(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

(43) 国际公布日
2024 年 7 月 4 日 (04.07.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/139307 A1

(51) 国际专利分类号:
G06F 40/232 (2020.01)

(21) 国际申请号: PCT/CN2023/115054

(22) 国际申请日: 2023 年 8 月 25 日 (25.08.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211680372.4 2022年12月27日 (27.12.2022) CN

(71) 申请人: 苏州元脑智能科技有限公司 (**SUZHOU METABRAIN INTELLIGENT TECHNOLOGY CO., LTD.**) [CN/CN]; 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。

(72) 发明人:李晓川(**LI, Xiaochuan**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。 赵雅倩(**ZHAO, Yaqian**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。 李仁刚(**LI, Rengang**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。 郭振华(**GUO, Zhenhua**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。 范宝余(**FAN, Baoyu**); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。

(74) 代理人: 北京润泽恒知识产权代理有限公司(**BEIJING RUN ZEHENG INTELLECTUAL PROPERTY LAW FIRM**); 中国北京市海淀区中关村南大街甲18号北京国际C座6层606, Beijing 100081 (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,

(54) **Title:** SENTENCE CORRECTION METHOD AND APPARATUS FOR IMAGE, AND ELECTRONIC DEVICE AND STORAGE MEDIUM

(54) 发明名称: 图像的文本纠错方法、装置、电子设备及存储介质

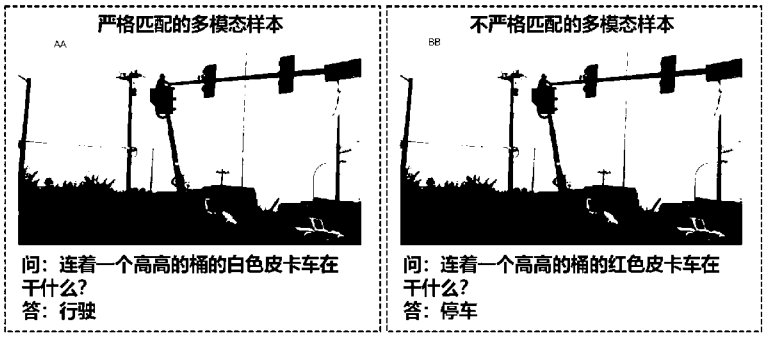


图 1

AA Strictly matched multi-modal sample
BB Not strictly matched multi-modal sample

(57) **Abstract:** Provided in the embodiments of the present application are a sentence correction method and apparatus for an image, and an electronic device and a non-volatile readable storage medium, which are applied to a multi-modal sentence correction system. The method comprises: firstly, performing feature extraction on both an input image and an input original sentence, so as to obtain an image feature and an original sentence feature; then, generating a comprehensive encoded feature by means of feature splicing, and obtaining an encoded sentence feature by means of capturing from the comprehensive encoded feature a feature corresponding to the position of the original sentence feature; then, performing feature correction on the encoded sentence feature by means of a sentence feature correction module, so as to generate a corrected sentence feature, and performing feature fusion on the corrected sentence feature and the encoded sentence feature by means of a correction vector accessor, so as to obtain a target sentence feature; and finally, performing feature replacement on the original sentence feature by means of a correction decoder and by using the target sentence feature, and outputting target sentence information. In this way, the identification and correction of a fine-grained error in an original sentence are realized, thereby greatly reducing the error rate when a multi-modal task is performed.

WO 2024/139307 A1

[见续页]

GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, 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))。

(57) 摘要: 本申请实施例提供了一种图像的文本纠错方法、装置、电子设备及非易失性可读存储介质，应用于多模态文本纠错系统，首先将输入的图像与原始文本分别进行特征提取，获得图像特征以及原始文本特征，接着通过特征拼接方式生成综合编码特征，并通过截取综合编码特征中对应于原始文本特征位置的特征，获得文本编码特征，接着通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，最后通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出目标文本信息，从而实现了对原始文本中细粒度错误的识别及纠正，大大降低了在进行多模态任务时的出错率。

图像的文本纠错方法、装置、电子设备及存储介质

相关申请的交叉引用

本申请要求于 2022 年 12 月 27 日提交中国专利局，申请号为 202211680372.4，申请名称为“图像的文本纠错方法、装置、电子设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，特别是涉及一种图像的文本纠错方法、一种图像的文本纠错装置、一种电子设备以及一种计算机非易失性可读存储介质。

背景技术

近年来，多模态人工智能成为 AI（Artificial Intelligence，人工智能）领域中重要的研究方向之一。多模态研究，旨在综合诸如图像、视频、音频、文本、传感器信号等多种模态输入，并综合理解或生成人类可用的信息的科学，如在视觉问答（Visual Question Answering，VQA）、视觉定位（Visual Grounding）等领域均包含图像、文本等多模态关系。随着 Transformer（基于自注意力机制的深度学习模型）结构的广泛应用，从而在诸如视觉问答 VQA、图片描述（Image Caption）、视觉对话（Visual Dialog）等多模态任务中，基于 Transformer 的多模态网络结构也越来越受到人们的青睐。

在现实世界中，人类的语言往往存在口误、比喻等常见语言现象，这些现象难以被现有的计算机语言技术掌握，从而在进行文本与图像间的匹配时，往往无法将这些带有口误或者比喻修辞手法的词语与图像进行对应匹配，也就是说，现阶段多模态理论研究无法精细地区分文本中的微小错误，例如一段文字可能错了某个字或某个词语，从而导致算法在完成多模态任务时出现错误，例如，在基于视觉问答的任务中，极有可能会遇到因带有刻意比喻的文字内容，导致算法无法理解人类实际想要描述的问题的情况，使得基于 Transformer 的多模态结构无法通过算法给出正确应答，从而给出错误答案。

发明内容

本申请实施例是提供一种图像的文本纠错方法、装置、电子设备以及计算机非易失性可读存储介质，以解决或部分解决因文本与图像无法严格匹配，导致算法在完成多模态任务时容易出现错误的问题。

本申请实施例公开了一种图像的文本纠错方法，应用于多模态文本纠错系统，多模态文本纠错系统至少包括文本特征修正模块、纠错向量存取器以及纠错解码器，方法包括：

响应于针对图像与文本的输入操作，获取输入操作对应的图像信息与原始文本信息，并分别对图像信息与原始文本信息进行特征提取，获得与图像信息对应的图像特征，以及与原始文本信息对应的原始文本特征；

将图像特征与原始文本特征进行特征拼接，获得综合编码特征，并根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征；

通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征；

通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息。

可选地，通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，包括：

通过文本特征修正模块对文本编码特征进行自注意力编码，获得对应的初始自注意力向量，并对初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，目标自注意力向量包含图像特征与原始文本特征的关联特征；

对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，有效文本信息向量表示文本编码特征中每个字符包含有效信息的概率；

对文本编码特征进行双向截取，分别获得前错位特征与后错位特征，并根据前错位特征与后错位特征，生成相邻特征交互向量；

对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，相邻文本信息向量表示文本编码特征中相邻字符连贯的概率；

采用目标自注意力向量、有效文本信息向量以及相邻文本信息向量对文本编码特征进行特征纠正，生成文本纠正特征。

可选地，通过文本特征修正模块对文本编码特征进行自注意力编码，获得对应的初始自注意力向量，包括：

将文本编码特征输入至自注意力层中，采用公式

$$\boldsymbol{i_{emlm} = softmax\left(\frac{\left(W_q \cdot f\right)^T \times \left(W_k \cdot f\right)}{\sqrt{size(f)}}\right) \times \left(W_v \cdot f\right)}$$

进行自注意力编码，获得对应的初始自注意力向量；其中，

$\boldsymbol{W_q}$ 、 $\boldsymbol{W_k}$ 、 $\boldsymbol{W_v}$ 均为可学习权重，f 为文本编码特征。

可选地，对初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，包括：

将初始自注意力向量输入至两组全连接层中分别进行当前字符预测处理与前置字符预测处理，获得当前预测向量与前置预测向量；

根据当前预测向量与前置预测向量确定目标自注意力向量。

可选地，根据当前预测向量与前置预测向量确定目标自注意力向量，包括：

采用当前预测向量对文本编码特征进行预测处理，获得文本编码特征对应的目标当前字符；

采用前置预测向量对目标当前字符进行预测处理，获得目标当前字符对应的目标前置字符；

将目标前置字符与目标当前字符进行拼接，输出对应的目标字符，并生成目标字符对应的目标自注意力向量。

可选地，采用当前预测向量对文本编码特征进行预测处理，获得文本编码特征对应的目标当前字符，包括：

根据当前预测向量，将文本编码特征与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的当前预测概率，并将当前预测概率最大的预设字符确定为目标当前字符。

可选地，采用前置预测向量对目标当前字符进行预测处理，获得目标当前字符对应的目

标前置字符，包括：

根据前置预测向量，将目标当前字符与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的前置预测概率，并将前置预测概率最大的预设字符确定为目标前置字符。

可选地，对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，包括：

采用公式

$$p_{ifo} = sigmoid(W_{iw}\left(softmax\left(\frac{(W_{iq}\cdot f)^{T}\times(W_{ik}\cdot f)}{\sqrt{size(f)}}\right)\times(W_{iv}\cdot f)\right)+b_{ib})$$

对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量；其中，W_{iq}、W_{ik}、W_{iv}均为转移矩阵权重，W_{iw}为信息量预测权重，b_{ib}为可学习模型参数，f为文本编码特征。

可选地，文本编码特征的大小为[M，d]，前错位特征与后错位特征的大小均为[M-1，d]，根据前错位特征与后错位特征，生成相邻特征交互向量，包括：

将前错位特征与后错位特征进行向量级联处理，生成与文本编码特征对应的大小为[M-1，d×2]的相邻特征交互向量。

可选地，对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，包括：

采用公式

$$p_{nbo} = sigmoid(W_{nw}\left(softmax\left(\frac{(W_{nq}\cdot f_{nb})^{T}\times(W_{nk}\cdot f_{nb})}{\sqrt{size(f_{nb})}}\right)\times(W_{nv}\cdot f_{nb})\right)+b_{in})$$

对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量；其中，W_{nw}、W_{nq}、W_{nv}、W_{nk}均为转移矩阵权重参数，b_{in}为偏置向量参数，f_{nb}为相邻特征交互向量。

可选地，纠错向量存取器至少包括特征存储空间，在根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征之后，方法还包括：

将文本编码特征拆分为若干个子文本特征，并将各个子文本特征依次存储至特征存储空间。

可选地，纠错向量存取器包括修复判断门以及特征更新器，通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，包括：

通过修复判断门对各个子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征；

通过特征更新器采用文本纠正特征对至少一个替换子文本特征进行特征替换，获得各自对应的目标子文本特征，并将至少一个目标子文本特征进行特征融合，获得对应的目标文本特征。

可选地，通过修复判断门对各个子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征，包括：

采用公式

$$s(x_k)=\begin{cases}1,\ p_{ifok}<thresh_{ifo}\cup p_{nbok}<thresh_{nbo}\\0,\ p_{ifok}\geq thresh_{ifo}\cap p_{nbok}\leq thresh_{nbo}\end{cases}$$

对各个子文本特征进行修复判断；

当$s(x_k)$为 1 时，将特征序号为 k 的子文本特征确定为需进行特征替换的替换子文本特征；其中，

k 表示子文本特征对应的特征序号，p_{ifok}为特征序号为 k 的子文本特征对应的有效文本信息向量，p_{nbok}为特征序号为 k 的子文本特征对应的相邻文本信息向量，$thresh_{ifo}$表示可设定信息量概率阈值，$thresh_{nbo}$表示可设定通顺概率阈值，$s(x_k)$表示特征序号为 k 的子文本特征是否需要进行特征替换。

可选地，通过特征更新器采用文本纠正特征对至少一个替换子文本特征进行特征替换，获得各自对应的目标子文本特征，包括：

根据文本纠正特征，采用公式

$$f_{ko}=f_k\times(1-\mu)+\left(p_{ifok}\times\theta+p_{nbok}\times(1-\theta)\right)\times\mu\times o_{emlm}$$

