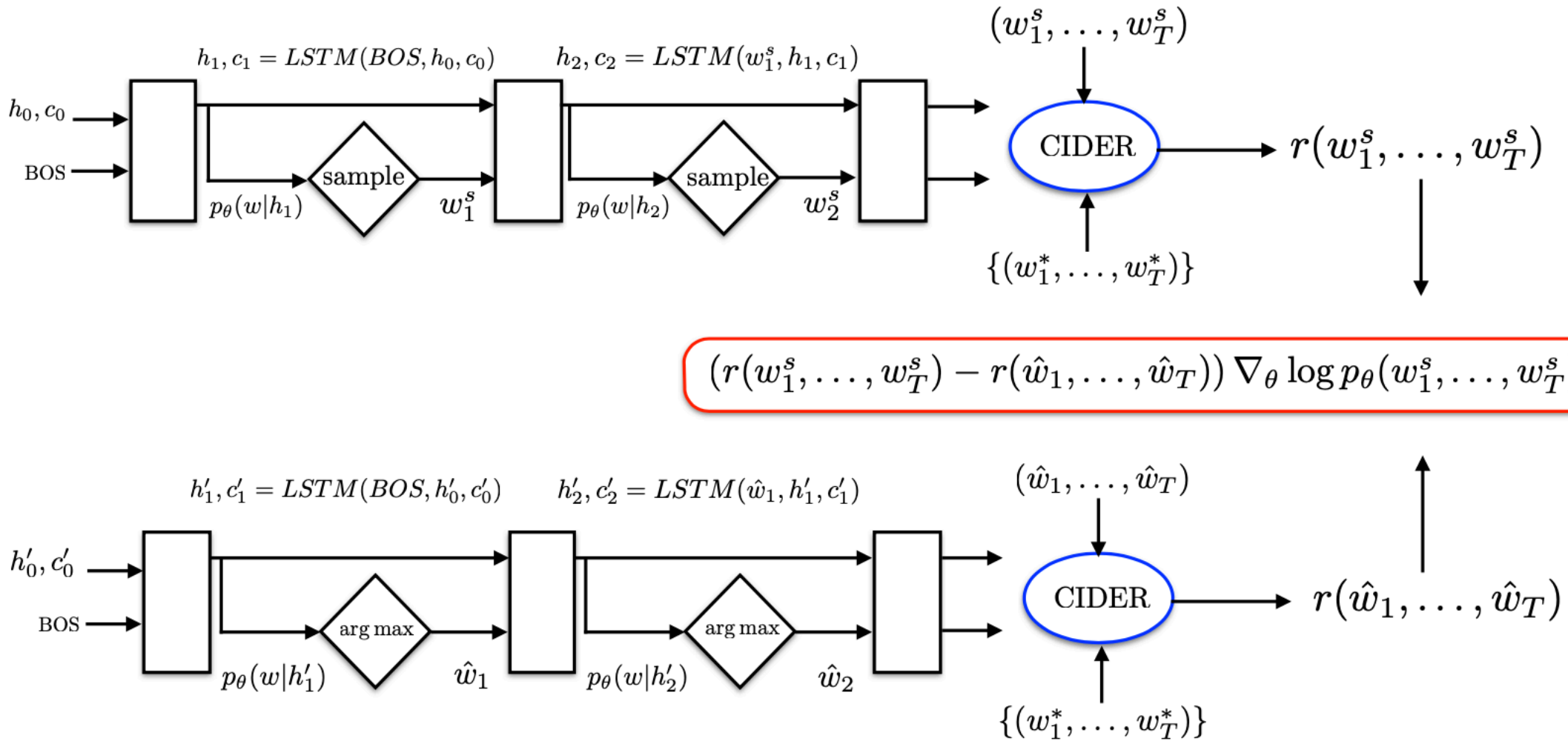
- Mismatch of training and testing objectives
  - Seq-to-seq models are trained with cross entropy loss
  - Evaluated at test time using discrete and non-differentiable metrics such as BLEU , ROUGE, METEOR or CIDEr

Training with Reinforce Algorithm

- Cross-Entropy Training
  - maximizes the probability of the observed sequence word by word:
  - $L = -logp(w_1, w_2, \dots, w_T) = -\sum_{t=1}^{T} logp(w_t | w_1, \dots, w_{t-1})$
- REINFORCE
  - Loss is negative expected reward:
  - $L_\theta = -\sum_{w_1^g, \dots, w_T^g} p_\theta(w_1^g, \dots, w_T^g) r(w_1^g, \dots, w_T^g) = -\mathbb{E}_{[w_1^g, \dots, w_T^g] \sim p_\theta} r(w_1^g, \dots, w_T^g)$
  - REINFORCE trains model at the sequence level using *any* user-defined reward.

Ranzato, et al., Sequence level training with recurrent neural networks ,ICLR2016

Method –Self-critical Sequence Training



Ranzato, et al., Sequence level training with recurrent neural networks ,ICLR2016

Method –Mixed Incremental Cross-Entropy Reinforce

- Incremental learning
- A hybrid loss function combining both REINFORCE and Cross-Entropy

Data: a set of sequences with their corresponding context.
Result: RNN optimized for generation.
Initialize RNN at random and set N^{XENT}, $N^{\text{XE+R}}$ and Δ;
for $s = T, 1, -\Delta$ **do**
  if $s == T$ **then**
    train RNN for N^{XENT} epochs using XENT only;
  else
    train RNN for $N^{\text{XE+R}}$ epochs. Use XENT loss in the first s steps, and REINFORCE (sampling from the model) in the remaining $T - s$ steps;
  end
end

Algorithm 1: MIXER pseudo-code.

Ranzato, et al., Sequence level training with recurrent neural networks ,ICLR2016

Adversarial Reward Learning for RL

- This is an adversarial example with an average METEOR score as high as 40.2:

We had a great time to have a lot of the. They were to be a of the. They were to be in the. The and it were to be the. The, and it were to be the.

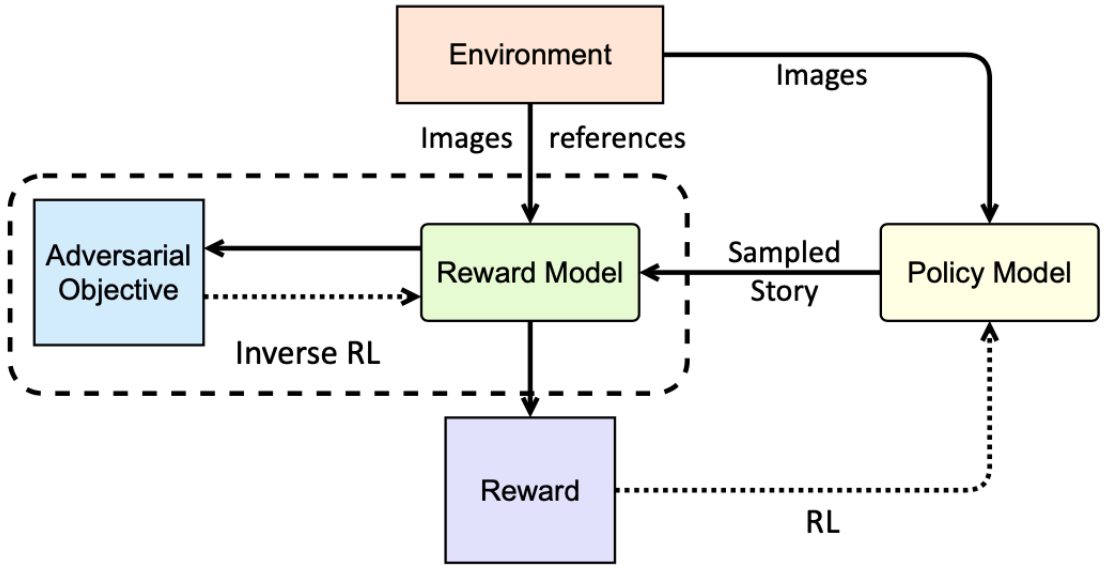


Figure 2: AREL framework for visual storytelling.

Wang, et al., No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling , ACL2018

Outline

- Cross Vision and Language Matching
- Vision-based Text Generation
- **Cross Vision and Language Reasoning**
- Language-based Vision Navigation

- Cross-modality Pretraining

Cross Vision and Language Reasoning

- Visual Question Answering (VQA)
- Visual Commonsense Reasoning (VCR)
- Visual Commonsense Reasoning in Time (Visual COMET)

Visual Question Answering (VQA)

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

How many children are in the bed?

2

1

VQA2.0

Question : Is there a clock on the tower?

Original Image | **yes**

Complementary Image | **no**

Question : How many sheep are there?

Original Image | **2**

Complementary Image | **8**

Question : What is the first letter of the third word on this sign?

Original Image | **w**

Complementary Image | **r**

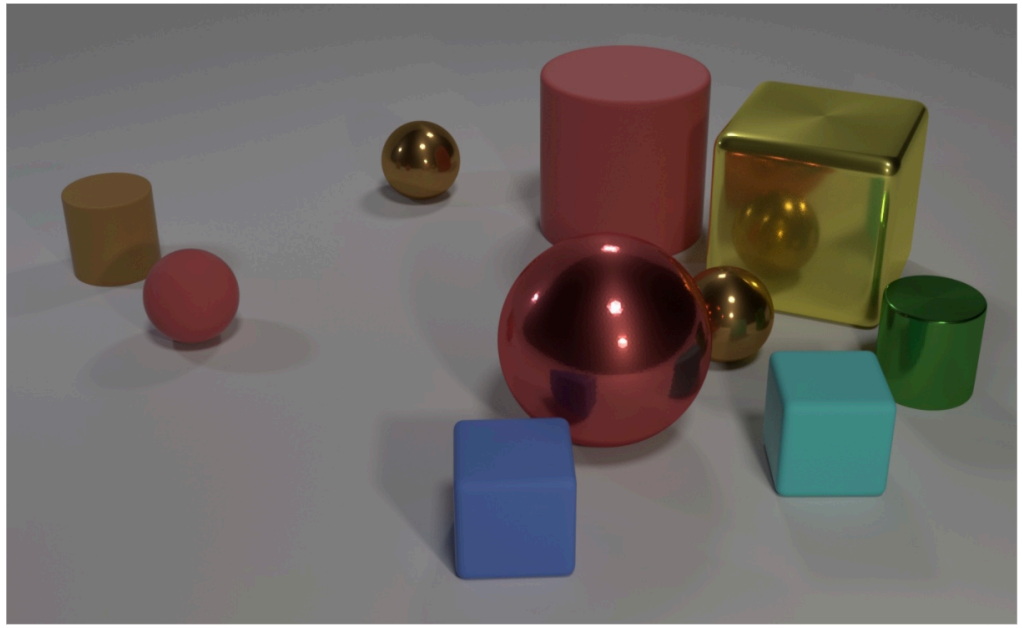
Question : What is everyone sitting down doing?

Original Image | **using computer**

Complementary Image | **talking**

Goyal, Yash, et al. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

CLEVR



Q: Are there an **equal number** of **large things** and **metal spheres**?

Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?

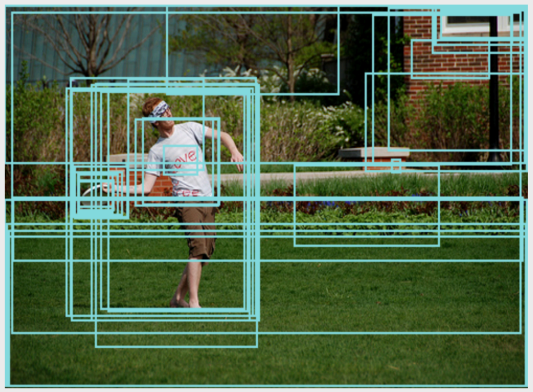
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?

Q: **How many** objects are **either** **small cylinders** or **red** things?

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In ICCV, 2017b.

Visual Genome

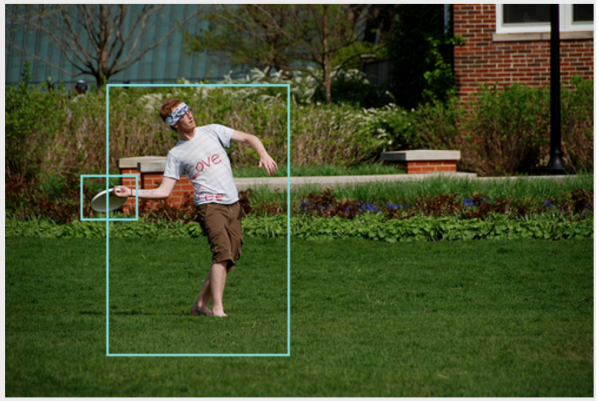
Regions	Attributes	Relationships
window has a white frame	window is white	man playing frisbee
green plants on side green grass	frame is white	building behind player
the letter LOVE on a tan shirt	plants is green	home OF bricks
man is bend backward	grass is green	window has frame
The grass is green.	shirt is tan	plants on side of grass
The grass is short.	grass is short	love ON shirt
The man is holding a frisbee.	tree is leafy	man holding frisbee
The frisbee is white.	bushes is green	man throwing frisbee
The frisbee is round.	flowers is purple	man standing in grass
	brick is red	
	brick is white	
	shirt is gray	



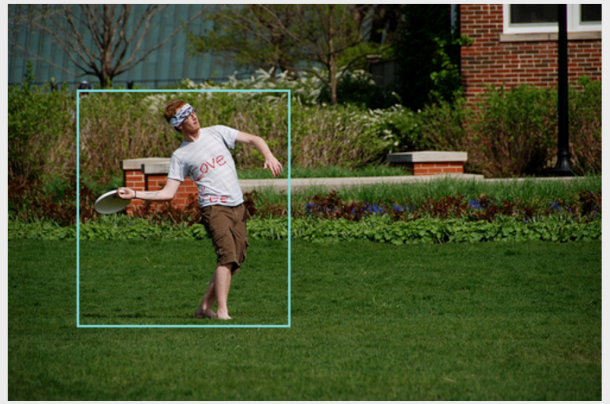
Question Answers

Why is the man leaning back?	He is getting ready to throw the frisbee.
What is on the man's forehead?	A bandana.
Where is the black pole?	The middle right of the photo.
What is the man holding?	A frisbee.
How many people are in the photo?	1.

