电子科技大学
University of Electronic Science and Technology of China

SMILE LAB
Statistical Machine Intelligence & Learning

Deep Attention Neural Tensor Network for Visual Question Answering

Yalong Bai[1,2], Jianlong Fu[3], Tiejun Zhao[1], and Tao Mei[2]

[1] Harbin Institute of Technology, Harbin, China
[2] JD AI Research, Beijing, China
[3] Microsoft Research Asia, Beijing, China
{baiyalong,tmei}@jd.com, jianf@microsoft.com, tjzhao@hit.edu.cn

Speaker: 贺丽荣
Supervisor: 徐增林教授
2018.09.28

Outline

- Review
  - Classification based Methods
  - Image-question-answer Triplet based Reasoning
- Deep Attention Neural Tensor Network for Visual Question Answering
- Conclusion

Outline

- Review
  - **Classification based Methods**
  - Image-question-answer Triplet based Reasoning
- Deep Attention Neural Tensor Network for Visual Question Answering
- Conclusion

Classification based Methods

- First order interactions like concatenation or element-wise product

$$q_i^l = \phi(W_l * (q^{l-1} \boxed{\otimes} v_i) + b_l)$$

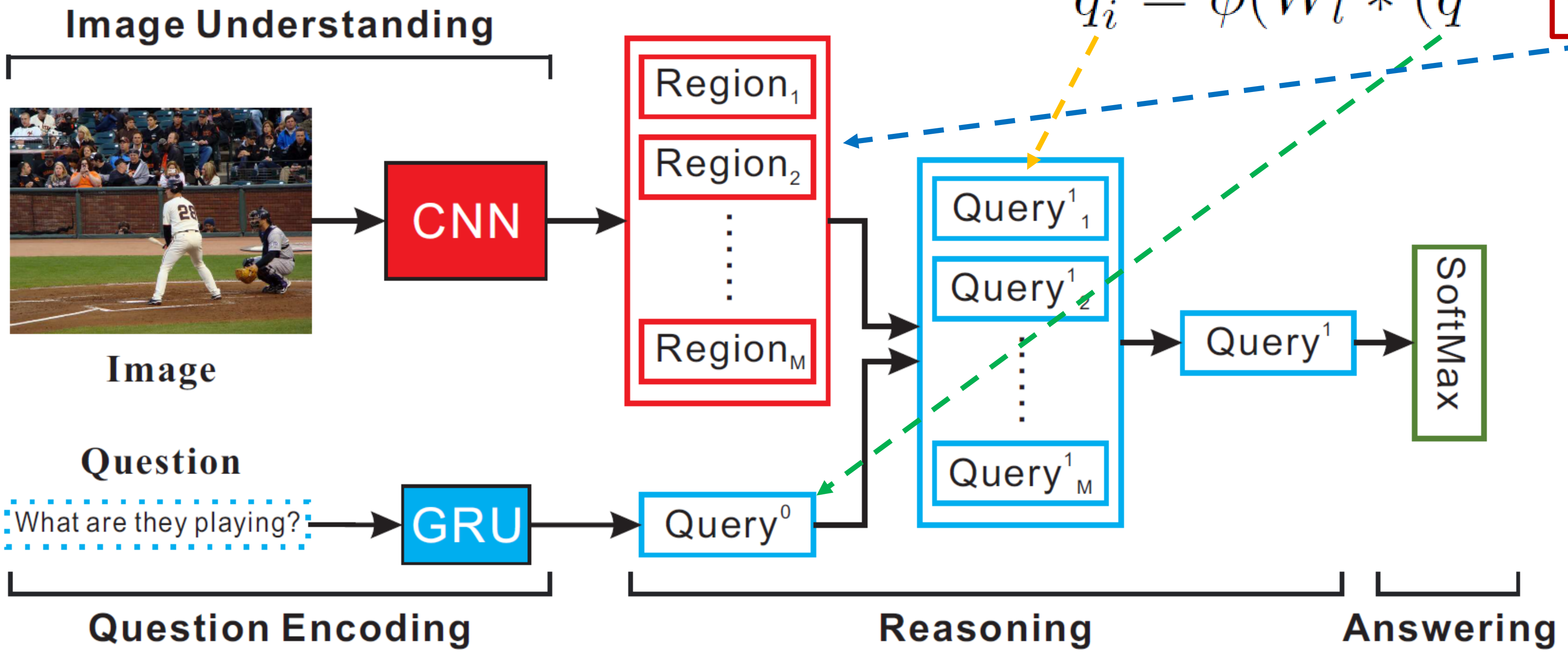


Figure 2: The overall architecture of our model with single reasoning layer for VQA.

[Li, R., Jia, J.: Visual question answering with question representation update (qru). In: Advances in NIPS. pp. 4655–4663 (2016)

Classification based Methods

- Second order pooling -- the outer product of image and question feature vectors