计算替换子文本特征对应的文本特征值，并根据文本特征值对替换子文本特征进行特征替换，获得对应的目标子文本特征；其中，

f_k为特征序号为 k 的子文本特征，o_{emlm}为目标自注意力向量，θ与μ均为大小为 0~1 的预设参数。

可选地，根据文本特征值对替换子文本特征进行特征替换，获得对应的目标子文本特征，包括：

采用文本特征值通过覆盖原值方式，对替换子文本特征的原有文本特征值进行替换，获得对应的目标子文本特征。

可选地，将图像特征与原始文本特征进行特征拼接，获得综合编码特征，包括：

将图像特征与原始文本特征进行特征拼接，并进行跨模态编码处理，获得综合编码特征。

可选地，根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征，包括：

对综合编码特征中与原始文本特征位置对应的特征进行截取，获得与原始文本特征对应的文本编码特征。

本申请实施例还公开了一种图像的文本纠错装置，应用于多模态文本纠错系统，多模态文本纠错系统至少包括文本特征修正模块、纠错向量存取器以及纠错解码器，装置包括：

特征提取模块，用于响应于针对图像与文本的输入操作，获取输入操作对应的图像信息与原始文本信息，并分别对图像信息与原始文本信息进行特征提取，获得与图像信息对应的图像特征，以及与原始文本信息对应的原始文本特征；

文本编码特征生成模块，用于将图像特征与原始文本特征进行特征拼接，获得综合编码特征，并根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征；

目标文本特征生成模块，用于通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获

得目标文本特征；

文本特征替换模块，用于通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息。

可选地，文本编码特征生成模块包括：

目标自注意力向量生成模块，用于通过文本特征修正模块对文本编码特征进行自注意力编码，获得对应的初始自注意力向量，并对初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，目标自注意力向量包含图像特征与原始文本特征的关联特征；

有效文本信息向量生成模块，用于对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，有效文本信息向量表示文本编码特征中每个字符包含有效信息的概率；

相邻特征交互向量生成模块，用于对文本编码特征进行双向截取，分别获得前错位特征与后错位特征，并根据前错位特征与后错位特征，生成相邻特征交互向量；

相邻文本信息向量生成模块，用于对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，相邻文本信息向量表示文本编码特征中相邻字符连贯的概率；

文本纠正特征生成模块，用于采用目标自注意力向量、有效文本信息向量以及相邻文本信息向量对文本编码特征进行特征纠正，生成文本纠正特征。

可选地，目标自注意力向量生成模块包括：

初始自注意力向量生成模块，用于将文本编码特征输入至自注意力层中，采用公式

$$i_{emlm} = softmax\left(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{size(f)}}\right) \times (W_v \cdot f)$$

进行自注意力编码，获得对应的初始自注意力向量；其中，

W_q、W_k、W_v均为可学习权重，f为文本编码特征。

可选地，目标自注意力向量生成模块包括：

字符预测处理模块，用于将初始自注意力向量输入至两组全连接层中分别进行当前字符预测处理与前置字符预测处理，获得当前预测向量与前置预测向量；

目标自注意力向量确定子模块，用于根据当前预测向量与前置预测向量确定目标自注意力向量。

可选地，目标自注意力向量确定子模块包括：

目标当前字符生成模块，用于采用当前预测向量对文本编码特征进行预测处理，获得文本编码特征对应的目标当前字符；

目标前置字符生成模块，用于采用前置预测向量对目标当前字符进行预测处理，获得目标当前字符对应的目标前置字符；

目标字符输出模块，用于将目标前置字符与目标当前字符进行拼接，输出对应的目标字符，并生成目标字符对应的目标自注意力向量。

可选地，目标当前字符生成模块具体用于：

根据当前预测向量，将文本编码特征与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的当前预测概率，并将当前预测概率最大的预设字符确定为目标当前字符。

可选地，目标前置字符生成模块具体用于包括：

根据前置预测向量，将目标当前字符与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的前置预测概率，并将前置预测概率最大的预设字符确定为目标前置字符。

可选地，有效文本信息向量生成模块具体用于：

采用公式

$$\boldsymbol{p_{ifo} = sigmoid(W_{iw}\left(softmax\left(\frac{\left(W_{iq}\cdot f\right)^{T}\times\left(W_{ik}\cdot f\right)}{\sqrt{size(f)}}\right)\times\left(W_{iv}\cdot f\right)\right)+b_{ib})}$$

对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量；其中，$\boldsymbol{W_{iq}}$、$\boldsymbol{W_{ik}}$、$\boldsymbol{W_{iv}}$均为转移矩阵权重，$\boldsymbol{W_{iw}}$为信息量预测权重，$\boldsymbol{b_{ib}}$为可学习模型参数，f为文本编码特征。

可选地，文本编码特征的大小为[M，d]，前错位特征与后错位特征的大小均为[M-1，d]，相邻特征交互向量生成模块具体用于：

将前错位特征与后错位特征进行向量级联处理，生成与文本编码特征对应的大小为[M-1，d×2]的相邻特征交互向量。

可选地，相邻文本信息向量生成模块具体用于：

采用公式

$$\boldsymbol{p_{nbo} = sigmoid(W_{nw}\left(softmax\left(\frac{\left(W_{nq}\cdot f_{nb}\right)^{T}\times\left(W_{nk}\cdot f_{nb}\right)}{\sqrt{size(f_{nb})}}\right)\times\left(W_{nv}\cdot f_{nb}\right)\right)+b_{in})}$$

对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量；其中，$\boldsymbol{W_{nw}}$、$\boldsymbol{W_{nq}}$、$\boldsymbol{W_{nv}}$、$\boldsymbol{W_{nk}}$均为转移矩阵权重参数，$\boldsymbol{b_{in}}$为偏置向量参数，$\boldsymbol{f_{nb}}$为相邻特征交互向量。

可选地，纠错向量存取器至少包括特征存储空间，装置还包括：

子文本特征拆分模块，用于将文本编码特征拆分为若干个子文本特征，并将各个子文本特征依次存储至特征存储空间。

可选地，纠错向量存取器包括修复判断门以及特征更新器，目标文本特征生成模块包括：

替换子文本特征确定模块，用于通过修复判断门对各个子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征；

目标子文本特征确定模块，用于通过特征更新器采用文本纠正特征对至少一个替换子文本特征进行特征替换，获得各自对应的目标子文本特征，并将至少一个目标子文本特征进行特征融合，获得对应的目标文本特征。

可选地，替换子文本特征确定模块具体用于：

采用公式

$$\boldsymbol{s(x_k)=\begin{cases}1,\ p_{ifok}<thresh_{ifo}\cup p_{nbok}<thresh_{nbo}\\0,\ p_{ifok}\geq thresh_{ifo}\cap p_{nbok}\leq thresh_{nbo}\end{cases}}$$

对各个子文本特征进行修复判断；

当$s(x_k)$为 1 时，将特征序号为 k 的子文本特征确定为需进行特征替换的替换子文本特征；其中，

k 表示子文本特征对应的特征序号，p_{ifok}为特征序号为 k 的子文本特征对应的有效文本信息向量，p_{nbok}为特征序号为 k 的子文本特征对应的相邻文本信息向量，$thresh_{ifo}$表示可设定信息量概率阈值，$thresh_{nbo}$表示可设定通顺概率阈值，$s(x_k)$表示特征序号为 k 的子文本特征是否需要进行特征替换。

可选地，目标子文本特征确定模块包括：

文本特征值计算模块，用于根据文本纠正特征，采用公式

$$f_{ko} = f_k \times (1-\mu) + \left(p_{ifok} \times \theta + p_{nbok} \times (1-\theta)\right) \times \mu \times o_{emlm}$$

计算替换子文本特征对应的文本特征值，并根据文本特征值对替换子文本特征进行特征替换，获得对应的目标子文本特征；其中，

f_k为特征序号为 k 的子文本特征，o_{emlm}为目标自注意力向量，θ与μ均为大小为 0~1 的预设参数。

可选地，文本特征值计算模块具体用于：

采用文本特征值通过覆盖原值方式，对替换子文本特征的原有文本特征值进行替换，获得对应的目标子文本特征。

可选地，文本编码特征生成模块包括：

跨模态编码处理模块，用于将图像特征与原始文本特征进行特征拼接，并进行跨模态编码处理，获得综合编码特征。

可选地，文本编码特征生成模块包括：

对应特征位置截取模块，用于对综合编码特征中与原始文本特征位置对应的特征进行截取，获得与原始文本特征对应的文本编码特征。

本申请实施例还公开了一种电子设备，包括处理器、通信接口、存储器和通信总线，其中，处理器、通信接口以及存储器通过通信总线完成相互间的通信；

存储器，用于存放计算机程序；

处理器，用于执行存储器上所存放的程序时，实现如本申请实施例的方法。

本申请实施例还公开了一种计算机非易失性可读存储介质，其上存储有指令，当由一个或多个处理器执行时，使得处理器执行如本申请实施例的方法。

本申请实施例包括以下优点：

在本申请实施例中，提供了一种应用于多模态文本纠错系统的基于图像的文本纠错方法，首先将输入的图像与原始文本分别进行特征提取，获得图像特征以及原始文本特征，接着将图像特征以及原始文本特征通过特征拼接方式生成综合编码特征，并通过截取综合编码特征中对应于原始文本特征位置的特征，获得文本编码特征，接着通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，最后通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息，从而根据图像实现了对原始文本中细粒度错误进行识别并纠正，大大降低了在进行多模态任务时的出错率。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 是一种当前不严格匹配的多模态样本对现有方法的干扰示意图；

图 2 是本申请实施例中提供的一种基于图像的文本纠错多模态样本示意图；

图 3 是本申请实施例中提供的一种基于视觉弹性掩膜的文本纠错系统示意图；

图 4 是本申请实施例中提供的一种图像的文本纠错方法的步骤流程图；

图 5 是本申请实施例中提供的一种视觉弹性掩膜示意图；

图 6 是本申请实施例中提供的一种相邻词汇关系预测示意图；

图 7 是本申请实施例中提供的一种纠错向量存取器的结构框架示意图；

图 8 是本申请实施例中提供的一种图像的文本纠错装置的结构框图；

图 9 是本申请实施例中提供的一种计算机非易失性可读介质的示意图；

图 10 是本申请实施例中提供的一种电子设备的框图。

具体实施方式

为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

为了使本领域技术人员更好地理解本申请实施例中的技术方案，下面对本申请实施例中涉及的部分技术特征进行解释、说明：

文本纠错（Sentence Correction，SC）：检测出文本错误并进行对应纠正。

多模态（Multi Modal，MM）：即多种异构模态数据协同推理，在人工智能领域中，往往指感知信息，如图像、文本、视频、音频等协同，帮助人工智能更准确地理解外部世界。

掩膜文本预测（Masked Language Modeling，MLM）：掩膜，常用于图像处理场景，可以通过掩膜技术实现对图像中对应的文本进行预测。

弹性掩膜文本预测（Masked Language Modeling Elastic，EMLM）:比掩膜文本预测更为精确的文本预测方式，与现有掩膜方式不同的是，本申请中采用弹性掩膜文本预测可以实现不定长文本的词汇替换。

作为一种示例，在现实世界中，人类的语言往往存在口误、比喻等常见语言现象，这些现象难以被现有的计算机语言技术掌握，从而在进行文本与图像间的匹配时，往往无法将这些带有口误或者比喻修辞手法的词语与图像进行对应匹配，也就是说，现阶段多模态理论研究无法精细地区分文本中的微小错误，例如一段文字可能错了某个字或某个词语，从而导致算法在完成多模态任务时出现错误，例如，在基于视觉问答的任务中，极有可能会遇到因带有刻意比喻的文字内容，导致算法无法理解人类实际想要描述的问题的情况，使得基于 Transformer 的多模态结构无法通过算法给出正确应答，从而给出错误答案。

为更好地进行说明，参照图 1，示出了一种当前不严格匹配的多模态样本对现有方法的干扰示意图，如图 1 的右边框内为不严格匹配的多模态样本，其中显示情景以问答形式展示，具体为：问的一方提出“连着一个高高的桶的红色皮卡车在干什么？”，答的一方则回复“停车”，而由图像内容，可以得出实际情况是：连着一个高高的桶的皮卡车的车身颜色实为白色，且该白色皮卡车在红绿灯下，且前方有一辆车，则很明显可以看出，该白色皮卡车实

