Structured Triplet Learning with POS-tag Guided Attention for VQA

Zhe Wang[1], Xiaoyi Liu[2], Liangjian Chen[1], Limin Wang[4], Yu Qiao[3], Xiaohui Xie[1], Charless Fowlkes[1]
[1] CS UC Irvine, [2]Microsoft, [3]SIAT CAS, [4]CVL ETH

Visual Question Answering

Question: Why was the hand of the woman over the left shoulder of the man?

A: They were together and engaging in affection
A: The woman was trying to get the man's attention
A: The woman was trying to scare the man
A: The woman was holding on to the man for balance

Visual question answering (VQA) tasks are of significant interest due to their potential as a strong test of image understanding systems and in probing the connection between language and vision. Despite much recent innovation, general VQA is far from a solved problem.

Our Approach

We explore four mechanisms for improving VQA performance

(i) POS Tag Guided Attention:
(1) Some words (i.e., nouns, verbs and adjectives) matter more than others
(2) We use seven POS categories (numbers, nouns, adjectives, verbs, wh-pronouns, wh-adverbs, other)

(ii) Convolutional N-Gram:
(1) We use a convolutional n-gram to integrate contextual information across word vectors.
(2) Contextual features for different window sizes are pooled to obtain a new word representation
(3) The final question / answer sentence is represented by an average of word representations $x_Q = \frac{1}{M}\sum_{i=1}^{M} \tilde{e}_i.$

(iii) Triplet Attention:
We derive a spatial attention weight from the question and answer representations.

$\boldsymbol{att}_I = \mathrm{norm}(\lambda \times \boldsymbol{att}_{Q-I} + \boldsymbol{att}_{A_i-I})$ where $\mathrm{norm}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\sum(\boldsymbol{x})}$

We use affinity matrix and max pooling to get both the attention from Question-Image and Answer-Image

(iv) Structured Triplet Learning:
We formulate VQA as a binary classification problem.
For each candidate triplet $\{I, Q, A_i, t_i\}$, where $t_1 = 1$ and $t_i = 0$ for I = 2,…,N,

The output for the i th candidate answer is $p_i = \mathrm{sigmoid}\left(\boldsymbol{W}_{QIA}\boldsymbol{x}_{QIA_i} + \boldsymbol{b}_{QIA}\right)$

