1. Implementation of Tree LSTM

We simplified the implementation of tree LSTM (Eq. (4)) in the main paper as:

\[ c^t_j, h^t_j = \text{TreeLSTM}(e_t, \{c^t_{ij}\}, \{h^t_{ij}\}), \quad j \in C_t, \tag{1} \]

where \( c^t_j, h^t_j \) denote the cell and hidden vectors of the \( j \)-th child of node \( t \). Specifically, our tree LSTM transition equations are:

\[
\begin{align*}
\tilde{h}_t &= \sum_{j \in C_t} h^t_{ij}, \\
i_t &= \sigma(W^{(i)}e_t + U^{(i)}\tilde{h}_t + b^{(i)}), \\
o_t &= \sigma(W^{(o)}e_t + U^{(o)}\tilde{h}_t + b^{(o)}), \\
u_t &= \tanh(W^{(u)}e_t + U^{(u)}\tilde{h}_t + b^{(u)}), \\
f_{ij} &= \sigma(W^{(f)}e_t + U^{(f)}h^t_{ij} + b^{(f)}), \\
c^t_i &= i_t \odot u_t + \sum_{j \in C_t} f_{ij} \odot c^t_{ij}, \\
h^t_i &= o_t \odot \tanh(c^t_i),
\end{align*}
\]

where \( \odot \) is the element-wise multiplication, \( \sigma(\cdot) \) is the sigmoid function, \( W, U, b \) are trainable parameters.

2. More Qualitative Results

In this section, we provide more qualitative results to demonstrate the internal reasoning steps of NMTREE. In Figure 1, we visualize the reasoning process inside Comp modules. In Figure 2, Figure 3, and Figure 4, we visualize the tree structures, the module assembly, the attention map at each intermediate step, and the final results. Specifically, Figure 2 are qualitative results with ground-truth bounding boxes. As comparison, we also show some failure cases. Figure 3 are qualitative results with detected bounding boxes. Figure 4 are qualitative results with detected masks.
Figure 1. The compositional reasoning inside Comp. Each example contains the original image (left), the contextual attention map (middle) and the output attention map (right). We represent partial tree structure by colors: red for the current node, blue for children and green for parent.
Figure 2. Qualitative results with ground-truth bounding boxes. Words in different colors indicate corresponding modules: black for Single, red for Comp, and blue for Sum. The bottom right corner is the original image with a green bounding box as ground-truth and a red bounding box as our result.
Figure 3. Qualitative results with detected bounding boxes. Words in different colors indicate corresponding modules: black for Single, red for Comp, and blue for Sum. The bottom right corner is the original image with a green bounding box as ground-truth and a red bounding box as our result.

Figure 4. Qualitative results with detected masks. Words in different colors indicate corresponding modules: black for Single, red for Comp, and blue for Sum. The blue masks indicate the regions with maximum score, and the green masks indicate the ground-truth.