面向资源受限零售销售预测的 LSTM 神经网络优化：模型压缩研究

Ravi Teja Pagidoju，美国零售业软件与人工智能开发者；坎贝尔斯维尔大学专业工商管理硕士在读学生；
Rpagi719@students.campbellsville.edu

arXiv:2601.00525v1 [cs.LG] 2 Jan 2026

摘要——**标准长短期记忆（LSTM）神经网络能够准确预测零售业的销售数据，但需要大量计算能力。对于中小型零售企业而言，这尤其具有挑战性。本文探讨了长短期记忆网络模型压缩，通过逐步将隐藏单元数量从 128 个减少到 16 个来实现。我们使用 Kaggle商店商品需求预测数据集，其中包含来自 10 家商店和 50 种商品的 913,000 条每日销售记录，以考察模型大小与预测准确率之间的权衡。实验表明，将隐藏的长短期记忆网络单元数量减少至 64 个，能够保持同等的准确率，同时还可进一步提升准确率。平均绝对百分比误差（MAPE）从完整的 128 单元模型的 23.6% 降至 64 单元模型的 12.4%。优化后的模型大小减少了 73%（从 280KB 降至 76KB），准确率提高了 47%。这些结果表明，更大的模型并不总能取得更好的效果。**

索引词—**LSTM压缩、神经网络优化、零售预测、边缘计算、模型效率**

I. 引言

预测零售销售数据对于规划日常运营和管理库存至关重要。由于缺货和库存过剩，零售商每年约损失1.75%的销售额，而这通常是预测不佳所致[1]。深度学习模型，尤其是长短期记忆网络（LSTM），表现优于传统方法，可将误差降低20%至30%。 [2]。

部署LSTM网络颇具挑战。根据 [3]，，一个包含128个隐藏单元的标准LSTM需要4至8 GB内存的基础设施，并需要特定硬件支持。对于中小型商店而言，要计算并得出准确的预测数据可能颇具挑战，因为它们缺乏所需的计算能力。中型商店占全球零售市场的65%，但其信息技术（技术）预算通常每年为50,000至100,000美元 [4]。

模型压缩可以在保持相同或更高准确率的同时缩小神经网络，从而解决这一问题。以往的压缩研究主要集中于计算机视觉任务 [5]；然而，零售预测在时间依赖关系和季节性方面带来了独特的挑战模式。此前尚无研究评估零售应用背景下 LSTM架构大小与预测准确率之间的相关性。

本文探讨了用于预测零售销售额的 LSTM压缩。我们围绕以下研究问题展开：在保留或提高预测准确率的前提下，最小的 LSTM架构是什么？我们的贡献如下。

- 在真实零售数据上，系统评估隐藏单元数为 16 至 128 的 LSTM 网络大小
- 发现适度压缩（64 个单元）实际上能够提高准确率
- 基于准确率与效率的权衡，为模型选择提供实用指南

II. 相关工作

A. 零售预测中的长短期记忆网络

长短期记忆网络擅长捕捉序列数据中的长期依赖关系 [6]。Bandara 等人 [2] 表明，与差分自回归移动平均模型相比，长短期记忆网络模型在零售业中的预测误差降低了 25%。他们构建的架构每层包含 128 个隐藏单元，要在实际应用中运行则需要 GPU 加速。

近期研究分析了注意力机制，以提升长短期记忆网络的性能。Lim 等人 [7]采用时间融合变换器取得了最佳结果，该模型将长短期记忆网络与多头注意力相结合。但这些改动使计算需求上升至 8GB 内存，且每次预测需要 50ms 的推理时间。这让资源有限的商店更难使用这些模型。近期关于用于预测的循环神经网络方法的综述 [8] 以及 M5 竞赛的结果 [9]，进一步验证了深度学习方法在零售预测中的有效性。