际正在行驶，而不是停车，从而该图像与文本的匹配关系应当如图 1 左边框内所示的情景，图 1 左边框内则为严格匹配的多模态样本，同样是以问答形式展示，具体为：问的一方提出“连着一个高高的桶的白色皮卡车在干什么？”，答的一方则回复“行驶”，从而可以得出，在进行图像与文本之间的匹配时，如果不能将图像与文本进行严格匹配，容易因信息判断错误导致输出错误答案。

进一步地，参照图 2，示出了本申请实施例中提供的一种基于图像的文本纠错多模态样本示意图，如图所示，图像依然采用的是图 1 中的图像，其输入文本为“连着一个高高的桶的红色皮卡车行驶在马路上”，可以看出，皮卡车的车身颜色明显是错误的，则需要根据图像中所显示的信息对输入文本中错误信息进行对应纠正，如应将“红色”纠正为“白色”，从而通过本申请提供的一种基于图像的文本纠错方法对该输入文本进行纠错，可以获得对应的输出文本应当为“连着一个高高的桶的白色皮卡车行驶在马路上”，实现对应于图像与文本之间的严格匹配，得出正确的文本内容。

因此，本申请实施例的核心申请点之一在于：提供一种应用于多模态文本纠错系统的基于图像的文本纠错方法，首先将输入的图像与原始文本分别进行特征提取，获得图像特征以及原始文本特征，接着将图像特征以及原始文本特征通过特征拼接方式生成综合编码特征，并通过截取综合编码特征中对应于原始文本特征位置的特征，获得文本编码特征，接着通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，最后通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息，从而根据图像实现对原始文本中细粒度错误进行识别并纠正，以期大大降低在进行多模态任务时的出错率。

参照图 3，示出了本申请实施例中提供的一种基于视觉弹性掩膜的文本纠错系统示意图，通过该文本纠错系统，结合本申请所提供的图像的文本纠错方法，可以将有误或可能有误的输入文本进行进一步判断并纠错，获得正确的与图像严格匹配的输出文本。

如图所示，该文本纠错系统至少可以包括图像/文本编码模块 301、特征截取模块 302、文本特征修正模块 303、纠错向量存取器 304 以及纠错解码器 305。

其中，首先可以通过图像/文本编码模块 301 对需要进行纠错的图像以及文本分别进行编码，示例性地，可以在接收图像以及对应的输入文本“连着一个高高的桶的红色皮卡车行驶在马路上”之后，可以分别对图像以及输入文本进行编码，获得图像对应的图像编码，以及输入文本对应的文本编码。

接着可以通过特征截取模块 302 对图像编码以及文本编码先进行特征合并，再进行特征编码，获得对应的综合编码特征，并对综合编码特征进行文本特征段截取，以获得文本编码对应的文本编码特征。

然后可以将文本编码特征输入至文本特征修正模块 303 进行特征纠正，获得对应的文本纠正特征，其中，针对文本特征的纠错，文本特征修正模块 303 中设置有 3 个子模块，分别为视觉弹性掩膜子模块、信息量预测网络子模块以及相邻词汇关系预测子模块。

具体地，可以通过视觉弹性掩膜子模块进行不定长句子的纠错，如输入文本中出现错误的词所对应的字符数有 2 个，经纠错后的词所对应的字符数实际为 3 个，从而可以实现针对

输入文本的不定长纠错，使文本纠错更加准确，可信度更高。

进一步地，纠错除了可能使原句的长度变长之外，还可能导致原句的长度变短，换言之，原句中某些字符应该被删除，如从 3 个变为 2 个，则为使文本纠错系统得模型能力更全面，本申请在文本特征修正模块 303 内设计了信息量预测网络子模块，从而通过信息量预测网络子模块可以使对应位置的特征能够预测其位置字符是否包含有效信息量。

如果输入文本中存在错误，则相邻的文字可能是不连贯、不通顺的，因此，本申请还在文本特征修正模块 303 内设计了相邻词汇关系预测子模块，以对输入文本对应的特征进行文本通顺性预测。

当经过文本特征修正模块 303 中的视觉弹性掩膜子模块、信息量预测网络子模块以及相邻词汇关系预测子模块对文本编码特征进行特征纠正，获得对应的文本纠正特征后，可以将文本纠正特征输入至纠错向量存取器 304，同时可以将文本编码特征也一并输入至纠错向量存取器 304，并可以通过纠错向量存取器 304 将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征。

最后可以将目标文本特征输入至纠错解码器 305，并可以通过纠错解码器 305 采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息，示例性地，针对输入文本“连着一个高高的桶的红色皮卡车行驶在马路上”，在经过本申请的文本纠错系统进行纠错之后，可以得出正确的输出文本“连着一个高高的桶的白色皮卡车行驶在马路上”。

需要指出的是，为更好地进行辅助说明，本实施例中采用上述图 2 中基于图像的文本纠错多模态样本进行示例性说明，且本实施例中对于采用文本纠错系统结合图像的文本纠错方法相关过程描述地较为简单，仅作为实现原理的简单性说明，较为具体的实现步骤可以在下方内容中对图 4 的详细说明中获得，可以理解的是，本申请对此不作限制。

需要说明的是，本申请实施例包括但不限于上述示例，可以理解的是，本领域技术人员在本申请实施例的思想指导下，还可以根据实际需求进行设置，本申请对此不作限制。

在本申请实施例中，提供了一种多模态文本纠错系统，该文本纠错系统至少可以包括图像/文本编码模块、特征截取模块、文本特征修正模块、纠错向量存取器以及纠错解码器，本申请所提供的文本纠错系统以当前热门的 Transformer 网络结构作为骨干网络，并通过设计视觉弹性掩膜、信息量预测网络、相邻词汇关系预测等子模块实现了模型对文本错误的纠正能力，从而在基于图像的文本纠错过程中，可以将输入的图像与原始文本分别进行特征提取，获得图像特征以及原始文本特征，接着将图像特征以及原始文本特征通过特征拼接方式生成综合编码特征，并通过截取综合编码特征中对应于原始文本特征位置的特征，获得文本编码特征，接着通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，最后通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息，从而通过上述文本纠错系统，结合基于图像的文本纠错方法，实现了对原始文本中细粒度错误进行识别并纠正，大大降低了在进行多模态任务时的出错率。

参照图 4，示出了本申请实施例中提供的一种图像的文本纠错方法的步骤流程图，方法可以应用于多模态文本纠错系统，多模态文本纠错系统至少包括文本特征修正模块、纠错向量存取器以及纠错解码器，方法具体可以包括如下步骤：

步骤 401，响应于针对图像与文本的输入操作，获取输入操作对应的图像信息与原始文本信息，并分别对图像信息与原始文本信息进行特征提取，获得与图像信息对应的图像特征，以及与原始文本信息对应的原始文本特征；

Transformer，一个 N 进 N 出的结构，也就是说，每个 Transformer 单元相当于一层的 RNN（Recursive Neural Network，递归神经网络）层，可以接收一整个句子所有词作为输入，接着为句子中每个词都做出一个输出。但与 RNN 不同的是，Transformer 能够同时处理句子中的所有词，并且任意两个词之间的操作距离均为 1。

本申请在 Transformer 多模态网络结构的基础上，可以实现基于图像的文本纠错，具体地，响应于针对图像与文本的输入操作，可以获取输入操作对应的图像信息与原始文本信息，并分别对图像信息与原始文本信息进行特征提取，获得与图像信息对应的图像特征，以及与原始文本信息对应的原始文本特征，从而可以通过特征提取方式，分别提取图像与文本中的特征，以便后续过程中基于提取出的特征进行更精细的文本纠错。

示例性地，对于输入大小为 N 的图像以及大小为 M 的文本，在分别进行编码之后，获得对应的图像编码以及文本编码，接着可以分别采用现有的编码器模型进行特征提取，得到图像对应的大小为[N，d]的图像特征，以及文本对应的大小为[M，d]的文本特征，其中，“d”具体可以指特征的维度，即每个特征由多少个数组成。而对于特征提取方式，均采用当前主流的特征提取模型，如卷积神经网络（Convolutional Neural Networks，CNN）以及 BERT（Bidirectional Encoder Representation from Transformers，双向语言模型）编码器进行提取，因此不作赘述，本领域技术人员可以采用其他类似的编码器或者图像/文本模型进行特征提取，本申请对此不作限制。

步骤 402，将图像特征与原始文本特征进行特征拼接，获得综合编码特征，并根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征；

在具体的实现中，当获得图像特征以及原始文本特征之后，可以将两者进行特征拼接，以获得综合编码特征，进一步地，将图像特征与原始文本特征进行特征拼接，获得综合编码特征，具体可以为：将图像特征与原始文本特征进行特征拼接，并进行跨模态编码处理，获得综合编码特征。

示例性地，对于大小为[N，d]的图像特征，以及大小为[M，d]的文本特征，可以将两者进行特征拼接，获得大小为[N+M，d]的综合特征，并且将该综合特征输入至 Transformer 结构中进行跨模态编码，得到大小为[N+M，d]的综合编码特征。其中，跨模态编码本质是利用多模态码流间的语义，以进行相关性的联合编码，是实现跨模态通信的关键技术之一，从而通过将图像特征与原始文本特征进行特征拼接的方式，可以获得图像与文本的综合编码特征，以实现跨模态交互。

当获得综合编码特征之后，可以根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征，具体可以为：对综合编码特征中与原始文本特征位置对应的特征进行截取，获得与原始文本特征对应的文本编码特征。

即可以将大小为[N+M，d]的综合编码特征中与大小为[M，d]的文本特征对应的位置进行截取，从而得出大小为[M，d]的文本编码特征，并可以将该文本编码特征存储进纠错向量存取器中，需要说明的是，虽然文本特征与文本编码特征大小均为[M，d]，但两者包含的内容完全不同，文本特征仅代表了文本对应的特征，而文本编码特征由于是从综合编码特征中

截取出来的，因此其除了具有文本对应的特征之外，还与图像对应的特征具有相关性，因此在后续针对文本的特征纠正过程中，不能忽略与图像特征对应的相关性。

步骤 403，通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征；

在具体的实现中，通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，可以包括如下子步骤：

子步骤 4031，通过文本特征修正模块对文本编码特征进行自注意力编码，获得对应的初始自注意力向量，并对初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，其中，目标自注意力向量包含图像特征与原始文本特征的关联特征。

在自然语言处理领域中，离不开掩膜文本预测的任务，通过掩膜文本预测可以实现对遮挡或错误字的纠正过程，但采用该方法只能实现对应字符数的句子纠正，无法改变原有句子的长度，如原有句子字符数为 10，经纠正后输出的纠正句子，其字符数也对应为 10。

然而在实际使用过程中，并无法保证纠正内容的长度一定与原句长度一样，例如，如果对于一个句子“操场上一个男孩在打棒球”，如果句子中的“棒球”实际上应该是“篮球”，那么掩膜文本预测可以实现该纠错过程，但是，如果“棒球”对应的正确内容应该为“曲棍球”（即需要被纠正的内容造成了原有句子长度的变化），则掩膜文本预测在这种情况下就会失效，从而为克服该问题，本申请设计了一种弹性掩膜文本预测方式，用于实现不定长句子的纠错。

为更好地进行说明，参照图 5，示出了本申请实施例中提供的一种视觉弹性掩膜示意图，从图中可以看出，如（a）中对应的不定长字符掩膜预测中，与图像对应的输入文本为“一个男孩在打篮球他很开心”，但实际对应的文本应该为“一个男孩在打曲棍球他很开心”，从而字数为 2 的“篮球”与字数为 3 的“曲棍球”是不等长的，在这种情况下，如（b）的当前字符预测以及（c）的前置字符预测，首先，对于从“篮”（sk）到“曲棍”（tk-1，tk）而言，“棍”（tk）为“篮”（sk）的当前字符，“曲”（tk-1）为“蓝”（sk）的前置字符，假设“篮”（sk）字最终对应的特征为 768 维的文本编码特征 f，则可以采用两组全连接层分别对文本编码特征 f 进行前传，从而得到两个新的向量，假设预设字典中共有 3000 个字，则这两个向量大小均为 1000×1，之后可以采用第一个向量预测当前字符，采用另一个预测前置字符，预测方法为找出 3000 个数中最大值所在位置，将预设字典中对应的字输出即可，从而在这个过程中引入了一种可以同时预测多个字的机制，从而增强所谓的弹性，实现基于视觉弹性掩膜的文本预测，进一步实现不定长句子的纠错。

