**基于Torch平台的神经网络压缩**

**研究与应用**

**需求文档**

Version 2.3

小组成员：

陈伟民

付强

曹进

李恬霖

**版本变更记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本 | 变更时间 | 修改人 | 审核人 | 备注 |
| 1.0 | 20170324 | 陈伟民 | 付强 曹进  李恬霖 | 初稿 |
| 1.1 | 20170331 | 李恬霖 | 付强 曹进  陈伟民 | 二稿 |
| 1.2 | 20170331 | 曹进 | 付强 李恬霖 陈伟民 | 三稿 |
| 1.3 | 20170331 | 陈伟民 | 付强 李恬霖 曹进 | 四稿 |
| 1.4 | 20170331 | 付强 | 陈伟民 李恬霖 曹进 | 五稿 |
| 1.5 | 20170407 | 陈伟民 | 付强 李恬霖 曹进 | 增加业务需求与扩展需求 |
| 1.6 | 20170419 | 曹进 | 付强 李恬霖 陈伟民 | 调整RUCM |
| 2.0 | 20170420 | 陈伟民 付强 李恬霖 曹进 | 陈伟民 付强 李恬霖 曹进 | 根据评审意见进行了修改和完善 |
| 2.1 | 20170426 | 付强 | 陈伟民 李恬霖 曹进 | 根据评审意见修改文章格式，添加参考资料 |
| 2.2 | 20170504 |  |  | 根据刘超老师评审意见修改和完善文档内容 |
| 2.3 | 20170608 |  |  | 根据刘超老师评审意见修改和完善文档内容 |

**目 录**

[1 前言 5](#_Toc1826)

[1.1 目的 5](#_Toc30769)

[1.2 系统概述 5](#_Toc31982)

[1.3 文档概述 5](#_Toc5163)

[1.4 术语和缩略语 6](#_Toc5769)

[2 引用文档 6](#_Toc19933)

[3 需求分析 7](#_Toc26757)

[3.1 业务需求分析 7](#_Toc8910)

[3.1.1 初学者 7](#_Toc12453)

[3.1.2 科研人员 7](#_Toc32448)

[3.1.3 深度学习领域 7](#_Toc25722)

[3.2 功能性需求分析 7](#_Toc24149)

[3.2.1 丰富的工具包 8](#_Toc24091)

[3.2.2 模块化搭建神经网络 8](#_Toc2938)

[3.3 扩展需求 9](#_Toc24381)

[3.3.1 修改卷积层RUCM 10](#_Toc27264)

[3.3.2 修改线性层RUCM 11](#_Toc9855)

[3.4 非功能性需求分析 12](#_Toc17627)

[3.4.1 高效性 12](#_Toc28584)

[3.4.2 用户友好性 12](#_Toc9937)

[3.4.3 可修改性 12](#_Toc16155)

[3.4.4 鲁棒性 13](#_Toc20382)

[3.5 需求识别 13](#_Toc6625)

[3.5.1 载入数据 14](#_Toc28643)

[3.5.2 定义神经网络模型 14](#_Toc5868)

[3.5.3 训练网络 16](#_Toc4180)

[3.5.4 测试网络 16](#_Toc23425)

[3.6 RUCM模型 16](#_Toc2813)

[3.6.1 文件数据读取 17](#_Toc25582)

[3.6.2 导入工具包 18](#_Toc5022)

[3.6.3 搭建MLP神经网络 19](#_Toc17142)

[3.6.4 搭建卷积神经网络 20](#_Toc28647)

[3.6.5 训练神经网络 21](#_Toc5530)

[3.6.6 测试神经网络 22](#_Toc18079)

[4 运行要求 23](#_Toc5670)

[4.1 硬件要求 23](#_Toc25852)

[4.2 软件要求 23](#_Toc20652)

# 前言

## 目的

为了便于协调组内成员进行后期的工作，对项目进行跟踪和监控，对任务的进度进行安排与调控，故对后期工作进行计划。提出需求，指导后续工作。

## 系统概述

Torch是一个用Lua编写的支持机器学习算法的计算框架。其中的一些版本被Facebook、Twitter这样的大型科技公司使用，为内部团队专门化其深度学习平台。Lua是一种在上世纪90年代早期在巴西开发出来的多范式的脚本语言。