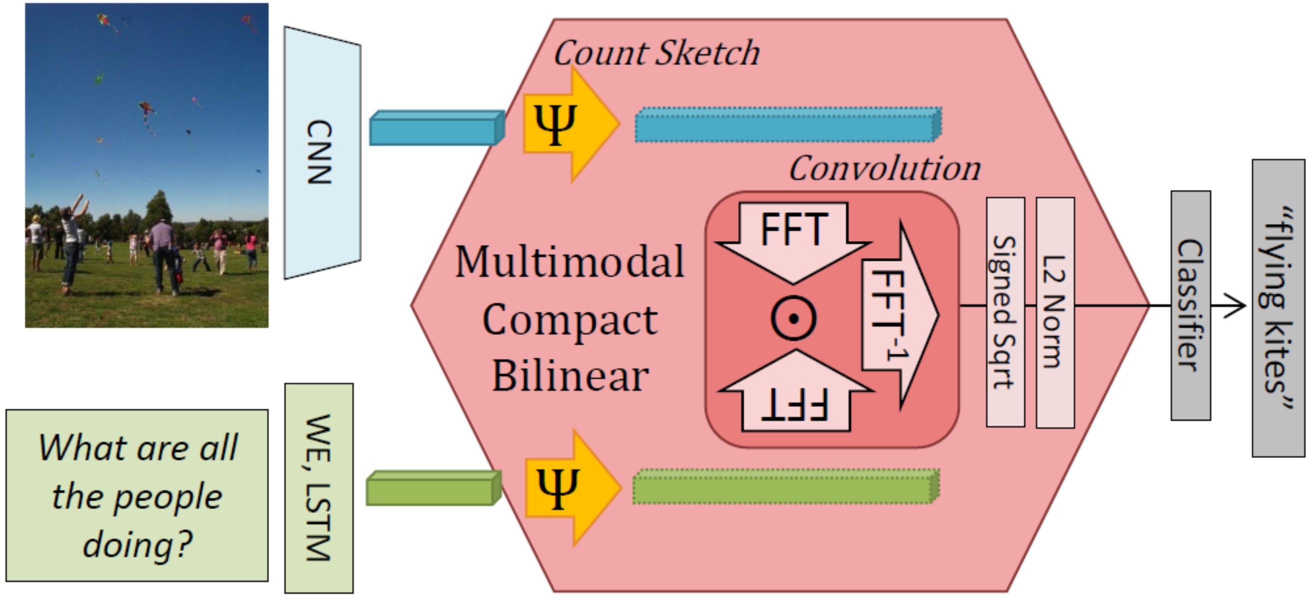


Figure 1: Multimodal Compact Bilinear Pooling for visual question answering.

[Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. 2016]

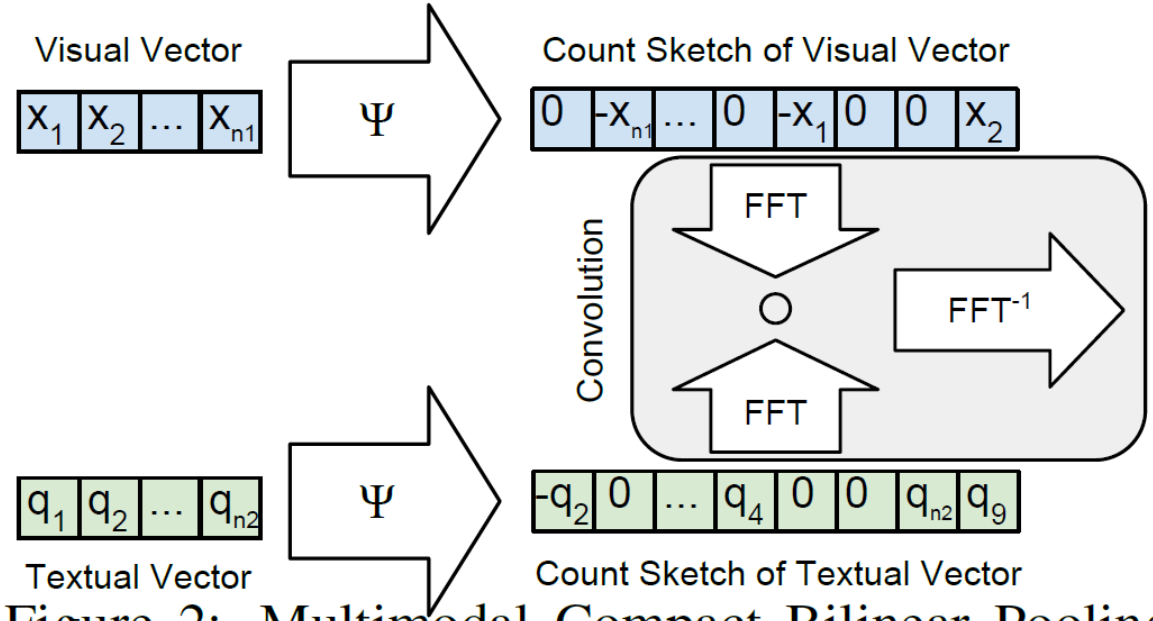


Figure 2: Multimodal Compact Bilinear Pooling (MCB)

$$\Psi(x \otimes q, h, s) = \Psi(x, h, s) * \Psi(q, h, s)$$

Algorithm 1 Multimodal Compact Bilinear

input: $v_1 \in \mathbb{R}^{n_1}, v_2 \in \mathbb{R}^{n_2}$
output: $\Phi(v_1, v_2) \in \mathbb{R}^d$
procedure MCB(v_1, v_2, n_1, n_2, d)
  for $k \leftarrow 1 \dots 2$ **do**
    if h_k, s_k not initialized **then**
      for $i \leftarrow 1 \dots n_k$ **do**
        sample $h_k[i]$ from $\{1, \dots, d\}$
        sample $s_k[i]$ from $\{-1, 1\}$
    $v'_k = \Psi(v_k, h_k, s_k, n_k)$
  $\Phi = \text{FFT}^{-1}(\text{FFT}(v'_1) \odot \text{FFT}(v'_2))$
  return Φ
procedure $\Psi(v, h, s, n)$
  $y = [0, \dots, 0]$
  for $i \leftarrow 1 \dots n$ **do**
    $y[h[i]] = y[h[i]] + s[i] \cdot v[i]$
  return y

$s \in \{-1, 1\}^n$

$h \in \{1, \dots, d\}^n$

Count Sketch projection function

Classification based Methods

- Second order pooling -- Hadamard product (MLB)

Inputs are a question embedding vector $\mathbf{q}$ and a set of visual feature vectors $\mathbf{F}$ over $S \times S$ lattice space. Attended visual feature $\hat{\mathbf{v}}$ is a linear combination of $\mathbf{F}_i$.

$$p(a|\mathbf{q}, \mathbf{F}; \Theta) = \text{softmax}\left(\mathbf{P}_o^T\left(\sigma(\mathbf{W}_\mathbf{q}^T \mathbf{q}) \boxed{\circ} \sigma(\mathbf{V}_{\hat{\mathbf{v}}}^T \hat{\mathbf{v}})\right)\right)$$

