Towards Good Practice for Visual Question Answering

Zhe Wang[1], Xiaoyi Liu[2], Liangjian Chen[1], Limin Wang[4], Yu Qiao[3], Xiaohui Xie[1], Charless Fowlkes[1]

[1] ICS UC Irvine, [2]EECS UC Irvine, [3]SIAT CAS, [4]EE 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 three mechanisms for improving VQA performance

(i) POS Tag Guided Attention:

(1) Some words (i.e., nouns, verbs and adjectives) should matter more than others (e.g., the conjunctions)

(2) We use a small set of seven POS categories (numbers, nouns, adjectives, verbs, wh-pronouns, wh-adverbs, other)

(ii) Convolutional N-Gram:

We propose using a convolutional n-gram to combine contextual information over multiple words represented as vectors.

Contextual features for different window sizes are pooled to obtain a new word representation: $\tilde{e}_i = \texttt{maxpool}(F_L, F_{L-1}, \ldots, F_1)$.

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 an spatial attention weight from the question and answer representations.

$att_I = \texttt{norm}(\lambda \times att_{Q-I} + att_{A_i-I})$ where $\texttt{norm}(x) = \frac{x}{\sum(x)}$, att_{Q-I} att_{A_i-I} are attention weights from questions/answers to images.

Given image features $X_I = \texttt{relu}(W_I X_{I,\text{raw}} + b_I)$

and question features $X_Q = [\tilde{e}_1, \ldots, \tilde{e}_M]$

We compute an affinity matrix $A = \texttt{softmax}(X_Q^T \times X_I)$

and a Question-Image attention vector $att_{Q-I} = \texttt{maxpool}(A)$

References

[1] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[3] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016.

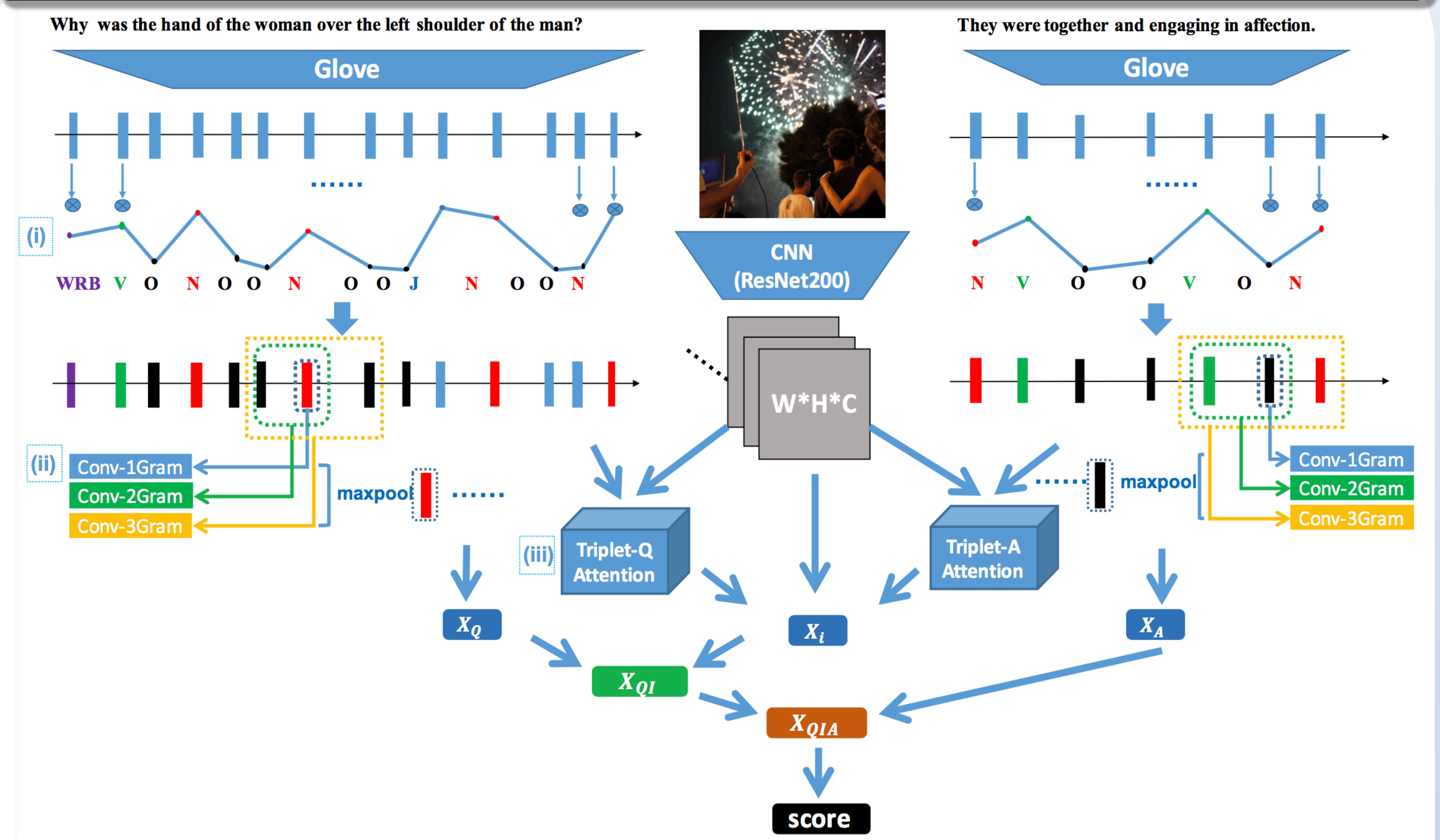
[4] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016

