南京航空航天大學
Nanjing University of Aeronautics and Astronautics

LLM-AutoDA: Large Language Model-Driven Automatic Data Augmentation for Long-tailed Problems

Pengkun Wang[1,2]*, Zhe Zhao[1,3]* Haibin Wen[5],
Fanfu Wang[6], Binwu Wang[1], Qingfu Zhang[3]†, Yang Wang[1,2]†

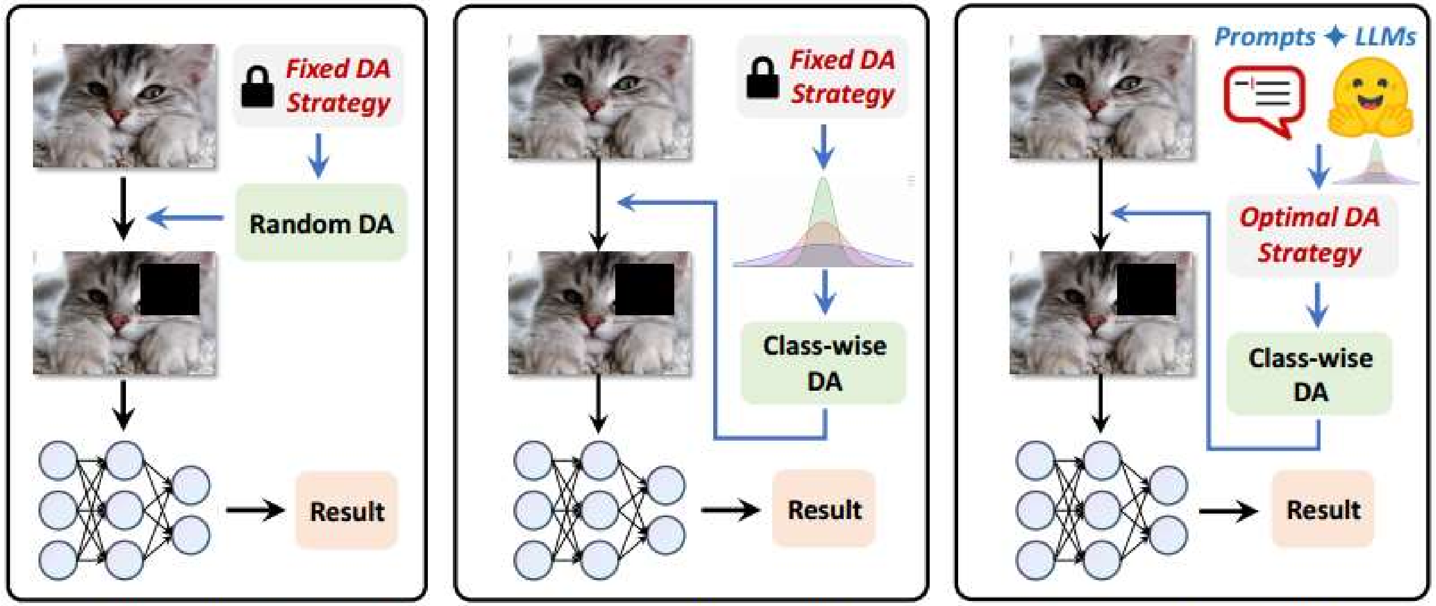
[1]University of Science and Technology of China (USTC), Hefei, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3]City University of Hong Kong, Hong Kong, China
[5]MorongAI, Suzhou, China
[6]Lanzhou University, Lanzhou, China
{pengkun@ustc.edu.cn, zz4543@mail.ustc.edu.cn, haibin65535@gmail.com, wangff21@lzu.edu.cn, wbw2024@ustc.edu.cn, qingfu.zhang@cityu.edu.hk, angyan@ustc.edu.cn}

NIPS 2024

Introduction

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Different long-tailed data augmentation paradigms



(a) Traditional Long-tailed DA (e.g., AutoAugment)
(b) Strategy Fixed Long-tailed DA (e.g., CUDA, DODA)
(c) LLM-driven Long-tailed DA (e.g., LLM-AutoDA)

(i) based on manually designed human knowledge and experience
(ii) the search space of these strategies is often limited
(iii) lack flexibility, can not adapt to changes in the data distribution during the training process

Introduction

LLM × LTL: Can LLMs Provide DA Strategies for Long-tailed Learning?