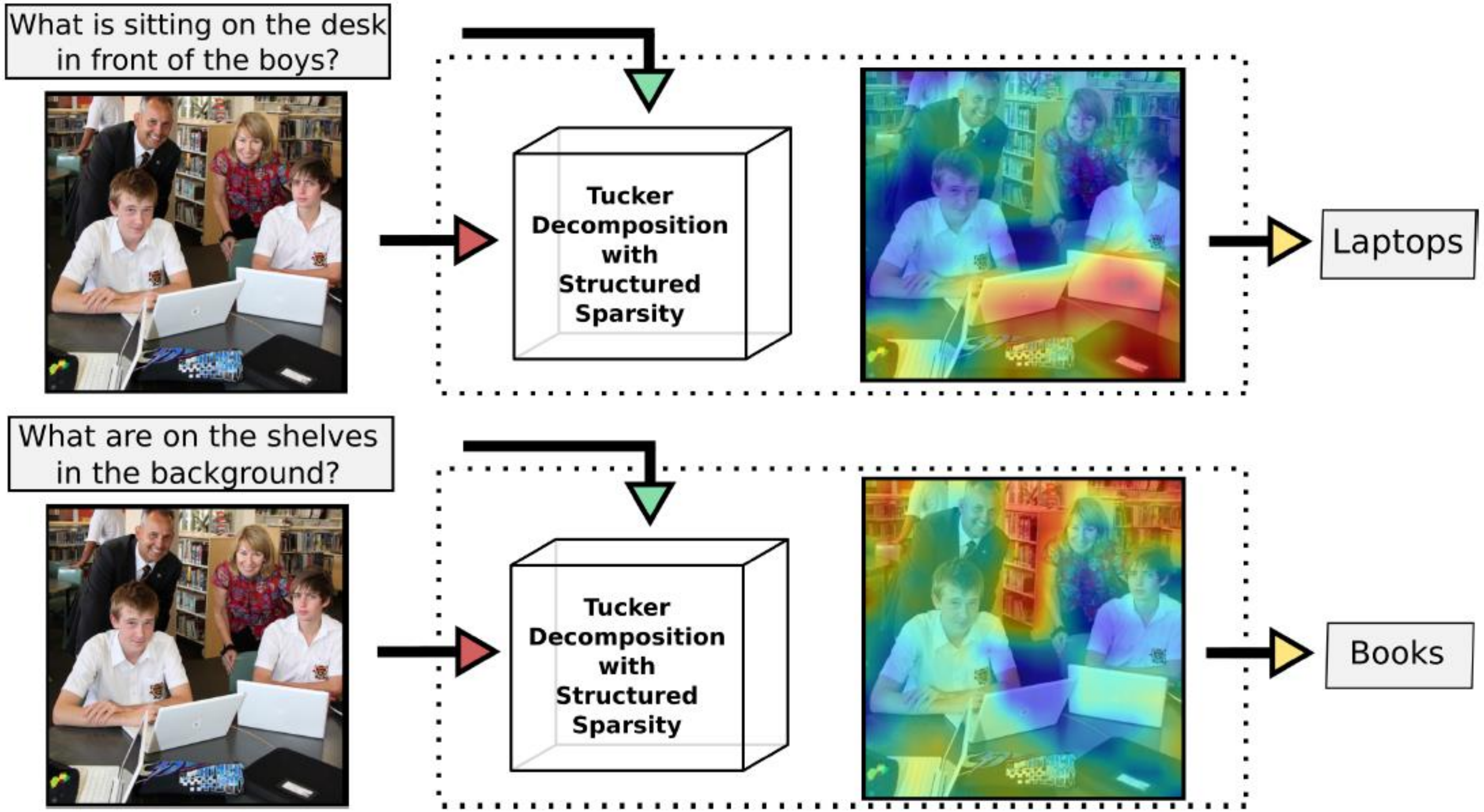
$$\hat{a} = \arg\max_{a \in \Omega} p(a|\mathbf{q}, \mathbf{F}; \Theta)$$

where $\hat{a}$ denotes a predicted answer, Ω is a set of candidate answers and Θ is an aggregation of entire model parameters.

[Kim, J.H., On, K.W., Lim,W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. (ICLR 2017)]

Classification based Methods

- Relying on a low-rank Tucker tensor-based decomposition (MUTAN)



[Ben-younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. ICCV. 2017]

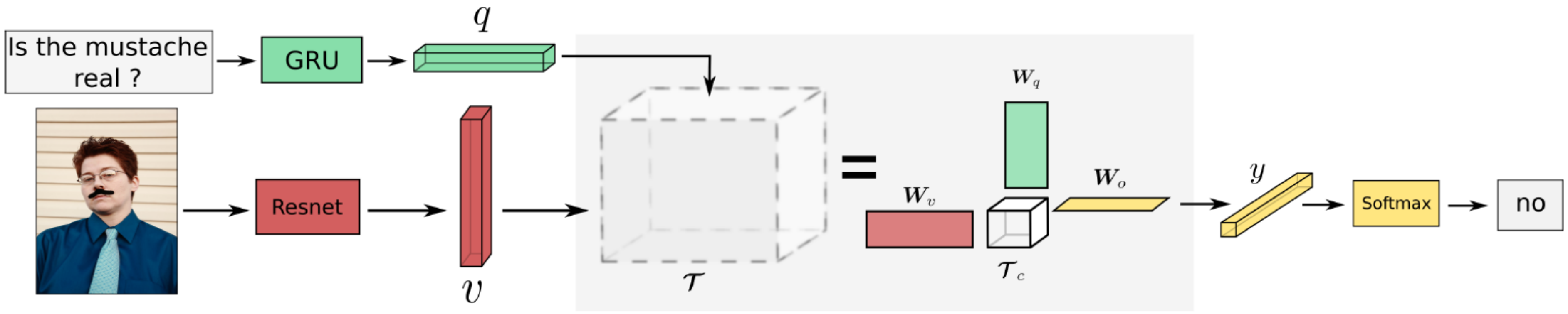


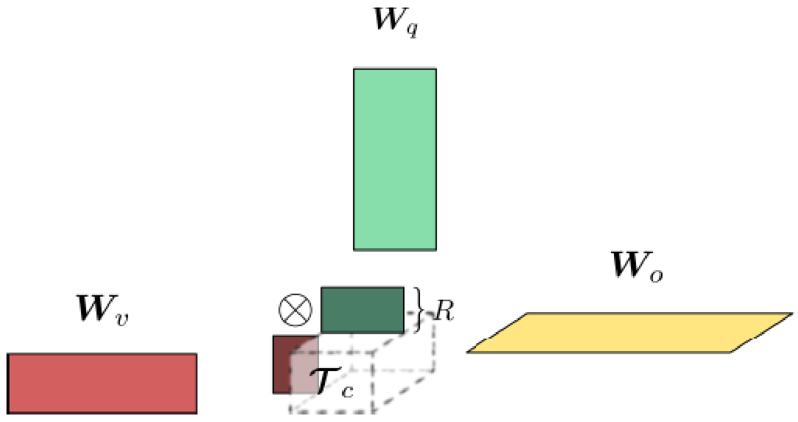
Figure 2: MUTAN fusion scheme for global Visual QA. The prediction is modeled as a bilinear interaction between visual and linguistic features, parametrized by the tensor $\mathcal{T}$. In MUTAN, we factorise the tensor $\mathcal{T}$ using a Tucker decomposition, resulting in an architecture with three intra-modal matrices $\boldsymbol{W}_q$, $\boldsymbol{W}_v$ and $\boldsymbol{W}_o$, and a smaller tensor $\mathcal{T}_c$. The complexity of $\mathcal{T}_c$ is controlled *via* a structured sparsity constraint on the slice matrices of the tensor.

