

(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
2020 年 6 月 18 日 (18.06.2020)



(10) 国际公布号
WO 2020/119631 A1

(51) 国际专利分类号:
G06K 9/00 (2006.01)

(21) 国际申请号: PCT/CN2019/124008

(22) 国际申请日: 2019 年 12 月 9 日 (09.12.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811518735.8 2018 年 12 月 12 日 (12.12.2018) CN

(71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(72) 发明人: 王磊 (WANG, Lei); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong

518055 (CN)。 赖坤耀 (LAI, Kunyao); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 程俊 (CHENG, Jun); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(74) 代理人: 北京市诚辉律师事务所 (BEIJING CHENGHUI LAW FIRM); 中国北京市朝阳区朝阳北路 99 号 2 号楼 905, Beijing 100123 (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, IR, IS, 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,

(54) Title: LIGHTWEIGHT VISUAL QUESTION-ANSWERING SYSTEM AND METHOD

(54) 发明名称: 一种轻量视觉问答系统及方法

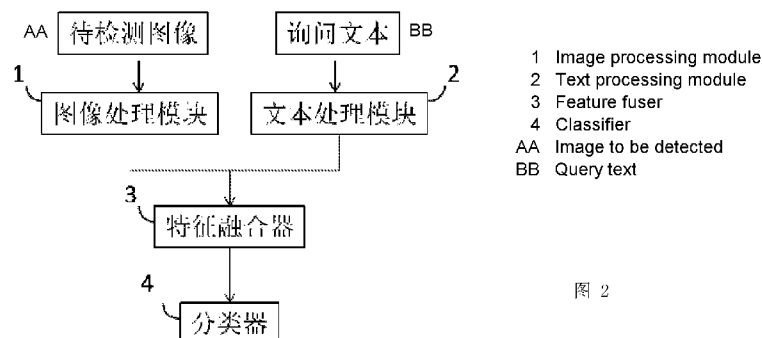


图 2

(57) Abstract: Disclosed are a lightweight visual question-answering system and method. The system comprises an image processing module (1), a text processing module (2), a feature fuser (3) and a classifier (4), wherein the image processing module (1) uses a convolutional neural network to extract an image feature and converts same into an image feature vector; the text processing module (2) extracts a text feature to form a text feature vector; and the image feature vector and the text feature vector are both sent to the feature fuser (3) for fusion, and a fusion result is sent to the classifier (4) to form a final answer. The method can reduce the complexity of a model in two aspects comprising image feature extraction and question text feature extraction, so as to transplant a question-answering system to a mobile terminal.

(57) 摘要: 一种轻量视觉问答系统和方法, 包括图像处理模块 (1)、文本处理模块 (2), 特征融合器 (3), 和分类器 (4), 其中, 所述图像处理模块 (1) 采用卷积神经网络提取图像特征, 并转化为图像特征向量; 所述文本处理模块 (2) 提取文本特征形成文本特征向量; 所述图像特征向量和所述文本特征向量均被送入所述特征融合器 (3) 进行融合, 并将融合的结果送入分类器 (4), 形成最终答案。本方法可以从图像特征提取和问题文本特征提取两方面对降低模型的复杂度, 便于将问答系统移植到移动端。

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, 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))。

一种轻量视觉问答系统及方法

技术领域

本发明涉及计算机视觉领域，尤其涉及视觉问答技术领域。

背景技术

深度学习以其强大的特征学习能力，在计算机视觉（CV）和自然语言处理（NLP）中获得广泛应用。卷积神经网络（CNN）可以抽取并压缩图像信息，多在图像处理中应用；而递归神经网络（RNN）在自然语言处理领域，尤其是在语音识别，机器翻译，语言模型与文本生成等方面取得很大的成功。

视觉问答是计算机视觉领域里最具挑战性的问题之一。视觉问答的任务就是利用计算机自动地分析图片与问题，从而对提出的问题给出回答。由于视觉问答涉及到计算机视觉和自然语言处理两个领域的内容，那么很自然的一种解决方案就是将在计算机视觉和自然语言处理中应用非常成功的卷积神经网络和递归神经网络结合构造组合模型。而其中最常使用的卷积神经网络是 Res-net 和 VGG-net，最常使用的递归神经网络是 LSTM 和 GRU。但视觉问答因为需要同时处理图像和问题，往往计算较慢，在算力不足时，比如移动端中，得出答案的时间会比较长。

