

ABSTRACT

A multi-modal pre-training method and device. The multi-modal pre-training method comprises : sampling a video in a video-text pair to obtain a first video frame sequence, and performing word segmentation processing on text in the video-text pair to obtain a first word segmentation sequence (101); performing mask processing on the first video frame sequence to obtain a second video frame sequence, and performing mask processing on the first word segmentation sequence to obtain a second word segmentation sequence (102); encoding the first video frame sequence to obtain a first video feature, and encoding the first word segmentation sequence to obtain a first word segmentation feature (103); encoding the second video frame sequence to obtain a second video feature, and encoding the second word segmentation sequence to obtain a second word segmentation feature (104); determining a pre-trained objective function by using the first video feature, the first word segmentation feature, the second video feature, and the second word segmentation feature (105); and performing multi-modal pre-training by using the objective function (106).