文章编号:1001-9081(2017)S2-0089-03

基于 Caffe 的并行绘制系统帧绘制时间预测

丁祝祥[1*],应三丛[2]
(1. 四川大学 计算机学院,成都 610065; 2. 四川大学 集成计算实验室,成都 610065)
(* 通信作者电子邮箱 2286618896@qq.com)

摘 要: 在具有多个绘制节点的并行绘制系统中,负载均衡算法起到至关重要的作用,直接影响到并行绘制系统的工作效率和系统资源的利用率。传统方法利用基于绘制历史所预测下一帧的绘制时间来作为负载均衡算法的负载划分依据,这种方法在突变场景中负载估计准确度较低。针对基于绘制历史算法中的问题,采用深度学习的方法来预测下一帧绘制的时间。该方法通过采用 Caffe 深度学习框架,首先采集并行绘制系统中影响负载的数据,然后输入采集到的数据集训练和测试设计的神经网络模型,最后在预测下一帧绘制时间时,调用训练好的模型实时获得下一帧的绘制时间。通过实验验证,深度学习方法提高了突变场景中下一帧绘制时间预测的准确度,并且预测所需时间满足了并行绘制系统的实时需求。该方法在预测准确度和预测耗时方面都达到的并行绘制系统负载平衡能够接受的范围。

关键词: 帧绘制时间预测; Caffe; 负载均衡; 并行绘制系统
中图分类号: TP391.41 **文献标志码:** A

Rendering time prediction in parallel rendering system based on Caffe

DING Zhuxiang[1*], YING Sancong[2]
(1. *School of Computer Science, Sichuan University, Chengdu Sichuan* 610065, *China*;
2. *Integrated Computing Technology Laboratory, Sichuan University, Chengdu Sichuan* 610065, *China*)

Abstract: Load balance algorithm acts as a significant role in parallel rendering system with multi-rendering nodes, which affects the work efficiency of parallel rendering system and the utilization rate of computing source directly. The ability of load balance depends on the rendering time of next frame's prediction algorithm straightly. The old method based on the rendering time of forward frame to predict the rendering time of next frame, has a low accuracy of prediction in sudden change scene. In order to solve the problem caused by pre-post frame prediction algorithm, a method of deep learning was put forward to predict the rendering time of next frame in parallel rendering system. First, the data which affect the work load of parallel rendering system were gathered, and input into the deep learning framework to train and test the designed neural network model. Finally, the trained model was used to get the rendering time of next frame in real-time. The experimental results show that the approach of deep learning improves the accuracy of predicting the rendering time of next frame in sudden change scene, and meets the needs of real-time parallel rendering. The method satisfies the requirements of parallel rendering system both in the accuracy of prediction and the prediction time.

Key words: rendering time prediction; Caffe; load balance; parallel rendering system

并行绘制系统的动态负载平衡方法[1],基于时间反馈来平衡负载。其下一帧绘制时间预测方法是前后帧相关性预测算法,该算法是基于前后帧相关性而采用计算上一帧绘制时间作为下一帧的参考绘制时间[2-3]。但在突变场景中前后帧绘制任务量突变时,打破了绘制节点前后帧绘制时间的相关性,容易造成负载任务各绘制节点划分的不合理,从而影响整个并行绘制系统的工作效率。使用 K-Dimensional(KD)树划分数据的方法[4],在多个绘制节点时也会增加调整 KD 树和遍历的时间开销。Regression forest[5]的方法也类似 KD 树,利用建树的方法解决高维数据回归问题,这种方法同样存在建树和划分节点时间开销较大的问题。采用 KD 树和 Regression forest 方法负载估计是人为选取绘制算法部分因素作为节点划分的依据,需要算法设计人员具有较强的图形学算法理论分析能力,这也就限制了负载估计算法的泛化能力。由于影响绘制时间的因素较多,如:视点位置、光源位置等,使用具体的函数量化绘制的时间代价往往不现实的,而且影响绘制时间的主次因素也因绘制算法不同而异。通过分析负载估计影响因素的复杂性,负载估计属于高维函数拟合问题。人工神经网络在高维函数拟合上表现出了其强大的能力,尤其是深度学习和开源工具方面的研究进展,使得神经网络能更好地解决并行绘制负载估计问题。

1 研究背景与研究现状

随着大批企业和科研机构投入到人工智能和深度学习研究和应用中,大规模训练数据较易获得以及高性能计算硬件 Graphics Processing Unit(GPU)加速的广泛应用,深度学习得到了快速发展。深度学习和神经网络工具被运用到各个领域,并取得很好的效益。深度学习在函数逼近方面的运用,预