在将图像信息与文本信息融合方面，Hedi Ben-younes 等在论文 MUTAN: Multimodal Tucker Fusion for Visual Question Answering 中提出了 MUTAN 融合模型，如图 1 所示，基于 Tucker 张量，分解为三个内模矩阵和核心张量，且通过约束核心张量进一步控制模型参数的数量，在训练期间能够防止过度拟合，而且能够更灵活地调整输入/输出预测。本发明基于 MUTAN 模型，使用 shuffle-net 处理图像，使用卷积神经网络 TextCNN 来处理问题语句，可以有效降低模型的复杂度，便于将问答系统移植到移动端。

发明内容

本发明的目的在于提出一种对算力要求低，便于移植到移动端的问答系统和方法。所采用的技术方案如下：

一种轻量视觉问答系统，包括图像处理模块 1、文本处理模块 2，特征融合器 3，和分类器 4，其中，所述图像处理模块 1 采用卷积神经网络提取图像特征，并转化为图像特征向量；所述文本处理模块 2 提取文本特征形成文本特征向量；所述图像特征向量和所述文本特征向量均被送入所述特征融合器 3 进行融合，并

将融合的结果送入分类器 4，形成最终答案。

进一步地，所述图像处理模块 1 采用 shuffle-net 模型提取图像特征。

进一步地，所述文本处理模块 2 采用 TextCNN 提取文本特征。

进一步地，所述文本处理模块 2 包括输入层 21、卷积层 22、池化层 23 和全连接层 24，所述输入层 21 将句子中每个单词预训练好的词向量排在一起，得到一个 $n \times k$ 矩阵，其中 n 是预先设定句子长度，不足时用 0 补充， k 为词向量的长度；所述输入层 21 与所述卷积层 22 连接，所述卷积层 22 对输入的矩阵进行卷积神经网络处理，卷积层包括多个层；卷积层 22 与所述池化层 23 连接，所述池化层 23 与所述全连接层 24 连接，最终由所述全连接层 24 得到文本的特征。

进一步地，所述特征融合器 3 采用 MUTAN 模型进行 Tucker 分解，对各分量进行融合，得到融合的结果。

进一步地，所述分类器 4 为 SoftMax 分类器，采用的损失函数为交叉熵损失函数。

进一步地，所述系统嵌入移动终端中使用。

一种轻量视觉问答方法，采用预先训练的 shuffle-net 模型提取图像特征，采用 TextCNN 提取文本特征，然后利用 MUTAN 模型将所述图像特征与所述文本特征融合，得到答案。

进一步地，所述融合的方法为，将从文本特征提取器得到的向量 q ，图像特征提取器得到的向量 v 融合，得到张量 T ，并对 T 进行 Tucker 分解，得到参数核心张量 τ_c ，以及三个内模矩阵 W_q 、 W_v 、 W_o ，计算获得融合的特征 y ：

$$y = ((\tau_c \times_1 (q^T W_q)) \times_2 (v^T W_v)) \times_3 W_o,$$

其中， $\times_i$ 代表向量在第 i 维同张量相乘，将 y 送入分类器中即可得到最终答案。

进一步地，在移动终端中应用该方法。

本发明的轻量视觉问答系统和方法的优势在于：从图像特征提取和问题文本特征提取两方面对降低模型的复杂度，便于将问答系统移植到移动端。

附图说明

图 1 为 MUTAN 融合模型架构图。

图 2 为轻量视觉问答系统框图。

图 3 为文本处理模块结构图。

具体实施方式

如图 2 所示，本发明的轻量化视觉问答系统，包括图像处理模块 1、文本处理模块 2，特征融合器 3，和分类器 4，其中，待检测图像进入所述图像处理模块 1 处理，图像处理模块 1 采用卷积神经网络提取图像特征，并转化为图像特征向量；询问文本进入所述文本处理模块 2 处理，在文本处理模块 2 中对文本的特征进行提取，形成文本特征向量；所述图像特征向量和所述文本特征向量均被送入特征融合器 3 进行融合，并将融合的结果送入分类器 4，形成最终答案。

在图像处理模块 1 中选取预训练好的 shuffle-net 模型提取特征，其中，shuffle-net 最后一个卷积层的特征将送入特征融合器中。

文本处理模块 2 采用 TextCNN 处理询问文本，其结构如图 3 所示，在输入层 21 中，将预先句子中每个单词对应预训练好的词向量排在一起，得到一个 $n \times k$ 矩阵。其中 n 是预先设定句子长度，不足时用 0 补充， k 为词向量的长度。然后按照卷积神经网络处理，即输入层 21 连接卷积层 22，在多个卷积层 22 中对特征进行提取。卷积层 22 与所述池化层 23 连接，池化层 23 中采用最大池化法的将特征池化，池化层 23 与所述全连接层 24 连接，最终由所述全连接层 24 得到文本的特征。

