# 结合随机梯度下降优化深度神经网络的基因改造狼优化

1st Manuel Bradicic计算机科学系 萨里大学吉尔福德，英国
mb01761@surrey.ac.uk

2nd Michal Sitarz计算机科学系 萨里大学吉尔福德，英国 ms02264@surrey.ac.uk

3rd Felix Sylvest Olesen计算机科学系 萨里大学吉尔福德，英国
fo00150@surrey.ac.uk

arXiv:2301.08950v1 [cs.NE] 21 Jan 2023

摘要—**训练卷积神经网络（CNN）时，人们非常重视构建高效的优化算法和高准确率的网络。目前最先进的网络优化方法是使用梯度下降算法，例如随机梯度下降（SGD）。然而，使用梯度下降方法存在一些局限性。其主要缺点是探索不足，且过度依赖开发。因此，本研究旨在分析一种优化神经网络（NN）权重的替代方法，即使用基于种群的元启发式算法。本文探讨了灰狼优化算法（GWO）与遗传算法（GA）的混合方法，并将其与随机梯度下降结合，提出了一种由随机梯度下降增强的基因改造狼优化算法（GMW-SGD）。该算法兼顾开发与探索，同时应对影响标准元启发式算法性能的高维性问题。所提出的算法在 CIFAR-10 上进行训练和测试，其表现与随机梯度下降算法相当，达到了较高的测试准确率，并显著优于标准元启发式算法。**

索引词—**灰狼优化器、随机梯度下降、SL-PSO、遗传算法、元启发式算法、优化算法、双目标优化、CIFAR-10**

## I. 引言

本文旨在研究和探讨如何使用进化算法，为神经网络寻找最优权重，从而在 Krizhevsky 等人提出的 CIFAR-10 数据集上进行图像分类 [1]。第 2 节首先回顾与卷积神经网络及其发展、进化算法社会学习粒子群优化（SL-PSO），以及一般用于神经网络权重优化的基于种群的元启发式算法相关的研究。第 3 节介绍本文神经网络的详细结构，并说明所选架构的依据。第 4 节详细探讨所提出的混合 GWO-SGD 算法 [2]。第 4 节还探讨了收集到的结果。第 5 节报告将所提出的算法作为双目标问题进行训练的情况，并讨论其整体性能。最后，第 6 节总结本文工作。

## II. 相关工作

### A. 卷积神经网络

典型的前馈神经网络（FFNN），也称为多层感知机（MLP），由输入层、隐藏层和输出层组成。每一层神经元都通过一组实数权重与相邻层相连。此外，除输入层外，每一层神经元还设有一个额外的激活阈值，称为偏置。本文使用卷积神经网络。卷积神经网络相较于前馈神经网络更适用于这一问题，其主要原因是输入为图像，而卷积神经网络能够对图像执行卷积运算。之所以更适合处理此类问题，是因为卷积神经网络所包含的卷积层能够有效捕获图像中的时空依赖关系。卷积是一种线性运算，它接受两个函数 $f$ 和 $h$，并生成另一个函数 $g$。对于卷积神经网络，$f$ 是一个多维数组（图像像素），$h$ 是一个预定义的矩阵或数组（也称为卷积核或滤波器），而 $g$ 则是卷积运算的结果，也称为特征图。前馈神经网络和卷积神经网络主要包含两个阶段：前馈和反向传播，在这两个阶段中，权重和偏置会持续优化。在训练过程中，选择合适的优化器至关重要。通常使用梯度下降方法（GDs） [3], [4] 进行图像分类。根据网络在输入数据上的表现，梯度下降通过代价函数更新网络参数。其目标是沿负梯度方向小步前进，从而达到该代价函数的极小值。其缺点之一是容易收敛到局部最优解 [5]。随机梯度下降（SGD） [6] 是训练神经网络最受欢迎的梯度类算法之一，但同样存在过早收敛的问题。为避免过早收敛，人们开发了多种自适应梯度算法，以高效的方式调整学习率，其中值得一提的是 Adam。以往的文献尝试通过改进梯度下降方法的全局搜索能力来解决过早收敛问题 [5], [7]。所有已提出的

这些相关论文中的解决方案，都致力于将梯度下降快速而高效的收敛速度，与元启发式优化算法无需梯度的全局搜索能力相结合。本文旨在通过将灰狼优化算法（GWO） [2], [8] 与随机梯度下降（SGD）结合以最小化损失，进一步拓展这种神经网络训练方法。此外，还使用了遗传算法（GA） [9] 中的一些元素来提高灰狼优化算法的多样性。

