Paper presentation:

“*Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering*”

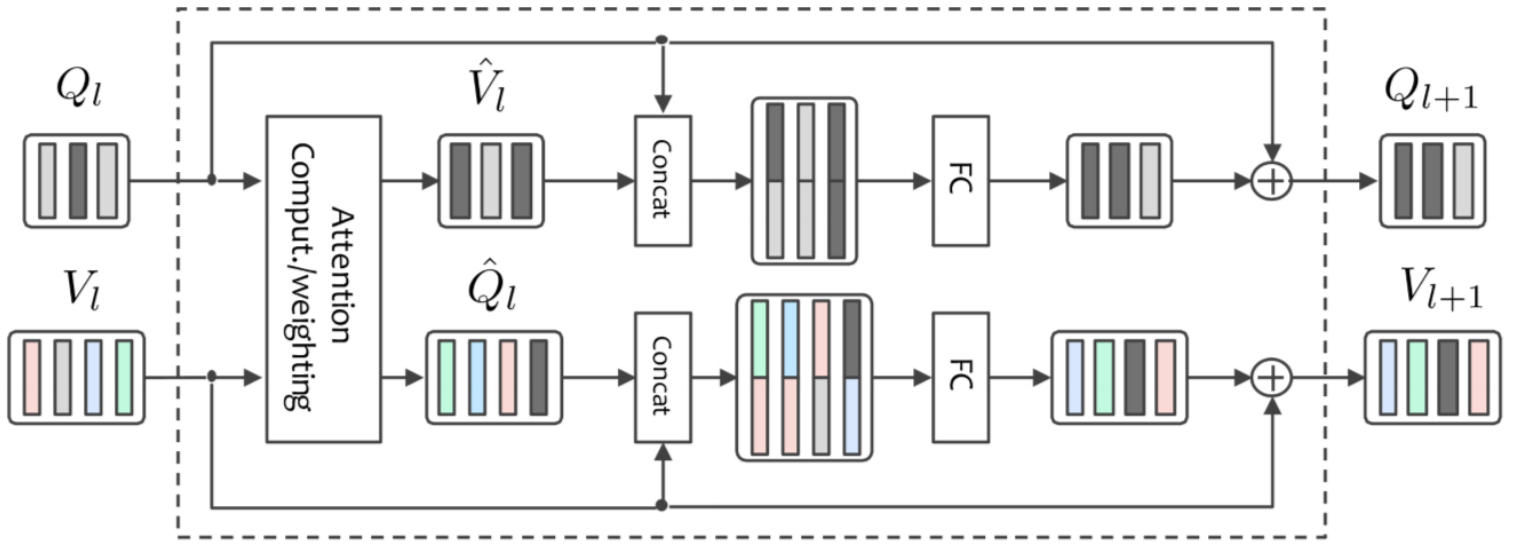
February 8, 2023,
Wednesday

Xiaona Zhou

COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH

What problem is the paper addressing

It proposed an attention mechanism, Dense Co-attention Network (DCN), that improves accuracy on VAQ task.



VIRGINIA TECH

Motivation: VQA is an AI-complete task

Table 1 Computer vision sub-tasks required to be solved by VQA

CV task	Representative VQA question
Object recognition	What is in the image?
Object detection	Are there any dogs in the picture?
Attribute classification	What color is the umbrella?
Scene classification	Is it raining?
Counting	How many people are there in the image?
Activity recognition	Is the child crying?
Spatial relationships among objects	What is between cat and sofa?
Commonsense reasoning	Does this person have 20/20 vision?
Knowledge-base reasoning	Is this a vegetarian pizza?

VIRGINIA TECH

Source: Manmadhan, Sruthy, and Binsu C. Kovoor. "Visual question answering: a state-of-the-art review." Artificial Intelligence Review 53 (2020): 5705-5745.

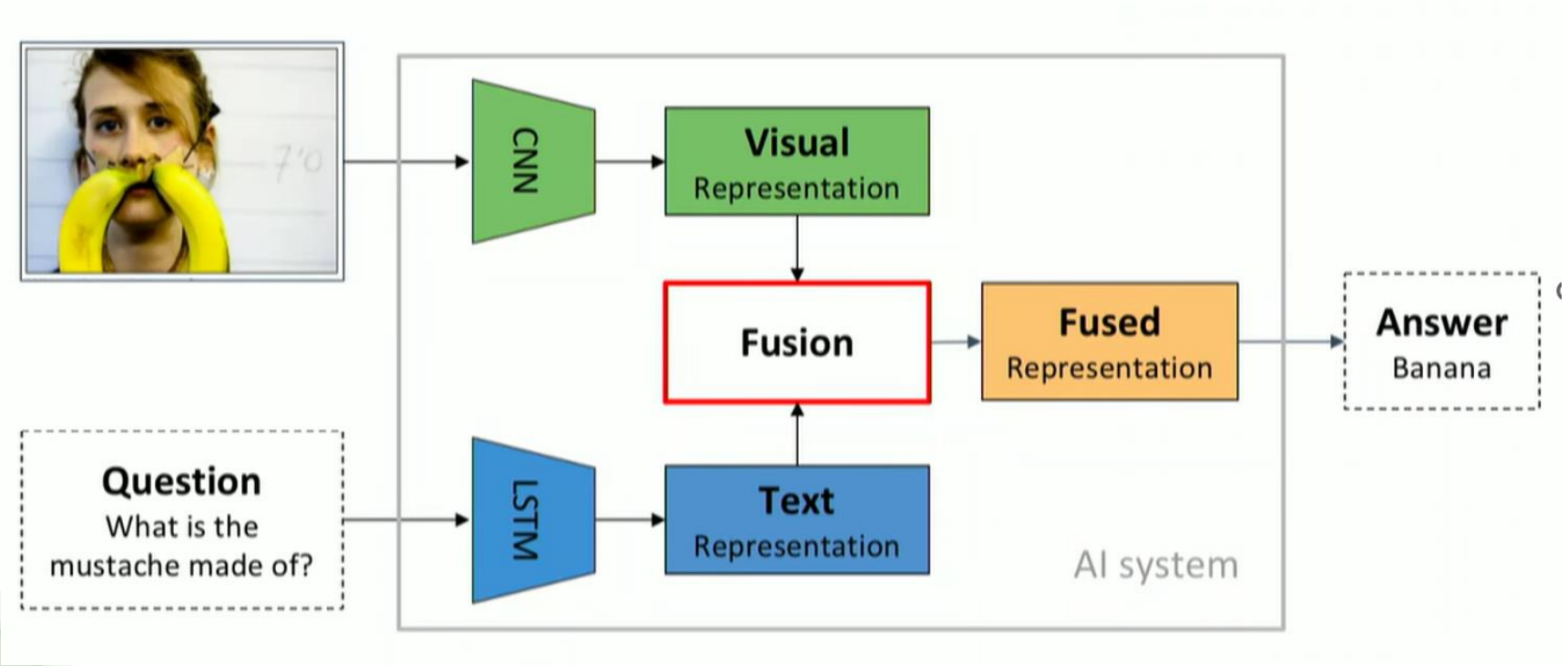
Motivation: many potential applications

- Help visually impaired
- Attract online shoppers
- Increase the popularity of online educational
- Summarize surveillance data

VIRGINIA TECH

Source: Manmadhan, Sruthy, and Binsu C. Kovoor. "Visual question answering: a state-of-the-art review." Artificial Intelligence Review 53 (2020): 5705-5745.

Multi-Modal Network for VQA



VIRGINIA TECH

Prior related work

Two lines of research on VQA (developed independently)

- Attention mechanism
- Fusion of extracted features

VIRGINIA TECH

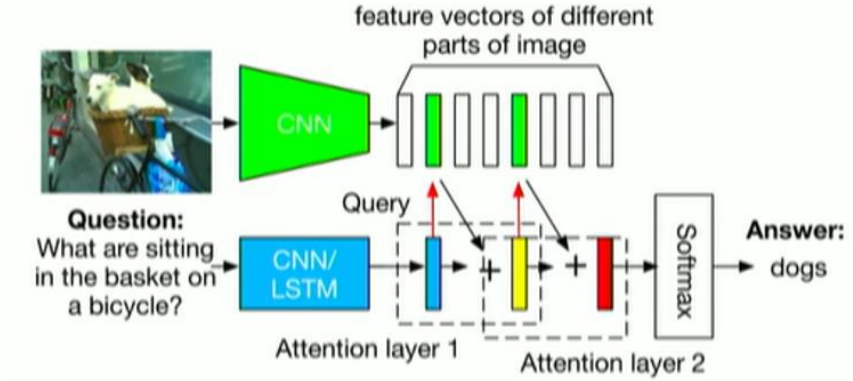
Prior related work: attention mechanism

Two classes:

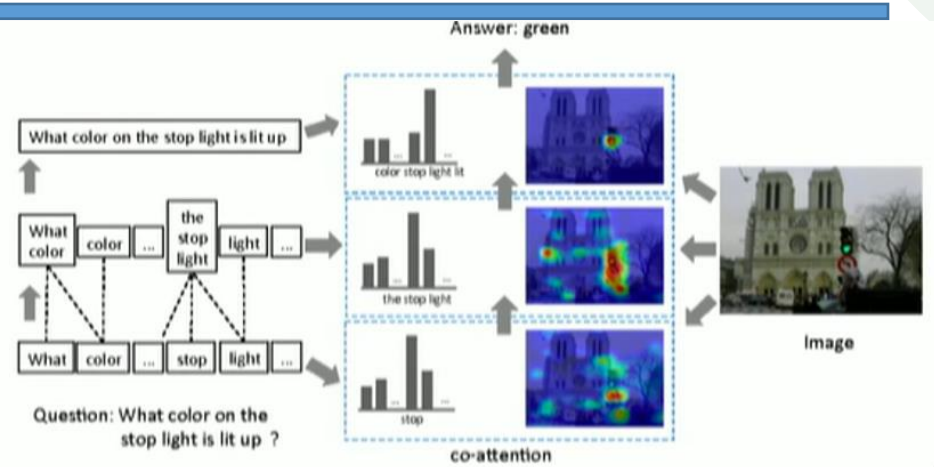
- Use visual features from some region proposals
- Use convolutional features

VIRGINIA TECH

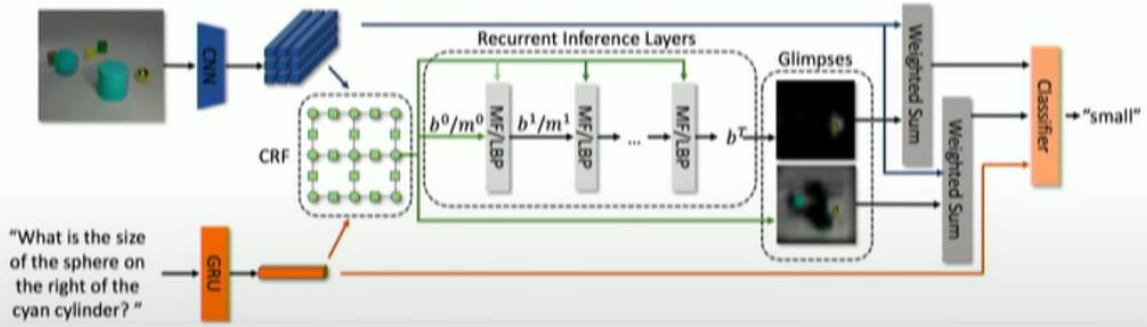
Attention models



Stacked Attention Network [Yang et al. 2016]



Hierarchical Co-Attention Network [Lu et al. 2016]



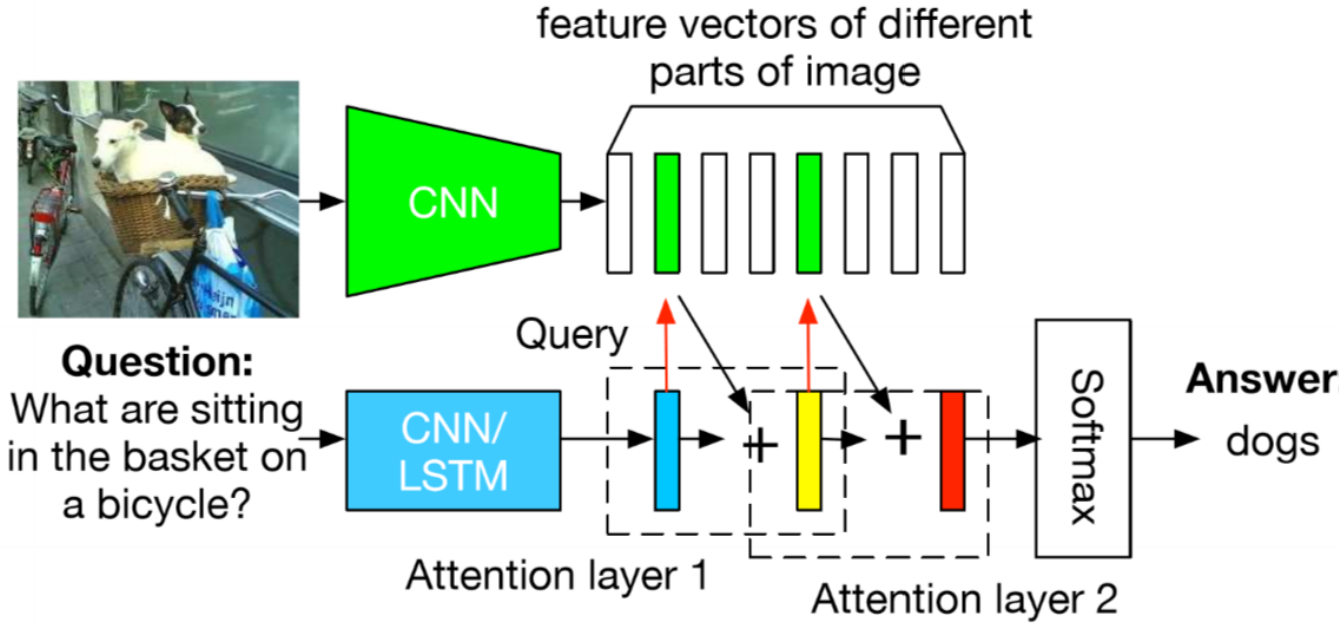
Structured Attention Network [Chen et al. 2017]

Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Stacked attention network (SAN)

SAN consists of three components:

1. Image model
2. Question model
3. Stacked attention model



(a) Stacked Attention Network for Image QA

Yang, Zichao, et al. "Stacked attention networks for image question answering."

