

(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
2021 年 5 月 27 日 (27.05.2021)



(10) 国际公布号
WO 2021/098620 A1

(51) 国际专利分类号:
G06F 16/16 (2019.01) *G06N 3/04* (2006.01)

(21) 国际申请号: PCT/CN2020/128860

(22) 国际申请日: 2020 年 11 月 13 日 (13.11.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201911146348.0 2019 年 11 月 21 日 (21.11.2019) 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)。

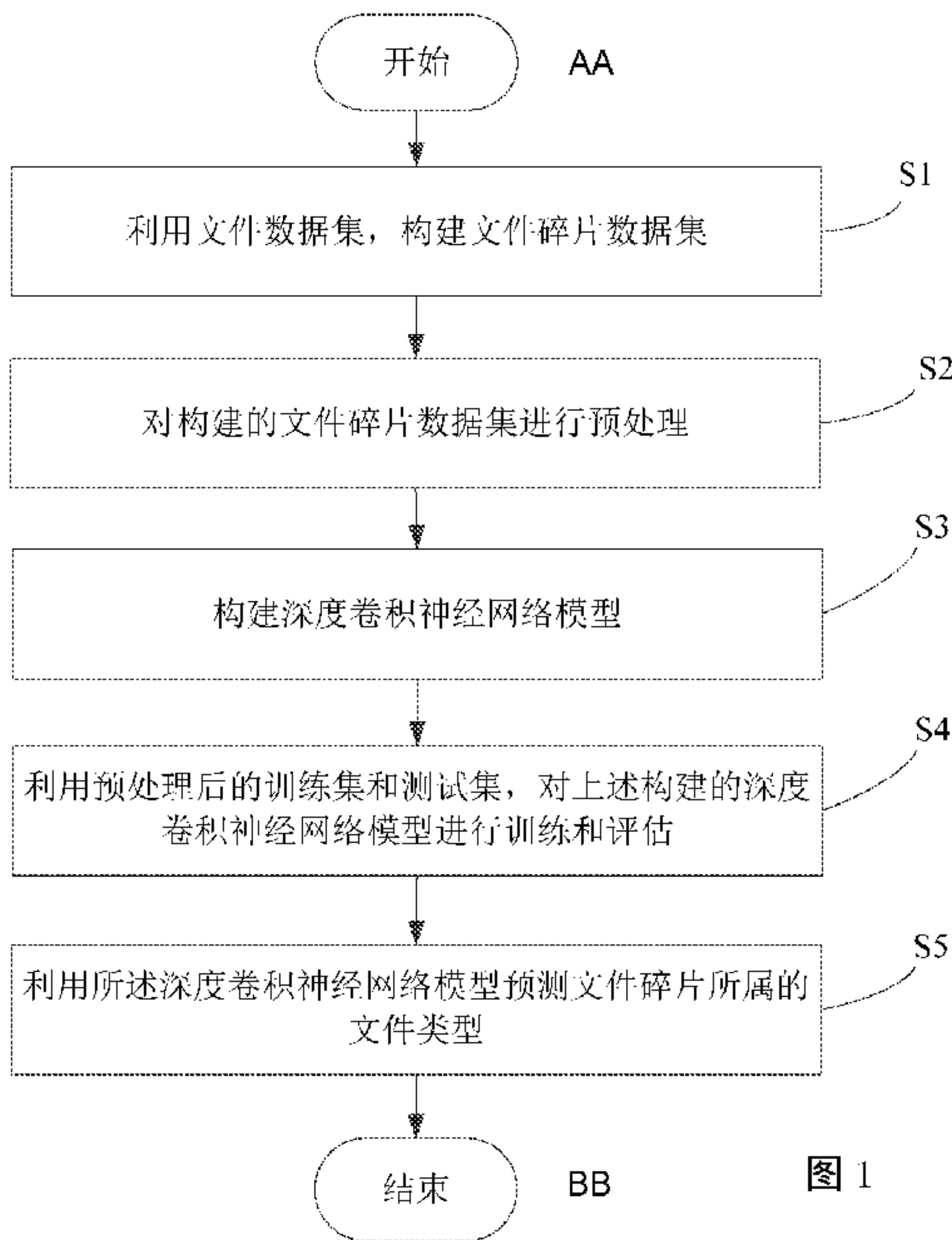
(72) 发明人: 尹凌(YIN, Ling); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。 奚桂锴(XI, Guikai); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(74) 代理人: 北京中巡通大知识产权代理有限公司 (BEIJING ZHONG XUN TONG DA INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市海淀区杏石口路 50 号中间建筑 1 区 148 号, Beijing 100195 (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,

(54) Title: FILE FRAGMENT CLASSIFICATION METHOD AND SYSTEM

(54) 发明名称: 一种文件碎片分类方法及系统



S1 Use a file data set to construct a file fragment data set
S2 Preprocess the constructed file fragment data set
S3 Construct a deep convolutional neural network model
S4 Use a preprocessed training set and test set to train and evaluate the deep convolutional neural network model constructed above
S5 Use the deep convolutional neural network model to predict the file type to which a file fragment belongs
AA Start
BB End

图 1

(57) Abstract: A file fragment classification method and a file fragment classification system. The method comprises: using a file data set to construct a file fragment data set (S1), the file fragment data set comprising a training set and a test set; preprocessing the constructed file fragment data set (S2); constructing a deep convolutional neural network model (S3); using the preprocessed training set and test set to train and evaluate the deep convolutional neural network model constructed above (S4); and using the deep convolutional neural network model to predict the file type to which a file fragment belongs (S5). The method and system do not require that features be

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, IT, 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))。

manually designed, do not require other prior knowledge, and may automatically learn the features of inputted file fragments. Moreover, the designed deep convolutional neural network may be suitable for classification tasks of file fragments of different sizes, and has a better classification effect.

