使用深度双向长期短期内存网络预测CloudWorkload的资源使用情况

摘要

资源使用预测是在云中实现最佳资源配置的重要方面。云工作负载中存在长期依赖关系，这使得常规时间序列资源使用情况预测模型不适合进行预测。在本文中，我们建议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。我们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。已对提出的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。实验结果表明，所提算法优于最新算法。

1.简介

云计算使您可以快速访问共享的托管服务池。不同数量的用户在不同的时刻动态进入和离开云环境，以访问由云提供的服务[2]。由于观察到的在不同时间点访问云服务的用户数量存在不确定性，因此服务提供商很难有效地向其客户端分配资源。这使资源供应成为一个挑战性的问题。提供最佳数量的资源的效率低下导致服务提供商最终分配过多或过少的资源的过度/不足。预测客户的需求有助于服务提供商更好地进行容量规划，并动态扩展活动机器的数量，以实现可感知服务水平协议（SLA）的节能资源调配系统。

将未来资源使用预测表述为时间序列分析问题，其中分析资源使用中的过去趋势以预测未来的预期资源使用趋势。已经探索了不同的统计方法来预测未来的资源使用模式。例如，已经探索了马尔可夫模型[3]，[4]，自回归综合移动平均模型（ARIMA）[5]和神经网络[6]来预测未来的资源需求。这些方法中的大多数都假设相隔很长时间的观测值彼此无关。但是已经观察到，在遥远的观察之间确实存在依赖性。在[7]中，基于分数差分的时间序列预测模型用于分析和捕获云工作负载中的远程依赖性。 [8]中使用了长期短期内存（LSTM）模型来对云工作负载中的长期依赖性进行建模。在[8]中，单变量LSTM模型用于预测资源使用的未来趋势，其中仅使用CPU使用率值来预测CPU负载。多元时间序列预测利用不同特征之间的相互作用来生成未来的预测。多元时间序列预测有助于分析其他功能对所需资源指标的影响。据我们所知，文献中尚未报道使用多元LSTM来预测云工作负载中未来的资源使用情况。在我们的工作中，我们探索多变量分析并分析其他功能对CPU资源使用的影响。使用不同的功能预测使用情况，这些功能包括正在运行的作业数，内存使用情况，分配的内存使用情况，未映射的页面缓存，总页面缓存，在一个测量间隔内观察到的最大内存使用情况，磁盘I / O时间，使用的磁盘空间，最大CPU在一个测量间隔内观察到的使用情况，在一个测量间隔内观察到的最大磁盘I / O时间，每条指令的周期和每条指令的内存访问以及CPU使用率本身[1]。

前瞻性预测模型分析过去历史中涉及的趋势以预测未来模式。双向预测体系结构由于过去的依赖关系与未来的依赖关系的耦合而导致减少了预测误差[9]。据我们所知，文献中尚未报道使用双向LSTM（BLSTM）来预测云中未来的资源使用情况。在这项工作中，我们建议建立一个BLSTM模型来预测资源使用的未来趋势，并且我们还建议使用单变量和多变量LSTM和BLSTM模型来分析和比较基于不同分数差异的方法的预测，以预测资源的未来趋势。用法。我们在工作中使用Google集群跟踪[1]对云工作负载进行建模，并进行分析以预测资源使用价值。

本文的主要贡献是双重的。第一个贡献是用于未来云资源使用预测的多元LSTM模型。这与[8]相反，后者使用了单变量LSTM模型。我们的第二个贡献是BLSTM模型，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。

本文的结构如下。在第2节中，我们回顾了用于预测云网络中资源使用情况的时间序列模型。第3节介绍了LSTM和BLSTM模型的体系结构。在第4节中，介绍了为验证所提出的方法而进行的实验结果。结论和未来的工作在第5节中介绍。

2.审查现有模型以预测云中的资源使用情况

云中资源使用情况预测的最重要目标之一是提高预测质量，以实现更实际的预测。已经探索了不同的方法来预测不同资源的未来使用。在本节中，我们回顾一些重要的方法，这些方法使用时间序列预测方法探索了云中资源使用率预测的领域。

龚等。 [3]使用马尔可夫模型预测Google集群跟踪中的资源。将所需资源度量的使用划分为n个bin，并计算大小为n x 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（1），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。小波变换可在时域和频域中提供多分辨率分析，可捕获时间序列中的所有频率及其位置。 AGILE的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。张等。 [5]使用称为ARIMA方法的线性模型来预测资源消耗值。将时刻t x 0 t的值预测为先前滞后的加权总和。 x 0 t =a0xt􀀀1+：：：+am􀀀1xt􀀀m+b0\_t􀀀1+：：+bn􀀀1\_t􀀀n（2）其中xi是时刻i，a的实际值； b是ARIMA模型的权重。 m和n是使用的滞后数，\_是与模型关联的误差项。 ARIMA模型是一种广泛使用的时间序列预测模型，但是它不能有效地捕获时间序列数据中的非线性模式。