### *B.* 训练算法

元启发式算法是一个迅速发展的领域，包含许多适用于不同问题的变体。无免费午餐定理 [10] 从逻辑上证明了，不存在能够解决所有优化问题的元启发式算法。因此，探索算法变体及其混合方法，或许有助于找到适用于 CIFAR-10 上卷积神经网络的最优权重集合。作为元启发式算法的一个分支，群体智能算法已被证明在此类任务中相当有效 [11]。因此，我们决定将 SL-PSO 和灰狼优化算法应用于这一问题，以探索它们寻找最优解的效果。社会学习粒子群优化（SL-PSO）是原始粒子群优化算法（ PSO）的一个变体 [12], ，由 Cheng 和 Jin [13] 提出，旨在将动物的社会学习特性融入传统 PSO。这样做的目的是将非社会性学习的试错过程转变为一种更具社会性的过程，使个体能够向各自群体中所有表现更好的个体（示范者）学习。SL-PSO 采用群体排序和行为学习 [13]，根据适应度值对群体进行排序，除最优粒子外，每个粒子都会向各自的示范者学习。除最优粒子外的每个粒子都被视为模仿者。此外，每个粒子还会通过随机惯性分量进行更新，这有助于增加群体的多样性，并提升标准 PSO 算法的全局搜索能力 [14]。SL-PSO 的设计也使其能够高效优化维度较高的问题 [13], ，因此适合作为参数数量较多的网络的基准算法。

灰狼优化算法 [2] 由 Mirjalili 提出，其灵感源自狼在野外互动和狩猎时所遵循的等级结构。这种结构将狼按适应度从高到低依次划分为阿尔法狼、贝塔狼、德尔塔狼（优势狼）和欧米伽狼（非优势狼）。在算法中，这一结构通过一种类似于 SL-PSO 社会学习机制的方法进行数学建模：非优势欧米伽狼从优势狼的均值中学习。关于灰狼优化算法实现的进一步说明见第 4 节。由于 SL-PSO 和灰狼优化算法在社会学习方面具有相似性，对二者进行比较有助于了解它们对优化问题的相对影响。灰狼优化算法以全局搜索效率高、收敛能力强著称 [2]，而 SL-PSO 的开发能力则有所欠缺能力 [14]。因此，灰狼优化算法是本研究中一种合理的实现方案。

### *C.* 神经网络群体元启发式算法

如前所述，可供选择的元启发式算法种类繁多。一项研究发现，基于种群的元启发式算法可能比精确优化算法更高效，因为后者难以解决高维空间中的问题 [15]。此外，作者认为，为了在每个问题上取得良好效果，这些算法必须采用不同的假设。基于种群的算法通常只需采用少得多的假设，甚至完全不对问题作任何假设，因此能够避开这一问题。正如上一节所述，与梯度下降方法不同，这些算法能够搜索较大的决策空间，也不容易陷入局部极小值。有人对使用元启发式算法训练前馈神经网络的性能进行了评估综述 [16]。他们分析了多种算法，在某些情况下，这些算法在低维神经网络上的表现与基于梯度的方法相当，甚至更胜一筹 [17], [18]。许多基于种群的元启发式算法具有很强的探索和开发能力，而基于梯度的方法则只专注于开发。然而，当元启发式算法用于训练深度神经网络时，问题便随之而来：现代架构中的决策变量数量会从几十个增加到数百万个。关于深度神经网络，尤其是卷积神经网络优化的大量文献，主要聚焦于优化神经网络架构 [19], [20]。这种方法能够针对给定问题生成最优网络；然而，本研究旨在设计一种用于训练网络权重的算法。文献中有多个实例 [5], [7], [11] 展示了不同研究者将基于梯度的方法与元启发式算法相结合。这种做法通过在两类方法之间交换信息来训练神经网络。例如，Albeahdili 等人 [11] 将粒子群优化和遗传算法结合，构建了一种混合算法，并将其与随机梯度下降结合，在 CIFAR-10 上取得了极具竞争力的结果。该方法着重于训练元启发式算法，然后对这些个体使用随机梯度下降。运行多个随机梯度下降算法存在一个问题：在更复杂的网络上，这些算法并行运行需要很长时间，且计算成本高昂。不过，这种方法有一项显著优势：粒子先探索空间，然后对其应用梯度下降，以不同参数开发该空间。此外，遗传算法在不同个体之间共享信息具有巨大优势，因为这有助于更充分地探索搜索空间，并引导粒子避开局部极小值 [21]。文献有力地证明，将某些基于种群的算法与遗传改进相结合能够提升性能 [11], [22], [23]。回到 [11], 一个缺点，正如 [16], 的荟萃分析所指出的，尽管粒子群优化的收敛速度很快，但找到全局极小值的能力较弱。相反，正如前文所述