Tucker decomposition:

$$\mathcal{T} = ((\mathcal{T}_c \times_1 \boldsymbol{W}_q) \times_2 \boldsymbol{W}_v) \times_3 \boldsymbol{W}_o$$

Multimodal Tucker Fusion:

$$y = ((\mathcal{T}_c \times_1 (\mathbf{q}^\top \boldsymbol{W}_q)) \times_2 (\mathbf{v}^\top \boldsymbol{W}_v)) \times_3 \boldsymbol{W}_o$$



(c) MUTAN

Outline

- Review
  - Classification based Methods
  - **Image-question-answer Triplet based Reasoning**
- Deep Attention Neural Tensor Network for Visual Question Answering
- Conclusion

Image-question-answer Triplet based Reasoning

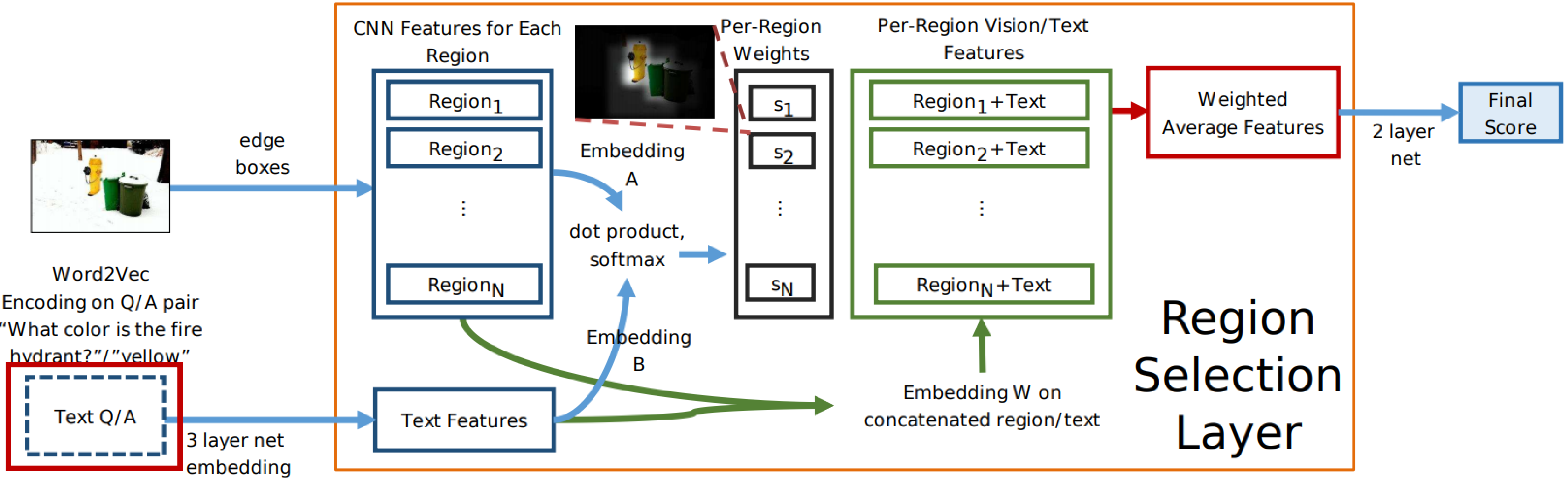


Figure 3. Overview of our network for the example question-answer pairing: "What color is the fire hydrant? Yellow." Question and answer representations are concatenated, fed through the network, then combined with selectively weighted image region features to produce a score.

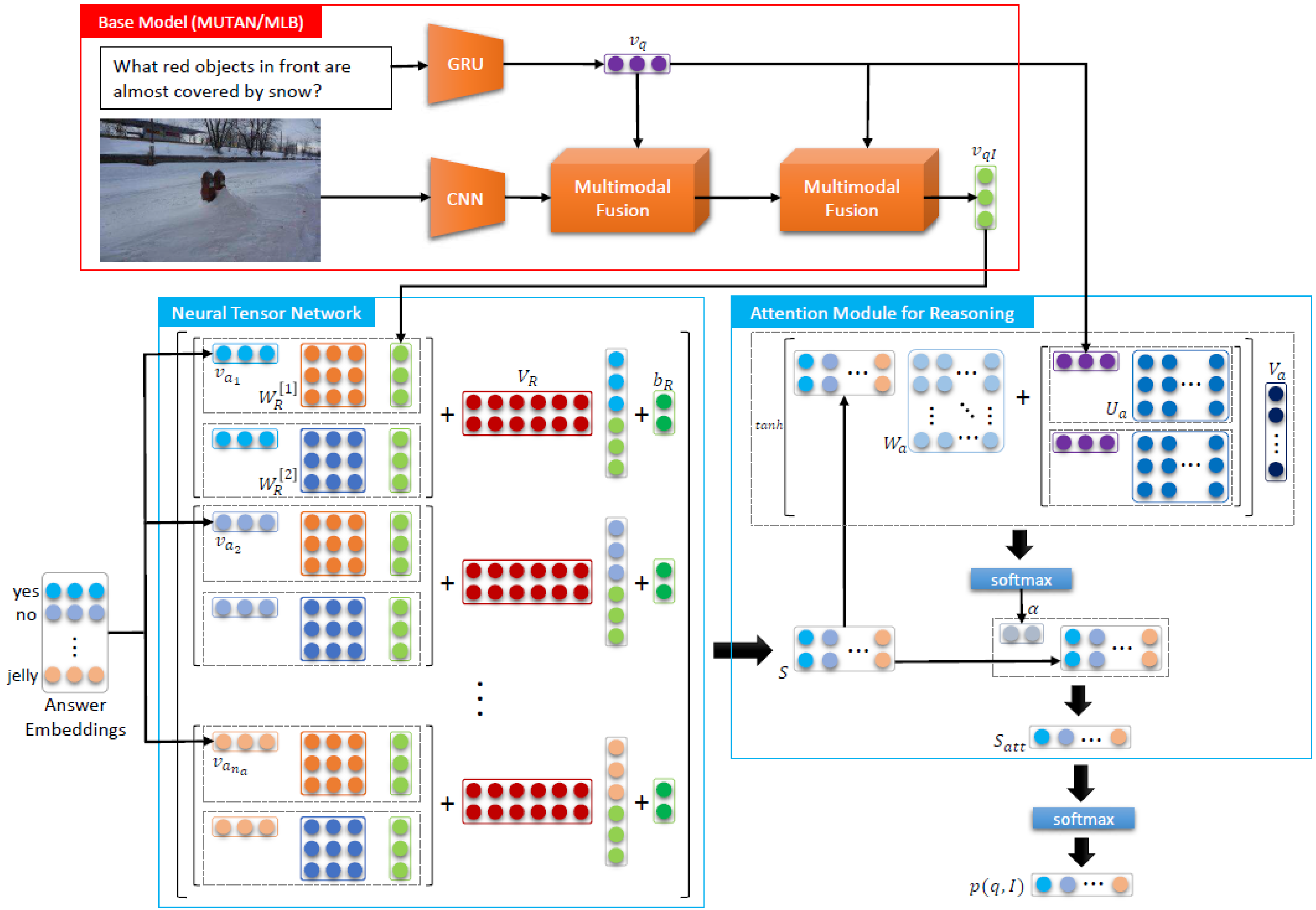
[Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. (CVPR 2016)]