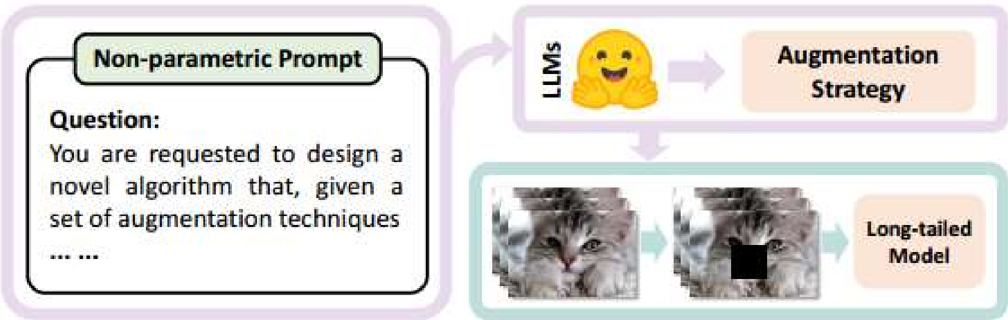


Figure 2: Strategy generation paradigm of SimpleLLM.

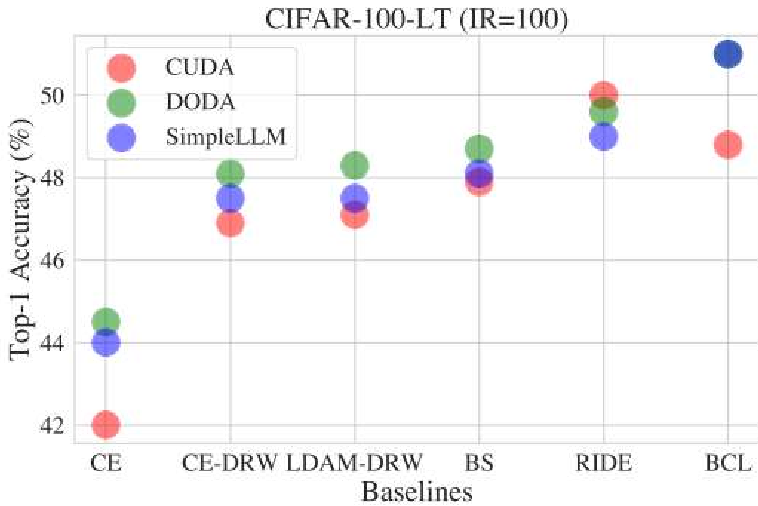


Figure 4: Average accuracy (%) on CIFAR-100-LT dataset (Imbalance ratio=100) with CUDA and DODA. SimpleLLM is comparable to the effectiveness of CUDA and DODA when combined with long-tailed learning baselines.

