Advances in Applied Mathematics 应用数学进展, 2023, 12(10), 4248-4254
Published Online October 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1210418

Hans汉斯

一种关于有效步长约束的自适应算法

姜文翰[1]，刘曜齐[2]，姜志侠[1*]

[1]长春理工大学，数学与统计学院，吉林 长春
[2]吉林大学，汽车工程学院，吉林 长春

收稿日期：2023年9月13日；录用日期：2023年10月8日；发布日期：2023年10月16日

摘 要

鉴于Adam算法在迭代后期因有效步长过大而导致算法的收敛性能下降，本研究提出了一种名为MAXGrad的优化算法。MAXGrad通过修改二阶矩的迭代公式以限制有效步长的增长。为深入评估MAXGrad算法的实际应用和性能，本文扩展了实验范围，采用三个较大规模的数据集，并与SGDM、Adam以及AMSGrad等算法进行了详细比较。实验结果清晰表明，在多个数据集上，MAXGrad算法相对于Adam和AMSGrad等自适应算法均取得了显著的性能改进。这些结果充分验证了MAXGrad算法作为一种全新的有效步长迭代算法的可行性和卓越性能。

关键词

MAXGrad算法，自适应算法，收敛性能，机器学习

An Adaptive Algorithm on Effective Step Size Constraints

Wenhan Jiang[1], Yaoqi Liu[2], Zhixia Jiang[1*]

[1]School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun Jilin
[2]School of Automotive Engineering, Jilin University, Changchun Jilin

Received: Sep. 13th, 2023; accepted: Oct. 8th, 2023; published: Oct. 16th, 2023

Abstract

In view of the fact that the convergence performance of Adam's algorithm is degraded at the later stage of iteration due to the excessively large effective step size, an optimization algorithm named

*通讯作者。

文章引用: 姜文翰, 刘曜齐, 姜志侠. 一种关于有效步长约束的自适应算法[J]. 应用数学进展, 2023, 12(10): 4248-4254.
DOI: 10.12677/aam.2023.1210418

MAXGrad is proposed in this study. MAXGrad limits the growth of the effective step size by modifying the iterative formulation of the second-order moments. In order to evaluate the practical application and performance of the MAXGrad algorithm in depth, this paper extends the experimental scope by using three larger-scale datasets and compares them in detail with the algorithms of SGDM, Adam, and AMSGrad. The experimental results clearly show that the MAXGrad algorithm achieves significant performance improvements over adaptive algorithms such as Adam and AMSGrad on multiple datasets. These results fully validate the feasibility and superior performance of the MAXGrad algorithm as a new effective step-size iterative algorithm.

Keywords

MAXGrad Algorithm, Adaptive Algorithm, Convergence Performance, Machine Learning

1. 引言

近些年，随着神经网络在图像处理[1] [2]、自然语言处理[3] [4]、医学治疗[5] [6]等领域取得突破进展，作为优化神经网络参数主要方法的梯度下降算法，已逐渐受到学者的广泛关注[7]。一个主流的优化算法是随机梯度下降算法(SGD)，它以固定步长将模型参数朝着损失函数的负梯度方向更新。然而，SGD算法存在一个问题，它只考虑了当前步的梯度作为参数更新方向，容易导致算法陷入局部最小值。为了克服这个问题，带动量的随机梯度下降算法(SGDM)采用了梯度的指数加权移动平均值(EMA)作为迭代方向。另一方面，SGD 算法中梯度的均匀缩放会导致优化效果不佳，AdaGrad 算法[8]使用历史梯度平方和的信息来实现自适应步长。在梯度稀疏或较小时，AdaGrad 算法的性能明显优于 SGD 算法。