(57) 摘要：一种文件碎片分类方法和文件碎片分类系统。所述方法包括：利用文件数据集，构建文件碎片数据集（S1），所述的文件碎片数据集包括：训练集和测试集；对构建的文件碎片数据集进行预处理（S2）；构建深度卷积神经网络模型（S3）；利用预处理后的训练集和测试集，对上述构建的深度卷积神经网络模型进行训练和评估（S4）；利用所述深度卷积神经网络模型预测文件碎片所属的文件类型（S5）。所述方法和系统无需手动设计特征，无需其他先验知识，能够自动学习到输入的文件碎片的特征，并且设计的深度卷积神经网络能够适用于不同大小的文件碎片的分类任务，具有更好的分类效果。

一种文件碎片分类方法及系统

技术领域

本发明涉及一种文件碎片分类方法及系统。

背景技术

当犯罪嫌疑人删除存储在磁盘中的文件后，磁盘中往往还会有残留的文件内容。如果司法取证人员想要通过磁盘中的文件碎片寻找证据，就需要对这些文件碎片进行重组进而拼接成文件。

如果直接对大量的文件碎片进行两两拼接，则需要耗费巨大的计算量。如果能够提前知道各文件碎片所属文件的文件类型（即，文件碎片的类型），则可以大大减少所需要尝试的组合数量。

现有文件碎片分类方法中的一类是利用魔幻数字等来识别不同文件类型的文件。这些魔幻数字一般出现在文件头和文件尾，并且不同文件类型的文件会在不同的位置出现不同数值的魔幻数字。由于磁盘中的文件往往以碎片化的形式存储，同属一个文件的多个文件碎片并不总是顺序相连，故通常难以利用文件的文件头信息和文件尾信息来识别不同文件类型的文件碎片。

另一类文件碎片分类方法是基于内容的文件碎片分类方法。基于内容的文件碎片分类方法是直接通过对文件碎片内容的分析进而预测该文件碎片的文件类型。该方法不需要依赖于文件签名或者魔幻数字等。

现有基于内容的文件碎片分类方法主要是从统计学角度出发，通过提取各文件碎片的统计学特征，如 unigram 和 bigram 的频率分布，以及熵等，建立传统的机器学习模型，如 LDA、SVM 和 KNN 等，进而识别出各文件碎片所对应的类型。在基于内容的文件碎片分类方法中，通过提取文件碎片的统计学特征进而建立传统机器学习模型的方法严重依赖于特征的设计，是耗时的并且要求具备大量的专业知识。而且，这类方法目前并不能达到比较好的分类效果。

在基于内容的文件碎片分类方法中，现有基于深度学习的文件碎片分类方法还没有成熟，相应的分类效果不好，低于基于传统机器学习模型的文件碎片分类方法。现有基于深度学习的研究还需针对不同大小的文件碎片设计不同的神经网络架构，因此这类现有方法的适用性也受到了一定的限制。

发明内容

有鉴于此，有必要提供一种文件碎片分类方法及系统。

本发明提供一种文件碎片分类方法，该方法包括如下步骤：a.利用文件数据集，构建文件碎片数据集，所述的文件碎片数据集包括：训练集和测试集；b.对构建的文件碎片数据集进行预处理；c.构建深度卷积神经网络模型；d.利用预处理后的训练集和测试集，对上述构建的深度卷积神经网络模型进行训练和评估；e.利用所述深度卷积神经网络模型预测文件碎片所属的文件类型。

其中，所述的步骤 a 具体包括：

对公开文件数据集 govdocs1 包含的所有 zip 压缩包文件进行解压，

将解压后文件夹中的文件按照所属的文件类型划分到不同的类别；

将对应对研究的文件类型所选取的文件划分成两类，以生成分别用于训练集和测试集的文件碎片；

对每个文件根据所选的文件碎片大小进行切片以生成大量文件碎片，并删除每个文件的头一个文件碎片，及每个文件最后一个小于指定文件碎片大小的文件碎片。

所述的步骤 b 具体包括：

对生成的训练集和测试集中的每一个文件碎片都进行转换，通过简单的形状变化将一维的文件碎片转换为二维灰度图像；

对每个所述二维灰度图像进行归一化处理，计算训练集中每个位置像素点的最大值和最小值，将训练集和测试集中对应的二维灰度图像，依据训练集中求得的所述最大值和最小值将对应的像素点进行缩放，使得所述像素点的灰度值落在-1 到 1 之间。

所述的深度卷积神经网络模型包含 L 个卷积块，一个全局平均池化层以及两个全连接层。

所述卷积块包括：卷积层、残差单元和最大池化层三个部分；

卷积块的数量 L 受转换后的灰度图像的大小限制：

$$L_{max} = \min(\log_2 \max(w, h) - 1, \log_2 \min(w, h))$$

在该式中， L_{max} 指的是所述模型中允许堆叠的卷积块的最大数量，w 和 h 分别指的是转换后的二维灰度图像的宽和高。

所述卷积层使用 d 个 1x1 的卷积核，假设卷积块输入了 C 个 IxJ 的特征图，则卷积层对输入特征图的通道数进行上采样。

所述残差单元包含两个卷积层，采用残差学习的方法进行跳跃连接。

所述最大池化层对每个输入特征图进行空间上的下采样，减小为原来的 $\frac{1}{2}$ ，即 $\frac{I}{2} \times \frac{J}{2}$ 。

所述的步骤 d 具体包括：

利用预处理后的测试集对所述的深度卷积神经网络进行评估，评估指标包括多个文件碎片类别的平均分类准确率，宏平均的 F1 分数和微平均的 F1 分数。

本发明提供一种文件碎片分类系统，该系统包括碎片数据集构建模块、预处理模块、模型构建模块、训练评估模块以及文件类型预测模块，其中：所述碎片数据集构建模块用于利用文件数据集，构建文件碎片数据集，所述的文件碎片数据集包括：训练集和测试集；所述预处理模块用于对构建的文件碎片数据集进行预处理；所述模型构建模块用于构建深度卷积神经网络模型；所述训练评估模块用于利用预处理后的训练集和测试集，对上述构建的深度卷积神经网络模型进行训练和评估；所述文件类型预测模块用于利用所述深度卷积神经网络模型预测文件碎片所属的文件类型。

本申请提供了一种文件碎片分类方法及系统，只需将输入的文件碎片先转换为二维的灰度图像，再输入到模型中即可进行预测。本发明在将文件碎片转换成二维的灰度图像时，并不需要耗费额外的计算量。本发明在预测文件碎片的类型时，完全基于该文件碎片的内容进行判断，无需其他先验知识。本发明可直接从输入的文件碎片中自动学习到特征，不需要先从文件碎片中手动提取特征再进行建模。另外，本发明设计的深度卷积神经网络能够适用于不同大小的文件碎片的分类任务。本发明设计的深度卷积神经网络采用残差结构设计，能够搭建更深的网络模型，

适用于处理不同大小的文件碎片分类任务，有效地提高了文件碎片的分类准确率，具有更好的分类效果。

附图说明

图 1 为本发明文件碎片分类方法的流程图；

图 2 是本发明实施例将文件碎片转换为灰度图像的过程示意图；

图 3 是本发明实施例深度卷积神经网络模型的示意图；

图 4 是本发明实施例深度卷积神经网络模型中卷积块的示意图；

图 5 是本发明实施例深度卷积神经网络模型中残差单元的示意图。

图 6 为本发明文件碎片分类系统的硬件架构图。

具体实施方式

下面结合附图及具体实施例对本发明作进一步详细的说明。

参阅图 1 所示，是本发明文件碎片分类方法较佳实施例的作业流程图。

步骤 S1，利用文件数据集，构建文件碎片数据集。所述的文件碎片数据集包括：训练集和测试集。具体而言：

在本实施例中，利用公开的文件数据集 govdocs1 生成所述的文件碎片数据集。所述文件数据集包含 1000 个 zip 压缩包文件。对该文件数据集包含的所有 zip 压缩包文件进行解压，并将解压后文件夹中的文件按照所属的文件类型划分到不同的类别。

