In this paper, we present a simple visual grounding framework that outperforms more complicated baselines such as MSRR and MAC. First, an object detector extracts region proposals from the input image. We utilize centernet for this. Second, we match the region proposals with the command. In particular, we compute the feature cosine similarity between encodings from the region proposals and the command. The region proposal encodings are extracted using efficientnet-b2 and the sentence is encoded with a state of the art sentence transformer RoBERTa. The cosine similarity values can be appraised as a score for how well the contents of the bounding box fit the command. The model is trained by mapping the cosine similarities to zero or one, depending on whether the proposed region has an IOU overlap of at least 0.5 with the ground-truth bounding box.