AdaGrad 算法是最早提出的自适应算法，它根据每个参数的历史梯度信息来调整学习率。虽然该算法在稀疏环境下表现良好，但在梯度密集的情况下，学习率会迅速减小，导致收敛速度下降。RMSProp 算法改进了 AdaGrad 算法，通过使用梯度平方的指数移动平均来减轻学习率的快速衰减问题。这使得 RMSProp 算法在更广泛的问题中表现更好。Adam [9]算法进一步综合了动量项和 RMSProp 算法的思想。它引入了动量项来增加梯度更新的惯性，同时使用梯度平方的指数移动平均来自适应地调整学习率。Adam 算法在很多应用中取得了良好的性能，但在迭代后期可能出现有效步长过大($\alpha/\sqrt{\boldsymbol{v}_t}$)的问题[10]。AMSGrad 算法是对 Adam 算法的改进，它使用历史上所有梯度平方的最大值来更新学习率，以解决 Adam 算法可能出现学习率问题的情况。虽然 AMSGrad 算法在理论上提供了一些保证，但在实际应用中表现不一定出色。因此有必要进一步改进 Adam 算法和 AMSGrad 算法。

本文贡献如下：

1) 通过对 $\boldsymbol{v}_t$ 项修正提出 MAXGrad 算法，使得 Adam 算法在不附加额外条件的情况下具有更好的收敛性。

2) 与 AMSGrad 算法相比，MAXGrad 算法能够维持对过去梯度的长期记忆，而且不需要对新变量 $\hat{v}_t$ 进行额外初始化。

3) 本研究还进行了初步实验，结果表明 MAXGrad 算法在机器学习任务中，如图像分类和语言建模问题上，表现出色。这些结果暗示了 MAXGrad 算法在实际应用中的潜在价值。

2. MAXGrad 算法

2.1. 符号说明

本文用 S_d^+ 表示所有 $d\times d$ 正定矩阵的集合。设 $\boldsymbol{a}\in\mathbb{R}^d$、$A\in S_d^+$，用 $\boldsymbol{a}/A$ 表示 $A^{-1}\boldsymbol{a}$，$\sqrt{A}$ 表示 $A^{1/2}$。对于 $\mathbb{R}^d$ 中的任意向量 $\boldsymbol{a}$，$\boldsymbol{b}$，使用 $\sqrt{\boldsymbol{a}}$，$\boldsymbol{a}^2$，$\max(\boldsymbol{a},\boldsymbol{b})$ 表示按向量中的元素进行平方根、平方、最大值运算。对于 $A\in S_d^+$ 时，将投影操作[11]定义为 $\Pi_{\mathcal{F},A}(\boldsymbol{y})$，即对于任意的 $\boldsymbol{y}\in\mathbb{R}^d$，$\mathcal{F}\subset\mathbb{R}^d$，$\arg\min_{x\in\mathcal{F}}\left\|A^{1/2}(\boldsymbol{x}-\boldsymbol{y})\right\|$。最后将损失函数定义为 $f_t(\boldsymbol{x}_t)$，其中 $\boldsymbol{x}_t$ 是需要优化的参数。

2.2. MAXGrad 算法的提出

Luo 等人[12]指出 AMSGrad 算法虽然给出 Adam 算法有效步长的修正方案，但在实际应用中，该算法可能并不会有较好的表现。AMSGrad 算法通过引入一个新的变量 $\hat{\boldsymbol{v}}_t$ 用来更新有效步长($\alpha/\sqrt{\hat{\boldsymbol{v}}_t}$)。但对于二阶矩 $\boldsymbol{v}_t$ 的迭代过程，AMSGrad 算法并未进行更改，这可能会导致 $\boldsymbol{v}_t$ 的值仍然会出现较大情况，进而影响算法的收敛性。本文提出 MAXGrad 算法对 $\boldsymbol{v}_t$ 的迭代方式进行修改，见表 1。

Table 1. AMSGrad algorithm and MAXGrad algorithm pseudo-code
表 1. AMSGrad 算法与 MAXGrad 算法伪代码