Method

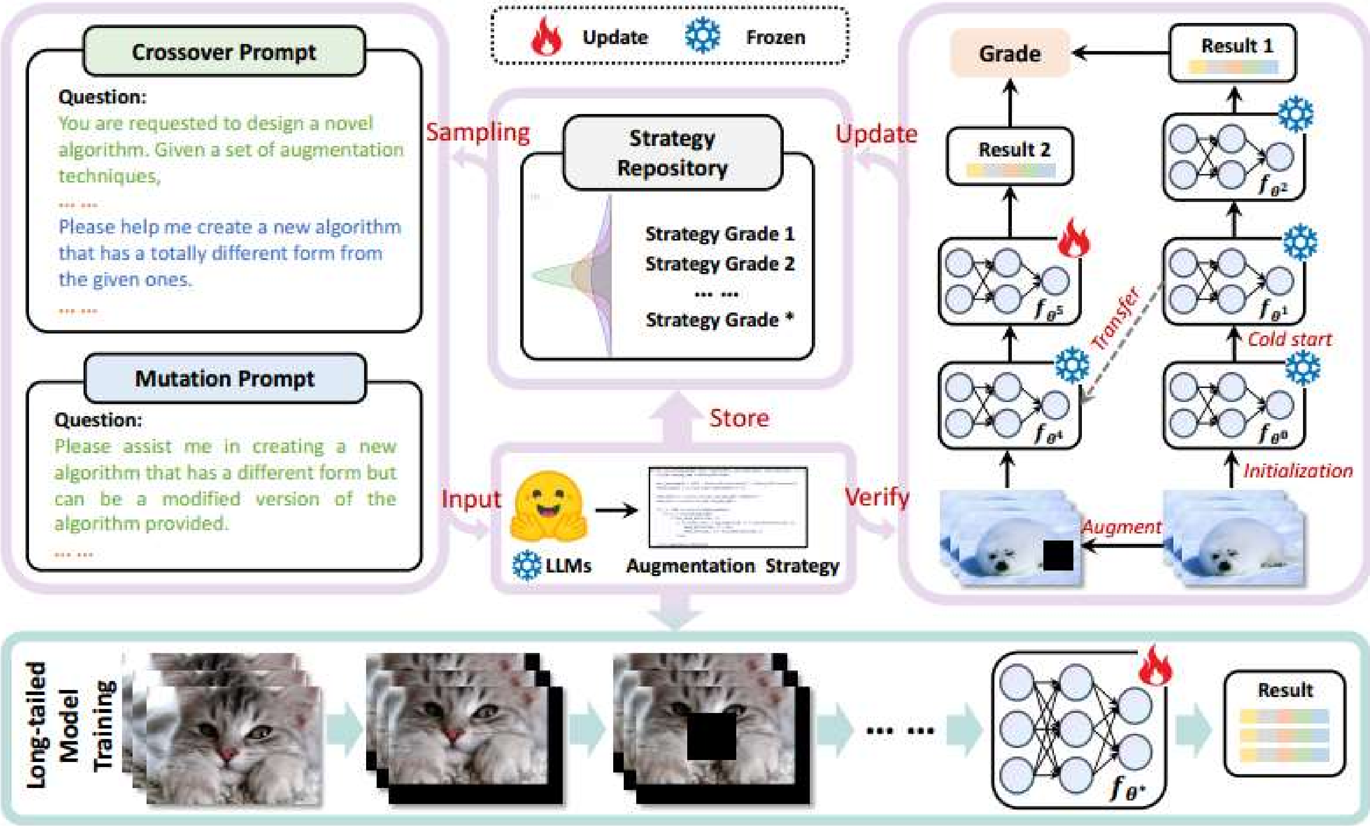


Figure 3: Overview of LLM-AutoDA. LLM-AutoDA leverages large-scale pretrained models to automatically search for the optimal augmentation strategies suitable for long-tailed data distributions.

Method

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Algorithm 1: LLM-AutoDA: Automatic Data Augmentation with Language Models

Input:
- Long-tailed datasets $\mathcal{D}_{train}, \mathcal{D}_{val}$
- Pretrained language model $\mathcal{L}$
- Initial data augmentation policies $\mathcal{K}$
- Long-tailed learning model f_θ with parameters θ