灰狼优化算法在高维空间中不仅收敛速度很快，而且具备寻找全局最优解的能力。此外，正如 [22] 所指出的，将灰狼优化算法与遗传算法结合，可以弥补该算法的一些不足。

### D. 总结

总而言之，本研究提出了一种算法，将元启发式算法与梯度下降法结合起来，充分发挥两者的优势。这使算法既能探索搜索空间，也能更好地开发搜索空间。由于深度神经网络模型的维度较高，随着模型复杂度增加，元启发式算法的开发环节将变得越来越困难。该算法以个体的形式在两种训练方式之间交换信息。不同于现有文献，本研究提出将灰狼优化算法与遗传算法相结合，并将其与一种梯度下降法（如随机梯度下降）结合。选择随机梯度下降而非 Adam，是因为随机梯度下降收敛更快，而且当网络权重不断切换时，其自适应特性效果不佳 [24]。这种方法具有新颖性，因为此前尚无人尝试使用这种混合算法优化深度神经网络权重。为避免过早收敛，本研究提出了遗传算法改进方案 [11], [22], ，仅在适应度停滞时才使用交叉和变异遗传算子。

## III. 所选架构及其合理性说明

卷积神经网络架构也是优化领域的热门话题，因为要找到适合特定问题的架构，可能是一个漫长而艰难的过程 [25]。在考察了一系列流行架构 [25], [26],后，我们发现，即便是其中最小的架构，特征数量仍达数十万。这种网络规模会使寻找最优权重集这一优化问题的代价非常高，因为每个特征都是一个需要训练的决策变量。为克服这一计算难题，我们改用规模较小的自定义网络架构，在 CIFAR-10 数据集上进行训练。该网络共有 58,685 个特征。图 1 展示了该架构的示意图。还值得一提的是，我们曾考虑采用迁移学习和分块训练，即使用预训练模型来提升模型的准确率，并在冻结的模型上附加若干未冻结层，供训练算法进行优化。这些方案可以解决维度过高的问题，但会偏离本研究的目的，因为它们无法有意义地发挥灰狼优化算法的全局搜索能力。

## IV. 训练算法

所实现的算法 GMW-SGD 显示在算法 IV-C 的伪代码中。所提出的算法是灰狼优化算法 [2] 与随机梯度下降相结合的混合算法，并偶尔进行变异或交叉。

首先，初始化个体种群。如 [2],所定义，个体表示狼群中的狼。每个个体代表卷积神经网络中的参数，可以是权重或偏置 [27], ，因此个体是网络的展平表示。对于包含数千个参数和许多层的大型卷积神经网络，编码策略会变得复杂，因此个体的简化编码表示如下：

$$ind(i) = [w_0, ..., w_{Nw}, b_0, ..., b_{Nb}], \tag{1}$$

$$population = [ind(1); ...; ind(Np)], \tag{2}$$

其中，$Np$为种群规模，$i$ = 1、.....、$Np$，$Nw$为网络中的权重数量，$Nb$ 为网络中的偏置数量。个体的适应度由式（9）所示的分类交叉熵损失函数计算，目标是将其最小化。按照原始灰狼优化算法，算法遵循一种以数学方式表示的灰狼社会等级体系。适应度最高的三个个体依次保存为阿尔法（$\alpha$）、贝塔（$\beta$）和德尔塔（$\delta$）狼，分别代表 $dominating$ 狼。其余个体则保存为欧米伽狼（$\omega$）。

### A. 灰狼优化算法

灰狼优化算法的基础是模拟野生环境中灰狼捕猎的各个阶段。该算法分为不同阶段：包围猎物、捕猎和攻击猎物。

在算法的每次迭代中，所有 $omega$狼都会根据 $dominant$ 狼的位置改变自身位置。假设这三只狼对猎物（即最优解）的潜在位置了解得更多。因此，狼群中的其余成员彼此分散，以搜索猎物；如式（6）和式（7）所示，利用 $A$和$C$ 实现这一过程。找到猎物后，它们便汇聚起来发起攻击。 $omega$个体的位置按以下公式更新：

$$\vec{D}_\alpha = |\vec{C_1} \cdot \vec{X}_\alpha(t) - \vec{X}(t)|, \tag{3a}$$
$$\vec{D}_\beta = |\vec{C_2} \cdot \vec{X}_\beta(t) - \vec{X}(t)|, \tag{3b}$$
$$\vec{D}_\delta = |\vec{C_3} \cdot \vec{X}_\delta(t) - \vec{X}(t)|, \tag{3c}$$