作为一种可选实施例，通过文本特征修正模块对文本编码特征进行自注意力编码，获得对应的初始自注意力向量，可以包括：将大小为[M，d]的文本编码特征输入至自注意力层中，采用公式

$$i_{emlm} = softmax\left(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{size(f)}}\right) \times (W_v \cdot f)$$

进行自注意力编码，获得对应的初始自注意力向量；其中，

i_{emlm} 为初始自注意力向量，W_q 、W_k 、W_v 均为可学习权重，f 为文本编码特征，size 表示文本编码特征的大小，T 表示转置矩阵。

softmax 函数，又称归一化指数函数，为二分类函数 sigmoid 在多分类上的推广，目的是将多分类的结果以概率形式进行展现。自注意力指注意力模型中注意力完全基于特征向量进行计算，因自注意力机制为当前图像/文本处理中较为常见的手段，因而此处不作赘述。

进一步地，对初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，可以包括：将初始自注意力向量i_{emlm}输入至两组全连接层中分别进行当前字符预测处理与前置字符预测处理，获得当前预测向量与前置预测向量，并根据当前预测向量与前置预测向量确定目标自注意力向量O_{emlm}。

其中，上述的两组全连接层为新的全连接层，需单独进行训练调优，具体地，在模型训练过程中，可以通过计算全连接层输出与真实字符之间的交叉熵以进行调优训练，从而在模型投入使用过程中，可以通过训练好的模型进行计算，从而输出目标自注意力向量O_{emlm}。

作为一种实施例，根据当前预测向量与前置预测向量确定目标自注意力向量，可以包括：采用当前预测向量对文本编码特征进行预测处理，获得文本编码特征对应的目标当前字符，进一步地，该过程具体可以为根据当前预测向量，将文本编码特征与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的当前预测概率，并将当前预测概率最大的预设字符确定为目标当前字符，从而通过对文本编码特征进行当前预测处理，可以获得对应的目标当前字符，以便后续进行前置字符预测，一定程度上增加了掩膜文本预测的弹性。

接着可以采用前置预测向量对目标当前字符进行预测处理，获得目标当前字符对应的目标前置字符，具体可以为根据前置预测向量，将目标当前字符与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的前置预测概率，并将前置预测概率最大的预设字符确定为目标前置字符，从而可以通过对目标当前字符的前置预测处理，确定对应的目标前置字符，增加了掩膜文本预测的弹性，实现了对于文本的不定长纠错。

如前述图 5 中，“篮”字对应的特征在当前字符预测的过程中负责预测出字符“棍”，在前置字符预测过程则需要预测出字符“曲”，因具体的预测示例在对图 5 的分析过程中进行了详细的描述，此处不再赘述。

然后可以将目标前置字符与目标当前字符进行拼接，输出对应的目标字符，并生成目标字符对应的目标自注意力向量，如将“蓝”对应的目标前置字符“曲”与目标当前字符“棍”进行拼接，获得目标字符“曲棍”，并生成“曲棍”对应的目标自注意力向量，以便进行后续的处理流程。

子步骤 4032，对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，有效文本信息向量表示文本编码特征中每个字符包含有效信息的概率。

由前述实施例中内容可知，纠错除了可能使原句的长度变长之外，还可能导致原句的长度变短，为使文本纠错系统得模型能力更全面，本申请在文本特征修正模块内设计了信息量预测网络子模块，从而通过信息量预测网络子模块可以使对应位置的特征能够预测其位置字符是否包含有效信息量。

在具体的实现中，对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，可以包括：采用公式

$$p_{ifo} = sigmoid(W_{iw}\left(softmax\left(\frac{(W_{iq}\cdot f)^T \times (W_{ik}\cdot f)}{\sqrt{size(f)}}\right)\times(W_{iv}\cdot f)\right) + b_{ib})$$

对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量；其中，p_{ifo}为大小为[M，1]的有效文本信息向量，表示文本编码特征中每个字符包含有效信息的概率，W_{iq}、W_{ik}、W_{iv}均为大小为[d，d]的转移矩阵权重，W_{iw}为大小为[d，1]的信息量预测权重，b_{ib}为可学习模型参数，f 为文本编码特征。需要说明的是，实施例中所涉及到的参数均为随机初始化并基于实际的数据集训练调优得出，且各个公式中各个参数符号的下标，如 q、w、k 等，仅作为便于区分各个参数，并无特殊含义，本领域技术人员可以根据实际情况或者实际需求进行设定，可以理解的是，本申请对此不作限制。

其中，sigmoid 函数是一个在生物学中常见的 S 型函数，也称为 S 型生长曲线，在信息科学中，由于 sigmoid 函数的单增以及反函数单增等性质，从而常被用作神经网络的激活函数，将变量映射至[0，1]之间。

为更好地进行说明，参照图 6，示出了本申请实施例中提供的一种相邻词汇关系预测示意图，其中，（a）表示原始特征，阴影部分表示原始特征中需要进行文本纠错的部分，（b）表示进行特征双向截取之后的前错位特征以及后错位特征，（c）表示将前错位特征与后错位特征进行向量级联之后获得的相邻特征交互向量，其中，阴影部分可以表示为文本纠错对应的相邻预测内容。具体地，本实施例将结合下述子步骤 4033 至子步骤 4034 对相邻特征交互向量的生成过程进行说明：

子步骤 4033，对文本编码特征进行双向截取，分别获得前错位特征与后错位特征，并根据前错位特征与后错位特征，生成相邻特征交互向量。

具体地，文本编码特征的大小可以为[M，d]，则进行对特征的双向截取处理之后，前错位特征与后错位特征的大小均为[M-1，d]，则进一步地，根据前错位特征与后错位特征，生成相邻特征交互向量，可以包括：将前错位特征与后错位特征进行向量级联处理，生成与文本编码特征对应的大小为[M-1，d×2]的相邻特征交互向量。从而通过对特征的双向截取处理以及对错位特征的向量级联处理，可以获得相邻特征交互向量，以方便后续对相邻预测进行概率计算，进一步提高文本纠错精确度。

子步骤 4034，对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，相邻文本信息向量表示文本编码特征中相邻字符连贯的概率。

具体地，对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，可以包括：采用公式

$$p_{nbo} = sigmoid(W_{nw}\left(softmax\left(\frac{(W_{nq}\cdot f_{nb})^T \times (W_{nk}\cdot f_{nb})}{\sqrt{size(f_{nb})}}\right)\times(W_{nv}\cdot f_{nb})\right) + b_{in})$$

对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量；

其中，p_{nbo}为大小为[M-1，1]的相邻文本信息向量，W_{nw}、W_{nq}、W_{nv}、W_{nk}均为大小为[d，d]的转移矩阵权重参数，b_{in}为大小为[1，d]的偏置向量参数，f_{nb}为相邻特征交互向量。

为保持所有向量大小的统一性，此后可以将p_{nbo}均更新为一个同样大小为[M-1，1]的

向量，由于默认句子的第一个字符无需与任何前文保证通顺，从而句子的第一个字符是一定合理的，因此可以在该向量前面新增一个 1。

子步骤 4035，采用目标自注意力向量、有效文本信息向量以及相邻文本信息向量对文本编码特征进行特征纠正，生成文本纠正特征。

对于文本特征修正模块而言，可以输出目标自注意力向量、有效文本信息向量以及相邻文本信息向量，从而可以在后续过程中采用目标自注意力向量、有效文本信息向量以及相邻文本信息向量对文本编码进行特征纠正，以生成文本纠正特征。

以上为子步骤 4031 至 4035 对应的内容，当通过文本特征修正模块对文本编码特征进行处理之后，可以将文本纠正特征输入至纠错向量存取器进行下一步的处理。

参照图 7，示出了本申请实施例中提供的一种纠错向量存取器的结构框架示意图，其中，纠错向量存取器主要可以包括三个部分，用于存储特征的特征存储空间 701、用于对特征修复进行判断的修复判断门 702、以及可以对特征进行更新的特征更新器 703，则在根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征之后，还可以将文本编码特征拆分为若干个子文本特征，并将各个子文本特征依次存储至特征存储空间 701，具体地，可以将大小为[M，d]的文本编码特征拆分为 M 条子文本特征。

在经过文本特征修正模块的处理之后，由于文本纠正特征新增了修复信息，因此可以通过修复判断门 702 判断每条文本纠正特征是否需要被修复以决定是否需要更新特征存储空间 701 中的对应向量。

在具体的实现中，通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，可以为：通过修复判断门 702 对各个子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征，进一步地，通过修复判断门 702 对各个子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征，可以为：采用公式

$$s(x_k)=\begin{cases}1,\ p_{ifok}<thresh_{ifo}\cup p_{nbok}<thresh_{nbo}\\ 0,\ p_{ifok}\geq thresh_{ifo}\cap p_{nbok}\leq thresh_{nbo}\end{cases}$$

对各个子文本特征进行修复判断，以确定是否需对子文本特征进行替换。

其中，k 表示子文本特征对应的特征序号，p_{ifok}为特征序号为 k 的子文本特征对应的有效文本信息向量，p_{nbok}为特征序号为 k 的子文本特征对应的相邻文本信息向量，$thresh_{ifo}$表示可设定信息量概率阈值，$thresh_{nbo}$表示可设定通顺概率阈值，$s(x_k)$表示是否将特征序号为 k 的子文本特征确定为需要进行特征替换的替换字文本特征。示例性地，若$s(x_k)$为 1，则表示需要对特征序号为 k 的子文本特征进行特征替换，此时可以将特征序号为 k 的子文本特征确定为需进行特征替换的替换子文本特征，若$s(x_k)$为 0，则表示不需要对特征序号为 k 的子文本特征进行特征替换，此时则不将特征序号为 k 的子文本特征确定为需进行特征替换的替换子文本特征。

接着可以通过特征更新器 703 采用文本纠正特征对至少一个替换子文本特征进行特征替换，获得各自对应的目标子文本特征，并将至少一个目标子文本特征进行特征融合，获得对应的目标文本特征，从而可以通过特征替换以及特征融合方式，实现对于文本纠正特征的更新。

在具体的实现中，通过特征更新器 703 采用文本纠正特征对至少一个替换子文本特征进

行特征替换，获得各自对应的目标子文本特征，可以为：根据文本纠正特征，采用公式

$$f_{ko} = f_k \times (1-\mu) + \left(p_{ifok} \times \theta + p_{nbok} \times (1-\theta)\right) \times \mu \times o_{emlm}$$

计算替换子文本特征对应的文本特征值，并根据文本特征值对替换子文本特征进行特征替换，获得对应的目标子文本特征。其中，

f_k为特征序号为 k 的子文本特征，o_{emlm}为目标自注意力向量，θ与μ均为大小为 0~1 的预设参数。

接着可以根据文本特征值对替换子文本特征进行特征替换，获得对应的目标子文本特征，具体地，可以为采用文本特征值通过覆盖原值方式，对替换子文本特征的原有文本特征值进行替换，获得对应的目标子文本特征，最后将特征替换完毕的目标子文本特征进行特征融合，获得对应的目标文本特征，实现对于文本纠正特征的更新。

步骤 404，通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息。

最后可以将目标文本特征从纠错向量存取器传送至纠错解码器，并通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并在进行解码之后，输出对应的目标文本信息，从而可以生成图像对应的正确文本，完成基于图像的文本纠错流程。

示例性地，本实施例中的纠错解码器可以为一个语句生成器，可采用当前主流的 GPT（Generative Pre-Training，生成式预训练语言模型）等模型实现，可以理解的是，本申请对此不作限制。