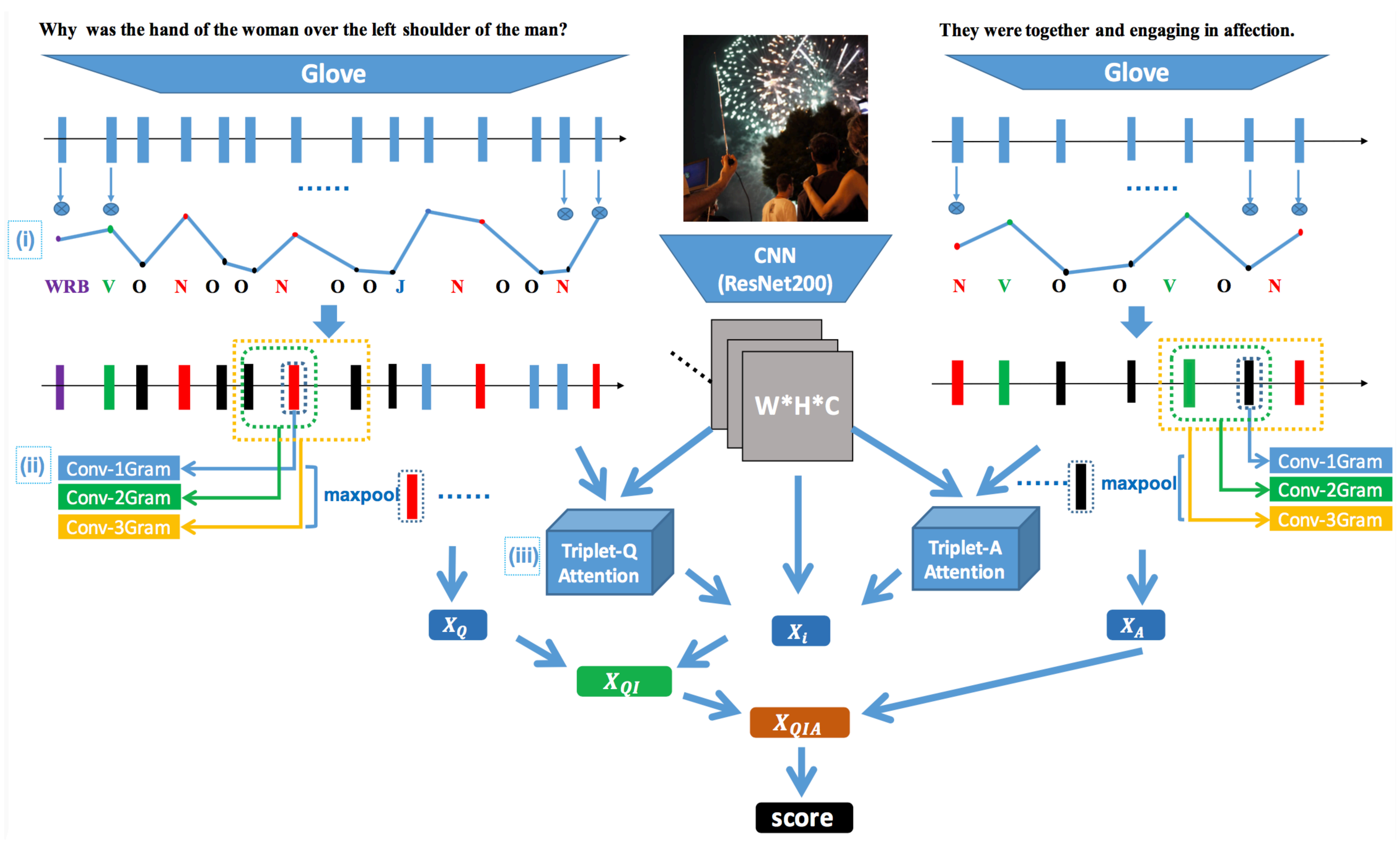
The full loss is $L = L_b + \lambda_2 L_s$ where $L_b = -\sum_{i=1}^{N} t_i \log p_i$

and $L_s = \max_i(max(\mathrm{margin} + p_i - p_1, 0))$

References

[1] Fukui et al. *EMNLP*, 2016.
[3] Jabri et al. *ECCV*, 2016.
[4] Lu et al. *NIPS*, 2016
[5] Teney et al. 2016. arXiv.1611.05546.
[6] Zhu et al. *CVPR*, 2016.
[7] Krishna, et al.. 2016. arXiv.1602.07332.
[8] Gan, et al. ICCV, 2017