$$\vec{X}_m(t+1) = \vec{X}_l(t) - A_m \cdot D_l, \tag{4}$$

$$\vec{X}(t+1) = \frac{(\vec{X}_1 + \vec{X}_2 + \vec{X}_3)}{3}, \tag{5}$$

其中，$t$ 表示当前代，$X_l$ 是由 $l$ = {$\alpha$, $\beta$, $\delta$}指示的猎物位置。($\vec{X}$)是当前狼的新位置，$m$ = {1,2,3}。最后，$\vec{A}$ 和 $\vec{C}$ 是按如下方式计算的常数向量：

$$\vec{A} = 2\vec{a}\vec{r_1} - \vec{a}, \tag{6}$$

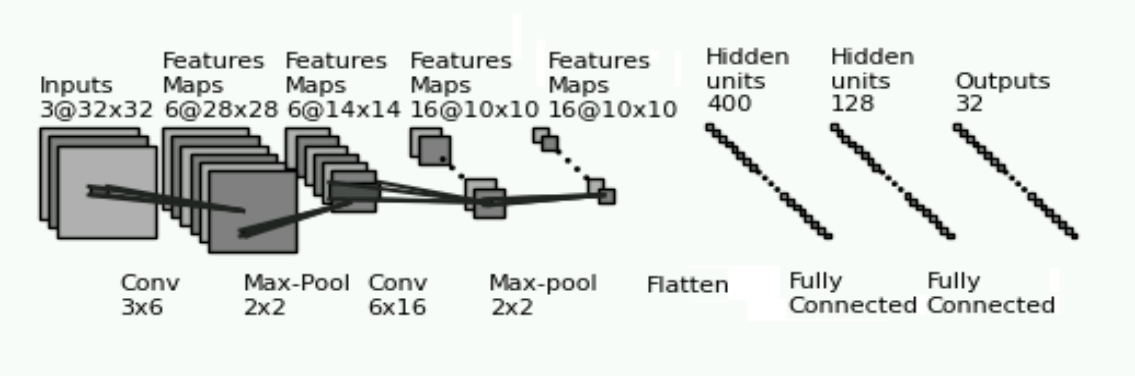


图 1. 所选神经网络架构

$$\vec{C} = 2\vec{r_2}, \tag{7}$$

其中， $r_1$、$r_2$ 是在范围 [0,1]内随机生成的向量，分量 $\vec{a}$ 在各代中从 2 线性递减至 0。

尽管相关工作中提到了灰狼优化算法的优点，但由于维度较高，该算法容易过早收敛并陷入局部极小值。为了解决这一问题，我们将采用遗传算法改进。

*B.* 遗传改进：灰狼优化算法与遗传算法的混合

通过利用遗传交叉和变异 [21],，个体在停滞时可以得到改进。因此，只有在达到 $patience$ 时，才会运行遗传算法，因为 $omega$ 狼的适应度没有提高。此外，种群是否会通过 $p_{mut}$进行变异或交叉也存在一定概率。

首先，为遵循该算法，只对 $omega$ 狼进行变异，因为 $dominant$ 狼知道猎物的位置，不应对其进行改动。每个个体都按以下公式进行变异：

$$p' \begin{cases} p + ((2u)^{\frac{1}{1+\eta_m}} - 1)(p - x_i^{(L)}), u \leq 0.5 & \text{(8a)} \\ p + (1 - (2(1-u))^{\frac{1}{1+\eta_m}})(x_i^{(U)} - p), u > 0.5 & \text{(8b)} \end{cases}$$

其中， $u$ 是在 0 到 1 范围内随机生成的， $\eta$ 是范围为 [20,100], 的常数，用于决定输出与亲代的相似程度；变异后的个体（$p'$）受 $x_i^{(L)}$ 和 $x_i^{(U)}$ 限制。较差的个体会被设置为进行更多变异，因为它们的表现较差。

其次，在所有 $omega$狼之间进行交叉，每只狼都会与一只随机的 $dominant$狼结合，如图 3 所示。与变异类似，随着个体适应度的提高，被修改的决策变量数量会减少，因此较优的个体不会受到太多改变。子代会从亲代双方各取一部分，而 $omega$ 狼则会从 $dominant$ 狼那里获取部分决策变量，从而更好地引导自身。