Regions	Attributes	Relationships
window has a white frame	window is white	man playing frisbee
green plants on side green grass	frame is white	building behind player
the letter LOVE on a tan shirt	plants is green	home OF bricks
man is bend backward	grass is green	window has frame
The grass is green.	shirt is tan	plants on side of grass
The grass is short.	grass is short	love ON shirt
The man is holding a frisbee.	tree is leafy	man holding frisbee
The frisbee is white.	bushes is green	man throwing frisbee
The frisbee is round.	flowers is purple	man standing in grass
	brick is red	
	brick is white	
	shirt is gray	



Regions	Attributes	Relationships
window has a white frame	window is white	man playing frisbee
green plants on side green grass	frame is white	building behind player
the letter LOVE on a tan shirt	plants is green	home OF bricks
man is bend backward	grass is green	window has frame
The grass is green.	shirt is tan	plants on side of grass
The grass is short.	grass is short	love ON shirt
The man is holding a frisbee.	tree is leafy	man holding frisbee
The frisbee is white.	bushes is green	man throwing frisbee
The frisbee is round.	flowers is purple	man standing in grass
	brick is red	
	brick is white	
	shirt is gray	



Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.

GQA

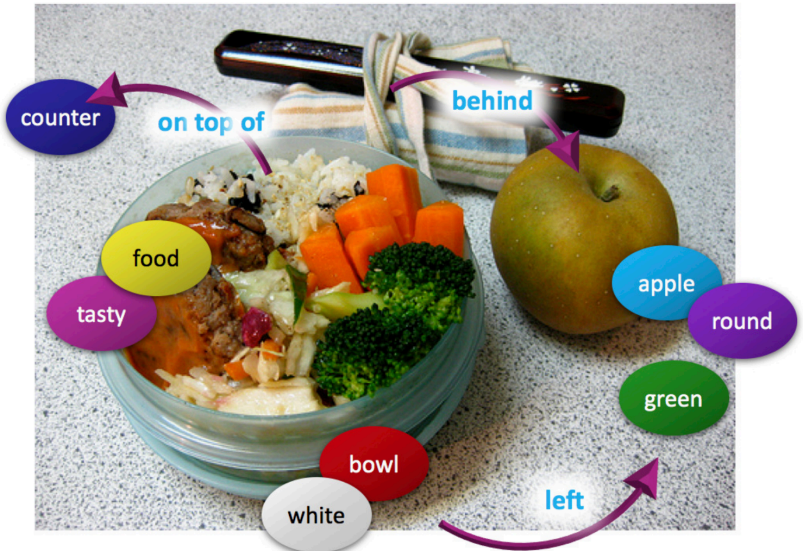
1. Is the **tray** on top of the **table** black or light brown? light brown
2. Are the **napkin** and the **cup** the same color? yes
3. Is the small **table** both oval and wooden? yes
4. Is the **syrup** to the left of the **napkin**? yes
5. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
6. Are there any **cups** to the left of the **tray** that is on top of the **table**? no
7. Could this room be a **living room**? yes

VQA

1. Does this man need a haircut?
2. What color is the guy's tie?
3. What is different about the man's suit that shows this is for a special occasion?

GQA

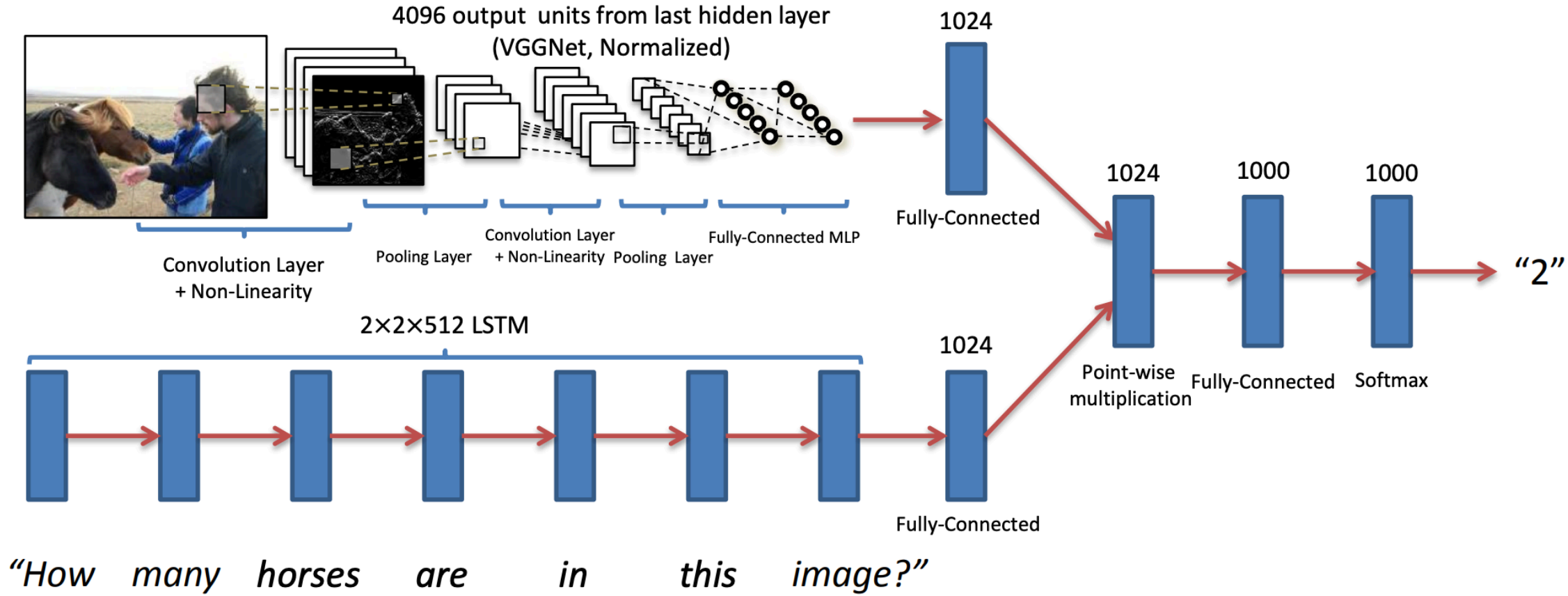
1. Is the person's hair long and brown?
2. What appliance is to the left of the man?
3. Who is in front of the refrigerator on the left?
4. Is there a necktie in the picture that is not red?
5. Is the color of the vest different than shirt?



Datasets

Datasets	Image	Text	Characteristic
VQA2.0(2015)	204,721(coco)	1,105,904	10 persons annotate answers; Eval: yes/no, number, other
CLEVR(2016)	100,000	864,968	Synthetic; Reason about relationships between objects of different shapes, colors and sizes
Visual Genome(2016)	108,077(coco,flickr)	1,445,322	Region based qa-pair and region based caption, scene graph, object detection with annotated attribute
GQA(2019)	113,018(coco,flickr, visual genome)	22,669,678	Unbalanced data; scene graph based; full answer; word-object mapping

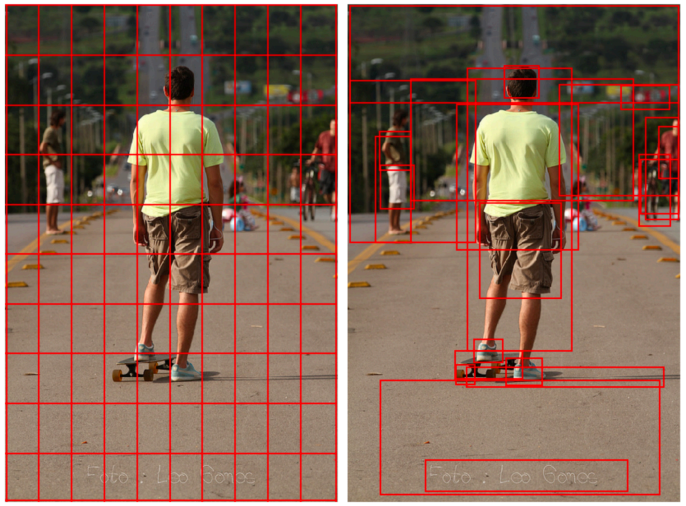
Baseline



Antol S, Agrawal A, Lu J, et al. VQA: Visual question answering[C] //Proceedings of the IEEE International Conference on Computer Vision. 2015: 2425-2433.

Bottom-Up and Top-Down Attention

- Top-down: pixel based vision representation
- Bottom-up: object based vision representation



Top-down Bottom-up

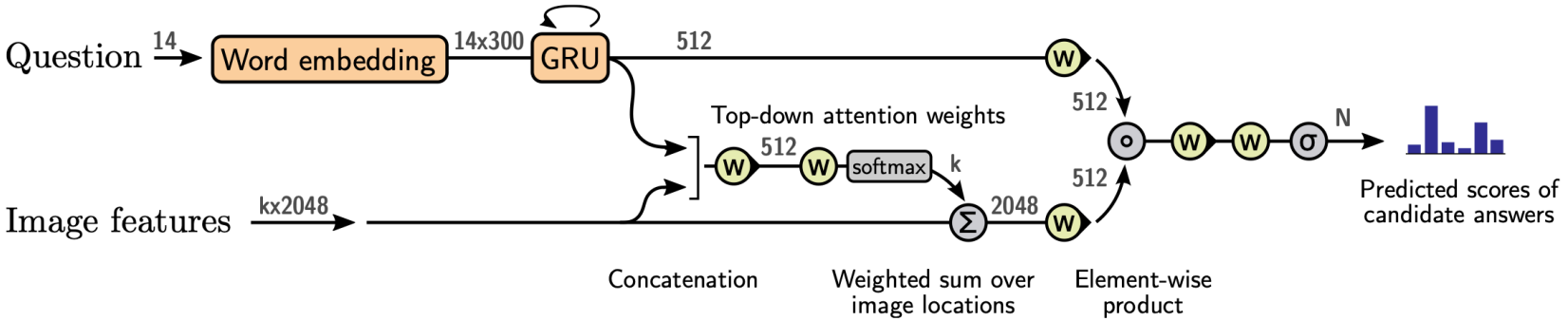


Figure 4. Overview of the proposed VQA model. A deep neural network implements a joint embedding of the question and image features $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_k\}$. These features can be defined as the spatial output of a CNN, or following our approach, generated using bottom-up attention. Output is generated by a multi-label classifier operating over a fixed set of candidate answers. Gray numbers indicate the dimensions of the vector representations between layers. Yellow elements use learned parameters.

Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Bottom-Up and Top-Down Attention

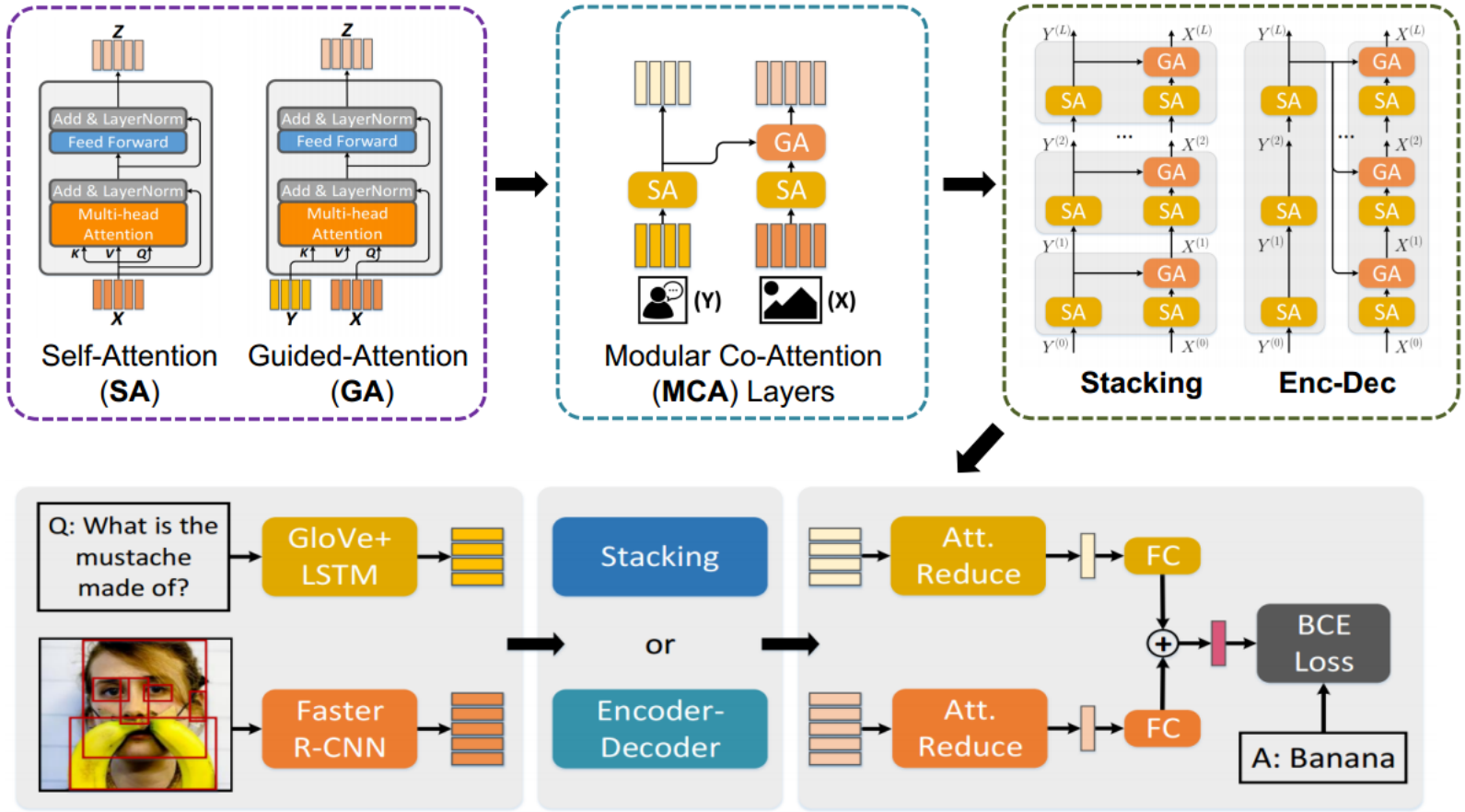
	Yes/No	Number	Other	Overall
d-LSTM+n-I [26, 12]	73.46	35.18	41.83	54.22
MCB [11, 12]	78.82	38.28	53.36	62.27
UPMC-LIP6	82.07	41.06	57.12	65.71
Athena	82.50	44.19	59.97	67.59
HDU-USYD-UNCC	84.50	45.39	59.01	68.09
Ours: Up-Down	**86.60**	**48.64**	**61.15**	**70.34**

Table 5. VQA v2.0 test-standard server accuracy as at 8 August 2017, ranking our submission against published and unpublished work for each question type. Our approach, an ensemble of 30 models, outperforms all other leaderboard entries.