针对需要研究的文件碎片类型，均选取一定数量的文件用于实验。将对有待研究的文件类型所选取的文件分别按照 6：4 的比例，划分成

两类，以生成分别用于训练集和测试集的文件碎片。

对每个文件根据所选的文件碎片大小进行切片以生成大量文件碎片。为了避免文件头中包含可用于识别文件类型的文件签名，删除每个文件的头一个文件碎片，同时，将每个文件最后一个小于指定文件碎片大小的文件碎片删除。针对所述训练集和所述测试集，通过随机抽样的方式限制各文件类型对应的文件碎片数量，以使得数据集尽可能平衡，得到对应不同文件类型分别用于训练和测试的大量文件碎片。

步骤 S2，对构建的文件碎片数据集进行预处理，也即，对所述训练集和所述测试集进行预处理。具体而言：

对生成的训练集和测试集中的每一个文件碎片都进行转换，通过简单的形状变化即可将一维的文件碎片转换为二维灰度图像，请参考图2。其中，所述文件碎片由字节序列组成；每个字节对应所述二维灰度图像中的每个像素点。在将文件碎片（一维的字节序列）转换为二维的灰度图像时，应使得灰度图像的形状尽可能接近方形，以利于构建足够深的模型来进行文件碎片的分类。

在本实施例中，将 512 字节的文件碎片转换为 16×32 ($16 \times 32 = 512$) 的二维灰度图像；将 4096 字节的文件碎片转换为 64×64 ($64 \times 64 = 4096$) 的二维灰度图像。

最后，对每个所述二维灰度图像进行归一化处理，计算训练集中每个位置像素点的最大值和最小值，将训练集和测试集中对应的二维灰度图像，依据训练集中求得的所述最大值和最小值将对应的像素点进行缩放，使得所述像素点的灰度值落在-1 到 1 之间。

步骤 S3，构建深度卷积神经网络模型。具体而言：

如图3所示，所述的深度卷积神经网络模型包含L个卷积块，一个全局平均池化层以及两个全连接层。图3中所述的ReLU (Rectified Linear Unit) 均指的是修正线性单元，是一种激活函数。

其中，每个卷积块的结构如图4所示，包含三个部分：卷积层、残差单元和最大池化层。其中：所述卷积层使用d个1x1的卷积核，假设卷积块输入了C个I_xJ的特征图，卷积层对输入特征图的通道数进行上采样（从C增大到d）；所述残差单元进行特征学习，而所述最大池化层对每个输入特征图进行空间上的下采样，减小为原来的 $\frac{1}{2}$ ，即 $\frac{I}{2} \times \frac{J}{2}$ ，特征图的数量则保持不变。

卷积块的数量L受到转换后的灰度图像的大小限制，如下式：

$$L_{max} = \min(\log_2 \max(w, h) - 1, \log_2 \min(w, h))$$

在该式中， L_{max} 指的是所述模型中允许堆叠的卷积块的最大数量，w和h分别指的是转换后的二维灰度图像的宽和高。

其中，所述残差单元的结构如图5所示，所述残差单元包含两个卷积层，采用残差学习的方法进行跳跃连接。所述两个卷积层均采用d个3x3的卷积核，用于学习输入特征图的特征。输入特征图在输入到所述两个卷积层之前，都先经过ReLU激活函数进行计算。

所述模型的两个全连接层均具有2048个神经元。

尽管本申请在一定的实践基础上构建了如图3、图4、图5的模型结构，给出了模型相关部分的参数，但本发明的模型结构不应仅限于此，也不应局限于所述的模型结构参数。

步骤 S4，利用预处理后的训练集和测试集，对上述构建的深度卷积

神经网络模型进行训练和评估。评价指标包括多个文件碎片类别的平均分类准确率，宏平均的 F1 分数和微平均的 F1 分数。具体而言：

在本实施例中：

采用基于 Adam 的梯度下降法对所述的深度卷积神经网络进行训练。其中，初始学习率设为 0.001，每 5 个轮次降低学习率为原先的 0.2，训练的总轮次设为 40。此外，还采用 earllystop 技术训练所述的深度卷积神经网络。当所述的深度卷积神经网络在测试集上的评价指标连续 5 轮没有改进，就提前停止训练，取当前的模型参数作为所述的深度卷积神经网络的最佳参数。

步骤 S5：利用所述深度卷积神经网络模型预测文件碎片所属的文件类型。具体包括：

给定待预测的文件碎片后，按照步骤 S2 所示，先将文件碎片转换成二维灰度图像，再将转换后的灰度图像进行归一化处理。

具体的，依据训练集中灰度图像对应位置像素点的最大值和最小值，将该灰度图像对应位置像素点的灰度值缩放至 -1 到 1 之间，再将归一化后的二维灰度图像输入所述深度卷积神经网络模型中，以预测所述文件碎片所属的文件类型。

参阅图 6 所示，是本发明文件碎片分类系统 10 的硬件架构图。该系统包括：碎片数据集构建模块 101、预处理模块 102、模型构建模块 103、训练评估模块 104 以及文件类型预测模块 105。

所述碎片数据集构建模块 101 用于利用文件数据集，构建文件碎片数据集。所述的文件碎片数据集包括：训练集和测试集。具体而言：

在本实施例中，所述碎片数据集构建模块 101 利用公开的文件数据集 govdocs1 生成所述的文件碎片数据集。所述文件数据集包含 1000 个 zip 压缩包文件。对该文件数据集包含的所有 zip 压缩包文件进行解压，并将解压后文件夹中的文件按照所属的文件类型划分到不同的类别。

针对需要研究的文件碎片类型，均选取一定数量的文件用于实验。将对应待研究的文件类型所选取的文件分别按照 6:4 的比例，划分成两类，以生成分别用于训练集和测试集的文件碎片。

所述碎片数据集构建模块 101 对每个文件根据所选的文件碎片大小进行切片以生成大量文件碎片。为了避免文件头中包含可用于识别文件类型的文件签名，删除每个文件的头一个文件碎片，同时，将每个文件最后一个小于指定文件碎片大小的文件碎片删除。针对所述训练集和所述测试集，通过随机抽样的方式限制各文件类型对应的文件碎片数量，以使得数据集尽可能平衡，得到对应不同文件类型分别用于训练和测试的大量文件碎片。