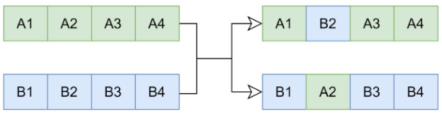


图 2. $omega$狼与 $dominant$ 狼之间的交叉

为进一步解决进化算法难以充分利用高维决策空间的问题，将上述元启发式算法与梯度下降法相结合，以便在探索之后进一步利用该空间。

*C.* 随机梯度下降

最后，该算法采用随机梯度下降，在元启发式步骤更新所有个体后，对三个最优个体（$\alpha$，$\beta$，$\delta$）进行运算。这与灰狼优化算法自然契合，因为最优个体会用来引导狼群中的其他个体。这模拟了 $dominant$ 狼群捕猎猎物的过程，因为这些狼处于狼群中的最佳位置。训练所用的损失函数为分类交叉熵（$CE$），如下所示：

$$CE = -\log(\frac{e^{s_p}}{\Sigma_j^c e^{s_j}}) \tag{9}$$

其中， $C$ 表示类别， $s$ 表示某个类别的分数， $p$ 表示正确/正向类别。

更新 $dominant$ 狼后，其余狼仍会探索搜索空间，但随着时间推移会开始朝最优狼移动。随着它们逐渐靠近，可能会发现另一个猎物（即更优的极小值）。该算法最后添加的优化特性是缓慢降低学习率，并通过随机梯度下降实现，使优化器采取更小的步长，从而更接近极小值。

GMW-SGD 将灰狼优化算法带来的强大探索能力与遗传改进相结合，进一步增强了算法，以克服原始 GWO 的一些缺陷。最后，进化部分与梯度下降部分通过

$dominant$ 狼之间交换信息，并利用随机梯度下降帮助开发已找到的最优解。

**Algorithm 1** Pseudocode of GMW-SGD

**Input:** Maximum number of evolutions ($NEvol$), maximum number of generations ($NGen$), maximum number of epochs ($NEpoch$). The population ($P$) of Np individuals ($i$), with the top three individuals ($alpha(i)$, $beta(i)$ and $delta(i)$) called $dominators$ and the rest, $omegas$. Patience ($pat$) defines the period before genetic modifications, probability of mutation/crossover happening ($p_{mut}$) , neural network model ($m$), with a learning rate ($lr$).

**Output:** The best individual, $alpha(i)$.

Initialise random population $P$ with $Np$ individuals according to Eqs. (1) and (2).
Initialise the constants $A$ and $C$ for $i^{th}$ according to Eqs. (6) and (7), respectively.
Evaluate the fitness of $i$ in $P$ according to Eq. (9).
Sort the individuals based on their fitness and select $dominators$ and $omegas$.
**while** $t \leq NEvol$ **do**
  **for** $g$ in $NGen$ **do**
    **for** individual $j$ in $omega$ **do**
      Update $j$ based on Eqs. (3-5),
    **end for**
    Update the constants $A$ and $C$ for $i^{th}$ according to Eqs. (6) and (7), respectively.
    Repeat Step 3.
    Repeat Step 4.
    **if** average loss across past generations didn't improve for $pat$ **then**
      **for** individual $j$ in $omega$ **do**
        Randomly generate $r$ between [0,1].
        **if** $r < p_{mut}$ **then**
          Mutate $j$ according to Eq. 8a and 8b.
        **else**
          Crossover $j$ with random $dominant(i)$ according to Fig. 2.
        **end if**
      **end for**
    **end if**
  **end for**
  **for** individual $j$ in $dominant$ **do**
    Update $m$ with the weights from $j$.
    **for** $e$ in $NEpoch$ **do**
      Train and validate $m$ with SGD.
      Update $j$ with the new weights generated.
    **end for**
  **end for**
  **if** $t\%2 == 0$ **then**
    $lr \leftarrow lr \cdot 0.1$
  **end if**
**end while**

## V. 结果

通过与另外两种训练算法进行基准比较，测试了所提出的算法的性能。选用的算法如下：

1) SL-PSO——一种基于种群的算法，与该算法中的元启发式部分密切相关。2) SGD——一种配有学习率调度器的基于梯度的优化器。

训练过程中使用的参数如表 I 所示。

表 I 算法的参数

| 算法 | 参数 | 值 |
|---|---|---|
| SGD | 学习率（$\alpha$） | 0.01 |
| | $\alpha$ 调度器因子 | 0.1 |
| SL-PSO | 种群规模（$Np$） | 60 |
| | 维度（$d$） | 58,685 |
| | 迭代次数（$NEvol$） | 36 |
| | 位置 | [-0.1, 0.1] |
| | 速度 | [-0.01, 0.01] |
| | 常数 $\alpha$ | 0.5 |
| | 常数 $\beta$ | 0.0001 |
| | 常数 $c1$和 $c2$ | 2 |
| GMW-SGD | $NEvol$ | 10 |
| | $d$ | 58,685 |
| | 迭代轮数（$NEpoch$） | 2 |
| | 代数（$NGen$） | 14 |
| | $Np$ | 15 |
| | 常数 $a$ | [1,0] |
| | 耐心值 | 4 |
| | 变异/交叉概率（$p_{mut}$） | 0.7 |
| | 修改率 | [0.6, 0.1] |

