

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号
WO 2023/040306 A1

(43) 国际公布日
2023 年 3 月 23 日 (23.03.2023)

(51) 国际专利分类号:
G06K 9/00 (2022.01) *G06F 16/483* (2019.01)
G06K 9/62 (2022.01) *G06F 40/289* (2020.01)

(21) 国际申请号: PCT/CN2022/092680

(22) 国际申请日: 2022 年 5 月 13 日 (13.05.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111078728.2 2021年9月15日 (15.09.2021) CN

(71) 申请人: 北京京东尚科信息技术有限公司 (BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO.,LTD.) [CN/CN]; 中国北京市海淀区知春路 76 号 8 层, Beijing 100086 (CN)。

(72) 发明人: 李业豪 (LI, Yehao); 中国北京市海淀区知春路 76 号 8 层, Beijing 100086 (CN)。 潘滢炜 (PAN, Yingwei); 中国北京市海淀区知春路 76 号 8 层, Beijing 100086 (CN)。 姚霆 (YAO, Ting); 中国北京市海淀区知春路 76 号 8 层, Beijing 100086 (CN)。 梅涛 (MEI, Tao); 中国北京市海淀区知春路 76 号 8 层, Beijing 100086 (CN)。

(74) 代理人: 中国贸促会专利商标事务所有限公司 (CCPIT PATENT AND TRADEMARK LAW OFFICE); 中国北京市复兴门内大街 158 号远洋大厦 F10 层, Beijing 100031 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS,

(54) Title: MULTI-MODAL PRE-TRAINING METHOD AND DEVICE

(54) 发明名称: 多模态预训练方法和装置

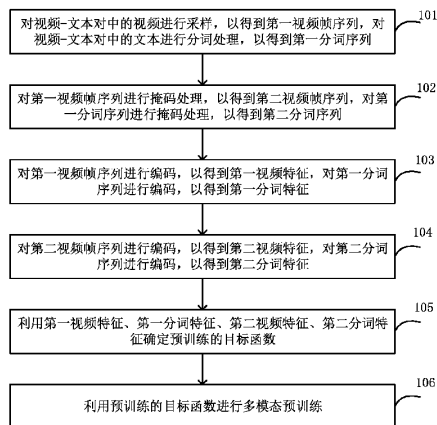


图 1

- 101 Sample a video in a video-text pair to obtain a first video frame sequence, and perform word segmentation processing on text in the video-text pair to obtain a first word segmentation sequence
- 102 Perform mask processing on the first video frame sequence to obtain a second video frame sequence, and perform mask processing on the first word segmentation sequence to obtain a second word segmentation sequence
- 103 Encode the first video frame sequence to obtain a first video feature, and encode the first word segmentation sequence to obtain a first word segmentation feature
- 104 Encode the second video frame sequence to obtain a second video feature, and encode the second word segmentation sequence to obtain a second word segmentation feature
- 105 Determine 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
- 106 Perform multi-modal pre-training by using the pre-trained objective function

(57) 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).

[见续页]

IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种多模态预训练方法和装置。多模态预训练方法包括: 对视频-文本对中的视频采样得到第一视频帧序列, 对视频-文本对中的文本进行分词处理得到第一分词序列 (101); 对第一视频帧序列进行掩码处理得到第二视频帧序列, 对第一分词序列进行掩码处理得到第二分词序列 (102); 对第一视频帧序列进行编码得到第一视频特征, 对第一分词序列进行编码得到第一分词特征 (103); 对第二视频帧序列进行编码得到第二视频特征, 对第二分词序列进行编码得到第二分词特征 (104); 利用第一视频特征、第一分词特征、第二视频特征、第二分词特征确定预训练的目标函数 (105); 利用目标函数进行多模态预训练 (106)。