

(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
2023 年 10 月 26 日 (26.10.2023)



(10) 国际公布号
WO 2023/201773 A1

- (51) 国际专利分类号:
G06N 3/063 (2006.01) **G06N 3/08** (2006.01)
- (21) 国际申请号: PCT/CN2022/090124
- (22) 国际申请日: 2022 年 4 月 29 日 (29.04.2022)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202210427530.9 2022年4月22日 (22.04.2022) CN
- (71) 申请人: 浙江大学 (ZHEJIANG UNIVERSITY) [CN/CN]; 中国浙江省杭州市西湖区余杭塘路 866 号, Zhejiang 310058 (CN)。
- (72) 发明人: 尹勋钊 (YIN, Xunzhao); 中国浙江省杭州市西湖区余杭塘路 866 号, Zhejiang 310058 (CN)。

叶文文 (YE, Wenwen); 中国浙江省杭州市西湖区余杭塘路 866 号, Zhejiang 310058 (CN)。卓成 (ZHUO, Cheng); 中国浙江省杭州市西湖区余杭塘路 866 号, Zhejiang 310058 (CN)。

- (74) 代理人: 杭州浙科专利事务所 (普通合伙) (HANGZHOU ZHEKE PATENT AGENCY (GENERAL PARTNERSHIP)); 中国浙江省杭州市西湖区余杭塘路 866 号沈渊琪, Zhejiang 310058 (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, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC,

(54) **Title:** NEURAL NETWORK RETRAINING AND GRADIENT SPARSIFICATION METHOD BASED ON MEMRISTOR AGING PERCEPTION

(54) 发明名称: 基于忆阻器老化感知的神经网络重训练和梯度稀疏方法

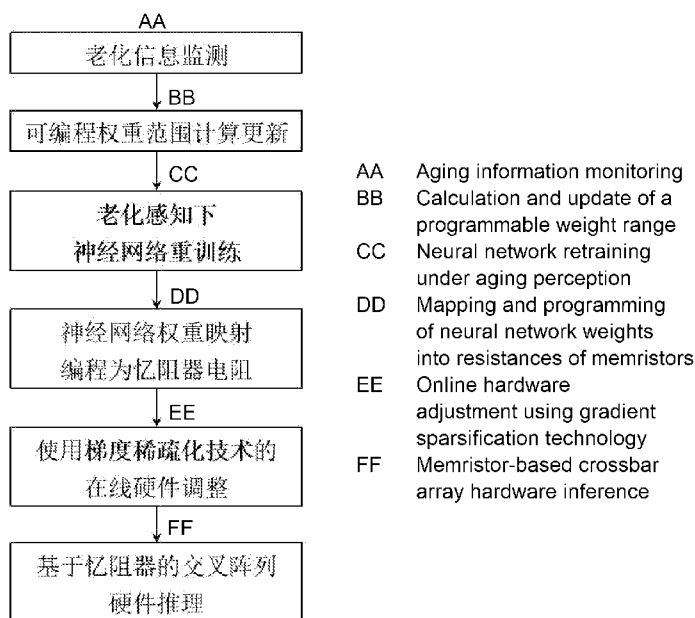


图 1

(57) **Abstract:** A neural network retraining and gradient sparsification method based on memristor aging perception, which relates to a design for prolonging the operating life of a memristor and increasing the inference accuracy for a crossbar array. With regard to the problem of the accuracy of online hardware inference being decreased after a crossbar array is aged, the maximum and minimum values of a programmable weight under the current aging condition are calculated using known memristor aging information, and a neural network model is retrained on this basis, thereby increasing the accuracy of the current online hardware inference. During the process of retraining, network weights exceeding the maximum and minimum values of the programmable weight are automatically truncated.

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, 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, 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)。

根据细则4.17的声明:

- 发明人资格(细则4.17(iv))

本国际公布:

- 包括国际检索报告(条约第21条(3))。

With regard to the problem of the service life of memristors being limited, during the process of hardware adjustment, derivatives having small absolute values are discarded by means of the sparsity of neural network derivatives, thereby ensuring that no voltage is applied to memristors corresponding to the small derivatives, stopping the aging process of the memristors and prolonging the service life of the memristors.

(57) 摘要: 一种基于忆阻器老化感知的神经网络重训练和梯度稀疏方法, 涉及延长忆阻器工作寿命和提高交叉阵列推理正确率的设计。对于交叉阵列老化后硬件在线推理正确率降低的问题, 通过利用已知的忆阻器老化信息, 计算得到当前老化情况下的可编程权重的最值, 依此对神经网络模型进行重训练, 从而提高当前硬件在线推理的正确率; 重训练过程中, 将超出可编程权重最值的网络权重进行自动截断; 对于忆阻器寿命有限的问题, 利用神经网络导数的稀疏性, 在硬件调整过程中, 将绝对值小的导数抛弃, 从而保证小导数所对应的忆阻器不被施加电压, 阻止忆阻器的老化过程, 延长其使用寿命。

基于忆阻器老化感知的神经网络重训练和梯度稀疏方法

技术领域

本发明涉及存算一体化领域，特别涉及一种基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，对提高老化忆阻器交叉阵列的硬件推理正确率并延长其使用寿命具有帮助。

背景技术

随着大数据时代的到来，非易失性器件，如忆阻器等，为实现高效、低功耗的存内计算带来了可能性。现有研究已经证实了基于忆阻器的交叉阵列电路在神经形态加速计算上的巨大潜力。将训练好的神经网络突触权值映射为交叉阵列的电阻时，需要施加高电压脉冲实现编程。由于编程映射时的量化会造成硬件推理正确率下降，还需要施加微调电压调整忆阻器所存储的电阻以得到满足需求的推理正确率。重复施加电压会造成忆阻器内部细丝发生不可逆的变化，这种现象称为老化。老化后的忆阻器可编程范围缩小，从而目标权重的值极有可能在编程映射后落在忆阻器实际的有效范围之外，造成极大的编程误差，进一步造成神经网络的正确率下降，进而需要更多次的施加电压来进行硬件在线调整，进一步加剧忆阻器的老化。