这些算法分别在 CIFAR-10 上独立运行了一次，结果已纳入表 II 和图 3。两个元启发式算法均设置为运行 2160 次函数评估。不过，随机梯度下降运行的次数较少，因为它很快就收敛并陷入了极小值。因此，当其约 20 个迭代轮数内没有提升时，便提前停止，并假设模型不会随时间变化（即使发生变化，也会在训练数据集上过拟合）。

通过测试模型在 CIFAR-10 训练集和测试集上的准确率，评估了模型的性能。此外，还使用交叉熵评估测试数据集上的损失。

分析结果后可以得出以下结论。图 3 中可以注意到的首要区别是 SL-PSO 与随机梯度下降之间的差异。这体现了这两类算法的主要区别。随机梯度下降会立即利用其初始所在的决策空间，并成功更新权重，从而迅速达到更高的

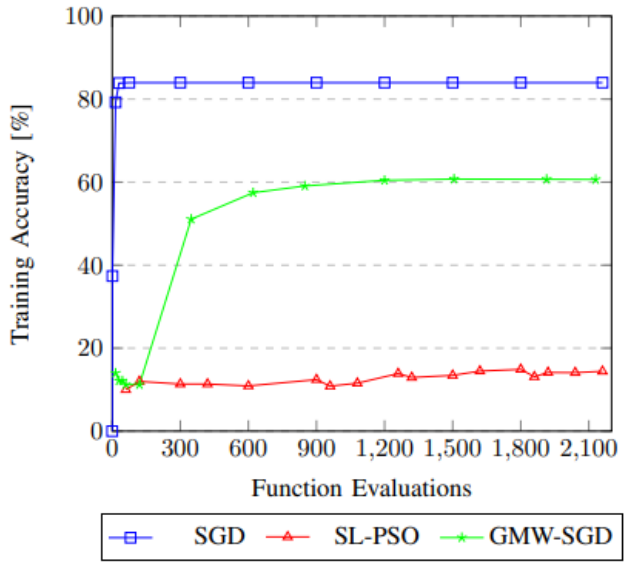


图 3. CIFAR-10 上最佳训练准确率的收敛情况

表 II CIFAR-10 上算法的结果比较

| 已评估算法 | 评估结果 | | |
|---|---|---|---|
| | 训练准确率 | 测试准确率 | *CE* |
| SGD | 83.97% | 62.99% | 1.3921 |
| SL-PSO | 14.42% | 14.42% | 2.2795 |
| GMW-SGD | 60.65% | 60.63% | 1.1170 |

准确率。另一方面，SL-PSO 随时间稳步提升，但其带来的改进幅度极小。

使用 GMW-SGD 可以看出，该算法兼具两类算法的特点。与纯粹的梯度下降方法不同，它的收敛时间较长，而且在整个训练过程中，个体都会探索搜索空间，以寻找可供利用的最佳解。可以看出，该算法在训练进行到一半左右时开始收敛，即便如此，之后的提升幅度仍然很小。此外，与标准群体式元启发式算法相比，它显然更适合这种训练。SL-PSO 可能会因过早收敛而无法达到较高的准确率；即使能够达到，也需要很长时间。

尽管 GMW-SGD 的训练准确率可能不如随机梯度下降那么高，但表 II 表明，随机梯度下降在训练数据上存在严重过拟合。随机梯度下降与 GMW-SGD 在测试数据集上的分类结果差异相当，前者的表现仅略好一些。此外，GMW-SGD 在测试数据集上的损失小于使用随机梯度下降训练的模型。这些结果表明，所提出的算法几乎可以达到标准基于梯度的方法的性能，甚至可能不相上下。正如表 II 所示，训练后的模型对测试数据集的泛化能力极强，因此完全没有过拟合。

所提出的算法兼顾了开发与探索，因此性能非常相近。因此，原本预期它会优于随机梯度下降，但它最终只能达到相当的指标。两种算法都无法表现得更好的一个可能原因是所使用的架构。如前所述，为了降低元启发式算法的维度，采用了一个简单的网络。如今大多数解决方案会针对这类任务使用更深、更复杂的神经网络，而所提出的神经网络容量不足，无法取得同样好的表现。因此，这可能会影响随机梯度下降和 GMW-SGD 的性能。其次，尽管按照深度学习的标准，该网络的维度相对较小（例如 EfficientNet 或 MobileNet 拥有超过一百万个参数），但对于元启发式算法而言，这些维度仍然极高，因为它会大幅增加需要调整的决策变量数量。因此，该算法可能无法取得更好的表现，因而可能主要依赖随机梯度下降。如果决策空间更小，该算法或许能更多地受益于对其他个体的探索。