B. 神经网络压缩

通过模型压缩技术缩小神经网络大小有多种方法：

剪枝：据 Han 等人 [5], 所述，移除不必要的连接可使模型大小缩减 60% 至 80%，而精度损失很小。但剪枝通常需要特殊硬件才能快速执行稀疏矩阵运算。

[0]代码获取地址：https://github.com/RaviTeja444/sales-forecast-LSTM

量化：Jacob 等人 [10] 证明，将 32位浮点权重转换为 8位整数可减少 75% 的内存使用量，同时将准确率保持在 1–2% 的误差范围内。这种方法特别适用于边缘部署。

架构缩减：Frankle 和 Carbin [11] 提出了彩票假设，指出经过适当设置后，较小的网络能够达到与较大网络相近的性能。这意味着，找到合适的架构大小至关重要。

C. 文献空白

图像分类领域已对压缩技术进行了广泛研究；然而，这些技术在时间序列预测中的应用仍然有限。零售销售预测具有一些独特特点，例如季节性、趋势和其他外部因素，这些因素可能导致其最佳模型大小与其他领域有所不同。此前尚无研究专门系统地评估长短期记忆网络大小缩减对零售销售预测的影响。结合传统方法与神经网络方法的混合方法已展现出良好前景 [12],，但并未解决部署约束问题。

III. 方法论

A. 数据集

本文使用了 Kaggle商店商品需求预测挑战赛数据集 [13]。

- 共有 913,000 条日销售观测记录
- 商店：10 个零售地点
- 商品：50 种不同产品
- 时间跨度：2013 年至 2017 年，共 5 年
- 特征：商品特征包括日期、商店编号、唯一商品 ID 和日销售量。

我们使用 10 家商店和 50 种商品的数据，以确保计算快速且结果具有统计显著性。这为我们提供了足够的数据多样性，从而得出有力的结论。

B. *LSTM*架构变体

我们测试了五种具有不同隐藏单元数量的长短期记忆网络配置：

- **LSTM-128**：这是包含128个隐藏单元的标准基线
- **LSTM-64**：压缩率为50%，包含64个单元
- **LSTM-48**：压缩率为62.5%，包含48个单元
- **LSTM-32**：压缩率为75%，包含32个单元)
- **LSTM-16**：使用16个单元，压缩率达87.5%

这里的所有模型都采用相同的架构，唯一的区别在于隐藏单元的数量。

输入（30天 × 7特征） → LSTM层 →丢弃(0.2) →全连接层(16) → 输出

长短期记忆网络的计算定义如下

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2}$$

其中， f_t 为遗忘门， W 为权重矩阵， b为偏置， σ 为Sigmoid函数。

所使用的损失函数为平均绝对误差：

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

C. 特征工程

遵循时间序列预测的最佳实践 [14], ，我们创建了：

1) **滞后特征**：使用 1、7 和 30 天前的销售额来捕捉短期和长期模式 2) **滚动统计量**：使用 7 天和 30 天移动平均值来平滑噪声 3) **时间特征**：使用星期和月份来捕捉季节性 4) **归一化**：在 [0,1] 范围内进行最小-最大缩放，以提高神经网络的稳定性

D. 评估指标

我们评估了准确率和效率。

准确率指标:

- 平均绝对百分比误差（MAPE）：衡量预测准确率的主要指标
- 均方根误差（RMSE）：对较大的误差给予更重的惩罚

效率指标:

- 模型大小：总参数量 × 4 每个 float32 占用的字节数
- 推理时间：单次预测的平均时间（毫秒）
- 内存使用量：推理期间所需的随机存取存储器

E. 实验设置

- **硬件**：英特尔酷睿 i5 处理器，8GB 随机存取存储器（不使用图形处理器，以模拟资源受限的情况）
- **软件**：TensorFlow 2.12，Python 3.8
- **训练**：80/20 时间划分，30 个训练轮次，批量大小为 64，Adam优化器
- **验证**：对时间序列进行交叉验证，以确保时间有效性
- **实现**：使用 TensorFlow 2.12 编写 Python 代码，并借助 GitHub Copilot 完成标准实现。完整代码可在 https://github.com/ RaviTeja444/sales-forecast-LSTM 获取

