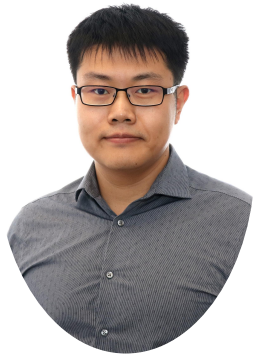
ECCV 2018
European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany
SiVL Workshop

Stacked Co-Attention for VisDial1.0

Tianhao Yang
USTC

Zheng-Jun Zha
USTC



Hanwang Zhang
NTU
Alibaba-NTU Joint Research Institute

VQA□Visual Dialog

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Co-reference

Longer free-form answers

VisDial Dataset v1.0

Training set :

- 1,23,287 images from COCO
- 1 dialog / image
- 10 rounds of question-answers / dialog

Validation set :

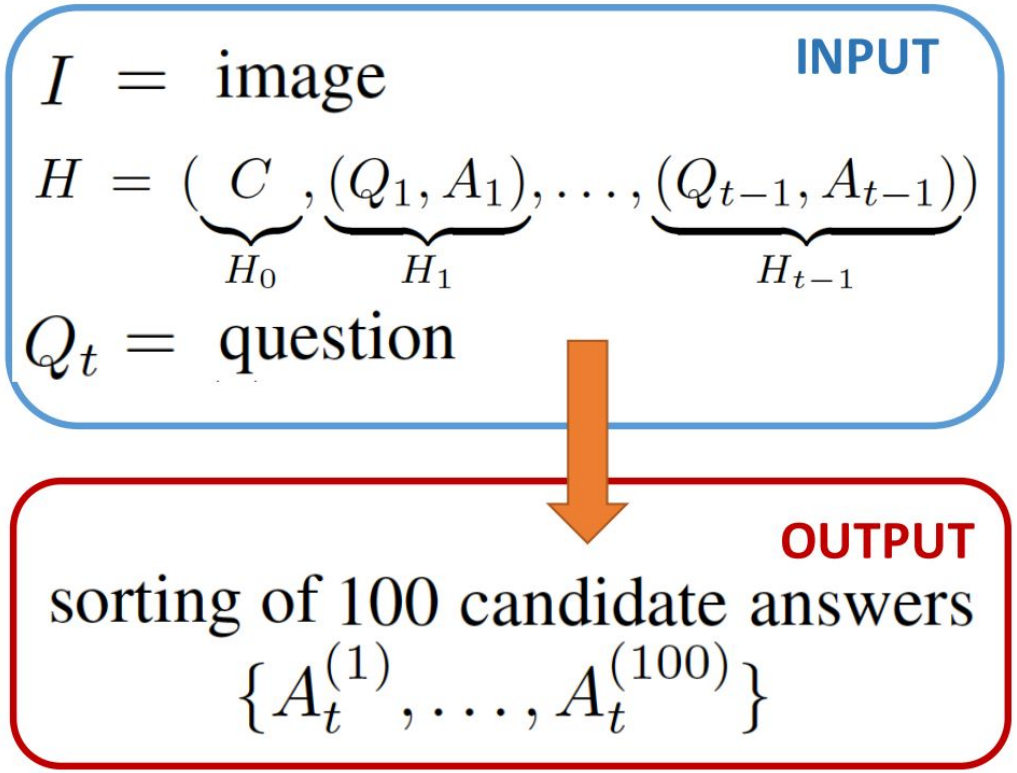
- 2,000 images from Flickr
- 1 dialog / image
- 10 rounds of question-answers / dialog

Test set :