SAN: image model

Image model:

1. Extract features VGGNet (last pooling layer)
2. Transform features to the same dimension as question vector.

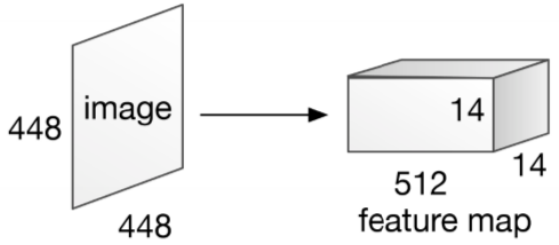


Figure 2: CNN based image model

$$f_I = \text{CNN}_{vgg}(I).$$

$$v_I = \tanh(W_I f_I + b_I),$$

Yang, Zichao, et al. "Stacked attention networks for image question answering."

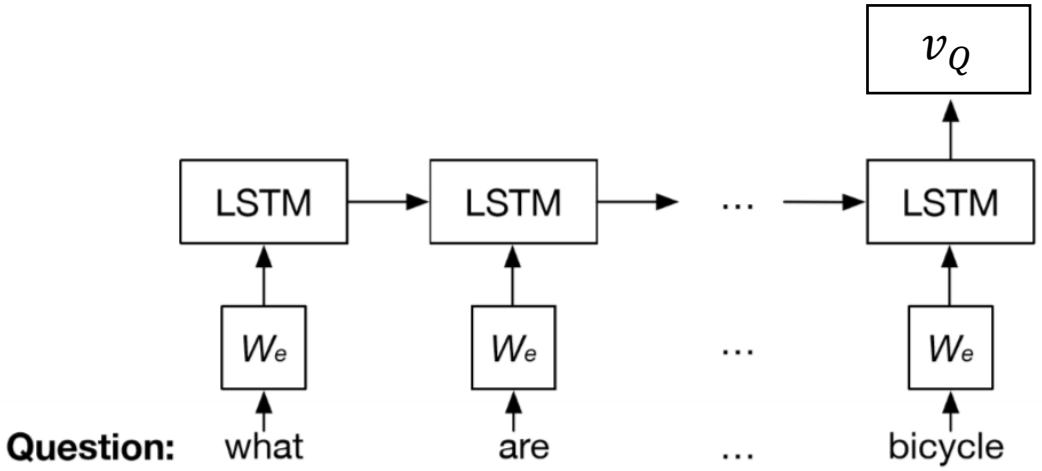
SAN: question model

Question model (LSTM):

1. Get one hot vector representation of words.
2. Get embedding matrix of word vector
3. Feed embedding to LSTM
4. Repeat 2, 3 for every time step.

$$x_t = W_e q_t, t \in \{1, 2, ...T\}, \quad (8)$$

$$h_t = \text{LSTM}(x_t), t \in \{1, 2, ...T\}. \quad (9)$$

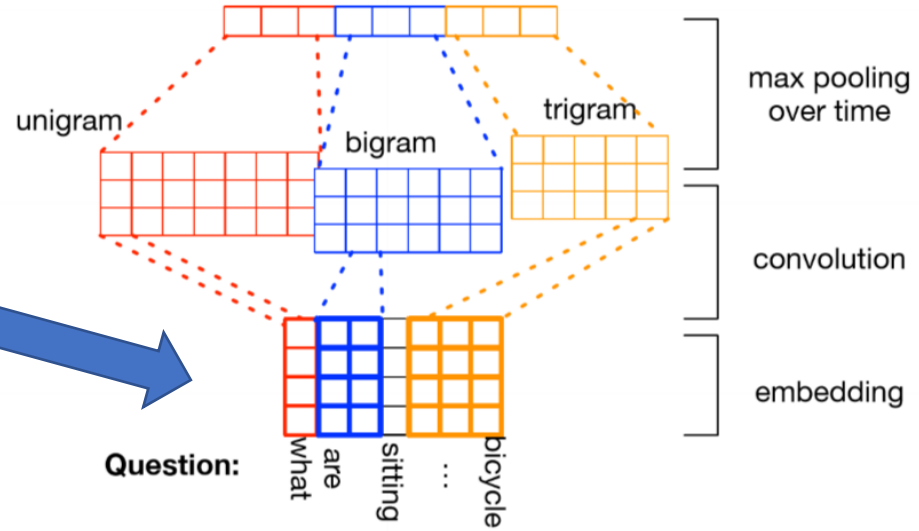


Yang, Zichao, et al. "Stacked attention networks for image question answering."

SAN: question model

Question model (CNN):

1. Get one hot vector representation of words.
2. Get embedding matrix of word vector
3. Concatenate the word vectors



$$x_{1:T} = [x_1, x_2, ..., x_T]. \qquad (10)$$

Figure 4: CNN based question model

Yang, Zichao, et al. "Stacked attention networks for image question answering."

SAN: question model

Question model (CNN):

1. Get one hot vector representation of words.
2. Get embedding matrix of word vector.
3. Concatenate the word vectors.
4. Apply convolution filters.

$$h_{c,t} = \tanh(W_c x_{t:t+c-1} + b_c). \quad (11)$$

$$h_c = [h_{c,1}, h_{c,2}, ..., h_{c,T-c+1}]. \quad (12)$$

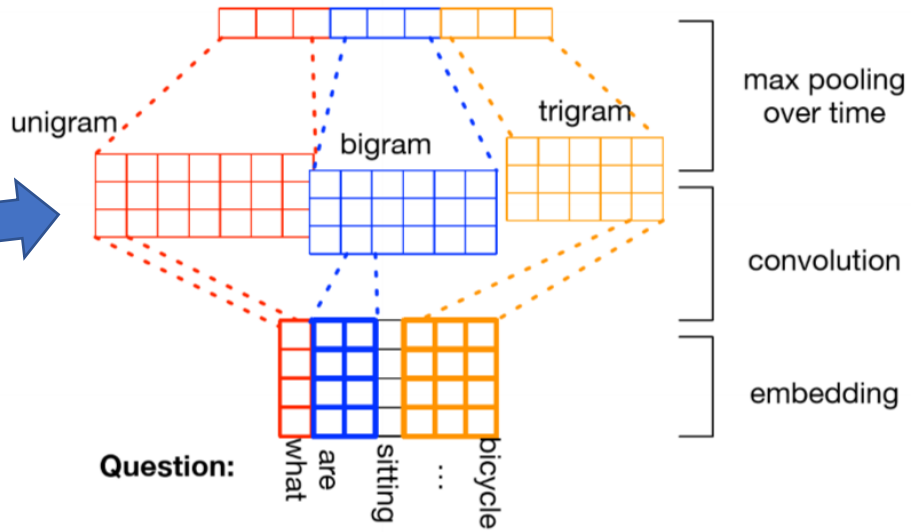


Figure 4: CNN based question model

Yang, Zichao, et al. "Stacked attention networks for image question answering."

SAN: question model

Question model (CNN):

1. Get one hot vector representation of words.
2. Get embedding matrix of word vector.
3. Concatenate the word vectors
4. Apply convolution filters
5. Max-pooling over the feature maps

$$\tilde{h}_c = \max_t [h_{c,1}, h_{c,2}, ..., h_{c,T-c+1}]. \quad (13)$$

$$h = [\tilde{h}_1, \tilde{h}_2, \tilde{h}_3], \quad (14)$$

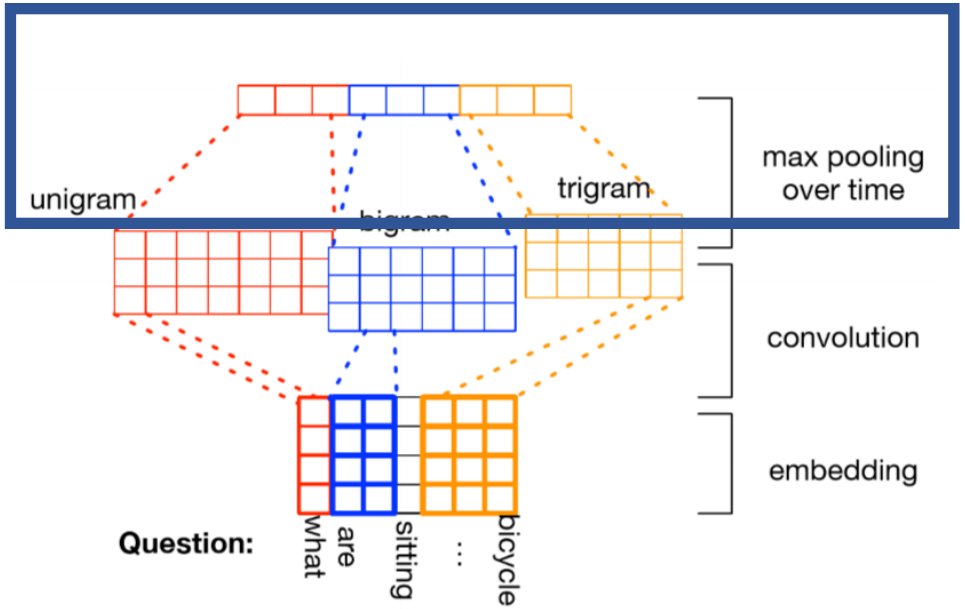
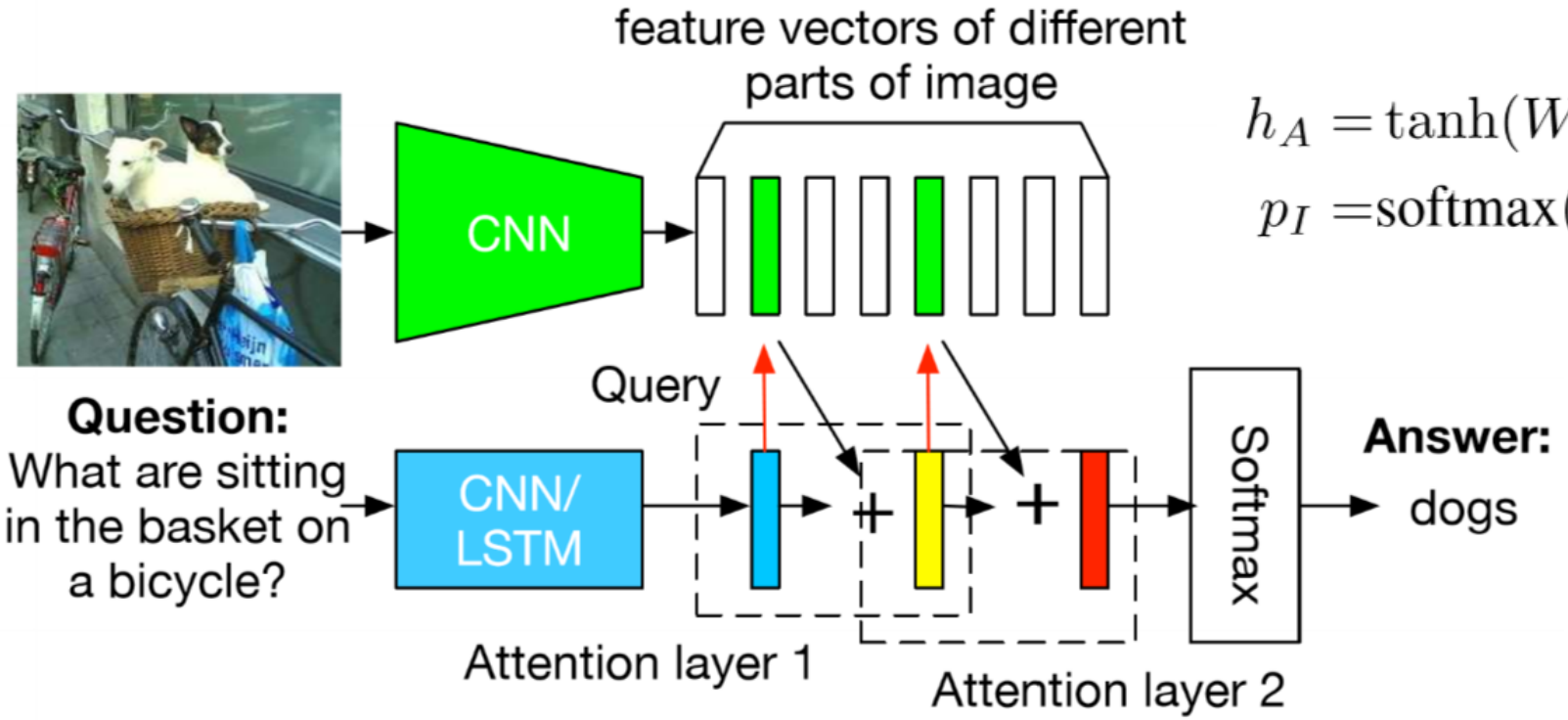


Figure 4: CNN based question model

Yang, Zichao, et al. "Stacked attention networks for image question answering."