Outline

- Review
  - Classification based Methods
  - Image-question-answer Triplet based Reasoning
- **Deep Attention Neural Tensor Network for Visual Question Answering**
- Conclusion

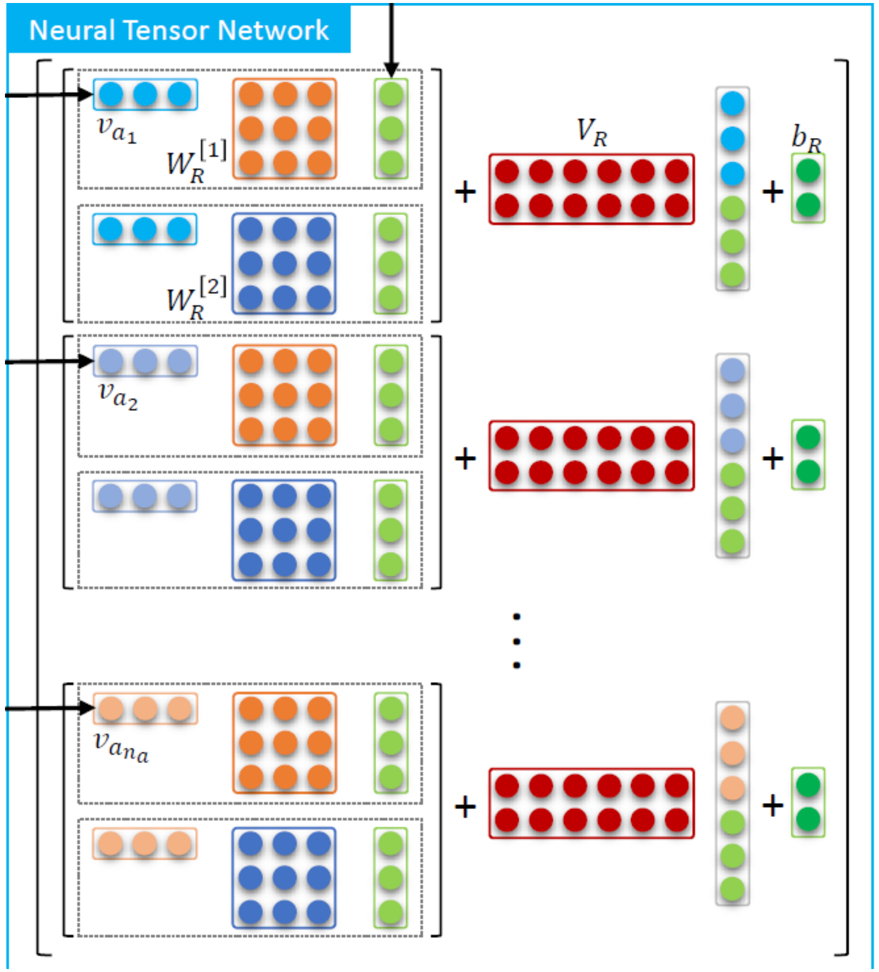
Deep Attention Neural Tensor Network for Visual Question Answering

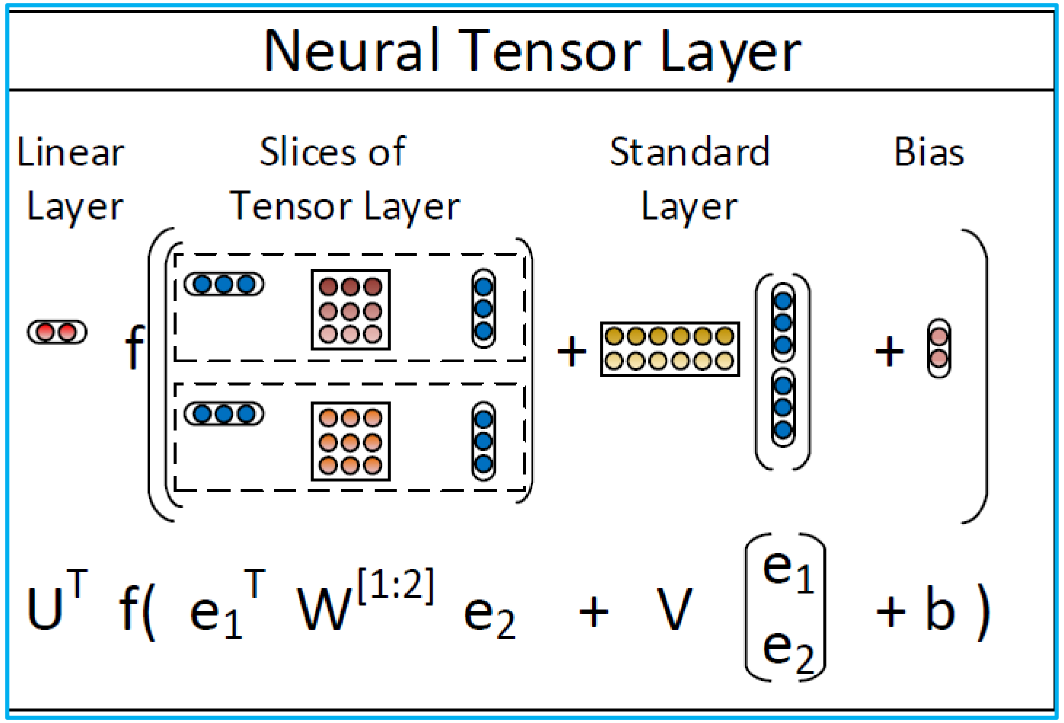
SMILE LAB Statistical Machine Intelligence & Learning