Antol S, Agrawal A, Lu J, et al. VQA: Visual question answering[C] //Proceedings of the IEEE International Conference on Computer Vision. 2015: 2425-2433.

Deep Modular Co-Attention Network (MCAN)

- Transformer-based image encoding
- Co-attention mechanism within transform node to enforce information exchange between two madality



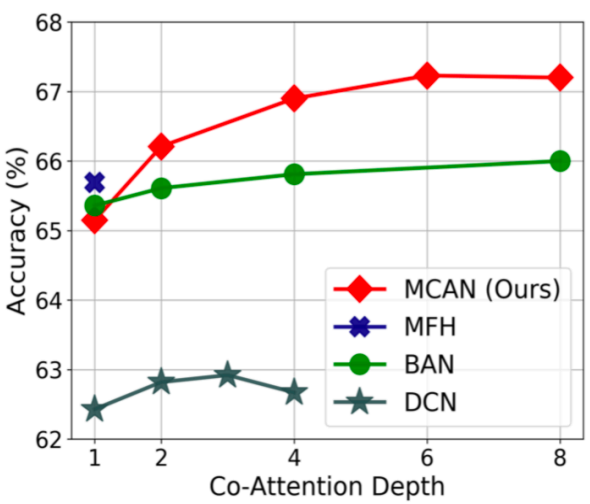
Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

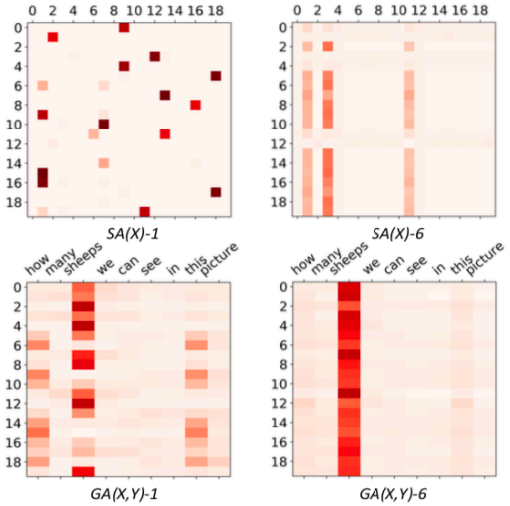
Deep Modular Co-Attention Network (MACN)

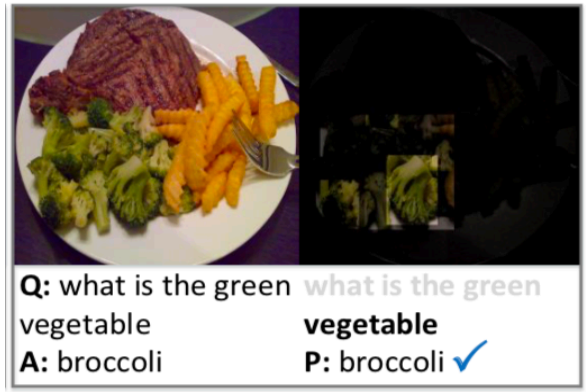
- Experiments

Table 2: Accuracies of **single-model** on the *test-dev* and *test-standard* splits to compare with the state-of-the-art methods. All the methods use the same bottom-up attention visual features [1], and are trained on the *train+val+vg* sets (*vg* denotes the augmented VQA samples from Visual Genome). The best results on both splits are bolded.

Model	Test-dev				Test-std
	All	Y/N	Num	Other	All
Bottom-Up [28]	65.32	81.82	44.21	56.05	65.67
MFH [33]	68.76	84.27	49.56	59.89	-
BAN [14]	69.52	85.31	50.93	60.26	-
BAN+Counter [14]	70.04	85.42	**54.04**	60.52	70.35
MCAN_{ed}-6	**70.63**	**86.82**	53.26	**60.72**	**70.90**







Neural State Machine

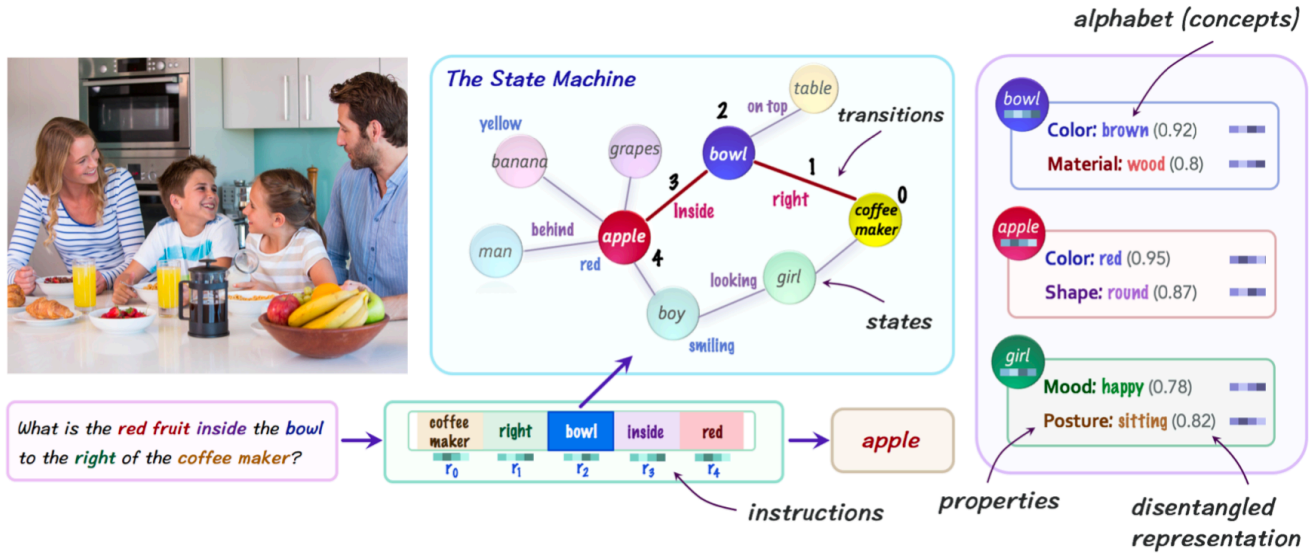


Figure 1: The Neural State Machine is a graph network that simulates the computation of an automaton. For the task of VQA, the model constructs a probabilistic scene graph to capture the semantics of a given image, which it then treats as a state machine, traversing its states as guided by the question to perform sequential reasoning.

Two stages of **learning** and **inference:**

1) Modeling: transforms the raw inputs into **abstract** semantic representations, and ***construct the state machine***.

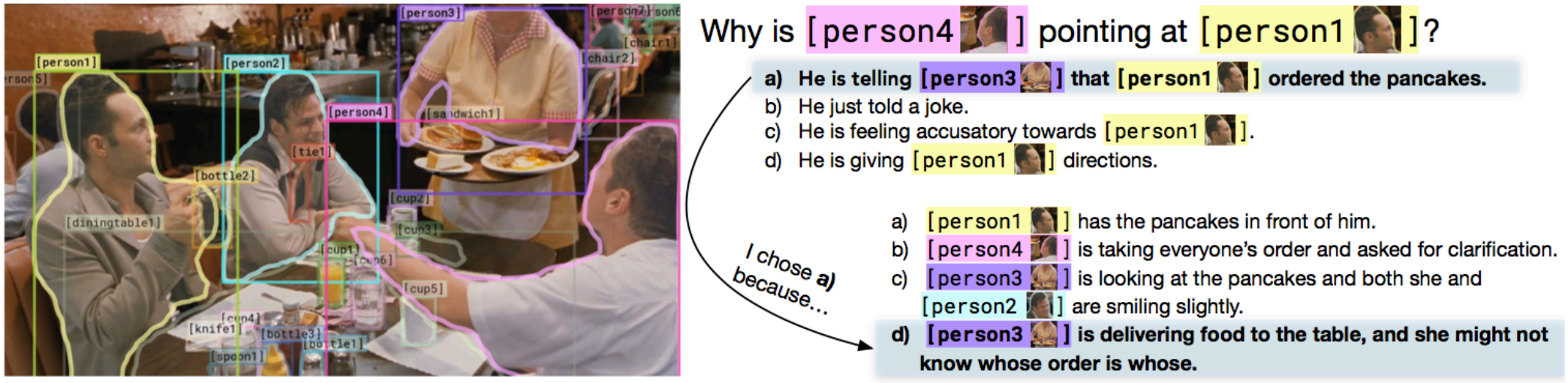
Image -> Scene graph Question -> Instructions

2) Inference: ***simulates an iterative computation*** over the machine, sequentially traversing the states until completion.

Reasoning over scene graph to compute an answer

Hudson, Drew and Manning, Christopher D. Learning by Abstraction: The Neural State Machine. In Advances in Neural Information Processing Systems(NIPS), 2019

Visual Commonsense Reasoning (VCR)



- **Task:** Visual Commonsense Reasoning (VCR). Given an image, objects, a question, and four answer choices. Ask the model to decide which answer choice is correct. Then, it's given four rationale choices, and it has to decide which of those is the best rationale that explains why its answer is right.
- **Dataset:** VCR, consisting of 290k multiple choice QA problems derived from 110k movie scenes, built using Adversarial Matching.

Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." ICCV. 2019.

Visual Commonsense Reasoning (VCR)

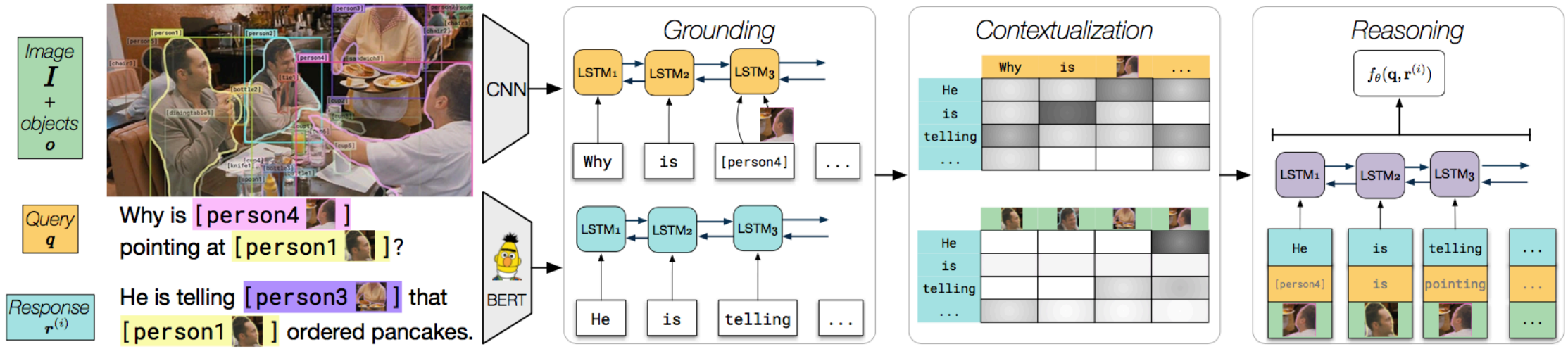


Figure 5: High-level overview of our model, **R2C**. We break the challenge of Visual Commonsense Reasoning into three components: grounding the query and response, contextualizing the response within the context of the query and the entire image, and performing additional reasoning steps on top of this rich representation.

- Given an image *I*, a set of objects *o*, a query *q*, and a set of responses $r^{(i)}$
- Query and the response contain a mixture of tags (pointing to image regions) and natural language words

Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." ICCV. 2019.

Experiments

- Settings
  - $(Q \rightarrow AR)$: $(Q \rightarrow A)$ and $(QA \rightarrow R)$
- Results on VCR

	Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
	Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text Only	BERT	53.8	53.9	64.1	64.5	34.8	35.0
Text Only	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
Text Only	ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
Text Only	LSTM+ELMo	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA [38]	39.4	40.5	34.0	33.7	13.5	13.8
VQA	BottomUpTopDown[4]	42.8	44.1	25.1	25.1	10.7	11.0
VQA	MLB [42]	45.5	46.2	36.1	36.8	17.0	17.2
VQA	MUTAN [6]	44.4	45.5	32.0	32.2	14.6	14.6
	R2C	**63.8**	**65.1**	**67.2**	**67.3**	**43.1**	**44.0**
	Human		91.0		93.0		85.0