所述预处理模块 102 用于对构建的文件碎片数据集进行预处理，也即，对所述训练集和所述测试集进行预处理。具体包括：

所述预处理模块 102 对生成的训练集和测试集中的每一个文件碎片都进行转换，通过简单的形状变化即可将一维的文件碎片转换为二维灰度图像，请参考图 2。其中，所述文件碎片由字节序列组成；每个字节对应所述二维灰度图像中的每个像素点。在将文件碎片（一维的字节序列）转换为二维的灰度图像时，应使得灰度图像的形状尽可能接近方形，以利于构建足够深的模型来进行文件碎片的分类。

在本实施例中，所述预处理模块 102 将 512 字节的文件碎片转换为

16x32 (16x32=512) 的二维灰度图像；将 4096 字节的文件碎片转换为 64x64 (64x64=4096) 的二维灰度图像。

最后，所述预处理模块 102 对每个所述二维灰度图像进行归一化处理，计算训练集中每个位置像素点的最大值和最小值，将训练集和测试集中对应的二维灰度图像，依据训练集中求得的所述最大值和最小值将对应的像素点进行缩放，使得所述像素点的灰度值落在-1 到 1 之间。

所述模型构建模块 103 用于构建深度卷积神经网络模型。具体而言：

如图3所示，所述的深度卷积神经网络模型包含L个卷积块，一个全局平均池化层以及两个全连接层。图3中所述的ReLU (Rectified Linear Unit) 指的是修正线性单元，是一种激活函数。

其中，每个卷积块的结构如图4所示，包含三个部分：卷积层、残差单元和最大池化层。其中：所述卷积层使用d个1x1的卷积核，假设卷积块输入了C个I_xJ的特征图，卷积层对输入特征图的通道数进行上采样（从C增大到d）；所述残差单元进行特征学习，而所述最大池化层对每个输入特征图进行空间上的下采样，减小为原来的 $\frac{1}{2}$ ，即 $\frac{I}{2} \times \frac{J}{2}$ ，特征图的数量则保持不变。

卷积块的数量L受到转换后的灰度图像的大小限制，如下式：

$$L_{max} = \min(\log_2 \max(w, h) - 1, \log_2 \min(w, h))$$

在该式中， L_{max} 指的是所述模型中允许堆叠的卷积块的最大数量，w和h分别指的是转换后的二维灰度图像的宽和高。

其中，所述残差单元的结构如图5所示，所述残差单元包含两个卷积层，采用残差学习的方法进行跳跃连接。所述两个卷积层均采用d个3x3的卷积核，用于学习输入特征图的特征。输入特征图在输入到所述两个卷积层之前，都先经过ReLU激活函数进行计算。

所述模型的两个全连接层均具有2048个神经元。

尽管本申请在一定的实践基础上构建了如图3、图4、图5的模型结构，给出了模型相关部分的参数，但本发明的模型结构不应仅限于此，也不应局限于所述的模型结构参数。

所述训练评估模块104用于利用预处理后的训练集和测试集，对上述构建的深度卷积神经网络模型进行训练和评估。评价指标包括多个文件碎片类别的平均分类准确率，宏平均的F1分数和微平均的F1分数。具体而言：

在本实施例中：

所述训练评估模块104采用基于Adam的梯度下降法对所述的深度卷积神经网络进行训练。其中，初始学习率设为0.001，每5个轮次降低学习率为原先的0.2，训练的总轮次设为40。此外，还采用earlystop技术训练所述的深度卷积神经网络。当所述的深度卷积神经网络在测试集上的评价指标连续5轮没有改进，就提前停止训练，取当前的模型参数作为所述的深度卷积神经网络的最佳参数。

所述文件类型预测模块105用于利用所述深度卷积神经网络模型预测文件碎片所属的文件类型。具体包括：

所述文件类型预测模块105在给定待预测的文件碎片后，先将文件碎片转换成二维灰度图像，再将转换后的灰度图像进行归一化处理。

具体的，所述文件类型预测模块 105 依据训练集中灰度图像对应位置像素点的最大值和最小值，将该灰度图像对应位置像素点的灰度值缩放至-1 到 1 之间，再将归一化后的二维灰度图像输入所述深度卷积神经网络模型中，预测所述文件碎片所属的文件类型。

虽然本发明参照当前的较佳实施方式进行了描述，但本领域的技术人员应能理解，上述较佳实施方式仅用来说明本发明，并非用来限定本发明的保护范围，任何在本发明的精神和原则范围之内，所做的任何修饰、等效替换、改进等，均应包含在本发明的权利保护范围之内。

权利要求书

1.一种文件碎片分类方法，其特征在于，该方法包括如下步骤：

a.利用文件数据集，构建文件碎片数据集，所述的文件碎片数据集包括：训练集和测试集；

b.对构建的文件碎片数据集进行预处理；

c.构建深度卷积神经网络模型；

d.利用预处理后的训练集和测试集，对上述构建的深度卷积神经网络模型进行训练和评估；

e.利用所述深度卷积神经网络模型预测文件碎片所属的文件类型。

2.如权利要求1所述的方法，其特征在于，所述的步骤a具体包括：

对公开文件数据集 govdocs1 包含的所有 zip 压缩包文件进行解压，将解压后文件夹中的文件按照所属的文件类型划分到不同的类别；

将对应对研究的文件类型所选取的文件划分成两类，以生成分别用于训练集和测试集的文件碎片；

对每个文件根据所选的文件碎片大小进行切片以生成大量文件碎片，并删除每个文件的头一个文件碎片，及最后一个小于指定文件碎片大小的文件碎片。

3.如权利要求2所述的方法，其特征在于，所述的步骤b具体包括：

对生成的训练集和测试集中的每一个文件碎片都进行转换，通过简单的形状变化将一维的文件碎片转换为二维灰度图像；

对每个所述二维灰度图像进行归一化处理，计算训练集中每个位置像素点的最大值和最小值，将训练集和测试集中对应的二维灰度图像，

依据训练集中求得的所述最大值和最小值将对应的像素点进行缩放，使得所述像素点的灰度值落在-1 到 1 之间。

4.如权利要求 3 所述的方法，其特征在于，所述的深度卷积神经网络模型包含 L 个卷积块，一个全局平均池化层以及两个全连接层。

5.如权利要求 4 所述的方法，其特征在于，所述卷积块包括：卷积层、残差单元和最大池化层三个部分；

卷积块的数量L受转换后的灰度图像的大小限制：

$$L_{max} = \min(\log_2 \max(w, h) - 1, \log_2 \min(w, h))$$

在该式中， L_{max} 指的是所述模型中允许堆叠的卷积块的最大数量，w 和 h 分别指的是转换后的二维灰度图像的宽和高。

6.如权利要求 5 所述的方法，其特征在于，所述卷积层使用 d 个 1x1 的卷积核，假设卷积块输入了 C 个 IxJ 的特征图，则卷积层对输入特征图的通道数进行上采样。

