

(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 12 月 10 日 (10.12.2020)



(10) 国际公布号
WO 2020/244183 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2019/123406

(22) 国际申请日: 2019 年 12 月 5 日 (05.12.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910487643.6 2019年6月5日 (05.06.2019) CN

(71) 申请人: 北京三快在线科技有限公司 (BEIJING SANKUAI ONLINE TECHNOLOGY CO., LTD) [CN/CN]; 中国北京市海淀区北四环西路 9 号 2106-030, Beijing 100080 (CN)。

(72) 发明人: 刘宇达 (LIU, Yuda); 中国北京市海淀区北四环西路 9 号 2106-030, Beijing 100080 (CN)。

(74) 代理人: 北京曼威知识产权代理有限公司 (BEIJING MAVAIPI INTELLECTUAL PROPERTY LAW FIRM); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 411 室, Beijing 100085 (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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(54) Title: DATA ANNOTATION

(54) 发明名称: 数据标注

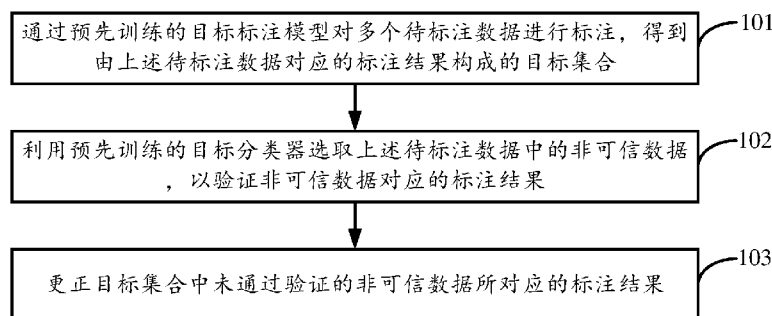


图 1

- 101 Annotate, by means of a pre-trained target annotation model, a plurality of pieces of data to be annotated to obtain a target set composed of annotation results corresponding to the data to be annotated
- 102 Use a pre-trained target classifier to select untrusted data in the data to be annotated so as to verify annotation results corresponding to the untrusted data
- 103 Correct an annotation result, which corresponds to the untrusted data and does not pass verification, in the target set

(57) Abstract: A data annotation method and apparatus, and an electronic device. The method comprises: annotating, by means of a pre-trained target annotation model, a plurality of pieces of data to be annotated to obtain a target set composed of annotation results corresponding to the data to be annotated (101); using a pre-trained target classifier to select untrusted data in the data to be annotated so as to verify annotation results corresponding to the untrusted data (102); and correcting an annotation result, which corresponds to the untrusted data and does not pass verification, in the target set (103).

(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: 一种数据标注的方法、装置及电子设备, 所述方法包括: 通过预先训练的目标标注模型对多个待标注数据进行标注, 得到由上述待标注数据对应的标注结果构成的目标集合 (101); 利用预先训练的目标分类器选取上述待标注数据中的非可信数据, 以验证非可信数据对应的标注结果 (102); 更正目标集合中未通过验证的非可信数据所对应的标注结果 (103)。

数据标注

技术领域

[01]本申请涉及机器学习技术领域，特别涉及机器学习过程中的数据标注。

背景技术

[02]随着人工智能技术的不断发展,人工智能技术在各个领域中已经得到了广泛地应用。人工智能技术通常涉及到机器学习,在机器学习的过程中,对训练样本数据的需求量巨大,并且需要对大量的训练样本数据进行标注。目前来说,通常需要人工对大量的训练样本数据进行标注,使得标注工作过分依赖于人,并且工作量巨大。从而耗费了大量的人力资源,并且标注的效率较低。

发明内容

[03]为了解决上述技术问题之一,本申请提供一种数据标注的方法、装置及电子设备。

[04]根据本申请实施例的第一方面,提供一种数据标注的方法,包括:

[05]通过预先训练的目标标注模型对多个待标注数据进行标注,得到由所述待标注数据对应的标注结果构成的目标集合;

[06]利用预先训练的目标分类器选取所述待标注数据中的非可信数据,以验证所述非可信数据对应的标注结果;

[07]更正所述目标集合中未通过验证的非可信数据所对应的标注结果。

[08]根据本申请实施例的第二方面,提供一种数据标注的装置,包括:

[09]标注模块,用于通过预先训练的目标标注模型对多个待标注数据进行标注,得到由所述待标注数据对应的标注结果构成的目标集合;

[10]选取模块,用于利用预先训练的目标分类器选取所述待标注数据中的非可信数据,以验证所述非可信数据对应的标注结果;

[11]更正模块,用于更正所述目标集合中未通过验证的非可信数据所对应的标注结果。

[12]根据本申请实施例的第三方面,提供一种计算机可读存储介质,所述存储介质存储有计算机程序,所述计算机程序被处理器执行时实现上述第一方面中任一项所述的方法。

[13]根据本申请实施例的第四方面，提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述第一方面中任一项所述的方法。

[14]应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

[15]此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

[16]图 1 是本申请根据一示例性实施例示出的一种数据标注的方法的流程图；

[17]图 2 是本申请根据一示例性实施例示出的另一种数据标注的方法的流程图；

[18]图 3 是本申请根据一示例性实施例示出的另一种数据标注的方法的流程图；

[19]图 4 是本申请根据一示例性实施例示出的一种数据标注的装置的框图；

[20]图 5 是本申请根据一示例性实施例示出的另一种数据标注的装置的框图；

[21]图 6 是本申请根据一示例性实施例示出的另一种数据标注的装置的框图；

[22]图 7 是本申请根据一示例性实施例示出的一种电子设备的结构示意图。

具体实施方式

[23]这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[24]在本申请使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本申请。在本申请和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[25]应当理解，尽管在本申请可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不