需要说明的是，本申请实施例包括但不限于上述示例，可以理解的是，本领域技术人员在本申请实施例的思想指导下，还可以根据实际需求进行设置，本申请对此不作限制。

在本申请实施例中，提供了一种应用于多模态文本纠错系统的基于图像的文本纠错方法，首先将输入的图像与原始文本分别进行特征提取，获得图像特征以及原始文本特征，接着将图像特征以及原始文本特征通过特征拼接方式生成综合编码特征，并通过截取综合编码特征中对应于原始文本特征位置的特征，获得文本编码特征，接着通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征，最后通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息，从而根据图像实现了对原始文本中细粒度错误进行识别并纠正，大大降低了在进行多模态任务时的出错率。

需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请实施例并不受所描述的动作顺序的限制，因为依据本申请实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例属于优选实施例，所涉及的动作并不一定是本申请实施例所必须的。

参照图 8，示出了本申请实施例中提供的一种图像的文本纠错装置的结构框图，应用于多模态文本纠错系统，多模态文本纠错系统至少包括文本特征修正模块、纠错向量存取器以及纠错解码器，装置具体可以包括如下模块：

特征提取模块 801，用于响应于针对图像与文本的输入操作，获取输入操作对应的图像信息与原始文本信息，并分别对图像信息与原始文本信息进行特征提取，获得与图像信息对

应的图像特征，以及与原始文本信息对应的原始文本特征；

文本编码特征生成模块 802，用于将图像特征与原始文本特征进行特征拼接，获得综合编码特征，并根据综合编码特征与原始文本特征进行特征截取，获得文本编码特征；

目标文本特征生成模块 803，用于通过文本特征修正模块对文本编码特征进行特征纠正，生成文本纠正特征，并通过纠错向量存取器将文本纠正特征与文本编码特征进行特征融合，获得目标文本特征；

文本特征替换模块 804，用于通过纠错解码器采用目标文本特征对原始文本特征进行特征替换，并输出对应的目标文本信息。

在一种可选实施例中，文本编码特征生成模块 802 包括：

目标自注意力向量生成模块，用于通过文本特征修正模块对文本编码特征进行自注意力编码，获得对应的初始自注意力向量，并对初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，目标自注意力向量包含图像特征与原始文本特征的关联特征；

有效文本信息向量生成模块，用于对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，有效文本信息向量表示文本编码特征中每个字符包含有效信息的概率；

相邻特征交互向量生成模块，用于对文本编码特征进行双向截取，分别获得前错位特征与后错位特征，并根据前错位特征与后错位特征，生成相邻特征交互向量；

相邻文本信息向量生成模块，用于对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，相邻文本信息向量表示文本编码特征中相邻字符连贯的概率；

文本纠正特征生成模块，用于采用目标自注意力向量、有效文本信息向量以及相邻文本信息向量对文本编码特征进行特征纠正，生成文本纠正特征。

在一种可选实施例中，目标自注意力向量生成模块包括：

初始自注意力向量生成模块，用于将文本编码特征输入至自注意力层中，采用公式

$$i_{emlm} = softmax\left(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{size(f)}}\right) \times (W_v \cdot f)$$

进行自注意力编码，获得对应的初始自注意力向量；其中，

W_q、W_k、W_v 均为可学习权重，f 为文本编码特征。

在一种可选实施例中，目标自注意力向量生成模块包括：

字符预测处理模块，用于将初始自注意力向量输入至两组全连接层中分别进行当前字符预测处理与前置字符预测处理，获得当前预测向量与前置预测向量；

目标自注意力向量确定子模块，用于根据当前预测向量与前置预测向量确定目标自注意力向量。

在一种可选实施例中，目标自注意力向量确定子模块包括：

目标当前字符生成模块，用于采用当前预测向量对文本编码特征进行预测处理，获得文本编码特征对应的目标当前字符；

目标前置字符生成模块，用于采用前置预测向量对目标当前字符进行预测处理，获得目标当前字符对应的目标前置字符；

目标字符输出模块，用于将目标前置字符与目标当前字符进行拼接，输出对应的目标字

符，并生成目标字符对应的目标自注意力向量。

在一种可选实施例中，目标当前字符生成模块具体用于：

根据当前预测向量，将文本编码特征与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的当前预测概率，并将当前预测概率最大的预设字符确定为目标当前字符。

在一种可选实施例中，目标前置字符生成模块具体用于包括：

根据前置预测向量，将目标当前字符与预设字典中各个预设字符进行概率匹配，获得各个预设字符对应的前置预测概率，并将前置预测概率最大的预设字符确定为目标前置字符。

在一种可选实施例中，有效文本信息向量生成模块具体用于：

采用公式

$$\boldsymbol{p_{ifo} = sigmoid(W_{iw}\left(softmax\left(\frac{(W_{iq}\cdot f)^{T}\times(W_{ik}\cdot f)}{\sqrt{size(f)}}\right)\times(W_{iv}\cdot f)\right)+b_{ib})}$$

对文本编码特征进行有效信息量预测，获得对应的有效文本信息向量；其中，$\boldsymbol{W_{iq}}$、$\boldsymbol{W_{ik}}$、$\boldsymbol{W_{iv}}$均为转移矩阵权重，$\boldsymbol{W_{iw}}$为信息量预测权重，$\boldsymbol{b_{ib}}$为可学习模型参数，f 为文本编码特征。

在一种可选实施例中，文本编码特征的大小为[M，d]，前错位特征与后错位特征的大小均为[M-1，d]，相邻特征交互向量生成模块具体用于：

将前错位特征与后错位特征进行向量级联处理，生成与文本编码特征对应的大小为[M-1，d×2]的相邻特征交互向量。

在一种可选实施例中，相邻文本信息向量生成模块具体用于：

采用公式

$$\boldsymbol{p_{nbo} = sigmoid(W_{nw}\left(softmax\left(\frac{(W_{nq}\cdot f_{nb})^{T}\times(W_{nk}\cdot f_{nb})}{\sqrt{size(f_{nb})}}\right)\times(W_{nv}\cdot f_{nb})\right)+b_{in})}$$

对相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量；其中，$\boldsymbol{W_{nw}}$、$\boldsymbol{W_{nq}}$、$\boldsymbol{W_{nv}}$、$\boldsymbol{W_{nk}}$均为转移矩阵权重参数，$\boldsymbol{b_{in}}$为偏置向量参数，$\boldsymbol{f_{nb}}$为相邻特征交互向量。

在一种可选实施例中，纠错向量存取器至少包括特征存储空间，装置还包括：

子文本特征拆分模块，用于将文本编码特征拆分为若干个子文本特征，并将各个子文本特征依次存储至特征存储空间。

在一种可选实施例中，纠错向量存取器包括修复判断门以及特征更新器，目标文本特征生成模块 803 包括：

替换子文本特征确定模块，用于通过修复判断门对各个子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征；

目标子文本特征确定模块，用于通过特征更新器采用文本纠正特征对至少一个替换子文本特征进行特征替换，获得各自对应的目标子文本特征，并将至少一个目标子文本特征进行特征融合，获得对应的目标文本特征。

在一种可选实施例中，替换子文本特征确定模块具体用于：

采用公式

$$s(x_k)=\begin{cases}1,\ p_{ifok}<thresh_{ifo}\cup p_{nbok}<thresh_{nbo}\\ 0,\ p_{ifok}\geq thresh_{ifo}\cap p_{nbok}\leq thresh_{nbo}\end{cases}$$

对各个子文本特征进行修复判断；

当$s(x_k)$为 1 时，将特征序号为 k 的子文本特征确定为需进行特征替换的替换子文本特征；其中，

k 表示子文本特征对应的特征序号，p_{ifok}为特征序号为 k 的子文本特征对应的有效文本信息向量，p_{nbok}为特征序号为 k 的子文本特征对应的相邻文本信息向量，$thresh_{ifo}$表示可设定信息量概率阈值，$thresh_{nbo}$表示可设定通顺概率阈值，$s(x_k)$表示特征序号为 k 的子文本特征是否需要进行特征替换。

在一种可选实施例中，目标子文本特征确定模块包括：

文本特征值计算模块，用于根据文本纠正特征，采用公式

$$f_{ko}=f_k\times(1-\mu)+\left(p_{ifok}\times\theta+p_{nbok}\times(1-\theta)\right)\times\mu\times o_{emlm}$$

计算替换子文本特征对应的文本特征值，并根据文本特征值对替换子文本特征进行特征替换，获得对应的目标子文本特征；其中，

f_k为特征序号为 k 的子文本特征，o_{emlm}为目标自注意力向量，θ与μ均为大小为 0~1 的预设参数。

在一种可选实施例中，文本特征值计算模块具体用于：

采用文本特征值通过覆盖原值方式，对替换子文本特征的原有文本特征值进行替换，获得对应的目标子文本特征。

在一种可选实施例中，文本编码特征生成模块 802 包括：

跨模态编码处理模块，用于将图像特征与原始文本特征进行特征拼接，并进行跨模态编码处理，获得综合编码特征。

在一种可选实施例中，文本编码特征生成模块 802 包括：

对应特征位置截取模块，用于对综合编码特征中与原始文本特征位置对应的特征进行截取，获得与原始文本特征对应的文本编码特征。

对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

另外，本申请实施例还提供了一种电子设备，包括：处理器，存储器，存储在存储器上并可在处理器上运行的计算机程序，该计算机程序被处理器执行时实现上述图像的文本纠错方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

如图 9 所示，本申请实施例还提供了一种计算机非易失性可读存储介质 901，计算机非易失性可读存储介质 901 上存储有计算机程序，计算机程序被处理器执行时实现上述图像的文本纠错方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。其中，的计算机非易失性可读存储介质 901，如只读存储器（Read-Only Memory，简称 ROM）、随机存取存储器（Random Access Memory，简称 RAM）、磁碟或者光盘等。

图 10 为实现本申请各个实施例的一种电子设备的硬件结构示意图。

该电子设备 1000 包括但不限于：射频单元 1001、网络模块 1002、音频输出单元

1003、输入单元 1004、传感器 1005、显示单元 1006、用户输入单元 1007、接口单元 1008、存储器 1009、处理器 1010、以及电源 1011 等部件。本领域技术人员可以理解，本申请实施例中所涉及的电子设备结构并不构成对电子设备的限定，电子设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。在本申请实施例中，电子设备包括但不限于手机、平板电脑、笔记本电脑、掌上电脑、车载终端、可穿戴设备、以及计步器等。

应理解的是，本申请实施例中，射频单元 1001 可用于收发信息或通话过程中，信号的接收和发送，具体的，将来自基站的下行数据接收后，给处理器 1010 处理；另外，将上行的数据发送给基站。通常，射频单元 1001 包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器、双工器等。此外，射频单元 1001 还可以通过无线通信系统与网络和其他设备通信。

电子设备通过网络模块 1002 为用户提供了无线的宽带互联网访问，如帮助用户收发电子邮件、浏览网页和访问流式媒体等。

音频输出单元 1003 可以将射频单元 1001 或网络模块 1002 接收的或者在存储器 1009 中存储的音频数据转换成音频信号并且输出为声音。而且，音频输出单元 1003 还可以提供与电子设备 1000 执行的特定功能相关的音频输出(例如，呼叫信号接收声音、消息接收声音等等)。音频输出单元 1003 包括扬声器、蜂鸣器以及受话器等。

输入单元 1004 用于接收音频或视频信号。输入单元 1004 可以包括图形处理器（Graphics Processing Unit，GPU）10041 和麦克风 10042，图形处理器 10041 对在视频捕获模式或图像捕获模式中由图像捕获装置（如摄像头）获得的静态图片或视频的图像数据进行处理。处理后的图像帧可以显示在显示单元 1006 上。经图形处理器 10041 处理后的图像帧可以存储在存储器 1009（或其它存储介质）中或者经由射频单元 1001 或网络模块 1002 进行发送。麦克风 10042 可以接收声音，并且能够将这样的声音处理为音频数据。处理后的音频数据可以在电话通话模式的情况下转换为可经由射频单元 1001 发送到移动通信基站的格式输出。

