Contents

[Contents 1](#_Toc1177704121)

[常用工具 2](#_Toc1846582306)

[命令 2](#_Toc2052655395)

[标注工具 2](#_Toc992817754)

[公开数据集 3](#_Toc2144126482)

[图形化DL训练工具DIGITS 3](#_Toc1696330371)

[硬件及平台 3](#_Toc1627639288)

[GPU 3](#_Toc411613513)

[GPU SDK 5](#_Toc1302158161)

[deploy优化 6](#_Toc1614373641)

[TensorRT 6](#_Toc1649102074)

[基本概念 6](#_Toc1327944384)

[优化方法： 7](#_Toc443369484)

[1. 网络优化 7](#_Toc1372062334)

[Typical development workflow: 7](#_Toc335340344)

[安装 8](#_Toc2037532373)

[方法1: Command Line Wrapper 9](#_Toc1811200453)

[方法2: python接口 9](#_Toc453149681)

[方法2.1: TensorRT Lite （高度抽象层） 9](#_Toc164103273)

[方法2.2: Using TensorRT to Optimize Caffe Models in Python 10](#_Toc93014553)

[方法3： C++接口 （主要用于自定义网络） 11](#_Toc45017978)

[方法4： TensorRT + Caffe 14](#_Toc1297457843)

[模型压缩 16](#_Toc2139728169)

[模型剪枝 （有损） 16](#_Toc1117068454)

[模型量化 （有损） 16](#_Toc326060133)

[Deep learning framework 17](#_Toc1677267924)

[Caffe 19](#_Toc120384474)

[Install 19](#_Toc1088003391)

[tutorial 20](#_Toc574164323)

[Blob storage and communication 20](#_Toc2043257149)

[Layer computation and connections 20](#_Toc800736088)

[Net = layers + layers + ... 25](#_Toc1751868444)

[Model 26](#_Toc1742355807)

[Solver 27](#_Toc705907835)

[可视化.prototxt 28](#_Toc597202550)

[Interfaces--Command Line (train, test, time) 28](#_Toc1738998641)

[Interfaces--Python 30](#_Toc254754559)

[安装 30](#_Toc77358190)

[1. 用代码来生成train.prototxt, test.prototxt, solver.prototxt文件 30](#_Toc3128506)

[2. 剖析模型 32](#_Toc1556912720)

[3. 训练--绘制loss and accuracy曲线 34](#_Toc1691731831)

[4. 测试 34](#_Toc1652230581)

[Interfaces--C++ 35](#_Toc737373456)

[Caffe源代码 35](#_Toc2135101316)

[caffemodel 解析 35](#_Toc876809267)

[caffe 调试 36](#_Toc1072713800)

[Caffe的整体流程图 37](#_Toc2025150041)

[自定义层 42](#_Toc540526072)

[Caffe2 44](#_Toc1525863481)

[Install 44](#_Toc41769666)

[Tutorial 46](#_Toc633540625)

[Keras 53](#_Toc1570881459)

[安装 54](#_Toc1339227510)

[重点概念 56](#_Toc625785146)

[概念 58](#_Toc540466265)

[Keras模型= Sequential模型 / 函数式模型Model 58](#_Toc1665287643)

[张量Tensor 61](#_Toc155569422)

[层Layer 61](#_Toc660850739)

[常用层 62](#_Toc605807387)

[数据预处理 68](#_Toc729733745)

[性能评估 70](#_Toc556624240)

[正则项，约束项 72](#_Toc1406543475)

[回调函数Callbacks 73](#_Toc334118541)

[Keras后端 74](#_Toc151496400)

[workflow 81](#_Toc2112451310)

[Image denoising 81](#_Toc931321092)

[数据预处理 81](#_Toc1890495041)

[tensorflow 81](#_Toc219722221)

常用工具

命令

Monitor GPU resource

$ watch -n 1 nvidia-smi

标注工具

ROI

<https://github.com/tzutalin/labelImg>

sudo apt-get install pyqt4-dev-tools

sudo pip install lxml

$git clone <https://github.com/tzutalin/labelImg.git>

labelImg$ make qt4py2

labelImg$ python labelImg.py

导入图像文件夹 -> Create ROI -> drag mouse to draw ROI -> Ctrl+s

xml 转json

<https://pypi.python.org/pypi/xmljson>

$ pip install xmljson

from xmljson import parker, Parker

from xml.etree.ElementTree import fromstring, fromstringlist

with open(root + '/Case-3-D-6-1.xml', 'r') as fh:

contents = fh.readlines()

anno = parker.data(fromstringlist(contents)) #json数据

print(json.dumps(anno, indent=4))