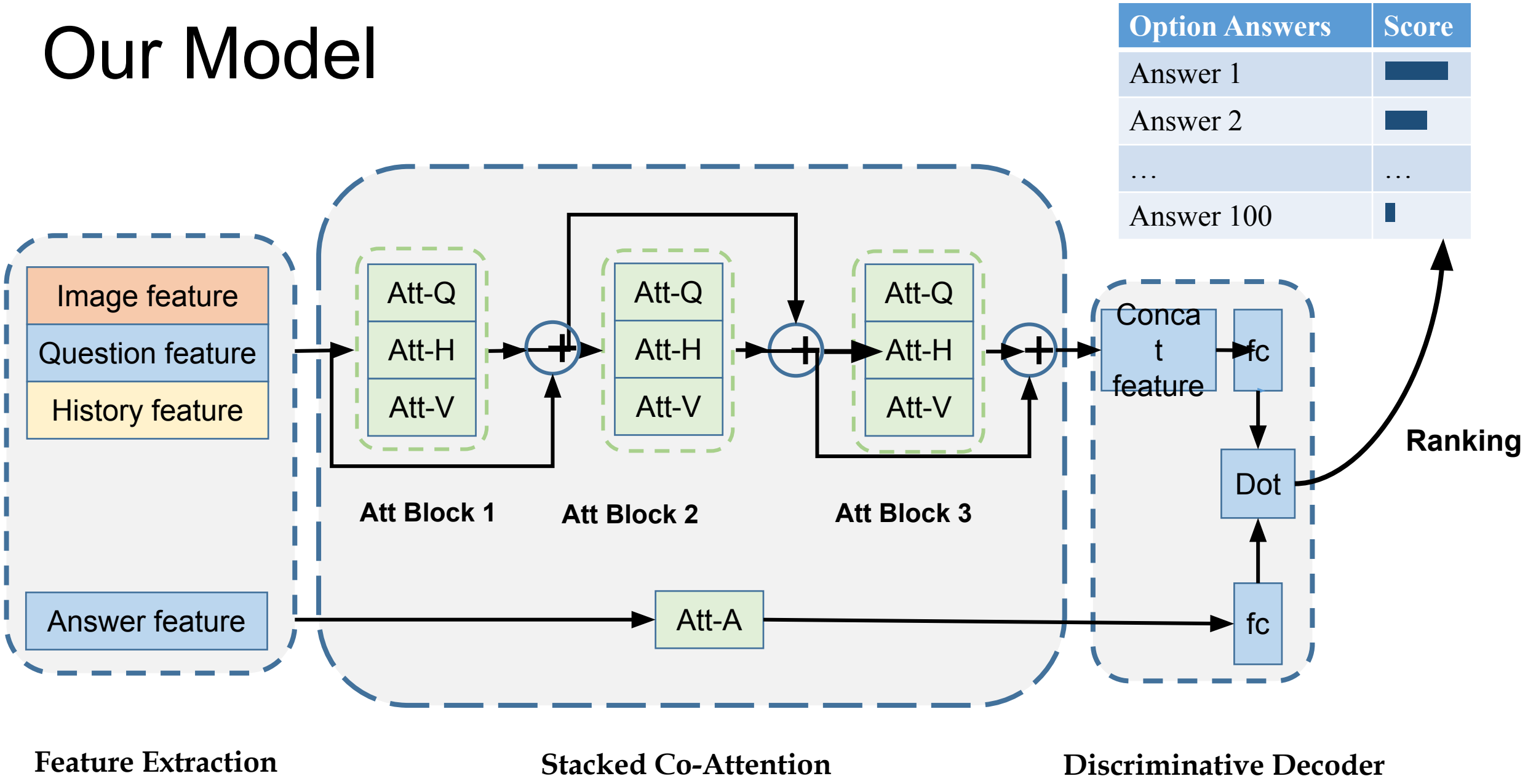
- 8,000 images from Flickr
- given ‘n’ rounds / dialog (‘n’ anywhere in 1 to 10)
- 1 follow-up question + 100 candidate answers

Evaluation Protocol

- Evaluate individual responses independently at each round (t = 1, 2, …, 10) in a retrieval or multiple-choice setup
- The model is evaluated on retrieval metrics:
  - rank of human response
  - recall@k, i.e. existence of the human response in top-k ranked responses
  - mean reciprocal rank (MRR) of the human response
  - **NDCG (new)**
- Candidate Answers: 1 ground-truth, answers to 50 similar questions, 30 most popular answers, 19 random answers. (1+50+30+19=100)