脱离本申请范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[26]如图 1 所示，图 1 是根据一示例性实施例示出的一种数据标注的方法的流程图，该方法可以应用于终端设备中，也可以应用于服务器中。该方法包括以下步骤：

[27]在步骤 101 中，通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合。

[28]在本实施例中，首先可以通过预先训练的目标标注模型对多个待标注数据进行标注，得到目标集合，该目标集合由上述待标注数据对应的标注结果构成。其中，待标注数据可以是图像类型的数据，也可以是声音类型的数据，还可以是文本类型的数据等，可以理解，待标注数据还可以是其它任意类型的数据，本申请对待标注数据的具体类型方面不限定。

[29]在本实施例中，对待标注数据进行标注可以是将待标注数据中的标注目标进行标记，并针对标记设置相应的标签（例如，类别标签、属性标签、ID（identifier）标签等）。以图像类型的待标注数据为例，对图像类型的数据进行标注可以是将图像类型的数据中的目标物体（即标注目标）用标注框标记下来，并设置目标物体的标签（如，物体类别、物体属性、物体 ID 等）。

[30]在本实施例中，目标标注模型可以是一个模型，也可以是多个结构不同的模型，可选地，如果待标注数据是图像类型的数据，则目标标注模型可以是神经网络类型的模型。在初始训练目标标注模型时，可以采用人工的方式标注训练样本数据，然后，利用人工标注的训练样本数据训练得到目标标注模型。在后续使用过程中，还可以不断对目标标注模型进行优化和更新，使目标标注模型标注的准确度更高。

[31]在本实施例中，如果目标标注模型包括多个结构不同的单元模型，且待标注数据为待标注的图像数据，则可以通过如下方式标注任一待标注的图像数据：首先，可以分别采用每个单元模型在该图像数据中进行标注；然后，基于每个单元模型在该图像数据中进行标注的结果，利用 NMS（non maximum suppression，非极大值抑制）算法确定该图像数据中的一个或多个标注目标，并确定每个标注目标对应的标签，从而完成对该图像数据的标注。其中，标注目标为该图像数据中需要标注的对象。由于采用多个结构不同的单元模型进行标注，可以提高标注的准确度。

[32]在步骤 102 中，利用预先训练的目标分类器选取上述待标注数据中的非可信数据，以验证非可信数据对应的标注结果。

[33]在本实施例中，可以利用预先训练的目标分类器从上述待标注数据中选取非可信数据，以验证非可信数据对应的标注结果。其中，非可信数据可以是具有较强迷惑性、特征不显著、难以分辨、难以标注且容易被标注错误的的数据。以图像类型的数据为例，非可信数据可以是图像模糊，或者图像中具有较大干扰物体，或者图像中目标物体不明显的图像数据等。以声音类型的数据为例，非可信数据可以是环境噪声较大，或者目标声音太小的声音数据等。以文本类型的数据为例，非可信数据可以是语义特征模糊的文本数据等。

[34]在本实施例中，目标分类器可以是预先训练好的分类器，本领域中已知的以及将来可能出现的任何能够应用于此的分类器都可以应用于本申请，本申请对目标分类器的具体类型方面不限定。在初始训练目标分类器时，可以采用人工分类筛选的方式，将训练样本数据分为可信样本数据和非可信样本数据。然后，利用人工分类筛选的训练样本数据训练得到目标分类器。在后续使用过程中，还可以不断对目标分类器进行优化和更新，使目标分类器分类的准确度更高。

[35]在本实施例中，利用目标分类器从上述待标注数据中选取容易被标注错误的非可信数据，然后，对非可信数据对应的标注结果进行验证，从而判断该非可信数据对应的标注结果是否正确。可以采用人工的方式进行验证，也可以采用其它任意合理的方式进行验证，可以理解，本申请对验证的具体方式方面不限定。针对任意非可信数据，如果确定该非可信数据对应的标注结果是错误的，则该非可信数据未通过验证，如果确定该非可信数据对应的标注结果是正确的，则该非可信数据通过验证。

[36]在步骤 103 中，更正目标集合中未通过验证的非可信数据所对应的标注结果。

[37]在本实施例中，对于选取出的非可信数据，如果该非可信数据通过验证，则不进行任何操作。如果该非可信数据未通过验证，则需要重新对该非可信数据进行正确的标注，从而得到正确的标注结果。并利用该非可信数据所对应的正确的标注结果，对目标集合中该非可信数据所对应的错误的标注结果进行替换，从而更正目标集合中未通过验证的非可信数据所对应的标注结果。

[38]在本实施例中，可以预先建立标注数据库，标注数据库用于存放标注好的数据，利用标注数据库存放的数据，进行机器学习的训练。可以将经过更正的目标集合作为标注

好的数据，存放入标注数据库中。

[39]本申请的上述实施例提供的标注数据的方法，通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合，利用预先训练的目标分类器选取上述待标注数据中的非可信数据，以验证非可信数据对应的标注结果，并更正目标集合中未通过验证的非可信数据所对应的标注结果。由于本实施例利用预先训练的目标标注模型对待标注数据进行标注后，又利用目标分类器筛选出有可能标注错误的待标注数据进行抽检，并对错误的标注结果进行更正。使得标注工作无需完全依赖于人，节省了大量的人力资源，提高了标注的效率。同时，能够更有针对性的验证被标注数据的结果，提高了标注的准确度。

[40]如图2所示，图2是根据一示例性实施例示出的另一种标注数据的方法的流程图，该实施例描述了更新目标分类器的过程，该方法可以应用于终端设备中，也可以应用于服务器中。该方法包括以下步骤：