## VI. 双目标问题

本文还展示了通过非支配排序将训练作为双目标问题处理的方法。Kalyanmoy 等人在 [28] 中提出了非支配排序遗传算法 II（NSGA-II），以减轻常规 NSGA 的一些缺陷，例如计算复杂度高和采用非精英策略。在所提出的方法中，子代种群生成后，算法会将父代种群和子代种群中的个体分别排序到不同的非支配前沿中。该策略用于确定哪些个体应被选入下一代。这个双目标训练问题考虑了两个目标：准确率最大化，以及高斯正则化项最小化（即权重平方和最小化）。本实验所用参数与上一节相同，见表 I。

总体来看，由于多种原因，GMW-SGD 优化器算法在以双目标问题进行训练时未能取得成功。本研究试图解决的图像分类问题，优先考虑较高的准确率，而非网络权重总和。非支配排序算法的特性会导致精英保留策略失效，使准确率高且高斯正则化值较高的个体被淘汰，转而优先保留相关性较低的个体。灰狼优化算法是一种社会学习算法，依靠引领狼群的 3 个最成功个体。在这种情况下，这些个体并不能很好地代表最成功的狼，且往往是高斯正则化器值较低、收敛于局部最小值的个体。这种选择方式阻碍了个体向表现更好的个体学习，也使算法无法达到全局最小值，而是收敛到某个局部最小值。总而言之，这个优化问题并不适合进行双目标

目标优化，作为单一优化问题时表现则显著更好。表 III 展示了一组拥挤距离最大的个体，它们最能代表帕累托最优前沿。

表 III 双目标优化中的帕累托前沿所选个体

| 帕累托前沿 | 评估指标 | |
|---|---|---|
| 个体 | 准确率 | 高斯正则化项 |
| 个体1 | 10.69 | 7.4533 |
| 个体2 | 24.85 | 11.7551 |
| 个体3 | 23.49 | 11.3150 |
| 个体4 | 11.14 | 7.6071 |
| 个体5 | 16.28 | 9.8632 |
| 个体6 | 13.4 | 9.3054 |

## VII. 结论与未来工作

尽管存在许多经典的元启发式优化技术，但可以得出结论：它们的收敛效率不如基于梯度的方法。然而，借助充足的计算资源，它们有可能达到相同的最优解。本文提出了一种与随机梯度下降相结合的混合进化算法，采用了一种新颖的变异实现方式，使粒子在空间中呈现异质行为。GMW-SGD 将探索与开发结合起来，其性能优于基准元启发式算法，同时达到了标准随机梯度下降的准确率。此外，GMW-SGD 还具有改进的基于种群的全局搜索能力，从而更有效地避免局部极小值。

未来可以通过让每种算法执行更多次函数评估来进一步改进本研究。因为在梯度下降法收敛到最优解后，元启发式算法将获得更多机会来展现其全局搜索能力。

除了遗传算法的变异和交叉外，还可以引入更多随机性，使欧米伽狼朝优势狼均值方向迈出的步伐更加多样。与 GMW-SGD 的标准实现相比，这应能以较不系统化的方式进一步探索。

为了解决使用元启发式算法优化深度神经网络权重的问题，可以将训练过程拆分为多个训练块。也就是说，训练过程将按层训练权重，而不是将整个网络作为一个整体进行训练。这一特性使 GWO-SGD 能够在维持低维度的同时训练深度神经网络。