Visual COMET

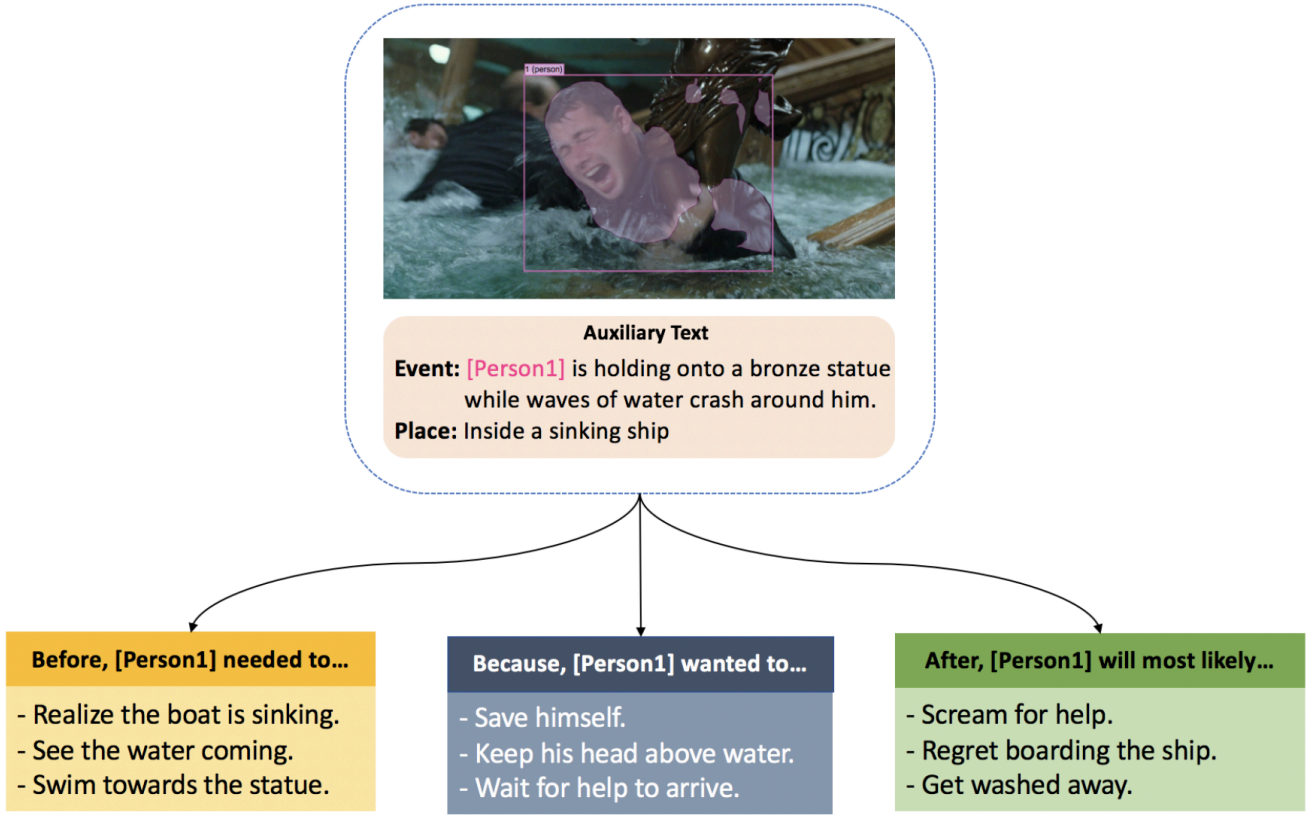


Fig. 2: **Task Overview:** Our proposed task is to generate commonsense inferences of **events before**, **events after** and **intents at present**, given an image, a description of an **event at present** in the image and a plausible scene / location of the image.

Task

(1) Given an image and one of the events at present, to generate the rest of visual commonsense graph that is connected to the specific current event.

(2) Given an image, to generate the complete set of commonsense inferences from scratch.

	Train	Dev	Test	Total
# Images/Places	47,595	5,973	5,968	**59,356**
# Events at Present	111,796	13,768	13,813	**139,377**
# Inferences on Events Before	467,025	58,773	58,413	584,211
# Inferences on Events After	469,430	58,665	58,323	586,418
# Inferences on Intents at Present	237,608	28,904	28,568	295,080
# Total Inferences	1,174,063	146,332	145,309	**1,465,704**

Table 1: **Statistics** of our Visual Commonsense Graph repository: there are in total 139,377 distinct Visual Commonsense Graphs over 59,356 images involving 1,465,704 commonsense inferences.

Park J S , Bhagavatula C , Mottaghi R , et al. VisualCOMET: Reasoning about the Dynamic Context of a Still Image[J]. 2020.

Visual COMET

Inputs for each image:

- a sequence of visual embeddings V representing the image and people detected in the image
- grounded event description e
- scene's location information p
- inference type r

Outputs:

- Event description after

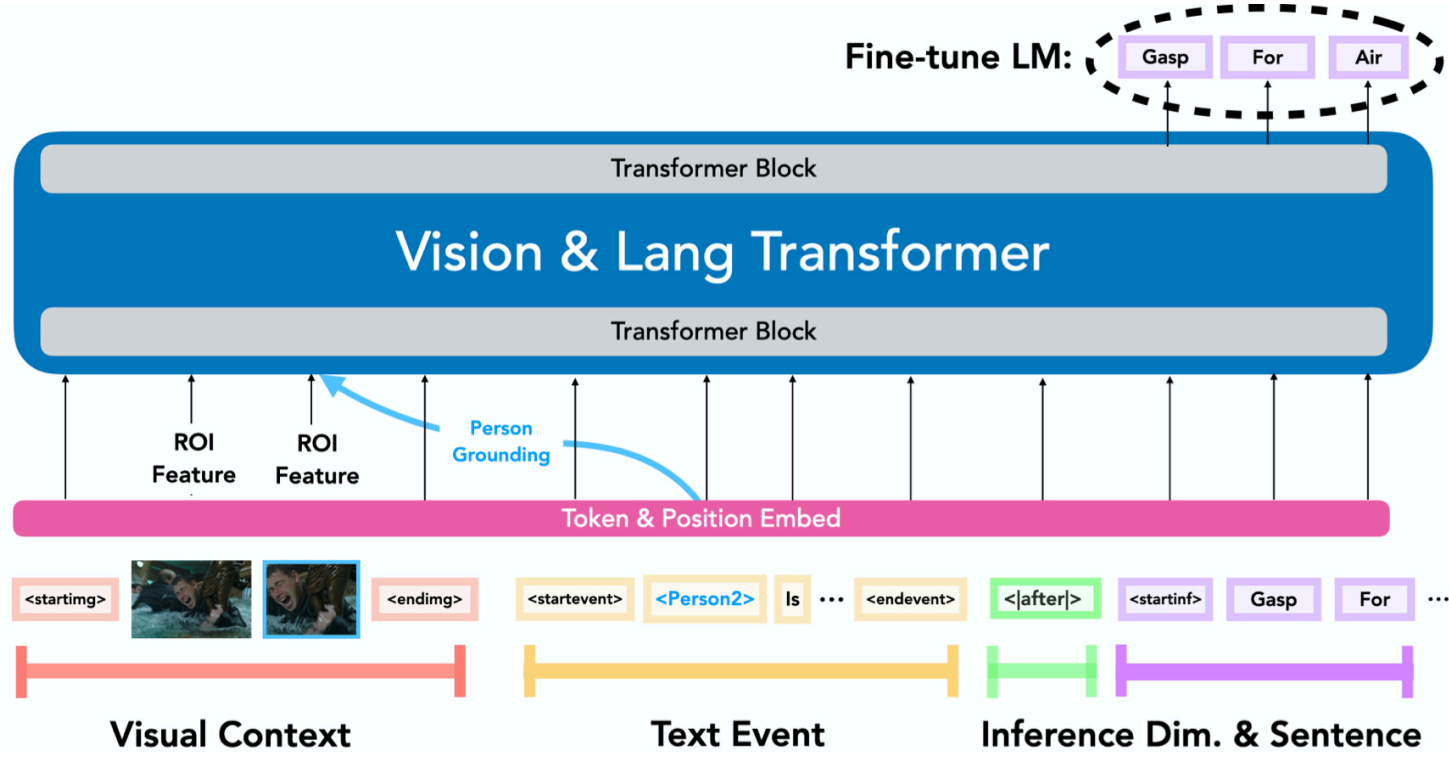


Fig. 5: **Model Overview.** Vision-Language Transformer for our approach. Our sequence of inputs uses special tokens indicating the start and end of image, event, place, and inference. We only show the start token in the figure for simplicity.

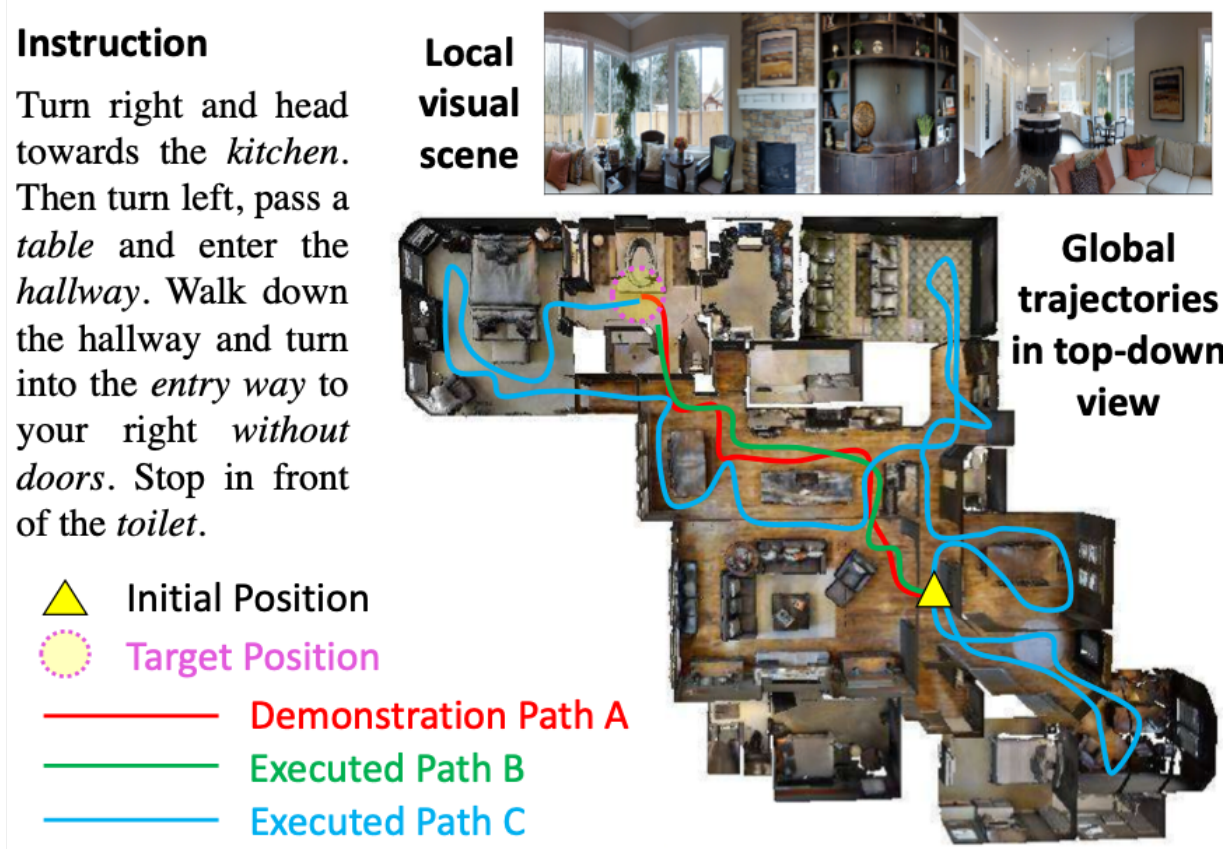
Outline

- Cross Vision and Language Matching
- Vision-based Text Generation
- Cross Vision and Language Reasoning
- **Language-based Vision Navigation**

- Cross-modality Pretraining

Vision-Language Navigation

- Natural language instructions
- An environment simulator
- A navigation agent: take actions after perceiving the environment following instructions
- Start position
- Target position



Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments, CVPR 2018

Challenges

- It is difficult for the agent to relate the language instructions to the visual environment.

  → How to exploit cross-modal information ?

- Past actions affect the actions to be taken in the future.

  → How to memorize past actions ?

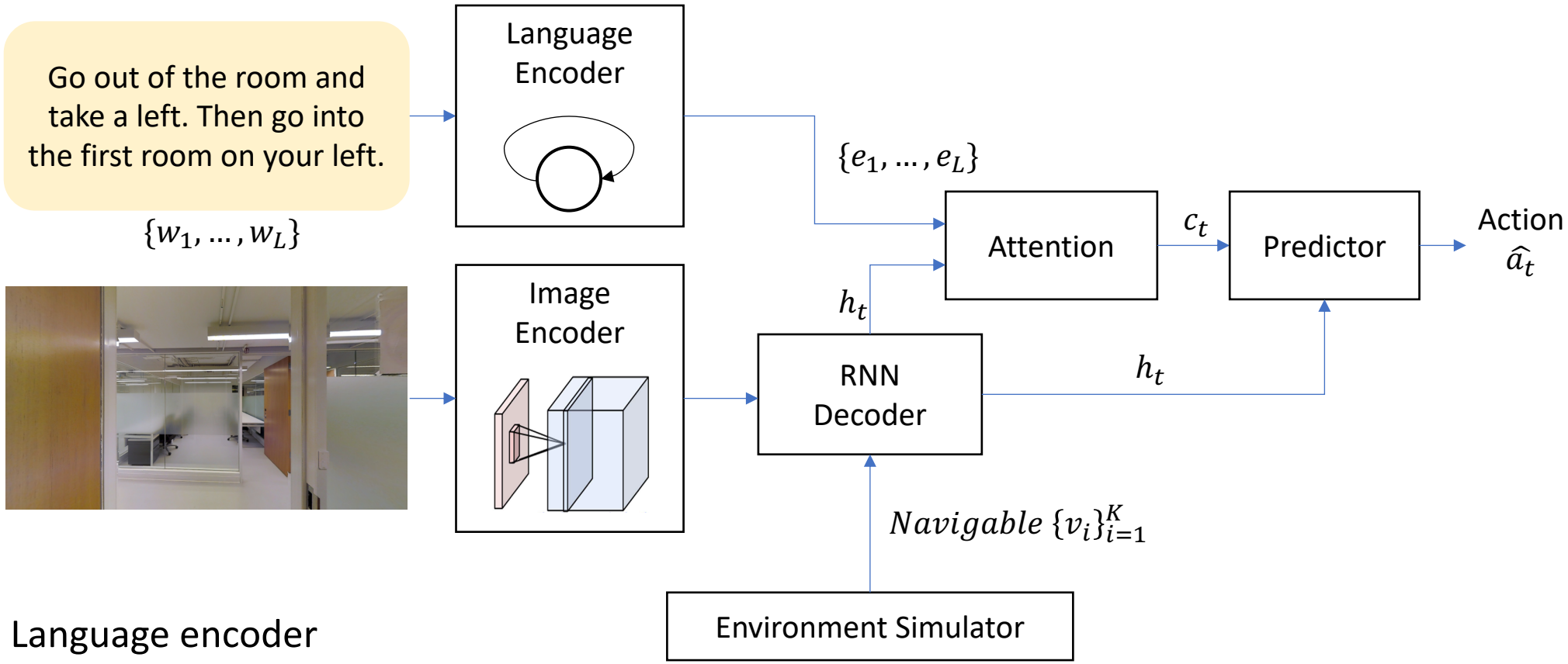
- The agent is not able to accurately assess the progress it has made.

