Structured Triplet Learning with POS-tag Guided Attention 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] CS UC Irvine, [2]Microsoft, [3]SIAT CAS, [4]CVL ETH

Multiple Choice Visual Question Answering (VQA)

Q: What endangered animal is featured on the truck?

A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?

A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?

A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Y Zhu, O, Groth, M. Bernstein, Fei-Fei Li. "Visual7W: Grounded Question Answering in Images", CVPR, 2016

Our contributions

The (spatial) attention only depends on the image-question pair input

- We consider the image-answer interactions when computing attention

Our contributions

The (spatial) attention only depends on the image-question pair input

- We consider the image-answer interactions when computing attention

The sentence representation is limited: either LSTM encoding or simple average of word vectors

- We propose to integrate Part-of-speech tags and convolutional n-gram processing to better encode query and answer sentences.

Our contributions

The (spatial) attention only depends on the image-question pair input

- We consider the image-answer interactions when computing attention

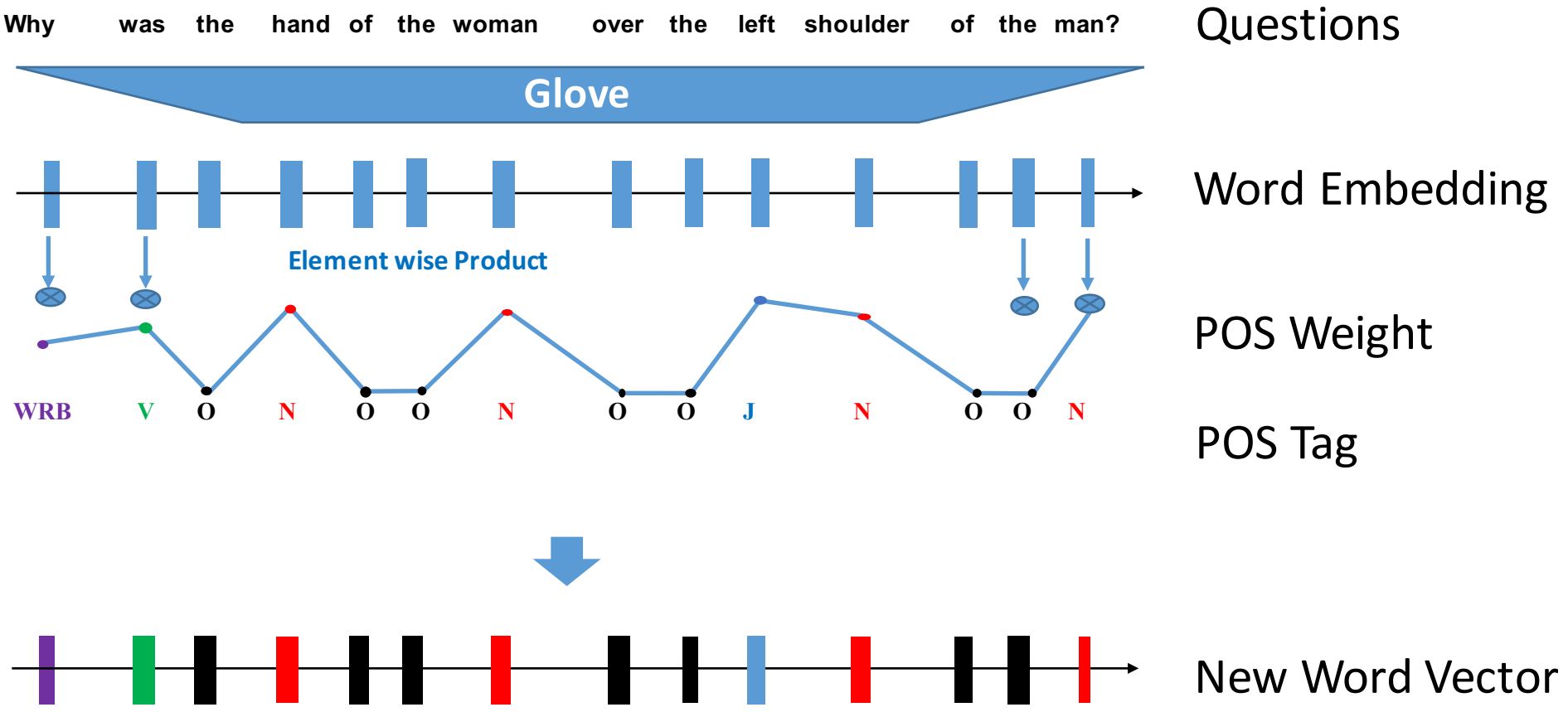
The sentence representation is limited: either LSTM encoding or simple average of word vectors

- We propose to integrate Part-of-speech tags and convolutional n-gram processing to better encode query and answer sentences.