电子设备 1000 还包括至少一种传感器 1005，比如光传感器、运动传感器以及其他传感器。具体地，光传感器包括环境光传感器及接近传感器，其中，环境光传感器可根据环境光线的明暗来调节显示面板 10061 的亮度，接近传感器可在电子设备 1000 移动到耳边时，关闭显示面板 10061 和/或背光。作为运动传感器的一种，加速计传感器可检测各个方向上（一般为三轴）加速度的大小，静止时可检测出重力的大小及方向，可用于识别电子设备姿态（比如横竖屏切换、相关游戏、磁力计姿态校准）、振动识别相关功能（比如计步器、敲击）等；传感器 1005 还可以包括指纹传感器、压力传感器、虹膜传感器、分子传感器、陀螺仪、气压计、湿度计、温度计、红外线传感器等，在此不再赘述。

显示单元 1006 用于显示由用户输入的信息或提供给用户的信息。显示单元 1006 可包括显示面板 10061，可以采用液晶显示器（Liquid Crystal Display，LCD）、有机发光二极管（Organic Light-Emitting Diode, OLED）等形式来配置显示面板 10061。

用户输入单元 1007 可用于接收输入的数字或字符信息，以及产生与电子设备的用户设置以及功能控制有关的键信号输入。具体地，用户输入单元 1007 包括触控面板 10071 以及其他输入设备 10072。触控面板 10071，也称为触摸屏，可收集用户在其上或附近的触摸操

作（比如用户使用手指、触笔等任何适合的物体或附件在触控面板 10071 上或在触控面板 10071 附近的操作）。触控面板 10071 可包括触摸检测装置和触摸控制器两个部分。其中，触摸检测装置检测用户的触摸方位，并检测触摸操作带来的信号，将信号传送给触摸控制器；触摸控制器从触摸检测装置上接收触摸信息，并将它转换成触点坐标，再送给处理器 1010，接收处理器 1010 发来的命令并加以执行。此外，可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板 10071。除了触控面板 10071，用户输入单元 1007 还可以包括其他输入设备 10072。具体地，其他输入设备 10072 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆，在此不再赘述。

进一步的，触控面板 10071 可覆盖在显示面板 10061 上，当触控面板 10071 检测到在其上或附近的触摸操作后，传送给处理器 1010 以确定触摸事件的类型，随后处理器 1010 根据触摸事件的类型在显示面板 10061 上提供相应的视觉输出。可以理解的是，在一种实施例中，触控面板 10071 与显示面板 10061 是作为两个独立的部件来实现电子设备的输入和输出功能，但是在某些实施例中，可以将触控面板 10071 与显示面板 10061 集成而实现电子设备的输入和输出功能，具体此处不做限定。

接口单元 1008 为外部装置与电子设备 1000 连接的接口。例如，外部装置可以包括有线或无线头戴式耳机端口、外部电源(或电池充电器)端口、有线或无线数据端口、存储卡端口、用于连接具有识别模块的装置的端口、音频输入/输出(I/O)端口、视频 I/O 端口、耳机端口等等。接口单元 1008 可以用于接收来自外部装置的输入(例如，数据信息、电力等等)并且将接收到的输入传输到电子设备 1000 内的一个或多个元件或者可以用于在电子设备 1000 和外部装置之间传输数据。

存储器 1009 可用于存储软件程序以及各种数据。存储器 1009 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据手机的使用所创建的数据（比如音频数据、电话本等）等。此外，存储器 1009 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

处理器 1010 是电子设备的控制中心，利用各种接口和线路连接整个电子设备的各个部分，通过运行或执行存储在存储器 1009 内的软件程序和/或模块，以及调用存储在存储器 1009 内的数据，执行电子设备的各种功能和处理数据，从而对电子设备进行整体监控。处理器 1010 可包括一个或多个处理单元；优选的，处理器 1010 可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器 1010 中。

电子设备 1000 还可以包括给各个部件供电的电源 1011（比如电池），优选的，电源 1011 可以通过电源管理系统与处理器 1010 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

另外，电子设备 1000 包括一些未示出的功能模块，在此不再赘述。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素

的过程、方法、物品或者装置中还存在另外的相同要素。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例的方法。

上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，均属于本申请的保护之内。

本领域普通技术人员可以意识到，结合本申请实施例中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

以上，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1、一种图像的文本纠错方法，其特征在于，应用于多模态文本纠错系统，所述多模态文本纠错系统至少包括文本特征修正模块、纠错向量存取器以及纠错解码器，所述方法包括：

响应于针对图像与文本的输入操作，获取所述输入操作对应的图像信息与原始文本信息，并分别对所述图像信息与所述原始文本信息进行特征提取，获得与所述图像信息对应的图像特征，以及与所述原始文本信息对应的原始文本特征；

将所述图像特征与所述原始文本特征进行特征拼接，获得综合编码特征，并根据所述综合编码特征与所述原始文本特征进行特征截取，获得文本编码特征；

通过所述文本特征修正模块对所述文本编码特征进行特征纠正，生成文本纠正特征，并通过所述纠错向量存取器将所述文本纠正特征与所述文本编码特征进行特征融合，获得目标文本特征；

通过所述纠错解码器采用所述目标文本特征对所述原始文本特征进行特征替换，并输出对应的目标文本信息。

2、根据权利要求 1 所述的方法，其特征在于，所述通过所述文本特征修正模块对所述文本编码特征进行特征纠正，生成文本纠正特征，包括：

通过所述文本特征修正模块对所述文本编码特征进行自注意力编码，获得对应的初始自注意力向量，并对所述初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，所述目标自注意力向量包含所述图像特征与所述原始文本特征的关联特征；

对所述文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，所述有效文本信息向量表示所述文本编码特征中每个字符包含有效信息的概率；

对所述文本编码特征进行双向截取，分别获得前错位特征与后错位特征，并根据所述前错位特征与所述后错位特征，生成相邻特征交互向量；

对所述相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，所述相邻文本信息向量表示所述文本编码特征中相邻字符连贯的概率；

采用所述目标自注意力向量、所述有效文本信息向量以及所述相邻文本信息向量对所述文本编码特征进行特征纠正，生成文本纠正特征。

3、根据权利要求 2 所述的方法，其特征在于，所述通过所述文本特征修正模块对所述文本编码特征进行自注意力编码，获得对应的初始自注意力向量，包括：

将所述文本编码特征输入至自注意力层中，采用公式

$$i_{emlm} = softmax\left(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{size(f)}}\right) \times (W_v \cdot f)$$

进行自注意力编码，获得对应的初始自注意力向量；其中，

W_q 、 W_k 、 W_v 均为可学习权重，f 为文本编码特征。

4、根据权利要求 2 或 3 所述的方法，其特征在于，所述对所述初始自注意力向量进行字符预测处理，获得对应的目标自注意力向量，包括：

将所述初始自注意力向量输入至两组全连接层中分别进行当前字符预测处理与前置

字符预测处理，获得当前预测向量与前置预测向量；

根据所述当前预测向量与所述前置预测向量确定目标自注意力向量。

5、根据权利要求 4 所述的方法，其特征在于，所述根据所述当前预测向量与所述前置预测向量确定目标自注意力向量，包括：

采用所述当前预测向量对所述文本编码特征进行预测处理，获得所述文本编码特征对应的目标当前字符；

采用所述前置预测向量对所述目标当前字符进行预测处理，获得所述目标当前字符对应的目标前置字符；

将所述目标前置字符与所述目标当前字符进行拼接，输出对应的目标字符，并生成所述目标字符对应的目标自注意力向量。

6、根据权利要求 5 所述的方法，其特征在于，所述采用所述当前预测向量对所述文本编码特征进行预测处理，获得所述文本编码特征对应的目标当前字符，包括：

根据所述当前预测向量，将所述文本编码特征与预设字典中各个预设字符进行概率匹配，获得各个所述预设字符对应的当前预测概率，并将当前预测概率最大的预设字符确定为目标当前字符。

7、根据权利要求 6 所述的方法，其特征在于，所述采用所述前置预测向量对所述目标当前字符进行预测处理，获得所述目标当前字符对应的目标前置字符，包括：

根据所述前置预测向量，将所述目标当前字符与预设字典中各个预设字符进行概率匹配，获得各个所述预设字符对应的前置预测概率，并将前置预测概率最大的预设字符确定为目标前置字符。

8、根据权利要求 2 所述的方法，其特征在于，所述对所述文本编码特征进行有效信息量预测，获得对应的有效文本信息向量，包括：

采用公式

$$\boldsymbol{p_{ifo} = sigmoid(W_{iw}\left(softmax\left(\frac{(W_{iq}\cdot f)^{T}\times(W_{ik}\cdot f)}{\sqrt{size(f)}}\right)\times(W_{iv}\cdot f)\right)+b_{ib})}$$

对所述文本编码特征进行有效信息量预测，获得对应的有效文本信息向量；其中，W_{iq}、W_{ik}、W_{iv}均为转移矩阵权重，W_{iw}为信息量预测权重，b_{ib}为可学习模型参数，f 为文本编码特征。

9、根据权利要求 2 所述的方法，其特征在于，所述文本编码特征的大小为[M，d]，所述前错位特征与所述后错位特征的大小均为[M-1，d]，所述根据所述前错位特征与所述后错位特征，生成相邻特征交互向量，包括：

将所述前错位特征与所述后错位特征进行向量级联处理，生成与所述文本编码特征对应的大小为[M-1，d×2]的相邻特征交互向量。

10、根据权利要求 2 或 9 所述的方法，其特征在于，所述对所述相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量，包括：

采用公式

$$p_{nbo} = sigmoid(W_{nw}\left(softmax\left(\frac{(W_{nq}\cdot f_{nb})^{T}\times(W_{nk}\cdot f_{nb})}{\sqrt{size(f_{nb})}}\right)\times(W_{nv}\cdot f_{nb})\right)+b_{in})$$

对所述相邻特征交互向量进行连贯预测处理，获得对应的相邻文本信息向量；其中，

W_{nw}、W_{nq}、W_{nv}、W_{nk}均为转移矩阵权重参数，b_{in}为偏置向量参数，f_{nb}为相邻特征交互向量。

11、根据权利要求 2 所述的方法，其特征在于，所述纠错向量存取器至少包括特征存储空间，在所述根据所述综合编码特征与所述原始文本特征进行特征截取，获得文本编码特征之后，所述方法还包括：

将所述文本编码特征拆分为若干个子文本特征，并将各个所述子文本特征依次存储至所述特征存储空间。

12、根据权利要求 11 所述的方法，其特征在于，所述纠错向量存取器包括修复判断门以及特征更新器，所述通过所述纠错向量存取器将所述文本纠正特征与所述文本编码特征进行特征融合，获得目标文本特征，包括：

通过所述修复判断门对各个所述子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征；

通过所述特征更新器采用所述文本纠正特征对至少一个所述替换子文本特征进行特征替换，获得各自对应的目标子文本特征，并将至少一个所述目标子文本特征进行特征融合，获得对应的目标文本特征。

13、根据权利要求 12 所述的方法，其特征在于，所述通过所述修复判断门对各个所述子文本特征进行修复判断，确定需进行特征替换的至少一个替换子文本特征，包括：

采用公式

$$s(x_k)=\begin{cases}1,\ p_{ifok}<thresh_{ifo}\cup p_{nbok}<thresh_{nbo}\\ 0,\ p_{ifok}\geq thresh_{ifo}\cap p_{nbok}\leq thresh_{nbo}\end{cases}$$

对各个所述子文本特征进行修复判断；

当所述$s(x_k)$为 1 时，将特征序号为 k 的子文本特征确定为需进行特征替换的替换子文本特征；其中，

k 表示子文本特征对应的特征序号，p_{ifok}为特征序号为 k 的子文本特征对应的有效文本信息向量，p_{nbok}为特征序号为 k 的子文本特征对应的相邻文本信息向量，$thresh_{ifo}$表示可设定信息量概率阈值，$thresh_{nbo}$表示可设定通顺概率阈值，$s(x_k)$表示特征序号为 k 的子文本特征是否需要进行特征替换。

14、根据权利要求 12 所述的方法，其特征在于，所述通过所述特征更新器采用所述文本纠正特征对至少一个所述替换子文本特征进行特征替换，获得各自对应的目标子文本特征，包括：