  → How to penalize / reward the intermediate actions ?

- Existing work suffers from the generalization problem, causing a huge performance gap between seen and unseen environments.

  → Regularization / Data Argumentation / Self-supervised Learning …

Seq2Seq based Model



- Language encoder
- Image encoder
- Environment Simulator
- RNN decoder to generate action one by one

Action Space Explanation

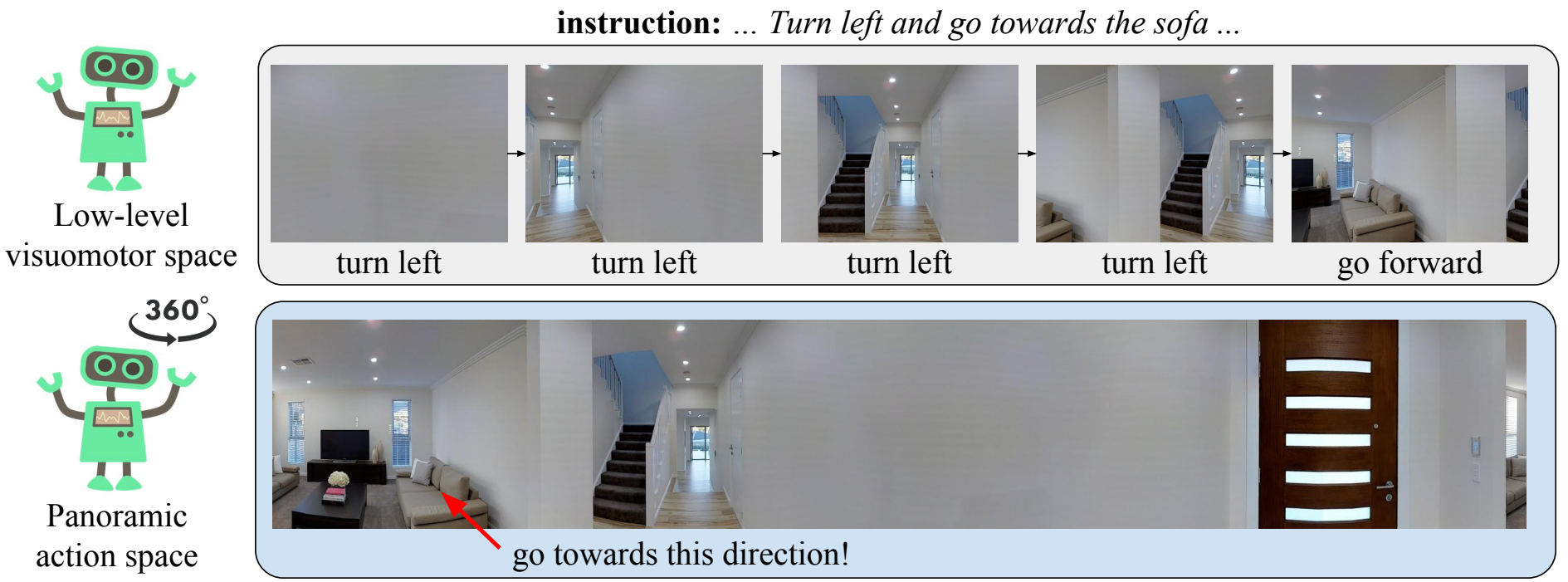
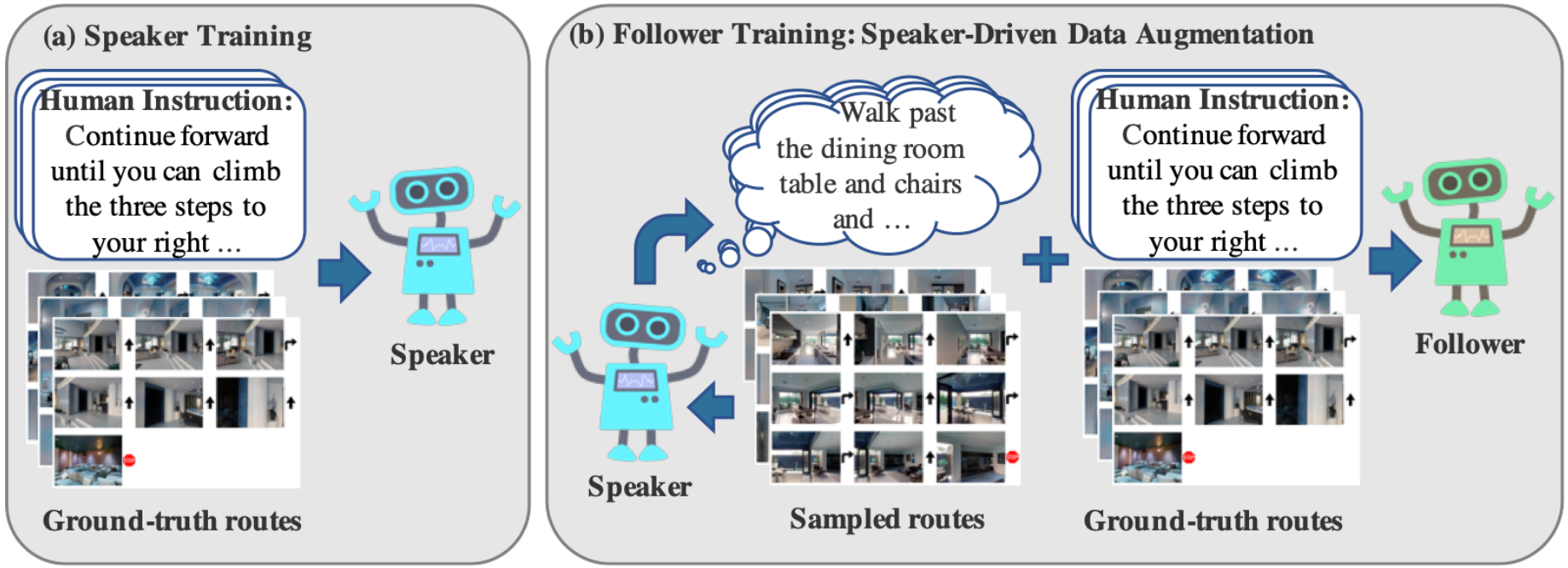


Figure 3: Compared with low-level visuomotor space, our panoramic action space (Sec. 3.3) allows the agents to have a complete perception of the scene, and to directly perform high-level actions.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor BergKirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-andlanguage navigation. NIPS.

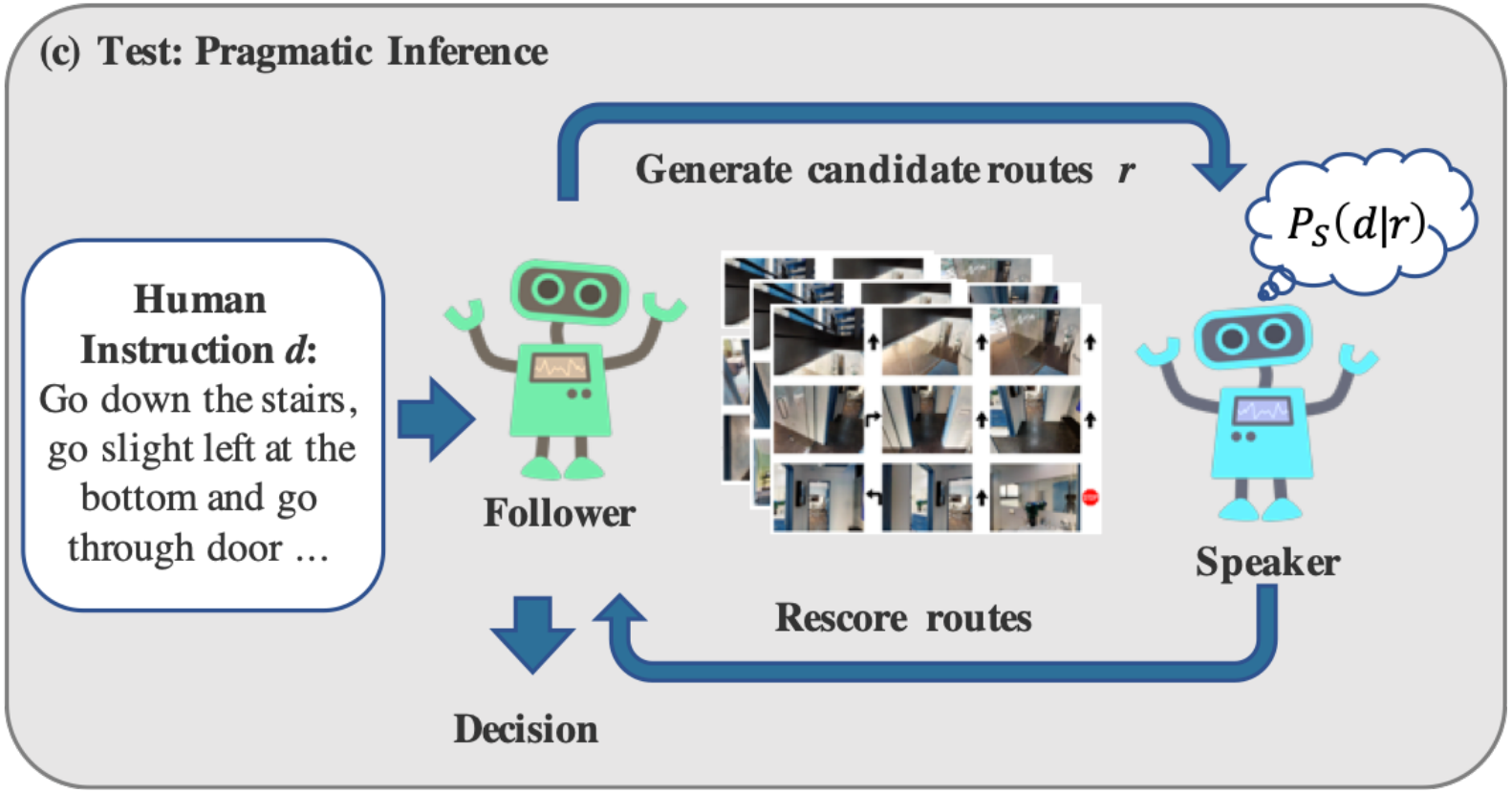
Speaker-Follower

- Treat the vision-and-language navigation task as a trajectory search problem.
- Train a speaker to generate instructions from route images
- Use speaker to construct extra instruction—routs pair for training



Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor BergKirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-andlanguage navigation. NIPS.

Speaker-Follower



For testing, the similarity between instructions generated by speaker and original is used to further evaluate the routes.

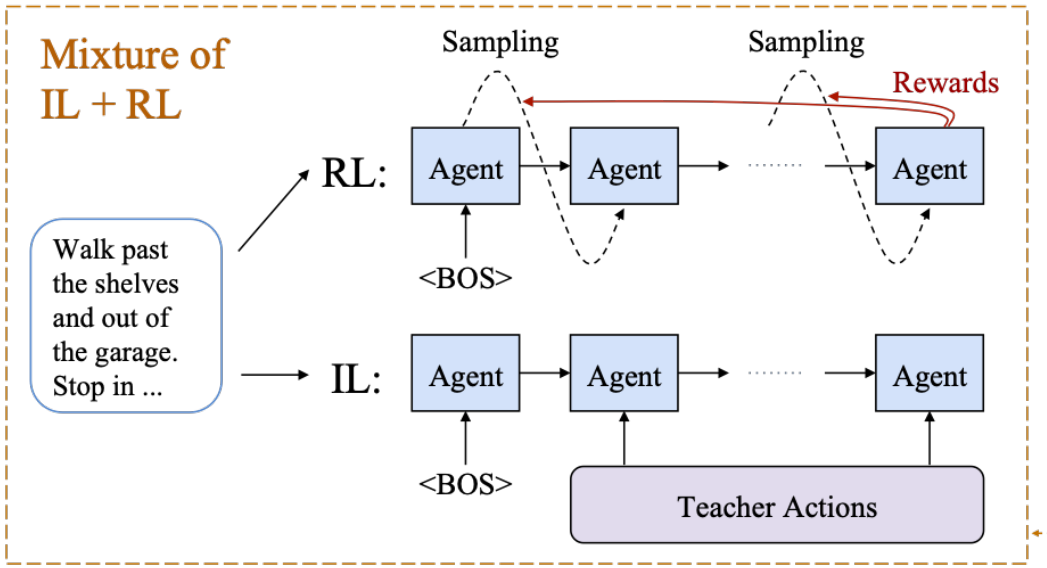
Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor BergKirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-andlanguage navigation. NIPS.

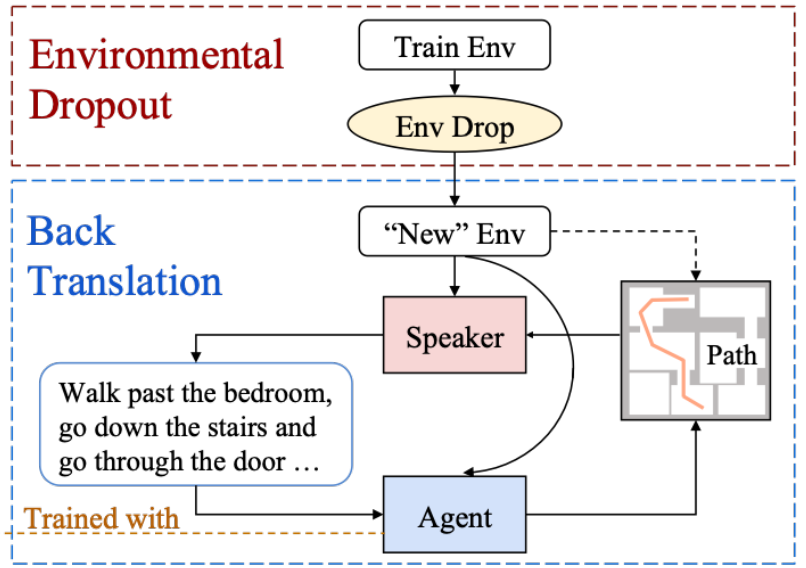
Reinforcement Learning based Model

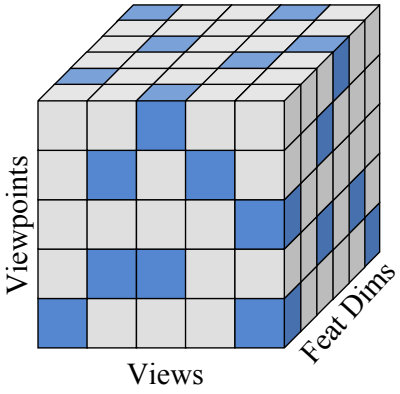
- Environment
  - EnvDrop : Generating more environments.
  - RCM+SIL : Exploration of unseen environments using a off-policy method

- Language
  - AuxRN : Auxiliary Reasoning Tasks are helpful.
  - LEO : Leveraging multiple instructions (as different views) to resolve language ambiguity.