Output:
- Optimal data augmentation algorithm $\mathcal{A}^*$
- Final model f_θ

Define task description prompt $\mathcal{P}_{task}$
Define exploration operator prompt template $\mathcal{P}_E$
Define mutation operator prompt template $\mathcal{P}_M$
Initialize algorithm population $\{\mathcal{A}_i^{(0)}\}_{i=1}^{N} = \textsc{Initialize}(\mathcal{L}, \mathcal{P}_{task}, \mathcal{K})$
for each generation t **do**
  $\{\mathcal{A}_i^{(t)}\}_{i=1}^{N_p} = \textsc{Select}(\{\mathcal{A}_i^{(t-1)}\}_{i=1}^{N})$ {Select parents}
  $\{\mathcal{A}_j^{(t)}\}_{j=1}^{N_e} = \textsc{Explore}(\mathcal{L}, \mathcal{P}_E, \{\mathcal{A}_i^{(t)}\}_{i=1}^{N_p}, \mathcal{K})$ {Explore}
  $\{\hat{\mathcal{A}}_i^{(t)}\}_{i=1}^{N_m} = \textsc{Mutate}(\mathcal{L}, \mathcal{P}_M, \{\mathcal{A}_i^{(t)}\}_{i=1}^{N_m}, \mathcal{K})$ {Mutate}
  $\{\mathcal{A}_i^{(t)}\}_{i=1}^{N} = \{\mathcal{A}_i^{(t)}\}_{i=1}^{N_p} \cup \{\mathcal{A}_j^{(t)}\}_{j=1}^{N_e} \cup \{\hat{\mathcal{A}}_i^{(t)}\}_{i=1}^{N_m}$
  for each $\mathcal{A}_i^{(t)}$ **do**
    $fitness_i^{(t)} = \boxed{\textsc{Evaluate}(\mathcal{A}_i^{(t)}, f_\theta, \mathcal{D}_{train}, \mathcal{D}_{val}, e_{ckp}, N_{ft})}$
  end for
end for
$\mathcal{A}^* =_{\mathcal{A}_i} fitness_i$ {Select algorithm with highest fitness}
return $\mathcal{A}^*, f_\theta$
=0

function $\textsc{Evaluate}(\mathcal{A}, f_\theta, \mathcal{D}_{train}, \mathcal{D}_{val}, e_{ckp}, N_{ft})$:
  $f'_\theta = \textsc{Finetune}(\mathcal{A}, f_\theta, \mathcal{D}_{train}, e_{ckp}, N_{ft})$
  return $\textsc{Accuracy}(f'_\theta, \mathcal{D}_{val})$ =0

function $\textsc{Finetune}(\mathcal{A}, f_\theta, \mathcal{D}_{train}, e_{ckp}, N_{ft})$:
  Load checkpoint of f_θ at epoch e_{ckp}
for $e = e_{ckp}$ to $e_{ckp} + N_{ft}$ **do**
  for each class c **do**
    $\mathcal{A}_c^{(e)}, \mathcal{E}_c^{(e)} = \mathcal{A}(\mathbf{W}_c^{(e-1)}, \mathbf{H}_c^{(e-1)}, acc_c^{(e-1)}, \mathcal{A}_c^{(e-1)}, \mathcal{E}_c^{(e-1)}, e)$
  end for
  Train f_θ for one epoch on $\mathcal{D}_{train}$ using $\{\mathcal{A}_c^{(e)}, \mathcal{E}_c^{(e)}\}$ for augmentation
end for
return f_θ =0

Method

Strategy Evaluation

Injecting candidate algorithms into the real training process and evaluating them on the validation set.

$$Fitness(f_{aug}) = Acc_{val}(Finetune(f_{aug}, \mathcal{D}_{train}, e_{ckp}, N_{ft}), \mathcal{D}_{val})$$

function EVALUATE($\mathcal{A}, f_\theta, \mathcal{D}_{train}, \mathcal{D}_{val}, e_{ckp}, N_{ft}$):
  $f'_\theta = $ FINETUNE($\mathcal{A}, f_\theta, \mathcal{D}_{train}, e_{ckp}, N_{ft}$)
  return ACCURACY($f'_\theta, \mathcal{D}_{val}$) =0

function FINETUNE($\mathcal{A}, f_\theta, \mathcal{D}_{train}, e_{ckp}, N_{ft}$):
  Load checkpoint of f_θ at epoch e_{ckp}
for $e = e_{ckp}$ to $e_{ckp} + N_{ft}$ **do**
  for each class c **do**
    $\mathcal{A}_c^{(e)}, \mathcal{E}_c^{(e)} = \mathcal{A}(\mathbf{W}_c^{(e-1)}, \mathbf{H}_c^{(e-1)}, acc_c^{(e-1)}, \mathcal{A}_c^{(e-1)}, \mathcal{E}_c^{(e-1)}, e)$
  end for
  Train f_θ for one epoch on $\mathcal{D}_{train}$ using $\{\mathcal{A}_c^{(e)}, \mathcal{E}_c^{(e)}\}$ for augmentation
end for
return f_θ =0

Method

LLM-based Search Operator

Initialization operator *I*: Based on the task description prompt P_{task} and the knowledge base of data augmentation K.

