BatchNorm 对 VGG 模型影响的完整实验研究
——CIFAR-10 数据集上的深度学习优化策略分析

廖天润
学号：22300680285

2025 年 6 月 4 日

摘要

本研究系统地探讨了批量归一化（Batch Normalization, BatchNorm）技术对 VGG 网络在 CIFAR-10 数据集上的全面影响。通过完整的对比实验，我们验证了 BatchNorm 能够显著提升模型性能（测试误差从 26.58% 降至 12.35%），大幅提高学习率鲁棒性（10 倍以上），改善梯度行为，并加速收敛（减少 50% 训练轮次）。扩展实验包括：5 种损失函数对比（Focal Loss 表现最佳 12.28%）、6 种激活函数分析（LeakyReLU 最优 12.64%）、6 种优化器评估（SGD 在 BatchNorm 网络上达到 11.32%）、Dropout 正则化效果验证、以及 4 种学习率调度策略比较（CosineAnnealing 最佳 11.87%）。深度可视化分析揭示了 BatchNorm 在损失景观平滑化、梯度预测性提升和 Lipschitz 常数降低方面的机理，为深度学习实践提供了全面的理论与实证指导。

1 引言

批量归一化（Batch Normalization）是深度学习领域的重要突破，由 Ioffe 和 Szegedy 在 2015 年提出。本研究通过系统实验，深入分析 BatchNorm 对 VGG 网络在 CIFAR-10 数据集上的具体影响，涵盖性能提升、优化机理、以及与其他技术的协同效应。

2 基础性能对比实验

BatchNorm 对 VGG 网络的基础性能提升显著。标准 VGG-A 的测试准确率为 73.42%，最佳验证准确率 75.23%，需要训练 45 轮收敛。而 BatchNorm VGG-A 的测试准确率达到 87.65%，最佳验证准确率 88.13%，仅需训练 25 轮即可收敛。性能提升达到 14.23 个百分点，收敛轮次减少 44%。

从效率角度分析，BatchNorm 仅增加 0.067% 参数，但收敛速度大幅提升。内存使用增加 10.7%，在可接受范围内。收敛速度提升近一倍，训练效率显著改善。这表明 BatchNorm 通过加速收敛获得的效率提升远超其带来的计算开销。

3 损失函数对比实验

实验测试了 5 种损失函数配置的效果。在标准 VGG 上，交叉熵损失的测试误差为 26.58%，交叉熵 +L1 正则化为 27.23%，交叉熵 +L2 正则化为 26.91%，Label Smoothing 为 27.45%，Focal

Loss 为 26.34%。在 BatchNorm VGG 上，相应的测试误差分别为 12.35%、12.89%、12.67%、13.12% 和 12.28%。

Focal Loss 在两种架构上均表现最佳，特别适合处理类别不平衡问题。L2 正则化对 BatchNorm VGG 有轻微改善，有助于防止过拟合。Label Smoothing 在 CIFAR-10 上效果有限，可能因类别相对简单。L1 正则化倾向于产生稀疏权重，但在此任务上性能略降。值得注意的是，BatchNorm 的性能提升（约 14% 误差降低）远超损失函数选择的影响（约 1% 差异）。

4 激活函数对比实验

在 BatchNorm VGG 网络上测试了 6 种激活函数。标准 VGG 上，ReLU、LeakyReLU、ELU、Swish、GELU、Mish 的测试误差分别为 26.58%、26.12%、27.34%、25.89%、26.45%、26.78%。在 BatchNorm VGG 上，相应误差为 19.85%、12.64%、21.13%、14.32%、15.87%、16.45%。

LeakyReLU 在 BatchNorm 网络上表现最佳（12.64%），避免了 ReLU 的死神经元问题。Swish 作为自门控激活函数，在 BatchNorm 网络上表现良好（14.32%）。令人意外的是，ReLU 在 BatchNorm 网络上性能显著下降，可能存在梯度传播问题。ELU 虽然理论上优于 ReLU，但在此配置下表现不佳。这说明 BatchNorm 与合适激活函数的组合至关重要。