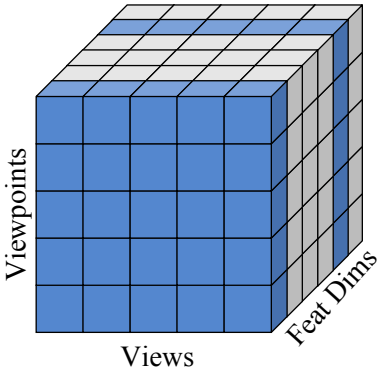
EnvDrop : Generating more environments







(a) Feature dropout



(b) Environmental dropout

Environmental Dropout

$$f'_{t,i} = f_{t,i} \odot \boldsymbol{\xi}^{\boldsymbol{E}}$$

$$\xi_e^E \sim \frac{1}{1-p}\,\mathrm{Ber}(1-p)$$

Hao Tan, Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In NAACL.

RCM+SIL : Exploration of unseen environments

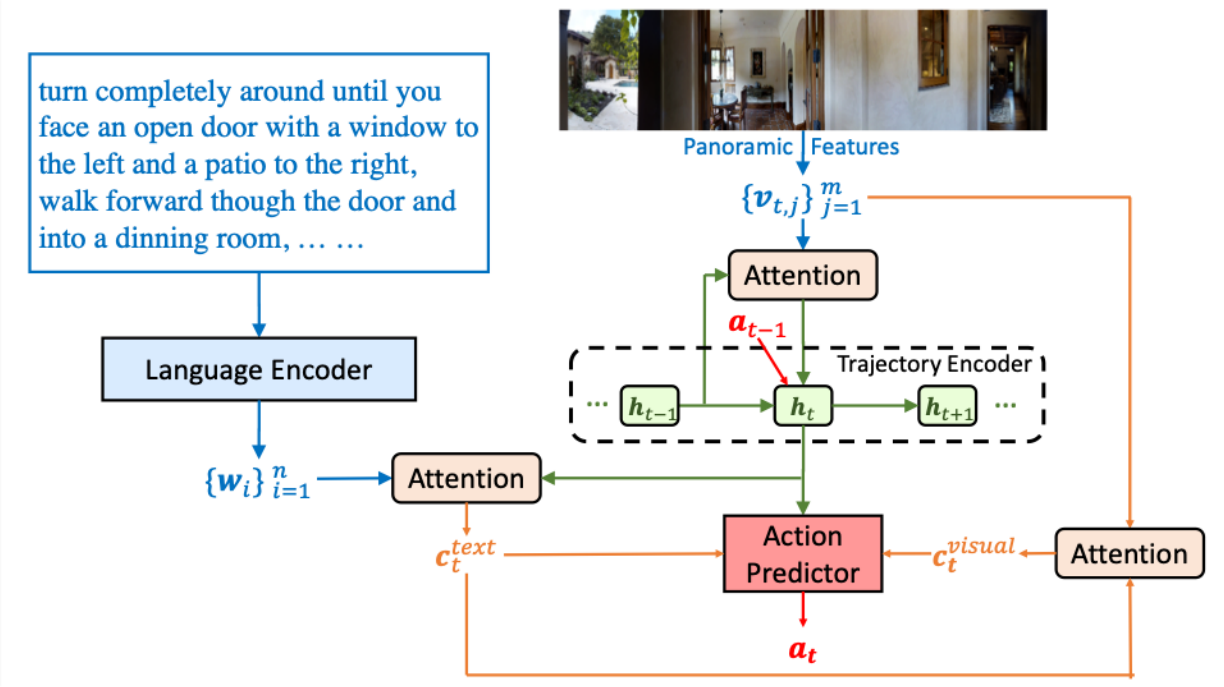
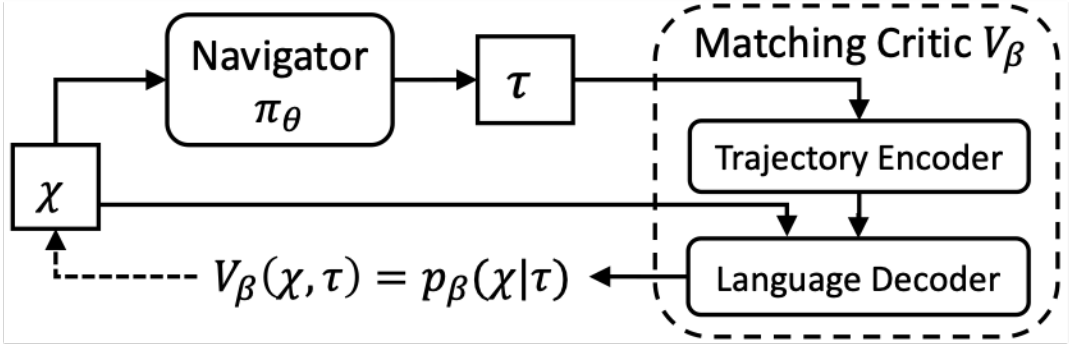


Figure 3: Cross-modal reasoning navigator at step t.



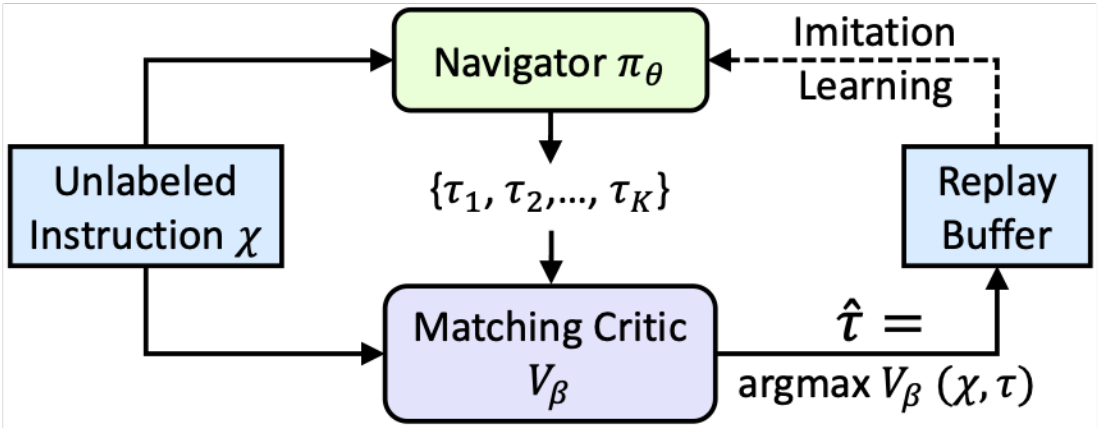


Figure 5: SIL for exploration on unlabeled data.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. CVPR.

AuxRN : Auxiliary Reasoning Tasks

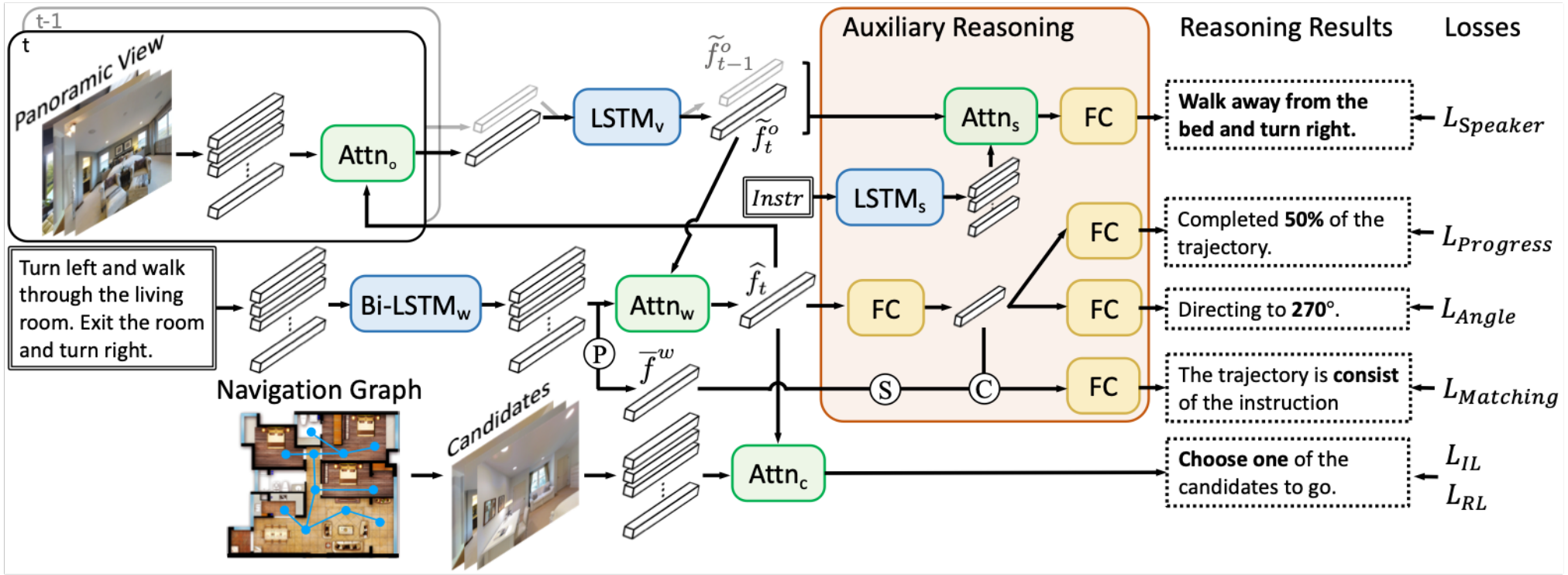


Figure 2. An overview of AuxRN. The agent embeds vision and language features respectively and performs co-attention between them. The embedded features are given to reasoning modules and supervised by auxiliary losses. The feature produced by vision-language attention is fused with the candidate features to predict a action. The "P", "S", and "C" in the white circles stand for the mean pooling, random shuffle and concatenate operations respectively.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2019. Vision-language navigation with self-supervised auxiliary reasoning tasks. arXiv preprint arXiv:1911.078

LEO : Leveraging multiple instructions

- Leveraging multiple instructions (as different views) to resolve language ambiguity.

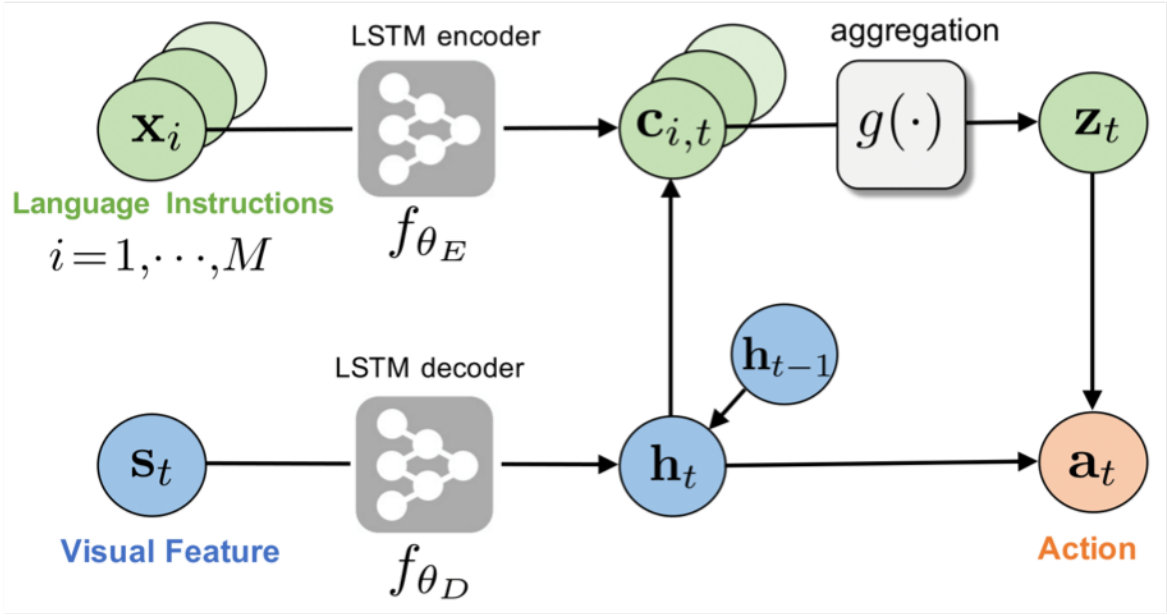
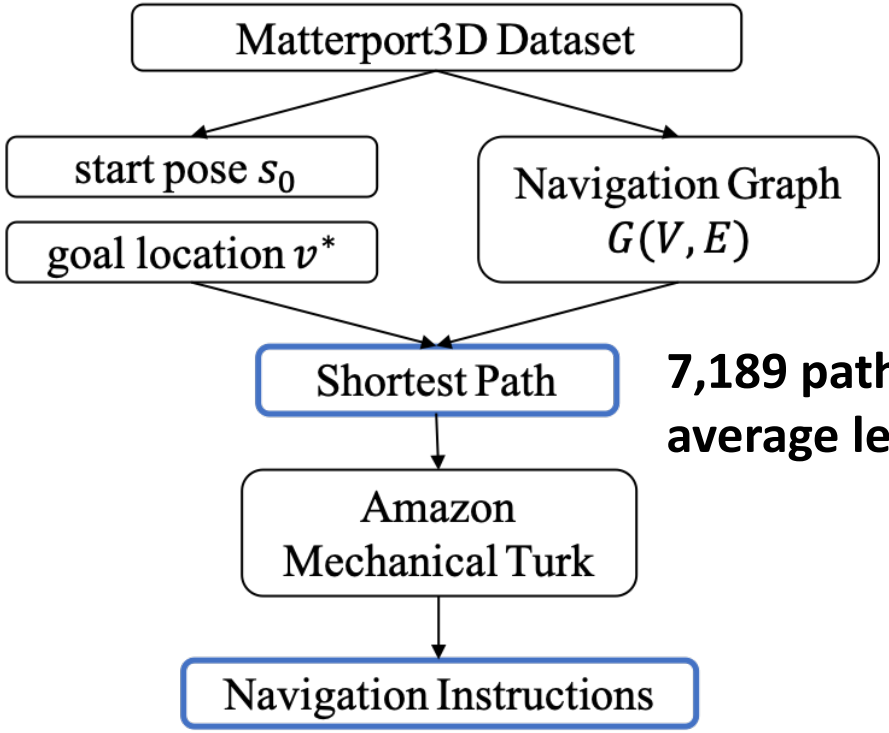


Figure 2: Illustration of learning to navigate with LEO. The action (red circle) is selected, based on both the visual scene states (blue circle) and textual language instructions (green circle). At time step t, LEO aggregates multiple instructions $\{\boldsymbol{x}_i\}_{i=1}^{M}$ together to generate one action $\boldsymbol{a}_t$.