算法 1 AMSGrad	算法 2MAXGrad
输入： $\boldsymbol{x}_1\in\mathcal{F}$，$\{\beta_{1t}\}_{t=1}^T$，$\beta_2$，$\{\alpha_t\}_{t=1}^T$，$\epsilon$	**输入：** $\boldsymbol{x}_1\in\mathcal{F}$，$\{\beta_{1t}\}_{t=1}^T$，$\beta_2$，$\{\alpha_t\}_{t=1}^T$，$\epsilon$
初始化： $\boldsymbol{m}_0=0$，$\boldsymbol{v}_0=0$，$\hat{\boldsymbol{v}}_0=0$	**初始化：** $\boldsymbol{m}_0=0$，$\boldsymbol{v}_0=0$，
1: **for** $t=1$ to T **do**	1: **for** $t=1$ to T **do**
2: $\boldsymbol{g}_t=\nabla_x f_t(\boldsymbol{x}_t)$	2: $\boldsymbol{g}_t=\nabla_x f_t(\boldsymbol{x}_t)$
3: $\boldsymbol{m}_t=\beta_{1t}\boldsymbol{m}_{t-1}+(1-\beta_{1t})\boldsymbol{g}_t$	3: $\boldsymbol{m}_t=\beta_{1t}\boldsymbol{m}_{t-1}+(1-\beta_{1t})\boldsymbol{g}_t$
4: $\boldsymbol{v}_t=\beta_2\boldsymbol{v}_t+(1-\beta_2)\boldsymbol{g}_t^2$ 5: $\hat{\boldsymbol{v}}_t=\max(\hat{\boldsymbol{v}}_{t-1},\boldsymbol{v}_t)$ and $V_t=\mathrm{diag}(\hat{\boldsymbol{v}}_t+\epsilon)$	4: $\boldsymbol{v}_t=\max(\boldsymbol{v}_{t-1},\beta_2\boldsymbol{v}_{t-1}+(1-\beta_2)\boldsymbol{g}_t^2)$ and $V_t=\mathrm{diag}(\boldsymbol{v}_t+\epsilon)$
6: $\boldsymbol{x}_{t+1}=\Pi_{\mathcal{F},\sqrt{V_t}}(\boldsymbol{x}_t-\alpha_t V_t^{-1/2}m_t)$	5: $\boldsymbol{x}_{t+1}=\Pi_{\mathcal{F},\sqrt{V_t}}(\boldsymbol{x}_t-\alpha_t V_t^{-1/2}\boldsymbol{m}_t)$
7: **end for**	6: **end for**

AMSGrad 算法维护一个历史梯度平方的移动平均 $\hat{\boldsymbol{v}}_t$，并在更新参数时使用它来调整学习率。在每个时间步 t，AMSGrad 首先计算当前梯度的平方 $\boldsymbol{g}_t^2$，然后将其与上一时间步的 $\hat{\boldsymbol{v}}_{t-1}$ 进行比较，较大者赋给 $\hat{\boldsymbol{v}}_t$。最后，参数更新采用了基于 $\hat{\boldsymbol{v}}_t$ 的学习率调整，以避免梯度震荡。MAXGrad 算法与 AMSGrad 算法类似，但有所不同之处在于历史梯度平方的比较。在 MAXGrad 中，历史梯度平方 $\boldsymbol{v}_t$ 的更新是通过比较当前梯度平方 $\boldsymbol{g}_t^2$ 和上一时间步的 $\boldsymbol{v}_{t-1}$ 进行的，选择较大者作为新的 $\boldsymbol{v}_t$。这一不同之处使得 $\boldsymbol{v}_t$ 呈现出不减趋势，这是 MAXGrad 算法的独特特点。