[5] Damien Teney and Anton van den Hengel. Zero-shot visual question answering. 2016. arXiv.1611.05546.

[6] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li FeiFei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016.

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S.Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. arXiv.1602.07332.

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 Images descriptor

$\odot$ Hadamard product(element-wise multiplication)

$$x_{QI} = x_Q \odot x_I.$$

$$x_{QIA_i} = \tanh(W_{QI} x_{QI} + b_{QI}) \odot x_{A_i}.$$

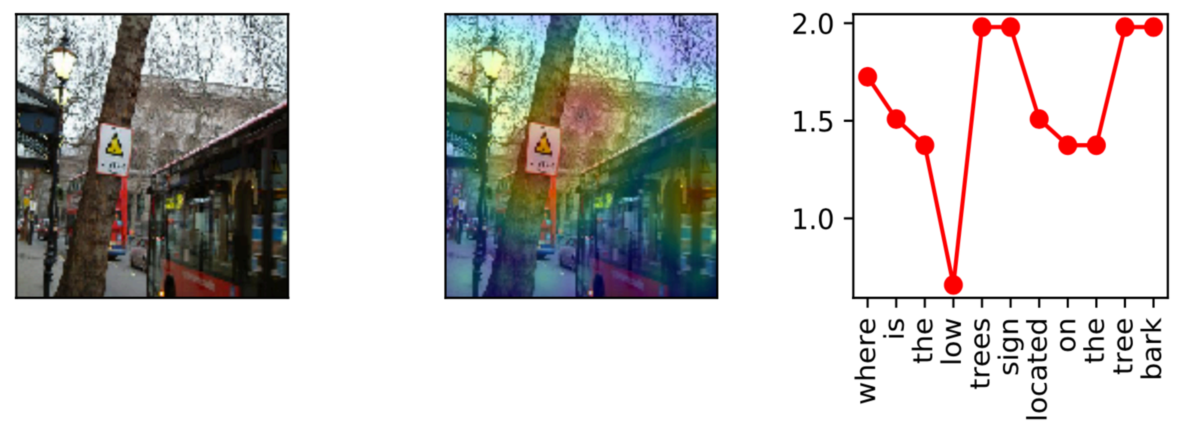
$$p_i = \texttt{sigmoid}(W_{QIA} x_{QIA_i} + b_{QIA}).$$