收稿日期:2016-12-19;修回日期:2017-02-27。
基金项目:国家 863 计划项目(2015AA016405);四川省科技厅科技支撑项目(2016GZ0097)。
作者简介:丁祝祥(1992—),男,广东韶关人,硕士研究生,CCF 会员,主要研究方向:集成计算、深度学习; 应三丛(1975—),男,四川简阳人,副教授,博士,主要研究方向:计算机应用。

示将深度学习工具用于并行绘制系统下一帧绘制时间的预测当中具有可行性。帧绘制时间受到包括相机位置、光源位置、三角面片数目等高维数据的影响，是一个高维函数的拟合问题，传统的数学方法处理起来比较困难，深度学习工具可以很好地解决这个问题。并行绘制对于实时性要求的苛刻，使得帧绘制时间预测要达到实时，即远小于大部分显示器 16 ms 的屏幕刷新时间。在比较了现有流行的深度学习开源框架后，选择 Caffe [4] 作为本文的并行绘制系统实时帧绘制时间的预测工具。Caffe 是由美国加州伯克利大学视觉与学习中心开发，开源后获得开源社区的支持和进一步扩展；其由 C ++ 实现，具有较高效率，而且支持 GPU 硬件加速，可以在较短的时间内训练得到较优的网络模型。在预测时，可以调用训练好的模型快速地获得预测的结果，作为负载均衡任务划分的依据。当预测绘制节点下一帧绘制时间大于阈值时，就将绘制任务划分到其他绘制节点或者增加绘制机器，平衡负载。该方法预测耗时较低，对并行绘制系统的工作效率的影响较小。

本文设计的深度学习网络模型是很深的神经网络，由简单的非线性模块构成，可以处理非常复杂的模型。神经网络通过学习过程从输入的数据中获得知识，并将其以连接权值参数的形式存储下来。

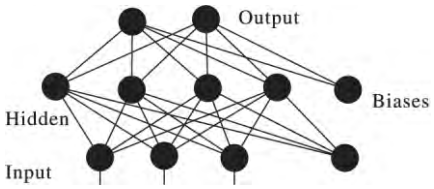

图 1 多层前馈神经网络

如图 1 所示为多层前馈式神经网络 [4] 模型，具有 1 个以上的隐藏层，具备高阶计算能力，可以用于处理高阶函数拟合问题。

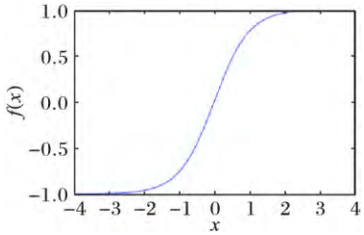

图 2 激励函数 tanh 图像

神经元采用非线性的激励函数双曲反正切函数 tanh 值域［－1，1］，函数图像如图 2，可以解决非线性模型问题。神经网络的能力与完成学习过程的学习算法息息相关，学习算法定义一系列修改连接权值的方法。卷积层采用局部互联，减少了参数的数量。核内共享权值，本文中每个神经元通过 1* 3 的卷积核提取局部特征，下采样层保证互联区域的不变性。全连接层是通常的多层前馈式网络，层间误差逆传播采用 BP 算法。利用 Caffe 深度学习框架处理高维函数回归问题，采用有监督学习训练方法离线训练网络模型。有监督学习通过误差函数计算预测值与真实值的误差，并将该误差逆传播，不断修改网络连接权值，最后使得整体误差最小，达到网络收敛。加速随机梯度下降算法，在有限的训练集条件下，能加快网络收敛，并使得网络不易陷入局部最优解。

2 实现方法

本文实验流程如图 3 所示。

实验所需的训练集、验证集和测试集数据均获取自实验室的并行绘制系统 PR-2015，实验所用软硬件设备为：Intel i5-3450 4 cores CPU，GTX1070 GPU，Windows7，VS2013。

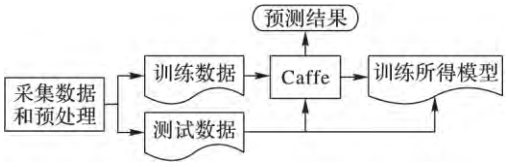

图 3 实验流程

2.1 数据采集和预处理

运行并行绘制系统，通过操作鼠标和键盘在绘制场景中选择一条漫游路径，保存路径位置信息。再次运行绘制程序读取保存的路径信息重新漫游，在这个过程中记录路径中相应的相机位置、光源位置等影响帧绘制时间的信息，共 309 维数据。计算并记录对应每帧的绘制时间，单位毫秒。为了减少计算机其他程序的干扰，每帧绘制 10 次，取 10 次绘制时间的中值作为该帧的绘制时间代价。通过上述方法获得数据集 11 000 条。