Image-question-answer triplets corresponding to the same image-question pair are treated independently

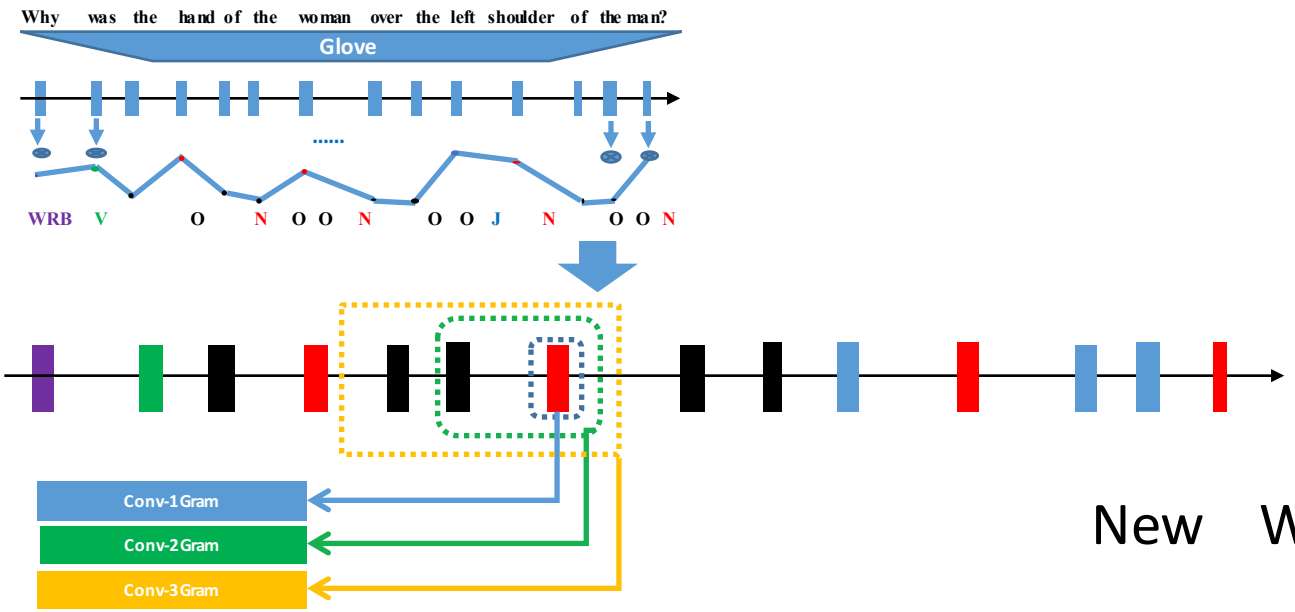
- We introduce structured triplet learning and mine “hard negative” triplets to improve the system

Part-of-speech-tag (POS) guided attention



Convolutional N-Gram

Convolutional filtering of word vectors encodes local sentence context



Word Vector

New Words/Phrases/Sentences

Structured Triplet Learning:

For a given question, the correct answer should score higher than incorrect (competing) answers by a specified margin

p_i = **Score[** Question**,** Image, Answer(i) **]**

ti = ground truth. {0,1}

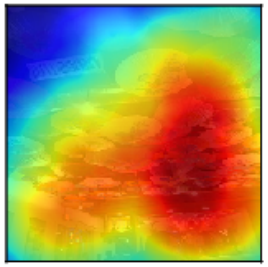
$$L_b = -\sum_{i=1}^{N} t_i \log p_i$$ **Logistic loss**

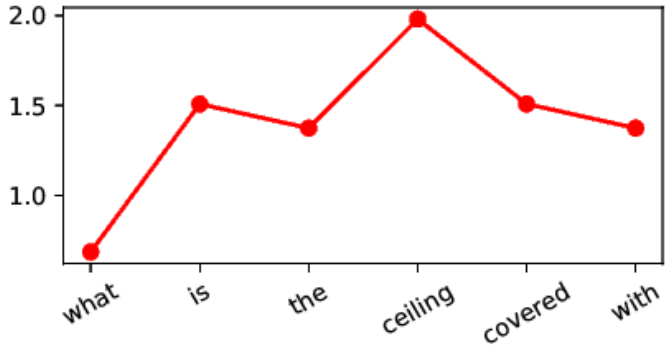
$$L_s = \max_i(max(\text{margin} + p_i - p_1, 0))$$ **Structured loss**