针对忆阻器的老化问题，目前已有研究能够在器件级通过调整编程电压或者改变级联电阻提高忆阻器的耐久力。也有研究为缓解基于忆阻器的交叉阵列的老化现象，提出交换行列，对神经网络进行偏斜训练的方法，从而提高整个器件的使用寿命。但以上方法均未能充分利用器件当前的老化情况这一信息，并且面对忆阻器老化造成的硬件推理正确率下降没有提出很好的改进。

综上，提供基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，利用已知的老化信息提高老化忆阻器交叉阵列的硬件推理正确率并延长其使用寿命，成为目前的重要技术问题。

发明内容

本发明的目的在于针对目前忆阻器交叉阵列的老化情况，提供了基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，通过利用忆阻器的老化信息，对神经网络重训练，并采用梯度稀疏化技术进行在线硬件调整，提高了老化情况下交叉阵列的硬件推理正确率和忆阻器交叉阵列的使用寿命。

本发明的目的是通过以下技术方案来实现的：

基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，该方法包括以下步骤：

S1 采集忆阻器的老化信息，计算老化后的忆阻器的有效可编程电阻范围：

$$R_{\max}^{\text{aged}} = R_{\max}^{\text{fresh}} - G(T, t)$$

$$R_{\min}^{\text{aged}} = R_{\min}^{\text{fresh}} - H(T, t)$$

其中， $R_{\max}^{\text{aged}}$ 和 $R_{\min}^{\text{aged}}$ 分别是交叉阵列的忆阻器老化后的可编程电阻的最大值和最小值， $R_{\max}^{\text{fresh}}$ 和 $R_{\min}^{\text{fresh}}$ 分别是全新忆阻器所能够存储的最大可编程电阻和最小可编程电阻， $G(T, t)$ 和 $H(T, t)$ 为 Arrhenius 老化方程，其参数可以通过测量数据提取得到；

S2 基于 S1 求得的可编程电阻范围，分别求倒数得到老化后的忆阻器可编程电导范围，最大值和最小值分别为 $G_{\max}^{\text{aged}}$ 和 $G_{\min}^{\text{aged}}$ ，并按照下述公式计算要执行的神经网络任务中可编程权重参数的最大值和最小值：

$$W_{\min}^{\text{aged}} = (G_{\min}^{\text{aged}} - G_{\min}^{\text{fresh}}) \frac{W_{\max} - W_{\min}}{G_{\max}^{\text{fresh}} - G_{\min}^{\text{fresh}}} + W_{\min}$$

$$W_{\max}^{\text{aged}} = (G_{\max}^{\text{aged}} - G_{\min}^{\text{fresh}}) \frac{W_{\max} - W_{\min}}{G_{\max}^{\text{fresh}} - G_{\min}^{\text{fresh}}} + W_{\min}$$

其中， $G_{\max}^{\text{fresh}}$ 和 $G_{\min}^{\text{fresh}}$ 分别是交叉阵列中全新忆阻器所能存储的最大电导和最小电导， $W_{\max}^{\text{aged}}$ 和 $W_{\min}^{\text{aged}}$ 分别为与交叉阵列相对应的可编程权重的最大值和最小值， $W_{\max}$ 和 $W_{\min}$ 分别是神经网络初始模型中突触权重的最大值和最小值；

S3 基于神经网络可编程权重参数的最值，对已有的神经网络模型进行重训练；

S4 将调整后的神经网络突触权重参数映射到基于忆阻器的交叉阵列，映射为基本记忆单元的电阻值；

S5 根据神经网络后向传播方法计算得到的导数的正负，将导数分为两组，分别求正负两组导数的平均值，并将绝对值低于该平均值的导数忽略，即置为 0；再按照当前导数值对对应忆阻器所存储的电阻进行硬件在线微调，即对忆阻器施加与导数符号相对应的正负电压：

$$V_{ij} = \begin{cases} +V_0 & \text{if } \text{grad}_{ij} < 0 \\ -V_0 & \text{if } \text{grad}_{ij} > 0 \end{cases}$$

其中， grad_{ij} 是神经网络损失函数值 Loss 对权重 w_{ij} 求导得到的导数，Loss 是硬件推理结果与实际的损失函数值， V_0 是在线调整时对忆阻器记忆单元施加的标准电压， V_{ij} 是交叉阵列第 i 行第 j 列的忆阻器记忆单元所对应的施加电压；

重复该过程，直到其在线硬件推理正确率满足要求，该要求由人为指定。

进一步地，S3 具体过程包括：

将超出可编程参数范围的权重进行截断，直接赋值为距离当前值最近的权重最大值或者最小权重值，并在当前截断后的参数的基础上对神经网络进行后向传播求导，并继续调整参

数；一直重复上述操作，直到参数满足可编程权重范围，并且仿真推理正确率并无明显降低时为止。

进一步地，S4 具体包括：

S4.1 按照权重和电导的映射关系，计算出神经网络所有突触权重 W 所对应的交叉阵列所有忆阻器应存储的电导值 G ，映射关系为：

$$G = \frac{G_{\max} - G_{\min}}{W_{\max} - W_{\min}}(W - W_{\min}) + G_{\min}$$

其中， $W_{\max}$ 和 $W_{\min}$ 分别是神经网络模型中突触权重的最大值和最小值， $G_{\max}$ 和 $G_{\min}$ 分别是忆阻器的可编程有效电导的最大值和最小值；

