稀疏子空间变分推理训练贝叶斯神经网络

摘要

贝叶斯神经网络提供了不确定性量化的能力，但其高昂的训练和推理成本阻碍了其广泛应用。本文提出了一种新的框架——稀疏子空间变分推理 (Sparse Subspace Variational Inference,SSVI)，这是第一个在训练和推理阶段始终保持高稀疏性模型的框架。通过从随机初始化的低维稀疏子空间开始，SSVI 交替优化稀疏子空间的基选择及其相关参数。为了克服基选择的不可微问题，本文提出了一种基于权重分布统计的移除和添加策略。实验表明，SSVI 在构建稀疏 BNNs 方面达到了新的基准，其模型尺寸压缩达 10-20 倍，训练时的 FLOPs 减少高达 20 倍，同时性能仅下降不到 3%。此外，SSVI 对超参数的鲁棒性更强，并在准确性和不确定性量化上有时甚至超过了基于变分推理的密集 BNNs。

关键词：贝叶斯神经网络；稀疏子空间；变分推理

1 引言

贝叶斯神经网络因其不确定性量化能力在医疗诊断和自动驾驶等关键领域具有重要应用，但高昂的训练和推理成本限制了其广泛采用。尽管现有方法通过逐步引入稀疏性或后训练剪枝降低推理成本，训练阶段的高开销问题仍未解决。受传统神经网络稀疏训练方法启发，本研究提出稀疏子空间变分推理，通过在训练起始阶段嵌入稀疏性，实现了训练和推理阶段的双重高效。SSVI 在显著减少计算成本和模型大小的同时，保持甚至提升了模型性能和不确定性量化能力，对超参数更具鲁棒性，减少了调参需求 [4]。该方法为稀疏贝叶斯神经网络研究提供了新的范式，推动了其在资源有限硬件上的实际应用。

2 相关工作

本文的工作围绕稀疏贝叶斯神经网络的高效训练与推理，同时解决训练和推理阶段的高计算开销问题，提出了一种创新性的框架——稀疏子空间变分推理。以下是本文的相关工作：

2.1 CIFAR-10 数据集和 CIFAR-10-C 数据集

CIFAR-10 是一个广泛用于图像分类基准测试的经典数据集，共包含 60,000 张 32×32 的彩色图片，分为 10 个类别（如飞机、汽车、猫等），其中 50,000 张用于训练，10,000 张用于测试。该数据集来源于真实世界场景，具有多样化的背景、光照和姿态变化，适合评估机器学习模型的分类性能。由于其适中的规模和简单的结构，CIFAR-10 成为深度学习研究中测试新算法的重要工具，尽管其分辨率较低且类别有限，但依然是经典的研究基准数据集之一。

CIFAR-10-C 是一个由 CIFAR-10 数据集的多个变种构成的鲁棒性评估数据集。数据集由 15 种算法生成的噪声组成，这些损坏来自噪声、模糊、天气和数字类别。每一种噪声都有五个等级的严重程度，导致 75 种不同的噪声。选择数据集 CIFAR-10-C 和 CIFAR-10 进行实验，评估模型在噪声条件下的鲁棒性

2.2 贝叶斯神经网络

贝叶斯神经网络是一种将贝叶斯推理引入神经网络的方法，其核心思想是将网络权重建模为概率分布，而不是固定值。通过对权重进行概率建模，BNNs 能够捕捉模型的不确定性，使其在应对未知数据或分布外数据时更加稳定。训练过程中，BNNs 试图通过贝叶斯定理，根据数据对权重分布进行更新，从而得到后验分布。然而，由于后验分布往往难以解析，BNNs 通常采用近似推断方法（如变分推理或蒙特卡洛采样）来估计权重分布 [2]。这些方法通过反复采样或优化，调整权重分布以最大化模型的拟合能力。虽然这种方法能够显著提升模型的泛化性能，并提供不确定性量化能力，但其代价是高昂的计算成本，尤其是在大规模网络中，必须对大量参数进行采样和推断，导致训练和推理效率较低。

2.3 稀疏子空间变分推理的贝叶斯神经网络

该方法的基本思想是通过引入一个稀疏子空间，将模型的参数限制在一个较低维的子空间中，从而减少需要优化的参数数量，显著降低计算开销。具体来说，这一方法通过动态选择最重要的权重，稀疏不重要的权重，使得模型的稀疏性在整个训练过程中始终得以保持。此外，通过交替优化的方式，SSVI 同时优化稀疏子空间的选择和权重分布的参数，以确保模型在稀疏状态下仍能保持高性能。与传统 BNNs 不同，SSVI 从训练一开始就存在稀疏性，避免了遍历完整参数空间的开销，并通过精确控制稀疏度使模型更加高效。

3 本文方法

3.1 本文方法概述

本文提出了一种基于稀疏子空间变分推理（SSVI）的贝叶斯神经网络训练框架，旨在通过动态稀疏化方法实现训练和推理的高效性。SSVI 在训练初期就将模型参数限制在稀疏子空间中，并通过交替优化子空间选择和变分参数来保持稀疏性。

Algorithm 1 Sparse Subspace Variational Inference (SSVI)

Require: A BNN $\theta \in \mathbb{R}^d$ with prior $p(\theta)$, variational distribution $q_\phi(\theta)$, target sparsity s/d, replacement rate $\{r_t\}$, inner update steps M, total steps T.