Torch的目标是在建立科学算法的同时，要有最大的灵活性和速度，而这一过程非常简单。Torch拥有一个大社区驱动包的生态系统，涉及机器学习、计算机视觉、信号处理、并行处理、图像、视频、音频和网络等，并建立在Lua社区基础之上。

Torch的核心是流行的神经网络和优化库，它们易于使用，同时在实现复杂的神经网络拓扑结构时具有最大的灵活性。可以建立任意的神经网络图，并在CPUs和GPUs上有效地并行化。

## 文档概述

文档用途：本文档主要介绍Torch系统需求及规格说明，内容如下：

⮚以用例图的形式给出 Torch系统功能需求的分解结构，并对用例模型中的参与者和用例进行详细的描述，其中主要包括软件系统的用例模型、系统的核心流程等；

⮚使用 RUCM 模板对用例图中的每个用例进行详细描述；

⮚描述了与Torch系统运行相关的硬件环境的一些要求；

⮚描述了与Torch系统运行相关的软件环境的要求；

## 术语和缩略语

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 术语 | 英文 | 说明 |
| 1 | UCM | UCM | 用例建模 |
| 2 | RUCM | RUCM | 限制性用例模型 |
| 3 | Torch | Torch | 一个用 Lua 编写的支持机器学习算法的计算框架，并拥有一个大社区驱动包的生态系统 |
| 4 | LuaJIT | LuaJIT | LuaJIT即采用C语言写的Lua代码的解释器 |
| 5 | Lua | Lua | Lua是一个小巧的脚本语言。由标准C编写而成，几乎在所有操作系统和平台上都可以编译，运行 |
| 6 | nn | nn | Torch中所包含的关于神经网络的一个包，主要提供了搭建神经网络的各种方法 |
| 7 | CUDA | CUDA | CUDA是一种由NVIDIA推出的通用并行计算架构，该架构使GPU能够解决复杂的计算问题 |
| 8 | Tensor | Tensor | 张量 |
| 9 | 容器 | Container | Torch中的一种模式，用于向其中添加层来构建复杂的神经网络 |
| 10 | 卷积层 | Convolution layer | 用一个过滤器对图像进行采样并处理 |
| 11 | 全连接层 | Linear layer | n-1层的任意一个节点，都和第n层所有节点有连接。即第n层的每个节点在进行计算的时候，激活函数的输入是n-1层所有节点的加权 |
| 12 | 神经网络 | Artificial Neural Network | 从信息处理角度对人脑神经元网络进行抽象， 建立某种简单模型，按不同的连接方式组成不同的网络。 |
| 13 | 深度学习 | Deep Learning | 深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。 |
| 14 | 深度压缩 | Deep Compress | 降低大型神经网络其存储和计算消耗，使得其甚至可以在移动设备上得以运行。 |

图表 1‑1 术语和缩略语

# 引用文档

[1] Torch官方关于nn包的文档[OL].https://github.com/torch/nn/tree/master/doc

[2] Han S, Mao H, Dally W J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding[J]. Fiber, 2015.

# 需求分析

## 业务需求分析

### 初学者

对于深度学习初学者（例如学生），需要应用抽象程度较高的编程工具，屏蔽底层细节，注重于实现简单的神经网络进行实验。因此不需要十分复杂的平台，只需提供基本的神经网络组件即可。

### 科研人员

作为一个深度学习工作者，通常需要即时、简便地构造一个神经网络（卷积神经网络、循环神经网络甚至更复杂的网络结构），另外用包装好的工具包（常见的损失函数和优化算法）对神经网络进行训练，并对网络进行测试，当网络结构过于庞大，需要迁移至GPU训练时，迁移过程需要尽可能简单，甚至不需要了解基于CUDA编程的细节与方法。

另外，工具包应当具有灵活性，源代码满足易读、易更改、易扩展等要求。开源平台可以实现使用者即时的扩展功能，并按自己的需求对模块进行修改。

### 深度学习领域

随着神经网络相关研究的发展，深度网络等复杂的网络结构对于解决某些实际问题往往有着更优秀的表现。搭建深度网络是一个复杂又耗时的工程，同时，巨大的网络结构造成在训练过程中需要调整大量的参数，这对硬件的存储和计算能力都提出了很高的要求。如果这些问题得以解决，势必将大大推进深度学习领域的研究。甚至在解决存储和计算问题后，深度网络可以迁移至移动端进行训练。