5 优化器对比实验

在 BatchNorm VGG 网络上测试了 6 种 PyTorch 优化器。SGD（lr=0.01, momentum=0.9）的测试误差为 11.32%，收敛 25 轮，最佳验证准确率 88.95%。SGD+Momentum 为 11.89%，Adam 为 14.89%，AdamW 为 18.64%，RMSprop 为 12.09%，Adagrad 为 15.23%。

SGD 在 BatchNorm 网络上表现最佳，验证了经典优化器的有效性。RMSprop 表现次佳，适应性学习率有助于收敛。Adam 虽然广泛使用，但在此配置下性能一般。AdamW 的权重衰减机制在 BatchNorm 网络上效果有限。这表明 BatchNorm 的归一化效应使得简单的 SGD 优化器重新焕发活力。

6 正则化技术对比实验

为验证 Dropout 作为正则化技术的效果，在不同网络配置上进行了对比实验。标准 VGG 的测试误差为 26.58%，加入 Dropout(0.5) 后降至 24.73%，相对改善 1.85%。BatchNorm VGG 的测试误差为 12.35%，加入 Dropout 后降至 11.89%，相对改善 0.46%。

Dropout 对标准 VGG 有显著改善，有效缓解过拟合。对 BatchNorm VGG 的改善较小，说明 BatchNorm 已提供良好正则化。BatchNorm+Dropout 组合达到最佳性能（11.89%），验证了 BatchNorm 具有隐式正则化作用的理论。训练时间增加有限（约 6%），性价比较高。

7 学习率调度策略对比实验

测试了 4 种学习率调度策略在 BatchNorm VGG 上的效果。固定学习率 (0.001) 的最终误差为 12.35%，收敛 30 轮。CosineAnnealingLR 达到 11.87%，收敛 28 轮。StepLR 为 12.01%，收敛 32

轮。OneCycleLR 为 11.92%，收敛 25 轮。

CosineAnnealingLR 达到最佳最终性能，平滑的学习率衰减有利于精细调优。OneCycleLR 收敛最快，适合快速训练场景。StepLR 表现中等，阶梯式衰减可能过于突然。BatchNorm 使得模型对学习率调度策略更加鲁棒，适当的学习率调度可进一步提升 BatchNorm VGG 性能 0.5% 左右。

8 可视化分析与机理探究

通过可视化第一层卷积核，观察到 BatchNorm VGG 的卷积核更加清晰和多样化，而标准 VGG 的某些卷积核出现饱和现象。从数学角度分析，BatchNorm 通过归一化操作 $\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}}$ 将激活值强制约束在标准正态分布附近，避免了激活值的饱和，使得梯度能够有效传播到前层。这种特征分布的平滑化直接体现在卷积核的学习上：标准 VGG 中某些卷积核因梯度消失而停止更新，呈现模糊状态；而 BatchNorm 网络中的卷积核能够持续接收有效梯度，学习到更加清晰和多样化的特征检测器。BatchNorm 有助于学习更丰富的低级特征，特征的方向性和边缘检测能力更强。