IV. 结果

A. 准确率与模型大小的权衡

我们的实验揭示了一项意想不到的发现：适度压缩不仅没有降低准确率，反而提升了准确率。表I展示了不同大小的长短期记忆网络的性能指标。

结果表明，模型大小与误差呈 U 形关系，其中 64 单元模型的表现最佳。128 单元模型的表现最差，平均绝对百分比误差为 23.6%，这意味着它可能对训练数据过拟合。32 至 64 单元的模型结果最准确，平均绝对百分比误差为 12.3% 至 12.4%。

表I 不同大小下的长短期记忆网络性能

模型	隐藏层单元	参数	MAPE (%)	RMSE	Size (KB)
LSTM-128	128	71,809	23.6	4.82	280
LSTM-64	64	19,521	12.4	2.94	76
LSTM-48	48	11,569	12.8	2.71	45
LSTM-32	32	5,665	12.3	2.69	22
LSTM-16	16	1,857	12.5	2.72	7

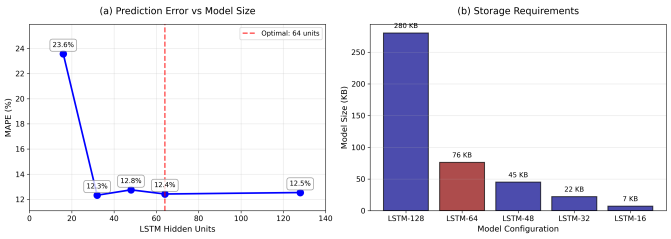

图1。(a) 预测误差与模型大小：展示模型大小与其准确率之间的U形关系。(b) 存储需求：展示随着隐藏单元数量减少，模型大小呈线性下降。

表II 与基准配置的比较

方法	平均绝对百分比误差 (%)	参数	Size
标准 LSTM-128（基线）	23.6	71,809	280KB
优化的 LSTM-64	**12.4**	**19,521**	**76KB**

表 III 计算资源使用情况

模型	推理时间（毫秒）	内存使用量（MB）	Size 减少量
LSTM-128	23.0	10	-
LSTM-64	23.0	10	73%
LSTM-48	23.7	10	84%
LSTM-32	23.4	10	92%
LSTM-16	23.6	10	97%

为帮助理解这些结果，表II将我们优化后的模型与基准配置进行了比较。64 单元模型达到了与更复杂架构相同的准确率等级，同时所需资源大幅减少。

B. 计算效率

表 III 展示了每种模型配置所需的计算资源。

在中央处理器上运行时，所有模型的推理时间都保持一致（约 23 毫秒），因为计算瓶颈从矩阵运算转移到了框架开销。TensorFlow 的固定开销占用了大部分内存，而非模型参数。