[41]在步骤201中，通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合。

[42]在步骤202中，利用预先训练的目标分类器选取上述待标注数据中的非可信数据，以验证非可信数据对应的标注结果。

[43]在步骤203中，更正目标集合中未通过验证的非可信数据所对应的标注结果。

[44]在步骤204中，将未通过验证的非可信数据确定为第一正样本，并将通过验证的非可信数据确定为第一负样本。

[45]在本实施例中，由于非可信数据具有较强迷惑性、特征不显著、难以分辨、难以标注且容易被标注错误。因此，如果目标分类器筛选出的非可信数据被标注错误，则说明目标分类器对该非可信数据的分类较为准确。如果目标分类器筛选出的非可信数据被标注正确，则说明目标分类器对该非可信数据的分类不够准确。所以，将未通过验证的非可信数据（即被标注错误的非可信数据）作为第一正样本，并将通过验证的非可信数据（即被标注正确的非可信数据）确定为第一负样本。

[46]在步骤205中，利用第一正样本和第一负样本对目标分类器进行更新。

[47]在本实施例中，可以利用第一正样本和第一负样本重新训练目标分类器，以对目标分类器进行优化更新，从而使目标分类器能够筛选出更合适的非可信数据。

[48]需要说明的是，对于与图 1 实施例中相同的步骤，在上述图 2 实施例中不再进行赘述，相关内容可参见图 1 实施例。

[49]本申请的上述实施例提供的数据标注的方法，通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合，利用预先训练的目标分类器选取上述待标注数据中的非可信数据，以验证非可信数据对应的标注结果，更正目标集合中未通过验证的非可信数据所对应的标注结果。将未通过验证的非可信数据确定为第一正样本，并将通过验证的非可信数据确定为第一负样本，利用第一正样本和第一负样本对目标分类器进行更新。不仅节省了大量的人力资源，提高了标注的效率和准确度，而且由于本实施例在标注过程中，不断对目标分类器进行优化更新，使得目标分类器筛选的非可信数据更有针对性。

[50]如图 3 所示，图 3 是根据一示例性实施例示出的另一种数据标注的方法的流程图，该实施例详细描述了更新目标标注模型的过程，该方法可以应用于终端设备中，也可以应用于服务器中。该方法包括以下步骤：

[51]在步骤 301 中，通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合。

[52]在步骤 302 中，利用预先训练的目标分类器选取上述待标注数据中的非可信数据，以验证非可信数据对应的标注结果。

[53]在步骤 303 中，更正目标集合中未通过验证的非可信数据所对应的标注结果。

[54]在步骤 304 中，将经过更正的目标集合存放入预先建立的标注数据库。

[55]在步骤 305 中，利用该标注数据库对目标标注模型进行更新。

[56]在本实施例中，可以预先建立标注数据库，标注数据库用于存放标注好的数据，利用标注数据库存放的数据，进行机器学习的训练。可以将经过更正的目标集合作为标注好的数据，存放入标注数据库中。

[57]在本实施例中，可以利用标注数据库中的数据作为训练样本，重新训练目标标注模型，以对目标标注模型进行优化更新，从而使目标标注模型对待标注数据进行标注的结果更加准确。

[58]需要说明的是，对于与图 1 和图 2 实施例中相同的步骤，在上述图 3 实施例中不再进行赘述，相关内容可参见图 1 和图 2 实施例。

[59] 本申请的上述实施例提供的标注数据的方法，通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合，利用预先训练的目标分类器选取上述待标注数据中的非可信数据，以验证非可信数据对应的标注结果，更正目标集合中未通过验证的非可信数据所对应的标注结果，将经过更正的目标集合存放入预先建立的标注数据库，并利用该标注数据库对目标标注模型进行更新。不仅节省了大量的人力资源，提高了标注的效率，而且由于本实施例在每次标注结束后，利用标注数据库对目标标注模型进行优化更新，使得目标标注模型不断完善，进一步提高了标注的准确度。

[60] 在一些可选实施方式中，目标标注模型可以通过如下方式训练：迭代执行对标注模型的更新操作，直至满足停止条件，将经过迭代更新后的标注模型作为目标标注模型。

[61] 在本实施例中，在初始训练标注模型时，可以采用人工的方式标注训练样本数据，并利用人工标注的训练样本数据训练标注模型。当训练的标注模型对待标注数据的标注结果达到准确度阈值后，可以采用本实施例的方式继续训练标注模型。具体来说，可以迭代执行对标注模型的更新操作，直至满足停止条件（例如，满足目标函数收敛，或者迭代次数超过预设次数等），将经过迭代更新后的标注模型作为目标标注模型。其中，上述更新操作可以包括以下步骤：

[62] 步骤 a：利用当前的目标分类器选取样本数据中的非可信样本。

[63] 在本实施例中，首先可以利用当前训练好的目标分类器选取样本数据中的非可信样本，该非可信样本可以是具有较强迷惑性、特征不显著、难以分辨、难以标注且容易被标注错误的样本数据。

[64] 步骤 b：获取非可信样本对应的人工标注结果，作为第一样本结果。

[65] 步骤 c：利用该第一样本结果更新当前的标注模型。

[66] 在本实施例中，由于非可信样本具有较强迷惑性，并且特征不显著，因此，利用标注模型对非可信样本进行标注更容易出现错误。所以，可以首先采用人工的方式对非可信样本进行标注，从而得到非可信样本对应的人工标注结果，作为第一样本结果。该第一样本结果的准确率较高，可以利用该第一样本结果不断更新当前的标注模型，使标注模型不断完善，直至满足停止条件，得到目标标注模型。从而有助于提高目标标注模型标注的准确度。

[67] 在另一些可选实施方式中，上述更新操作还可以包括以下步骤：

[68] 步骤 d: 利用当前的目标分类器选取样本数据中的可信样本。

[69] 在本实施例中, 还可以利用当前的目标分类器选取样本数据中的可信样本, 该可信样本与非可信样本相反。该可信样本可以是特征显著、容易分辨、容易标注且不易被标注错误的样本数据。