Stacked attention network



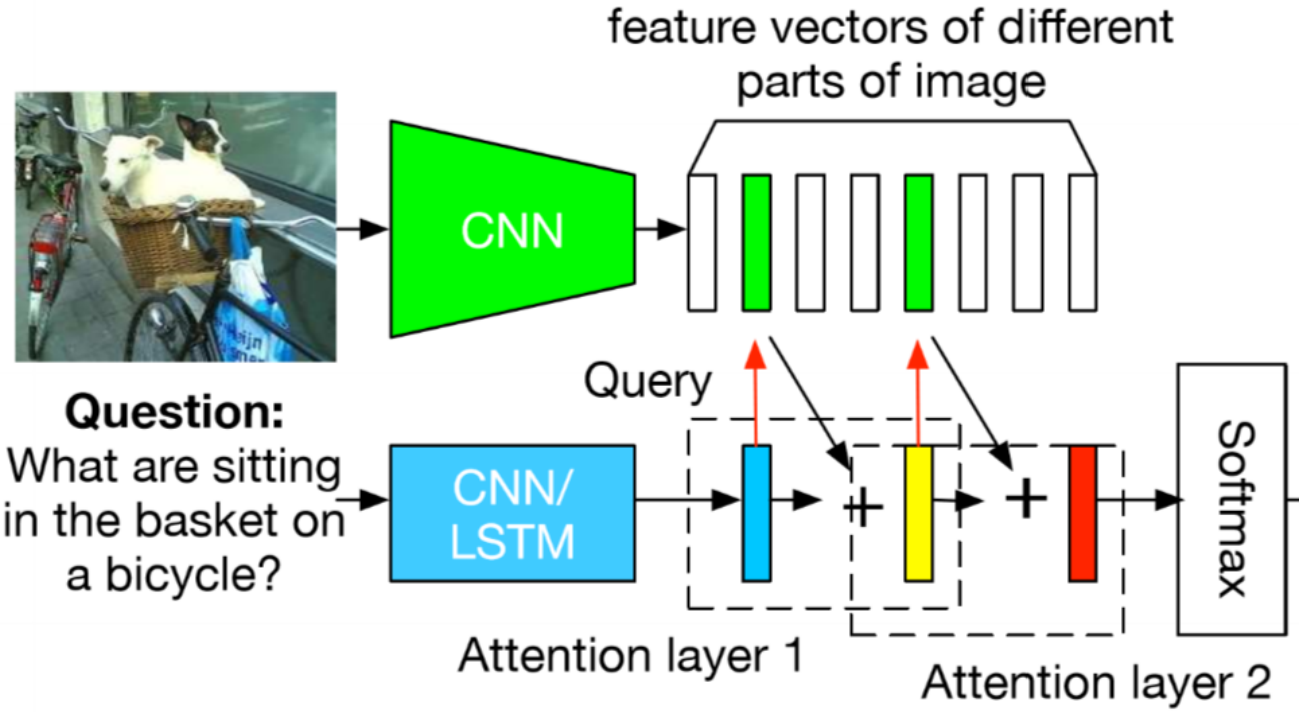
$$h_A = \tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A)), \quad (15)$$

$$p_I = \text{softmax}(W_P h_A + b_P), \quad (16)$$

(a) Stacked Attention Network for Image QA

Yang, Zichao, et al. "Stacked attention networks for image question answering."

Stacked attention network



(a) Stacked Attention Network for Image QA

$$h_A^k = \tanh(W_{I,A}^k v_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k)), \quad (19)$$

$$p_I^k = \text{softmax}(W_P^k h_A^k + b_P^k). \quad (20)$$

$$\tilde{v}_I^k = \sum_i p_i^k v_i, \quad (21)$$

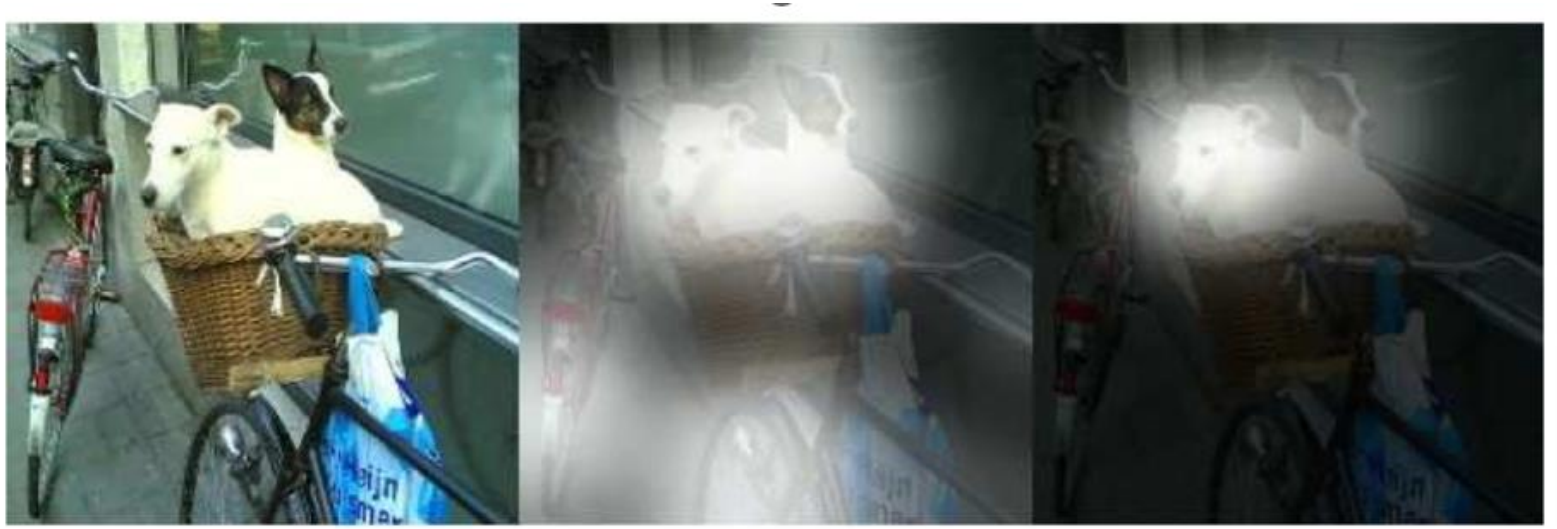
$$u^k = \tilde{v}_I^k + u^{k-1}. \quad (22)$$

$$p_{\text{ans}} = \text{softmax}(W_u u^K + b_u). \quad (23)$$

Yang, Zichao, et al. "Stacked attention networks for image question answering."

VIRGINIA TECH

Stacked attention network



Yang, Zichao, et al. "Stacked attention networks for image question answering."

SAN: performance on VQA

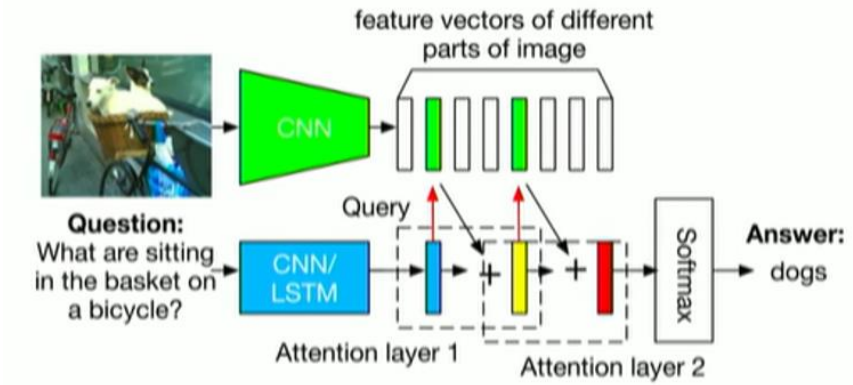
	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

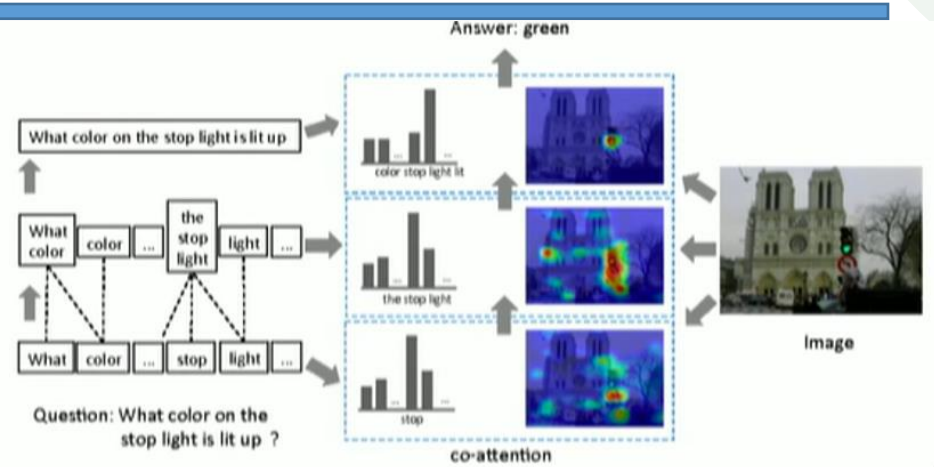
Yang, Zichao, et al. "Stacked attention networks for image question answering."

VIRGINIA TECH

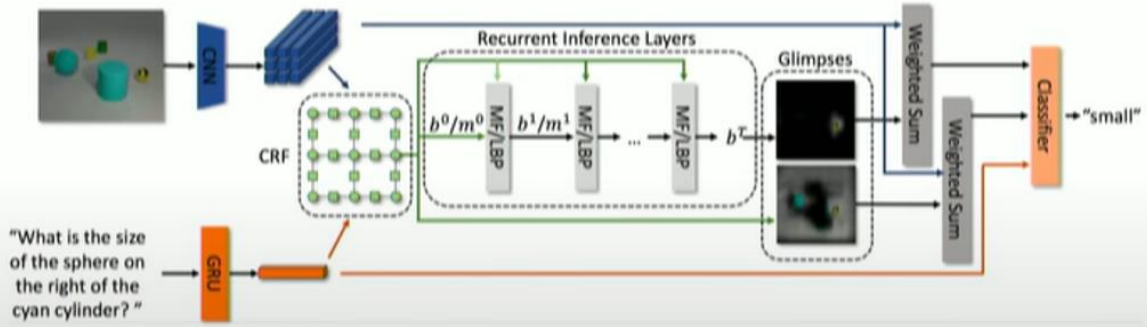
Attention models



Stacked Attention Network [Yang et al. 2016]



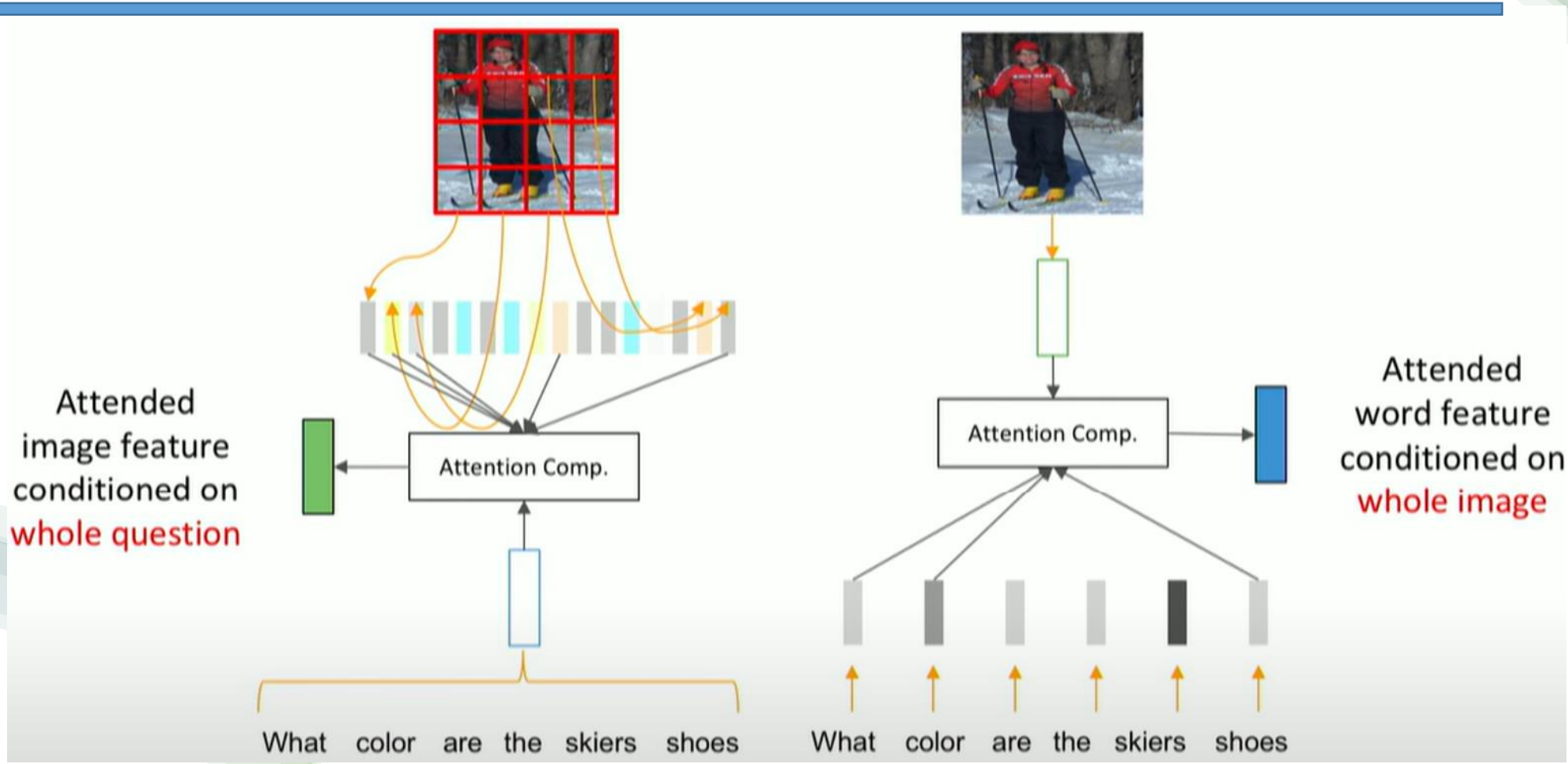
Hierarchical Co-Attention Network [Lu et al. 2016]