Xia, Q., Li, X., Li, C., Bisk, Y., Sui, Z., Choi, Y., & Smith, N.A. (2020). Multi-View Learning for Vision-and-Language Navigation. ArXiv, abs/2003.00857.

Dataset: Room2Room

Construction Process



7,189 paths with average length 10m

21,567 navigation instructions with an average length of 29 words.

Dataset Splits

	Scenes	Instructions
Train	61	14025
Validation (Seen)		1020
Validation (Unseen)	11	2349
Test	18	4173

Anderson, Peter et al. "Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018): 3674-3683.

Evaluation Metrics

- **Navigation Error (NE)**

  The shortest path distance between the agent's final position v_T and the goal location v_* .

- **Success Rate (SR)**

  Usually consider an episode to be a success if the navigation error is less than 3m.

- **Oracle Success Rate (OSR)**

  An oracle stopping rule, i.e. if the agent stopped at the closest point to the goal on its trajectory.

- **Path Length (PL)**

  The average length of the navigation trajectory.

- **Success rate weighted by Path Length (SPL)**

$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_i}{\max(p_i, \ell_i)}$$

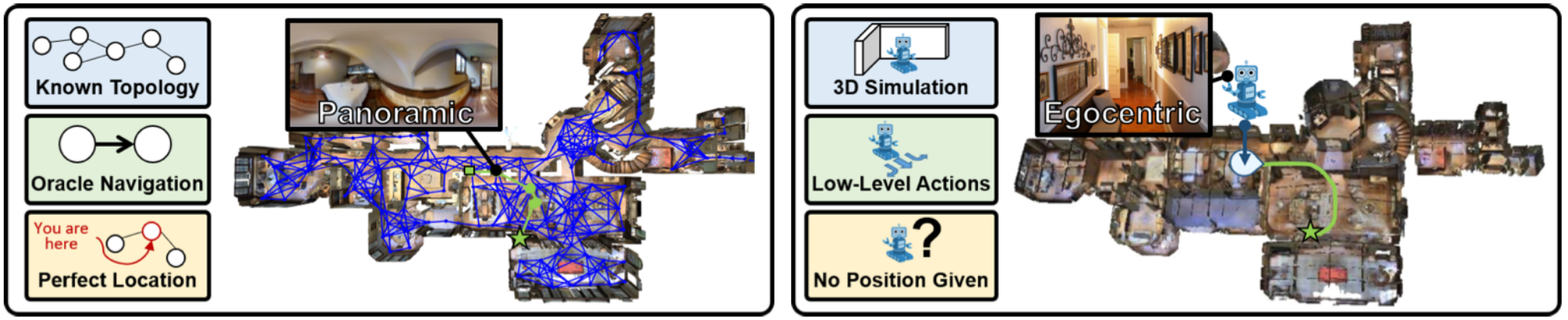
Current Leaderboard

Model	Validation Seen				Validation Unseen				Test Unseen			
	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑	TL ↓	NE ↓	SR ↑	SPL ↑
Approaches that **do not** *explore the* test *environments during training, and utilizes* **one** *instruction during testing*												
RANDOM (Anderson et al., 2018b)	**9.58**	9.45	16	-	9.77	9.23	16	-	9.93	9.77	13	12
S2S-ANDERSON (Anderson et al., 2018b)	11.33	6.01	39	-	**8.39**	7.81	22	-	**8.13**	7.85	20	18
RPA (Wang et al., 2018)	-	5.56	43	-	-	7.65	25	-	9.15	7.53	25	23
SPEAKER-FOLLOWER (Fried et al., 2018)	-	3.36	66	-	-	6.62	35	-	14.82	6.62	35	28
SMNA (Ma et al., 2019a)	-	-	-	-	-	-	-	-	18.04	5.67	48	35
RCM+SIL (Wang et al., 2019)	10.65	3.53	67	-	11.46	6.09	43	-	11.97	6.12	43	38
REGRETFUL (Ma et al., 2019b)	-	3.23	69	63	-	5.32	50	41	13.69	5.69	48	40
FAST (Ke et al., 2019)	-	-	-	-	21.17	4.97	56	43	22.08	5.14	54	41
ENVDROP (Tan et al., 2019)	11.00	3.99	62	59	10.70	5.22	52	48	11.66	5.23	51	47
PRESS (Li et al., 2019)	10.57	4.39	58	55	10.36	5.28	49	45	10.77	5.49	49	45
AUXRN(*) (Zhu et al., 2019)	-	3.33	70	67	-	5.28	54	50	-	5.15	55	51
PREVALENT (Hao et al., 2020)	10.32	3.67	69	65	10.19	4.71	58	53	10.51	5.30	54	51
Approaches that **do not** *explore the* test *environments during training, and utilizes* **three** *instruction during testing*												
PRESS (Li et al., 2019)	10.35	3.09	71	67	10.06	4.31	59	55	10.52	4.53	57	53
PREVALENT (Hao et al., 2020)	10.31	3.31	67	63	9.98	4.12	60	57	10.21	4.52	59	56
LEO+ (Ours)	10.41	**2.30**	**81**	**78**	10.06	**3.35**	**70**	**65**	10.24	**3.76**	**65**	**62**
Approaches that **do** *explore the* test *environments during training, and utilizes* **one** *instruction during testing*												
RCM+SIL (Wang et al., 2019)	10.13	2.78	73	-	9.12	4.17	61	-	9.48	4.22	61	59
ENVDROP (Tan et al., 2019)	9.92	4.84	55	52	9.57	3.78	65	61	9.79	3.97	64	61
AUXRN(*) (Zhu et al., 2019)	-	-	-	-	-	-	-	-	-	3.69	68	65
Human	-	-	-	-	-	-	-	-	11.85	1.61	86	76

Some New Directions

Can we

- Navigate in the continuous space?
- Introduce pre-training methods into VLN?
- Apply memory block into VLN?



(a) Vision-and-Language Navigation (VLN) (b) VLN in Continuous Environments (VLN-CE)

Outline

- Cross Vision and Language Matching
- Vision-based Text Generation
- Cross Vision and Language Reasoning
- Language-based Vision Navigation

- Cross-modality Pretraining

Two-Stream Cross-Modal Pre-training Model

- Two encoders to encode language and Image Respectively

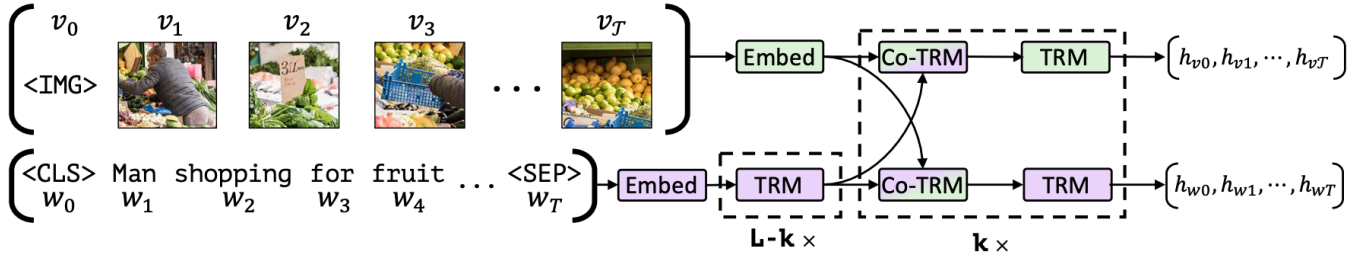


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

ViLBERT, NIPS 2019

- Cross-Modality Encoder to enforce the interaction of two modalities.

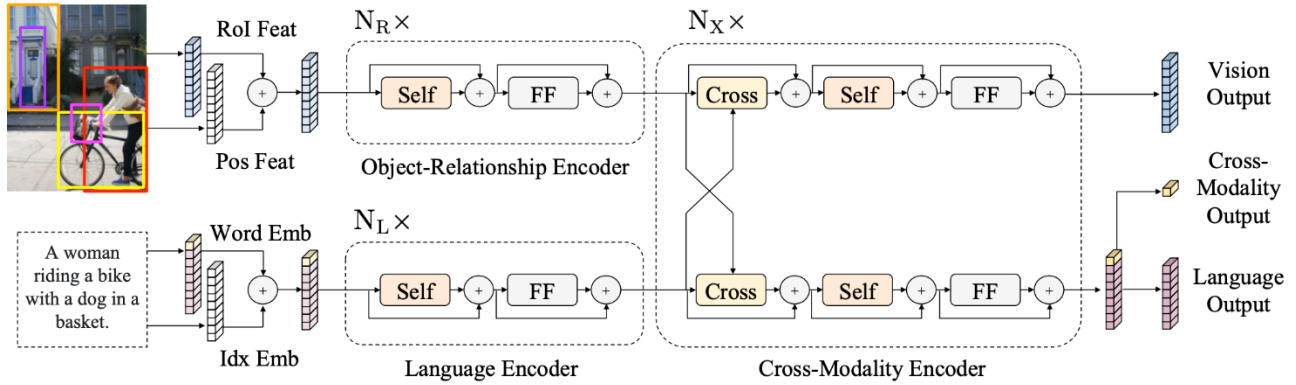


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

LXMBert, EMNLP 2019

Single-stream Cross-Modal Pretraining Model

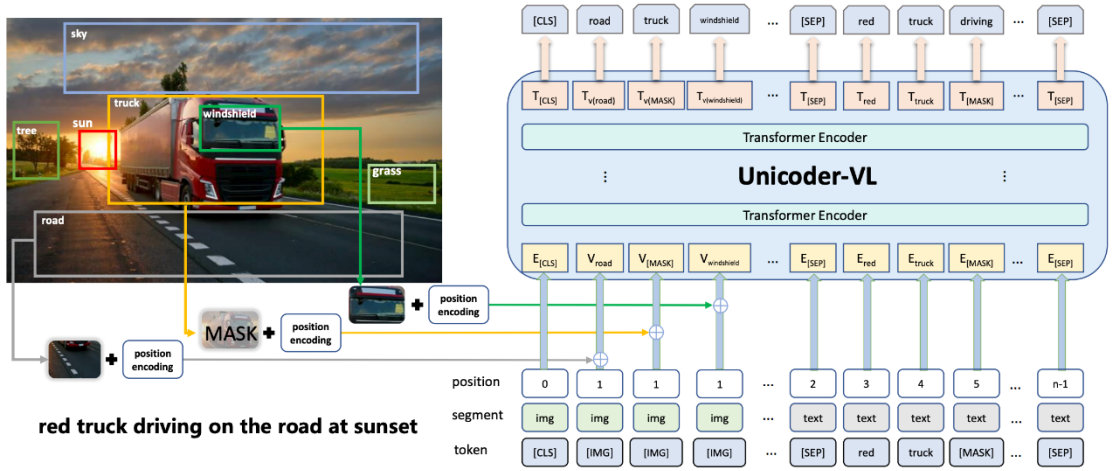


Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple Transformer encoders which are used to learn viusal and linguistic representation jointly.

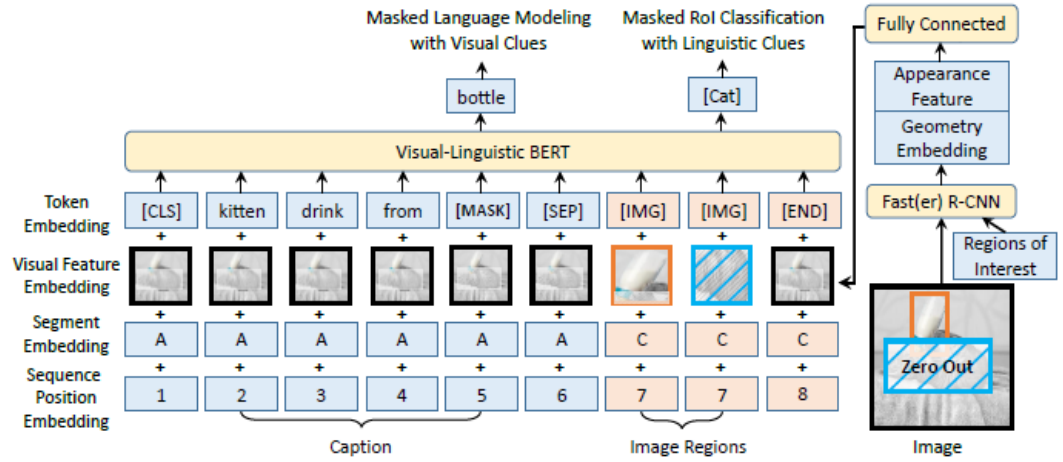


Figure 1: Architecture for pre-training VL-BERT. All the parameters in this architecture including VL-BERT and Fast R-CNN are jointly trained in both pre-training and fine-tuning phases.

Unicoder-VL, AAAI 2020 **VLBert, ICLR 2020**

- Token Embedding: word piece embedding for words; specific tokens for visual elements.
- Visual Feature Embedding: RoI features from Faster R-CNN; CNN feature of whole image for non-visual elements
- Segment Embedding: A and B for the words from the first and second input sentence respectively, and C for the RoIs from the input image.