[70] 步骤 e: 利用当前的标注模型对可信样本进行标注, 得到第二样本结果。

[71] 步骤 f: 确定针对第二样本结果进行验证的验证结果。

[72] 在本实施例中, 由于可信样本特征显著且容易分辨, 因此, 利用标注模型对可信样本进行标注不容易出现错误。可以利用当前的标注模型对可信样本进行标注, 得到第二样本结果。可以针对第二样本结果进行验证, 以确定当前的标注模型对可信样本进行标注的结果是否正确。

[73] 步骤 g: 基于该验证结果, 将未通过验证的第二样本结果确定为第二负样本, 并将通过验证的第二样本结果确定为第二正样本。

[74] 在本实施例中, 如果当前的目标分类器筛选出的可信数据被标注模型标注错误, 则说明目标分类器对该可信数据的分类不够准确。如果当前的目标分类器筛选出的可信数据被标注模型标注正确, 则说明当前的目标分类器对该可信数据的分类比较准确。所以, 可以将未通过验证的第二样本结果 (即被标注错误的可信数据) 确定为第二负样本, 并将通过验证的第二样本结果 (即被标注正确的可信数据) 确定为第二正样本。

[75] 步骤 h: 利用第二正样本和第二负样本对当前的目标分类器进行更新。

[76] 在本实施例中, 可以利用第二正样本和第二负样本对当前的目标分类器进行优化更新, 从而使目标分类器对可信数据和非可信数据的筛选更加准确。

[77] 应当注意, 尽管在上述实施例中, 以特定顺序描述了本申请方法的操作, 但是, 这并非要求或者暗示必须按照该特定顺序来执行这些操作, 或是必须执行全部所示的操作才能实现期望的结果。相反, 流程图中描绘的步骤可以改变执行顺序。附加地或备选地, 可以省略某些步骤, 将多个步骤合并为一个步骤执行, 和/或将一个步骤分解为多个步骤执行。

[78] 与前述数据标注的方法实施例相对应, 本申请还提供了数据标注的装置的实施例。

[79] 如图 4 所示, 图 4 是本申请根据一示例性实施例示出的一种数据标注的装置框图, 该装置可以包括: 标注模块 401, 选取模块 402 和更正模块 403。

[80]其中，标注模块 401，用于通过预先训练的目标标注模型对多个待标注数据进行标注，得到由上述待标注数据对应的标注结果构成的目标集合。

[81]选取模块 402，用于利用预先训练的目标分类器选取待标注数据中的非可信数据，以验证非可信数据对应的标注结果。

[82]更正模块 403，用于更正目标集合中未通过验证的非可信数据所对应的标注结果。

[83]如图 5 所示，图 5 是本申请根据一示例性实施例示出的另一种数据标注的装置框图，该实施例在前述图 4 所示实施例的基础上，该装置还可以进一步包括：确定模块 404 和第一更新模块 405。

[84]其中，确定模块 404，用于将未通过验证的非可信数据确定为第一正样本，并将通过验证的非可信数据确定为第一负样本。

[85]第一更新模块 405，用于利用第一正样本和第一负样本对目标分类器进行更新。

[86]在一些可选实施方式中，上述目标标注模型可以通过如下方式训练：迭代执行对标注模型的更新操作，直至满足停止条件，将经过迭代更新后的标注模型作为所述目标标注模型。其中，上述更新操作可以包括：利用当前的目标分类器选取样本数据中的非可信样本，获取非可信样本对应的人工标注结果，作为第一样本结果，并利用第一样本结果更新当前的标注模型。

[87]在另一些可选实施方式中，上述更新操作还可以包括：利用当前的目标分类器选取样本数据中的可信样本，利用当前的标注模型对可信样本进行标注，得到第二样本结果，确定针对第二样本结果进行验证的验证结果，基于该验证结果，将未通过验证的第二样本结果确定为第二负样本，并将通过验证的第二样本结果确定为第二正样本，并利用第二正样本和第二负样本对当前的目标分类器进行更新。

[88]在另一些可选实施方式中，上述目标标注模型可以包括多个结构不同的单元模型，待标注数据为待标注的图像数据。

[89]可以通过如下方式标注任一待标注的图像数据：分别采用每个单元模型在该图像数据中进行标注，基于每个单元模型在该图像数据中进行标注的结果，利用非极大值抑制 NMS 算法确定该图像数据中的一个或多个标注目标，并确定每个标注目标对应的标签。

[90]如图 6 所示，图 6 是本申请根据一示例性实施例示出的另一种数据标注的装置框图，该实施例在前述图 4 所示实施例的基础上，该装置还可以进一步包括：存放模块 406 和

第二更新模块 407。

[91]其中，存放模块 406，用于将经过更正的目标集合存放入预先建立的标注数据库中。

[92]第二更新模块 407，用于利用标注数据库对目标标注模型进行更新。

[93]应当理解，上述装置可以预先设置在终端设备或服务器中，也可以通过下载等方式而加载到终端设备或服务器中。上述装置中的相应模块可以与终端设备或服务器中的模块相互配合以实现数据标注的方案。

[94]对于装置实施例而言，由于其基本对应于方法实施例，所以相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例仅仅是示意性的，其中所述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模块来实现本申请方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

[95]本申请实施例还提供了一种计算机可读存储介质，该存储介质存储有计算机程序，计算机程序可用于执行上述图 1 至图 3 任一实施例提供的标注数据的方法。

[96]对应于上述的标注数据的方法，本申请实施例还提出了图 7 所示的根据本申请的一示例性实施例的电子设备的示意结构图。请参考图 7，在硬件层面，该电子设备包括处理器、内部总线、网络接口、内存以及非易失性存储器，当然还可能包括其他业务所需要的硬件。处理器从非易失性存储器中读取对应的计算机程序到内存中然后运行，在逻辑层面上形成标注数据的装置。当然，除了软件实现方式之外，本申请并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[97]本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本申请的真正范围和精神由下面的权利要求指出。