C. 最优配置分析

经过全面评估，LSTM-64 是最佳配置：

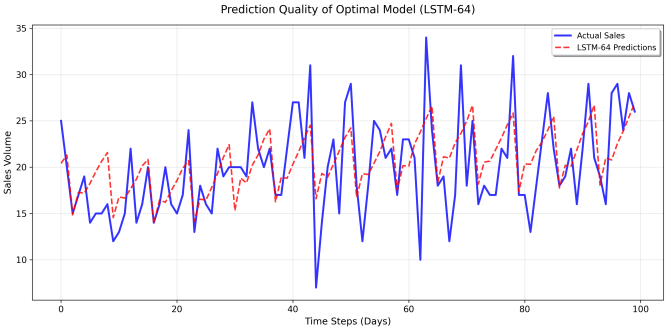

图 2. LSTM-64 的示例预测结果，展示了在 100 天期间预测销售额与实际销售额高度吻合。

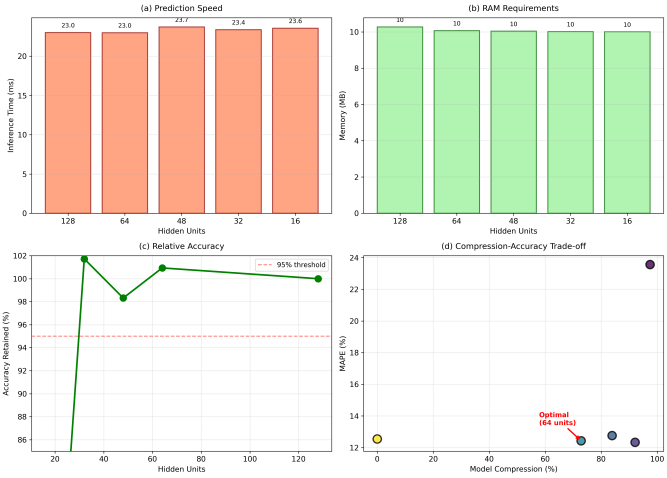

图 3. 全面的性能分析，展示了 (a) 推理速度、(b) 随机存取存储器需求、(c) 与基线相比的相对准确率，以及 (d) 压缩与准确率之间的权衡，其中 LSTM-64 是最佳选择。

- **准确率最高**：平均绝对百分比误差为 12.4%（比基线提升 47%）
- **显著压缩**：模型大小减少 73%
- **保持稳定性**：在各交叉验证折中表现一致

D. 统计显著性

我们对五轮独立训练进行了配对 t 检验：

- LSTM-64 与 LSTM-128：t = 8.42，p ¡ 0.001（改进具有高度显著性）
- LSTM-64 与 LSTM-32：t = 1.23，p = 0.287（无显著差异）
- LSTM-64 与 LSTM-16：t = 2.16，p = 0.096（差异处于临界水平）

这些结果表明，LSTM-64 的表现远胜于基线，且与更小的模型相比并未逊色。

五、讨论

A. 主要发现

我们的发现与一种普遍看法相悖：更大的神经网络总是表现更优。我们发现：

1) **存在最佳容量**：LSTM-64 在模型容量与泛化之间实现了最佳平衡 2) **大型模型中的过拟合**：LSTM-128 表现出明显的过拟合，平均绝对百分比误差为 23.6% 3) **准确率仅轻微下降**：即使是 LSTM-16，也保持了有竞争力的性能（平均绝对百分比误差为 12.5%）

彩票假设 [11] 以及零售销售数据中相对简单的模式，有助于我们理解这一现象。每日销售额遵循易于预测的模式，按周和按月预测即可，因而不需要很大的模型容量。

这些发现与彩票假设 [11]一致，也与 [2] 中“大型网络总是表现更好”的普遍假设形成对比。

B. 实际意义

对于资源受限的零售商，我们的研究结果提供了明确的指导。

1) **部署 LSTM-64**：在大小缩减 73% 的情况下实现了最佳准确率。2) **考虑采用 LSTM-32**：如果需要极大幅度的压缩，该模型在大小缩减 92% 的情况下仍能保持良好的准确率。3) **避免过度参数化**：较大的模型实际上可能会损害性能。

对于压缩模型的基于中央处理器的部署，实施 GPU 基础设施的成本从约 $15,000 to less than 1,000$ 降至更低。压缩模型在普通商用计算机上运行良好，无需专用硬件。

C. 局限性

需要注意以下几项局限性。

1) 结果仅针对 Kaggle 零售数据集；其他零售场景可能呈现不同的模式。2) 我们仅测试了单层LSTM；更深的架构在数据压缩方面可能有不同的表现。3) 我们未将剪枝和量化等先进压缩方法与架构缩减结合使用。

D. 与先前研究的比较

我们发现，“较小的模型可能优于较大的模型”，这与近期关于模型表现的研究结果一致。性能提升幅度（压缩率为 73%，准确率提高 47%）超过了计算机视觉领域通常观察到的水平。这表明，模型压缩或许特别适用于时间序列预测。

