Deep neural networks have achieved great successes on the image captioning task. However, most of the existing models depend heavily on paired image-sen