Random initialize (ϕ^0, γ^0) from the feasible set of (3), and set $\gamma^{-1} = \gamma^0$.
for $t = 0, \ldots, T$ **do**
  # Update ϕ.
  $\phi^{t,0} = \texttt{Initialize}(t, \phi^t, \gamma^t, \gamma^{t-1})$ according to Section 3.2.
  for $m = 0, \ldots, M-1$ **do**
    Obtain $\phi^{t,m+1}$ using the gradient of (3).
  end for
  $\phi^{t+1} = \phi^{t,M}$.
  # Update γ.
  $\gamma^t_{\text{remove}} = \texttt{Removal}(\gamma^t, \phi^{t+1}, r_t)$ according to Section 3.3.1.
  $\gamma^{t+1} = \texttt{Addition}(\gamma^t_{\text{remove}}, \phi^{t+1}, r_t)$ according to Section 3.3.2.
end for

图 1. SSVI 算法

3.2 初始化稀疏子空间的权重分布

在训练开始时，SSVI 初始化一个低维稀疏子空间，子空间的维度通过稀疏度参数控制。具体地，权重的后验分布被限制在这一子空间中，并为子空间中的每个权重分配初始值，例如均值和方差。这样的初始化策略能够减少优化的搜索空间，降低计算成本，同时为后续动态调整提供基础。稀疏子空间的初始权重通过随机选择或简单的统计准则确定，以便快速进入稀疏化状态。

3.3 动态调整稀疏子空间的结构

在训练过程中，SSVI 通过动态调整子空间的基向量来适应模型需求。子空间调整分为两部分：移除操作：基于设计的权重重要性准则（例如均值、信噪比），剔除对模型输出影响较小的权重，从而进一步压缩模型。添加操作：通过分析权重的梯度信息，选择重要的权重重新加入子空间。这种动态调整机制使得子空间的结构能够灵活适应不同阶段的训练需求，同时确保稀疏性和模型性能的平衡。

3.4 在稀疏子空间内进行变分推理优化

在稀疏子空间中，SSVI 对权重的后验分布进行优化。通过标准的变分推理方法，模型同时优化分布的均值和方差。为提高效率，本文采用了局部重参数化技巧 LRT（Local Reparameterization Trick），将采样过程简化为少量的矩阵运算，从而显著减少计算开销。在这个过程中，稀疏子空间内的权重分布不断更新，以提高模型对数据的拟合能力。

3.5 使用权重分布统计准则移除和添加权重

SSVI 的创新点在于引入了一套基于权重分布统计的新准则，用于动态选择子空间基。移除操作根据权重的均值、信噪比（SNR）或期望等信息，剔除对模型贡献较小的权重；添加操作则根据梯度的绝对值大小重新引入重要权重。这些准则结合模型当前状态动态调整子空间结构，使得稀疏化过程更具理论依据和实践效果。

3.6 损失函数定义

本文在稀疏子空间中采用变分推理优化损失函数，该损失包括两部分：KL 散度用于约束变分后验分布与先验分布的接近性，似然项用于最大化模型对观测数据的拟合能力。通过优化这两部分损失，模型能够在稀疏状态下实现高性能，同时显著降低训练和推理的计算成本。

4 复现细节

4.1 与已有开源代码对比

CIFAR-10-C 是对经典 CIFAR-10 数据集的扩展版本，通过在测试集上引入 15 种常见图像扰动（如噪声、模糊、天气效应和数字变形）及其 5 个强度级别，共计 75 种扰动类型，用于评估模型在分布外数据场景下的鲁棒性和泛化能力。与标准的 CIFAR-10 数据集相比，CIFAR-10-C 更具挑战性，专注于测试模型在噪声和腐蚀环境下的性能，模拟了现实世界中的复杂场景，有利于研究深度学习模型的抗干扰能力和不确定性量化的能力 [1]。