将采集到的数据都归一化到［－1，1］，原因是数据集通过归一化预处理操作后，能提高模型训练速度和神经网络收敛速度。将归一化的数据混合后，按 9∶1∶1 的比例划分得到实验所需的训练、验证和测试数据集，这样划分的目的是使训练样本数据尽可能多。将数据集转换为 Caffe 可以读取的 hdf5 数据格式文件。

2.2 Caffe 网络模型设计

对应流程图中的 Caffe 处理框，Caffe 读取设计好的网络模型文件，训练网络。如图 4 为实验所设计的神经网络模型的卷积层部分，便于提取特征。如图 5 所示，实验设计的神经网络模型，采用多个全连接层提高了神经网络的非线性划分能力。

通过分析实验的输入数据：309 维数据，1 维标签值，且数据具有三维一组的特征，即三维坐标数值组。在卷积层添加 1* 3 尺寸卷积核提取特征，通过最大值下采样的池化层使得数据维度降为 103 维。池化层后是 6 个全连接层，分别具有 50、40、30、20、10、1 个神经元。

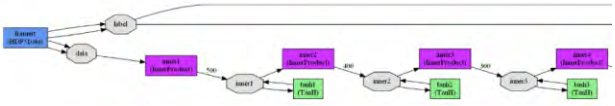

图 4 神经网络模型卷积层部分

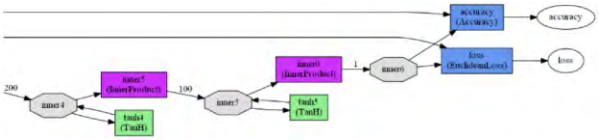

图 5 神经网络模型全连接层部分

损失函数采用欧氏距离作为预测值与真实标签值的误差计算，将计算误差逆向传播。采用式(1)反正切 Tanh 函数作为全连接层的激励函数。

$$f(x) = (\exp(2x) - 1) / (\exp(2x) + 1) \tag{1}$$

网络超参数设置，通过不断调整参数，使得预测结果更优，基本学习率设为 0.001，针对训练集数据具有较高的维度，网络设置较低的基本学习率，防止学习不充分。学习率衰减策略采用式(2)的 inv [4]：

$$base\_lr * (1 + gamma * iter)^{-power} \tag{2}$$

其中：*gamma* 为 0.000 1；*iter* 为 145；*power* 为 0.75。采用 inv 学习率衰减策略，在训练数据相对较少的情况下，能够使网络学习更加充分。学习算法采用加速梯度下降算法

NESTEROV [8-9] 使得网络快速收敛到全局最小。离线训练网络模型时,设置 GPU 加速模式,加速网络的训练和测试 [10-11]。

3 实验结果与分析

将采集到的数据集,按不同的混合方式,得到 1 ~ 5 组数据,每组数据分别离线训练循环 5 000 次后得到 Caffe 网络模型。输入测试集测试网络模型预测精确度,最后获得测试准确度。相同场景下采用前后帧相关性预测算法,获得相同测试集下的测试准确度,结果对比如图 6 所示。

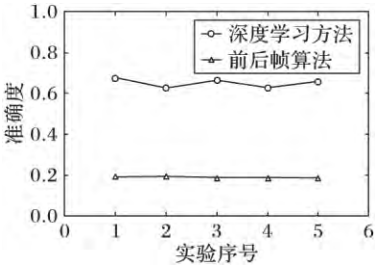

图 6　预测准确度对比

从图 6 中可以看到,在实验设计的突变场景中,1 ~ 5 号数据组的实验结果都是深度学习方法预测的准确度高于前后帧相关性预测算法。实验结果说明深度学习方法预测的准确度显著优于前后帧预测方法,分析原因在于: 在突变场景中前后帧绘制时间上的相关性就被打破了,这样造成使用前后帧预测方法预测准确度低,平均为 18%。相同的 5 组数据,使用深度学习方法,准确度平均为 65%,波动为 + -5%,说明深度学习方法比较稳定。这也说明实验所设计的 Caffe 神经网络模型和设置的超参数使得神经网络稳定收敛了。