分动分割

<https://github.com/wkentaro/labelme>

# Ubuntu 14.04

sudo apt-get install python-qt4 pyqt4-dev-tools

sudo pip install labelme # python2 works

公开数据集

Pascal VOC <http://host.robots.ox.ac.uk/pascal/VOC/>

coco <http://cocodataset.org/#home>

kitti <http://www.cvlibs.net/datasets/kitti/eval_object.php>

openimages <https://github.com/openimages/dataset>

图形化DL训练工具DIGITS

<https://github.com/NVIDIA/DIGITS>

深度学习GPU训练系统，支持Caffe, Torch and Tensorflow

硬件及平台

GPU

**Tesla** 人工智能，高性能计算和超大规模数据中心

Tesla V100（16/32GB），P100, P4/P40

**GEFORCE GTX** 游戏体验

**GTX 1080 Ti**， 1080， 1070， 1060, 1050

NVIDIA CUDA® Cores:3584 Boost Clock (MHz): 1582

Memory Speed: 11 Gbps Standard Memory Config:11 GB GDDR5X

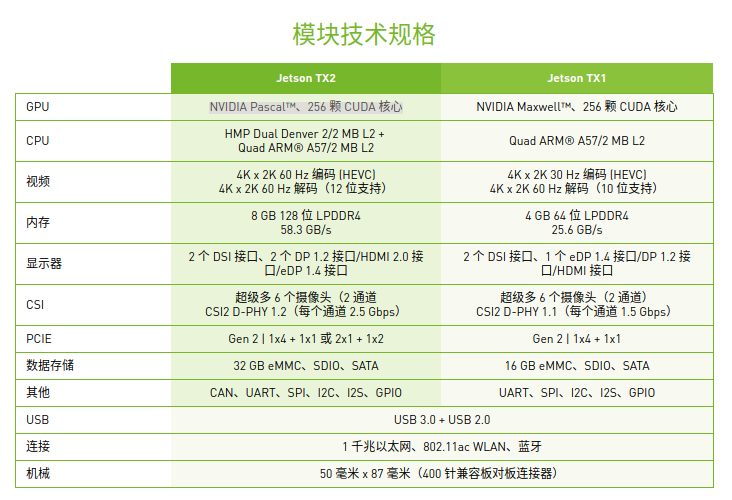
Memory Interface Width:352-bit Memory Bandwidth (GB/sec): 484

Maximum Digital Resolution:7680x4320@60Hz

Maximum GPU Temperature (in C): 91 Graphics Card Power (W):250 W

**JETSON TX2**

性能强大，但外形小巧，节能高效，非常适合机器人、无人机、智能摄像机和便携医疗设备等智能终端设备。



Jetson Development Pack (JetPack)

一体化软件包，捆绑并安装了适用于 NVIDIA Jetson 嵌入式平台的所有开发用软件工具。 JetPack 包括:

* 深度学习: TensorRT、cuDNN、NVIDIA DIGITS™ 工作流程
* 计算机视觉: NVIDIA VisionWorks、OpenCV
* GPU 计算: NVIDIA CUDA、CUDA 库
* 多媒体: ISP 支持、摄像头图像、视频 CODEC
* 同时，它还包括 ROS 兼容性、OpenGL、高级开发者工具等等。
* support for TensorFlow models, up to 15% perf/W improvement for DL applications, out-of-the-box kernel support for Docker, and support for Ubuntu 16.04 on your host PC.