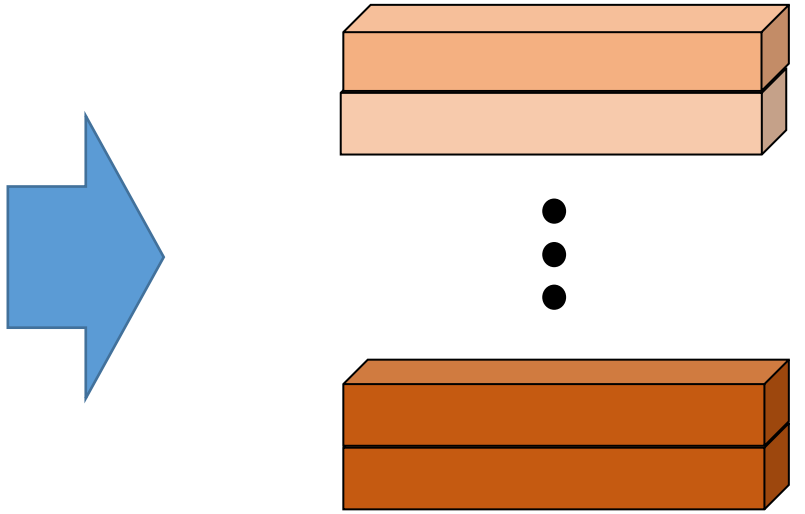


Our Model


Option Answers
Score
Answer 1
Answer 2
…
…
Answer 100
Image feature
Question feature
History feature
Answer feature
Att-Q
Att-H
Att-V
Att Block 1
Att Block 2
Att Block 3
Att-A
Concat feature
fc
Dot
fc
Ranking
Feature Extraction
Stacked Co-Attention
Discriminative Decoder

Feature Extraction

- Image feature ：Bottom-Up Attention Model [Anderson et al. 2018]

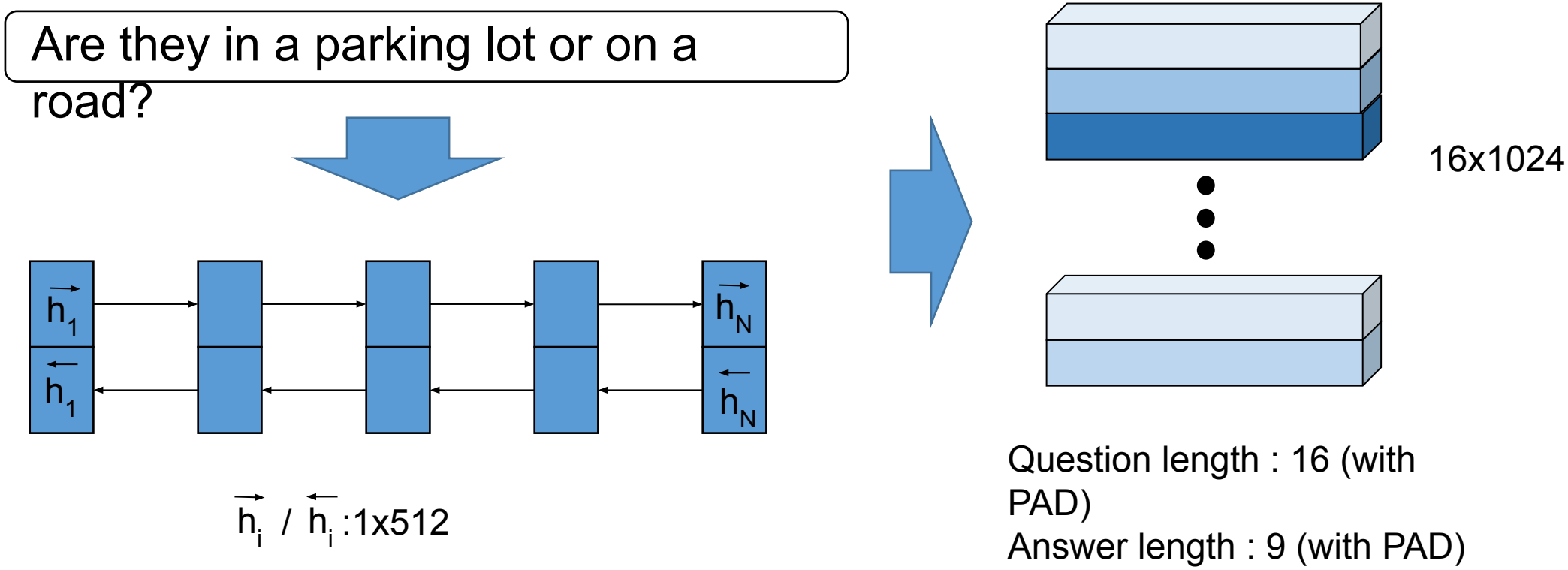


36x2048

Anderson et al. 2018 : P. Anderson , X. He, C. Buehler, D. Teney , M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image caption and vqa. In *CVPR 2018*.

