MULTI-MODAL PRE-TRAINING METHOD AND MULTI-MODAL PRE-TRAINING APPARATUS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application is based on and claims priority from the application of CN application No. 202111078728.2 filed on September 15, 2021, the disclosure of hereby incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of information processing, and in particular, to a multi-modal pre-training method and a multi-modal pre-training apparatus.

BACKGROUND

[0003] The multi-modal pre-training technology for visual language is one of recently emerging subjects in the multi-modal field, which aims at enabling a model to pre-train large-scale weakly labeled visual data (such as images and videos) and text data to obtain a better multi-modal feature representation, thereby improving the performance of various multi-modal task models.

[0004] Technologies related to the multi-modal pre-training for the visual language are basically methods of pre-training a model with reference to BERT (Bidirectional Encoder Representations From Transformer) in the field of natural language processing.