在特征融合器 3 中，采用采用 MUTAN 模型进行 Tucker 分解，对各分量进行融合，得到融合的结果。MUTAN 融合模型是由 Hedi Ben-younes 等在论文 MUTAN: Multimodal Tucker Fusion for Visual Question Answering 中提出的，其流程如图 1 所示。

从文本特征提取器得到的向量 q ，图像特征提取器得到的向量 v 融合，得到张量 T ，并对 T 进行 Tucker 分解，得到参数核心张量 τ_c ，以及三个内模矩阵 W_q 、 W_v 、 W_o ，计算获得融合的特征 y ：

$$y = ((\tau_c \times_1 (q^T W_q)) \times_2 (v^T W_v)) \times_3 W_o,$$

其中， $\times_i$ 代表向量在第 i 维同张量相乘，将 y 送入分类器中即可得到最终答案。

所述 Tucker 张量分解的方式为： $T = ((\tau_c \times_1 W_q) \times_2 W_v) \times_3 W_o$ ，其中， T

由文本特征向量 q 和图像特征向量 v 融合得到。

分类器 4 为 SoftMax 层，训练选取的损失函数为交叉熵损失，表示为：

$$\text{Loss} = \sum_{i=1}^{|A|} y_i \log \hat{y}_i,$$

其中 y_i 代表真实的答案索引， $\hat{y}_i$ 是预测的答案索引。 $i=1 \dots |A|$ ， $|A|$ 是不同的答案个数。

一种轻量视觉问答方法，采用预先训练的 shuffle-net 模型提取图像特征，采用 TextCNN 提取文本特征，然后利用 MUTAN 模型将所述图像特征与所述文本特征融合，得到答案。

经过实验证明，使用 shuffle-net 作为视觉问答的图像特征提取器，TextCNN 作为文本特征提取器，可以有效降低模型的复杂度，便于将问答系统移植到移动端。

权 利 要 求 书

1、一种轻量视觉问答系统，其特征在于，包括图像处理模块（1）、文本处理模块（2），特征融合器（3），和分类器（4），其中，所述图像处理模块（1）采用卷积神经网络提取图像特征，并转化为图像特征向量；所述文本处理模块（2）提取文本特征形成文本特征向量；所述图像特征向量和所述文本特征向量均被送入所述特征融合器（3）进行融合，并将融合的结果送入分类器（4），形成最终答案。

2、根据权利要求 1 所述的轻量视觉问答系统，其特征在于，所述图像处理模块（1）采用 shuffle-net 模型提取图像特征。

3、根据权利要求 1 或 2 所述的轻量视觉问答系统，其特征在于，所述文本处理模块（2）采用 TextCNN 提取文本特征。

4、根据权利要求 3 所述的轻量视觉问答系统，其特征在于，所述文本处理模块（2）包括输入层（21）、卷积层（22）、池化层（23）和全连接层（24），所述输入层（21）将句子中每个单词预训练好的词向量排在一起，得到一个 $n \times k$ 矩阵，其中 n 是预先设定句子长度，不足时用 0 补充， k 为词向量的长度；所述输入层（21）与所述卷积层（22）连接，所述卷积层（22）对输入的矩阵进行卷积神经网络处理，卷积层包括多个层；卷积层（22）与所述池化层（23）连接，所述池化层（23）与所述全连接层（24）连接，最终由所述全连接层（24）得到文本的特征。

5、根据权利要求 1-4 任一项所述的轻量视觉问答系统，其特征在于，所述特征融合器（3）采用 MUTAN 模型进行 Tucker 分解，对各分量进行融合，得到融合的结果。

6、根据权利要求 1-5 任一项所述的轻量视觉问答系统，其特征在于，所述分类器（4）为 SoftMax 分类器，采用的损失函数为交叉熵损失函数。

7、根据权利要求 1-6 任一项所述的轻量视觉问答系统，其特征在于，所述

系统嵌入移动终端中使用。

8、一种轻量视觉问答方法，其特征在于，采用预先训练的 shuffle-net 模型提取图像特征，采用 TextCNN 提取文本特征，然后利用 MUTAN 模型将所述图像特征与所述文本特征融合，得到答案。