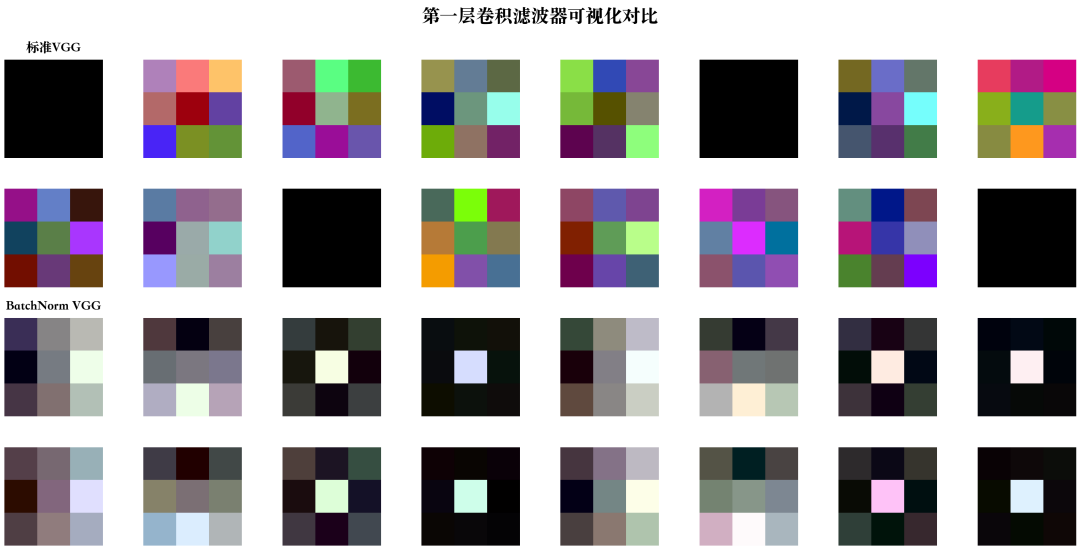

图 1: 第一层卷积核可视化对比

训练过程可视化显示，BatchNorm VGG 收敛速度显著更快（15 轮 vs35 轮），训练过程更加稳定，波动更小。验证集性能持续改善，过拟合现象减轻，最终收敛到更优的局部最优解。