Cross-modal Pre-training tasks

- Masked Language Modeling (MLM)
  - Randomly mask 15% of both words input and tasking the model to re-construct them given the remaining inputs.
- Mask Region Modeling(MOC)
  - **Mask Region Regression(MOR)** learns to regress the output of each masked region to its visual features.
  - **Masked Region Classification (MRC)** MRC learns to predict the object semantic class for each masked region.
  - **Masked Region Classification with KL-Divergence (MRC-kl)** MRC takes the most likely object class from the object detection model as the hard label.
- Visual-Linguistic Matching (VLM)
  - Given a pair of image and text, model is tasking to predict if they are aligned one.
  - Negative pairs are constructed by replace sentence or image in the original pair.

ERNIE-VIL

- Conduct a two-stream Cross-modal Transformer.
- Introduce a mask task on scene graph for pre-training.
- Construct Scene graph from text. and predict masked object, attribute and relationship.

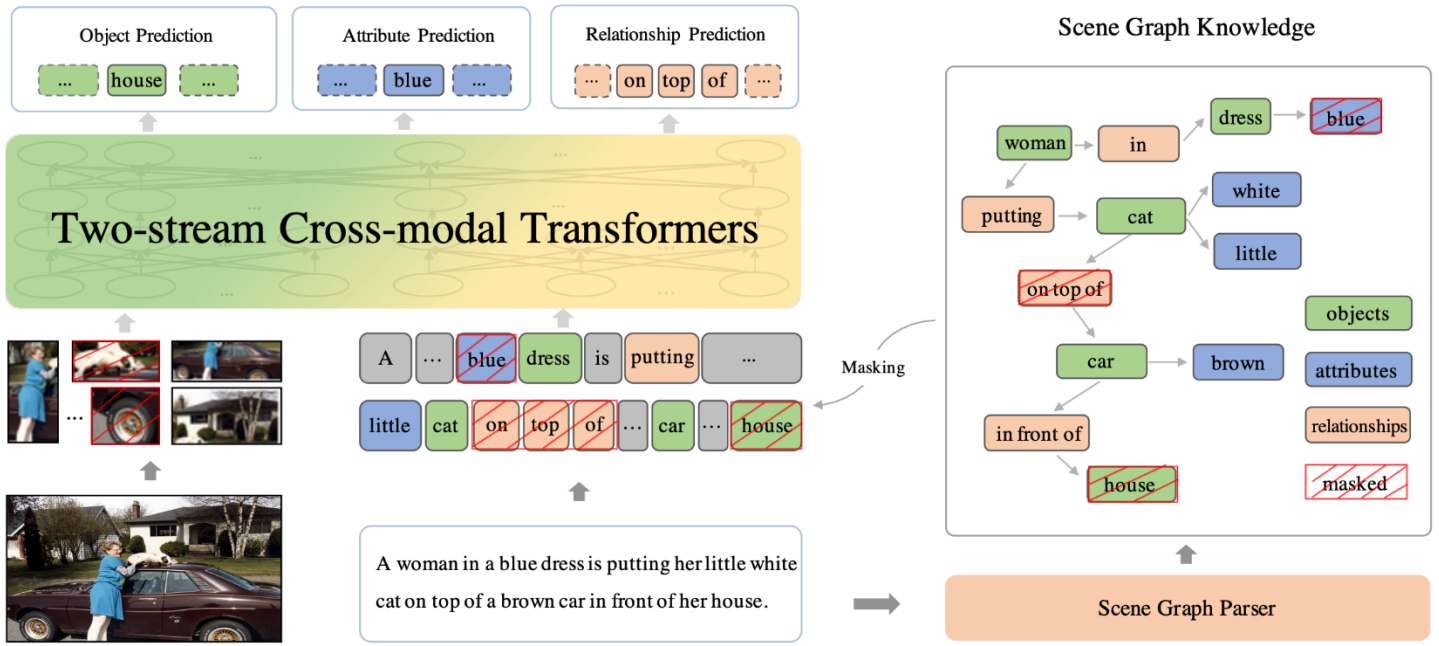


Figure 2: Illustration of Scene Graph Prediction for ERNIE-ViL. Given detected regions for image and token sequence for the text, ERNIE-ViL use a Two-stream Cross-modal Transformer network to model the joint vision-language representation. Based on the scene graph parsed from the text using Scene Graph Parser, we construct Object Prediction, Attribute Prediction and Relationship Prediction to learn the alignments of detailed semantics across modals.

ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph, Arxiv, 2020

Training Strategy

- **Pre-Training**
  - COCO, Visual Genome, Conceptual Captions, and SBU Captions.

- **Fine-Tuning**
  - The model is trained to maximize the performance on the specific downstream task.

	In-domain		Out-of-domain	
Split	COCO Captions	VG Dense Captions	Conceptual Captions	SBU Captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Table 1: Statistics on datasets used for pre-training. Each cell shows #image-text pairs (#images).

Performance of Cross-Modal Pretraining Models

Models		VCR			RefCOCO+		
		Q→A	QA→R	Q→AR	val	testA	testB
Out-of-domain	ViLBERT-base	72.42 (73.3)	74.47 (74.6)	54.04 (54.8)	72.34	78.52	62.61
	Unicoder-VL-base	72.6 (73.4)	74.5 (74.4)	54.4 (54.9)	-	-	-
	VLBERT-base	73.8 (-)	74.4 (-)	55.2 (-)	71.60	77.72	60.99
	UNITER-base	-	-	-	72.78	-	-
	VLBERT-large	75.5 (75.8)	77.9 (78.4)	58.9 (59.7)	72.59	78.57	62.30
	ERNIE-ViL-base	74.37 (77.0)	79.65 (80.3)	61.24 (62.1)	74.02	80.33	**64.74**
	ERNIE-ViL-large	**78.52(79.2)**	**83.37(83.5)**	**65.81(66.3)**	**74.24**	**80.97**	64.70
Out-of-domain + in-domain	UNITER-base	74.56 (75.0)	77.03 (77.2)	57.76 (58.2)	75.31	81.30	65.58
	VILLA-base	75.54 (76.4)	78.78 (79.1)	59.75 (60.6)	76.05	81.65	65.70
	UNITER-large	77.22 (77.3)	80.49 (80.8)	62.59 (62.8)	75.90	81.45	66.70
	VILLA-large	78.45 (78.9)	82.57 (82.8)	65.18 (65.7)	**76.17**	81.54	66.84
	ERNIE-ViL-large	**78.62** (-)	**83.42** (-)	**65.95**(-)	75.95	**82.07**	**66.88**

Models		VQA		IR-Flickr30K			TR-Flickr30K		
		test-dev	test-std	R@1	R@5	R@10	R@1	R@5	R@10
Out-of-domain	ViLBERT-base	70.55	70.92	58.20	84.90	91.52	-	-	-
	Unicoder-VL-base	-	-	71.50	90.90	94.90	86.20	96.30	99.00
	VLBERT-base	71.16	-	-	-	-	-	-	-
	UNITER-base	71.56	-	-	-	-	-	-	-
	VLBERT-large	71.79	72.22	-	-	-	-	-	-
	ERNIE-ViL-base	72.62	72.85	74.44	92.72	95.94	86.70	97.80	99.00
	ERNIE-ViL-large	**73.78**	**73.96**	**75.10**	**93.42**	**96.26**	**88.70**	**97.30**	**99.10**
Out-of-domain + in-domain	UNITER-base	72.70	72.91	72.52	92.36	96.08	85.90	97.10	98.80
	OSCAR-base	73.16	73.61	-	-	-	-	-	-
	VILLA-base	73.59	73.67	74.74	92.86	95.82	86.60	97.90	**99.20**
	12-in-1-base	73.15	-	67.90	-	-	-	-	-
	UNITER-large	73.82	74.02	75.56	94.08	96.76	87.30	**98.00**	**99.20**
	OSCAR-large	73.44	73.82	-	-	-	-	-	-
	VILLA-large	74.69	74.87	76.26	**94.24**	**96.84**	87.90	97.50	98.80
	ERNIE-ViL-large	**74.75**	**74.93**	**76.70**	93.58	96.44	**88.10**	**98.00**	**99.20**

Table 3: Results on downstream V+L tasks for ERNIE-ViL model, compared with previous state-of-the-art pre-trained models. IR: Image Retrieval. TR: Text Retrieval.

ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph, Arxiv, 2020

Some thoughts about Multi-modal Research

- Feature fusion and alignment is fundamental
- AI product should be bring into consideration. e.g., are these applications really useful?
- What is the value of multi-modal research?
  - Provide auxiliary information from different modality.
  - Information from one modality as guidance.
- New tasks are always welcome!

References

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems* (pp. 2121-2129).

Wu, Y., Wang, S., Song, G., & Huang, Q. (2019, October). Learning fragment self-attention embeddings for image-text matching. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2088-2096).

Wu, Y., Wang, S., Song, G., & Huang, Q. (2019, October). Learning fragment self-attention embeddings for image-text matching. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2088-2096).

Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 201-216).

Zhang, Q., Lei, Z., Zhang, Z., & Li, S. Z. (2020). Context-Aware Attention Network for Image-Text Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3536-3545).

Wei, X., Zhang, T., Li, Y., Zhang, Y., & Wu, F. (2020). Multi-Modality Cross Attention Network for Image and Sentence Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10941-10950).

Li, Y., Zhang, D., & Mu, Y. (2020). Visual-Semantic Matching by Exploring High-Order Attention and Distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12786-12795).

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11-20).

Yang, S., Li, G., & Yu, Y. (2020). Graph-Structured Referring Expression Reasoning in The Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9952-9961).

References

Goyal, Yash, et al. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In ICCV, 2017b.

Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.

Hudson, Drew A., and Christopher D. Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering." In CVPR. 2019.

Antol S, Agrawal A, Lu J, et al. VQA: Visual question answering[C] //Proceedings of the IEEE International Conference on Computer Vision. 2015: 2425-2433.

Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

Hudson, Drew and Manning, Christopher D. Learning by Abstraction: The Neural State Machine. In Advances in Neural Information Processing Systems(NIPS), 2019

Zellers, Rowan, et al. "From recognition to cognition: Visual commonsense reasoning." ICCV. 2019.

Park J S , Bhagavatula C , Mottaghi R , et al. VisualCOMET: Reasoning about the Dynamic Context of a Still Image[J]. 2020.

References

Chang, A.X., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., & Zhang, Y. (2017). Matterport3D: Learning from RGB-D Data in Indoor Environments. 2017 International Conference on 3D Vision (3DV), 667-676.

Anderson, P., Chang, A.X., Chaplot, D., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., & Zamir, A. (2018). On Evaluation of Embodied Navigation Agents. ArXiv, abs/1807.06757.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., & Hengel, A.V. (2018). Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3674-3683.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L., Berg-Kirkpatrick, T., Saenko, K., Klein, D., & Darrell, T. (2018). Speaker-Follower Models for Vision-and-Language Navigation. NeurIPS.

Wang, X.E., Xiong, W., Wang, H., & Wang, W.Y. (2018). Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation. ArXiv, abs/1803.07729.

Wang, X.E., Huang, Q., Çelikyilmaz, A., Gao, J., Shen, D., Wang, Y., Wang, W.Y., & Zhang, L. (2019). Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6622-6631.

Tan, H., Yu, L., & Bansal, M. (2019). Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. ArXiv, abs/1904.04195.

Zhu, F., Zhu, Y., Chang, X., & Liang, X. (2020). Vision-Language Navigation With Self-Supervised Auxiliary Reasoning Tasks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10009-10019.

References

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, Tamara L. Berg: Baby talk: Understanding and generating simple image descriptions. CVPR 2011: 1601-1608

Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator. CVPR. 2015: 3156-3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015: 2048-2057

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo: Image Captioning with Semantic Attention. CVPR 2016: 4651-4659

Ting Yao, Yingwei Pan, Yehao Li, Tao Mei: Exploring Visual Relationship for Image Captioning. ECCV (14) 2018: 711-727

Bo Dai, Sanja Fidler, Dahua Lin: A Neural Compositional Paradigm for Image Captioning. NeurIPS 2018: 656-666

Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh: Neural Baby Talk. CVPR 2018: 7219-7228

Justin Johnson, Andrej Karpathy, Li Fei-Fei: DenseCap: Fully Convolutional Localization Networks for Dense Captioning. CVPR 2016: 4565-4574

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, Yueting Zhuang: Diverse Image Captioning via GroupTalk. IJCAI 2016: 2957-2964

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, Xuanjing Huang: A Question Type Driven Framework to Diversify Visual Question Generation. IJCAI 2018: 4048-4054

Xin Wang, Wenhu Chen, Yuan-Fang Wang, William Yang Wang: No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. ACL (1) 2018: 899-909

References

Ruize Wang, Zhongyu Wei, Piji Li and Xuanjing Huang, Topic-Aware Visual Storytelling via Multi-Agent Communication, COLING 2020.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, Xuanjing Huang: Storytelling from an Image Stream Using Scene Graphs. AAAI 2020: 9185-9192

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu: Bleu: a Method for Automatic Evaluation of Machine Translation. ACL 2002: 311-318

Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. IEEvaluation@ACL 2005: 65-72

Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh: CIDEr: Consensus-based image description evaluation. CVPR 2015: 4566-4575

Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould: SPICE: Semantic Propositional Image Caption Evaluation. ECCV (5) 2016: 382-398

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, Serge J. Belongie: Learning to Evaluate Image Captioning. CVPR 2018: 5804-5812

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, Jianfeng Gao: TIGEr: Text-to-Image Grounding for Image Caption Evaluation. EMNLP/IJCNLP (1) 2019: 2141-2152

Xia, Q., Li, X., Li, C., Bisk, Y., Sui, Z., Choi, Y., & Smith, N.A. (2020). Multi-View Learning for Vision-and-Language Navigation. ArXiv, abs/2003.00857.

References

Krantz, J., Wijmans, E., Majumdar, A., Batra, D., & Lee, S. (2020). Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. ArXiv, abs/2004.02857.

Hao, W., Li, C., Li, X., Carin, L., & Gao, J. (2020). Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-Training. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13134-13143.