## 参考文献


[1] Alex Krizhevsky、Geoffrey Hinton 等。通过微小图像学习多层特征。2009 年。

[2] Seyedali Mirjalili、Seyed Mohammad Mirjalili 和 Andrew Lewis。灰狼优化器。《Advances in Engineering Software》，69:46–61，2014年。[3] Gao Huang、Zhuang Liu、Laurens van der Maaten 和 Kilian Q. Weinberger。密集连接卷积网络。发表于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，2017年7月。[4] Mingxing Tan 和 Quoc Le。EfficientNetV2：更小的模型与更快的训练。发表于国际机器学习会议，页码10096–10106。PMLR，2021年。[5] Yu Xue、Yiling Tong 和 Ferra nte Neri。差分进化与 Adam 相结合的集成方法，用于训练前馈神经网络。《Information Sciences》，608:453–471，2022年。[6] Herbert Robbins 和 Sutton Monro。一种随机逼近法。《The Annals of Mathematical Statistics》，页码400–407，1951年。[7] Xiaodong Cui、Wei Zhang、Zolt´an T¨uske 和 Michael Picheny。用于优化深度神经网络的进化随机梯度下降。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编，《神经信息处理系统进展》，第31卷。Curran Associates, Inc.，2018年。[8] Shima Amirsadri、Seyed Jalaleddin Mousavirad 和 Hossein Ebrahimpour-Komleh。一种基于莱维飞行的灰狼优化器，与反向传播算法结合用于神经网络训练。《Neural Comput. Appl.》，30(12):3707–3720，2018年12月。[9] John McCall。用于建模与优化的遗传算法。《Journal of Computational and Applied Mathematics》，184 222，2005 1 205( ): – 。免疫学应用数学特刊。[10] David H Wolpert 和 William G Macready。优化的无免费午餐定理。《IEEE Transactions on Evolutionary Computation》，1(1):67–82，1997年。[11] Hayder M Albeahdili、Tony Han 和 Naz E Islam。用于优化卷积神经网络训练的混合算法。《Int J Adv Comput Sci Appl》，1(6):79–85，2015年。[12] James Kennedy 和 Russell Eberhart。粒子群优化。发表于 ICNN'95-国际神经网络会议论文集，第4卷，页码1942–1948。IEEE，1995年。[13] Ran Cheng 和 Yaochu Jin。一种用于可扩展优化的社会学习粒子群优化算法。《Information Sciences》，291:43–60，2015年。[14] Xinming Zhang、Xia Wang、Qiang Kang 和 Jinfeng Cheng。差分变异与新型社会学习粒子群优化算法。《Information Sciences》，480:109–129，2019年。[15] Zahra Beheshti 和 Siti Mariyam Hj Shamsuddin。基于种群的元启发式算法综述。《Int. J. Adv. Soft Comput. Appl》，5(1):1–35，2013年。[16] Zhonghuan Tian 和 Simon Fong。深度学习训练的元启发式算法综述。载于 Ozgur Baskan 编，《Optimization Algorithms》，第9章。IntechOpen，里耶卡，2016年。[17] Seyedali Mirjalili。灰狼优化器在训练多层感知器方面的效果如何？《Applied Intelligence》，43(1):150–161，2015年。[18] Mohammad Reza Mosavi、Mohammad Khishe 和 Abolfazl Ghamgosar。使用灰狼优化训练的神经网络对声呐数据集进行分类。《Neural Network World》，26(4):393，2016年。[19] Raphael de Lima Mendes、Alexandre Henrick da Silva Alves、Matheus de Souza Gomes、Pedro Luiz Lima Bertarini 和 Laurence Rodrigues do Amaral。GACNN：结合卷积神经网络与遗传算法，构建优化的混合分类架构。发表于2021年 IEEE进化计算大会（CEC），页码79–86。IEEE，2021年。[20] Bin Wang、Bing Xue 和 Mengjie Zhang。通过演化和堆叠可迁移模块，利用粒子群优化演化用于图像分类的深度神经网络。发表于2020年 IEEE进化计算大会（CEC），页码1–8，2020年。[21] John H. Holland。遗传算法。《Scientific American》，267(1):66–73，1992年。[22] Ebenezer Daniel。采用混合遗传算法—灰狼优化算法的基于小波的同态医学图像融合优化方法。《IEEE Sensors Journal》，18(16):6804–6811，2018年。[23] Po-Han Chen 和 Seyed Mohsen Shahandashti。遗传算法与模拟退火相结合，用于具有多重资源约束的多项目调度。《Automation in Construction》，18(4):434–443，2009年。[24] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980，2014年。

[25] Asifullah Khan、Anabia Sohail、Umme Zahoora 和 Aqsa Saeed Qureshi。深度卷积神经网络最新架构综述。《人工智能评论》，53(8):5455–5516，2020年。[26] Raimi Karim。图解：10种卷积神经网络架构，2022年1月。[27] Jing-Ru Zhang、Jun Zhang、Tat-Ming Lok 和 Michael R. Lyu。一种用于前馈神经网络训练的混合粒子群优化—反向传播算法。《应用数学与计算》，185(2):1026–1037，2007年。智能计算理论与方法论特刊。[28] K. Deb、A. Pratap、S. Agarwal 和 T. Meyarivan。一种快速且精英化的多目标遗传算法：NSGA-II。《*IEEE Transactions on Evolutionary Computation*》，6(2):182–197，2002年。