[98]应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

权利要求书

1、一种数据标注的方法，包括：

通过预先训练的目标标注模型对多个待标注数据进行标注，得到由所述待标注数据对应的标注结果构成的目标集合；

利用预先训练的目标分类器选取所述待标注数据中的非可信数据，以验证所述非可信数据对应的标注结果；

更正所述目标集合中未通过验证的非可信数据所对应的标注结果。

2、根据权利要求 1 所述的方法，还包括：

将未通过验证的非可信数据确定为第一正样本，并将通过验证的非可信数据确定为第一负样本；

利用所述第一正样本和所述第一负样本对所述目标分类器进行更新。

3、根据权利要求 1 所述的方法，所述目标标注模型通过如下方式训练：

迭代执行对标注模型的更新操作，直至满足停止条件；

将经过迭代更新后的标注模型作为所述目标标注模型；

其中，所述更新操作包括：

利用当前的目标分类器选取样本数据中的非可信样本；

获取所述非可信样本对应的人工标注结果，作为第一样本结果；

利用所述第一样本结果更新当前的标注模型。

4、根据权利要求 3 所述的方法，所述更新操作还包括：

利用当前的目标分类器选取所述样本数据中的可信样本；

利用当前的标注模型对所述可信样本进行标注，得到第二样本结果；

确定针对所述第二样本结果进行验证的验证结果；

基于所述验证结果，将未通过验证的第二样本结果确定为第二负样本，并将通过验证的第二样本结果确定为第二正样本；

利用所述第二正样本和所述第二负样本对当前的目标分类器进行更新。

5、根据权利要求 1 所述的方法，所述目标标注模型包括多个结构不同的单元模型；
所述待标注数据为待标注的图像数据；

通过如下方式标注任一待标注的图像数据：

分别采用每个所述单元模型在该图像数据中进行标注；

基于每个所述单元模型在该图像数据中进行标注的结果，利用非极大值抑制 NMS 算法确定该图像数据中的一个或多个标注目标；并

确定每个所述标注目标对应的标签。

6、根据权利要求 1 所述的方法，还包括：

将经过更正的目标集合放入预先建立的标注数据库中；

利用所述标注数据库对所述目标标注模型进行更新。

7、一种数据标注的装置，包括：

标注模块，用于通过预先训练的目标标注模型对多个待标注数据进行标注，得到由所述待标注数据对应的标注结果构成的目标集合；

选取模块，用于利用预先训练的目标分类器选取所述待标注数据中的非可信数据，以验证所述非可信数据对应的标注结果；

更正模块，用于更正所述目标集合中未通过验证的非可信数据所对应的标注结果。

8、根据权利要求 7 所述的装置，还包括：

确定模块，用于将未通过验证的非可信数据确定为第一正样本，并将通过验证的非可信数据确定为第一负样本；

第一更新模块，用于利用所述第一正样本和所述第一负样本对所述目标分类器进行更新。