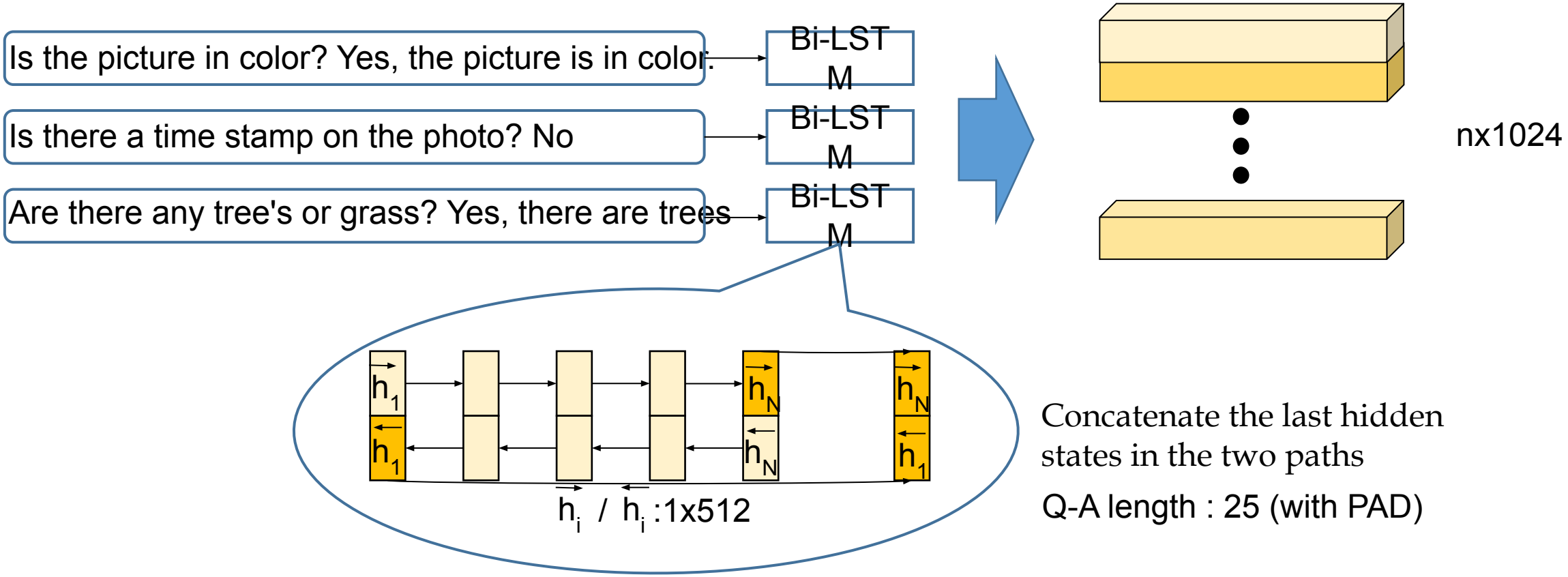
Feature Extraction

- Question feature/Answer feature :