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！ （3）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。

上面提出的方法假定时间序列是平稳的且无记忆。但是，已经观察到云工作负载中存在远程依赖性[7]。长期依赖性是时间序列分析中的一种现象，其中下一步值的增/减最有可能受到序列中过去几个时间滞后的影响。在长存储时间序列中，与正常时间序列相比，连续观察之间的依存关系衰减得更慢。在[7]中，使用重新缩放的范围分析方法和Hurst参数值研究了云工作负载中远程依赖的存在。为了捕获数据中长存储的影响，应用了分数差分方法。分数微分算子定义为无穷二项式级数展开x 0 t = xt􀀀dxt􀀀1+ d（d􀀀1）2！ xt􀀀2􀀀d（d􀀀1）（d􀀀2）3！ xt􀀀3:::（4）其中d是差参数，可以取小数值。分数差异数据将传递到资源预测模型，并对该模型生成的预测进行后处理。该方法比没有分数差分的方法执行得更好，但是包括预处理和后处理数据的额外开销。本质上，LSTM模型适合处理长期依赖关系。在[8]中，使用LSTM网络预测了资源的未来需求。但是，仅通过分析CPU资源过去的历史趋势就可以预测对CPU资源使用的未来需求。

在[10]中已经观察到西澳大利亚急诊部门需求的预测，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。在我们的工作中，我们建议使用多元分析来预测未来的CPU资源使用情况。我们考虑其他功能的影响，例如内存使用率，页面缓存，正在运行的作业数等，以及过去的CPU资源使用率，以预测未来的趋势。在当前的工作中，我们建议构建多变量LSTM，以预测云工作负载中的资源使用情况。

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。在这项工作中，探索了多元LSTM模型的双向扩展，我们评估了其对样本外资源使用预测的有效性。在[11]中，通过对激光数据执行时间序列预测来评估双向神经网络。从实验上可以看出，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具有很好的优势。另一项工作[9]使用双向预测体系结构对黑子数据执行时间序列预测，并通过结合过去和将来的依赖关系观察到减少的预测误差。一项工作[12]利用双向极限学习机机制进行风能预测，并在使用双向机制后观察到了更好的预测。据我们所知，尚未通过分析云工作负载来报告使用BLSTM进行资源的未来预测。在本文中，我们建议将BLSTM网络生成的预测与LSTM模型和其他分数差异资源预测方法进行分析和比较。

3.双向长短期存储器（LSTM）网络体系结构

常规时间序列预测模型的主要限制是它们只能记住最近的信息。在神经网络模型的训练中，误差函数的局部最小值是通过在梯度的负方向上迭代地采取小步骤来找到的。在反向传播阶段，将梯度信号迭代乘以递归隐藏层的权重。如果权重很小，则梯度会变得很小，以至于学习变得非常缓慢或完全停止工作。这称为消失梯度问题。这使得很难学习数据中的长期依赖性。长短期记忆[13]

（LSTM）已被证明可以有效地处理具有长期依赖性的时间序列。他们通过使恒定的误差流通过称为存储单元的特殊单元的内部状态来克服神经网络中消失的梯度问题[13]。单变量资源预测模型分析所需资源度量的过去历史值中涉及的时间关系，只是为了学习模式并生成将来的样本外预测。多元时间序列预测有助于分析其他功能对所需资源指标的影响。一起研究许多相关变量，而不只是研究一个变量，有助于通过分析除单个时间序列之外的其他时间序列特征的变化来更好地理解[14]。在这项工作中，我们探索多元LSTM，其中时间序列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。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）（5）其中，f1和f2是两个函数。建模多元数据的主要挑战在于捕获每个特征内的时间变化以及每个时间步长上不同特征之间的相关性。

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。在两个步骤中定义新状态，其中输入门确定要更新的值，并使用tanh层获得新候选值的向量。然后，将单元状态ct设为：ct =ftct􀀀1+ it〜ct（7）其中ct􀀀1表示时间t􀀀1时的先前单元状态，它是输入门的输出，〜ct是新的候选值给出为：〜ct = tanh（Wxcxt +Whcht􀀀1+ bc）（8）其中Wxc和Whc分别代表具有输入层和先前隐藏层的单元状态的权重。 bc代表相应的偏差。单元的输入门定义为：it = sig（Wxixt +Whiht􀀀1+ bi）（9）其中Wxi和Whi是与输入关联的权重以及输入门的上一个隐藏层。 bi表示输入门的偏置。单元的输出门使用先前的隐藏层和当前输入来计算单元的输出。输出门的结果用于计算时间t，ht处隐藏层的结果。 ot = sig（Wxoxt +Whoht􀀀1+ bo）（10）其中Wxo表示输出门中输入的权重，而Who表示输出门中过去隐藏层的权重。 bo表示门的偏置。隐藏层的输出计算如下：ht = g（ot tanh（ct））（11）其中g（:)是激活函数。输出Yt的计算公式为：Yt = Whyht + by（12）其中Why和by分别代表权重和偏差。 [15]。