## 功能性需求分析

Torch是一个广泛支持机器学习算法的科学计算框架。易于使用且高效，主要得益于一个简单的和快速的脚本语言Lua，和底层的C / CUDA实现。

### 丰富的工具包

Torch7提供了拥有十分详尽接口的库，Torch有一个在机器学习领域大型生态社区驱动库包，包括计算机视觉软件包、信号处理、并行处理、图像、视频、音频和网络等，是基于Lua社区建立的。

Torch可以看作是由Lua编写的一个package，本身的后台都是基于C和CUDA写的，并提供Lua接口。一般基于torch的简单开发，可以在Lua的级别下完成。如果考虑到功能的复杂性或者package的效率问题，则需要用C和CUDA来编写后台。

神经网络是Torch中的一个包，如(nn/cunn)，其核心是各个层（torch中称为Module），主要的代码包括Lua、C、CUDA三部分。若一个层所需运算可直接通过Tensor操作来完成，则仅一个Lua文件即可，如nn.Linear。若需要C/CUDA实现，则还需要写C/CUDA代码，如nn.SpatialConvolutionMM。

### 模块化搭建神经网络

图计算框架都有一个限制，就是需要用户把所有的计算全部都表示成一张图来高效运行。这么做不论是在逻辑上，调试上以及和运行环境的交互(python)上面都是有一定劣势。

而Torch采取了支持用户把计算拆分成多步来做的方式。用户可以直接利用Lua来选择下一步执行什么。用户可以比较简单地对计算进行模块分割，并且根据比如输入长度的不同来直接动态改变需要运行哪一个步骤。Torch为代表的过程式计算更加灵活。

Torch7和nn类库拥有清晰的设计和模块化的接口。Torch对卷积网络的支持非常好。Torch通过时域卷积的本地接口使得它的使用非常直观。Torch通过很多非官方的扩展支持大量的递归神经网络。Torch本质上是以图层的方式定义网络的，在Torch中定义新图层非常容易。

作为一个深度学习的框架，Torch允许自由地实现对已有模块逻辑复杂的调用，允许依据需要使用已有的模型（模型的参数自己训练得到），也可以在已有模型的基础上增加自己的layer，或者在顶端选择自己需要的分类器，Torch构建的是一个生态系统，安装新的模型实现模块只需要使用luarocks install package。

在需要编写系统没有提供的层时，如果所有运算均能通过Tensor自带的操作来完成，就只需要写一个Lua文件，通过require来使用。而使用Tensor操作无法完成或效率太低时，就需要使用C和CUDA来实现核心算法，通过Lua来调用。

## 扩展需求

神经网络功能强大。但是，其巨大的存储和计算代价也使得其实用性特别是在移动设备上的应用受到了很大限制。降低大型神经网络其存储和计算消耗，使得其甚至可以在移动设备上得以运行，即要实现“深度压缩”。

利用Torch平台提供的模块化实现，我们可以扩展自己的神经网络层（全连接层、卷积层），修改内部的权重计算方式，去掉不必要的连接，进而减少参数的数量或存储，缩短网络训练时间和存储消耗，但仍保证网络训练产生模型的准确性。