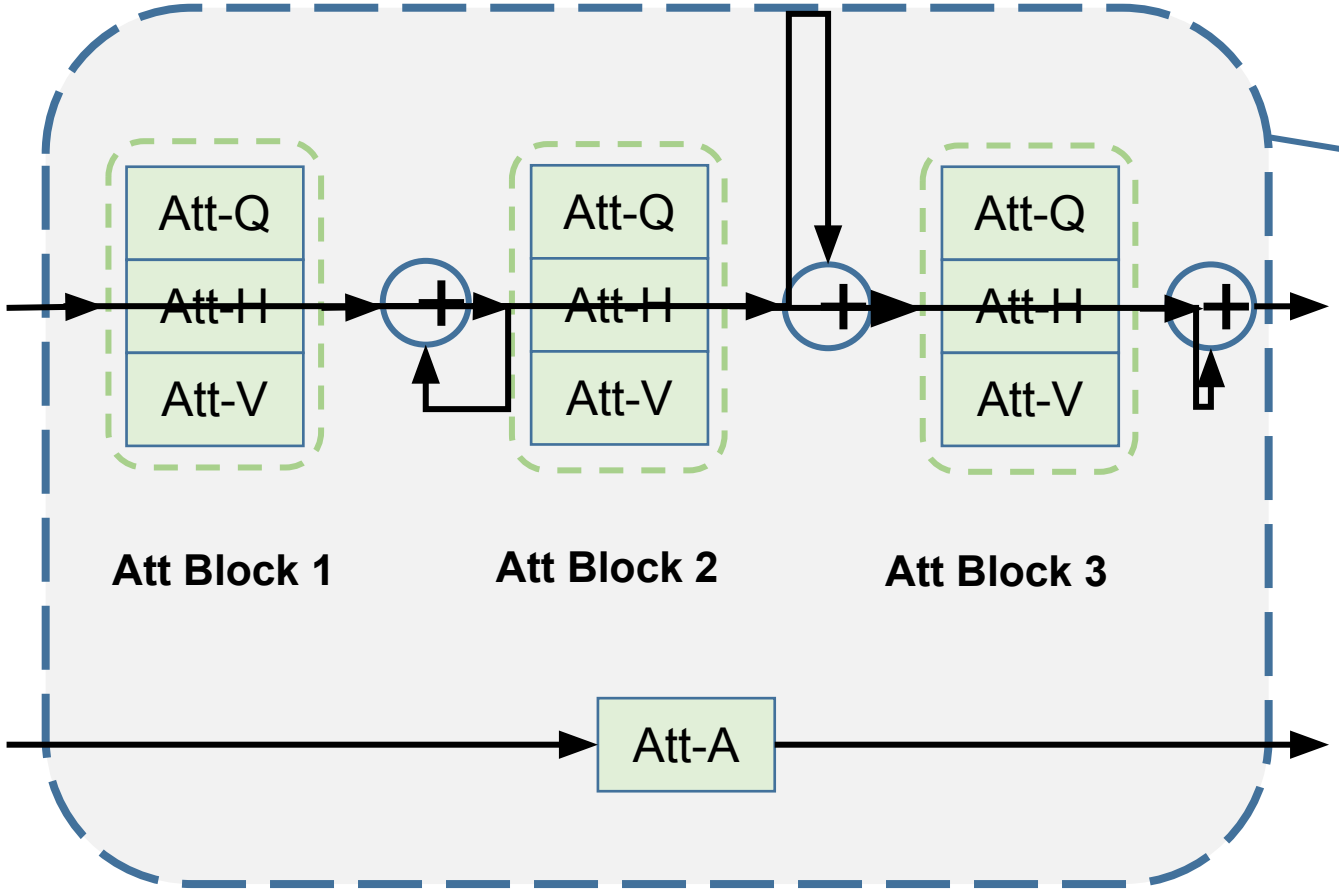


Feature Extraction

- History feature:



Stacked Co-Attention

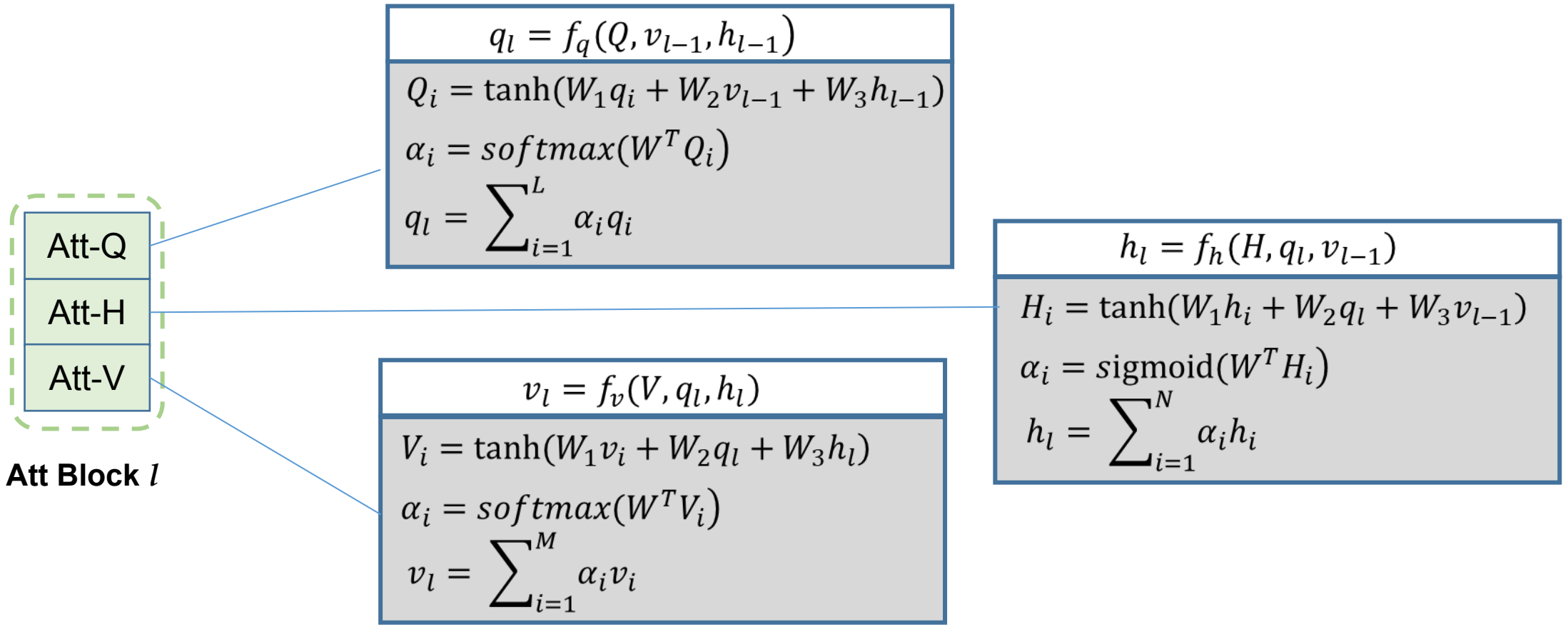


Why stacked? Multiple reasoning steps.

- **Q:"what color is A in B?"**
- **H: co-reference A & B**
- **V: re-attention A & B**

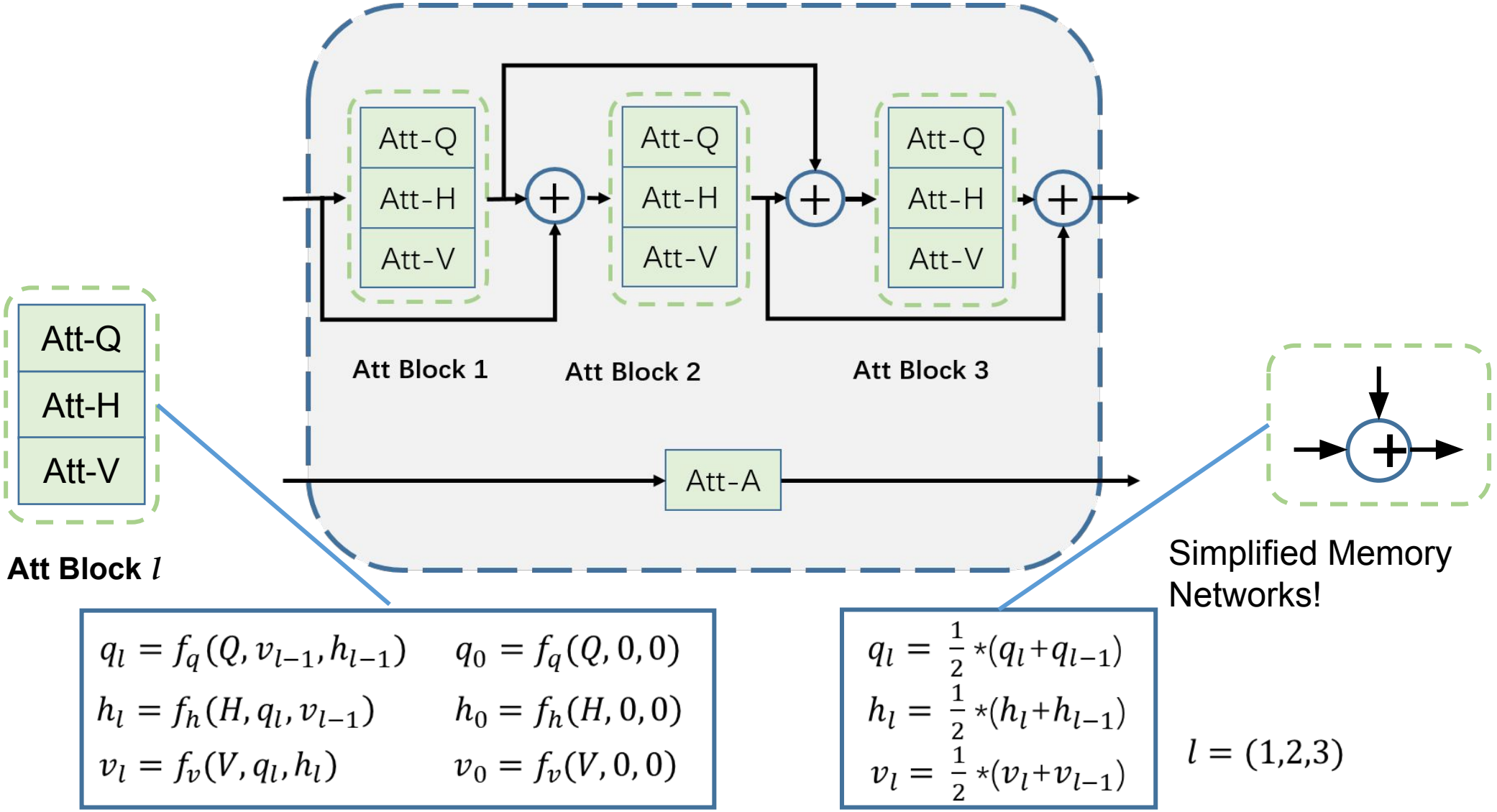
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola. Stacked Attention Networks for Image Question Answering. In *CVPR 2016*.

