摘要

本文提出使用多元长期短期记忆（LSTM）模型来预测云工作负载中的资源使用情况，并使用基于分数差异的方法分析和比较了LSTM模型和双向LSTM模型的预测。本文对提出的LSTM模型进行了评估，并与Google集群跟踪[1]上的现有技术进行了比较，结果表明所提算法优于最新算法。

简介

未来资源使用预测被归结为时间序列分析问题，即通过分析过去的资源使用趋势以预测未来的资源使用趋势。前瞻性预测模型通过分析过去涉及的趋势以预测未来资源使用模式。

本文的第一个贡献是提出了用于未来云资源使用预测的多元LSTM模型，可用于未来云资源使用预测，第二个贡献是提出了BLSTM模型，可用于分析前向和后向预测网络对未来资源使用预测的影响。

现有云资源使用情况预测模型回顾

Gong等人 [3]使用马尔可夫模型预测Google集群跟踪中的资源。该模型将所需资源度量的使用划分为n个bin，并计算大小为n \*n的转移概率矩阵P；使用Chapman-Kolmogorov方程预测下一个bin的概率为\_t =\_t􀀀1P，其中\_t和\_t􀀀1分别表示在时间t和t􀀀1处状态的概率。

AGILE [4]使用小波变换和马尔可夫模型来预测云中的资源，其基本思想是：将时间序列分解为基于小波的信号；将马尔可夫模型应用于每个信号以生成未来的预测；通过从小波重构原始信号可以生成最终的样本外预测。

Zhang等人 [5]使用称为ARIMA方法的线性模型来预测资源消耗值，该模型将时刻t ，x的值预测为先前滞后的加权总和。

iOverbook，[6]使用前馈神经网络进行资源使用预测。神经网络模型的解析方程表示为：x 0 t = Xn k = 1 \_kg Xm i = 1 \_ikxi！其中，x表示输入时间序列。 m是延迟的数量，n是隐藏层中神经元的数量，g（:)是激活函数，\_，\_表示在训练过程中将学习的连接权重。

上述方法均假定时间序列是平稳的且无记忆，但研究表明云工作负载中存在长期依赖性[7]。

[10]对西澳大利亚急诊部门需求的预测分析表明，与使用单变量自回归模型相比，使用多元矢量自回归模型可提供更好的预测。

大多数研究使用单向计算依赖性来根据过去的信息预测未来的事件。而从 [11] 可以看出，与单向模型相比，双向模型中过去-将来依赖性的集成在模型的训练能力和预测质量上表现更佳。

双向长短期存储器（LSTM）网络体系结构

本研究提出的多元LSTM模型的时间序列X表示为多元连续时间序列X =（x1; x2;：：：; xN）其中xi 2 RD 8i = 1; N描述了使用不同资源和不同性能指标的时间演变。这里N是时间序列观测值的数量，D是每个观测值的维数。这里N是时间序列观测值的数量，D是每个观测值的维数。多元LSTM的主要目标是在时间点t预测资源xt j的值，该值取决于其自身以及其他资源的使用情况和不同的性能指标xt j = f1（xt􀀀1j）+ XD k = 1； k6 = j f2（xt􀀀1k）其中，f1和f2是两个函数。

图1显示了LSTM块的结构。 LSTM体系结构由不同的单元组成，每个单元由三个门组成，这三个门用于调节进出单元存储器的信息流。输入门调节输入信息到存储器的流，（ii）遗忘门调节应该存储在存储器中的信息，（iii）输出门调节块的输出激活。图中的矩形框代表一层S型/双曲线正切神经元。 LSTM单元中的遗忘门可以表示为：ft = sig（Wxfxt +Whfht􀀀1+ bf）（6）其中ft是遗忘门的输出，而sig（:)表示S型激活函数。 Wxf和Whf分别表示输入（xt􀀀1）和忘记门之间的权重，隐藏层（ht􀀀1）和忘记门之间的权重。 bf代表遗忘门的偏见。将旧单元格状态乘以ft即可忘掉部分状态信息。

双向LSTM通过使用前向和后向依赖关系来学习时间序列中的模式。双向模型在两个方向上处理数据，其中前向层从t = 1开始迭代； ：：：; T和后向层从t = T进行迭代； ：：：; 1并了解反向时间序列的趋势。图2显示了BLSTM网络的结构。这些模型使用两个单独的隐藏层来分析正向和反向的时间序列。然后将这些隐藏层作为输入传递到同一输出层。然后，将输出Yt计算为：Yt = Wb hyhbt + Wf hyhf t +b。其中hbt，hf t表示在时间t处向后和向前隐藏层的输出。 Wb hy和Wf hy分别表示最终输出与后向和前向隐藏层之间的权重。b代表偏差。

结果与讨论

本研究的实验是在Google簇跟踪[1]上进行的，实验主要分析和预测CPU资源使用率指标。实验抽取了60480个样本（7天）来训练资源预测模型，然后将时间序列的下20个样本（3分钟）用作验证数据，以选择适当的参数；生成并分析了后续60（10分钟），120（20分钟）和180（30分钟）步骤的样本外预测。资源使用率值以10秒的时间间隔汇总。

图3显示了40个延迟的CPU资源使用率的自相关图，数据显示连续观测之间的相关性存在非常缓慢的衰减，表明存在长期依赖性。

实验使用单变量和多变量LSTM和BLSTM模型生成样本外CPU使用率预测。实验将数据分成L批，大小为bs，然后X =（X1;：：：; XL）XL =（x（（L􀀀1）\_bs + 1）;：：：; x（L\_bs））; xi 2 RD在有状态模型中，所有状态均传播到下一批。下一批XL的计算中使用了批次（L􀀀1）中的观测状态。图4展示了在不同神经元处使用一层LSTM进行的无状态模型与有状态模型的比较。结果表明，与无状态LSTM模型相比，有状态LSTM在RMSE中产生的误差较小。

本研究使用逐层构造方案来构建多层LSTM和BLSTM模型，并在模型中探索了多达五层。实验发现，三层膜可以达到最佳效果。两层以下该模型欠拟合，而在三层以上该模型会过度拟合。

表1列出了（i）PRESS（ii）AGILE（iii）ARIMA和（iv）NARNN v）LSTM-U（单变量LSTM）（vi）BLSTM-U（单变量）的样本外预测的多步RMSE BLSTM）（vii）LSTM-M（多元LSTM）和（viii）BLSTM-M（多元BLSTM）资源预测模型。结果显示，多元LSTM和BLSTM网络比所有单变量模型产生更好的预测，而多元BLSTM在所有资源预测模型中表现最好。

图5和图6分别显示了由多元LSTM模型和多元BLSTM模型生成的样本外预测。结果显示，基于LSTM的模型生成更好的样本外预测，但是在训练时间和计算资源需求方面有更多的训练开销。

结论与未来工作

本研究提出了用于资源使用预测的LSTM和BLSTM模型，实验观察到LSTM网络很好地在基于时间序列的资源使用数据中对远程依赖关系进行了建模，并生成了更好的样本外预测。

实验比较了有状态和无状态LSTM模型的预测，并观察到由于有状态LSTM可以在批次之间保留内存，因此它们在资源使用率预测方面比无状态模型要好。

实验将BLSTM和LSTM的预测在CPU资源使用率预测上进行比较，表明未来和过去的预测转换之间的耦合提供了优于单向模型的预测能力。

未来将扩展本研究，以利用资源预测模型生成的预测来动态调度云中的资源。