详细设计请见《基于Torch平台的扩展：神经网络压缩算法研究》。

### 修改卷积层RUCM

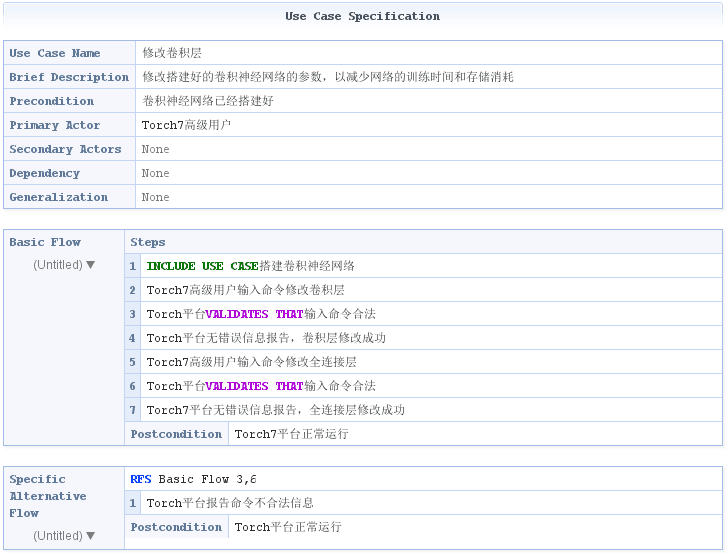




图3-1 修改卷积层RUCM

我们要扩展Torch平台，选择扩展卷积神经网络和MLP神经网络，并以Torch7高级用户的角色参与。要扩展卷积神经网络，选择修改卷积层以减少卷积神经网络的内部参数，缩短卷积神经网络的训练时间和存储消耗。首先，我们要作为Torch7用户，使用Lua语言来对Torch平台进行扩展。先建立卷积神经网络，然后输入命令修改卷积层和全连接层；每次输入命令后，Torch平台会验证输入的命令是否合法；若Torch 7平台对Torch用户的输入命令不报告任何错误信息，则说明Torch用户输入的命令运行成功。当Torch 7高级用户在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

### 修改线性层RUCM

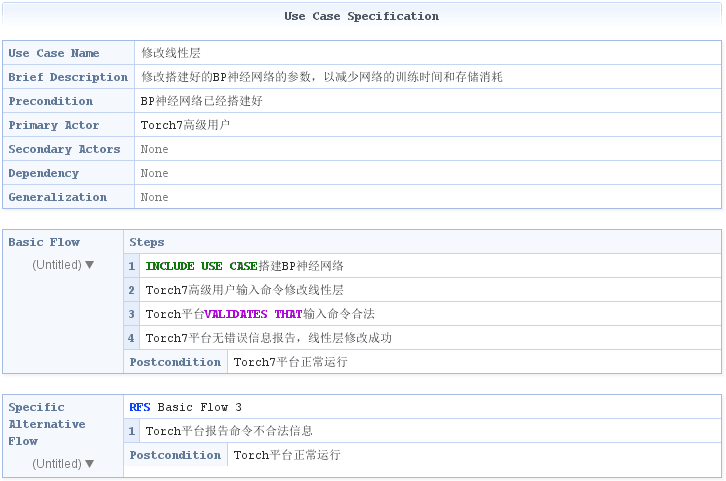




图3-2 修改线性层RUCM

我们要扩展MLP神经网络，选择修改线性层以减少MLP神经网络的内部参数，缩短MLP神经网络的训练时间和存储消耗。首先，我们要作为Torch7高级用户，使用Lua语言来对Torch平台进行扩展。先建立MLP神经网络，然后输入命令线性层；每次输入命令后，Torch平台会验证输入的命令是否合法；若Torch 7平台对输入的命令不报告任何错误信息，则说明Torch用户输入的命令运行成功。当Torch 7高级用户在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

## 非功能性需求分析

### 高效性

（1）轻量级系统

Torch平台与其他开源深度学习框架相比更加轻量级的系统架构，以保证高效运行。未来的发展目标是希望torch 7可以直接部署到手机上。

（2）高效底层语言支持

Torch应该使用效率高且平台兼容性好的语言编写底层代码，例如C或C++。这样不仅可以进一步实现核心计算单元的优化，也保证了Torch做高性能计算、异构计算、以及应用于嵌入式平台的可能性。

（3）支持并发运行

Torch应该支持多线程、多GPU运行，以保证大规模深度神经网络的高效训练和使用。

### 用户友好性

为了保证深度学习研究者和开发者可以尽快熟悉、使用Torch平台，平台应该具有更简洁易懂的操作界面，更简单易学的编程语法，在各个方面更加注重用户友好性。

### 可修改性

Torch作为一个内置丰富软件工具包的开发平台，为深度学习研究者和开发者提供了基础的深度学习模型框架和方法接口。但是由于其提供的框架和方法过于基础，故需要其具有可修改性，以便通过不断地研究，引入新的算法或技术，以实现平台搭建的深度学习网络或实现的其他功能具有更好地时间和空间效率。