9、根据权利要求 8 所述的轻量视觉问答方法，其特征在于，所述融合的方法为，将从文本特征提取器得到的向量 q ，图像特征提取器得到的向量 v 融合，得到张量 T ，并对 T 进行 Tucker 分解，得到参数核心张量 τ_c ，以及三个内模矩阵 W_q 、 W_v 、 W_o ，计算获得融合的特征 y ：

$$y = ((\tau_c \times_1 (q^T W_q)) \times_2 (v^T W_v)) \times_3 W_o,$$

其中， $\times_i$ 代表向量在第 i 维同张量相乘，将 y 送入分类器中即可得到最终答案。

10、根据权利要求 8 所述的轻量视觉问答方法，其特征在于，在移动终端中应用该方法。

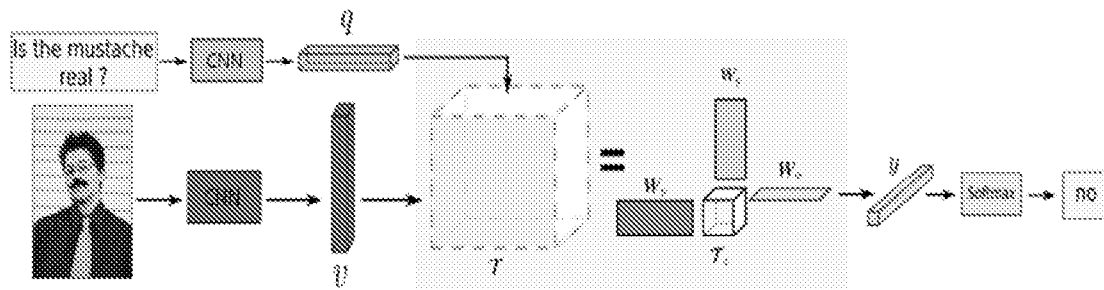


图 1

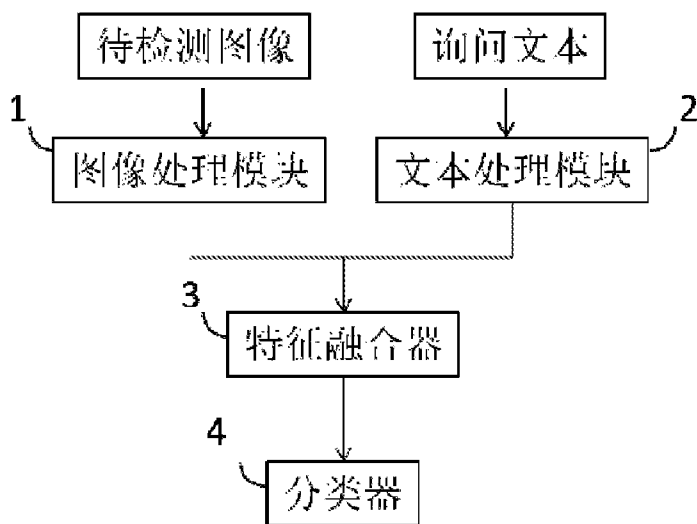


图 2

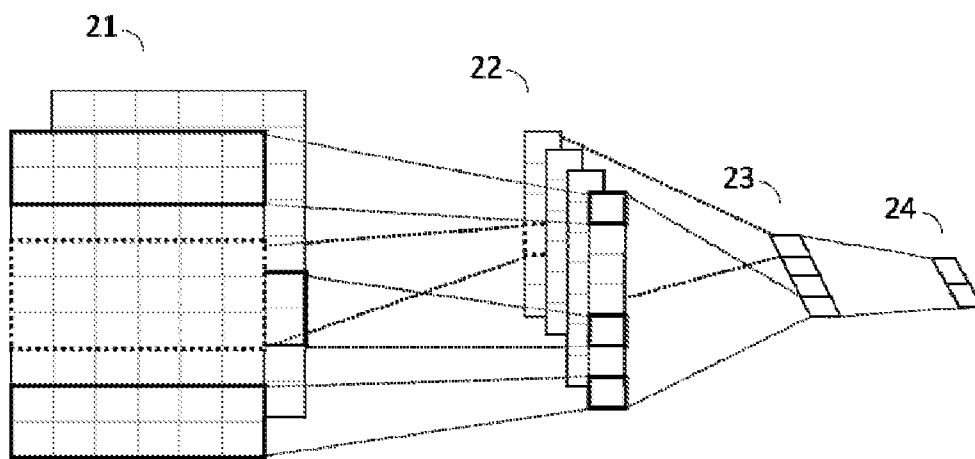


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/124008

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/00(2006.01)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)