在 GPU 加速模式下 [12],网络读取并测试 5 条数据所需时间如图 7 中所示。

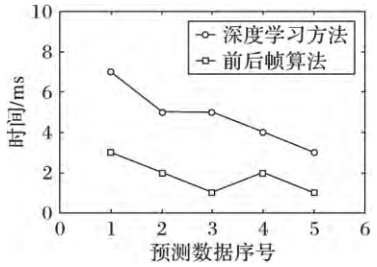

图 7　预测耗时对比

神经网络方法预测耗时平均为 5 ms,前后帧预测方法耗时平均为 1.5 ms。这是由于前后帧方法只需要机械地从变量中读取前一帧绘制时间,而不需要计算,但在神经网络方法中计算是不可避免的。虽然神经网络方法耗时都高于前后帧方法,但从图 7 中的数据可以看出深度学习方法预测每条数据花费的时间都达到了并行绘制系统实时绘制的要求,即远小于 16 ms。在整个系统中,并行绘制系统通过接口调用 Caffe 获得实时预测的下一帧绘制时间作为负载均衡的任务划分依据,动态调整负载,达到负载平衡。

基于 Caffe 预测并行绘制系统下一帧绘制时间的方法,预测准确度上较高,并且预测耗时在可接受的范围内,可以很好地解决在突变场景中,并行绘制系统对下一帧绘制时间预测困难的问题。

4 结语

在并行绘制系统中,利用深度学习工具强大的模式学习能力来预测并行绘制系统下一帧绘制的时间,有效地提高了并行绘制系统负载均衡的工作效率,提高了资源利用率,进而提高并行绘制系统的实时绘制能力。与传统的前后帧预测算法相比,深度学习方法采用从数据中学习的方法,有效地提高了在突变场景中预测下一帧绘制时间的准确度。深度学习中的卷积层提取出了训练数据的特征,池化层下采样减少神经元的数目,减少了训练所耗费的时间,更快地获得训练好的模型。实际使用的过程中采用 GPU 硬件加速,极大地提高了计算速度,预测耗时在并行绘制系统实时绘制要求的范围内。深度学习工具给并行绘制系统的绘制时间预测提供了新的解决方法,并且效果优于传统的方法。

参考文献:

[1] PENG H, XIONG H, SHI J. Parallel-SG: research of parallel graphics rendering system on PC-Cluster[C]// Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications. New York: ACM, 2006: 27 - 33.

[2] 付讯,杨红雨,叶庆. 基于 InfiniBand 的集群分布式并行绘制系统设计[J]. 四川大学学报(自然科学版),2015,52(1):39 - 44.

[3] 彭浩宇. 基于 PC 集群机的并行图形绘制系统研究[D]. 杭州: 浙江大学,2006:2 - 3.

[4] 孔明明. 基于 GPU 集群的并行体绘制[D]. 杭州: 浙江大学,2007:7 - 8.

[5] BRAEUCHLE C, RUENZ J, FLEHMIG F, et al. Situation analysis and decision making for active pedestrian protection using Bayesian networks[EB/OL]. [2017 - 12 - 19]. https://mediatum.ub.tum.de/doc/1187195/1187195.pdf.

[6] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]// Proceedings of the ACM International Conference on Multimedia. New York: ACM, 2014:675 - 678.

[7] ROBERT T F, KING A H, TU X, et al. Multi-contingency transient stability-constrained optimal power flow using multilayer feed forward neural networks[C]// Proceedings of the 2016 IEEE Canadian Conference on Electrical and Computer Engineering. Vancouver, Canada: IEEE, 2016:1 - 6.

[8] JAN H, MOHAMED O, RODRIGO B, et al. Taking a deeper look at pedestrians[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 4073 - 4082.

[9] PARK S, CHOI S. Online multi-label learning with accelerated non-smooth stochastic gradient descent[C]// Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013:3322 - 3326.

[10] RUI X, FU L, CHOI K, et al. Evaluation of convergence speed of a modified Nesterov gradient method for CT reconstruction[C]// Proceedings of the 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference. Piscataway: IEEE, 2012: 3667 - 3670.

[11] BOTTLESON J, KIM S, JEF, et al. clCaffe: openCL accelerated Caffe for convolutional neural networks[C]// Proceedings of the 2016 IEEE International Parallel and Distributed Processing Symposium Workshops. Piscataway: IEEE, 2016.:50 - 57.

[12] LI D, CHEN X, BECCHI M et al. Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs[C]// Proceedings of the 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom). Washington, DC: IEEE Computer Society, 2016: 477 - 484.