### 鲁棒性

系统应该在异常和危险情况下都能保持健壮的表现和稳定的性能，有健全的容错机制和方法。以导入工具包过程为例，每次输入命令（例如导入xx工具包），Torch平台会验证命令是否合法。若命令不合法（例如工具包导入错误信息或工具包不存在），Torch平台会报告命令不合法信息并使Torch保持正常使用状态。当Torch 7用户（包括Torch7初学者和Torch7高级用户）在Torch平台输入命令遇到软中断，Torch平台将等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

## 需求识别

将业务需求分解为功能性需求和非功能性需求，同时考虑到Torch7用户分为Torch7初学者和Torch7高级用户两种，最终得到的Torch的用例图如图3-3所示。

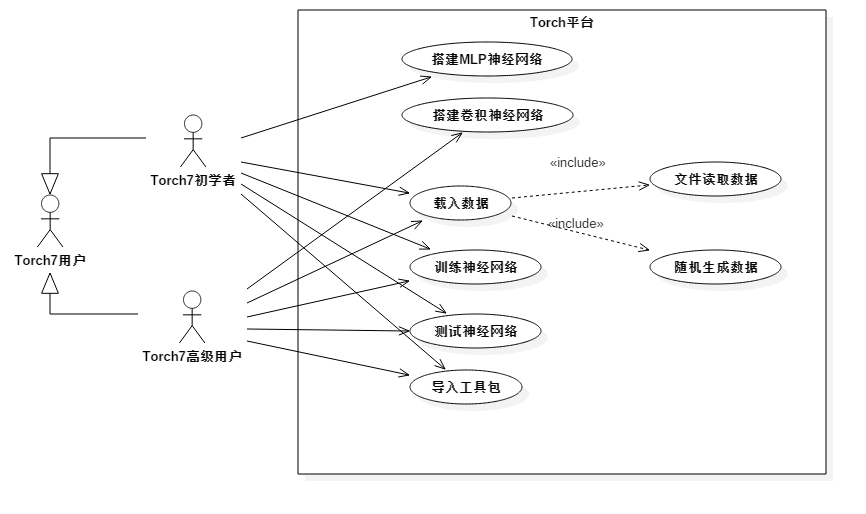


图3‑3 高级Torch用户的用例图

Torch7用户分为两类：Torch7初学者和Torch7高级用户（包括科研人员和深度学习领域研究者）。Torch7初学者在Torch平台上搭建MLP神经网络，Torch7高级用户在Torch平台上搭建卷积神经网络。

### 载入数据

载入数据主要包括两个步骤：

1.首先用户需要引入所需要的包，如cephes、randomkit、nn、cutorch等。Torch是一个工具包的集合，在使用工具前首先要引入工具所在的包，例如torch包提供了最基本的矩阵或表的操作，nn包提供了神经网络组件的实现，cutorch包提供了程序在GPU上运行的基础实现。引入包的操作由require [package]实现；

2.根据不同的需求载入数据，例如可从文件载入，也可使用torch自带的随机方法生成数据。

### 定义神经网络模型

生成模型的过程主要分为以下几个步骤（以建立卷积神经网络模型为例）：

1.引入相关的包。

2.选择容器存放模型，如sequential、concat、parallel、bottle等。容器是建立模型的基础，以神经网络为例，首先要定义神经网络的架构，torch提供了三种基本类型的容器：

Sequential：所有层组成一个序列化的神经网络，第一层连接所有输入。由[module]=nn.Sequential()实现。图3-4是Sequential容器结构的示意图。

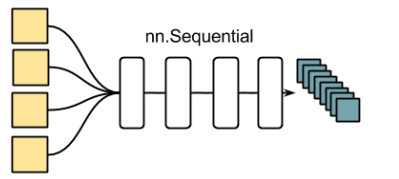


图3‑4 Sequential容器

Parallel：针对每一个输入，可以构建一个序列化的神经网络，最后连接输出的结果。由[module]=nn.Parallel(inputDimension, outputDimension)实现。图3-5是Parallel容器结构的示意图。