Crossover operator *E*: Building upon P_{task}, N_p parent algorithms $A_i^{(t)}i = 1^{N_p}$ from the current population are used as references, along with the incorporation of knowledge base K.

Mutation operator *M*: Based on P_{task}, N_m individuals $A_i^{(t)}{}_{i=1}^{N_m}$ are selected from the current population, and local improvement directions are provided.

The prompt P_E of the crossover operator E can be represented as follows:

$$P_E(P_{task}, \{A_i^{(t)}\}_{i=1}^{N_p}, \mathcal{K}, D_{func}) = P_{task} + P_{ref}(\{A_i^{(t)}\}_{i=1}^{N_p}) + P_{know}(\mathcal{K}) + P_{diff} + P_{format}(D_{func})$$

$$\{A_j^{(t)}\}_{j=1}^{N_e} = \mathcal{L}(P_E(P_{task}, \{A_i^{(t)}\}_{i=1}^{N_p}, \mathcal{K}, D_{func}))$$

Experiments

Table 1: Accuracy (%) on CIFAR-100-LT dataset (Imbalance ratio={50, 100}) with SOTA DA methods. **Blod** indicates the best performance while underline indicates the second best. (+) and (-) indicate the relative gain.

Method	IR = 50				IR = 100			
	Head	Medium	Tail	All	Head	Medium	Tail	All
CE [15]	63.9	36.2	15.2	43.8 (+0.0)	65.6	36.2	8.2	40.1 (+0.0)
CE + CUDA	68.3	38.4	13.7	46.2 (+2.4)	70.7	41.4	9.3	42.0 (+3.9)
CE + DODA	71.2	40.3	12.6	48.0 (+4.2)	74.8	43.8	10.0	44.5 (+6.4)
CE + SimpleLLM	71.4	39.9	13.1	48.0 (+4.2)	72.5	44.9	9.8	44.0 (+5.9)
CE + LLM-AutoDA	**72.3**	40.0	14.1	48.6 (+4.8)	**74.9**	45.3	9.6	45.0 (+6.9)
CE-DRW [5]	60.6	39.0	22.9	45.0 (+0.0)	63.4	41.2	15.7	41.4 (+0.0)
CE-DRW + CUDA	63.8	48.0	37.0	52.5 (+7.5)	63.5	48.9	25.3	46.9 (+5.5)
CE-DRW + DODA	63.4	47.4	38.9	52.5 (+7.5)	60.2	**51.9**	29.6	48.1 (+6.7)
CE-DRW + SimpleLLM	62.3	49.4	38.8	52.7 (+7.7)	62.1	49.6	27.9	47.5 (+6.1)
CE-DRW + LLM-AutoDA	63.1	48.4	39.3	52.8 (+7.8)	62.9	50.7	29.9	48.7 (+7.3)
LDAM-DRW [5]	63.0	41.2	25.1	47.2 (+0.0)	62.8	42.6	21.1	43.2 (+0.0)
LDAM-DRW + CUDA	66.2	46.2	26.4	50.8 (+3.6)	66.0	49.5	22.1	47.1 (+3.9)
LDAM-DRW + DODA	64.7	46.3	27.5	50.5 (+3.3)	65.4	50.8	25.5	48.3 (+5.1)
LDAM-DRW + SimpleLLM	65.1	45.2	27.3	50.1 (+2.9)	65.7	49.4	23.9	47.5 (+4.3)
LDAM-DRW + LLM-AutoDA	66.7	46.1	27.4	51.2 (+4.0)	66.7	50.1	26.3	48.8 (+5.6)
BS [33]	60.3	41.3	34.3	47.9 (+0.0)	59.6	42.3	23.7	42.8 (+0.0)
BS + CUDA	63.6	48.4	37.3	52.7 (+4.8)	62.5	49.1	29.4	47.9 (+5.1)
BS + DODA	62.2	**51.2**	41.5	54.0 (+6.1)	63.1	49.3	31.2	48.7 (+5.9)
BS + SimpleLLM	62.4	48.8	37.7	52.4 (+4.5)	62.4	48.8	30.6	48.1 (+5.3)
BS + LLM-AutoDA	63.3	50.5	40.2	53.9 (+6.0)	63.3	50.0	31.0	49.0 (+6.2)
RIDE [43]	65.7	47.7	31.8	52.2 (+0.0)	65.7	48.6	25.0	47.5 (+0.0)
RIDE + CUDA	67.8	47.0	33.4	53.1 (+0.9)	67.9	51.2	27.6	50.0 (+2.5)
RIDE + DODA	68.2	46.1	29.3	52.1 (-0.1)	68.7	50.9	25.7	49.6 (+2.1)
RIDE + SimpleLLM	67.3	46.8	30.8	52.3 (+0.1)	69.3	48.8	25.4	49.0 (+1.5)
RIDE + LLM-AutoDA	67.1	47.3	32.7	52.8 (+0.6)	69.1	50.2	28.1	50.2 (+2.7)
BCL [48]	61.6	43.1	34.3	49.1 (+0.0)	63.1	42.9	23.9	44.2 (+0.0)
BCL + CUDA	64.0	47.4	39.4	52.7 (+3.6)	64.7	49.7	29.1	48.8 (+4.6)
BCL + DODA	64.9	48.0	40.6	53.6 (+4.5)	66.0	50.7	**33.8**	**51.0** (+6.8)
BCL + SimpleLLM	65.0	49.2	39.9	54.0 (+4.9)	64.1	50.4	30.1	49.0 (+4.8)
BCL + LLM-AutoDA	64.9	49.2	**44.1**	**54.8** (+5.7)	66.6	50.6	33.1	**51.0** (+6.8)