2018京东价格：

Jetson TX2 核心板： 3799元

Jetson TX1 TX2 嵌入式开发套件 TX2套件： 5580

Quadro 视觉计算平台

Tegra 为便携设备提供高性能、低功耗体验

NVIDIA® NVS™ 多显示器商用显卡的标准

架构：

2017 NVIDIA Volta 架构: 如Tesla V100

2016 NVIDIA Pascal 架构： 如GTX 10XX系列， JETSON TX2

2014 NVIDIA Maxwell 架构： 如GTX 9XX系列

2012 NVIDIA Kepler™ 架构： 如GTX 7XX，

GeForce GTX 1080 vs Tesla P100

<https://alisha17.github.io/machine-learning/2017/12/15/benchmarks.html>

Larger mini-batches are more efficient to compute and lead to better convergence in fewer epochs

a larger memory in a GPU and take a larger batch size, donot mean that it will give better validation accuracy

the time taken for training with a larger batch also increases significantly

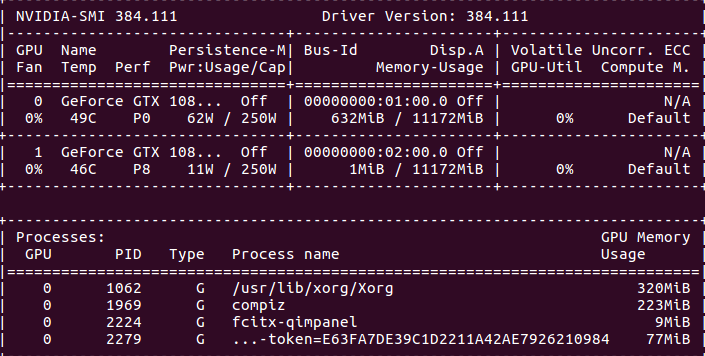
2000 steps and the batch size as 32 was 51 minutes, the time taken for training with same number of steps but with the batch size as 85, was 120 minutes

GPU SDK

TFLOPS： 万亿次浮点指令每秒

1T=1024G， floating point operations per second

$ watch -n 1 nvidia-smi



第一栏的Fan：N/A是风扇转速，从0到100%之间变动，这个速度是计算机期望的风扇转速，实际情况下如果风扇堵转，可能打不到显示的转速。有的设备不会返回转速，因为它不依赖风扇冷却而是通过其他外设保持低温（比如我们实验室的服务器是常年放在空调房间里的）。

第二栏的Temp：是温度，单位摄氏度。

第三栏的Perf：是性能状态，从P0到P12，P0表示最大性能，P12表示状态最小性能。

第四栏下方的Pwr：是能耗，上方的Persistence-M：是持续模式的状态，持续模式虽然耗能大，但是在新的GPU应用启动时，花费的时间更少，这里显示的是off的状态。

第五栏的Bus-Id是涉及GPU总线的东西，domain:bus:device.function

第六栏的Disp.A是Display Active，表示GPU的显示是否初始化。

第五第六栏下方的Memory Usage是显存使用率。

第七栏是浮动的GPU利用率。

第八栏上方是关于ECC的东西。

第八栏下方Compute M是计算模式。

下面一张表示每个进程占用的显存使用率。

显存占用和GPU占用是两个不一样的东西，显卡是由GPU和显存等组成的，显存和GPU的关系有点类似于内存和CPU的关系。我跑caffe代码的时候显存占得少，GPU占得多，师弟跑TensorFlow代码的时候，显存占得多，GPU占得少。