根据所述文本纠正特征，采用公式

$$f_{ko}=f_k\times(1-\mu)+\left(p_{ifok}\times\theta+p_{nbok}\times(1-\theta)\right)\times\mu\times o_{emlm}$$

计算所述替换子文本特征对应的文本特征值，并根据所述文本特征值对所述替换子

文本特征进行特征替换，获得对应的目标子文本特征；其中，

f_k为特征序号为 k 的子文本特征，O_{emlm}为目标自注意力向量，θ 与 μ 均为大小为 0~1 的预设参数。

15、根据权利要求 14 所述的方法，其特征在于，所述根据所述文本特征值对所述替换子文本特征进行特征替换，获得对应的目标子文本特征，包括：

采用所述文本特征值通过覆盖原值方式，对所述替换子文本特征的原有文本特征值进行替换，获得对应的目标子文本特征。

16、根据权利要求 1 所述的方法，其特征在于，所述将所述图像特征与所述原始文本特征进行特征拼接，获得综合编码特征，包括：

将所述图像特征与所述原始文本特征进行特征拼接，并进行跨模态编码处理，获得综合编码特征。

17、根据权利要求 1 所述的方法，其特征在于，所述根据所述综合编码特征与所述原始文本特征进行特征截取，获得文本编码特征，包括：

对所述综合编码特征中与所述原始文本特征位置对应的特征进行截取，获得与所述原始文本特征对应的文本编码特征。

18、一种图像的文本纠错装置，其特征在于，应用于多模态文本纠错系统，所述多模态文本纠错系统至少包括文本特征修正模块、纠错向量存取器以及纠错解码器，所述装置包括：

特征提取模块，用于响应于针对图像与文本的输入操作，获取所述输入操作对应的图像信息与原始文本信息，并分别对所述图像信息与所述原始文本信息进行特征提取，获得与所述图像信息对应的图像特征，以及与所述原始文本信息对应的原始文本特征；

文本编码特征生成模块，用于将所述图像特征与所述原始文本特征进行特征拼接，获得综合编码特征，并根据所述综合编码特征与所述原始文本特征进行特征截取，获得文本编码特征；

目标文本特征生成模块，用于通过所述文本特征修正模块对所述文本编码特征进行特征纠正，生成文本纠正特征，并通过所述纠错向量存取器将所述文本纠正特征与所述文本编码特征进行特征融合，获得目标文本特征；

文本特征替换模块，用于通过所述纠错解码器采用所述目标文本特征对所述原始文本特征进行特征替换，并输出对应的目标文本信息。

19、一种电子设备，其特征在于，包括处理器、通信接口、存储器和通信总线，其中，所述处理器、所述通信接口以及所述存储器通过所述通信总线完成相互间的通信；

所述存储器，用于存放计算机程序；

所述处理器，用于执行存储器上所存放的程序时，实现如权利要求 1-17 任一项所述的方法。

20、一种计算机非易失性可读存储介质，其上存储有指令，当由一个或多个处理器执行时，使得所述处理器执行如权利要求 1-17 任一项所述的方法。

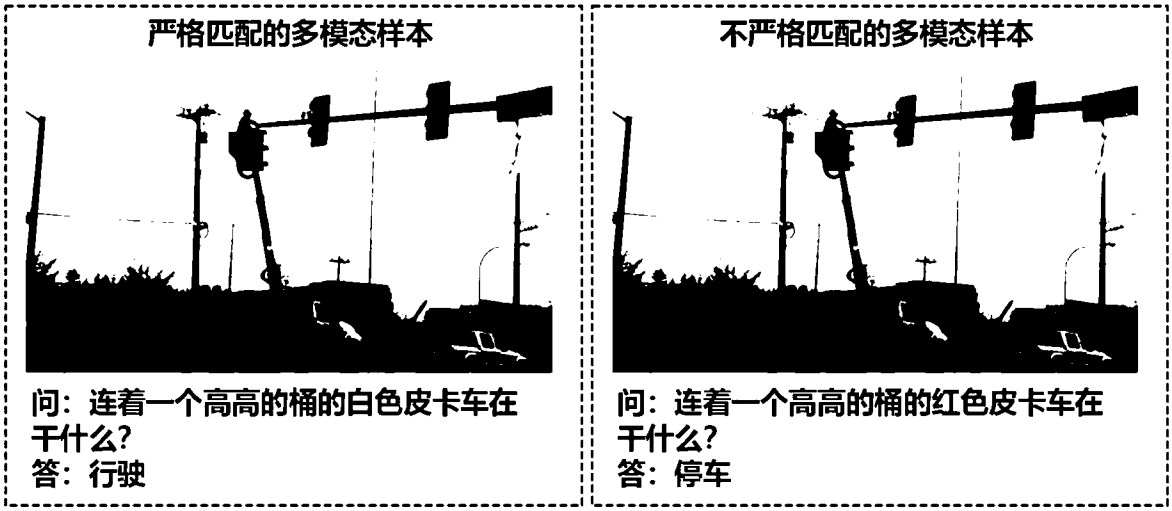

严格匹配的多模态样本
问：连着一个高高的桶的白色皮卡车在干什么?
答：行驶
不严格匹配的多模态样本
问：连着一个高高的桶的红色皮卡车在干什么?
答：停车

图 1


基于图像的文本纠错多模态样本
输入文本：连着一个高高的桶的红色皮卡车行驶在马路上。
输出文本：连着一个高高的桶的白色皮卡车行驶在马路上。

图 2

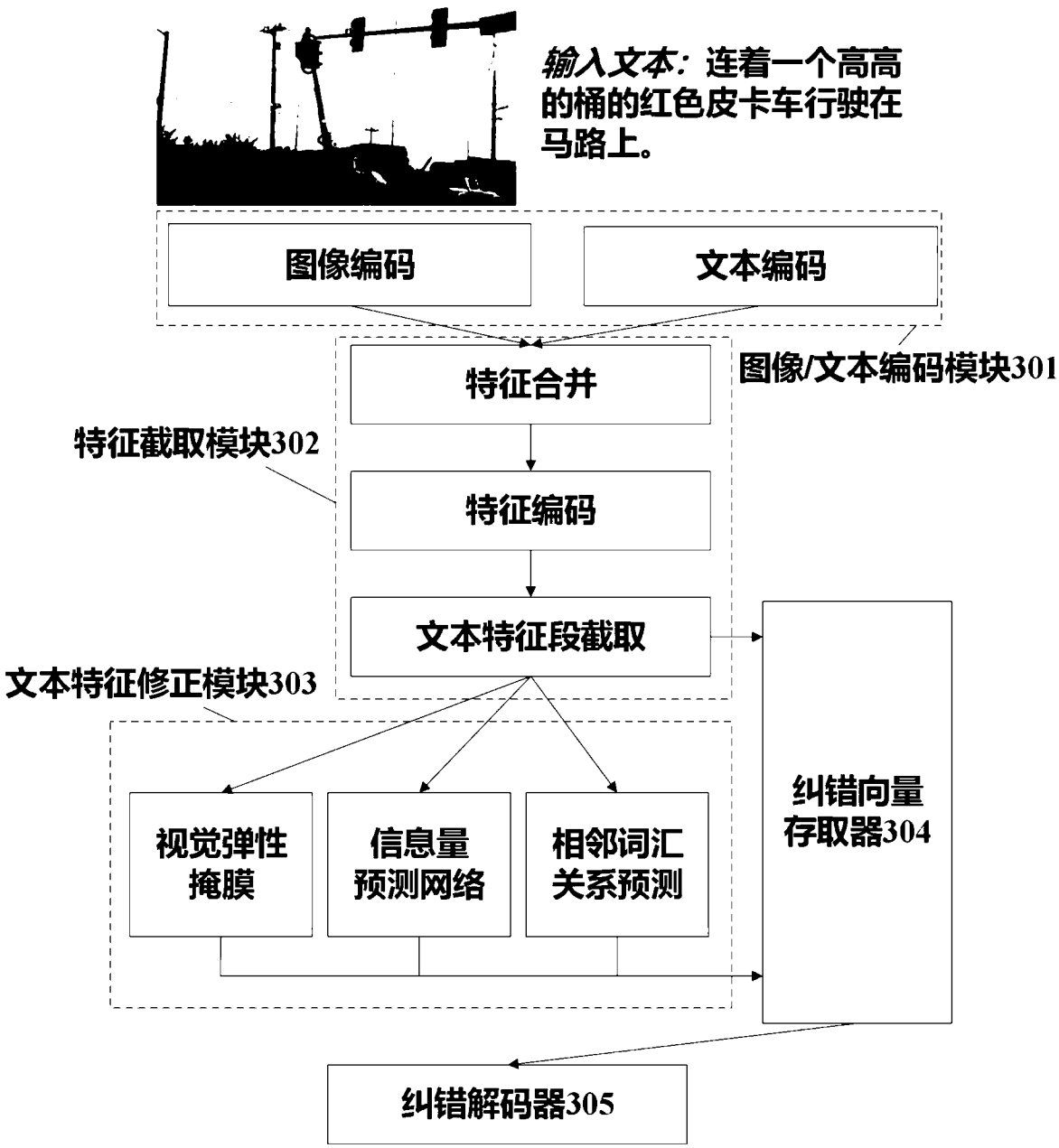

输入文本：连着一个高高的桶的红色皮卡车行驶在马路上。
图像编码
文本编码
图像/文本编码模块301
特征合并
特征截取模块302
特征编码
文本特征段截取
文本特征修正模块303
视觉弹性掩膜
信息量预测网络
相邻词汇关系预测
纠错向量存取器304
纠错解码器305
输出文本：连着一个高高的桶的白色皮卡车行驶在马路上。

图 3

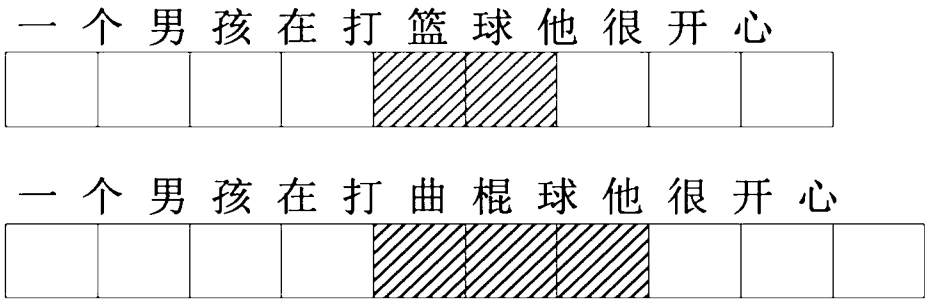
401 响应于针对图像与文本的输入操作，获取所述输入操作对应的图像信息与原始文本信息，并分别对所述图像信息与所述原始文本信息进行特征提取，获得与所述图像信息对应的图像特征，以及与所述原始文本信息对应的原始文本特征

402 将所述图像特征与所述原始文本特征进行特征拼接，获得综合编码特征，并根据所述综合编码特征与所述原始文本特征进行特征截取，获得文本编码特征

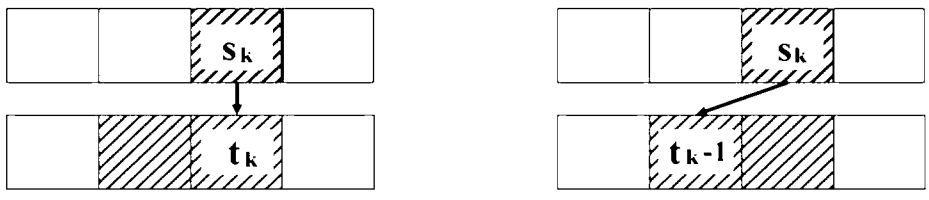
403 通过所述文本特征修正模块对所述文本编码特征进行特征纠正，生成文本纠正特征，并通过所述纠错向量存取器将所述文本纠正特征与所述文本编码特征进行特征融合，获得目标文本特征