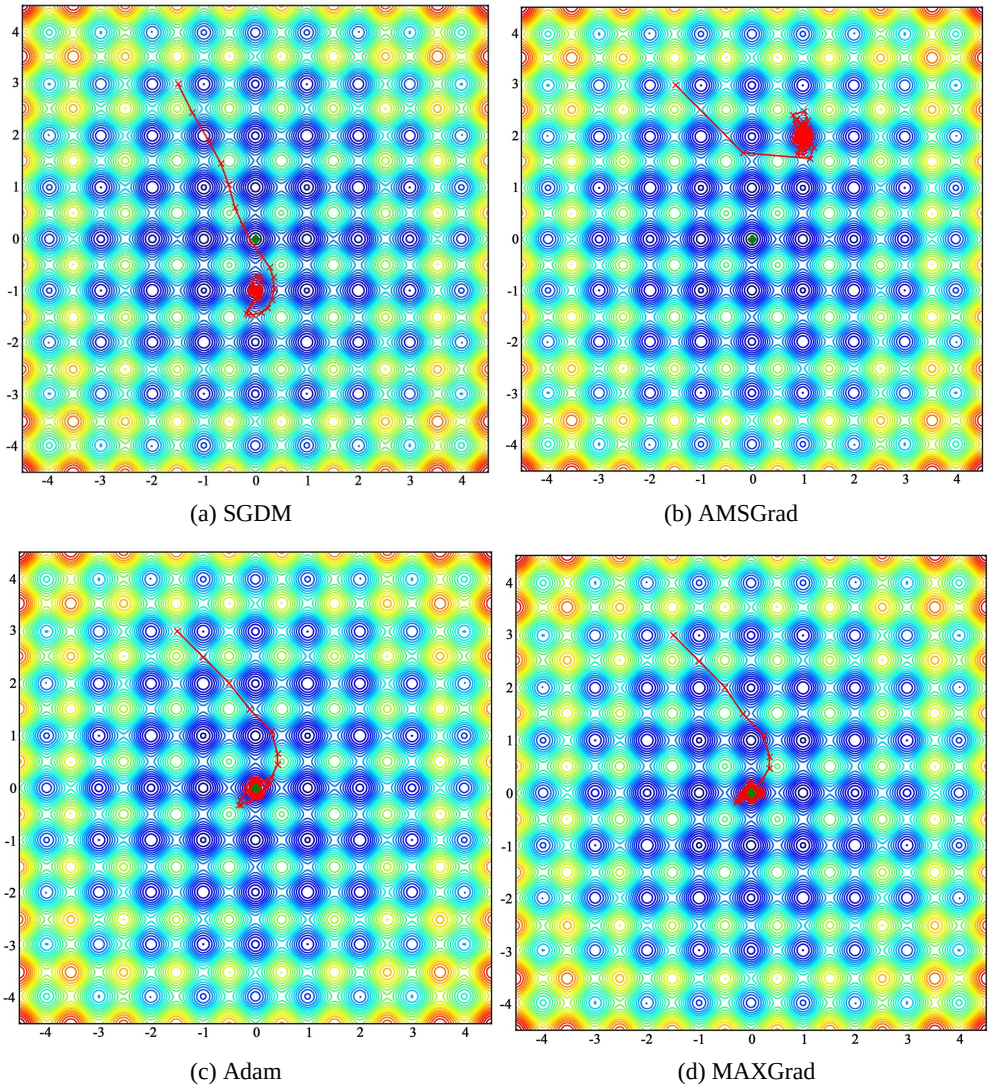
3. 实验结果

3.1. Rastrign 测试函数

Rastrign 测试函数是一个标准的测试函数，定义如下：

$$f\left(x_1,x_2\right)=10+x_1^2+x_2^2-10\left(\cos\left(2\pi x_1\right)+\cos\left(2\pi x_2\right)\right),\quad x_1,x_2\in\left(-5.12,5.12\right),$$

其中全局最小点是在$(0,0)$处取得。如图 1 所示，SGDM 算法在搜索过程中表现出对全局最小值的跳跃特性。这意味着它具有一定的探索性，有时会离开当前位置以寻找更好的解决方案。然而，这也可能导致它偶尔跳过全局最小值，因为它的动量可能使其跳出最小值区域。而 AMSGrad 算法设计目的是克服 Adam 算法中学习率下降过快的问题。AMSGrad 倾向于更稳定地收敛，通常会收敛到局部最小值。相对于 Adam 算法，MAXGrad 算法在搜索全局最小值时表现出更加平滑的特性。这种平滑性可能有助于它更可靠地找到全局最小值。



