Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering: trained with Adam (learning rate = 0.001) and L2 regularization on all the weights except for the bias with gamma = $5 * 10^{4}$ -6 Gated tanh layer Output size: 100 LSTM Multi-modal Embedding layer w/ dropout Gated tanh layer Premise w/ 300D GloVe fusion through Output size: 100 (keep prob. = 0.2)element-wise embeddings Output size: 100 multiplication Prem. top-down Gated tanh layer 32 2048D attention over Output size: 100 Concatenation visual vectors image regions extracted from 32 ReLU layer ReLU layer ReLU layer Softmax layer regions identified | L2 normalization Output size: 200 Output size: 200 Output size: 200 Output size: 3 through bottom-Hypo. top-down up attention on a Gated tanh layer attention over ResNet-101 Output size: 100 image regions Multi-modal **LSTM Embedding layer** fusion through Gated tanh layer w/ dropout Hypothesis w/ 300D GloVe element-wise (keep prob. = 0.2)Output size: 100 embeddings multiplication Output size: 100 Gated tanh layer Output size: 100