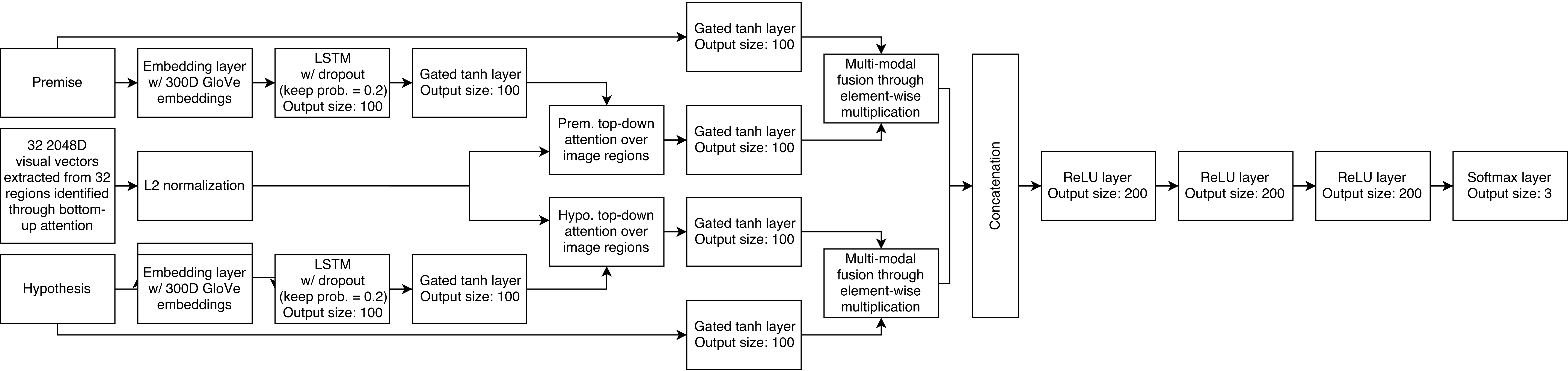
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering: trained with Adam (learning rate = 0.001) and L2 regularization on all the weights except for the bias with gamma = 5 * 10^-6



Premise
Embedding layer w/ 300D GloVe embeddings
LSTM w/ dropout (keep prob. = 0.2) Output size: 100
Gated tanh layer Output size: 100
Gated tanh layer Output size: 100
32 2048D visual vectors extracted from 32 regions identified through bottom-up attention
L2 normalization
Prem. top-down attention over image regions
Gated tanh layer Output size: 100
Multi-modal fusion through element-wise multiplication
Hypo. top-down attention over image regions
Gated tanh layer Output size: 100
Multi-modal fusion through element-wise multiplication
Hypothesis
Embedding layer w/ 300D GloVe embeddings
LSTM w/ dropout (keep prob. = 0.2) Output size: 100
Gated tanh layer Output size: 100
Gated tanh layer Output size: 100
Concatenation
ReLU layer Output size: 200
ReLU layer Output size: 200
ReLU layer Output size: 200
Softmax layer Output size: 3