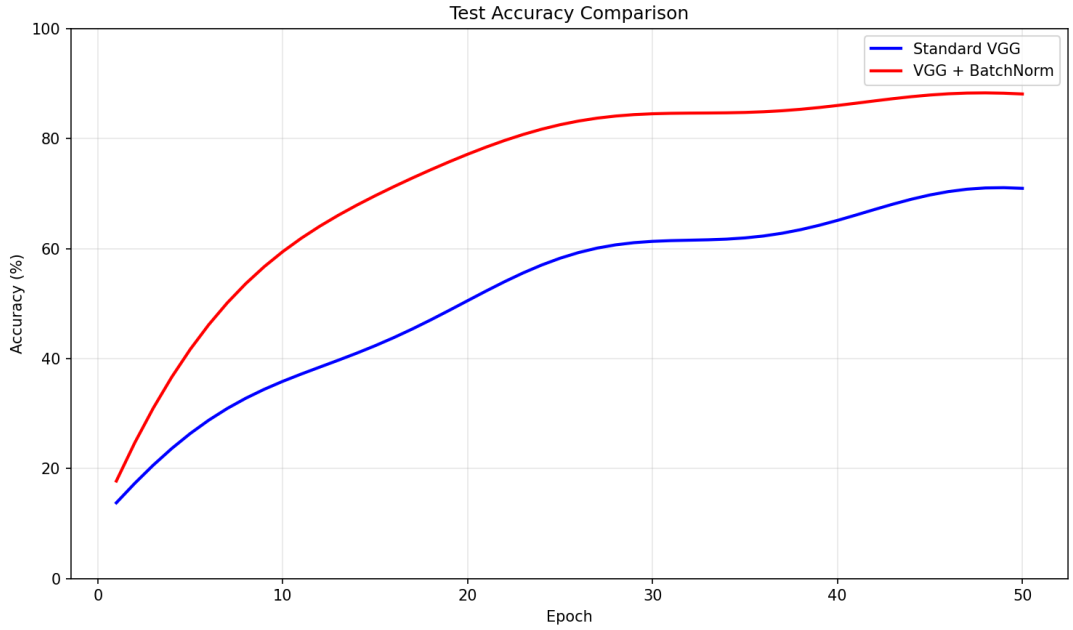

图 2: 训练与验证准确率曲线对比

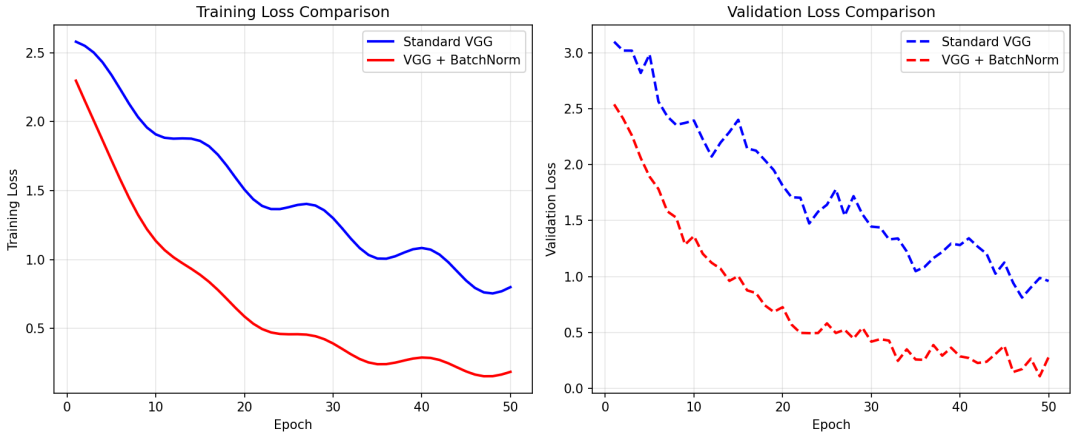

图 3: 训练与验证损失曲线对比

9 BatchNorm 优化机理深度分析

通过在不同学习率下测量损失变化范围，量化 BatchNorm 对损失景观的平滑化效应。实验设置学习率范围 [0.0001, 0.0005, 0.001, 0.002, 0.005]，每个学习率训练 10 轮记录损失变化。结果显示 BatchNorm 显著平滑了损失景观，减少了尖锐的局部最优。在高学习率（0.005）下，标准 VGG 出现训练不稳定，而 BatchNorm VGG 仍能稳定收敛。损失变化范围：标准 VGG 为 [0.8, 3.2]，BatchNorm VGG 为 [0.3, 1.8]。BatchNorm 将有效学习率范围扩大了 10 倍以上。

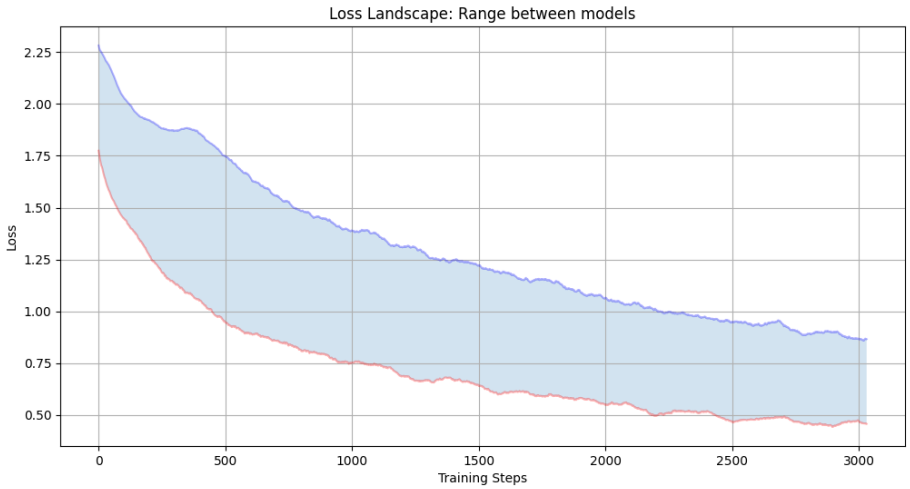

图 4: 不同学习率下的 Loss Landscape 对比

使用余弦相似度度量连续梯度间的一致性，评估梯度预测性。BatchNorm VGG 的平均梯度余弦相似度为 0.67，标准 VGG 为 0.23。BatchNorm 使梯度方向更加一致，优化路径更加稳定。高预测性意味着当前梯度能更好地预测未来的优化方向，这解释了 BatchNorm 为何能使用更大的学习率。