Experiments

Table 2: Accuracy (%) on ImageNet-LT and iNaturalist 2018 datasets with SOTA DA methods. **Blod** indicates the best performance while underline indicates the second best. (+) and (-) indicate the relative gain.

Method	ImageNet-LT				iNaturalist 2018			
	Head	Medium	Tail	All	Head	Medium	Tail	All
CE [15]	64.0	33.8	5.8	41.6 (+0.0)	73.9	63.5	55.5	61.0 (+0.0)
CE + CUDA	67.1	47.1	13.4	47.2 (+5.6)	74.6	65.0	57.2	62.5 (+1.5)
CE + DODA	67.4	47.5	13.9	48.1 (+6.5)	74.9	66.0	58.4	63.6 (+2.6)
CE + LLM-AutoDA	68.2	47.1	14.3	50.4 (+8.8)	75.1	66.3	58.9	64.0 (+3.0)
CE-DRW [5]	61.7	47.3	28.8	50.1 (+0.0)	68.2	67.3	66.4	67.0 (+0.0)
CE-DRW + CUDA	61.7	48.4	30.5	51.1 (+1.0)	68.8	67.9	66.5	67.4 (+0.4)
CE-DRW + DODA	62.4	48.5	31.3	52.2 (+2.1)	69.0	68.2	67.8	68.2 (+1.2)
CE-DRW + LLM-AutoDA	62.8	48.3	31.7	51.6 (+1.5)	68.8	68.8	68.1	68.7 (+1.7)
LDAM-DRW [5]	60.4	46.9	30.7	49.8 (+0.0)	-	-	-	66.1 (+0.0)
LDAM-DRW + CUDA	63.2	48.2	31.2	51.5 (+1.7)	68.0	67.5	66.8	67.3 (+1.2)
LDAM-DRW + DODA	63.7	48.6	31.9	52.4 (+2.6)	68.6	68.1	67.9	68.7 (+2.6)
LDAM-DRW + LLM-AutoDA	63.3	49.4	32.4	52.5 (+2.7)	68.0	69.4	68.6	69.5 (+3.4)
BS [33]	60.9	48.8	32.1	51.0 (+0.0)	65.7	67.4	67.5	67.3 (+0.0)
BS + CUDA	61.8	49.1	31.8	51.5 (+0.5)	67.6	68.2	68.3	68.2 (+0.9)
BS + DODA	61.9	49.5	32.4	52.0 (+1.0)	68.1	68.9	69.5	69.4 (+2.1)
BS + LLM-AutoDA	62.5	50.0	32.8	52.5 (+1.5)	68.0	69.1	69.9	69.8 (+2.5)
RIDE [43]	64.9	50.4	34.4	53.6 (+0.0)	70.4	71.8	71.8	71.6 (+0.0)
RIDE + CUDA	66.0	51.7	34.7	54.7 (+1.1)	70.6	72.6	72.7	72.4 (+1.4)
RIDE + DODA	66.6	51.9	35.9	55.8 (+2.2)	70.9	72.4	73.9	73.7 (+2.8)
RIDE + LLM-AutoDA	67.1	52.3	37.3	56.5 (+2.9)	70.9	72.8	73.8	73.9 (+3.0)
BCL [48]	65.3	53.5	36.3	55.6 (+0.0)	69.4	72.4	71.8	71.8 (+0.0)
BCL + CUDA	66.8	53.9	36.6	56.3 (+0.7)	70.8	72.7	72.0	72.2 (+0.4)
BCL + DODA	66.9	54.1	37.4	56.9 (+1.3)	71.2	73.2	73.4	73.7 (+1.9)
BCL + LLM-AutoDA	67.2	55.1	38.3	57.5 (+1.9)	70.9	73.6	74.7	74.2 (+2.4)

