CR-Drop：基于矫正正则化 Dropout 的多任务学习

肖洪湖[1] 黄成泉[1*]
XIAO Honghu HUANG Chengquan

摘 要 多任务学习（multi-task learning, MTL）在计算机视觉等应用场景中展现出卓越的性能，然而，在实践中由于模型在训练中所学习特征的随机性而导致推理能力低下。以往的多任务学习没有关心模型推理能力低下的问题，这增加了不必要的开销。为此，文章为 MTL 引入一种矫正网络，即 CR-Drop（corrected regularized dropout）。在这种方法中，STL（single-task learning）网络与 MTL 网络同时训练，以指导 MTL 训练过程中通用特征的学习。为了高效实现 CR-Drop，使用改进的线上知识蒸馏（IOKD）作为模型架构。同时，使用梯度操纵方法 FAMO（fast adaptive multitask optimization）平衡任务损失。在 NYUv2 数据集上的实验结果表明，所提方法比基准模型分别提高了 8.48% 和 8.64%，矫正网络 CR-Drop 可以学习到更具泛化性的特征，从而缩小训练与推理之间的差距，提高资源利用率。

关键词 推理能力；矫正网络；改进的线上知识蒸馏；多任务学习

doi：10.3969/j.issn.1672-9528.2026.01.034

0 引言

多任务学习（multi-task learning, MTL）是一种人工智能技术，涉及在同一系统中学习多个任务，以便在学习过程中共享信息，从而提高学习效率和任务性能。多任务学习作为一种颇具前景的人工智能技术，已经在许多领域得到应用，例如自然语言处理[1]，计算机视觉[2-8]和推荐系统[9]等领域。相比于单任务学习（single-task learning, STL），MTL 往往具有更好的泛化能力，减少资源消耗等优点。然而，深度 MTL 面临着一些挑战，如任务间的相互干扰、任务权重的确定和模型推理能力低下。解决任务间相互干扰和任务权重问题的方法有加权方案、梯度操纵和知识蒸馏。

在实践中，解决 MTL 问题最常用的方法是加权方案和知识蒸馏。在这种情况下，加权方案大多是动态调整用于训练损失的加权系数，例如，动态权重平均（dynamic weight average, DWA）[2]，以及 Liu 等[3]提出可与最先进的梯度操纵技术相媲美的快速自适应多任务优化算法（fast adaptive multitask optimization, FAMO）。相比之下，Li 等[7]提出知识蒸馏（knowledge distillation, KD），以及 Jacob 等[8]在 KD 基础上提出的线上知识蒸馏（online knowledge distillation, OKD），一定程度上提高了 MTL 的泛化性能，但 OKD 存在蒸馏缓慢的问题，所以这些方法仍然未解决 MTL 模型推理能力低的问题，即任务子集严重过拟合。这种优化失败背后的一个主要原因是深度 MTL 参数量庞大，而上述方法未在 MTL 训练过程中施加适当的正则化，致使 MTL 模型在训练过程中学习到的通用特征不足，导致泛化能力低下。

基于上述分析，本文提出一种方法来解决 MTL 训练过程中任务子集严重过拟合的问题，目的是在多任务模型训练过程中学习到更加通用的特征。为了实现这一点，本文参考了神经网络的正则化 Dropout（R-Drop）[10]。值得一提的是，在单任务设置中，R-Drop 通过最小化 Dropout 采样的两个子模型输出分布之间的双向 KL 散度，克服了 Dropout 引入的随机性而导致训练和推理之间出现的不一致问题。为此，本文将这一思想扩展到 MTL，提出矫正网络（corrected regularized dropout, CR-Drop），以矫正 MTL 训练过程中通用特征的学习。在这种情况下，CR-Drop 可以在每次迭代中对多任务模型学习的特征予以正确的矫正。这使得模型在优化的同时，也关注通用特征的学习。在 Cityscapes 和 NYUv2 数据集上的实验表明，矫正网络 CR-Drop 性能良好，可有效缓解多任务网络推理能力低下的问题。