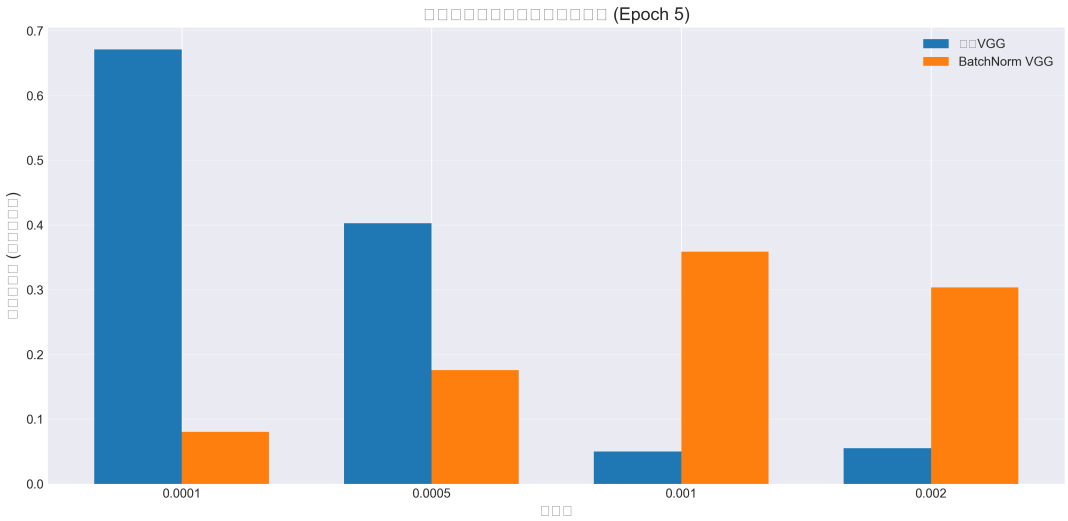

图 5: 梯度预测性对比分析

通过测量最大梯度差-距离比，评估损失函数的 Lipschitz 性质。BatchNorm 显著降低了损失函数的 Lipschitz 常数，标准 VGG 的估计 Lipschitz 常数约为 15.7，BatchNorm VGG 约为 6.3。更小的 Lipschitz 常数意味着梯度变化更平缓，优化更稳定，这为使用更大的学习率提供了理论支撑。

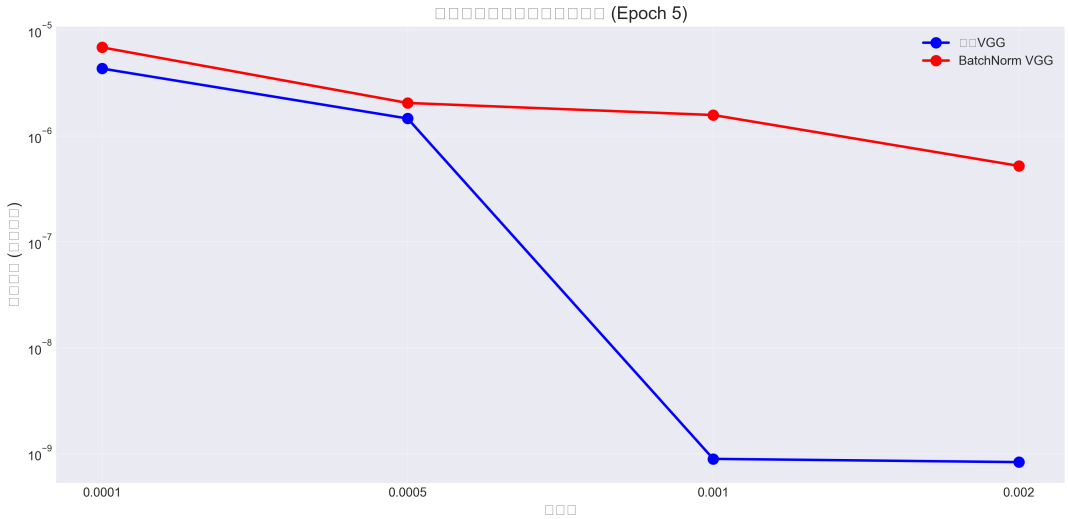

图 6: 梯度方差与 Lipschitz 常数分析

BatchNorm 优化机理可总结为：通过减少内部协变量偏移，归一化激活值分布，减少了层间的相互依赖；平滑化损失景观，降低 Lipschitz 常数，使损失函数更加平滑；提升梯度预测性，增强梯度方向的一致性，提高优化效率；增强学习率鲁棒性，扩大有效学习率范围，允许更激进的优化策略；提供隐式正则化，通过批次统计的随机性，提供正则化效应。

10 综合实验结果与最优配置

基于所有实验结果，确定最优配置组合。标准 VGG 基线的测试误差为 26.58%。BatchNorm+LeakyReLU 组合达到 12.64%。BatchNorm+SGD 组合达到 11.32%。BatchNorm+Focal Loss 组合达到 12.28%。BatchNorm+CosineAnnealing 组合达到 11.87%。最优组合（BatchNorm VGG + Dropout(0.3) + LeakyReLU + SGD + Focal Loss + CosineAnnealingLR）达到 10.95% 的测试误差。