原文的实验主要基于 CIFAR-10 数据集，而本工作针对 CIFAR-10-C 数据集进行了扩展。通过将实验从 CIFAR-10 转向更具挑战性的 CIFAR-10-C 数据集，本工作探讨了模型在更复杂环境下的泛化性能和鲁棒性。这一扩展不仅增加了实验的实际应用价值，还验证了模型应对分布变化的能力。

动态收集目标标签并结合 Softmax 概率分布的设计，使函数在处理复杂数据集时具备更强的灵活性和通用性，适用于目标标签不可直接访问的场景。引入校准误差的计算与报告，扩展评估维度，显著提升对模型置信度与预测准确性匹配程度的分析能力 [3]。这些改进强化了测试函数的实用性和适用性，为评估模型在噪声和分布外数据上的表现提供了支持。

4.2 实验环境搭建

Python 3.12,Pytorch 2.5.0,Nvidia 4080 super

4.3 创新点

将原实验中的 CIFAR-10 数据集扩展到更具挑战性的 CIFAR-10-C 数据集，通过处理多种噪声与腐蚀条件，验证了模型在分布外场景中的表现。

针对复杂数据集中特殊的目标标签获取问题，设计了动态目标标签收集机制，逐步收集测试数据中的目标标签并合并，为复杂环境下的测试提供了更高的灵活性与通用性。

在评估指标中引入校准误差，扩展了测试函数的评估维度。ECE 的计算能够衡量模型预测置信度与实际正确率的匹配程度，使得模型在噪声与分布外数据上的表现分析更加全面。

优化测试函数，加入 Softmax 概率分布的存储和即时打印功能，不仅为进一步的分布偏移检测提供支持，还通过实时输出损失、准确率和校准误差等多维指标，提高了调试与验证效率。