7.如权利要求 6 所述的方法，其特征在于，所述残差单元包含两个卷积层，采用残差学习的方法进行跳跃连接。

8.如权利要求 7 所述的方法，其特征在于，所述最大池化层对每个输入特征图进行空间上的下采样，减小为原来的 $\frac{1}{2}$ ，即 $\frac{I}{2} \times \frac{J}{2}$ 。

9.如权利要求 8 所述的方法，其特征在于，所述的步骤 d 具体包括：

利用预处理后的测试集对所述的深度卷积神经网络进行评估，评估指标包括多个文件碎片类别的平均分类准确率，宏平均的 F1 分数和微平均的 F1 分数。

10.一种文件碎片分类系统，其特征在于，该系统包括碎片数据集构建模块、预处理模块、模型构建模块、训练评估模块以及文件类型预测

模块，其中：

所述碎片数据集构建模块用于利用文件数据集，构建文件碎片数据集，所述的文件碎片数据集包括：训练集和测试集；

所述预处理模块用于对构建的文件碎片数据集进行预处理；

所述模型构建模块用于构建深度卷积神经网络模型；

所述训练评估模块用于利用预处理后的训练集和测试集，对上述构建的深度卷积神经网络模型进行训练和评估；

所述文件类型预测模块用于利用所述深度卷积神经网络模型预测文件碎片所属的文件类型。

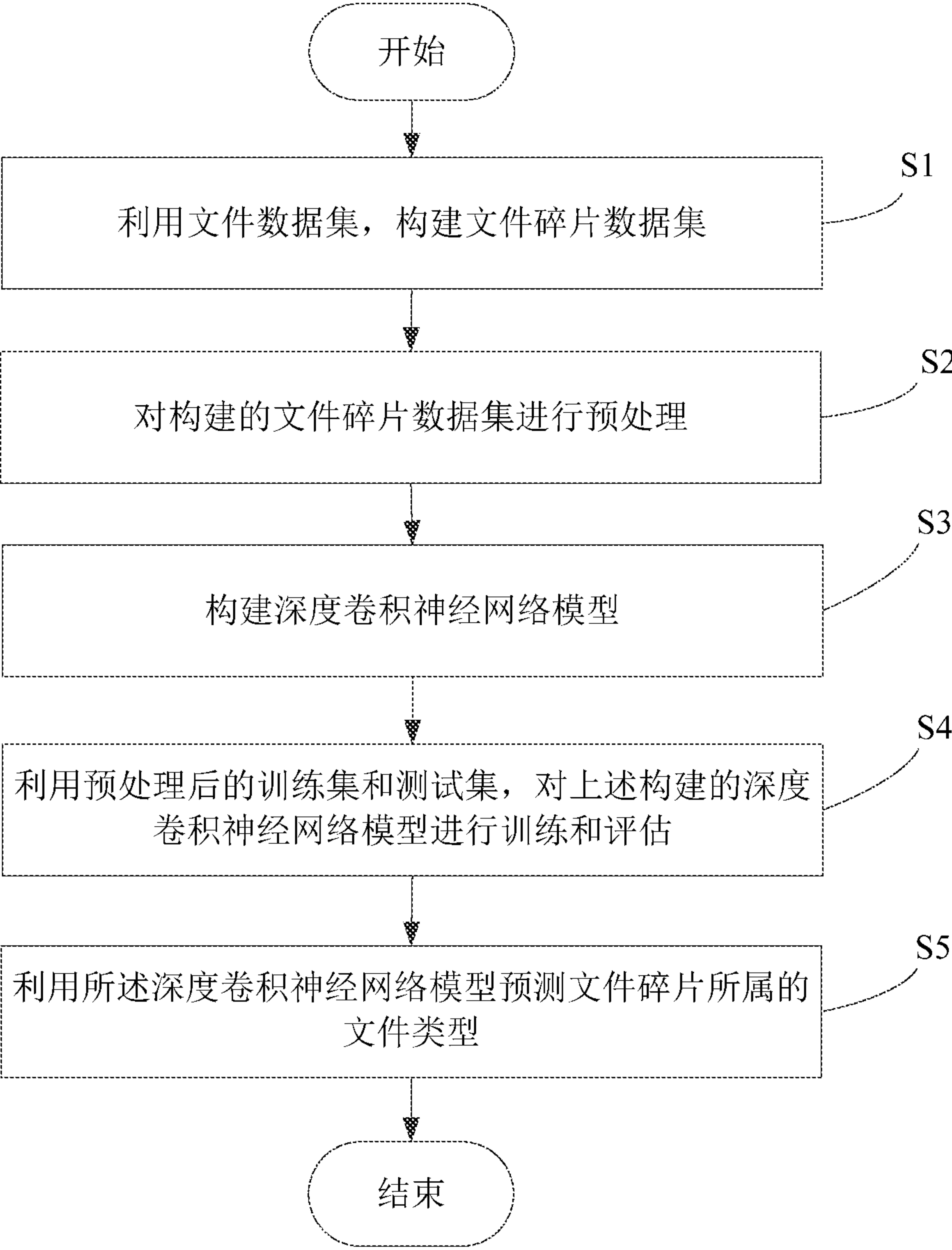


图 1

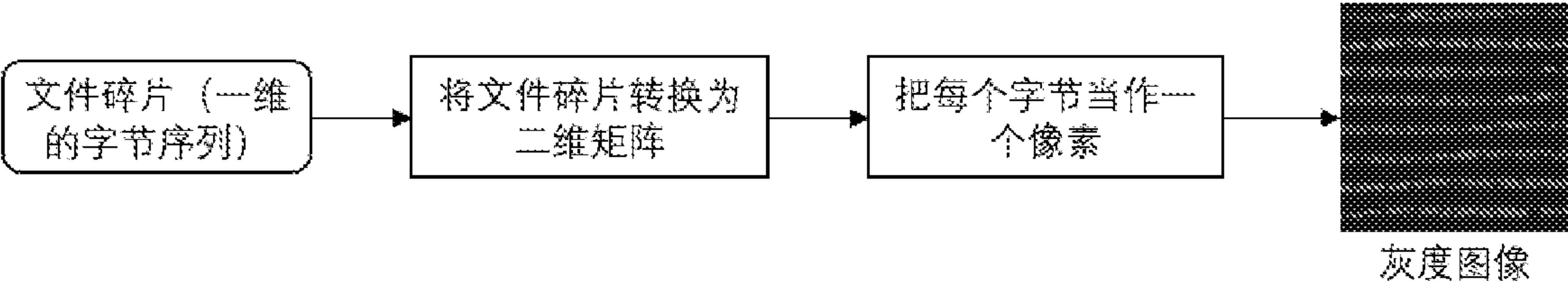


图 2

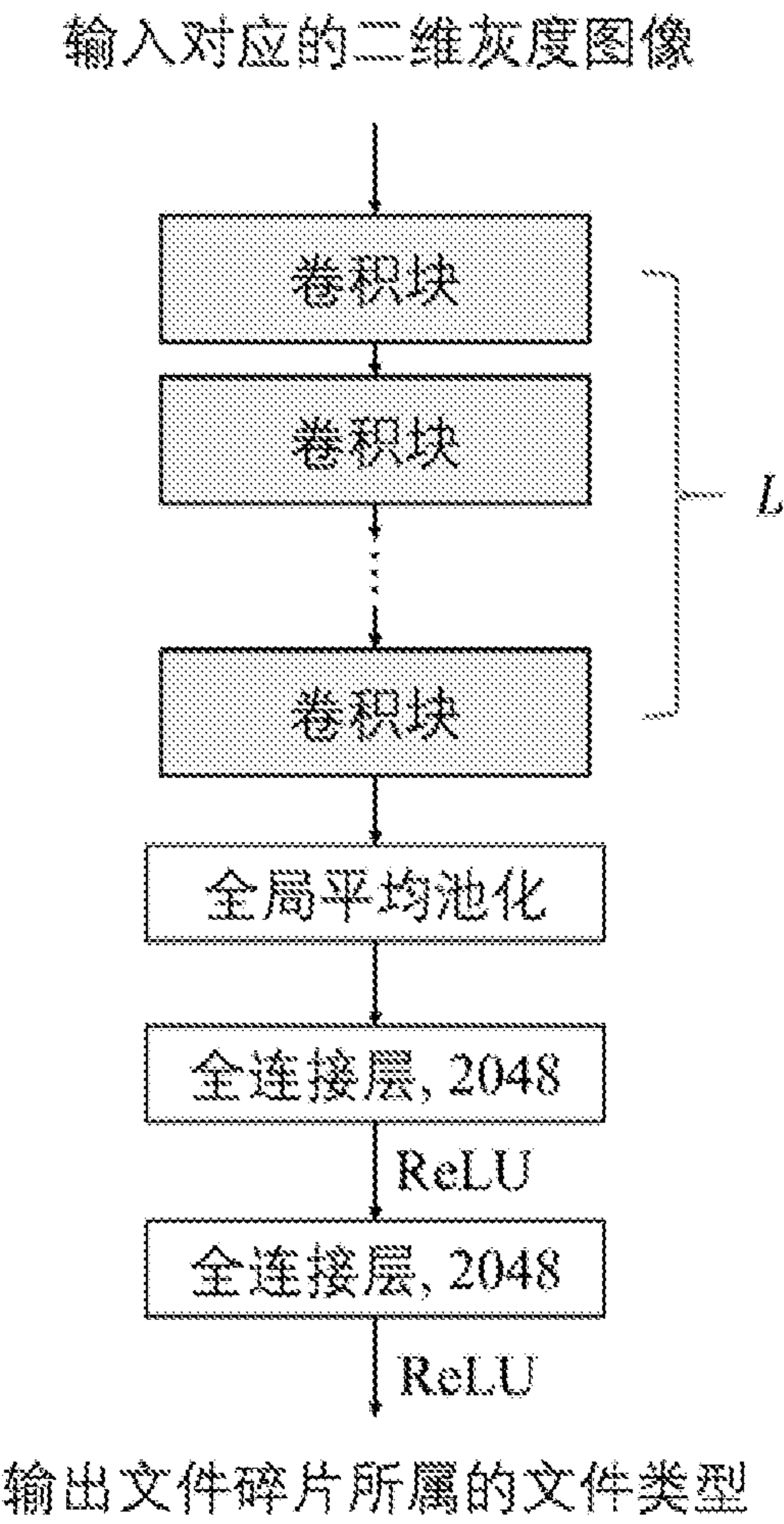


图 3

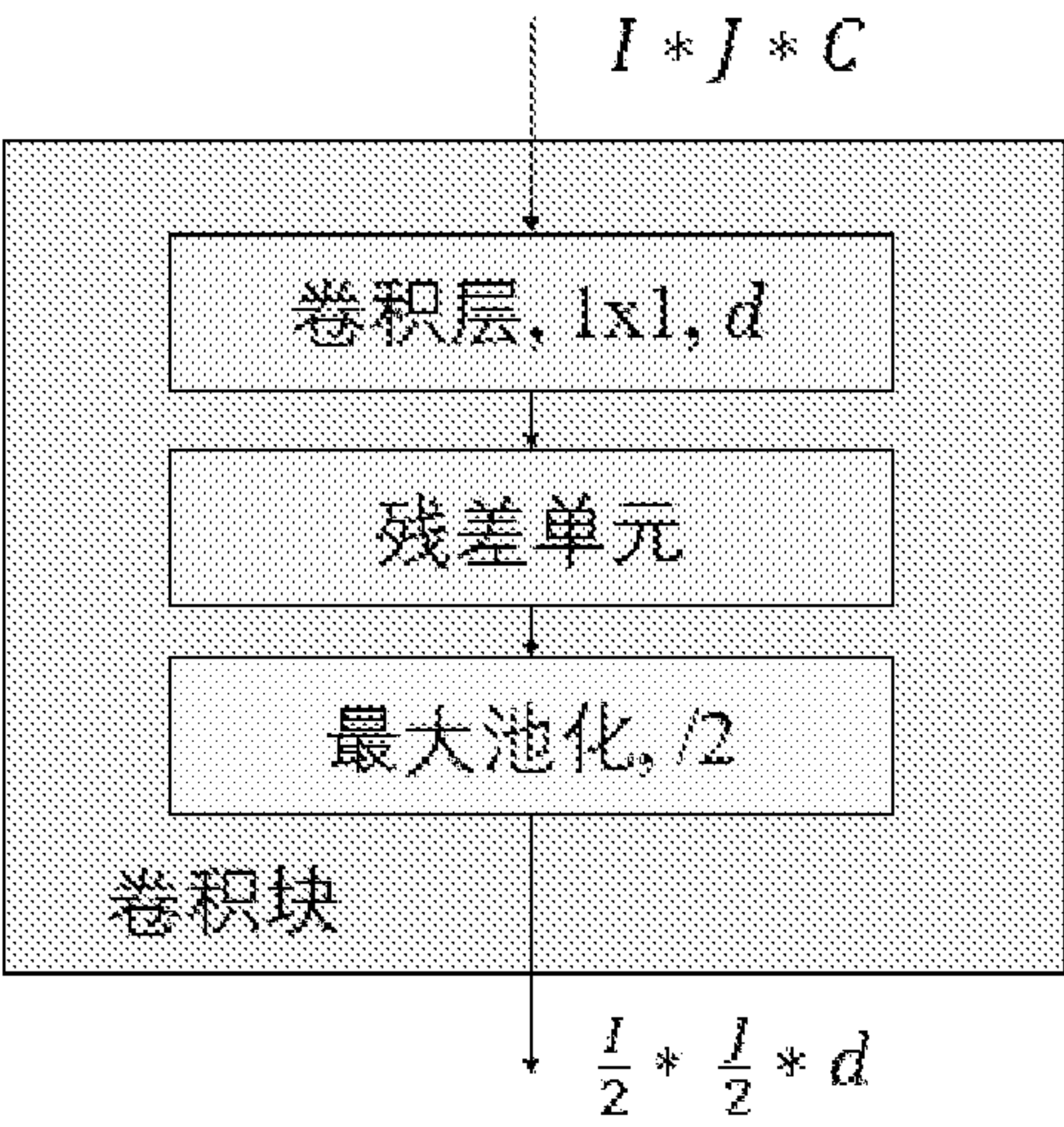


图 4

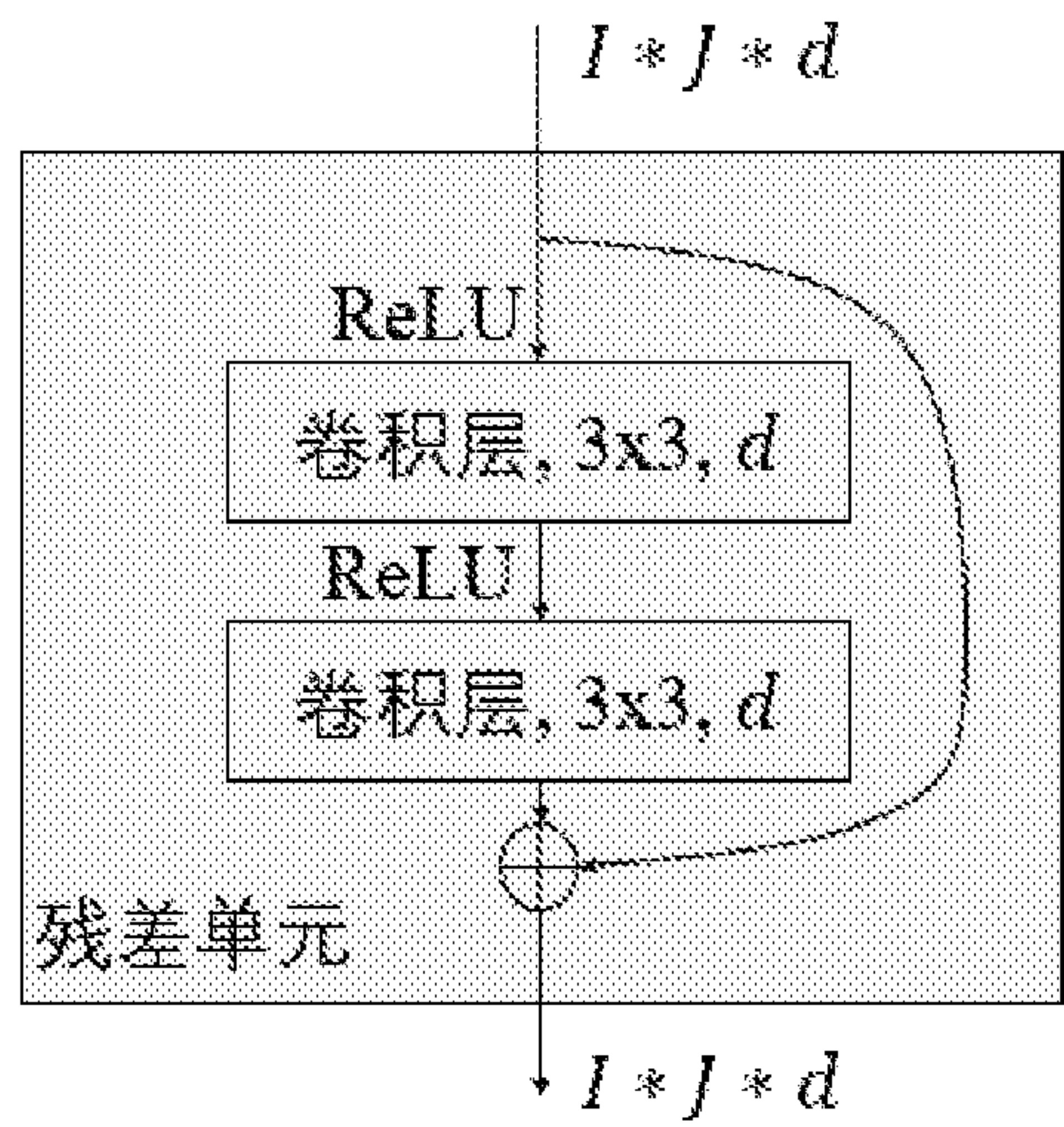


图 5

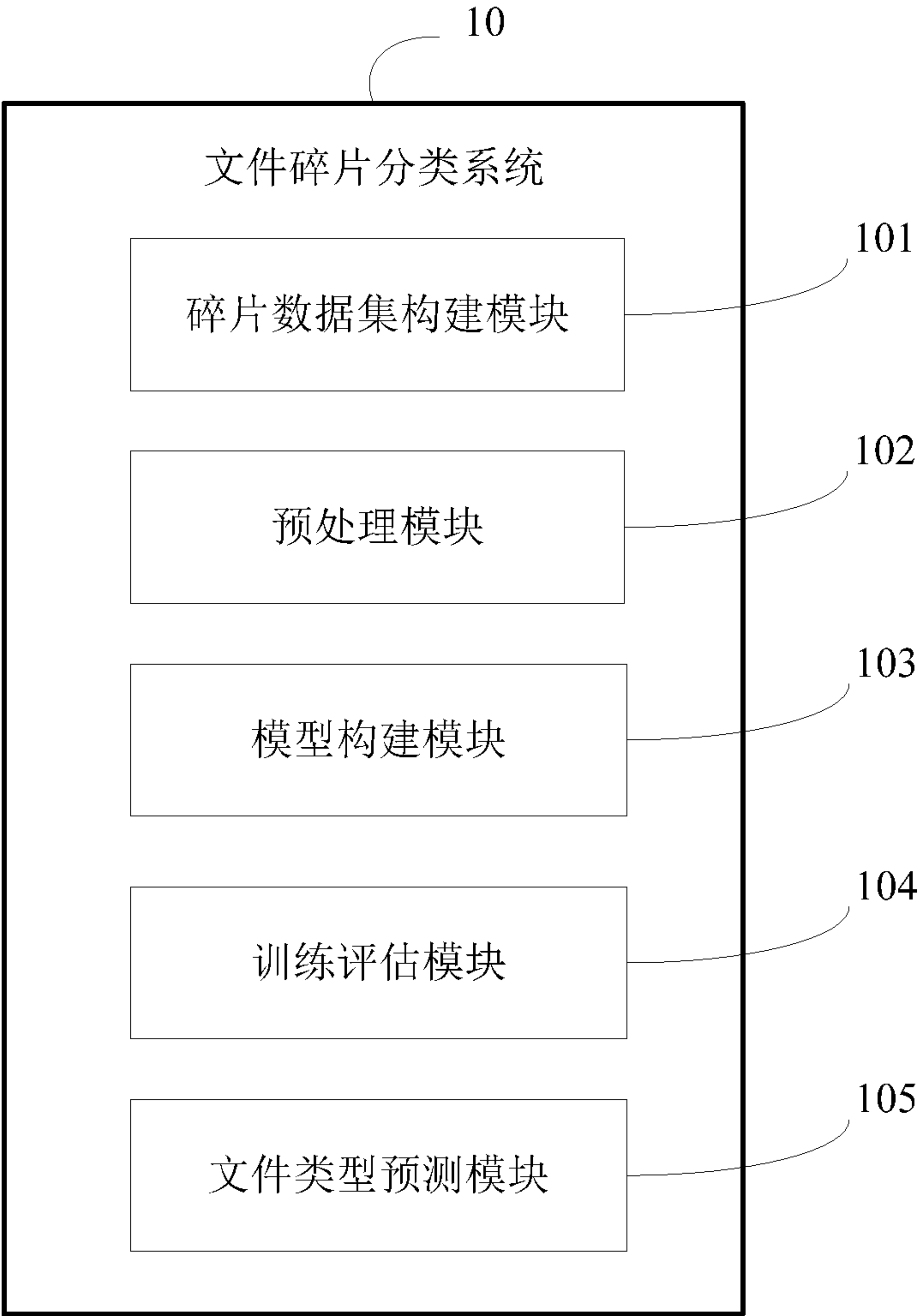


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/128860

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/16(2019.01)i; G06N 3/04(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)

G06F; G06N; G06T; 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: 神经网络, 深度学习, 训练, 卷积网络, 分类, 类型, 识别, 模型, 图像转换, 灰度, 文件, 数据, 碎片, 片段, 粉碎, 删除, 法证, 证物, 证据, 评估, CNN, deep learning, classification, distinguish, identify, recognize, model, image, conversion, gray, file, data, fragment, part, divide, shredder, delete, forensics