$$i^* = \arg\max_{i=1,\ldots,N} p_i.$$

Visualizing Attention

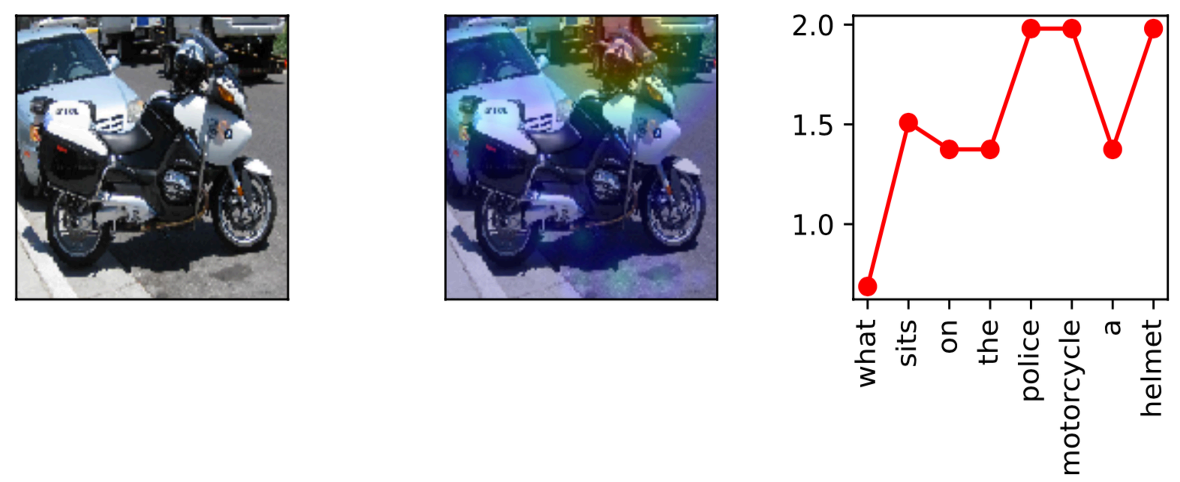
Question: Where is the "low trees" sign located?

A: On the tree bank



Question: What sits on the police motor cycle?

A: A Helmet



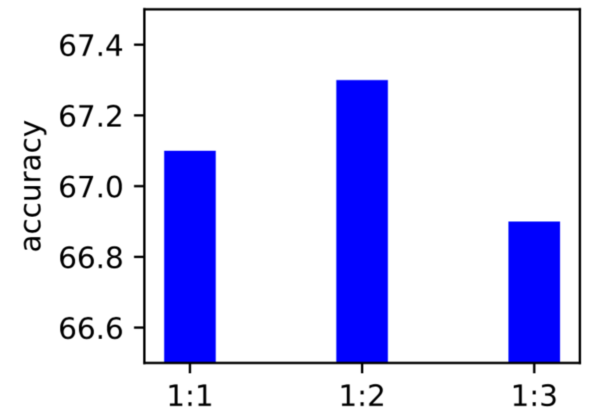
Input Image | Triplet attention | POS-tag guided attention

Triplet attention focuses on image region (e.g. sign on the tree bark) and relevant key words in the question and answer are assigned high weights