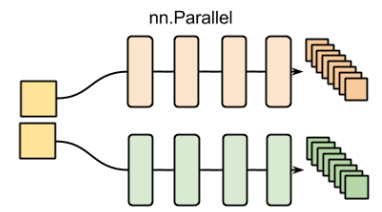


图3‑5 Parallel容器

Concat：构建若干个序列化的神经网络，每个网络的第一层连接所有输入。由[module]=nn.Concat(dim)实现。图3-6是Concat容器结构的示意图。

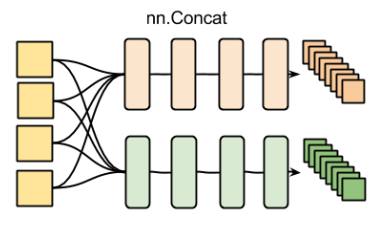


图3‑6 Concat容器

3.添加层。建立好容器后，就可以向容器中添加不同的层，以构建一个卷即神经网络为例，nn包提供了两个基本的层：

SpatialConvolution：一个二维的卷积层，可对输入数据进行卷积处理。由[module]=nn.SpatialConvolution(nInputPlane, nOutputPlane, kW, kH, [dW], [dH], [padW], [padH])实现。其中nInputPlane是输入图像的通道数，nOutputPlane表示卷积层的输出通道数，kW和kH表示卷积核的大小，dW和dH表示卷积核的移动步长，padW和padH表示对输入通道补0的情况。

Linear：线性全连接层，可对输入数据进行线性处理，如y=Ax+b。由[module]=nn.Linear(inputDimension, outputDimension, [bias=true])实现。其中inputDimension和outputDimension分别表示全连接层输入和输出的维度，偏置项默认存在。在向容器中添加层时，可通过[Container]:add(nn.[layer]([params])实现。

4.添加层后，还需要定义隐藏单元的激活函数，如tanh、logsoftmax等。

### 训练网络

训练一个神经网络，包括以下步骤：

1.定义损失函数，例如，MSECriterion用于计算回归问题的损失，ClassNLLCriterion用于计算分类问题的损失。通过[criterion] = nn.MSECriterion()实现。

2.选择训练方法，可以通过自行编写更新权重的算法，也可以调用torch提供的优化方法，例如，随机梯度下降算法等，通过[trainer]=StochasticGradient (module, criterion)实现。

3.训练神经网络，利用定义的损失函数和优化算法训练神经网络，通过trainer:train(dataset)实现。

### 测试网络

测试一个神经网络，可以通过output = [model]:forward(testData)来实现，其中testData是测试数据集。

## RUCM模型

RUCM 即限制性用例建模。它的目标是：

1．使 UCMs 更加可理解并且更精确。

2．从 UCMs 自动生成分析模型。

RUCM 有以下两部分组成：

1．一个用于系统组织 UCSs 的用例模板。

2．限制用户写 UCSs 的一系列规则。

通过 RUCM 模型能够对用例进行规范的描述，接下来将使用 RUCM 模型描述图3-1中的用例。

### 文件数据读取





图3‑7 文件读取数据RUCM

Torch 7用户（包括Torch7初学者和Torch7高级用户）选择从文件读取数据，先导入paths包，然后输入导入命令和文件路径，Torch平台会先验证输入命令是否合法，接着验证文件路径是否合法；若命令和文件路径都合法，Torch平台会将数据加载到内存中，用户再对输入的数据进行预处理。当Torch 7用户（包括Torch7初学者和Torch7高级用户）在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

### 导入工具包





图 3‑8 导入工具包RUCM

Torch 7用户（包括Torch7初学者和Torch7高级用户）需要工具包搭建神经网络时，可以从Torch平台的工具包库中引入想要的工具包；每次输入命令，Torch平台会验证命令是否合法，若命令不合法，Torch平台会报告命令不合法信息。当Torch 7用户（包括Torch7初学者和Torch7高级用户）在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