S4.2 根据电导和电阻的倒数关系，将计算得到的电导，求倒数，得到交叉阵列所有忆阻器应存储的电阻值： $R=1/G$ ，且有 $R_{\max} = 1/G_{\min}$ 和 $R_{\min} = 1/G_{\max}$ ， $R_{\max}$ 和 $R_{\min}$ 分别是忆阻器的可编程电阻的最大值和最小值；

S4.3 对与神经网络突触位置对应的忆阻器存储单元施加电压，使其存储与计算得到的电阻 R 所对应的最近的量化电阻；

进一步地，S2 中，计算神经网络权重可编程最大值和最小值时，是按照当前各层神经网络的权重分别查找各层的最大最小值，以及每层神经网络所对应的忆阻器当前可编程电导的最值，计算得到的。

本发明的有益效果如下：

1) 提高基于忆阻器的交叉阵列在老化情况下的硬件推理正确率。通过利用已知的老化信息，计算当前交叉阵列的可编程电阻范围，并依此对神经网络进行重训练，经此得到的网络权重参数能够更好得和当前老化情况下的硬件相匹配，降低了映射时因权重超出可编程范围而造成的部分误差，从而提高了硬件推理正确率。

2) 延长基于忆阻器的交叉阵列的使用寿命。利用老化信息的情况得到的重训练后的参数能够最好地匹配当前硬件的状态，因此，不必为了提高硬件推理的正确率而进行过多的硬件在线调整的迭代过程，避免了忆阻器的进一步老化；此外，梯度稀疏化的在线调整方式避免了冗余梯度调整时，对忆阻器施加电压从而老化进一步严重的问题，减少了忆阻器被施加电压的次数，也起到了延长忆阻器交叉阵列使用寿命的作用。

附图说明

图 1 是本发明的实现总流程图；

图 2 是本发明的神经网络突触权重参数映射到基于忆阻器的交叉阵列的示意图；

图 3 是本发明的梯度稀疏化技术的示意图；

图 4 是本发明方法与其他方法的正确率对比图；

图 5 是本发明方法与其他方法的忆阻器交叉阵列的使用寿命的对比图。

具体实施方式

下面结合附图和具体实施例对本发明作进一步详细说明。

如图 1 所示，本发明提供的基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，该方法包括以下步骤：

S1 使用非专利文献 1（S. Balatti *et al.*, “Voltage-controlled cycling endurance of hfo_x-based resistive-switching memory,” IEEE Trans. Electron Devices, vol. 62, no. 10, pp. 3365–3372, 2015.）中的 Arrhenius 老化模型对忆阻器老化过程进行建模，采集得到忆阻器的老化信息，通过下述公式计算老化后的忆阻器的有效可编程电阻范围：

$$R_{\max}^{\text{aged}} = R_{\max}^{\text{fresh}} - G(T, t)$$

$$R_{\min}^{\text{aged}} = R_{\min}^{\text{fresh}} - H(T, t)$$

其中， $R_{\max}^{\text{aged}}$ 和 $R_{\min}^{\text{aged}}$ 分别是交叉阵列的忆阻器经过反复执行任务，被多次施加电压，出现老化时的有效可编程电阻的最大值和最小值， $R_{\max}^{\text{fresh}}$ 和 $R_{\min}^{\text{fresh}}$ 分别是全新的还未执行任务的忆阻器交叉阵列所能够存储的最大可编程电阻和最小可编程电阻。 $G(T, t)$ 和 $H(T, t)$ 为 Arrhenius 老化方程，其参数可以通过测量数据提取得到。

S2 基于 S1 的可编程电阻范围，分别求倒数得到老化后的忆阻器可编程电导范围，最大值和最小值分别为 $G_{\max}^{\text{aged}}$ 和 $G_{\min}^{\text{aged}}$ ，并按照下述公式计算要执行的神经网络任务中可编程权重参数的最大值和最小值：

$$W_{\min}^{\text{aged}} = (G_{\min}^{\text{aged}} - G_{\min}^{\text{fresh}}) \frac{W_{\max} - W_{\min}}{G_{\max}^{\text{fresh}} - G_{\min}^{\text{fresh}}} + W_{\min}$$

$$W_{\max}^{\text{aged}} = (G_{\max}^{\text{aged}} - G_{\min}^{\text{fresh}}) \frac{W_{\max} - W_{\min}}{G_{\max}^{\text{fresh}} - G_{\min}^{\text{fresh}}} + W_{\min}$$

其中， $G_{\max}^{\text{fresh}}$ 和 $G_{\min}^{\text{fresh}}$ 分别是交叉阵列中全新忆阻器所能存储的最大电导和最小电导， $W_{\max}^{\text{aged}}$ 和 $W_{\min}^{\text{aged}}$ 分别为与交叉阵列相对应的可编程权重的最大值和最小值， $W_{\max}$ 和 $W_{\min}$ 分别是神经网络初始模型中突触权重的最大值和最小值。。神经网络模型中的权重参数所允许的范围与忆阻器交叉阵列中的每个记忆单元所能存储的电导范围一一对应。

S3 基于 S2 中神经网络可编程权重参数的最值，对已有的神经网络模型进行重训练。具体过程为：将每个超出可编程参数范围的权重进行截断，直接赋值为距离当前权重值最近的最大值或者最小值：

$$w_{ij} = \begin{cases} w_{\min-ij}^{\text{aged}}, & \text{如果 } w_{ij} < w_{\min-ij}^{\text{aged}}, \\ w_{\max-ij}^{\text{aged}}, & \text{如果 } w_{ij} > w_{\max-ij}^{\text{aged}}. \end{cases}$$

w_{ij} 为神经网络某一层第 i 行第 j 列的权重值, $w_{\min-ij}^{\text{aged}}$ 和 $w_{\max-ij}^{\text{aged}}$ 分别是所能保证被成功映射到忆阻器的神经网络某一层第 i 行第 j 列的最小和最大权重值, 分别对应交叉阵列第 i 行第 j 列的老化忆阻器所能存储的最小和最大电导值。在当前截断参数的基础上计算神经网络的损失函数值, 对神经网络进行后向传播求导, 继续调整参数; 一直重复上述操作, 直到参数满足可编程权重范围, 并且软件仿真推理正确率并无明显降低时为止。

