We derive a spatial attention weight from the question and answer representations.

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 $X_Q = \frac{1}{M} \sum_{i=1}^{M} e_i$.

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

$\mathbf{at}_i = \frac{\mathbf{a} \times \mathbf{d}_i}{\| \mathbf{a} \times \mathbf{d}_i \|}$ where $\mathbf{a} \times \mathbf{d}_i = \text{softmax}(\mathbf{a} \times \mathbf{d}_i)$

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 $\{Q, A, t_i\}$, where $t_i = 1$ and $t_i = 0$ for $i = 2, \ldots, N$.

The output for the $i$th candidate answer is $p_i = \text{sigmoid}(W_{QIA}x_{QIA} + b_{QIA})$

The full loss is $L = L_b + \lambda_2 L_a$

and $L_a = \max(n\text{margin}, p_i - n_1, 0)$

Visualizing Attention

Spatial Attention Question Word Attention Correct Answer

Wrong Answer

Visual Question Answering

Our Pipeline for VQA

Our Approach

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.

Results & Analysis

Method VQA VQA Test Dev
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 QA 67.3 60.2
POS Att + Conv N-Gram + Triplet attention QA + 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 777 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 $\lambda_2$ (f) find the optimal margin.

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.

References