### 搭建MLP神经网络

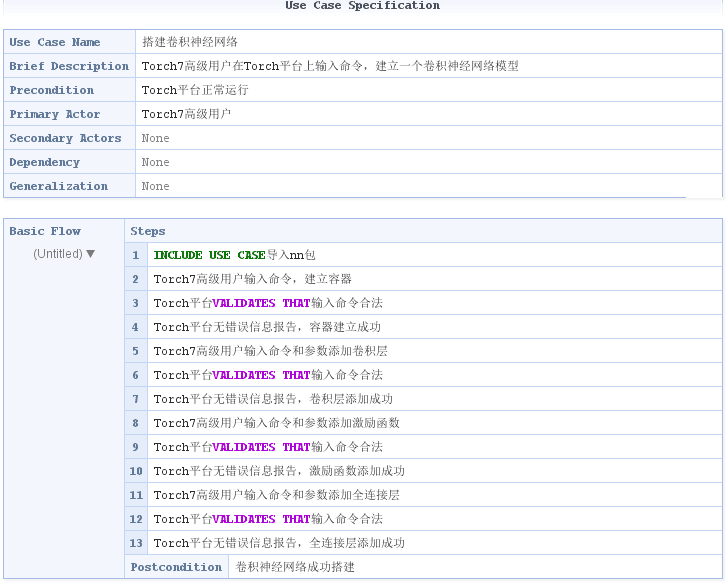


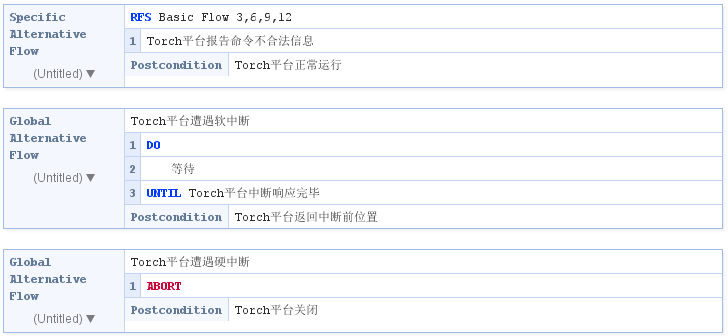


图 3‑9 搭建MLP神经网络RUCM

Torch 7初学者搭建MLP神经网络，要先建立容器，接着在容器中添加线性层；每次输入命令，Torch平台会验证命令是否合法，若命令不合法，Torch平台会报告命令不合法信息。当Torch 7初学者在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

### 搭建卷积神经网络





图表 3‑10 搭建卷积神经网络RUCM

Torch 7高级用户搭建卷积神经网络，要先建立容器，接着在容器中添加卷积层和全连接层，并建立激励函数；每次输入命令，Torch平台会验证命令是否合法，若命令不合法，Torch平台会报告命令不合法信息。当Torch 7高级用户在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

### 训练神经网络





图表 3‑11 训练神经网络RUCM

神经网络搭建好后，Torch 7用户（包括Torch7初学者和Torch7高级用户）选择训练网络，要输入命令选择训练算法，接着输入命令选择损失函数，每次输入命令，Torch平台会验证命令是否合法，若命令不合法，Torch平台会报告命令不合法信息。当Torch 7用户（包括Torch7初学者和Torch7高级用户）在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

### 测试神经网络





图表 3‑13 测试网络RUCM

神经网络搭建并训练好后，Torch 7用户（包括Torch7初学者和Torch7高级用户）会输入命令测试神经网络，每次输入命令，Torch平台会验证命令是否合法，若命令不合法，Torch平台会报告命令不合法信息。Torch平台运行神经网络后会报告测试完成，对于神经网络的测试结果，Torch 7用户（包括Torch7初学者和Torch7高级用户）可以将测试结果导出。当Torch 7用户（包括Torch7初学者和Torch7高级用户）在Torch平台输入命令遇到软中断，Torch平台等待中断响应完毕回到中断前位置，以保证用户不会丢失数据；当Torch平台遇到断电等硬中断时，Torch平台会直接关闭。

# 运行要求

## 硬件要求

* CPU: Intel等主流CPU
* 内存：2G内存及以上
* 硬盘：20G硬盘及以上
* GPU：能够使用cuda（只有G80、G92、G94、G96、GT200、GF100、GF110、GK100、GK104、GK107及以上平台的NVidia显卡）

## 软件要求

* 操作系统：linux（包括ubuntu14.04.以上和OS X10.0以上等）
* 编译环境：luaJit 2.0.0以上版本（安装时包括）

…