1 方法

所提方法 CR-Drop 框架如图 1 所示，多个单任务模型与多任务模型同时训练，任务特定损失用于训练单任务模型（L^i_{STL}）和多任务模型（L^i_{MTL}）的任务头。在单任务模型与

1. 贵州民族大学数据科学与信息工程学院 贵州贵阳 550025
[基金项目] 贵州省科技计划项目（黔科合基础 -ZK〔2021〕一般 342）；贵州省教育厅自然科学研究项目（黔教技〔2022〕015）

多任务模型的中间特征之间使用改进的自适应特征蒸馏损失（improved adaptive feature distillation, IAFD），MTL 损失与 FAMO 通过矫正网络 CR-Drop 连接起来。在矫正网络 CR-Drop 中，R-Drop 强制 Dropout 多任务模型生成的不同子模型的输出分布彼此一致。随后，矫正算法（correction algorithm, CorrA）督促模型特征学习以进一步提高模型推理能力。最后，通过 FAMO 平衡处理后的损失函数。

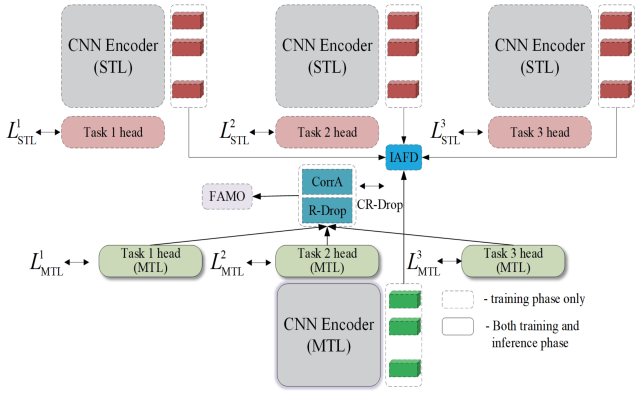


图 1 模型概览

1.1 改进的线上知识蒸馏 IOKD

线上任务权重（online task weight, OTW）假设单任务性能优于多任务性能。在整个 MTL 训练过程中，STL 都助益 MTL 模型的训练。在同时学习单任务与多任务时，使用多任务模型相对于单任务模型的性能作为任务权重。假设在第 t 次迭代中多任务模型第 i 个任务的损失为 $L^i_{\mathrm{MTL}}(t)$，而第 i 个任务的单任务模型损失为 $L^i_{\mathrm{STL}}(t)$。第 t 次迭代中任务 i 的任务权重 [8] 是由相应的多任务损失与单任务损失之比的温度缩放 Softmax 函数计算得出：

$$\lambda_i = N_t \exp(\frac{m^i_t}{T}) \Big/ \sum_{j=1}^{N_t} \exp(\frac{m^j_t}{T}),\ m^i_t = L^i_{\mathrm{MTL}}(t) / L^i_{\mathrm{STL}}(t) \tag{1}$$

式中：T 表示控制任务权重的温度项 [11]；N_t 表示任务数，且 $\sum_{i=1}^{N_t} \lambda_i = N_t$，在本文中 T 取 0.1。实际上，MTL 通过知识共享来提高模型泛化性能，应该具有比单任务模型更好的性能。随着模型训练迭代，MTL 性能会逐渐优于单任务性能。这时上述 OTW 也会反过来迫使 MTL 性能向 STL 性能对齐，从而限制最终的多任务性能。基于上述问题，文章修正了 OTW 公式：

$$\lambda^{'}_i = N_t \frac{\exp(\frac{m^i_t}{T})}{\sum_{j=1}^{N_t} \exp(\frac{m^j_t}{T})},\ m^i_t = \max(\frac{L^i_{\mathrm{MTL}}(t)}{L^i_{\mathrm{STL}}(t)}) \tag{2}$$