(a) SGDM (b) AMSGrad

(c) Adam (d) MAXGrad

Figure 1. Convergence trajectories of the four algorithms
图 1. 四种算法的收敛轨迹

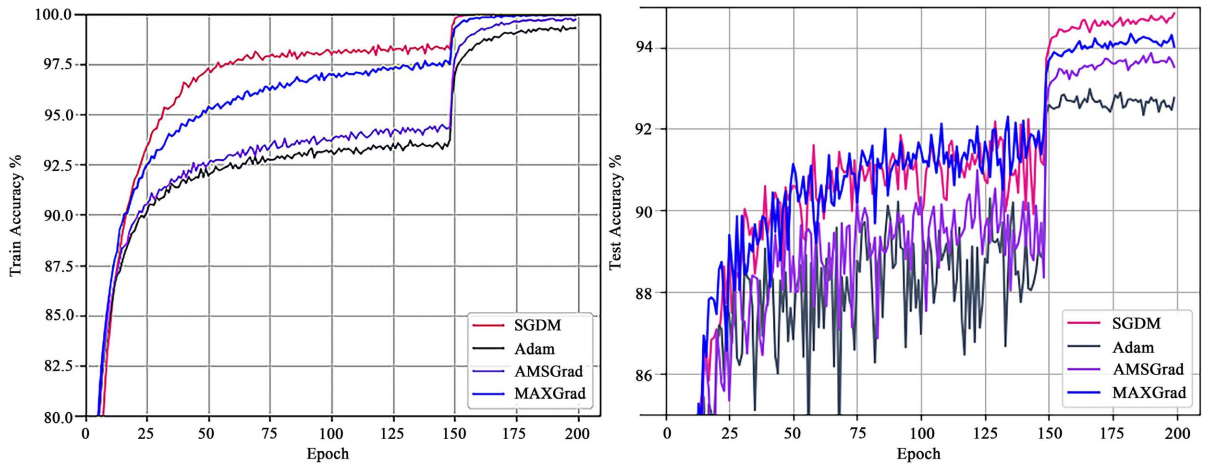
为了检验 MAXGrad 算法的性能，本文选取三个真实数据集与 SGDM、Adam、AMSGrad 算法进行对比。在表 2 中对优化算法的超参数进行设置。电脑配置如下，CPU 使用 AMD Ryzen 7 6800H，GPU 使用 NVIDIA GeForce RTX-3050。

Table 2. Hyperparameter settings in the field of image classification
表 2. 在图像分类领域的超参数设置

优化器	学习率	β_1	β_2	权重系数	迭代次数
SGDM	0.1	0.9	--	0.0005	200
Adam	0.001	0.9	0.999	0.0005	200
AMSGrad	0.001	0.9	0.999	0.0005	200
MAXGrad	0.001	0.9	0.999	0.0005	200

3.2. CIFAR-10 数据集

CIFAR-10 数据集是一个用于图像分类的标准数据集，它包含了 10 个不同的图像类别。该数据集分为两部分：训练集和测试集。在训练集中，共有 50,000 张图片作为样本进行模型训练。而测试集则包含 10,000 张图片，用于评估模型的性能。每个图像都被分配到这 10 个类别中的一个，任务是训练一个模型来自动识别图像并将其分类到正确的类别中。如图 2 所示，MAXGrad 算法虽然表现没有 SGDM 算法出色，但相比于同为自适应算法的 Adam 与 AMSGrad，不论在训练集中还是在测试集中都有更快的收敛速度和更高的分类准确率。



(a) Accuracy of the CIFAR-10 training set (b) Accuracy of the CIFAR-10 test set

Figure 2. CIFAR-10 image classification using ResNet-34
图 2. 使用 ResNet-34 进行 CIFAR-10 图像分类