S4 将经过 S3 重训练后的神经网络突触权重参数, 按照图 2 所示的位置对应关系, 分层一一映射到忆阻器交叉阵列的存储单元中, 映射为基本记忆单元的电阻值。

S4.1 按照下述公式中权重和电导的映射关系, 计算出神经网络突触权重 W 所对应的交叉阵列所有忆阻器应编程存储的电导值 G :

$$G = \frac{G_{\max} - G_{\min}}{W_{\max} - W_{\min}} (W - W_{\min}) + G_{\min}$$

其中, $W_{\max}$ 和 $W_{\min}$ 分别是当前神经网络模型中突触权重的最大值和最小值, $G_{\max}$ 和 $G_{\min}$ 分别是交叉阵列各个记忆单元可编程有效电导的最大值和最小值。

S4.2 根据电导和电阻的倒数关系, 将 S4.1 计算得到的电导, 求倒数, 得到交叉阵列所有忆阻器经过编程所应存储的电阻值: $R=1/G$, 且有 $R_{\max} = 1/G_{\min}$ 和 $R_{\min} = 1/G_{\max}$, $R_{\max}$ 和 $R_{\min}$ 分别是忆阻器的可编程电阻的最大值和最小值。

S4.3 对与神经网络突触位置所对应的忆阻器存储单元施加电压, 使其存储与 S4.2 计算得到的电阻 R 所对应的最近的量化电阻。

S5 在交叉阵列中完成神经网络任务。

S5.1 验证交叉阵列在线硬件推理的正确率。

S5.2 如果该正确率不满足要求, 则对忆阻器交叉阵列所存储的电阻值进行在线调整, 在线调整过程中采用图 3 所示的稀疏梯度技术: 根据神经网络后向传播方法计算得到的导数的正负, 将导数分为两组, 分别求正负两组导数的平均值, 并将绝对值低于该平均值的导数忽略, 即置为 0。再按照当前导数值对对应忆阻器所存储的电阻进行硬件在线微调。该微调过程中对忆阻器所施加的电压与上述经过稀疏梯度技术得到的导数符号相对应:

$$V_{ij} = \begin{cases} +V_0 & \text{如果 } \text{grad}_{ij} < 0 \\ -V_0 & \text{如果 } \text{grad}_{ij} > 0 \end{cases}$$

其中, grad_{ij} 是神经网络损失函数值 Loss 对权重 w_{ij} 经过后向传播求导得到的导数, Loss 是硬件推理结果与实际结果的损失函数值, V_0 是在线调整时对忆阻器记忆单元所施加的标准电压值, V_{ij} 是交叉阵列第 i 行第 j 列的忆阻器记忆单元所对应的施加电压。

S5.3 重复 S5.2 这一过程, 直到交叉阵列在线硬件推理正确率满足要求, 所要求的正确率按照实际情况和任务人为指定。

本发明的功能和效果通过以下仿真实验进一步说明展示:

仿真任务

在应用本实施例系统的具体案例中, 选取了 LeNet-5 和 ResNet-18 两个经典网络, 分别用于执行 Fashion-MNIST 和 CIFAR-10 两个数据集的分类任务。在基于忆阻器的交叉阵列中执行上述两个网络的模型, 并多次使用本发明提供的基于忆阻器老化感知的神经网络重训练和梯度稀疏方法的方法, 查看每次使用的效果以及累积使用对交叉阵列的影响。

除了对本发明提到的方法进行实验验证, 还将结果与不使用本方法的常用基本流程以及非专利文献 2 (S. Zhang *et al.*, “Aging-aware Lifetime Enhancement for Memristor-based Neuromorphic Computing,” 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2019, pp. 1751-1756.) 中的所提出的方法进行了比较。比较的指标主要是相同条件下权重参数映射后的硬件推理正确率, 以及为了得到满足需要的正确率所需要的硬件在线调整次数。

仿真结果

(1) 相同条件下权重参数映射后的硬件推理正确率的比较

图 4 给出了在忆阻器交叉阵列中重复执行神经网络时的推理正确率变化。通过纵向比较, 经过重复执行后, 忆阻器老化程度加剧, 正确率衰减严重; 横向与不使用本发明的基本方法相比, 在同等执行次数下, 使用本发明方法的硬件推理正确率能够得到更好的维持。

下表给出了本发明方法与其他方法在同等老化程度的忆阻器交叉阵列中执行推理的正确率比较。

网络	初次映射正确率			在线调整正确率			在线调整次数		
	本发明方法	基本方法	非专利文献 2	本发明方法	基本方法	非专利文献 2	本发明方法	基本方法	非专利文献 2
LeNet-5	85.9%	74.0%	66.9%	87.4%	87.4%	84.3%	5	69	*
ResNet-18	72.1%	69.2%	67.6%	73.6%	73.4%	73.5%	1	5	9

* 到达最大调整限制仍未得到所需要的正确率。

由上表可以看出，在忆阻器交叉阵列一定老化的程度下，执行 LeNet-5 任务时，本发明方法的初次映射正确率比基本方法和非专利文献 1 的方法分别提高了 16% 和 28%；执行 ResNet-18 任务时，分别提高了 4% 和 7%。当前老化情况不至于太过恶劣，所以经过在线调整后，都可以达到相差不多的硬件推理正确率；而达到此正确率所需要的在线调整次数，本发明方法明显减少，与基本方法和非专利文献 1 的方法相比，至多可达到 13.8 倍和 9 倍的调整次数节省。