Experiments

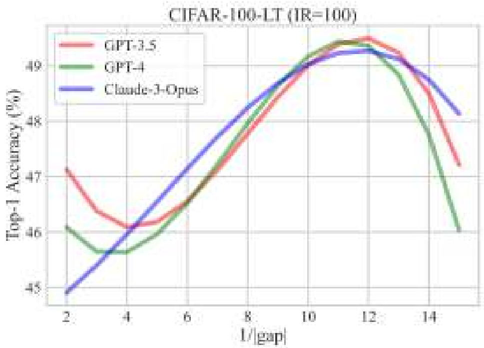


Figure 5: Impact of different LLMs on the performance of long-tailed learning models.

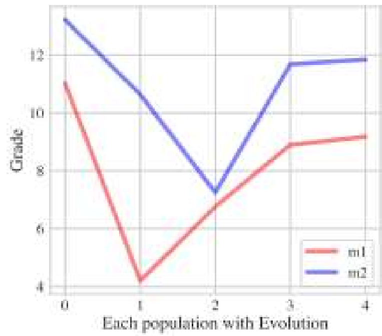


Figure 6: Impact of different population numbers in the mutation prompts.

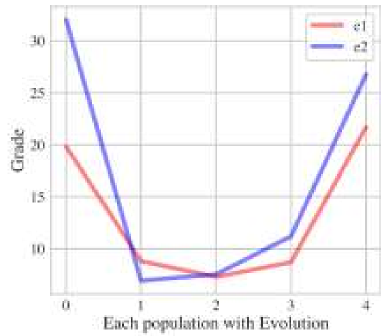


Figure 7: Impact of different population numbers in the crossover prompts.

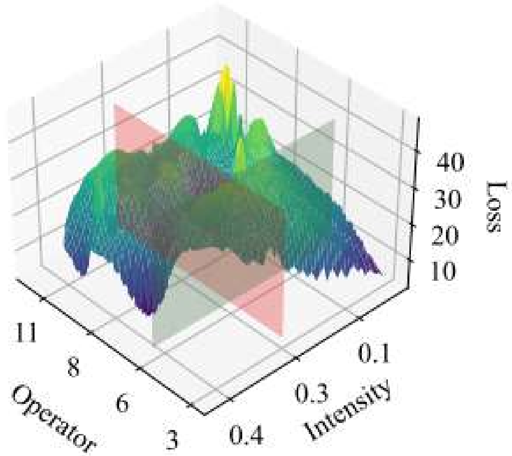


Figure 8: Visualization of the process of finding the optimal solution for different augmentation paradigms.

Thanks