Visualizing Attention Maps

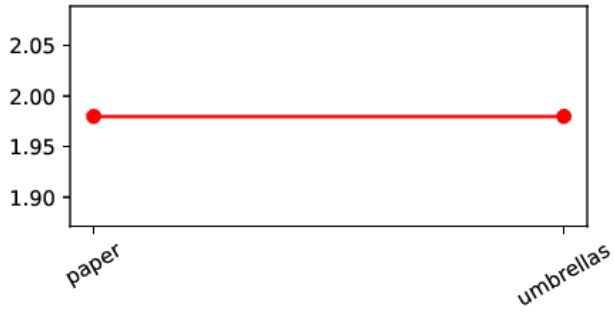
What is the ceiling covered with? Paper Umbrellas/ Tiles

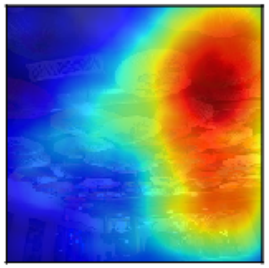
Input Image

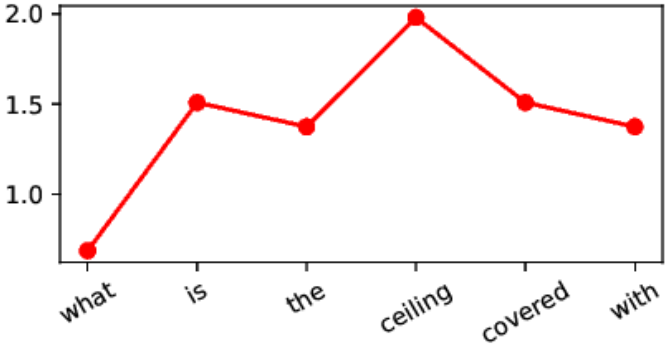




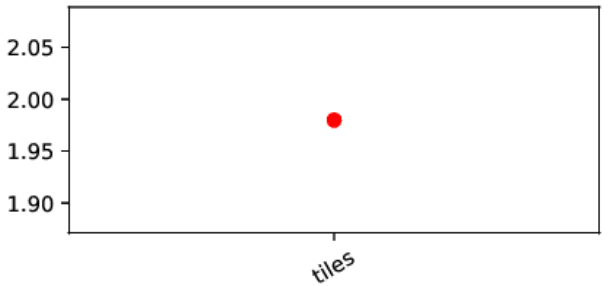
Correct Answer







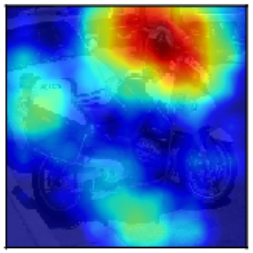
Wrong Answer

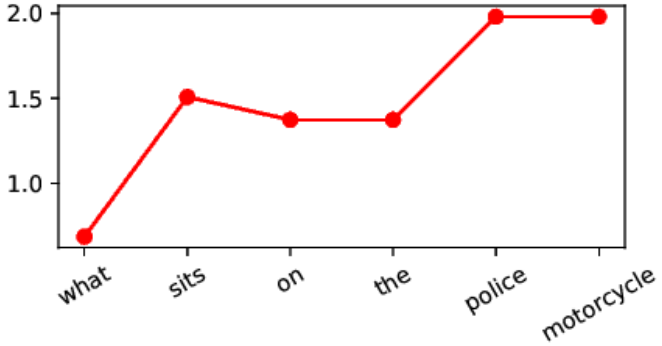


Question Word Attention

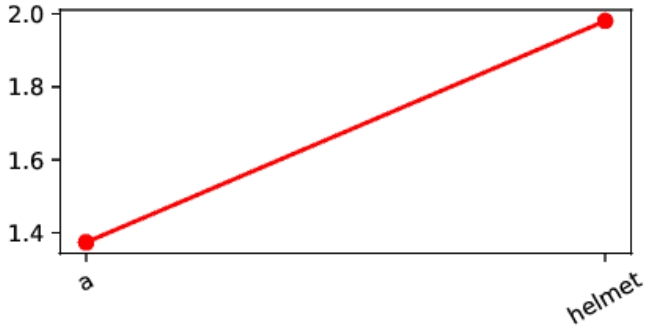
What sits on the police motorcycle? A helmet/ a pair of gloves

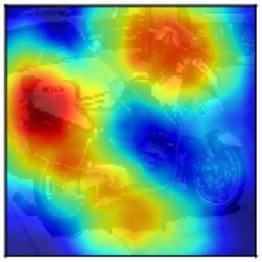
Input Image





Correct Answer





Wrong Answer