（2）硬件在线调整次数比较

图 5 给出了在重复执行网络任务的过程中，忆阻器交叉阵列为了达到要求的正确率所需要的在线调整次数的变化。将本发明方法与基本方法以及只使用部分本发明方法（重训练部分）进行比较发现，本发明方法能够执行最多次数的网络任务后，才需要比较多的硬件在线调整迭代次数以达到所要求的正确率，说明忆阻器交叉阵列的使用寿命得到延长。与基本方法相比，执行 LeNet-5 和 ResNet-18 分别将使用寿命延长了 2.7 和 2.5 倍。

从以上结果可以看出，本发明能够通过对网络进行重训练在一定程度老化下的忆阻器交叉阵列中实现更好的推理正确率，并且达到目标正确率所需要的硬件在线调整次数少，延长了基于忆阻器的交叉阵列的使用寿命。

本专利不局限于上述网络任务。上述实施例用来解释说明本发明，而不是对本发明进行限制。任何人在本专利的启示下都可以用其他各种网络任务在老化忆阻器交叉阵列下使用该方法，提高推理正确率并延长忆阻器交叉阵列使用寿命。在本发明的精神和权利要求的保护范围内，凡依照本发明申请专利范围所做的均等变化与修改，皆应属本发明的涵盖范围。

权利要求书

1. 基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，其特征在于，该方法包括以下步骤：

S1 采集忆阻器的老化信息，计算老化后的忆阻器的有效可编程电阻范围：

$$R_{\max}^{\text{aged}} = R_{\max}^{\text{fresh}} - G(T, t)$$

$$R_{\min}^{\text{aged}} = R_{\min}^{\text{fresh}} - H(T, t)$$

其中， $R_{\max}^{\text{aged}}$ 和 $R_{\min}^{\text{aged}}$ 分别是交叉阵列的忆阻器老化后的可编程电阻的最大值和最小值， $R_{\max}^{\text{fresh}}$ 和 $R_{\min}^{\text{fresh}}$ 分别是全新忆阻器所能够存储的最大可编程电阻和最小可编程电阻， $G(T, t)$ 和 $H(T, t)$ 为 Arrhenius 老化方程，其参数可以通过测量数据提取得到；

S2 基于 S1 求得的可编程电阻范围，分别求倒数得到老化后的忆阻器可编程电导范围，最大值和最小值分别为 $G_{\max}^{\text{aged}}$ 和 $G_{\min}^{\text{aged}}$ ，并按照下述公式计算要执行的神经网络任务中可编程权重参数的最大值和最小值：

$$W_{\min}^{\text{aged}} = (G_{\min}^{\text{aged}} - G_{\min}^{\text{fresh}}) \frac{W_{\max} - W_{\min}}{G_{\max}^{\text{fresh}} - G_{\min}^{\text{fresh}}} + W_{\min}$$

$$W_{\max}^{\text{aged}} = (G_{\max}^{\text{aged}} - G_{\min}^{\text{fresh}}) \frac{W_{\max} - W_{\min}}{G_{\max}^{\text{fresh}} - G_{\min}^{\text{fresh}}} + W_{\min}$$

其中， $G_{\max}^{\text{fresh}}$ 和 $G_{\min}^{\text{fresh}}$ 分别是交叉阵列中全新忆阻器所能存储的最大电导和最小电导， $W_{\max}^{\text{aged}}$ 和 $W_{\min}^{\text{aged}}$ 分别为与交叉阵列相对应的可编程权重的最大值和最小值， $W_{\max}$ 和 $W_{\min}$ 分别是神经网络初始模型中突触权重的最大值和最小值；

S3 基于神经网络可编程权重参数的最值，对已有的神经网络模型进行重训练；

S4 将调整后的神经网络突触权重参数映射到基于忆阻器的交叉阵列，映射为基本记忆单元的电阻值；

S5 根据神经网络后向传播方法计算得到的导数的正负，将导数分为两组，分别求正负两组导数的平均值，并将绝对值低于该平均值的导数忽略，即置为 0；再按照当前导数值对对应忆阻器所存储的电阻进行硬件在线微调，即对忆阻器施加与导数符号相对应的正负电压：

$$V_{ij} = \begin{cases} +V_0 & \text{if } \text{grad}_{ij} < 0 \\ -V_0 & \text{if } \text{grad}_{ij} > 0 \end{cases}$$

其中， grad_{ij} 是神经网络损失函数值 Loss 对权重 w_{ij} 求导得到的导数，Loss 是硬件推理结果与实际的损失函数值， V_0 是在线调整时对忆阻器记忆单元施加的标准电压， V_{ij} 是交叉阵列第 i 行第 j 列的忆阻器记忆单元所对应的施加电压；

重复该过程，直到其在线硬件推理正确率满足要求，该要求由人为指定。

2.根据权利要求 1 所述的基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，其特征在于，S3 具体过程包括：

将超出可编程参数范围的权重进行截断，直接赋值为距离当前值最近的权重最大值或者最小权重值，并在当前截断后的参数的基础上对神经网络进行后向传播求导，并继续调整参数；一直重复上述操作，直到参数满足可编程权重范围，并且仿真推理正确率并无明显降低时为止。

3.根据权利要求 2 所述的基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，其特征在于，S4 具体包括：

S4.1 按照权重和电导的映射关系，计算出神经网络所有突触权重 W 所对应的交叉阵列所有忆阻器应存储的电导值 G ，映射关系为：

$$G = \frac{G_{\max} - G_{\min}}{W_{\max} - W_{\min}} (W - W_{\min}) + G_{\min}$$

