We address the problem of phrase grounding by learning a multi-level common semantic space shared by the textual and visual modalities. We exploit multiple