9、一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述权利要求 1-6 中任一项所述的方法。

10、一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述权利要求 1-6 中任一项所述的方法。

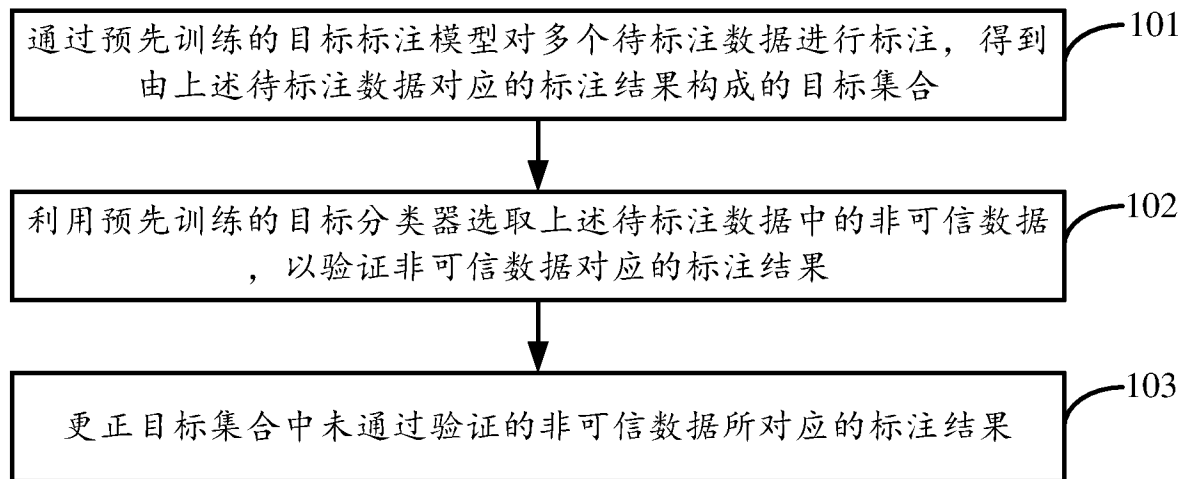


图 1

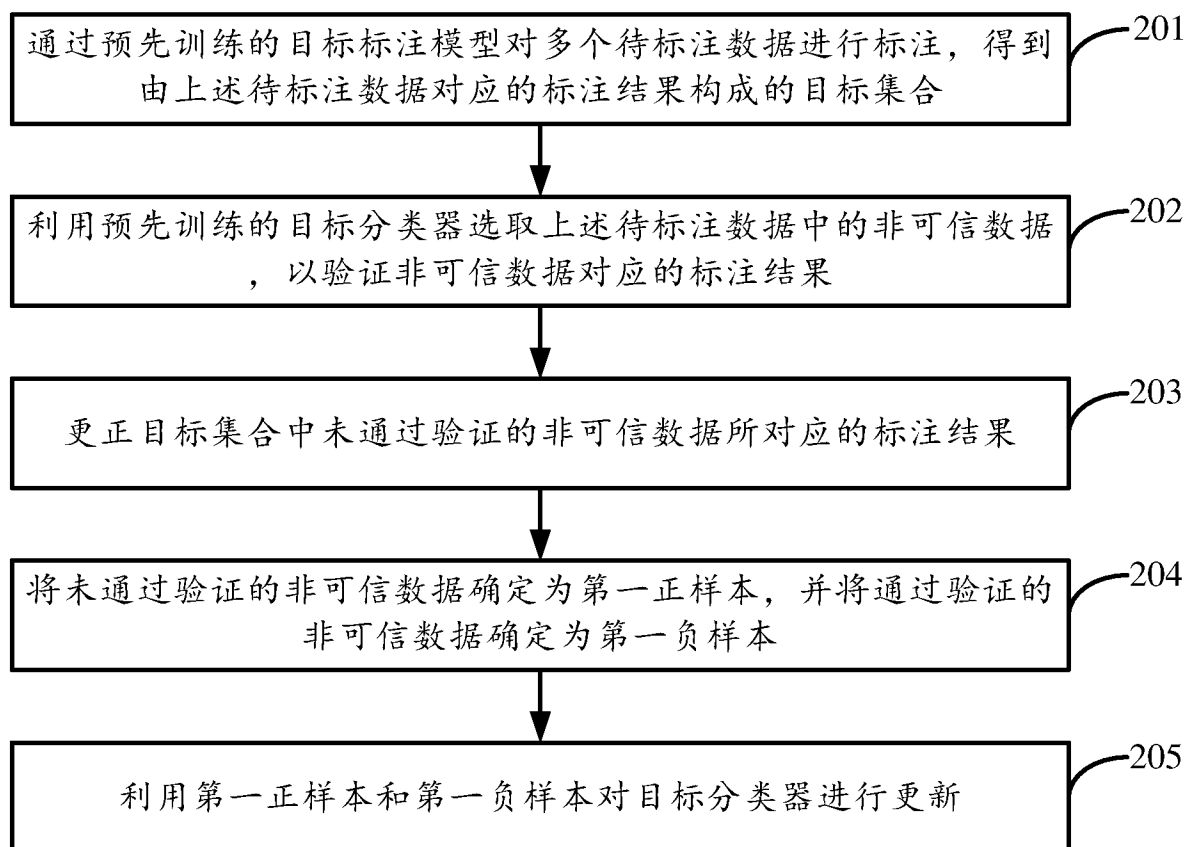


图 2

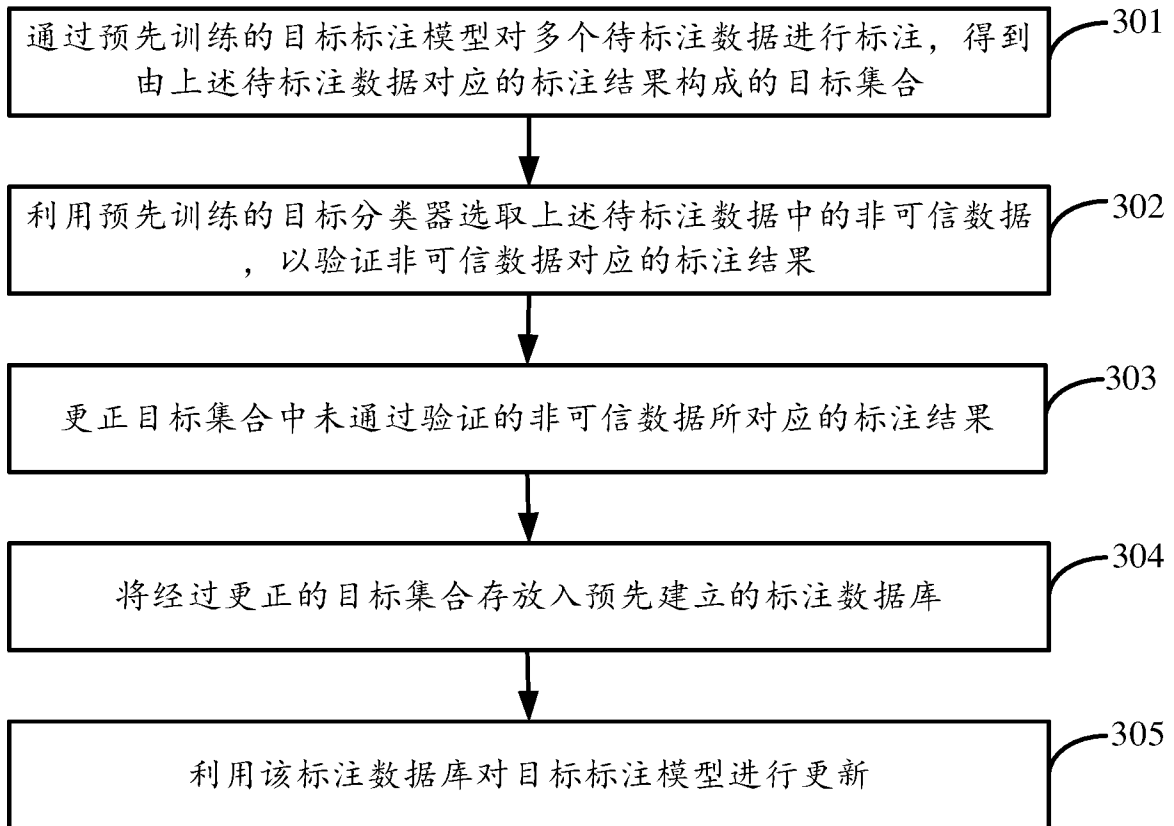


图 3

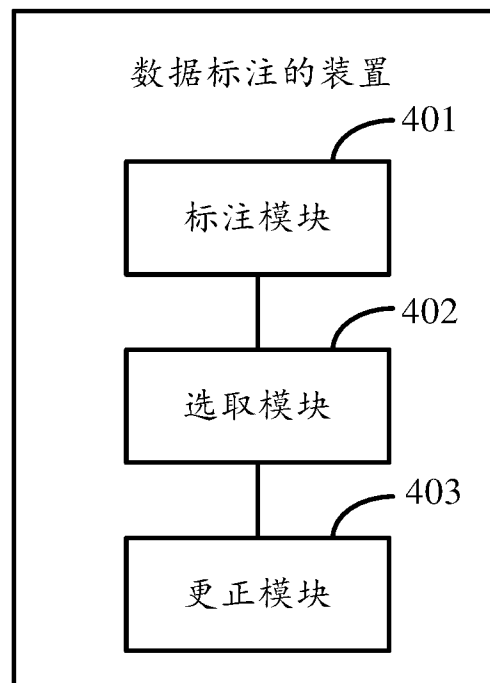


图 4

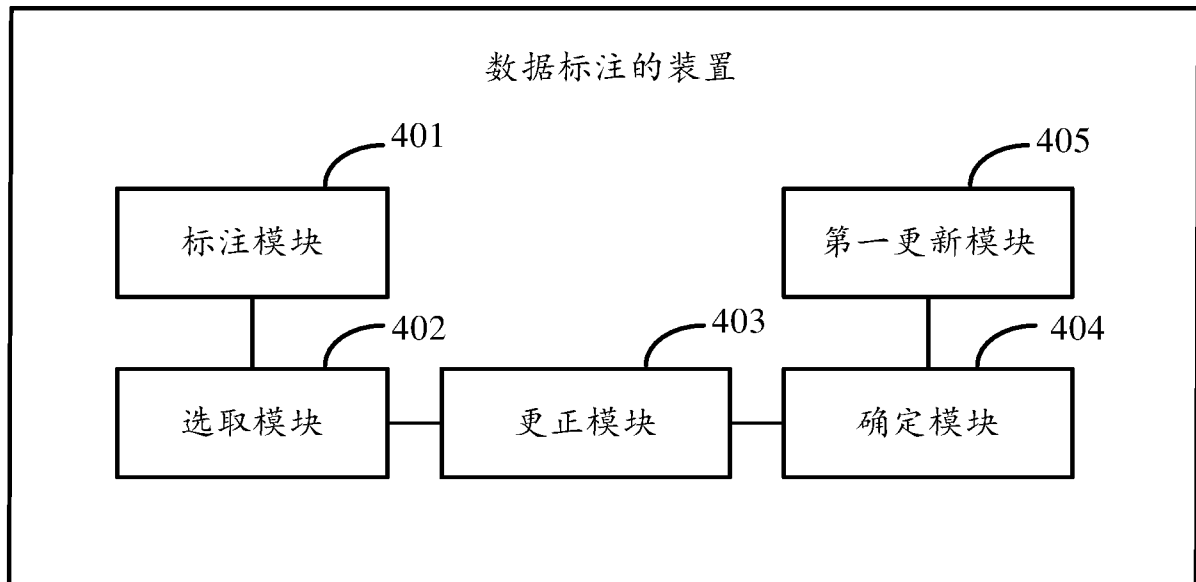


图 5

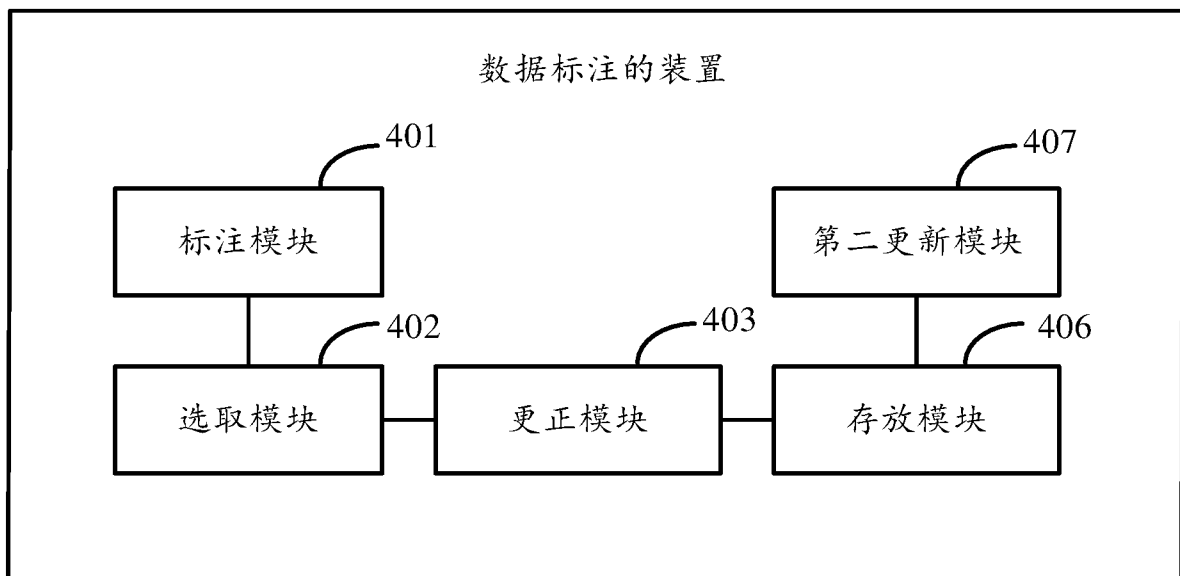


图 6

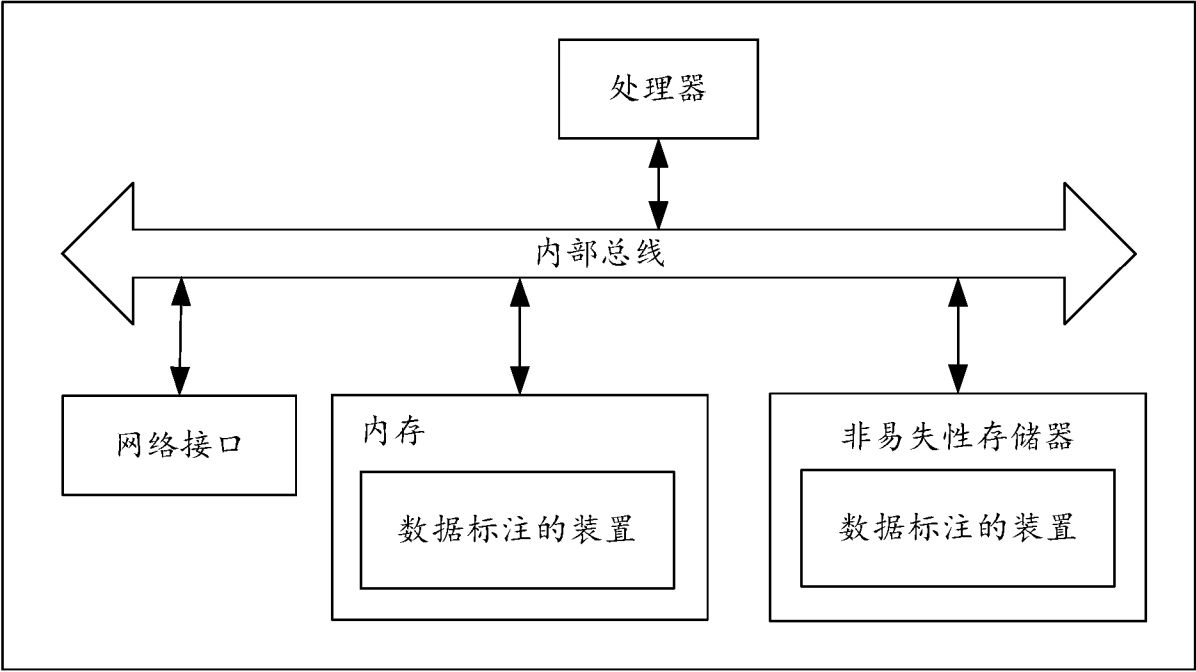


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/123406

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(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; G06N

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; CNKI; WPI; EPODOC: 数据, 图像, 标注, 样本, 训练, 模型, 目标, 分类器, 验证, 更正, 更新, data, image, label, sample, train, model, object, target, classifier, verify, correct, update

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110288007 A (BEIJING SCIENCE AND TECHNOLOGY CO., THREE FAST ONLINE) 27 September 2019 (2019-09-27) claims 1-10	1-10
X	CN 109241997 A (SHUO CHENG (XIAMEN) TECHNOLOGY CO., LTD.) 18 January 2019 (2019-01-18) description, paragraphs [0052]-[0076]	1-10
