

Existing methods for image captioning are usually trained by cross entropy loss, which leads to exposure bias and the inconsistency between the optimizing fun