其中， $W_{\max}$ 和 $W_{\min}$ 分别是神经网络模型中突触权重的最大值和最小值， $G_{\max}$ 和 $G_{\min}$ 分别是忆阻器的可编程有效电导的最大值和最小值；

S4.2 根据电导和电阻的倒数关系，将计算得到的电导，求倒数，得到交叉阵列所有忆阻器应存储的电阻值： $R=1/G$ ，且有 $R_{\max} = 1/G_{\min}$ 和 $R_{\min} = 1/G_{\max}$ ， $R_{\max}$ 和 $R_{\min}$ 分别是忆阻器的可编程电阻的最大值和最小值；

S4.3 对与神经网络突触位置对应的忆阻器存储单元施加电压，使其存储与计算得到的电阻 R 所对应的最近的量化电阻；

4.根据权利要求 1-3 中任一所述的基于忆阻器老化感知的神经网络重训练和梯度稀疏方法，其特征在于，S2 中，计算神经网络权重可编程最大值和最小值时，是按照当前各层神经网络的权重分别查找各层的最大最小值，以及每层神经网络所对应的忆阻器当前可编程电导的最值，计算得到的。

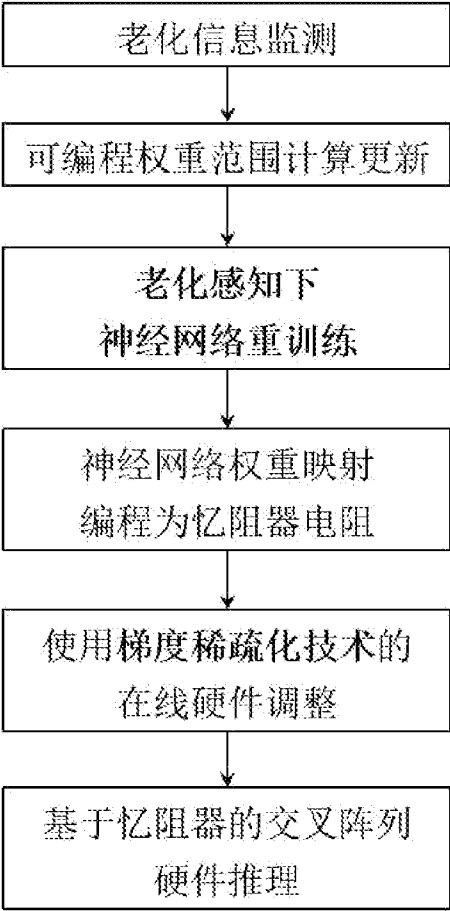


图 1

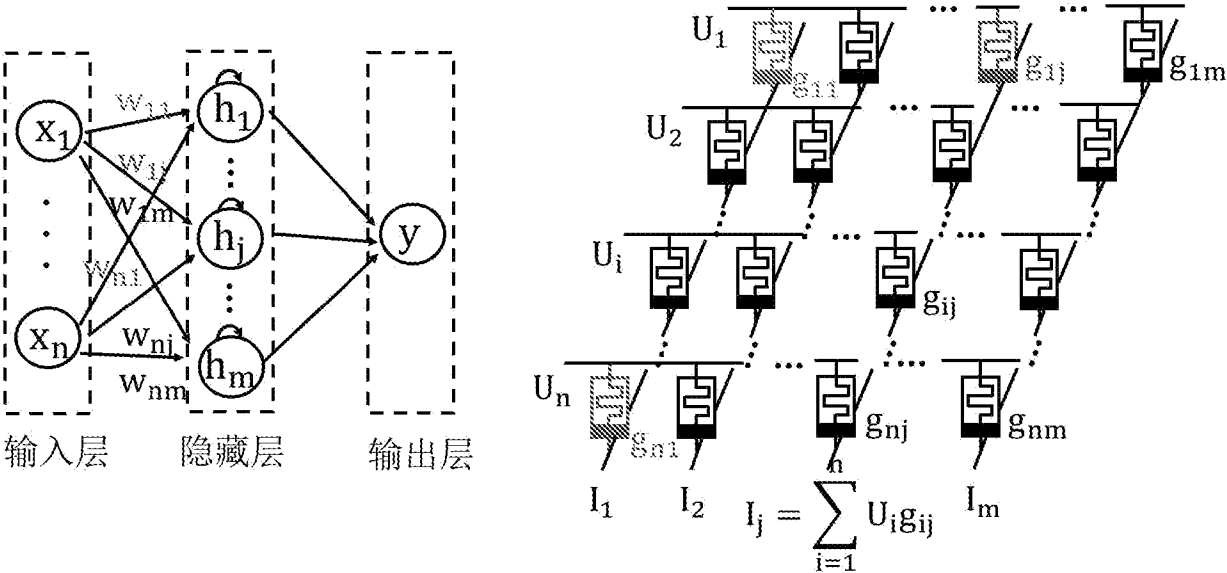


图 2

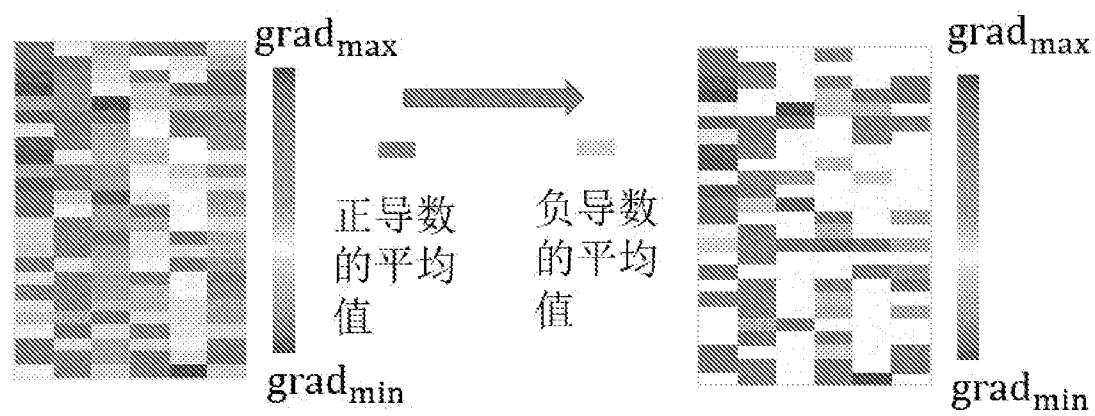


图 3