在任何一次迭代 t 时，m^i_t 通过 $\max(m^i_t, 1)$ 来选择，而不是一直取 $L^i_{\mathrm{MTL}}(t)/L^i_{\mathrm{STL}}(t)$ 值。有效避免了当 $L^i_{\mathrm{MTL}}(t) < L^i_{\mathrm{STL}}(t)$ 时出现 $m^i_t < 1$，单任务性能反过来限制多任务性能的情况。

自适应特征蒸馏（adaptive feature distillation, AFD）是一种对齐共享骨干模型特征的方法，L 表示共享网络编码器的层数。AFD 损失 [8] L_{AFD} 计算为：

$$L^i_{\mathrm{AFD}} = \sum_{l=1}^{L} \left\| A_\tau(f_{\mathrm{MTL}}(l)) - \sum_{i=1}^{N_t} w^l_i f^i_{\mathrm{STL}}(l) \right\|^2 \tag{3}$$

式中：$f_{\mathrm{MTL}}(l)$ 表示共享 MTL 骨干网络第 l 层的特征；$f^i_{\mathrm{STL}}(l)$ 表示第 i 个单任务网络第 l 层的特征；A_τ 表示任务特定适配器，将多任务特征编码器的输出映射到每个任务特定的编码器中；w^l_i 表示第 l 层第 i 个任务的可学习参数，决定每个任务的 STL 特征与 MTL 特征的对齐程度。

AFD 函数确保 MTL 网络的特征空间与 STL 网络的特征空间尽量保持一致，MTL 从 STL 特征中进行跨任务学习。OKD-MTL 设置两个优化器，一个针对适配器参数，另外一个是 MTL 模型参数和 w^l_i 参数。OKD-MTL 模型所采用的优化公式为 $\min \sum_{i=1}^{N_t} L^i_{\mathrm{MTL}}(t)\lambda_i + \lambda_\tau L^i_{\mathrm{AFD}}$，其中 λ_τ 是特定任务的权衡超参数。

OKD-MTL 优化公式存在不足，多任务损失 $L^i_{\mathrm{MTL}}(t)$ 和 AFD 损失 L_{AFD} 是通过线性加权进行反向传播，导致参数更新倾向于 MTL 模型参数，适配器参数 A_τ 和 w^l_i 优化缓慢。其次，AFD 中的超参数 λ_τ 难以确定，往往遵循前人工作的经验。例如对于 NYUv2 数据集上的 3 个任务 λ_τ 取 1、1 和 2，但损害了部分任务，限制了 MTL 最终的性能。基于 OKD-MTL 的不足，本文提出改进之后的 MTL 优化公式为：

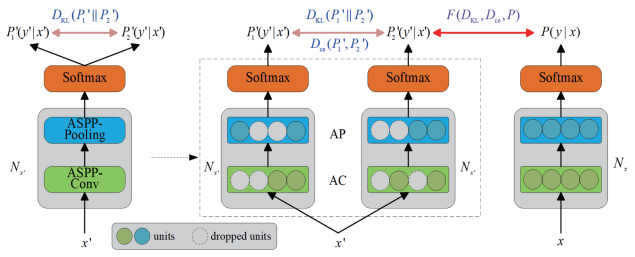
$$\min \sum_{i=1}^{N_t} L^i_{\mathrm{MTL}}(t)\lambda^{'}_i,\ \min \sum_{i=1}^{N_t} L^i_{\mathrm{AFD}}\lambda^{'}_i \tag{4}$$

式（2）将 MTL 损失与 AFD 损失分别优化，同样设置两个优化器。优化器 I 是多任务参数与 FAMO 优化参数，优化器 II 是适配器参数 A_τ 和参数 w^l_i。此外，AFD 的优化公式不再采用难以确定的超参数 λ_τ，而是与多任务优化公式一样采用修正之后的 OTW 系数 $\lambda^{'}_i$。