A	CN 109242013 A (BEIJING JIUHU TIMES INTELLIGENT TECHNOLOGY CO., LTD.) 18 January 2019 (2019-01-18) entire document	1-10
A	CN 109635838 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 16 April 2019 (2019-04-16) entire document	1-10
A	CN 108875768 A (BEIJING MEGVII TECHNOLOGY CO., LTD.) 23 November 2018 (2018-11-23) entire document	1-10
A	CN 103324937 A (NEC CHINA CO., LTD.) 25 September 2013 (2013-09-25) 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

20 February 2020

Date of mailing of the international search report

06 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/CN2019/123406

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2018204111 A1 (Z ADVANCED COMPUTING, INC.) 19 July 2018 (2018-07-19) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/123406

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110288007	A	27 September 2019	None	
CN	109241997	A	18 January 2019	None	
CN	109242013	A	18 January 2019	None	
CN	109635838	A	16 April 2019	None	
CN	108875768	A	23 November 2018	None	
CN	103324937	A	25 September 2013	None	
US	2018204111	A1	19 July 2018	None	

国际检索报告

国际申请号

PCT/CN2019/123406

A. 主题的分类

G06K 9/62 (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; G06N

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; CNKI; WPI; EPDOC; 数据, 图像, 标注, 样本, 训练, 模型, 目标, 分类器, 验证, 更正, 更新, data, image, label, sample, train, model, object, target, classifier, verify, correct, update

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 110288007 A (北京三快在线科技有限公司) 2019年 9月 27日 (2019 - 09 - 27) 权利要求1-10	1-10
X	CN 109241997 A (硕橙厦门科技有限公司) 2019年 1月 18日 (2019 - 01 - 18) 说明书第[0052]-[0076]段	1-10
A	CN 109242013 A (北京九狐时代智能科技有限公司) 2019年 1月 18日 (2019 - 01 - 18) 全文	1-10
A	CN 109635838 A (平安科技深圳有限公司) 2019年 4月 16日 (2019 - 04 - 16) 全文	1-10
A	CN 108875768 A (北京迈格威科技有限公司) 2018年 11月 23日 (2018 - 11 - 23) 全文	1-10
A	CN 103324937 A (日电中国有限公司) 2013年 9月 25日 (2013 - 09 - 25) 全文	1-10
A	US 2018204111 A1 (Z ADVANCED COMPUTING, INC.) 2018年 7月 19日 (2018 - 07 - 19) 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 2月 20日

国际检索报告邮寄日期

2020年 3月 6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

焦月

电话号码 86-(10)-53961306

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/123406

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	110288007	A	2019年 9月 27日	无	
CN	109241997	A	2019年 1月 18日	无	
CN	109242013	A	2019年 1月 18日	无	
CN	109635838	A	2019年 4月 16日	无	
CN	108875768	A	2018年 11月 23日	无	
CN	103324937	A	2013年 9月 25日	无	
US	2018204111	A1	2018年 7月 19日	无	