Neural Tensor Networks for VQA

$$s(q, I, a_i) = v_{qI} W_R^{[1:k]} v_{a_i} + V_R \begin{bmatrix} v_{qI} \\ v_{a_i} \end{bmatrix} + b_R$$





[Socher, R., Chen, D., Manning, C. D., & Ng, A. Reasoning with neural tensor networks for knowledge base completion. (NIPS 2013)]

Attention Module for Reasoning

For VQA task,

- the relationship of $\langle q, I, a_i \rangle$ triplet be decided by the type of question q.
- the responses of all candidate answers can provide more detail information about the question type.

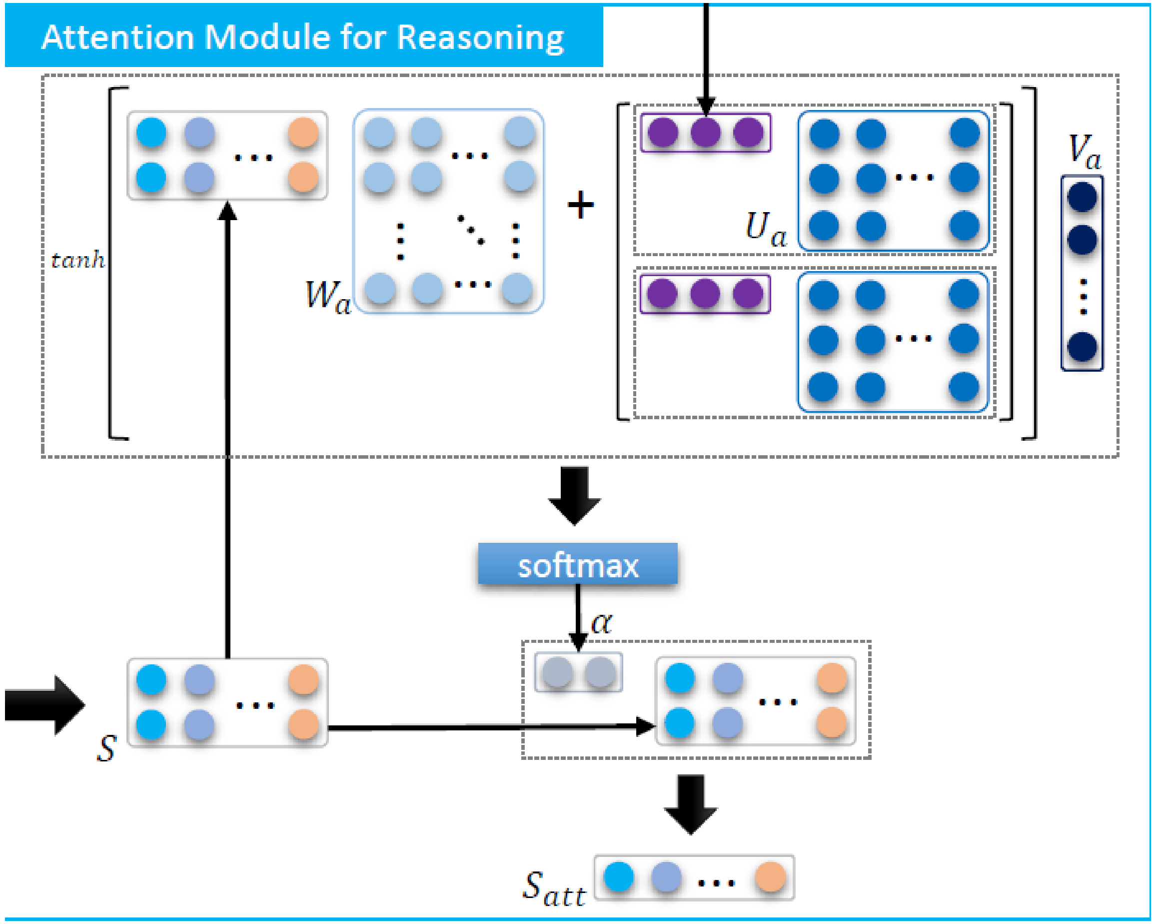
The output of the finally score

$$s_{att}(q, I, a_i) = \sum_{j=1}^{k} s_{i,j} \alpha_j$$

The attention score α_j is calculated by

$$\alpha_j = \frac{\exp(c_j)}{\sum_{e=1}^{k} \exp(c_e)}$$

$$c_j = V_a \cdot \tanh(W_a S_j + U_a v_q)$$



Label Distribution Learning with Regression

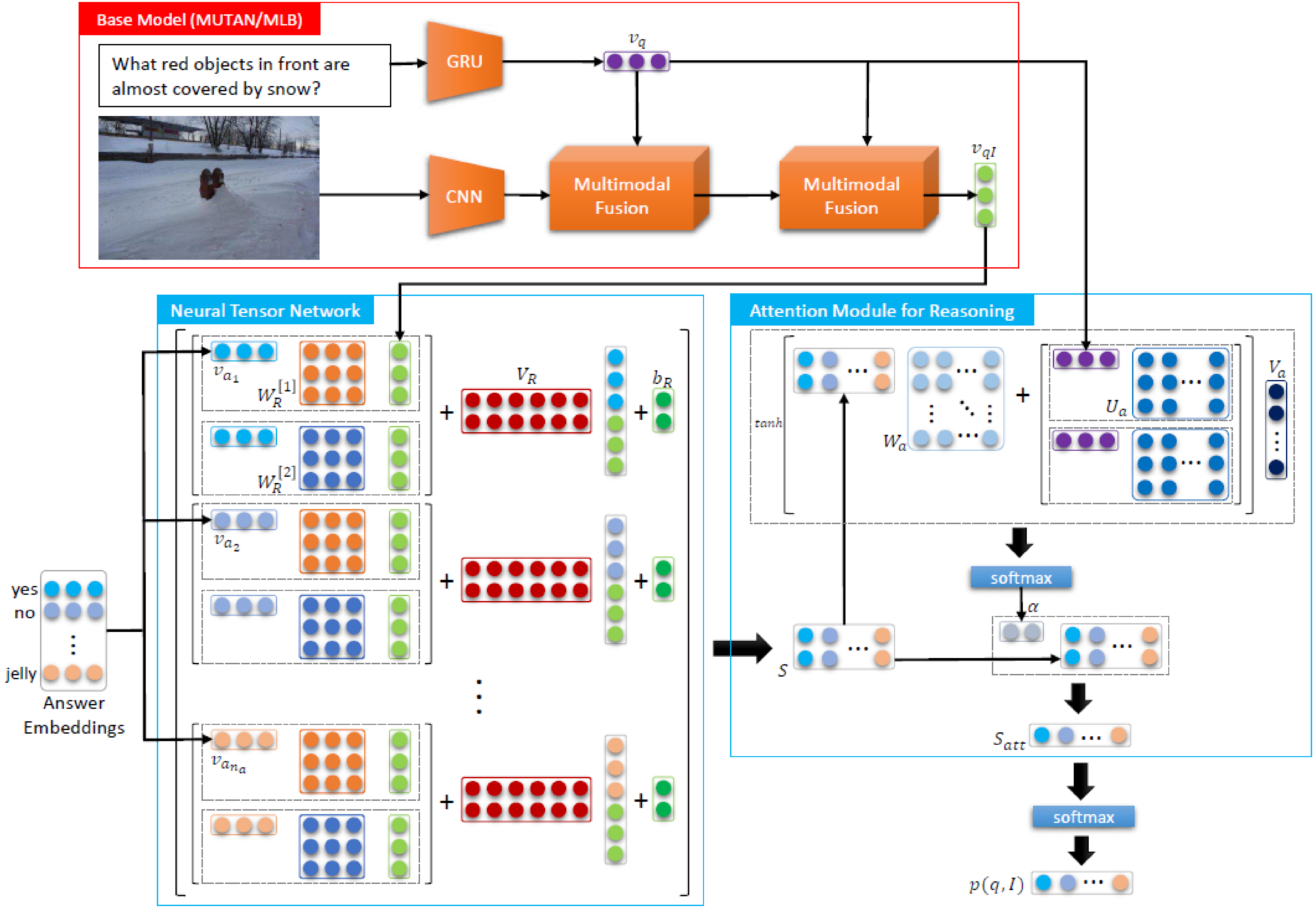
The answers for each sample can be represented as a distribution vector of all the possible answers $y \in \mathbb{R}^{n_a}$, where $y_i \in [0, 1]$ indicates the occurrence probability of the i-the answer in $\mathcal{A}$ across human labeled answers for this image-question pair.

For each image-question pair, we compute the regression score $s_{att}(q, I, a_i)$ for each answer a_i in overall answer candidate set $\mathcal{A}$. Then use a softmax regression to approach the answers distributions:

$$p_i(q, I) = \frac{\exp(s_{att}(q, I, a_i))}{\sum_{j=1}^{n_a} \exp(s_{att}(q, I, a_j))}$$

The KL-divergence loss function is applied to penalize the prediction $p_i \in \mathbb{R}^{n_a}$, our model is trained by minimizing

$$l = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{n_a} y_i \log \frac{y_i}{p_i(q_j, I_j)}$$



ILE LAB
Machine Intelligence & Learning

Experiments

➢ Dataset (question-answer pairs)

	Training set	Validation set	Testing set
VQA-1.0	248K	121k	244k
VQA-2.0	440k	214k	

➢ The accuracy of a predicted answer a_i is given by:

$$\min\left(1, \frac{\#\ \text{annotators the provided } a_i}{3}\right)$$

It means that if the predicted answer a_i appears greater than or equal to three times in human labeled answer list, the accuracy is calculated as 1.

