

(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 2025/123998 A1

(43) 国际公布日
2025 年 6 月 19 日 (19.06.2025)

WIPO | PCT

(51) 国际专利分类号:
G06F 16/583 (2019.01)

(21) 国际申请号: PCT/CN2024/129489

(22) 国际申请日: 2024 年 11 月 1 日 (01.11.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202311724117.X 2023 年 12 月 14 日 (14.12.2023) CN

(71) 申请人: 杭州阿里巴巴海外互联网产业有限公司 (HANGZHOU ALIBABA INTERNATIONAL INTERNET INDUSTRY CO., LTD.) [CN/CN]; 中国浙江省杭州市滨江区长河街道网商路 699 号 5 号楼 3 楼 303 室 310052 (CN)。

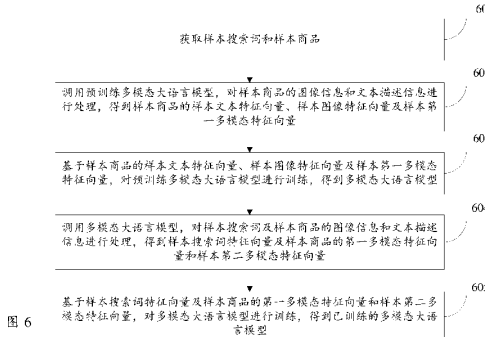
(72) 发明人: 郭元凯 (GUO, Yuankai); 中国北京市朝阳区广善路 21 号院-阿里巴巴北京朝阳科技园 A 区 100020 (CN)。刘凯 (LIU, Kai); 中国北京市朝阳区广善路 21 号院-阿里巴巴北京朝阳科技园 A 区 100020 (CN)。赵荣勋 (ZHAO, Rongxun); 中国北京市朝阳区广善路 21 号院-阿里巴巴北京朝阳科技园 A 区 100020 (CN)。

(74) 代理人: 北京同钧律师事务所 (BEIJING TONGJUN LAW FIRM); 中国北京市海淀区西直门北大街 32 号院 1 号楼 7 层 808 室 100082 (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, CV, 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,

(54) Title: MULTIMODAL LARGE LANGUAGE MODEL TRAINING METHOD, CORRELATION CALCULATION METHOD, AND LABEL GENERATION METHOD

(54) 发明名称: 多模态大语言模型训练方法、相关性计算及标签生成方法



601 ACQUIRE A SAMPLE SEARCH WORD AND A SAMPLE COMMODITY
602 CALL A PRE-TRAINED MULTIMODAL LARGE LANGUAGE MODEL TO PROCESS IMAGE INFORMATION AND TEXT DESCRIPTION INFORMATION OF THE SAMPLE COMMODITY, TO OBTAIN A SAMPLE TEXT FEATURE VECTOR, A SAMPLE IMAGE FEATURE VECTOR AND A SAMPLE FIRST MULTIMODAL FEATURE VECTOR OF THE SAMPLE COMMODITY
603 ON THE BASIS OF THE SAMPLE TEXT FEATURE VECTOR, THE SAMPLE IMAGE FEATURE VECTOR AND THE SAMPLE FIRST MULTIMODAL FEATURE VECTOR OF THE SAMPLE COMMODITY, TRAIN THE PRE-TRAINED MULTIMODAL LARGE LANGUAGE MODEL TO OBTAIN A MULTIMODAL LARGE LANGUAGE MODEL
604 CALL THE MULTIMODAL LARGE LANGUAGE MODEL TO PROCESS THE SAMPLE SEARCH WORD, AND THE IMAGE INFORMATION AND THE TEXT DESCRIPTION INFORMATION OF THE SAMPLE COMMODITY, TO OBTAIN A SAMPLE SEARCH WORD FEATURE VECTOR, AND A FIRST MULTIMODAL FEATURE VECTOR AND A SAMPLE SECOND MULTIMODAL FEATURE VECTOR OF THE SAMPLE COMMODITY
605 ON THE BASIS OF THE SAMPLE SEARCH WORD FEATURE VECTOR, AND THE FIRST MULTIMODAL FEATURE VECTOR AND THE SAMPLE SECOND MULTIMODAL FEATURE VECTOR OF THE SAMPLE COMMODITY, TRAIN THE MULTIMODAL LARGE LANGUAGE MODEL TO OBTAIN A TRAINED MULTIMODAL LARGE LANGUAGE MODEL

(57) Abstract: The present disclosure relates to the technical field of artificial intelligence, and provides a multimodal large language model training method, a correlation calculation method, and a label generation method. The multimodal large language model training method comprises: on the basis of a sample text feature vector, a sample image feature vector, and a sample first multimodal feature vector which are obtained by processing image information and text description information of a sample commodity by a pre-trained multimodal large language model, training the pre-trained multimodal large language model to obtain a multimodal large language

[见续页]

IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ,
LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN,
MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, 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, ME, 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))。

model; and processing a sample search word, the image information and the text description information on the basis of the multimodal large language model to obtain a sample search word feature vector, a first multimodal feature vector and a sample second multimodal feature vector, and training the multimodal large language model to obtain a trained multimodal large language model. The trained multimodal large language model of the present disclosure can simultaneously learn image features related to the search word and the text description information, and thus the generated second multimodal feature vector is more accurate.

(57) 摘要: 本公开提供了一种多模态大语言模型训练方法、相关性计算及标签生成方法，属于人工智能技术领域。所述方法包括：基于预训练多模态大语言模型对样本商品的图像信息和文本描述信息进行处理得到的样本文本特征向量、样本图像特征向量及样本第一多模态特征向量，对预训练多模态大语言模型进行训练，得到多模态大语言模型；基于多模态大语言模型对样本搜索词、图像信息及文本描述信息进行处理得到样本搜索词特征向量及第一多模态特征向量和样本第二多模态特征向量，对多模态大语言模型进行训练，得到已训练的多模态大语言模型。本公开所训练的多模态大语言模型能够同时学习到与搜索词和文本描述信息有关的图像特征，生成的第二多模态特征向量更准确。