Structured Attention Network [Chen et al. 2017]

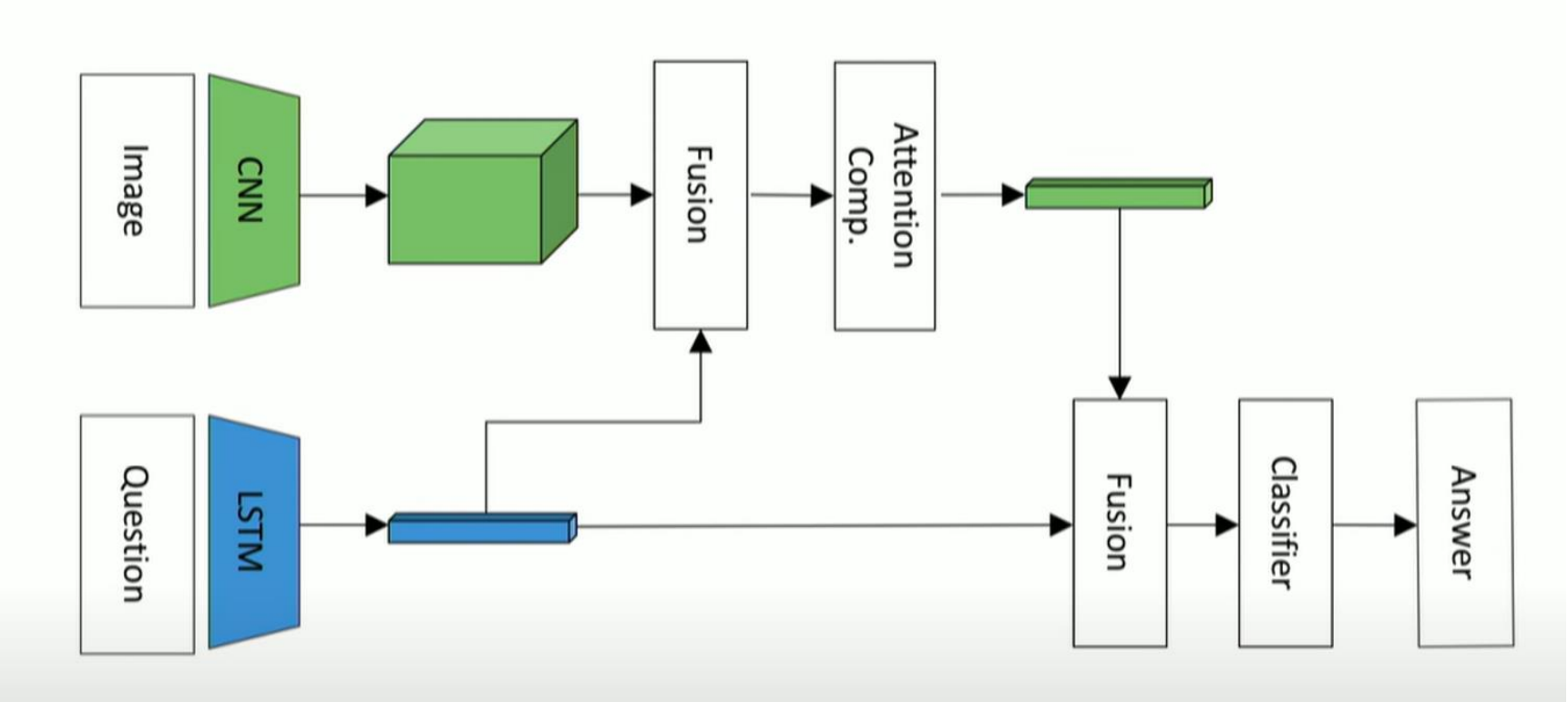
Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Attention models



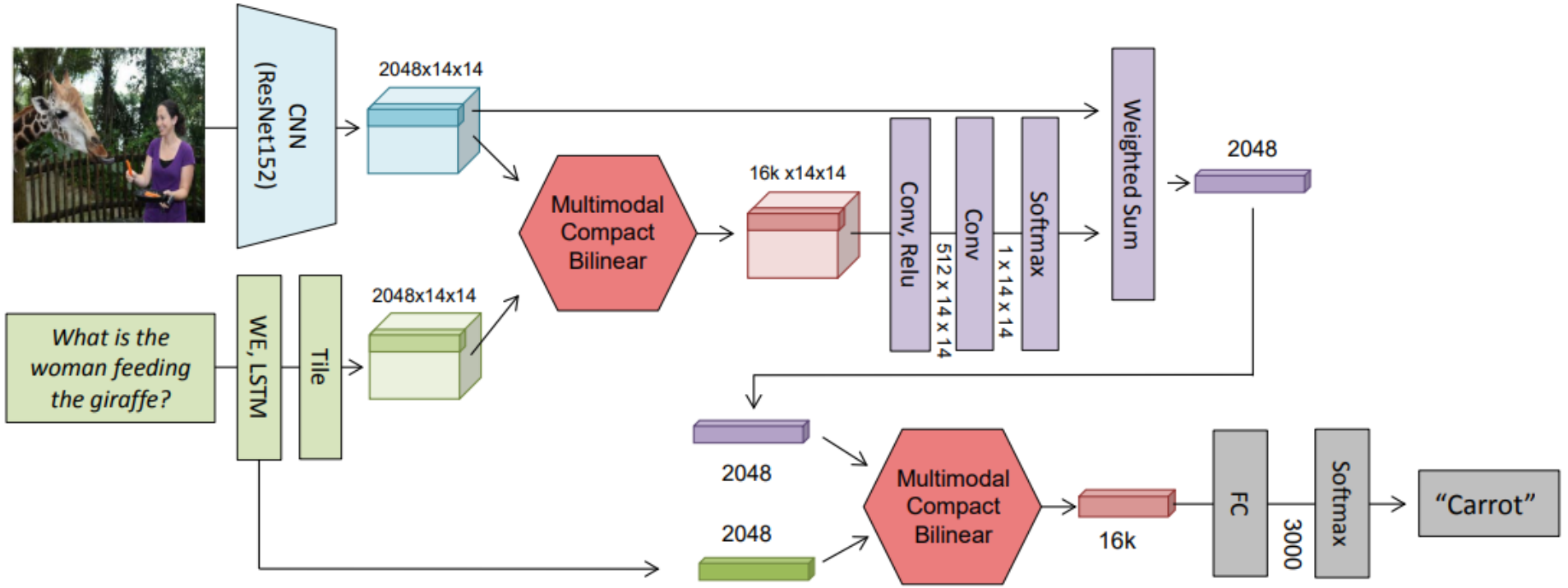
Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Prior related work: feature fusion methods



Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

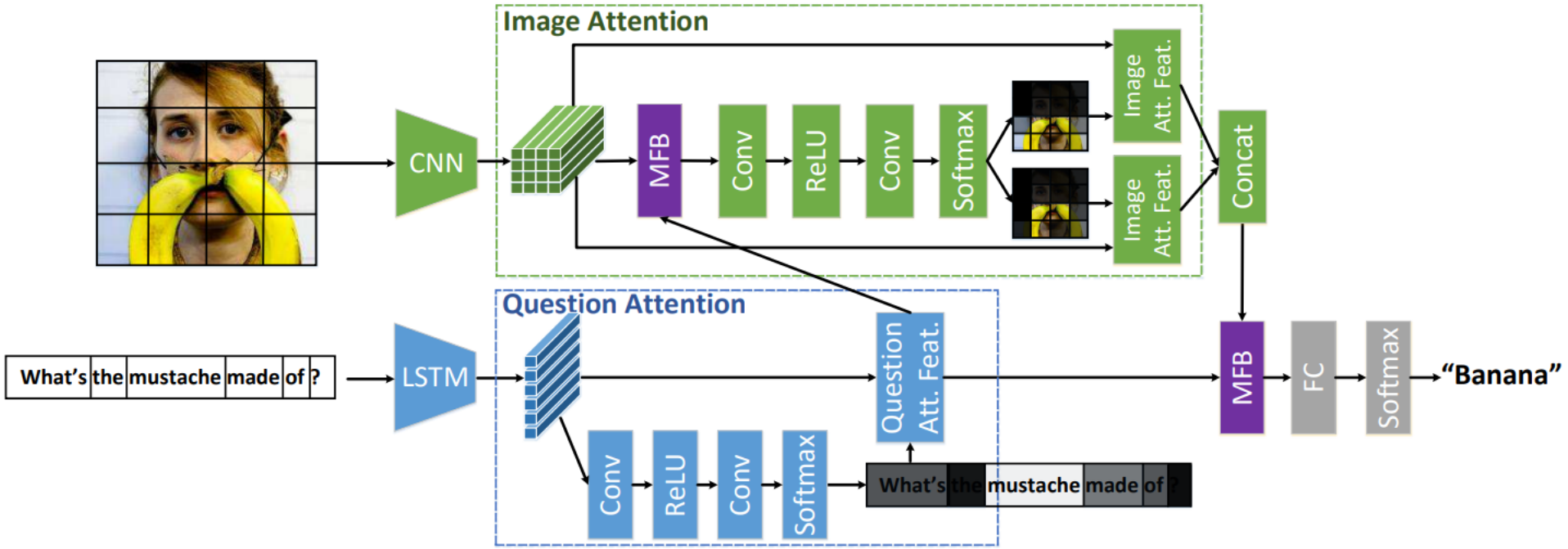
Prior related work: feature fusion models



Multimodal Compact Bilinear (MCB/MLB)[Fukui et al. 2016/ Kim et al. 2017]

VIRGINIA TECH

Prior related work: feature fusion models



Multimodal Factorized Bilinear (MFB)[Yu et al. 2017]

VIRGINIA TECH

Prior related work: limitations

Attention computation is sparse

- Global question representation to image regions
- Global image representation to question words

Bilinear fusion is complex

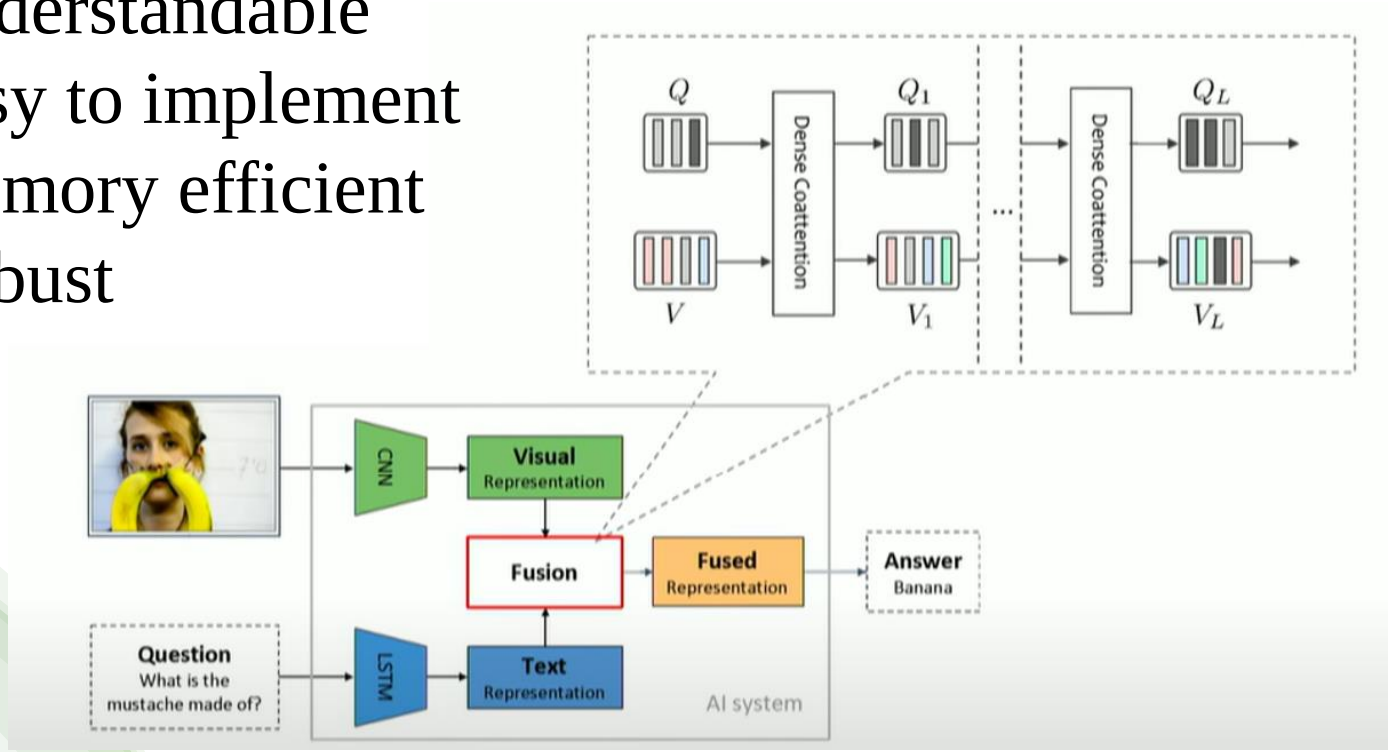
- Sensitive to hyperparameters
- Memory-inefficient
- May lead to unstable training

