

Region-Word Matching - Alignment

Models local cross-modal correlation between image regions and words for dense visual recognition tasks such as object detection.

Loss Formulation

$$\mathcal{L}_{RW} = p \log S_r(r^I, w^T) + (1 - p) \log(1 - S_r(r^I, w^T))$$