3.3. CIFAR-100 数据集

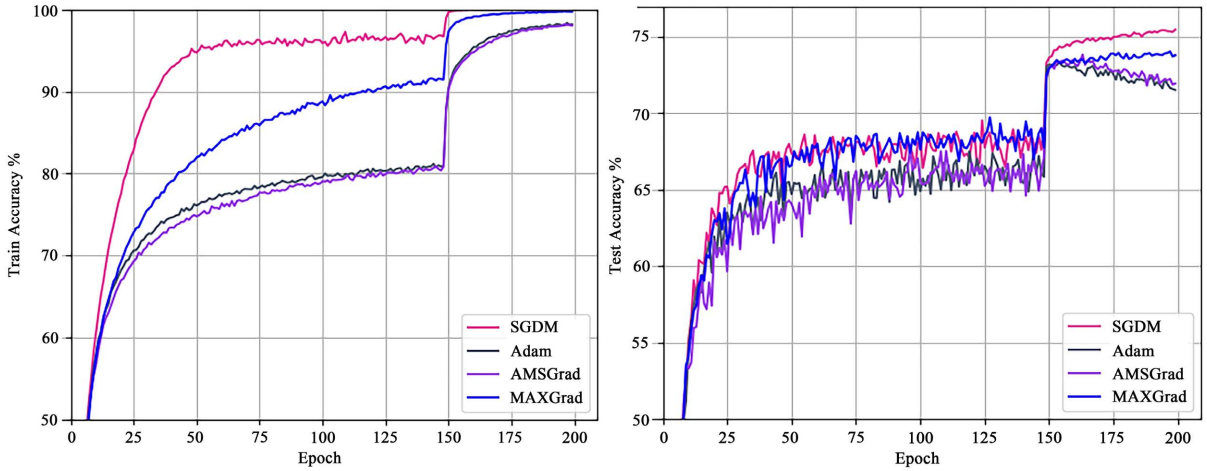
CIFAR-100 数据集与 CIFAR-10 数据集类似，区别在于它包含 100 个不同的图像类别。如图 3 所示，MAXGrad 算法相对于自适应算法的 Adam 与 AMSGrad 仍然具有更好的收敛结果。

3.4. PTB 数据集

PTB 数据集(Penn Treebank 数据集)在自然语言处理领域扮演着至关重要的角色，被广泛用于训练和评估语言模型以及其他自然语言处理任务(Natural Language Processing)。一种常用于评估语言模型性能的指标是“困惑度”，它衡量了模型对输入文本序列的概率估计质量。

本文对比 SGDM、Adam、AMSGrad 三个优化算法在 PTB 数据集上的性能进行了详尽的研究和量化

分析。如图 4 所示，对比三个优化算法的性能，MAXGrad 算法表现出了最低的困惑度，将其从 Adam 的 89.65 降至 88.03，这意味着它在估计语言序列的概率方面表现出色。这一发现进一步支持了 MAXGrad 算法在 NLP 中的潜在应用价值。



(a) Accuracy of the CIFAR-100 training set (b) Accuracy of the CIFAR-100 test set

Figure 3. CIFAR-100 image classification using ResNet-34
图 3. 使用 ResNet-34 进行 CIFAR-100 图像分类