1.2 基于 R-Drop 的矫正网络 CR-Drop

在深度神经网络的训练过程中通常采用 Dropout 正则化方法来提高模型的推理能力，Dropout 通过丢弃网络的每一层随机部分神经元避免模型过拟合。

图 2 展示了 CR-Drop 详细信息，（a）R-Drop 与（b）CorrA 算法细节，本文仅在 STL 与 MTL 的任务头使用 Dropout。



（a） （b）

图 2 CR-Drop 详细信息

图 2（a）显示增强后的训练数据 x' 被输入该模型两次，分别获得不同的分布 P_1' 和 P_2'，随后计算得到中间变量 D_{ce} 和 KL 散度 D_{KL}；图 2（b）显示训练数据 x 被输入不使用 Dropout 模型一次获得分布 P；综合（a）和（b）得到 $F(D_{KL}, D_{ce}, P)$。

具体而言，给定迭代 epoch，增强后的训练数据 $x_i^{'}$ 被输入网络两次，分别获得分布 $P_1^w(y_i^{'}, x_i^{'})$ 和 $P_2^w(y_i^{'}, x_i^{'})$。由于 Dropout 在模型中随机丢弃部分神经元，因此两次前向传播实际上是基于两个不同的子模型。如图 2（a）所示，输出分布 $P_1^w(y_i^{'}, x_i^{'})$ 左路径与输出分布 $P_2^w(y_i^{'}, x_i^{'})$ 右路径每层的丢弃单元不同。本文引入 R-Drop，通过最小化同一样本中两个输出分布之间的双向 Kullback-Leibler（KL）散度对模型预测进行正则化 [10]，即 $L_{KL}^i=\frac{1}{2}(D_{KL}(P_1^w\|P_2^w)+D_{KL}(P_2^w\|P_1^w))$，同时两个前向传播的基本负对数似然学习目标 L_{NLL}^i 为 $-\log P_1^w-\log P_2^w$，R-Drop 的训练目标是最小化数据 $(x_i^{'}, y_i^{'})$ 的 L_{RD}^i，即 $L_{RD}^i=L_{NLL}^i+\alpha\cdot L_{KL}^i$，其中，$\alpha$ 是控制 L_{RD}^i 的系数权重。

增强后的训练数据 $(x_i^{'}, y_i^{'})$ 由已有训练数据 (x_i, y_i) 随机缩放裁剪变换得到，在训练阶段 (x_i, y_i) 充当验证数据以构造 CR-Drop。本文在训练阶段多加载一个验证数据集，以及禁用梯度计算的前向传播，所以在训练时间上稍有增加，而推理时间没有变化。梯度计算是求导运算，且假设测试损失大于训练损失。CR-Drop 在训练损失上乘以验证损失与训练损失的比值 L_{val}^i/L_{tra}^i，用验证损失来惩罚模型。该思想迫使模型在训练过程中学习更泛化的特征，使得 L_{val}^i/L_{tra}^i 比值尽可能等于 1。但是存在一个问题，直接比值导致模型振荡，得到相反的效果。为此，提出改进的比值，即矫正算法 CorrA，定义为 $1+\beta\cdot(\max(L_{val}^i/L_{tra}^i, 1)-1)$，其中，$\max(L_{val}^i/L_{tra}^i, 1)$ 避免偶尔验证损失小于训练损失时，对模型的反向惩罚，β 是控制矫正力度的系数权重，合理设置 β 值，能有效避免模型反向记忆训练数据，从而稳定提高模型推理能力。最后，MTL 优化公式 $F(D_{KL}, D_{ce}, P)$，即矫正网络 CR-Drop 定义为：

$$L_{mtl}^i=L_{RD}^i(1+\beta.(\max(L_{val}^i/L_{NLL}^i,1)-1))\lambda_i^{'} \quad (5)$$