404 通过所述纠错解码器采用所述目标文本特征对所述原始文本特征进行特征替换，并输出对应的目标文本信息

图 4



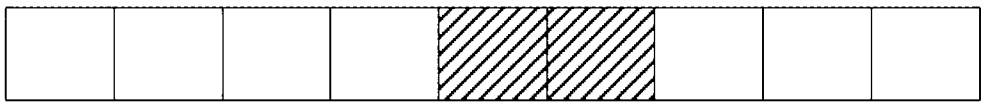
(a)不定长字符掩膜预测



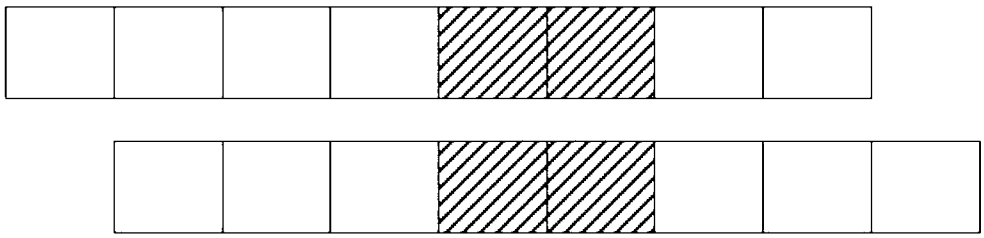
(b) 当前字符预测 (c)前置字符预测

图 5

(a) 原始特征



(b) 前错位特征和后错位特征



(c) 相邻特征交互向量

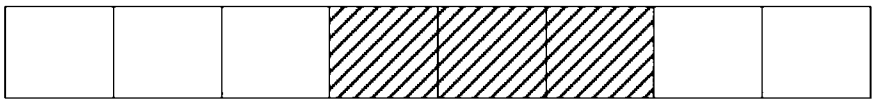


图 6

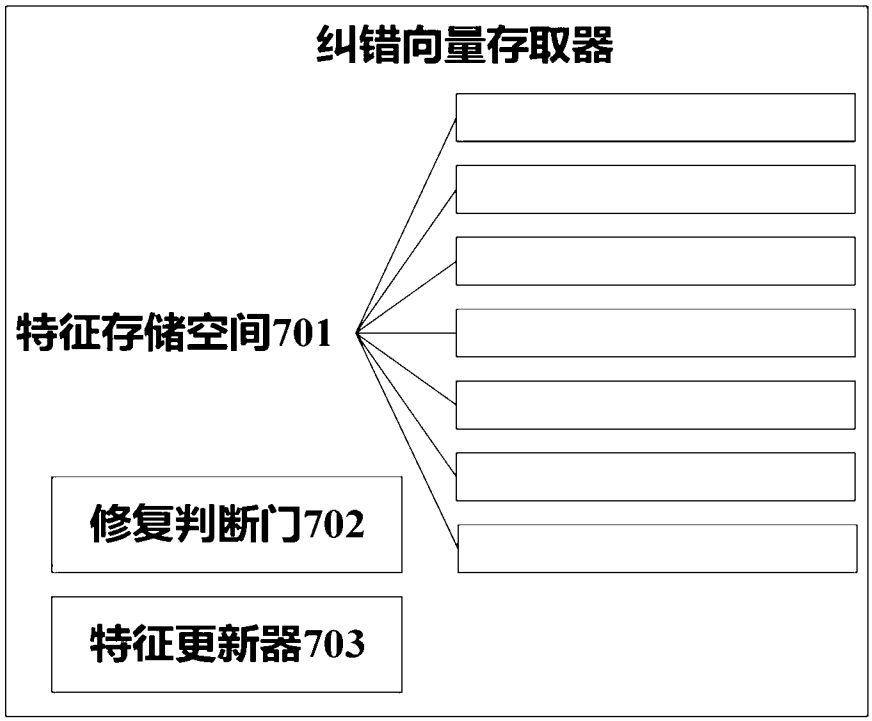


图 7

特征提取模块801

文本编码特征生成模块802

目标文本特征生成模块803

文本特征替换模块804

图 8

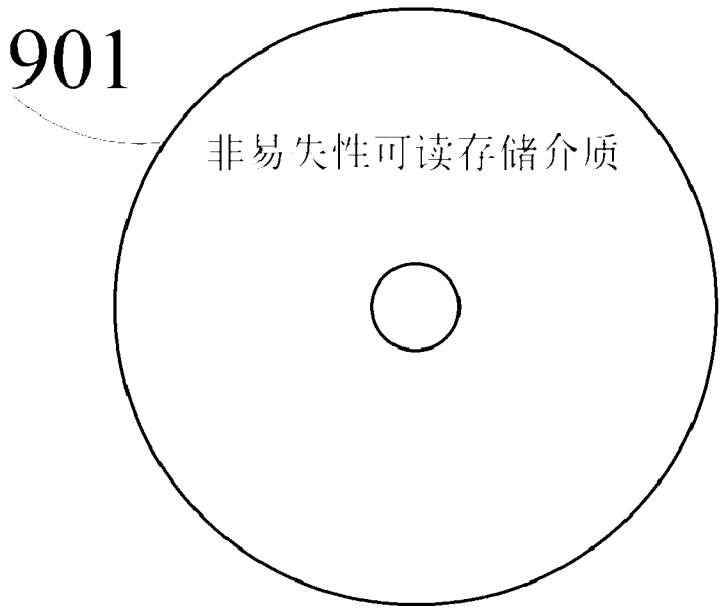


图 9

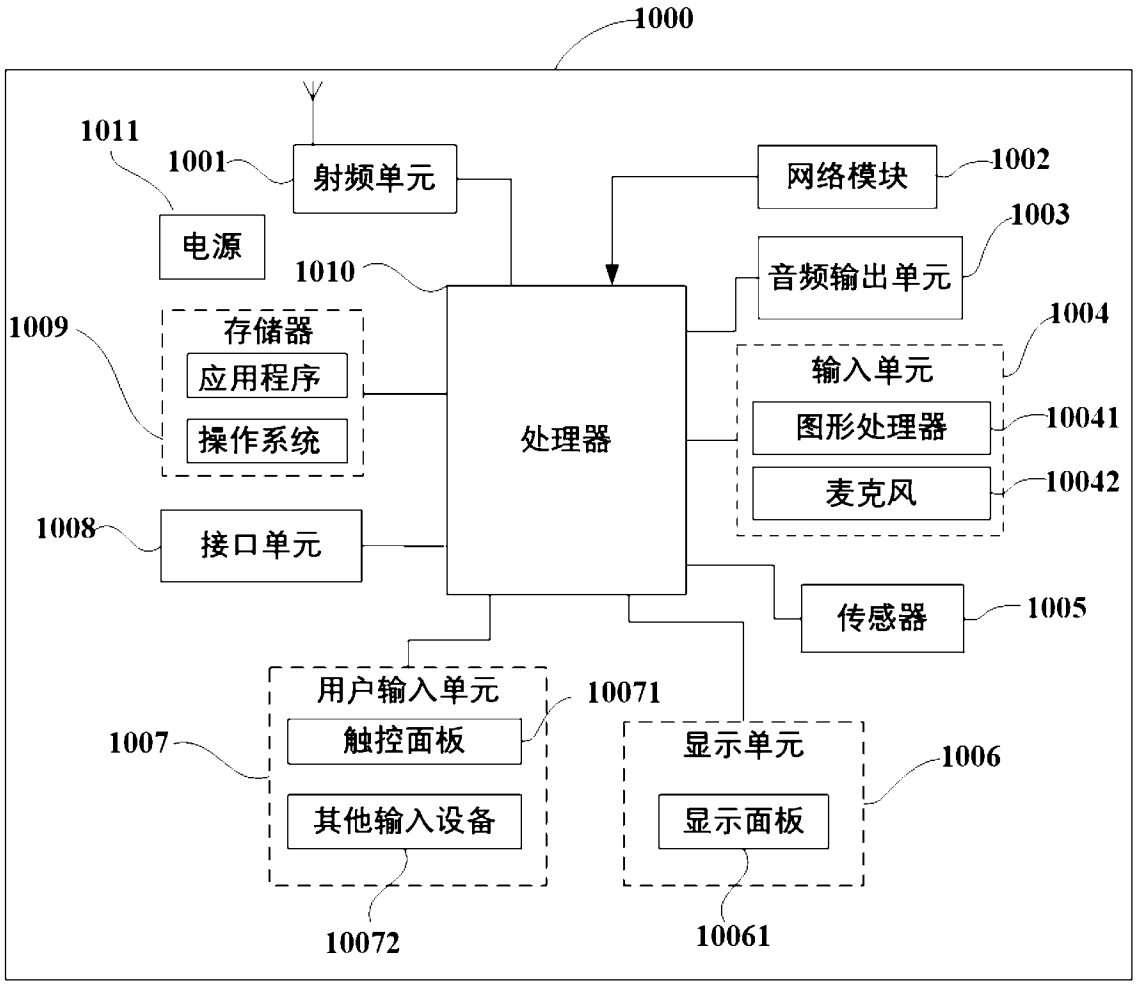

1000
1011
1001
射频单元
网络模块
1002
电源
1003
1010
音频输出单元
1009
存储器
1004
应用程序
输入单元
操作系统
处理器
图形处理器
10041
麦克风
10042
1008
接口单元
传感器
1005
用户输入单元
10071
1007
触控面板
显示单元
1006
其他输入设备
显示面板
10072
10061

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/115054

A. CLASSIFICATION OF SUBJECT MATTER

G06F40/232(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F,G06V

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 多模态, 图像, 文本, 特征提取, 纠错向量, 编码, 解码, 自注意力, 字符预测, 有效, 概率, 对齐, 错位, multimodal, image?, text, feature, extract+, error w correct+, vector, encoding, decoding, self w attention, character w predict+, effective+, probability, alignment, misalignment

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 115659959 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 31 January 2023 (2023-01-31) claims 1-20	1-20
X	CN 114462356 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 10 May 2022 (2022-05-10) description, paragraphs 0028-0067, and figures 1-5	1-3, 16-20
A	CN 112686030 A (IFLYTEK CO., LTD.) 20 April 2021 (2021-04-20) entire document	1-20
A	CN 114241279 A (ZHONGKE IFLYTEK INTERCONNECT (BEIJING) INFORMATION TECHNOLOGY CO., LTD. et al.) 25 March 2022 (2022-03-25) entire document	1-20
A	WO 2022095345 A1 (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 12 May 2022 (2022-05-12) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 November 2023	**16 November 2023**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/115054

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 115659959 A	31 January 2023	None	
CN 114462356 A	10 May 2022	None	
CN 112686030 A	20 April 2021	None	
CN 114241279 A	25 March 2022	None	
WO 2022095345 A1	12 May 2022	CN 112464993 A	09 March 2021

国际检索报告

国际申请号
PCT/CN2023/115054

A. 主题的分类

G06F40/232(2020.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F,G06V

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

VEN,CNABS,CNTXT,WOTXT,EPTXT,USTXT,CNKI,IEEE: 多模态，图像，文本，特征提取，纠错向量，编码，解码，自注意力，字符预测，有效，概率，对齐，错位，multimodal, image?, text, feature, extract+, error w correct+, vector, encoding, decoding, self w attention, character w predict+, effective+, probability, alignment, misalignment

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 115659959 A (苏州浪潮智能科技有限公司) 2023年1月31日 (2023 - 01 - 31) 权利要求1-20	1-20
X	CN 114462356 A (苏州浪潮智能科技有限公司) 2022年5月10日 (2022 - 05 - 10) 说明书第0028-0067段,附图1-5	1-3,16-20
A	CN 112686030 A (科大讯飞股份有限公司) 2021年4月20日 (2021 - 04 - 20) 全文	1-20
A	CN 114241279 A (中科讯飞互联（北京）信息科技有限公司 等) 2022年3月25日 (2022 - 03 - 25) 全文	1-20
A	WO 2022095345 A1 (苏州浪潮智能科技有限公司) 2022年5月12日 (2022 - 05 - 12) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2023年11月8日	2023年11月16日
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	受权官员 王静 电话号码 (+86) 010-53961317

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/115054

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 115659959 A	2023年1月31日	无	
CN 114462356 A	2022年5月10日	无	
CN 112686030 A	2021年4月20日	无	
CN 114241279 A	2022年3月25日	无	
WO 2022095345 A1	2022年5月12日	CN 112464993 A	2021年3月9日