Simple but Effective: CLIP Embeddings for Embodied AI Supplemental Material

Apoorv Khandelwal* Luca Weihs* Roozbeh Mottaghi Aniruddha Kembhavi
Allen Institute for AI
{apoorvk, lucaw, roozbehm, anik}@allenai.org

A. Baseline Architecture Details

We provide further details here for implementations of baseline architectures for tasks in Sec. 4. We have already given such a description for RoboTHOR OBJECTNAV in Sec. 3. Each of these architectures is a replica of the baseline provided by task authors, with the visual encoder substituted with a frozen CLIP ResNet-50 encoder. And, for our ImageNet baseline agents, we simply use a frozen ResNet-50 that is pretrained on ImageNet instead.

A.1. Room Rearrangement in iTHOR

The model for ROOMR in iTHOR receives two $3\times224\times224$ RGB images at every step i_1 and i_2. These RGB image are encoded into $2048\times7\times7$ tensors I_1 and I_2 by a CLIP ResNet-50 model whose weights are frozen and final attention pooling and classification layers have been removed. These feature maps are stacked to form a $6144\times7\times7$ tensor $s = [I_1, I_2, I_1 * I_2]$. An attention mask formed by applying a 1×1 convolution to s with an output dimension of 512. s is then convolved to 512 channels (resulting in a shape of $512\times7\times7$) with another 1×1 kernel. The attention mask is then used for attention pooling, resulting in a 512-dim embedding V. V is passed into a 1-layer GRU with 512 hidden units, along with any prior hidden state. An actor head (one linear layer) maps the GRU output to a 6-dimensional vector of logits and a critic head (another linear layer) maps it to a scalar.

A.2. Habitat OBJECTNAV and POINTNAV

We use the same architecture for our POINTNAV and OBJECTNAV baselines in Habitat, with the only difference being the input goal and how it is encoded (to a 32-dim encoding G). Like in RoboTHOR, the input goal for Habitat OBJECTNAV is an integer $g \in \{0, ..., 20\}$ indicating an object category. In this case, g is used to index an embedding matrix and form G. In Habitat POINTNAV, the input goal is a 2-dim polar coordinate (to the target position, expressed relatively to the agent's current position). Here, g is passed through a linear layer to form G. Weights for both the embedding matrix and linear layer are learned during training.

The model receives G, a $3\times224\times224$ RGB image i, the action from the previous step a (as an integer index in the action space). The RGB image is encoded into a $2048\times7\times7$ tensor I by a CLIP ResNet-50 model whose weights are frozen and final attention pooling and classification layers have been removed. I is average pooled spatially (from 7×7 to 1×1) and flattened to form V: a 2048-dim visual embedding. The previous action a is used to index an embedding matrix to form a 32-dim encoding A. G, V, and A are passed into a 2-layer GRU with 512 hidden units, along with any prior hidden state. An actor head (one linear layer) maps the GRU output to a 6-dimensional vector of logits and a critic head (another linear layer) maps it to a scalar.

*Equal contribution