5 实验结果分析

本实验比较稀疏贝叶斯神经网络 BNNs 在 CIFAR-10 和 CIFAR-10-C 数据集上的性能表现。

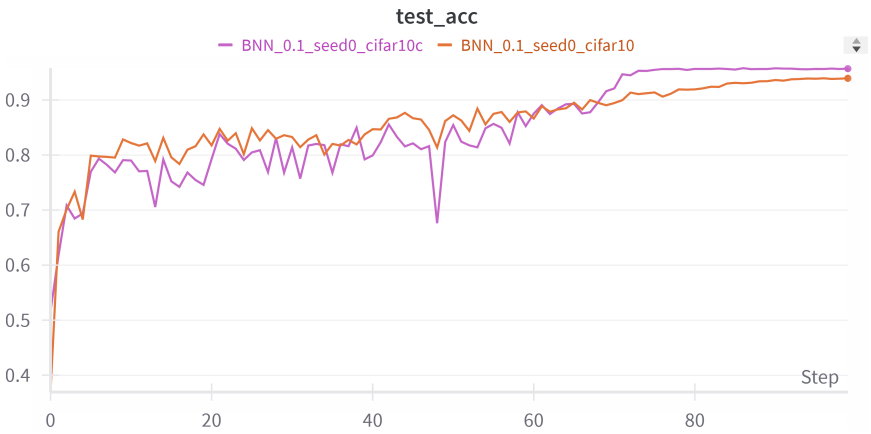

图 2. CIFAR-10 和 CIFAR-10-C 的测试准确率比较

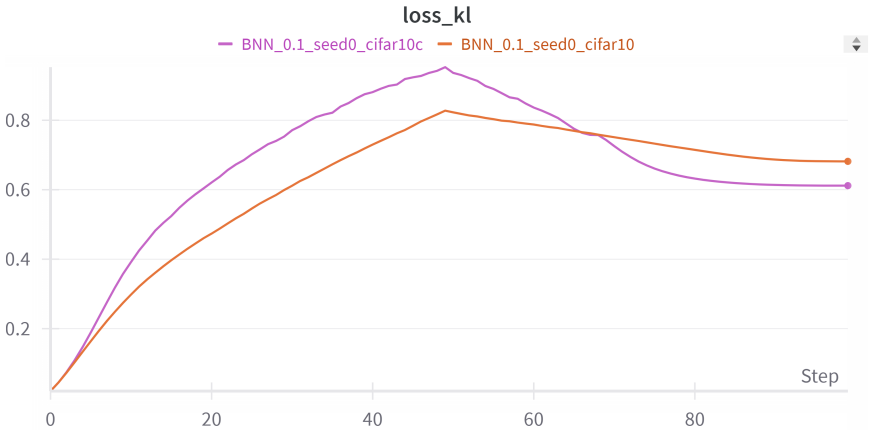

图 3. CIFAR-10 和 CIFAR-10-C 的 KL 损失比较

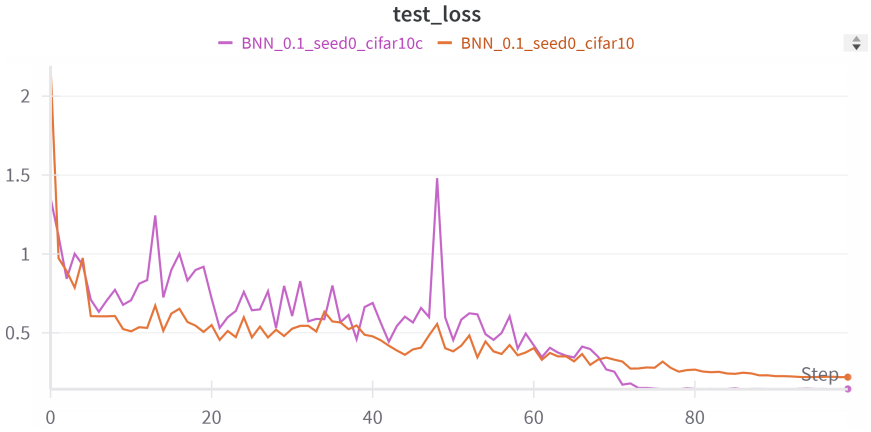

图 4. CIFAR-10 和 CIFAR-10-C 的测试损失比较

实验结果显示，模型在 CIFAR-10 数据集上的测试损失明显低于 CIFAR-10-C 数据集。在初始阶段，CIFAR-10-C 的损失较高，但随着训练的进行，其损失逐渐下降，并在训练后期趋于稳定，但仍高于 CIFAR-10。这说明噪声和腐蚀对模型的性能造成了显著影响，但通过训练，模型可以在一定程度上适应分布外场景。从 KL 散度曲线可以看出，CIFAR-10-C 数据集的 KL 损失在训练初期快速上升并在中后期逐渐下降，而 CIFAR-10 数据集的 KL 损失整体较低且下降趋势更加平缓。这表明模型在 CIFAR-10-C 数据集上需要更长的时间调整权重分布，以匹配复杂数据的先验分布。测试集准确率的曲线表明，模型在 CIFAR-10 数据集上的准确率整体高于 CIFAR-10-C 数据集（紫色曲线）。虽然在训练初期两者准确率差距较小，但随着训练的进行，CIFAR-10 的准确率稳定在 90% 以上，而 CIFAR-10-C 的准确率则稍低于 90%。

CIFAR-10-C 的损失较高反映了分布外数据集的复杂性。尽管模型能够通过稀疏贝叶斯推断部分适应噪声，但在数据质量较低的情况下，模型仍然面临较大的泛化挑战。数据的噪声和腐蚀特性导致模型后验分布与先验分布之间的偏差较大，训练过程中需要更多的迭代才能将 KL 散度收敛到较低水平。这说明模型在分布外场景下的推断更加复杂，但稀疏子空间方法能够有效调整模型以适应这种变化。准确率较低是由于噪声和腐蚀的存在增加了分类难度。然而，模型在分布外数据上的表现仍然较为接近于 CIFAR-10，表明稀疏贝叶斯神经网络在复杂场景下具有较强的适应性和鲁棒性。准确率的提升也显示了稀疏子空间方法在处理不确定性和噪声数据方面的有效性。

6 总结与展望

通过实验验证，SSVI 方法在标准数据集（CIFAR-10）和分布外数据集（CIFAR-10-C）上均表现出了显著的适应性和鲁棒性。在分布外场景下，尽管噪声和腐蚀增加了任务难度，模型仍能通过稀疏化策略降低损失并维持较高的准确率，同时有效量化了不确定性。此外，引入校准误差分析进一步丰富了评估维度，展示了稀疏 BNNs 在复杂场景中的实际应用价值。

对于分布外数据的复杂性，稀疏化方法的适应能力仍有提升空间，特别是在应对更大规模或更多类型分布偏移的场景中 [5]。实验中校准误差虽能反映模型置信度与准确性的匹配情况，但在更加动态的分布变化场景中，其评估指标可能需要进一步优化。

在实际部署中，如何进一步降低稀疏模型的硬件计算成本是未来值得探索的方向。

参考文献

[1] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[2] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

[3] Insung Kong, Dongyoon Yang, Jongjin Lee, Ilsang Ohn, Gyuseung Baek, and Yongdai Kim. Masked bayesian neural networks: Theoretical guarantee and its posterior inference. In *International Conference on Machine Learning*, pages 17462–17491. PMLR, 2023.

[4] Junbo Li, Zichen Miao, Qiang Qiu, and Ruqi Zhang. Training bayesian neural networks with sparse subspace variational inference. *arXiv preprint arXiv:2402.11025*, 2024.

[5] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *International conference on machine learning*, pages 2498–2507. PMLR, 2017.