式中：L_{RD}^i 表示通过 R-Drop 计算得到的损失；L_{NLL}^i 表示通过两次子模型得到的平均损失值；$\lambda_i^{'}$ 表示修正之后的 OTW。

2 实验与结果分析

为了进行公平比较，本文将所提出的方法与以下基线模型进行比较。首先，考虑单任务基线，分别为每个任务训练一个特定任务的主干与特定任务的头部。其次，使用多任务基线，所有任务共享相同的预训练骨干网络，且具有单独的任务特定头部。本文比较了几种经典的多任务学习方法，包括加权方案（DWA[2]、FAMO[3]、RLW[4] 和 UW[5]）、梯度操纵 GradNorm[6]，以及基于知识蒸馏的 OKD-MTL 技术 [8]。本文还比较了所提出的方法在 MTAN 基线 [2] 上的性能。与文献 [4] 中的实验类似，将这些方法与 MTAN（DeepLabV3-MTAN）和基线 CNN（DeepLabV3[12]）进行比较。

2.1 数据集和任务

在 NYUv2[13] 公共数据集上对所提出的方法进行了评估。NYUv2 数据集是一个常用的室内场景理解数据集，特别在计算机视觉领域的研究中。该数据集用于评估 3 个学习任务的性能：语义分割（13 个标签）、深度估计和表面法线估计。

2.2 训练细节和评估指标

本文使用 MTAN 代码库中提供的 NYUv2 的预处理数据集，同时使用在 ImageNet 数据上微调的 DeepLabV3 和 DeepLabV3-MTAN 模型作为骨干网。本文使用 AdamW 优化器和 CosineAnnealingLR 学习率调度器训练所有模型，初始化学习率设置为 10^{-4}，每个数据集训练模型 200 个 epoch。定量评估使用文献 [2-8] 中常用的评估指标，具体而言，语义分割采用平均交并联（mIoU）和像素精度（pAcc），深度估计用绝对误差（abs）和相对误差（rel），同时通过预测矢量的平均角偏差（Mean）和中值角偏差（Median）来评估表面法线估计。当前方法相对于基线的性能 [4] 计算为：

$$\Delta=\frac{1}{N_t}\sum_{n=1}^{N_t}\frac{1}{N_n}\sum_{m=1}^{N_n}\frac{(-1)p_{n,m}(M_{n,m}-M_{n,m}^B)}{M_{n,m}^B}\times 100\% \quad (6)$$

式中：N_t 是任务数量；N_n 是第 n 个任务的指标数量；$M_{n,m}$ 和 $M_{n,m}^B$ 分别是当前方法和基线的度量指标，下标表示第 n 个任务和第 m 个指标，上标 B 表示基线模型。符号通过 $p_{n,m}$ 控制，如果数值越大，表示性能越好，则设置为 1，反之则设置为 -1。Δ 值越大表示较基线改善越大。

2.3 数据集 NYUv2 的结果

为了验证 CR-Drop 方法的实际性能，表 1 呈现了使用在 ImageNet 数据集上预训练的网络 DeepLabV3 和 DeepLabV3-MTAN 在数据集 NYUv2 上对不同方法（STL、DWA、FAMO、UW、RLW、GradNorm、OKD、CR-Drop）进行实验所获得的结果。由表 1 可知，在 DeepLabV3 和 DeepLabV3-MTAN 网络骨干上，本文方法 CR-Drop 相对 MTL 基线方法表现最优，分别提高了 8.48% 和 8.64%；其次是 OKD 方法，分别提高了 2.59% 和 2.83%。在 DeepLabV3 网络骨干上，加权方案（除了 RLW 外）、梯度操纵和 STL 表现一般，而在 DeepLabV3-MTAN 网络骨干上，加权方案 FAMO 和 STL 表现出较好的性能改进。