VIRGINIA TECH

Slide adapted from the authors presentation at (7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube

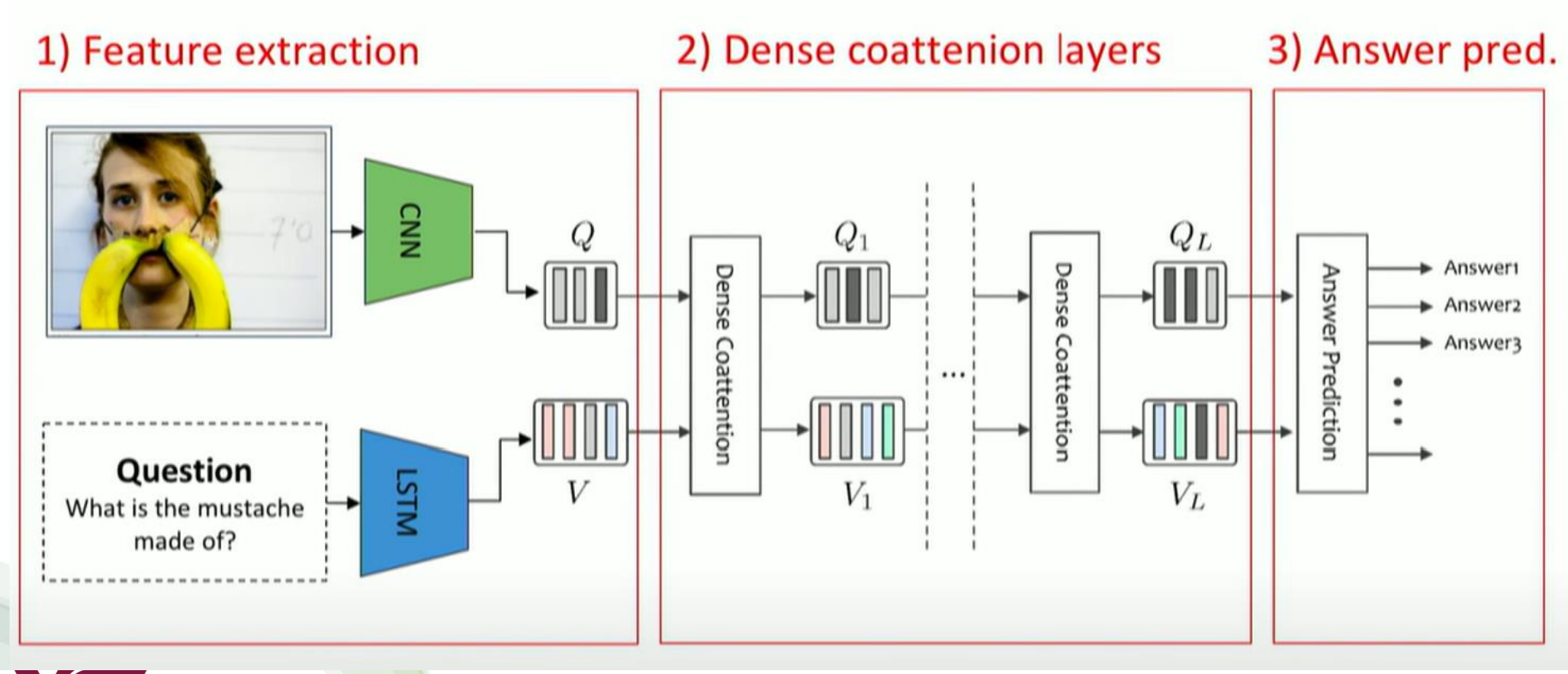
Goals of Dense Co-attention Network

1. Improve performance
2. Understandable
3. Easy to implement
4. Memory efficient
5. Robust



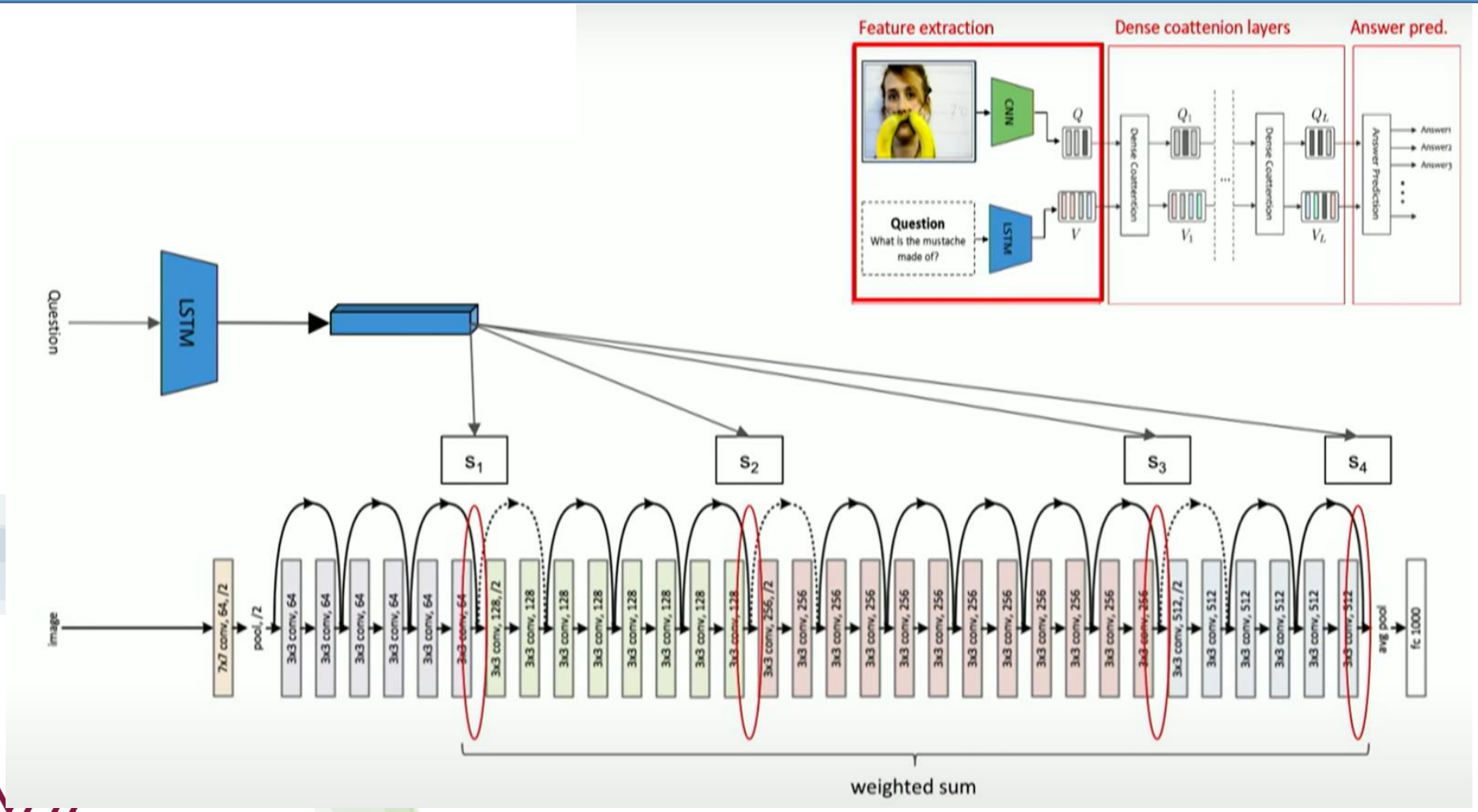
Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Model overview



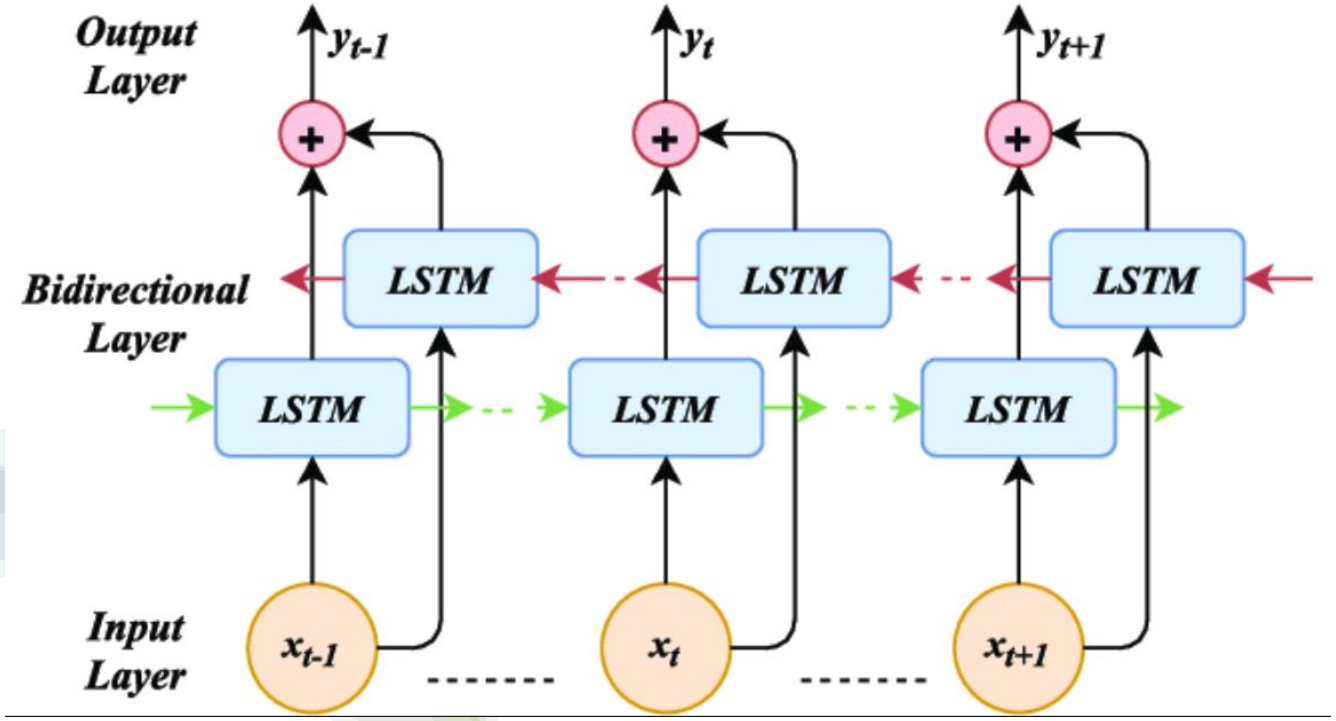
Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Feature extraction



Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Question and answer extraction



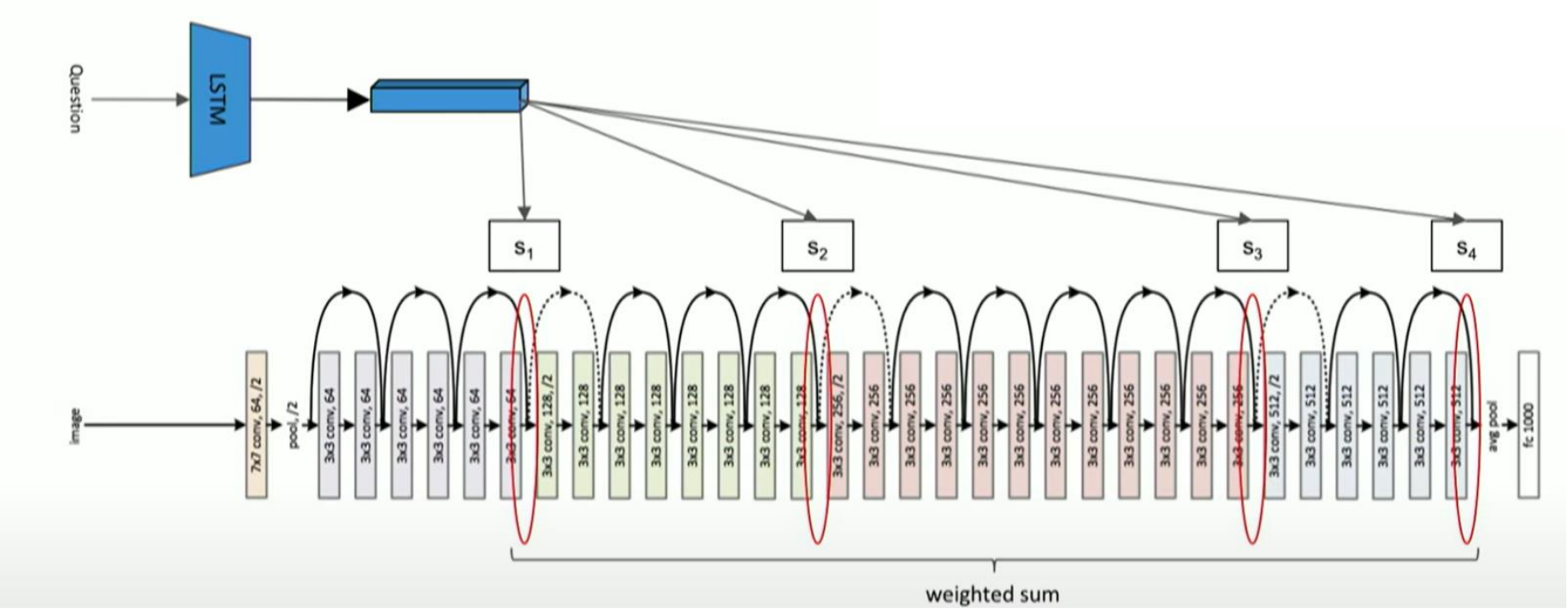
$$\overrightarrow{q_n} = \text{Bi-LSTM}(\overrightarrow{q_{n-1}}, e_n^Q), \quad (1)$$

$$\overleftarrow{q_n} = \text{Bi-LSTM}(\overleftarrow{q_{n+1}}, e_n^Q). \quad (2)$$

VIRGINIA TECH

Ihianle, Isibor Kennedy, et al. "A deep learning approach for human activities recognition from multimodal sensing devices." *IEEE Access* 8 (2020): 179028-179038.