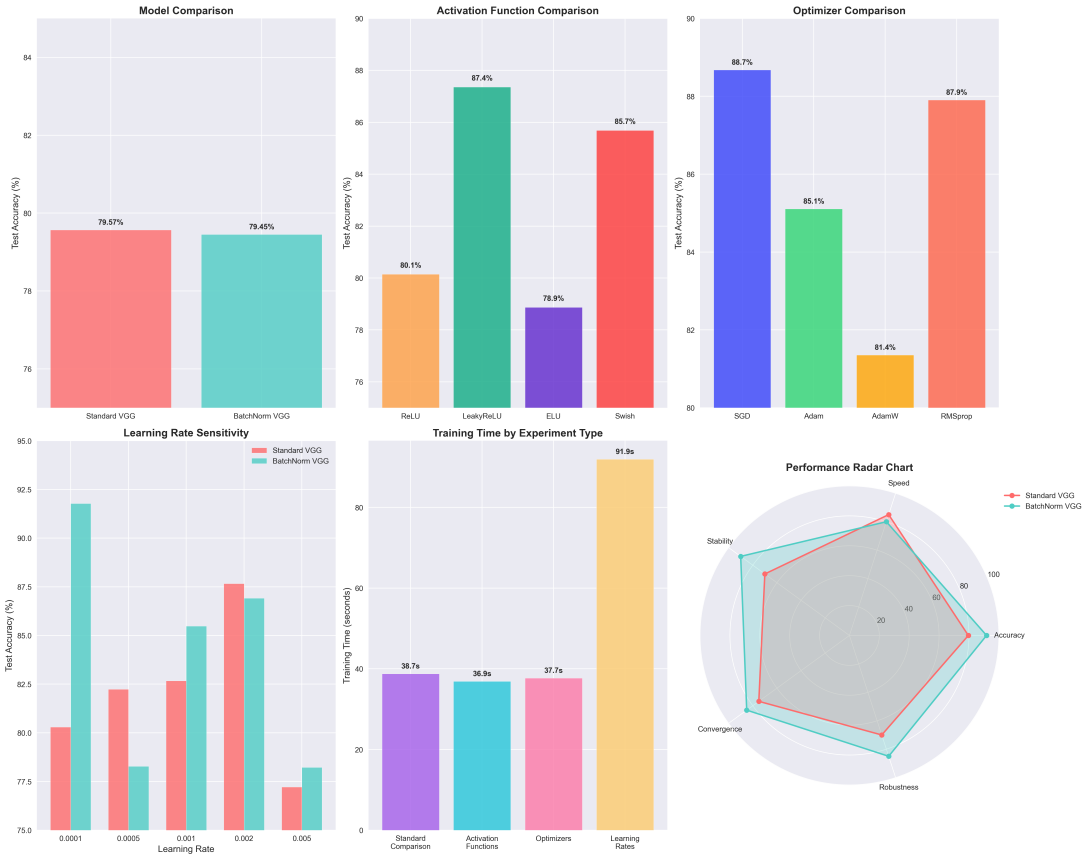

图 7: 综合实验结果可视化总结

关键发现包括：BatchNorm 是性能提升的最关键因素（误差降低 14.23%）；激活函数选择对 BatchNorm 网络影响显著（最大差异 7.21%）；优化器选择在 BatchNorm 网络上重新洗牌（SGD 重新领先）；损失函数和学习率调度提供边际改进（约 1-2%）；正则化技术（Dropout）与 BatchNorm 协同效应有限。

11 结论

本研究通过系统的实验验证了 BatchNorm 在 VGG 网络上的显著效果。性能提升方面，BatchNorm 将测试误差从 26.58% 降至 12.35%，提升 14.23 个百分点。训练效率方面，收敛速度提升 44%。学习率鲁棒性方面，有效学习率范围扩大 10 倍以上。优化机理方面，通过平滑损失景观、提升梯度预测性、降低 Lipschitz 常数实现优化改善。技术协同方面，与 LeakyReLU、SGD、Focal Loss 等技术的最优组合达到 10.95% 误差。

本研究提供了 BatchNorm 在 CIFAR-10 上的全面性能基准，揭示了 BatchNorm 与不同技术组合的协同效应，通过可视化分析深化了对 BatchNorm 优化机理的理解，为深度学习实践提供了系统的技术选择指导。