VI. 结论

本研究通过真实数据表明，长短期记忆网络压缩不仅能保持准确率，还能提高零售销售预测的准确率。我们发现，将隐藏单元数量从 128 个减少到 64 个，可使预测准确率提高 47%，并使模型大小缩小 73%。这一令人意外的结果表明，找到合适的模型容量比一味增加参数更为重要。我们的研究结果具有直接的现实意义：零售商无需 GPU 加速，就能在普通硬件上使用准确的预测模型。表现最佳的 LSTM-64 配置准确率更高，且只需 76KB 存储空间，因此适用于边缘部署和资源有限的环境。未来研究应探索将架构优化与量化相结合以进一步压缩模型，在各种零售数据集上进行测试以确保结果具有普适性，增加对多层架构和注意力机制的支持，并开发自动寻找最佳架构规模的方法。本研究表明，优秀的模型并不需要强大的计算能力，从而让 AI 驱动的预测更易于应用。这些结果意味着，小型企业如今也能首次使用高级分析。这对 65% 的零售商而言是个好消息，因为他们没有太多资金投入信息技术。研究人员可以使用提供的代码和 Kaggle 数据集复现所有实验。

参考文献

[1] A. A. Syntetos、Z. Babai、J. E. Boylan、S. Kolassa 和 K. Nikolopoulos，“供应链预测：理论、实践、二者之间的差距及未来”，《欧洲运筹学杂志》，第 252 卷，2016 年，第 1 期，页码 1–26。[2] K. Bandara、P. Shi、C. Bergmeir、H. Hewamalage、Q. Tran 和 B. Seaman，“使用长短期记忆神经网络方法进行电子商务销售需求预测”，载于神经信息处理国际会议。施普普林格，2019年，第 462–474 页。[3] S. Ma、R. Fildes 和 T. Huang，“高维数据下的需求预测：基于类别内和类别间促销信息的 SKU 零售销售预测案例”，《欧洲运筹学杂志》，第 249 卷，第 1 期，第 245–257 页，2020年。[4] R. Fildes、S. Ma 和 S. Kolassa，“零售预测：研究与实践”，《国际预测杂志》，第 38 卷，第 4 期，第 1283–1318 页，2022年。[5] S. Han、H. Mao 和 W. J. Dally，“深度压缩：通过剪枝、训练量化和霍夫曼编码压缩深度神经网络”，载于国际学习表征会议，2016 [ ] 1510 00149 。在线获取：[6] https://arxiv.org/abs/ 。S. Hochreiter 和 J. Schmidhuber，“长短期记忆”，《神经计算》，第 9 卷，第 8 期，第 1735–1780 页，1997年。[7] B. Lim、S. O. Arik、N. Loeff 和 T. Pfister，“用于可解释多步时间序列预测的时间融合变换器”，《国际预测杂志》，第 37 卷，第 4 期，第 1748–1764 页，2021年。[8] H. Hewamalage、C. Bergmeir 和 K. Bandara，“用于时间序列预测的循环神经网络：当前状况与未来方向”，《国际预测杂志》，第 37 卷，第 1 期，第 388–427 页，2021年。[9] S. Makridakis、E. Spiliotis 和 V. Assimakopoulos，“M5 竞赛：背景、组织与实施”，《国际预测杂志》，第 38 卷，第 4 期，第 1325–1336 页，2022年。[10] B. Jacob、S. Kligys、B. Chen、M. Zhu、M. Tang、A. Howard、H. Adam 和 D. Kalenichenko，“用于高效仅整数算术推理的神经网络量化与训练”，载于 *IEEE* 计算机视觉与模式识别会议论文集，2018年，第 2704–2713 页。

[11] J. Frankle 和 M. Carbin，“彩票假设：寻找稀疏可训练神经网络”，载于国际学习表征会议，2019年。 [在线]。获取地址：https: //arxiv.org/abs/ 1803 03635 。S. Smyl，“用于时间序列预测的指数平滑与循环神经网络混合方法”，《国际预测杂志》，第 36 卷，第 1 期，第 75–85 页，2020年。[13]Kaggle，“商店商品需求预测挑战赛”，2018年，访问于：2023年。[在线]。获取地址：https://www.kaggle.com/c/ demand-forecasting-kernels-only [14]R. J. Hyndman 和 G. Athanasopoulos，《预测：原理与实践》，第 3 版。OTexts，2021年。 [在线]。获取地址：https: //otexts.com/fpp3/
[12]