Our Pipeline for VQA

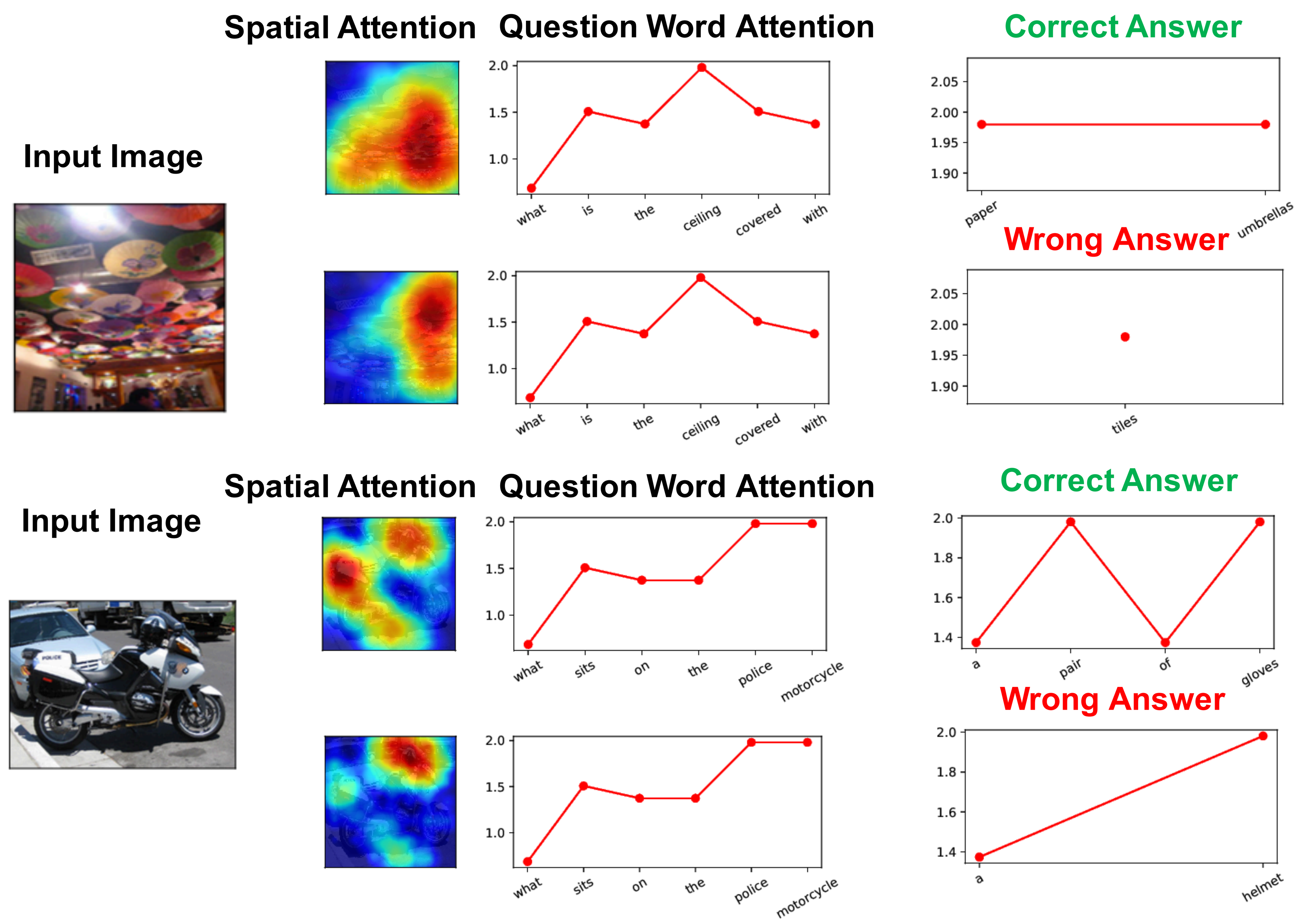


General framework:
(1) Extract vector representations of the image, question and candidate answer using deep neural network.
(2) Score the compatibility using a two layer network:

x_Q Question sentence descriptor
x_{A_i} Answer descriptor for ith answer
x_I Image descriptor
$\odot$ Hadamard product(element-wise multiplication)

$$\boldsymbol{x}_{QI} = \boldsymbol{x}_Q \odot \boldsymbol{x}_I.$$
$$\boldsymbol{x}_{QIA_i} = \mathrm{tanh}(\boldsymbol{W}_{QI}\boldsymbol{x}_{QI} + \boldsymbol{b}_{QI}) \odot \boldsymbol{x}_{A_i}.$$
$$p_i = \mathrm{sigmoid}\left(\boldsymbol{W}_{QIA}\boldsymbol{x}_{QIA_i} + \boldsymbol{b}_{QIA}\right).$$
$$i^* = \arg\max_{i=1,\dots,N} p_i.$$