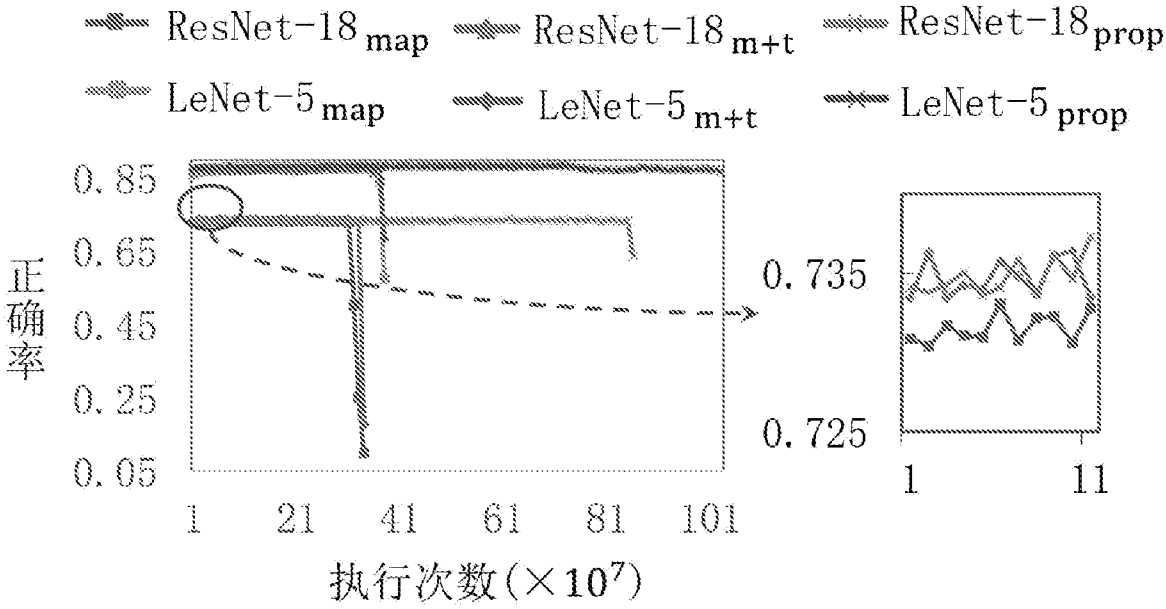


图 4

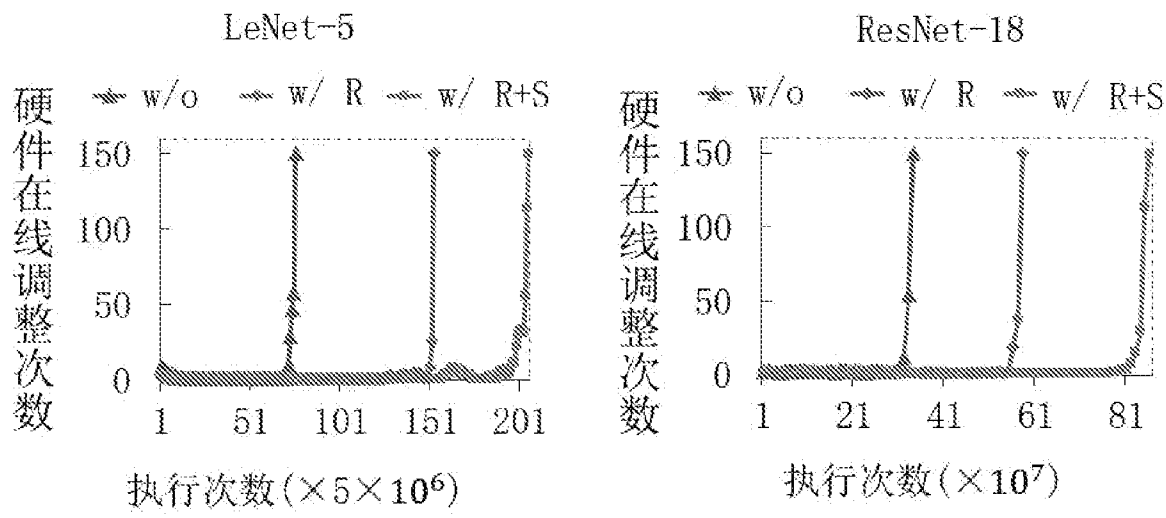


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/090124

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/063(2006.01)i; G06N 3/08(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)

G06N3

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: 训练, 学习, 梯度, 稀疏, 忆阻器, 电阻, 电导, 电压, 范围, 大, 小, 存储, 编程, 老化, 退化, 参数, 阵列, 神经网络, 导数, 符号, training, learning, gradient, sparse, memristor, resistance, conductance, voltage, range, large, small, storage, programming, degeneration, parameters, array, neural network, derivative, symbol