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110928848 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 27 March 2020 (2020-03-27) claims 1-10	1-10
X	CHEN, Qian et al. "File Fragment Classification Using Grayscale Image Conversion and Deep Learning in Digital Forensics" 2018 IEEE Symposium on Security and Privacy Workshops DOI 10.1109/SPW.2018.00029, 31 May 2018 (2018-05-31), pp. 140-146	1-10
X	CN 108694414 A (HARBIN INSTITUTE OF TECHNOLOGY SHENZHEN GRADUATE SCHOOL) 23 October 2018 (2018-10-23) abstract, claim 2, description paragraphs [0024]-[0051]	1-10
X	CN 109359090 A (INSTITUTE OF INFORMATION ENGINEERING, CHINESE ACADEMY OF SCIENCES) 19 February 2019 (2019-02-19) description, paragraphs [0010]-[0035]	1, 10
A	CN 102682024 A (INSTITUTE OF HIGH ENERGY PHYSICS CHINESE ACADEMY OF SCIENCES) 19 September 2012 (2012-09-19) 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

29 January 2021

Date of mailing of the international search report

18 February 2021

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/CN2020/128860

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2016071010 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 10 March 2016 (2016-03-10) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/128860

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110928848	A	27 March 2020	None			
CN	108694414	A	23 October 2018	None			
CN	109359090	A	19 February 2019	None			
CN	102682024	A	19 September 2012	CN	102682024	B	05 February 2014
US	2016071010	A1	10 March 2016	CN	105224984	A	06 January 2016
				CN	105224984	B	13 March 2018
				US	10296827	B2	21 May 2019
				WO	2015180397	A1	03 December 2015

国际检索报告		国际申请号 PCT/CN2020/128860
A. 主题的分类 G06F 16/16(2019.01)i; G06N 3/04(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) G06F; G06N; G06T; 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: 神经网络, 深度学习, 训练, 卷积网络, 分类, 类型, 识别, 模型, 图像转换, 灰度, 文件, 数据, 碎片, 片段, 粉碎, 删除, 法证, 证物, 证据, 评估, CNN, deep learning, classification, distinguish, identify, recognize, model, image, conversion, gray, file, data, fragment, part, divide, shredder, delete, forensics		
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 110928848 A (中国科学院深圳先进技术研究院) 2020年 3月 27日 (2020 - 03 - 27) 权利要求1-10	1-10
