MUREL: Multimodal Relational Reasoning for Visual Question Answering
Supplementary Materials

Remi Cadene, Hedi Ben-younes, Matthieu Cord, Nicolas Thome
1 Sorbonne Université, CNRS, LIP6, 4 place Jussieu, 75005 Paris
2 Heuritech, 110 avenue de la République, 75011 Paris
3 Conservatoire National des Arts et Métiers, 75003 Paris
remi.cadene@lip6.fr, hedi.ben-younes@lip6.fr, matthieu.cord@lip6.fr, nicolas.thome@cnam.fr

Experimental setup

Image We use the pretrained Faster-RCNN [10] by [1] on Visual Genome [9] to extract objects features from each image. Two setups have been proposed in the literature. A first one that extract 36 regions per image and a second one that extract 10 to 100 regions depending on a threshold. For the sake of simplicity, we choose the first setup in order to always represent our image as $R^{36 \times 2048}$. We do not fine tune any pf the Faster-RCNN parameters.

Question We use the same preprocessing as [4], which apply a lower case transformation and remove all the punctuation. We only consider the questions that are associated to the 3000 most occurring answers (1480 for the TDIUC dataset) while containing less than 26 words. We use a pre-trained Skip-thought encoder by [8] with a two glimpses self attention mechanism [11] to represent our question in a 4800-dimensional space. We fine tune every parameters of the Skip-thoughts including the embedding layer.

Optimization process We use the Adam optimizer [7] with a learning rate of $5 \times 10^{-5}$ and a batch size of 256. During the first 7 epochs, we linearly increase the learning rate to $2 \times 10^{-4}$. After the epoch 14, we decrease it by a factor 0.25 every two epochs until convergence. We also apply a gradient clipping of 0.25. We use early stopping based on the validation accuracy. This process is inspired from [12, 5].

Loss function We use the standard cross-entropy loss function for multi-class classification problems.

Comparison with classic attention

MuRel leverages the bilinear strategy in a different way than the classical VQA models [2, 3, 4, 6]. Instead of scalar question-guided visual attention maps, the fusion between question and each region is represented as a vector. This more expressive multidimensional representation allows MuRel to focus on specific features of a particular region given a textual context.

An other important aspect of MuRel lies in its pairwise module which models the relations between regions over multiple steps. Besides bringing more capacity, this pairwise modeling also allows to visualize the strongest region interactions, as we show in Figure 1, which is not possible with a classic attention model.

References