表 1 各方法使用不同主干网络在数据集 NYUv2 上相对基线

的改进

Method	DeepLabV3 Δ↑	DeepLabV3-MTAN Δ↑
STL	0.59	1.61
基线 （MTL）	0	0
DWA	0.56	0.29
FAMO	0.83	2.38
UW	0.17	0.87
RLW	1.03	0.67
GradNorm	0.74	0.37
OKD	2.59	2.83
CR-Drop	8.48	8.64

3 结语

本文提出一种矫正网络 CR-Drop，旨在缓解 MTL 模型推理能力低下的问题。CR-Drop 的核心是 R-Drop 和 CorrA，它们有助于在 MTL 训练过程中有效地避免模型过拟合。这些方法使训练能够在MTL环境中平衡任务优化与推理。此外，针对线上知识蒸馏 OKD 存在的蒸馏缓慢问题，本文已经展示了 CR-Drop 在基于改进的线上知识蒸馏 IOKD 的 MTL 架构中的应用。同时 CR-Drop 在场景理解任务的联合学习中得到了验证，单任务和多任务模型同步训练，并在多任务模型中应用 FAMO 算法，使得各个任务以相同的速率学习。实验结果表明，在两个基准数据集上，CR-Drop 方法比基准方法有所改进，超过了单任务模型的准确性。

参考文献：

[1]BONFIGLI A, BACCO L, PECCHIA L, et al. Efficient multi-task learning with instance selection for biomedical NLP [J]. Computers in biology and medicine, 2025, 190:110050.

[2]LIU S K, JOHNS E, DAVISON A J. End-to-end multi-task learning with attention [C]//IEEE/CVF conference on computer vision and pattern recognition. Piscataway:IEEE,2019: 1871-1880.

[3]LIU B, FENG Y H, STONE P, et al. FAMO: fast adaptive multitask optimization [EB/OL].(2023-10-30)[2025-12-30]. https://doi.org/10.48550/arXiv.2306.03792.

[4]LIN B J, YE F Y, ZHANG Y. A closer look at loss weighting in multi-task learning [EB/OL]. (2022-07-27)[2025-12-24]. https://doi.org/10.48550/arXiv.2111.10603.

[5] CIPOLLA R,GAL Y,KENDALL A . Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]/2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2018: 7482-7491.

[6]CHEN Z, BADRINARAYANAN V, LEE C Y, et al. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks [EB/OL].(2018-06-12)[2025-12-22]. https://doi.org/10.48550/arXiv.1711.02257.

[7]LI W H, BILEN H. Knowledge distillation for multi-task learning[EB/OL].(2020-09-24)[2025-12-30].https://doi.org/10.48550/arXiv.2007.06889.

[8]JACOB G M, AGARWAL V, STENGER B. Online knowledge distillation for multi-task learning [C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway:IEEE, 2023: 2359-2368.

[9]DENG Y, ZHANG W X, XU W W, et al. A unified multi-task learning framework for multi-goal conversational recommender systems[J]. ACM transactions on information systems, 2023, 41(3): 1-25.

[10]LIANG X B, WU L J, LI J T, et al. R-Drop: regularized dropout for neural networks [EB/OL].(2021-10-29)[2025-12-03]. https://doi.org/10.48550/arXiv.2106.14448.

[11]HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [EB/OL]. (2015-03-09)[2025-05-24]. https://doi.org/10.48550/arXiv.1503.02531.

[12]CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL].(2017-12-05)[2025-05-24]. https://doi.org/10.48550/arXiv.1706.05587.

[13]SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images [C]//ECCV'12 Proceedings of the 12th European conference on Computer Vision.Berlin:Springer,2012: 746-760.

【作者简介】

肖洪湖（1998—），男，贵州遵义人，硕士研究生，研究方向：统计建模与模式识别，email: 2143821719@qq.com。

黄成泉（1976—），通信作者（email:hcq@gzmu.edu.cn），男，贵州黔东南人，博士，教授，研究方向：深度学习与图像处理。

（收稿日期：2025-09-26 修回日期：2026-01-13）