单向网络通过仅从过去的输入中学习趋势来预测未来的输出。另一方面，双向网络集成了来自未来以及过去的依赖关系，以学习时变模式。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。组合时间上向前和向后的模型有助于获取捕获不同时间序列模式的更多好处。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。因此，从直觉上讲，它们有望比单向LSTM更好。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 +（13）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。代表偏见。

4.结果与讨论

在本节中，我们介绍为预测云工作负载中的CPU利用率而进行的研究的结果。在这项研究中进行的实验是在Google簇跟踪[1]上进行的。 Google集群跟踪基于约12500台计算机的集群，并提供29天内到达集群的不同作业的运行时信息。该工作负载提供有关不同作业的到达，执行和终止以及其时间戳的数据。此跟踪可以深入了解真实的云环境。在这项工作中，我们分析和预测CPU资源使用率指标。我们抽取了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。我们生成并分析了后续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。

为了可视化数据中是否存在长距离依赖关系，分析了40个延迟的CPU资源使用率的自相关图，如图3所示。时滞与自相关图给出了时间序列的连续观察之间的相关性[7]。我们可以看到，连续观测之间的相关性存在非常缓慢的衰减，表明存在远距离依赖性。由于存在长期依赖关系，传统的时间序列预测模型（ARIMA，Markov模型）不适用于未来的资源使用预测，因为它们只能以指数衰减的方式建模[16]。因此，我们在工作中使用LSTM和BLSTM模型来预测云中的资源使用情况。

在当前的工作中，我们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。为了构建模型，我们分析了模型的无状态和有状态变化。如果模型是无状态的，则在每个序列上重置单元状态。但是，在有状态模型中，所有状态都将传播到下一批。我们将数据分成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。可以观察到，与无状态LSTM模型相比，有状态LSTM在RMSE中产生的误差较小。由于有状态模型可以在批次之间保留内存，因此它们比无状态模型更好地工作。

我们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。尽管我们已经观察到在我们的情况下，三层膜可以达到最佳效果。最多两层，该模型似乎欠拟合，而在三层以上，我们观察到该模型本质上是过度拟合。深度网络是通过在彼此之上堆叠几个隐藏层而形成的。一层的输出作为下一层的输入传递。在这项研究中，隐藏层中LSTM单元的数量从16到128不等。LSTM块使用逻辑S形作为输入。 Keras [17]在后端使用基于Theano的框架[18]来实现模型。模型的训练在NVIDIA Tesla K80 GPU上进行。均方误差用作损失函数，其中误差et表示为：et = PT t = 1（yt ^ yt）2 T（14）其中，yt表示时刻t的实际资源使用值，而yt表示在时刻t预测的资源使用值。我们将这些时间序列预测与基于分数差异的时间序列预测模型进行比较和分析。在分数差分法的情况下，我们使用重新定标的范围和赫斯特指数值来计算分数差分参数（d）的值[7]。不同的资源预测模型。 PRESS，AGILE，ARIMA和NARNN已实现分数差分，以生成未来的多步提前时间序列预测。

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。从表中可以看出，NARNN从基于分数的四个差分模型的集合中生成了最佳预测。长期短期存储网络很好地模拟了数据中的长期依赖关系，其预测的RMSE低于分数差分时间序列预测模型。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。我们分析了在（a）60（b）120和（c）180步之前的资源预测模型的迭代多步预测。从图中可以看出，这些模型非常紧密地捕获了资源使用的趋势。 BLSTM模型利用前向和后向时间序列依赖性中的耦合，因此比LSTM模型产生更好的预测。比较分数差分模型和基于LSTM的模型，可以说基于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。

5.结论与未来工作

预测资源使用情况对于在云中更好地进行调度和负载平衡很有价值。尽管可以使用许多时间序列模型，但用于生成更好的资源使用情况预测的研究从未停止过。在我们的工作中，我们探索了用于资源使用预测的LSTM和BLSTM模型。我们观察到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。提议的LSTM和BLSTM模型的多元扩展比单变量模型产生更好的预测。我们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。将BLSTM和LSTM的预测在CPU资源使用率预测上进行比较，可以说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。将来，我们计划扩展我们的工作，以利用资源预测模型生成的预测来动态调度云中的资源。