Stacked Co-Attention



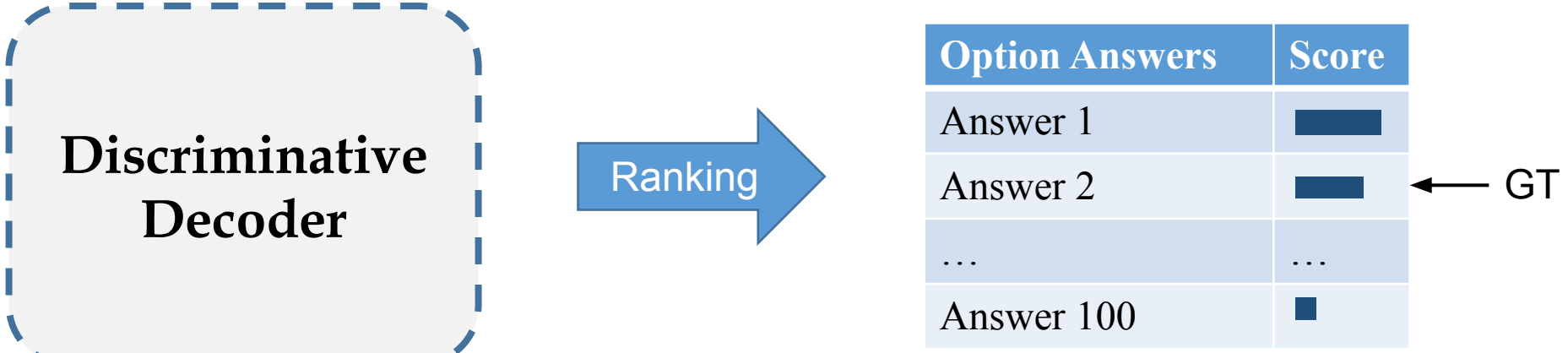
Q. Wu, P. Wang, C. Shen, I. Reid, and A. Hengel. Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In *CVPR 2018*.

Yan Zhang, Jonathon Hare, Adam Prügel-Bennett. Learning to Count Objects in Natural Images for Visual Question Answering. In ICLR 2018



Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, Leonid Sigal. Visual Reference Resolution using Attention Memory for Visual Dialog. In *NIPS 2017*.

Training Loss



$$\text{Loss} = \log(1 + \sum_{i=1}^{N} \exp(score_i^- - score^{gt}))$$

J. Lu, A Kannan, J. Yang, D. Parikh, and D. Batra. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. In *NIPS 2017*

Training Details

- **GPU : 1** NVIDIA GTX 1080Ti
- **Time :** ~2300 seconds/epoch, 25epoch
- **Optimizer :** Adam
- **Learning Rate :** $5 * 10^{-4}$ $(1\sim20\ \text{epoch})$, $2.5 * 10^{-4}$ (20~ epoch)
- **Dropout (keep): 0.5** for each full-connected layer in discriminative decoder;
- **0.5** for word embedding and in Bi-LSTM
- **0.7** for softmax/sigmoid in each Att Block
- Use **Weight Normalization** after embedding the input features in Att Blocks
- **Without** pretrained weight (word2vec, glove) for word embedding
- Using **Faster RCNN (ResNet 101)** pretrained on **COCO** for extracting image features

Results on VisDial0.9val

	MRR	R@1	R@5	R@10	Mean
Lu et al. [1]	62.22	48.48	78.75	87.59	4.81
Wu et al. [2]	63.98	50.29	80.71	88.81	4.47
USTC-YTH	64.03	50.27	80.85	88.96	4.43

[1] J. Lu, A Kannan, J. Yang, D. Parikh, and D. Batra. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. In *NIPS 2017*
[2] Q. Wu, P. Wang, C. Shen, I. Reid, and A. Hengel. Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In *CVPR 2018*.