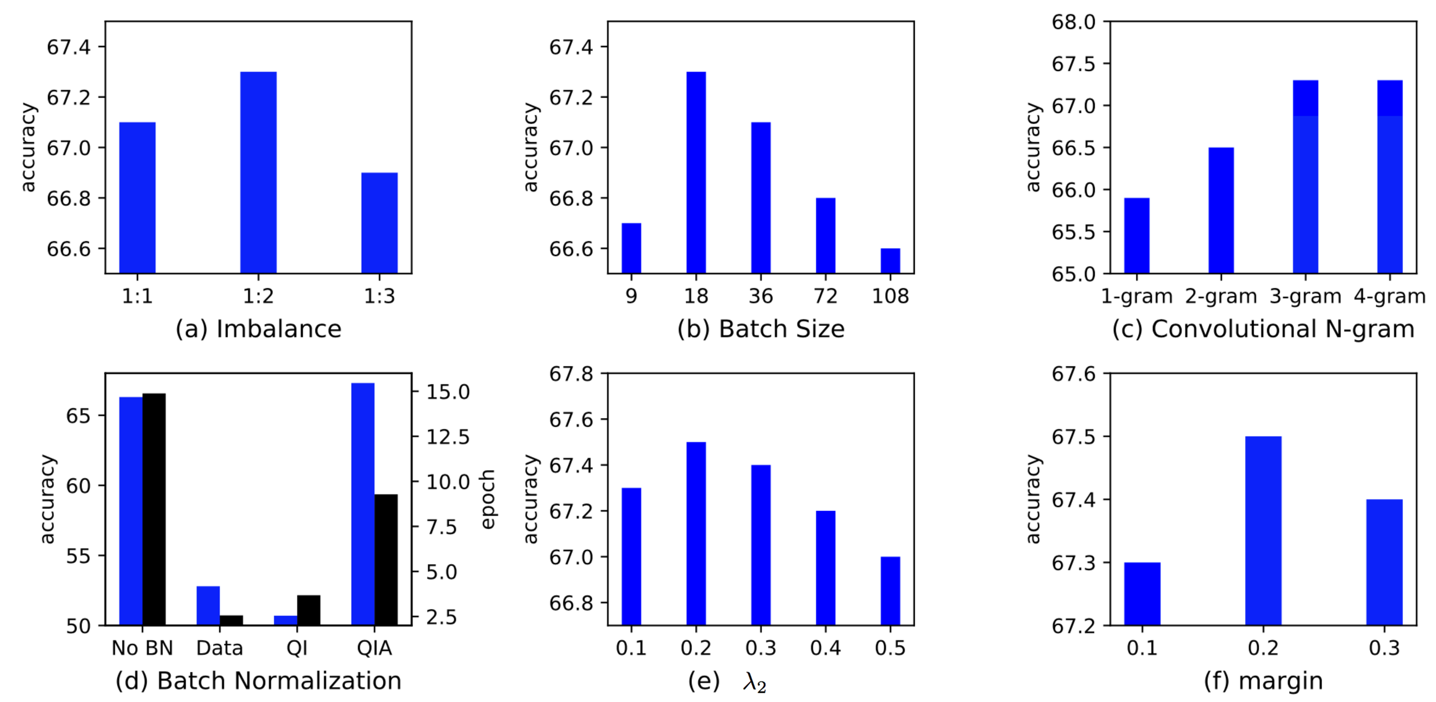
Visualizing Attention



Results & Analysis

Method	V7W	VQA
Our Baseline	65.6	58.3
+POS tag guided attention(POS-Att)	66.3	58.7
+Convolutional N-Gram (Conv N-Gram)	66.2	59.3
+POS-Att +Conv N-Gram	66.6	59.5
+POS-Att +Conv N-Gram +Triplet attention-Q	66.8	60.1
+POS-Att +Conv N-Gram +Triplet attention-A	67.0	60.1
+POS-Att +Conv N-Gram +Triplet attention-Q+A	67.3	60.2
+POS-Att +Conv N-Gram +Triplet attention-Q+A+ Structured Triplet Learning	67.5	60.3

Verification of our proposed (1) POS tag guided attention, (2) Conv N-Gram (3) Triplet Attention and (4) Structured Triplet Learning step by step. Integrating them all further improves the performance on Visual7W and VQA validation set. Notes: the feature is 7*7 on spatial resolution.



Exploration of good practice (a) Handling data imbalance (b) adjusting batch size (c) parameter to adjust convolutional n-gram and (d) where to add batch normalization (e) find the optimal λ_2 (f) find the optimal margin.

Method	Visual 7W	VQA Test Standard	VQA Test Dev
Co-Attention [4]	-	66.1	65.8
Attention-LSTM [6]	55.6	-	-
MCB + Att [1]	62.2	-	68.6
Zero-shot [5]	65.7	-	-
MLP [3]	67.1	68.9	65.2
VQS[8]	-	-	68.9
Full model (14*14 Resnet feature)	68.2	69.6	69.7

Benchmark comparison with previous work:
(1) We outperform the state-of-the-art performance on Visual7w and get competitive performance on VQA.
(2) Use POS-tagging to guide word attention, making pooled sentence vectors more meaningful and effective.
(3) Utilize hard-negative mining and the relationship among multiple answers corresponding to the same image-question pair during training to improve the system.