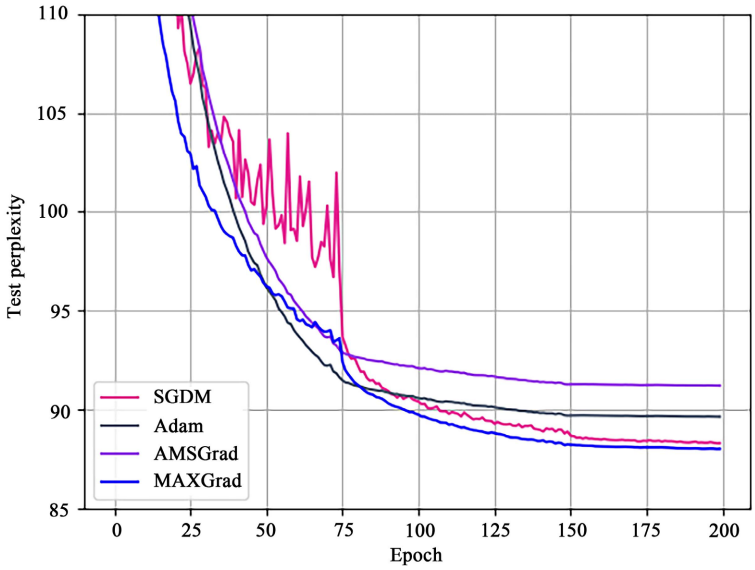


Figure 4. Linguistic modeling of PTB dataset using LSTM
图 4. 利用 LSTM 对 PTB 数据集进行语言建模

4. 结论

本文提出一种新的有效步长更新方式，即对历史中所有 v_t 取最大值作为本次更新的选择项。这种改进代码实现非常简单，并且相对于 Adam 也无须引入额外超参数就可获得更好的实验结果。在今后的工作中，可以考虑将这种有效步长的更新方式推广到 Adam W [13]、AdaBound [12]、RAdam [14]等优化器中。

基金项目

吉林省自然科学基金(No.YDZJ202201ZYTS519)，吉林省自然科学基金(No.YDZJ202201ZYTS585)，国家自然科学基金(No.11426045)。

参考文献

[1] 王斌, 罗莉, 刘金沧, 黄小川, 雷雳. 一种稀疏降噪自编码神经网络影像变化检测方法[J]. 测绘与空间地理信息, 2022, 45(1): 40-44.

[2] 曹中森. 基于卷积神经网络图像融合算法的电力巡检系统研究[J]. 安阳师范学院学报, 2022(5): 29-32.

[3] Liu, Z., Tian, Y. and Wang, Z. (2017) Improving Human Action Recognitionby Temporal Attention. 2017 *IEEE International Conference on Image Processing*, Beijing, 17-20 September 2017, 870-874. https://doi.org/10.1109/ICIP.2017.8296405

[4] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, December 2019, 5753-5763.

[5] Xu, Y., Verma, D., Sheridan, R.P., *et al*. (2020) Deep Dive into Machine Learning Models for Protein Engineering. *Journal of Chemical Information and Modeling*, **60**, 2773-2790. https://doi.org/10.1021/acs.jcim.0c00073

[6] Hu, L., Fu, C., Ren, Z., *et al*. (2023) SSELM-neg: Spherical Search-Based Extreme Learning Machine for Drug-Target Interaction Prediction. *BMC Bioinformatics*, **24**, 1471-2105. https://doi.org/10.1186/s12859-023-05153-y

[7] 史加荣, 王丹, 尚凡华. 随机梯度下降类算法研究进展[J]. 自动化学报, 2021, 47(9): 2103-2119.

[8] Duchi, J., Hazan, E. and Singer, Y. (2011) Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, **12**, 2121-2159.

[9] Kingma, D.P. and Ba, J. (2014) Adam: A method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980

[10] Reddi, S.J., Kale, S. and Kumar, S. (2018) On the Convergence of Adam and Beyond. https://openreview.net/forum?id=ryQu7f-RZ

[11] Zeng, K., Liu, J., Jiang, Z., *et al*. (2022) A Decreasing Scaling Transition Scheme from Adam to SGD. *Advanced Theory and Simulations*, **5**, 1-15. https://doi.org/10.1002/adts.202100599

[12] Luo, L., Xiong, Y., Liu, Y. and Sun, X. (2018) Adaptive Gradient Methods with Dynamic Bound of Learning Rate. https://openreview.net/forum?id=Bkg3g2R9FX

[13] Loshchilov, I. and Hutter, F. (2018) Decoupled Weight Decay Regularization. https://openreview.net/forum?id=Bkg6RiCqY7

[14] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J. (2019) On the Variance of the Adaptive Learning Rate and Beyond. https://openreview.net/forum?id=rkgz2aEKDr