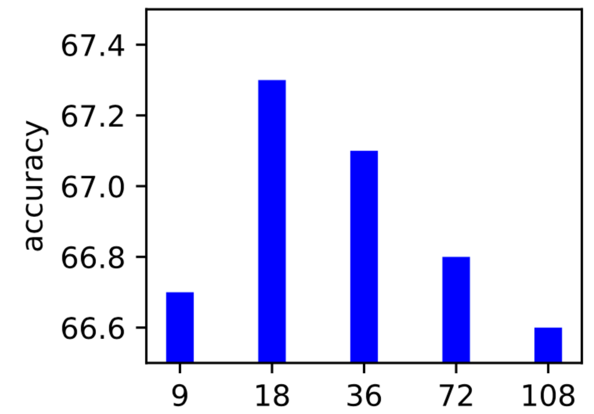
Results & Analysis

Method	Visual 7W
Our Baseline	65.6
+POS tag guided attention(POS-Att)	66.3
+Convolutional N-Gram (Conv N-Gram)	66.2
+POS-Att +Conv N-Gram	66.6
+POS-Att +Conv N-Gram +Triplet attention-Q	66.8
+POS-Att +Conv N-Gram +Triplet attention-A	67.0
+POS-Att +Conv N-Gram +Triplet attention-Q+A (Full model)	67.3

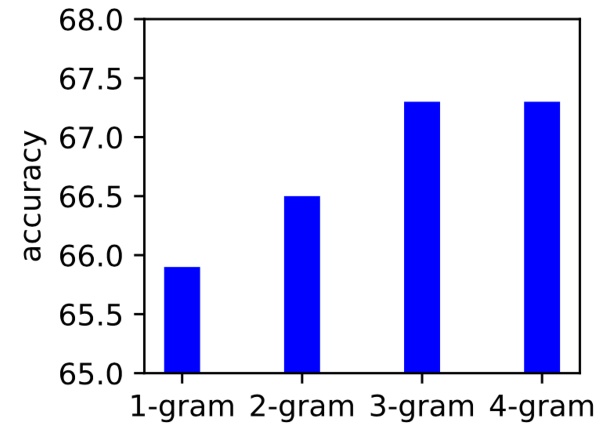
Verification of our proposed (1) POS tag guided attention, (2) Conv N-Gram and (3) Triplet Attention step by step. Integrating them all further improves the performance.



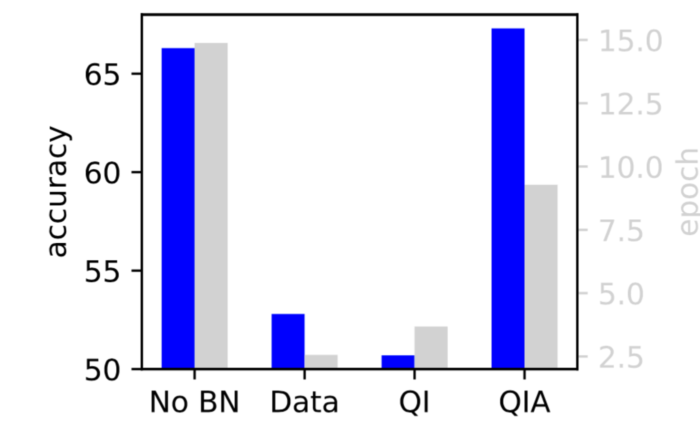
(a) Imbalance



(b) Batch Size



(c) Convolutional N-gram



(d) Batch Normalization

Exploration of good practice (1) Handling data imbalance (2) adjusting batch size (3) parameter to adjust convolutional n-gram and (4) where to add batch normalization [2].

Method	Visual 7W Telling	VQA Real Multi Choice
Co-Attention [4]	-	66.1
Attention-LSTM [6]	55.6	-
MCB [1]	-	65.4
MCB + Att [1]	62.2	-
Ensemble of 7 Att models [1]	-	70.1
Zero-shot [5]	65.7	-
MLP [3]	64.8	65.2
Full model (7*7 Resnet feature)	67.3	68.3
Full model (14*14 Resnet feature)	68.1	-

In comparison with the related work:

(1) We outperform the state-of-the-art performance on visual7w and get competitive performance on VQA.

(2) Recent state-of-the-art work in [1] used an ensemble of 7 models and trained with additional data (the Visual Genome dataset [7]), performing slightly better than our model on VQA suggesting our simpler model is still quite competitive in terms of computation trade-off.