Image extraction

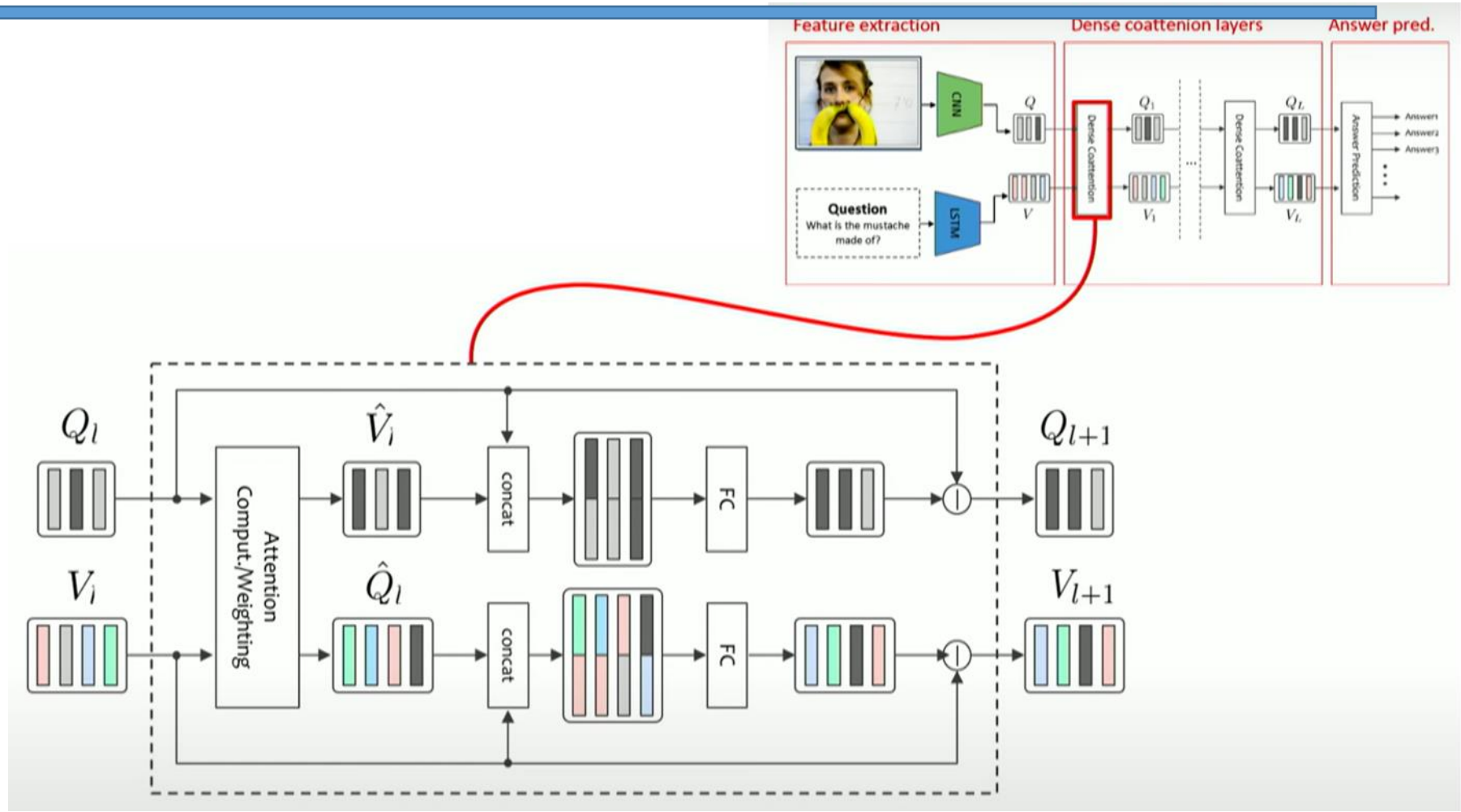


$$s_Q = [\overrightarrow{q_N}^\top, \overleftarrow{q_1}^\top]^\top$$

$$[s_1, s_2, s_3, s_4] = \text{MLP}(s_Q)$$

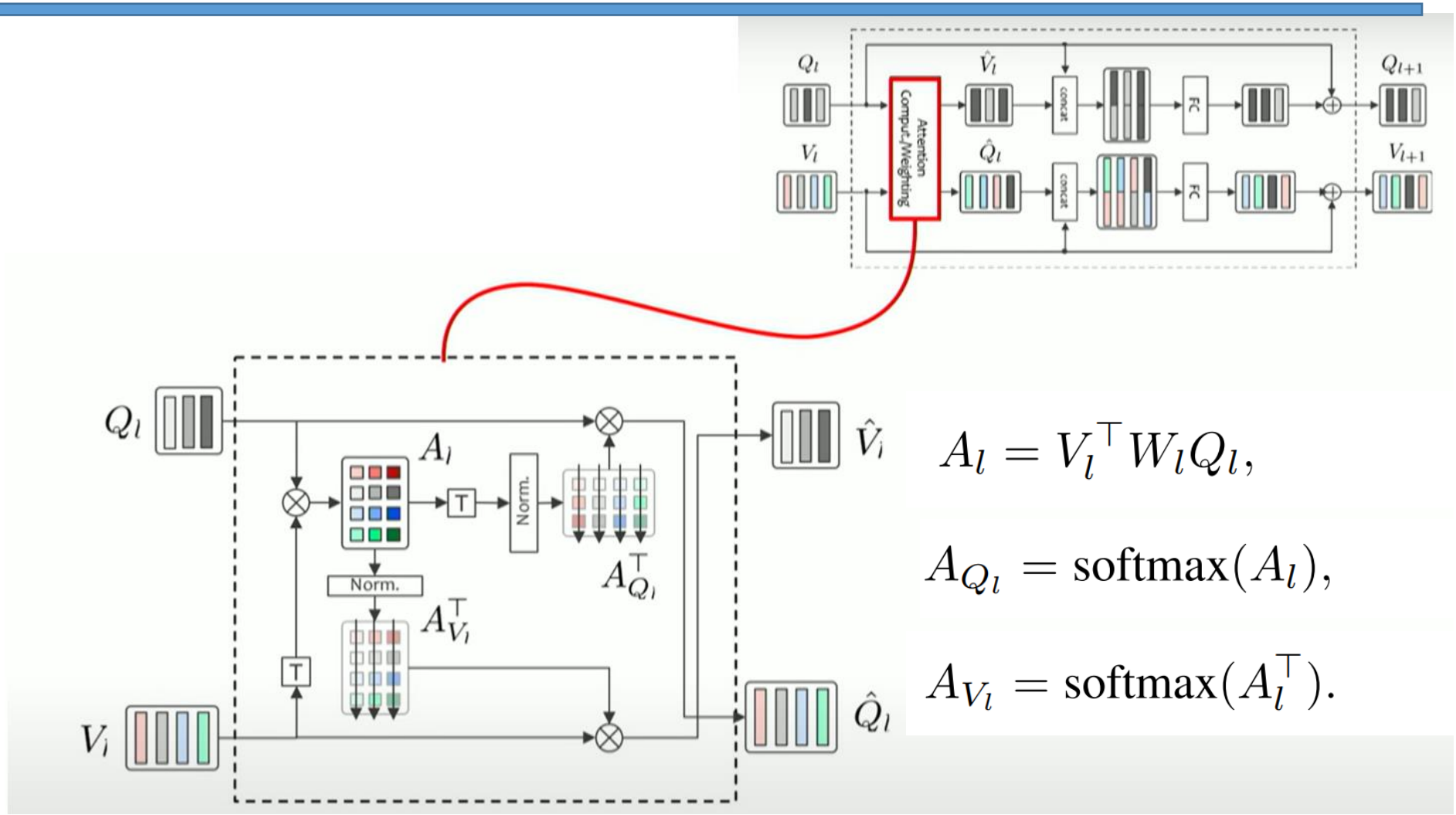
Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Dense co-attention layer



Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Dense co-attention layer: attention map



$$A_l = V_l^\top W_l Q_l, \tag{4}$$

$$A_{Q_l} = \text{softmax}(A_l), \tag{5}$$

$$A_{V_l} = \text{softmax}(A_l^\top). \tag{6}$$

Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

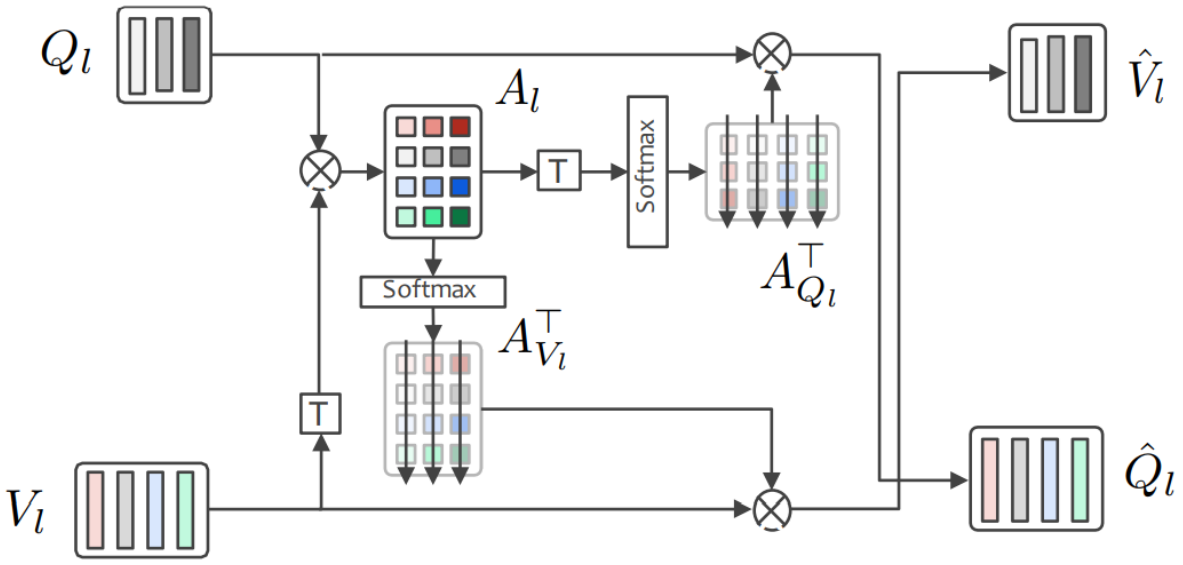
Dense co-attention layer: nowhere-to-attend

When creating attention map, there is no particular region or word the model should attend?