G06K

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)

CNPAT, EPODOC, WPI, CNKI, IEEE: 图像, 文本, 特征, 提取, 向量, 融合, 分类, 模型, 视觉, 问答, image, text, feature, character, extract, vector, fusion, classify, model, visual, question

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109784163 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 21 May 2019 (2019-05-21) claims 1-10	1-10
X	CN 108256549 A (BEIJING DAJIA INTERCONNECTION INFORMATION TECHNOLOGY CO., LTD.) 06 July 2018 (2018-07-06) description, paragraphs [0002]-[0014], [0027]-[0071]	1-10
X	CN 107066583 A (HUAQIAO UNIVERSITY) 18 August 2017 (2017-08-18) claims 1-5, description, paragraphs [0033]-[0047]	1-10
A	CN 107679582 A (SHENZHEN VISION TECHNOLOGY CO., LTD.) 09 February 2018 (2018-02-09) entire document	1-10
A	CN 108763325 A (BEIJING DAJIA INTERCONNECTION INFORMATION TECHNOLOGY CO., LTD.) 06 November 2018 (2018-11-06) entire document	1-10
A	CN 106777185 A (ZHEJIANG UNIVERSITY) 31 May 2017 (2017-05-31) entire document	1-10



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

“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

03 March 2020

Date of mailing of the international search report

12 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/124008

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017124432 A1 (BAIDU USA, LLC) 04 May 2017 (2017-05-04) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/124008

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109784163	A	21 May 2019	None			
CN	108256549	A	06 July 2018	None			
CN	107066583	A	18 August 2017	None			
CN	107679582	A	09 February 2018	None			
CN	108763325	A	06 November 2018	None			
CN	106777185	A	31 May 2017	None			
US	2017124432	A1	04 May 2017	EP	3166049	A1	10 May 2017
				CN	106649542	A	10 May 2017
				JP	2017091525	A	25 May 2017
				KR	20180038937	A	17 April 2018

国际检索报告

国际申请号

PCT/CN2019/124008

A. 主题的分类

G06K 9/00 (2006.01) 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)

G06K

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))

CNPAT, EPDOC, WPI, CNKI, IEEE: 图像, 文本, 特征, 提取, 向量, 融合, 分类, 模型, 视觉, 问答, image, text, feature, character, extract, vector, fusion, classify, model, visual, question

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 109784163 A (中国科学院深圳先进技术研究院) 2019年 5月 21日 (2019 - 05 - 21) 权利要求1-10	1-10
X	CN 108256549 A (北京达佳互联信息技术有限公司) 2018年 7月 6日 (2018 - 07 - 06) 说明书第[0002]-[0014], [0027]-[0071]段	1-10
X	CN 107066583 A (华侨大学) 2017年 8月 18日 (2017 - 08 - 18) 权利要求1-5, 说明书第[0033]-[0047]段	1-10
A	CN 107679582 A (深圳市唯特视科技有限公司) 2018年 2月 9日 (2018 - 02 - 09) 全文	1-10
A	CN 108763325 A (北京达佳互联信息技术有限公司) 2018年 11月 6日 (2018 - 11 - 06) 全文	1-10
A	CN 106777185 A (浙江大学) 2017年 5月 31日 (2017 - 05 - 31) 全文	1-10
A	US 2017124432 A1 (BAIDU USA, LLC) 2017年 5月 4日 (2017 - 05 - 04) 全文	1-10

☐ 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

“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

2020年 3月 3日

国际检索报告邮寄日期

2020年 3月 12日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

王莹

电话号码 86-(10)-53961411

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/124008

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	109784163	A	2019年 5月 21日	无			
CN	108256549	A	2018年 7月 6日	无			
CN	107066583	A	2017年 8月 18日	无			
CN	107679582	A	2018年 2月 9日	无			
CN	108763325	A	2018年 11月 6日	无			
CN	106777185	A	2017年 5月 31日	无			
US	2017124432	A1	2017年 5月 4日	EP	3166049	A1	2017年 5月 10日
				CN	106649542	A	2017年 5月 10日
				JP	2017091525	A	2017年 5月 25日
				KR	20180038937	A	2018年 4月 17日