X	CHEN, Qian 等. "File Fragment Classification Using Grayscale Image Conversion and Deep Learning in Digital Forensics" 2018 IEEE Symposium on Security and Privacy Workshops DOI 10.1109/SPW.2018.00029, 2018年 5月 31日 (2018 - 05 - 31), 第140-146页	1-10
X	CN 108694414 A (哈尔滨工业大学深圳研究生院) 2018年 10月 23日 (2018 - 10 - 23) 摘要、权利要求2、说明书第[0024]-[0051]段	1-10
X	CN 109359090 A (中国科学院信息工程研究所) 2019年 2月 19日 (2019 - 02 - 19) 说明书第[0010]-[0035]段	1, 10
A	CN 102682024 A (中国科学院高能物理研究所) 2012年 9月 19日 (2012 - 09 - 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 2021年 1月 29日		国际检索报告邮寄日期 2021年 2月 18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		授权官员 孙娟 电话号码 (86-10)53961383

国际检索报告

国际申请号

PCT/CN2020/128860

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
A	US 2016071010 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2016年 3月 10日 (2016 - 03 - 10) 全文	1-10

国际检索报告 关于同族专利的信息				国际申请号 PCT/CN2020/128860			
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	110928848	A	2020年 3月 27日	无			
CN	108694414	A	2018年 10月 23日	无			
CN	109359090	A	2019年 2月 19日	无			
CN	102682024	A	2012年 9月 19日	CN	102682024	B	2014年 2月 5日
US	2016071010	A1	2016年 3月 10日	CN	105224984	A	2016年 1月 6日
				CN	105224984	B	2018年 3月 13日
				US	10296827	B2	2019年 5月 21日
				WO	2015180397	A1	2015年 12月 3日