Solution: Add elements to the attention maps that serve as memory for storing information

VIRGINIA TECH

Dense co-attention layer: parallel attention



$$A_l^{(i)} = (W_{\tilde{V}_l}^{(i)} \tilde{V}_l)^\top (W_{\tilde{Q}_l}^{(i)} \tilde{Q}_l). \tag{7}$$

$$A_{Q_l}^{(i)} = \text{softmax}\left(\frac{A_l^{(i)}}{\sqrt{d_h}}\right), \tag{8}$$

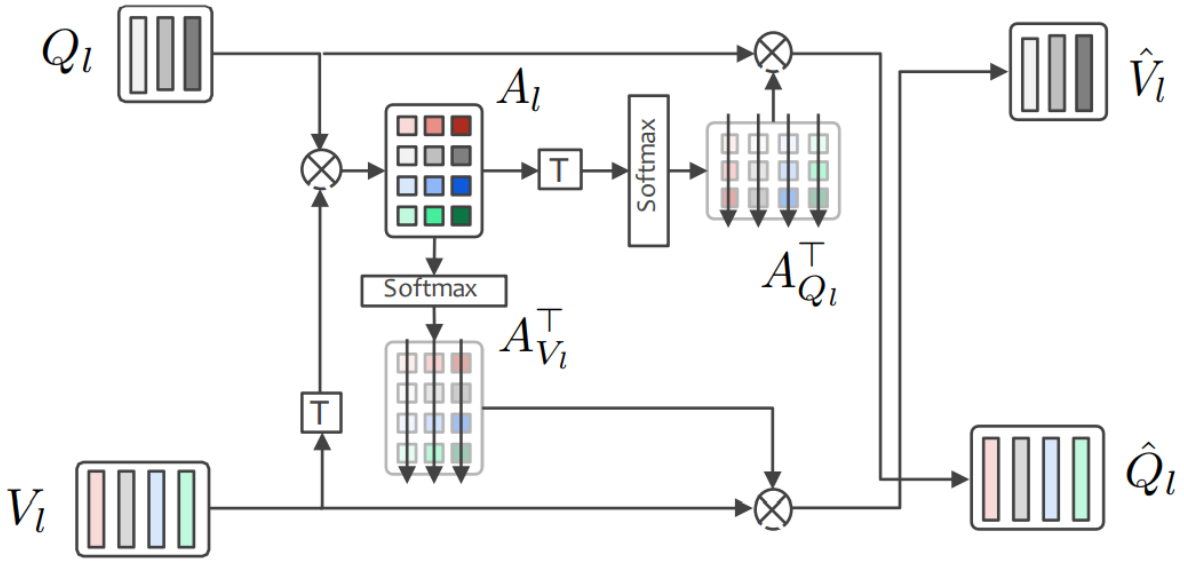
$$A_{V_l}^{(i)} = \text{softmax}\left(\frac{A_l^{(i)\top}}{\sqrt{d_h}}\right). \tag{9}$$

$$A_{Q_l} = \frac{1}{h}\sum_{i=1}^{h} A_{Q_l}^{(i)}, \tag{10}$$

$$A_{V_l} = \frac{1}{h}\sum_{i=1}^{h} A_{V_l}^{(i)}. \tag{11}$$

Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Dense co-attention layer: attended feature



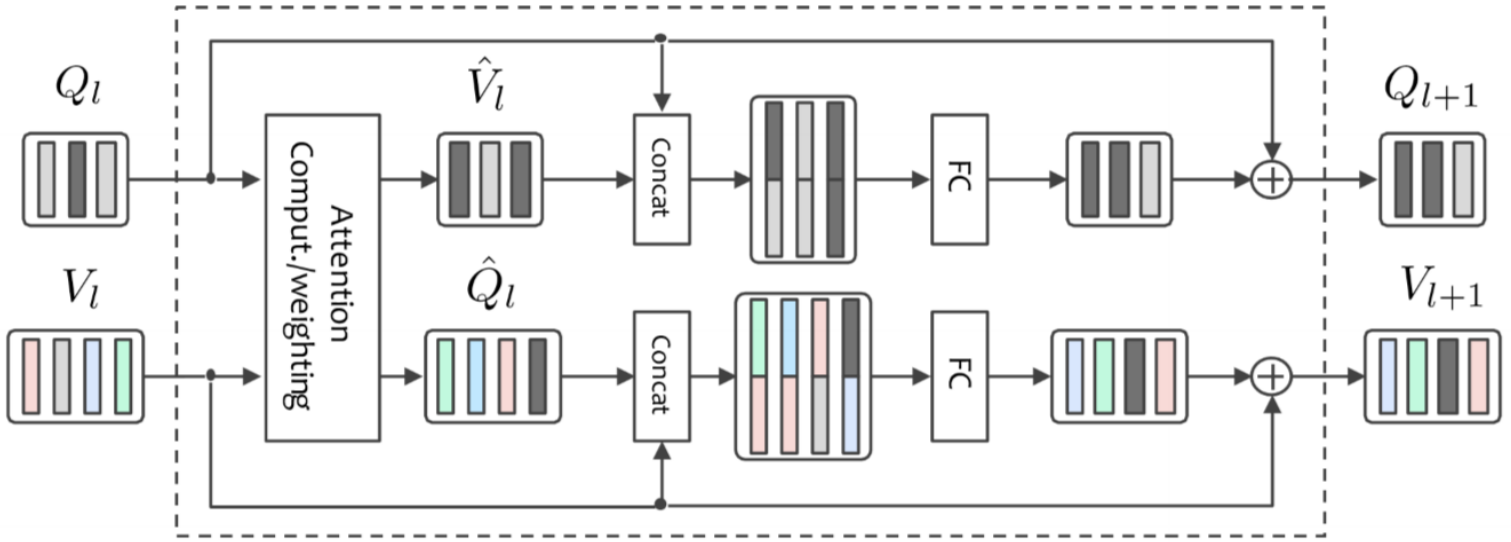
$$\hat{Q}_l = \tilde{Q}_l \, A_{Q_l}[1{:}\mathtt{T}, :]^\top, \tag{12}$$

$$\hat{V}_l = \tilde{V}_l \, A_{V_l}[1{:}\mathtt{N}, :]^\top, \tag{13}$$

Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

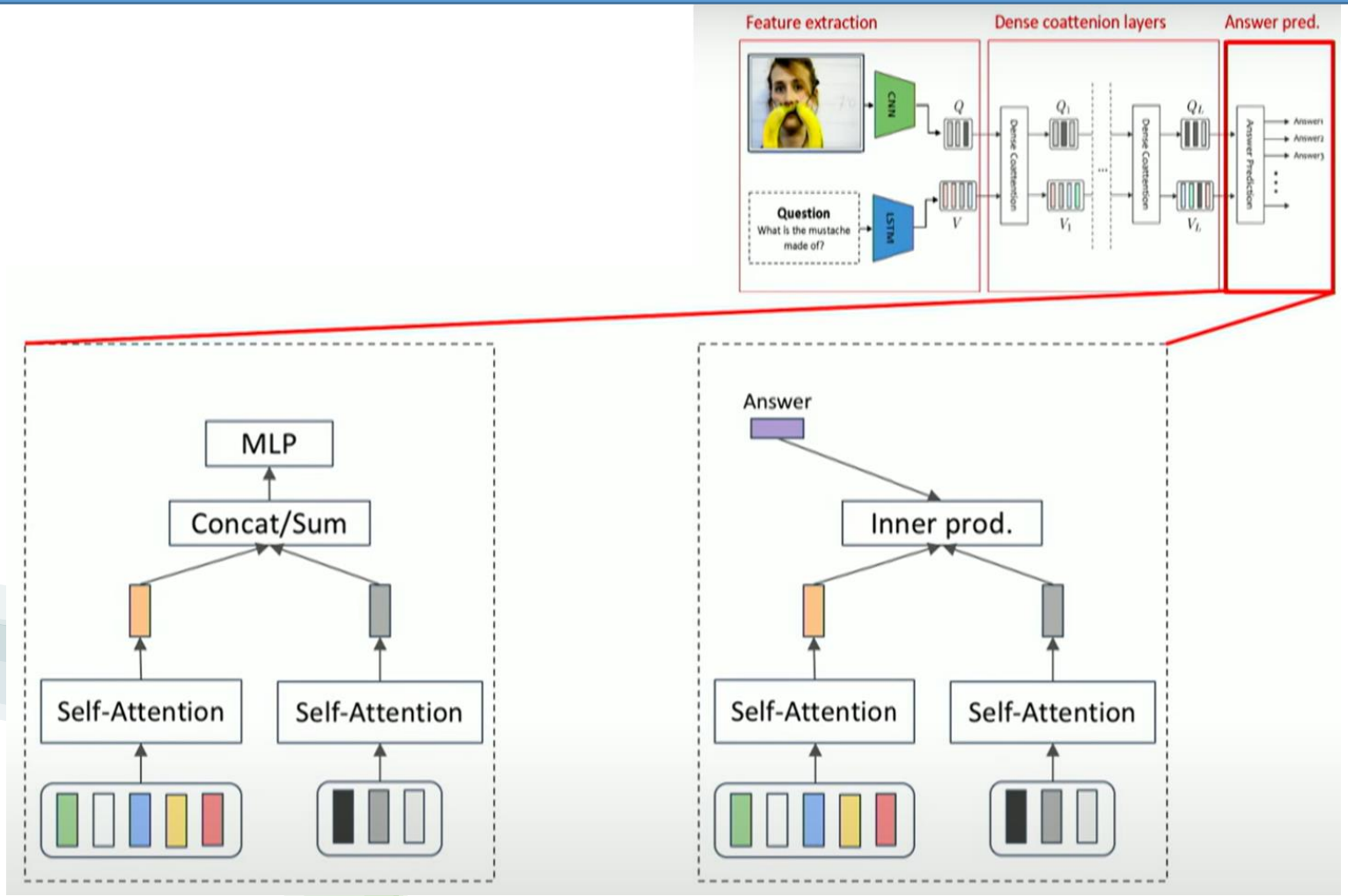
Fusing image and question representations

$$q_{(l+1)n} = \text{ReLU}\left(W_{Q_l}\begin{bmatrix} q_{ln} \\ \hat{v}_{ln} \end{bmatrix} + b_{Q_l}\right) + q_{ln}, \qquad (14)$$



VIRGINIA TECH

Answer prediction



$$\sigma\left(s_A^\top W\left(s_{Q_L} + s_{V_L}\right)\right), \quad (16)$$

$$\sigma\left(\mathrm{MLP}\left(s_{Q_L} + s_{V_L}\right)\right), \quad (17)$$

$$\sigma\left(\mathrm{MLP}\left(\begin{bmatrix} s_{Q_L} \\ s_{V_L} \end{bmatrix}\right)\right), \quad (18)$$

Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Answer prediction: self-attention function

Calculate aggregated representation of the whole question s_{QL}:

1. Apply an identical two-layer MLP with ReLU nonlinearity in its hidden layer
2. Apply softmax to get attention weights, $\alpha_1^Q, \ldots, \alpha_N^Q$
3. $s_{Q_L} = \sum_{n=1}^{N} \alpha_n^Q q_{Ln}$

VIRGINIA TECH

Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Model evaluation: dataset

- VQA 1.0
  - Human-annotated question-answer pairs on images from MS COCO.
- VQA 2.0
  - Larger and more balanced compared.

VIRGINIA TECH

Model evaluation: method

Candidate answer:

- VQA: more than 5 times
- VQA 2.0: more than 8 times

Training and evaluation:

- Trian on train + val splits
- Report on test-dev and test standard

Metric:

$$\text{accuracy} = \min\left(\frac{\#\text{ humans that provided that answer}}{3}, 1\right)$$

VIRGINIA TECH

Model evaluation: setup

- Adam optimizer with $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.99$
- Learning rate decay
- Dropouts
- Feature space dimension is 1024

VIRGINIA TECH™

Model evaluation: ablation study

Evaluate different components of the model with VQA 2.0

VIRGINIA TECH

Table 1: Ablation study on each module of DCNs using the validation set of the Open-Ended task (VQA 2.0). * indicates modules employed in the final model.

Category	Detail	Accuracy
Attention direction	I $\leftarrow$ Q	60.95
	I $\rightarrow$ Q	62.63
	I $\leftrightarrow$ Q*	62.94
Memory size (K)	1	62.53
	3*	62.94
	5	62.83
Number (h) of parallel attention maps	2	62.82
	4*	62.94
	8	62.81
Number (L) of stacked layers	1	62.43
	2	62.82
	3*	62.94
	4	62.67
Attention in answer prediction layer	Attention used*	62.94
	Avg of features	61.63
Attention in image extraction layer	Attention used*	62.94
	Only last conv layer	62.39

VIRGINIA TECH

Comparison with other methods on VQA

$$(\text{score of answers}) = \sigma\Big(\text{MLP}\big(s_{Q_L} + s_{V_L}\big)\Big), \qquad (17)$$

Table 2: Results of the proposed method along with published results of others on VQA 1.0 in similar conditions (i.e., a single model; trained without an external dataset).

Model	Test-dev				Test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
VQA team [2]	57.75	43.08	36.77	80.50	58.16	43.73	36.53	80.569
SMem [27]	57.99	43.12	37.32	80.87	58.24	43.48	37.53	80.80
SAN [28]	58.70	46.10	36.60	79.30	58.90	-	-	-
FDA [11]	59.24	45.77	36.16	81.14	59.54	-	-	-
DNMN [1]	59.40	45.50	38.60	81.10	59.40	-	-	-
HieCoAtt [17]	61.00	51.70	38.70	79.70	62.10	-	-	-
RAU [20]	63.30	53.00	39.00	81.90	63.20	52.80	38.20	81.70
DAN [19]	64.30	53.90	39.10	83.00	64.20	54.00	38.10	82.80
Strong Baseline [12]	64.50	55.20	39.10	82.20	64.60	55.20	39.10	82.00
MCB [5]	64.70	55.60	37.60	82.50	-	-	-	-
N2NMNs [10]	64.90	-	-	-	-	-	-	-
MLAN [31]	64.60	53.70	40.20	83.80	64.80	53.70	40.90	83.70
MLB [14]	65.08	54.87	38.21	84.14	65.07	54.77	37.90	84.02
MFB [32]	65.90	56.20	39.80	84.00	65.80	56.30	38.90	83.80
MF-SIG-T3 [4]	66.00	56.37	39.34	84.33	65.88	55.89	38.94	84.42
DCN (16)	66.43	56.23	42.37	84.75	66.39	56.23	41.81	84.53
DCN (17)	**66.89**	57.31	**42.35**	**84.61**	**67.02**	**56.98**	**42.34**	**85.04**
DCN (18)	66.83	**57.44**	41.66	84.48	66.66	56.83	41.27	84.61

VIRGINIA TECH

Comparison with other methods on VQA 2.0

Table 3: Results of the proposed method along with published results of others on VQA 2.0 in similar conditions (i.e., a single model; trained without an external dataset). DCN(number) indicates the DCN equipped with the prediction layer that uses equation (number) for score computation. *: trained with external datasets. ‡: the winner of VQA challenge 2017, unpublished.