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 113610220 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY OF PLA) 05 November 2021 (2021-11-05) description, paragraphs [0032]-[0088]	1-4
A	US 2015046382 A1 (QUALCOMM INC.) 12 February 2015 (2015-02-12) entire document	1-4
A	TW 202207093 A (POLYN TECHNOLOGY LTD.) 16 February 2022 (2022-02-16) entire document	1-4
A	US 2021143834 A1 (TECHNION RESEARCH & DEVELOPMENT FOUNDATION LTD.) 13 May 2021 (2021-05-13) entire document	1-4
A	US 10510003 B1 (CAPITAL ONE SERVICES, L.L.C.) 17 December 2019 (2019-12-17) entire document	1-4
A	CN 111931924 A (TSINGHUA UNIVERSITY) 13 November 2020 (2020-11-13) entire document	1-4
A	US 2020227635 A1 (UNIVERSITY OF MASSACHUSETTS) 16 July 2020 (2020-07-16) entire document	1-4

☒ 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

04 December 2022

Date of mailing of the international search report

03 January 2023

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/CN2022/090124

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019164046 A1 (ANAFASH INC.) 30 May 2019 (2019-05-30) entire document	1-4

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/090124

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113610220	A	05 November 2021	None			
US	2015046382	A1	12 February 2015	WO	2015020802	A2	12 February 2015
TW	202207093	A	16 February 2022	US	2021406666	A1	30 December 2021
				US	2021406661	A1	30 December 2021
				US	2021406663	A1	30 December 2021
				US	2022004861	A1	06 January 2022
				US	2021406664	A1	30 December 2021
				US	2021406665	A1	30 December 2021
				US	2021406667	A1	30 December 2021
US	2021143834	A1	13 May 2021	EP	3652680	A1	20 May 2020
				WO	2019012533	A1	17 January 2019
US	10510003	B1	17 December 2019	US	10510002	B1	17 December 2019
				US	2021232925	A1	29 July 2021
				US	2020265296	A1	20 August 2020
CN	111931924	A	13 November 2020	None			
US	2020227635	A1	16 July 2020	None			
US	2019164046	A1	30 May 2019	CN	111406265	A	10 July 2020
				US	2022261624	A1	18 August 2022
				WO	2019108458	A1	06 June 2019
				TW	201937414	A	16 September 2019
				US	2019164044	A1	30 May 2019
				EP	3718055	A1	07 October 2020
				KR	20200065079	A	08 June 2020

国际检索报告

国际申请号

PCT/CN2022/090124

A. 主题的分类 G06N 3/063 (2006.01)i; G06N 3/08 (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) G06N3 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, EP0DOC:训练, 学习, 梯度, 稀疏, 忆阻器, 电阻, 电导, 电压, 范围, 大, 小, 存储, 编程, 老化, 退化, 参数, 阵列, 神经网络, 导数, 符号, training, learning, gradient, sparse, memristor, resistance, conductance, voltage, range, large, small, storage, programming, degeneration, parameters, array, neural network, derivative, symbol				
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	CN 113610220 A (中国人民解放军国防科技大学) 2021年11月5日 (2021 - 11 - 05) 说明书第[0032]-[0088]段	1-4		
A	US 2015046382 A1 (QUALCOMM INC.) 2015年2月12日 (2015 - 02 - 12) 全文	1-4		
A	TW 202207093 A (POLYN TECHNOLOGY LIMITED) 2022年2月16日 (2022 - 02 - 16) 全文	1-4		
A	US 2021143834 A1 (TECHNION RESEARCH & DEVELOPMENT FOUNDATION LTD.) 2021年5月13日 (2021 - 05 - 13) 全文	1-4		
A	US 10510003 B1 (CAPITAL ONE SERVICES L.L.C.) 2019年12月17日 (2019 - 12 - 17) 全文	1-4		
A	CN 111931924 A (清华大学) 2020年11月13日 (2020 - 11 - 13) 全文	1-4		
A	US 2020227635 A1 (UNIV. MASSACHUSETTS) 2020年7月16日 (2020 - 07 - 16) 全文	1-4		
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。				
<table><tr><td>* 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</td><td>“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</td></tr></table>			* 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
* 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 2022年12月4日		国际检索报告邮寄日期 2023年1月3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 王玥 电话号码 86-(10)- 53962545		

国际检索报告

国际申请号	
-------	--

PCT/CN2022/090124

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 2019164046 A1 (ANAFLASH INC.) 2019年5月30日 (2019 - 05 - 30) 全文	1-4

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/090124

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	113610220	A	2021年11月5日	无			
US	2015046382	A1	2015年2月12日	WO	2015020802	A2	2015年2月12日
TW	202207093	A	2022年2月16日	US	2021406666	A1	2021年12月30日
				US	2021406661	A1	2021年12月30日
				US	2021406663	A1	2021年12月30日
				US	2022004861	A1	2022年1月6日
				US	2021406664	A1	2021年12月30日
				US	2021406665	A1	2021年12月30日
				US	2021406667	A1	2021年12月30日
US	2021143834	A1	2021年5月13日	EP	3652680	A1	2020年5月20日
				WO	2019012533	A1	2019年1月17日
US	10510003	B1	2019年12月17日	US	10510002	B1	2019年12月17日
				US	2021232925	A1	2021年7月29日
				US	2020265296	A1	2020年8月20日
CN	111931924	A	2020年11月13日	无			
US	2020227635	A1	2020年7月16日	无			
US	2019164046	A1	2019年5月30日	CN	111406265	A	2020年7月10日
				US	2022261624	A1	2022年8月18日
				WO	2019108458	A1	2019年6月6日
				TW	201937414	A	2019年9月16日
				US	2019164044	A1	2019年5月30日
				EP	3718055	A1	2020年10月7日
				KR	20200065079	A	2020年6月8日