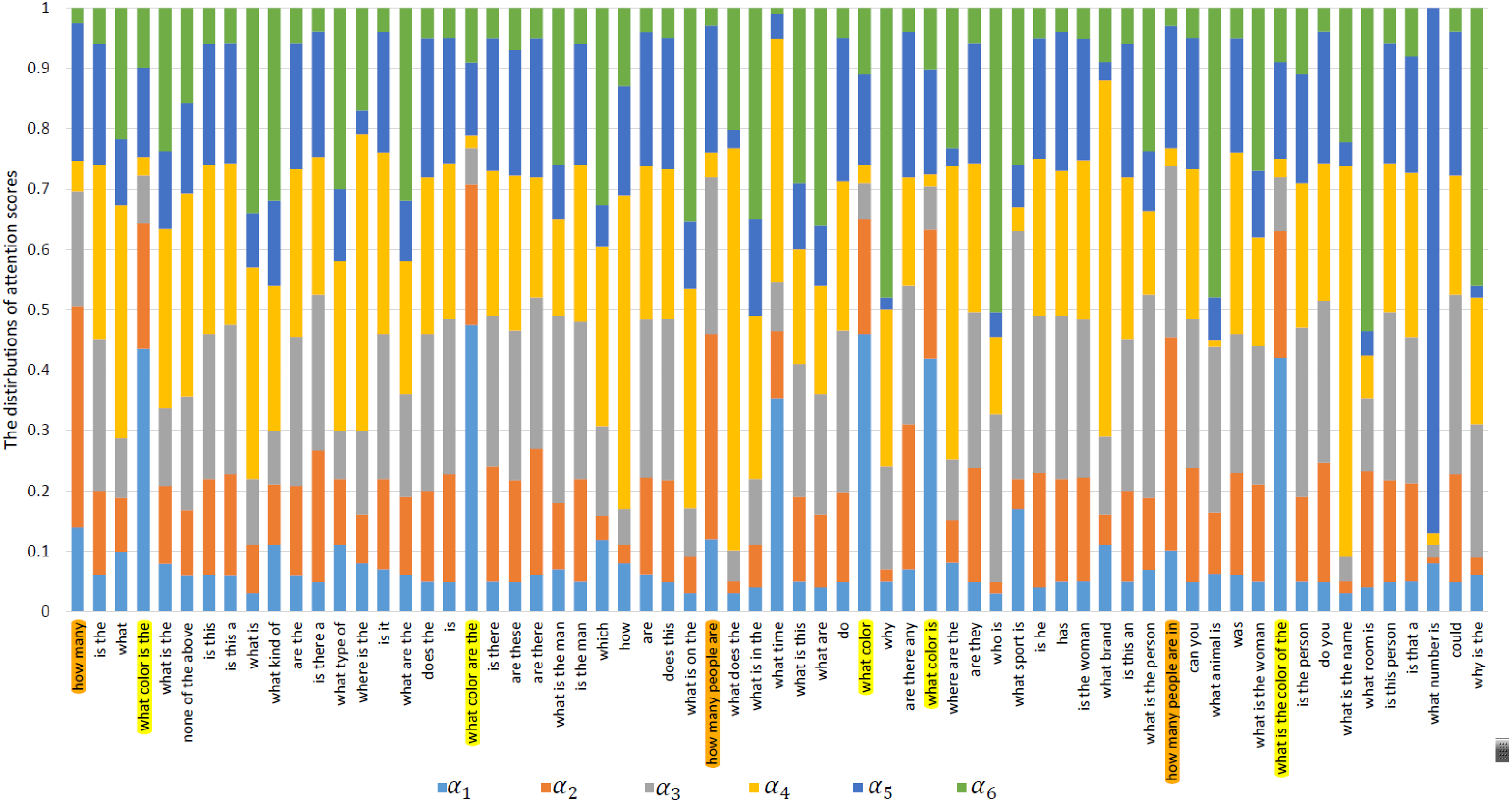
Model	Model Size	VQA-2.0 val set			
		Yes/no	Numb.	Other	All
MUTAN	38.0M	81.09	42.25	54.41	62.84
MUTAN + NTN ($k = 3$)	39.3M	81.69	43.88	55.35	63.74
MUTAN + NTN ($k = 6$)	39.9M	81.96	43.63	55.39	63.83
MUTAN + NTN ($k = 10$)	40.6M	82.23	43.34	55.33	63.86
MUTAN + DA-NTN ($k = 3$)	48.1M	81.96	44.59	55.63	64.07
MUTAN + DA-NTN ($k = 6$)	48.7M	81.98	44.85	55.72	64.16
MUTAN + DA-NTN ($k = 10$)	49.4M	82.24	44.55	55.43	64.07
MLB	67.2M	81.89	42.97	53.89	62.98
MLB + DA-NTN ($k = 6$)	87.5M	83.09	44.88	55.71	**64.58**

Table 2. The performance of different single model for open-ended VQA on the test-dev and test-stand set of VQA-2.0 dataset.

Model	VQA-2.0 Test-dev set				VQA-2.0 Test-standard set			
	Y/N	No.	Other	All	Y/N	No.	Other	All
Prior [10]	-	-	-	-	61.20	0.36	1.17	25.98
LSTM (blind) [10]	-	-	-	-	67.01	31.55	27.37	44.26
MCB [10]	-	-	-	-	78.82	38.28	53.36	62.27
MUTAN	82.88	44.54	56.50	66.01	83.06	44.28	56.91	66.38
MLB	83.58	44.92	56.34	66.27	83.96	44.77	56.52	66.62
MUTAN + DA-NTN	83.58	46.78	57.77	67.15	83.92	46.64	58.0	67.51
MLB + DA-NTN	84.29	47.14	57.92	**67.56**	84.60	47.13	58.20	**67.94**

Attention Module Analysis

SMILE LAB
Statistical Machine Intelligence & Learning



Answer Representations Analysis

Answers	DA-NTN	GloVe
0	1:-0.43, 2:-0.32	1:-0.60, 5:-0.53, 9:-0.51, 6:-0.51, 3:-0.50, 4:-0.50, 8:-0.50, etc.
orange	red:-0.39, yellow:-0.33, brown:-0.32	orange and yellow:-0.90, orange and blue-0.89, orange juice:-0.88, green and orange:-0.87, etc.
table	on table:-0.35, desk:-0.30	on table:-0.84, picnic table:-0.84, chairs:-0.62, dining room:-0.60, etc.
rectangle	square:-0.34	triangle:-0.64, squares:-0.61, circle:-0.60, oval:-0.59, etc.
glove	baseball glove: -0.34, mitt:-0.33	baseball glove:-0.82, gloves:-0.81, knee pads:-0.57, helmet:-0.56, etc.
playing frisbee	catching frisbee:-0.37, throwing frisbee:-0.35	frisbee:-0.81, throwing frisbee:-0.80, playing tennis:-0.80, playing:-0.80, etc.
river	lake:-0.32, pond:-0.32	lake:-0.72, shore:-0.63, railroad crossing:-0.58, bridge:-0.58, water:-0.58, etc.
middle	center:-0.30	end:-0.64, in corner:-0.64, right side:-0.63, left one:-0.63, etc.

Table 4. For query words, we show their most similar words based on our method and context based word embedding [21]. We also show the cosine similarity scores between query word and its nearest neighbors, only the words whose cosine similarity scores are smaller than -0.3 are shown in this table.

Outline

- Review
  - Classification based Methods
  - Image-question-answer Triplet based Reasoning
- Deep Attention Neural Tensor Network for Visual Question Answering
- **Conclusion**

Conclusion

- This paper proposes a novel deep attention neural tensor network (DA-NTN) for visual question answering, which can **discover the joint correlations over images, questions and answers with tensor-based representations**.

  - First, this paper models one of the pairwise interaction (e.g., image and question) by bilinear features, which is further encoded with the third dimension (e.g., answer) to be a triplet by **bilinear tensor product**.
  - Second, this paper decomposes the correlation of different triplets by different answer and question types, and further propose a **slice-wise attention module on tensor** to select the most discriminative reasoning process for inference.

Thanks!