Model	Test-dev				Test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
VQA team-Prior [7]	-	-	-	-	25.98	01.17	00.36	61.20
VQA team-Language only [7]	-	-	-	-	44.26	27.37	31.55	67.01
VQA team-LSTM+CNN [7]	-	-	-	-	54.22	41.83	35.18	73.46
MCB [5] reported in [7]	-	-	-	-	62.27	53.36	38.28	78.82
MF-SIG-T3 * [4]	64.73	55.55	42.99	81.29	-	-	-	-
Adelaide Model * ‡ [24]	62.07	52.62	39.46	79.20	62.27	52.59	39.77	79.32
Adelaide + Detector * ‡ [24]	65.32	56.05	44.21	81.82	65.67	56.26	43.90	82.20
DCN (16)	**66.87**	**57.26**	46.61	83.51	66.97	**57.09**	46.98	83.59
DCN (17)	66.72	56.77	**46.65**	**83.70**	**67.04**	56.95	**47.19**	83.85
DCN (18)	66.60	56.72	46.60	83.50	67.00	56.90	46.93	**83.89**

VIRGINIA TECH

Qualitative Evaluation

Figure 4: Typical examples of attended image regions and question words for complementary image-question pairs from VQA 2.0 dataset. Each row contains visualization for two pairs of the same question but different images and answers. The original image and question are shown along with their attention maps generated in the answer prediction layer. The brightness of image pixels and redness of words indicate the attention weights.

VIRGINIA TECH

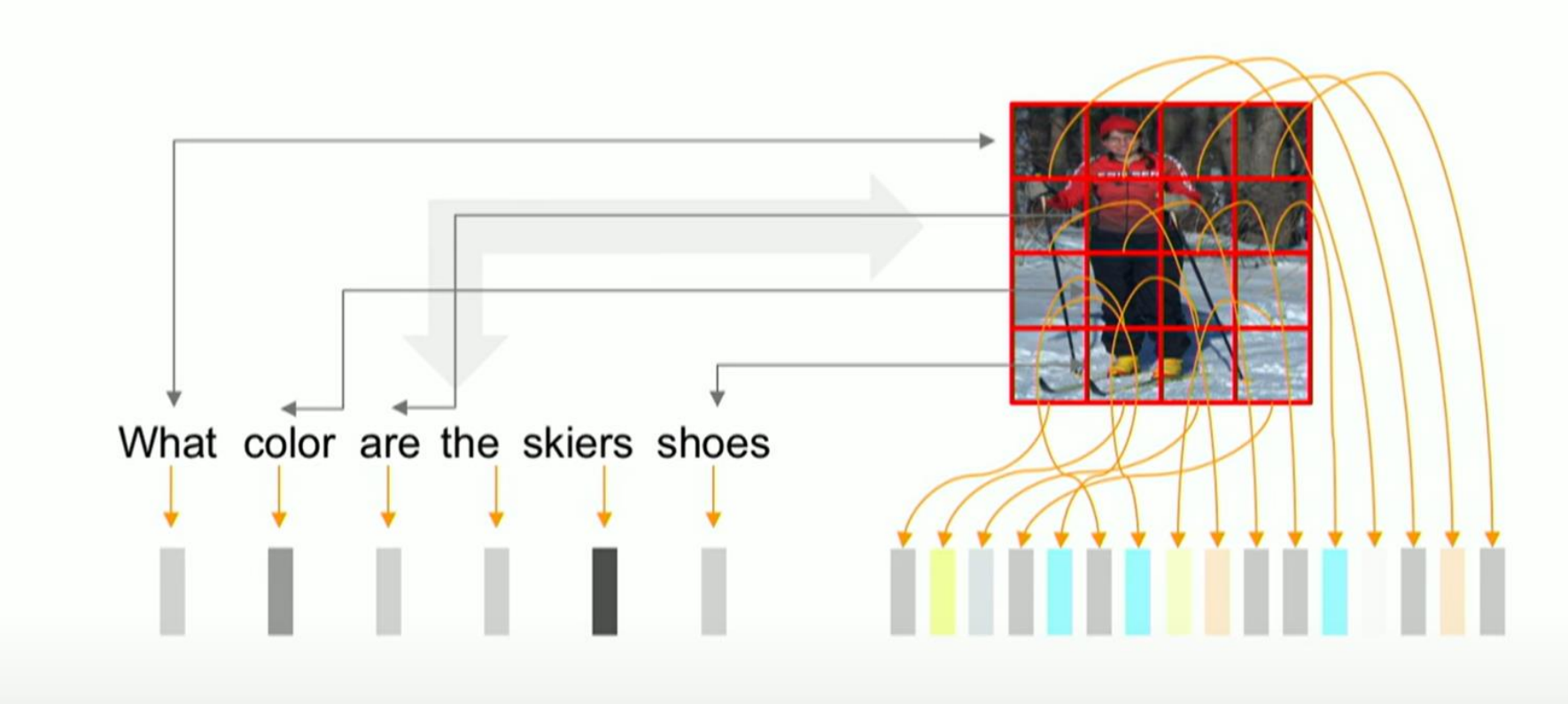
Strength: Memory efficient + fast convergence

Model	No. Params	Iteration
MLB [Kim et al. 2017]	25M	~250K
MFB [Yu et al. 2017]	46M	~100K
MF-SIG-T3 [Chen et al. 2017]	53M	~100K
DCN (concat)	32M	~45K
DCN (sum)	31M	~45K
DCN (inner)	28M	~45K

VIRGINIA TECH

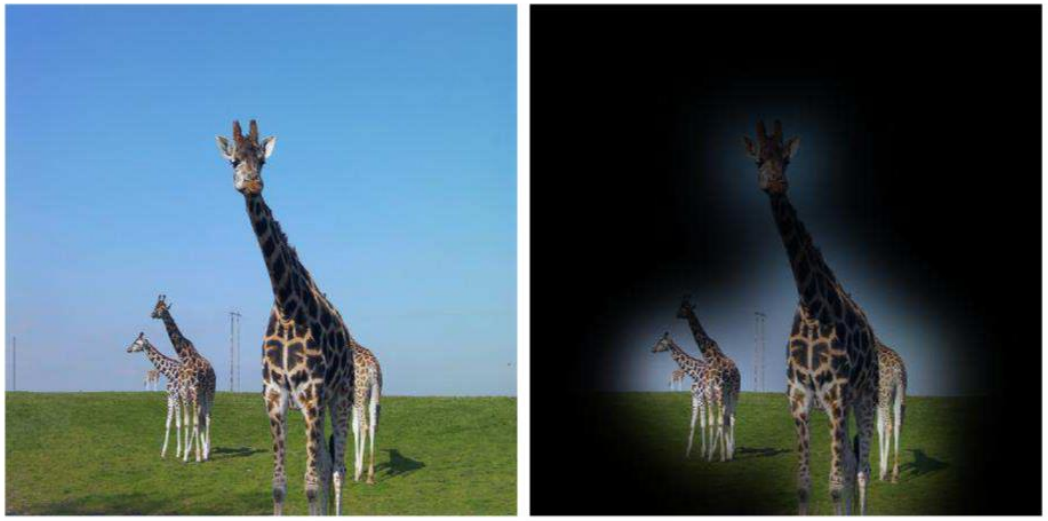
Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Strength: fully symmetrical attention



Slide adapted from the authors presentation at [(7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube]

Strength: understandable via attention map



What are these animals What are these animals

Pred: Giraffes, Ans: Giraffes

What are these animals What are these animals

Pred: Cows, Ans: Cows

Slide adapted from the authors presentation at (7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube

Weakness:

Table 1: Ablation study on each module of DCNs using the validation set of the Open-Ended task (VQA 2.0). * indicates modules employed in the final model.

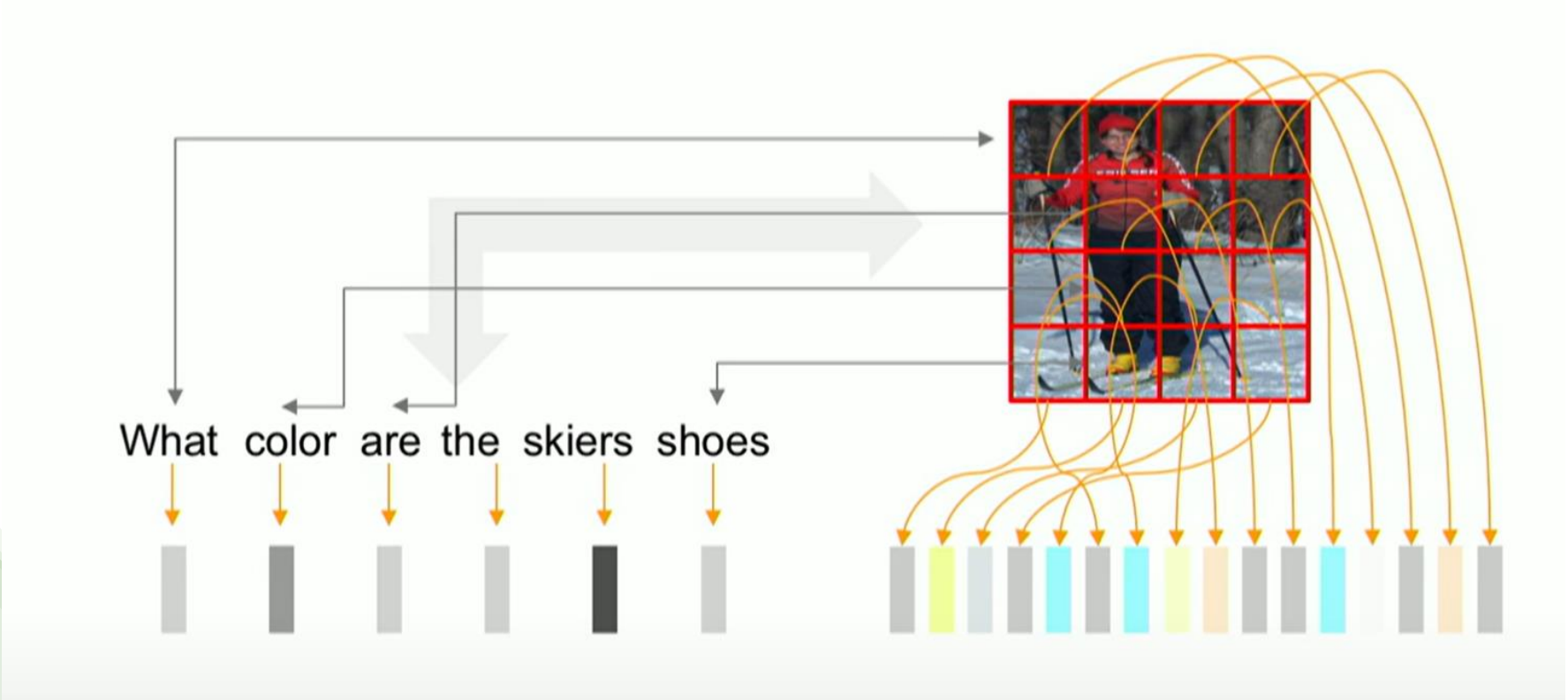
Category	Detail	Accuracy
Attention direction	$I \leftarrow Q$	60.95
	$I \rightarrow Q$	62.63
	$I \leftrightarrow Q$*	62.94
Memory size (K)	1	62.53
	3*	62.94
	5	62.83
Number (h) of parallel attention maps	2	62.82
	4*	62.94
	8	62.81
Number (L) of stacked layers	1	62.43
	2	62.82
	3*	62.94
	4	62.67
Attention in answer prediction layer	Attention used*	62.94
	Avg of features	61.63
Attention in image extraction layer	Attention used*	62.94
	Only last conv layer	62.39

What values other parameters (K, h, L) were set to when calculating the accuracy?

VIRGINIA TECH

Weakness:

Are we considering too many redundant relationship since the performance did not improve a lot?



Slide adapted from the authors presentation at (7) CVPR18: Session 2-2B: Object Recognition & Scene Understanding III - YouTube

Weakness:

$$A_{Q_l}^{(i)} = \text{softmax}\left(\frac{A_l^{(i)}}{\sqrt{d_h}}\right), \tag{8}$$

$$A_{V_l}^{(i)} = \text{softmax}\left(\frac{A_l^{(i)\top}}{\sqrt{d_h}}\right). \tag{9}$$

Why are we dividing affinity matrix by $\sqrt{d_h}$?

VIRGINIA TECH

Future work:

- Further investigate the effectiveness of dense co-attention layer. (e.g., visualize attention map on image after every attention layer)
- Incorporate other guidance, such as an object detector, into the framework, instead purely rely on the input image and text only.

VIRGINIA TECH

Reference:

Manmadhan, Sruthy, and Binsu C. Kovoor. "Visual question answering: a state-of-the-art review." Artificial Intelligence Review 53 (2020): 5705-5745.

Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." Advances in neural information processing systems 29 (2016).

Ihianle, Isibor Kennedy, et al. "A deep learning approach for human activities recognition from multimodal sensing devices." *IEEE Access* 8 (2020): 179028-179038.

Yang, Zichao, et al. "Stacked attention networks for image question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Fukui, Akira, et al. "Multimodal compact bilinear pooling for visual question answering and visual grounding." *arXiv preprint arXiv:1606.01847* (2016).

Kim, Jin-Hwa, et al. "Hadamard product for low-rank bilinear pooling." *arXiv preprint arXiv:1610.04325* (2016).

Yu, Zhou, et al. "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering." *Proceedings of the IEEE international conference on computer vision*. 2017.

VIRGINIA TECH

Discussions and questions

VIRGINIA TECH™

SAN: performance on VQA

	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

Yang, Zichao, et al. "Stacked attention networks for image question answering."

VIRGINIA TECH