Results on VisDial1.0val (our implementation)

	MRR	R@1	R@5	R@10	Mean
Lu et al. [1]	61.94	48.45	78.70	88.27	4.63
Wu et al. [2]	62.77	49.38	78.99	88.49	4.56
USTC-YTH	63.05	49.62	79.60	88.83	4.47

[1] J. Lu, A Kannan, J. Yang, D. Parikh, and D. Batra. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. In *NIPS 2017*

[2] Q. Wu, P. Wang, C. Shen, I. Reid, and A. Hengel. Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning. In *CVPR 2018*.

Ablative Results on VisDial1.0val

	MRR	R@1	R@5	R@10	Mean
w/o sigmoid	62.79	49.38	79.12	88.56	4.53
Block 1	62.22	48.66	78.93	88.12	4.61
Block 1,2	62.88	49.42	79.31	88.66	4.49
Block 1,2,3	63.05	49.62	79.60	88.83	4.47

Results on VisDial1.0test-std

Rank	Participant Team	NDCG (x 100)	MRR (x 100)	R@1	R@5	R@10	Mean
1	**DL-61**	57.88	63.42	49.30	80.77	90.68	3.97
2	**USTC-YTH**	56.47	61.44	47.65	78.13	87.88	4.65
3	**MS Conversational AI**	55.35	63.27	49.53	80.40	89.60	4.15
4	**Technion**	54.46	67.25	53.40	85.28	92.70	3.55

Q :

Are	they	in	a	parking	lot	or	on	a	road
0.174	**0.165**	**0.161**	**0.168**	**0.205**	0.069	0.025	0.012	0.002	0.001

Q : Are **they** in a parking lot or on a road?

(Round 10)

• Caption : Trailer tractor truck in parking lot with natural greenery	**0.352**
• Round 1: Does it look like the truck is picking up or delivering? It's just sitting in a parking lot	**0.400**
• Round 2: What color is the truck? The truck is orange, black, and white	0.309
• Round 3: Is the driver of the truck nearby? I can't see anyone in the picture	**0.610**
• Round 4: Is the picture in color? Yes, the picture is in color	0.221
• Round 5: Is there a time stamp on the photo? No	0.151
• Round 6: Can you tell about what time of day it is? It is maybe early morning	0.335
• Round 7: Are they parked near a store? No	0.126
• Round 8: Are there any tree's or grass? Yes, there are trees	0.202
• Round 9: Are there any animals around? No	0.070

- **GT : In a parking lot Prediction : In a parking lot**

Q : Are they in a parking lot or on a road?

GT :

In a parking lot

Prediction :

In a parking lot

Q :

What	color	are	the	bowls
0.127	0.124	0.117	0.109	0.211

Q : What color are the bowls?

(Round 3)

- Caption : Plates and bowls of food are sitting on top of a table
- Round 1: What kind of food? Looks like fried chicken and some rice and sauce
- Round 2: What color are the plates? They are orange

0.041
0.149
0.372

- **GT : Also orange Prediction: Also orange**

Q : What color are the bowls?

GT :

Also orange

Prediction:

Also orange

Q :

Are	there	any	people
0.137	0.133	0.131	0.132

Q : Are there any people ?

(Round 5)

- Caption : Plates and bowls of food are sitting on top of a table
- Round 1: What kind of food? Looks like fried chicken and some rice and sauce
- Round 2: What color are the plates? They are orange
- Round 3: What color are the bowls? Also orange
- Round 4: Are there any drinks? No drinks are visible

0.040
0.061
0.149
0.102
0.421

- **GT : No people are visible Prediction: No people in the picture**

Q : Are there any people ?

GT :

No people are visible

Prediction:

No people in the picture

Q :

Can	you	see	the	type	of	pizza
0.078	0.079	0.090	0.110	0.124	0.101	0.067

Q : Can you see the type of pizza ?

(Round 2)

- Caption : 6 people are sitting around a table with 2 pizzas and drinks
- Round 1: Is this in a public setting? Yes

0.204
0.113

- **GT : Appears to be cheese and a deep dish Prediction: It looks like pizza**

Q : Can you see the type of pizza ?

GT :

Appears to be cheese and a deep dish

Prediction:

It looks like pizza