cuda 版本

cat /usr/local/cuda/version.txt

cudnn 版本

cat /usr/local/cuda/include/cudnn.h | grep CUDNN\_MAJOR -A 2

deploy优化

主要指标：Power efficiency and speed of response

TensorRT

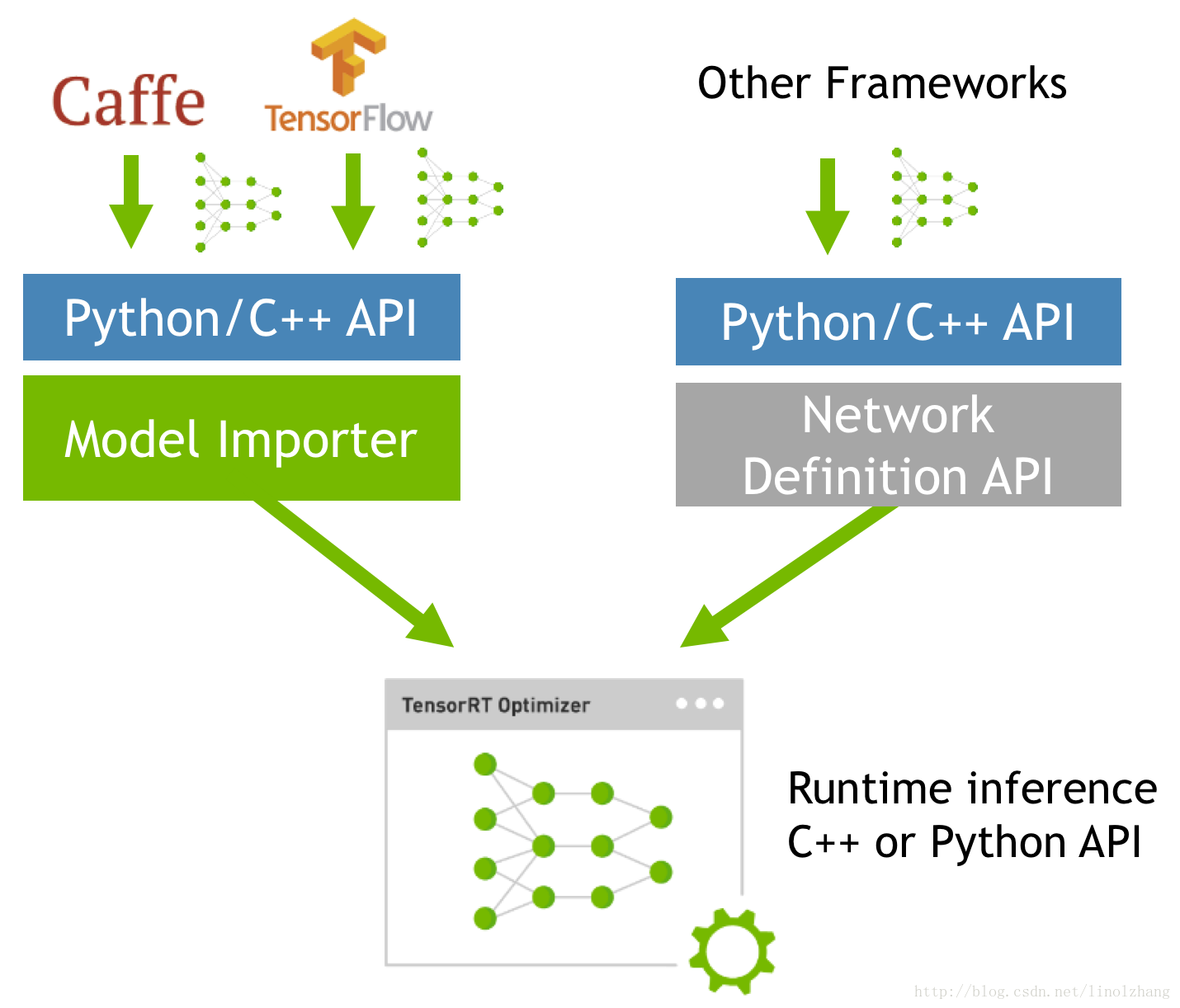
<https://developer.nvidia.com/tensorrt>

<https://blog.csdn.net/fengbingchun/article/details/78469551>

基本概念

TensorRT是为优化生产环境中部署的深度学习模型而创建的库(C++库), 自动优化训练过的神经网络。使用TensorRT，你无需在部署硬件上安装并运行深度学习框架。

TensorRT 4 provides an ONNX parser so you can easily import ONNX models from frameworks such as Caffe 2, Chainer, Microsoft Cognitive Toolkit, MxNet and PyTorch into TensorRT



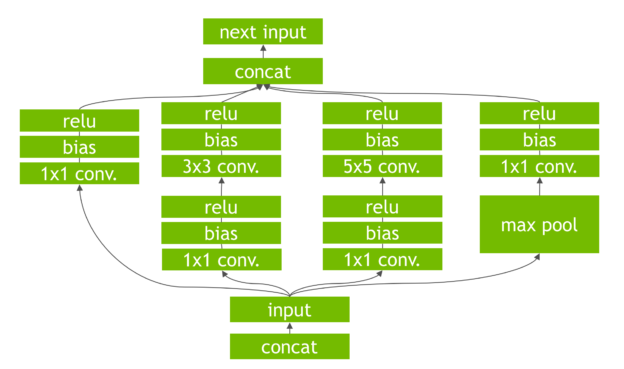
术语

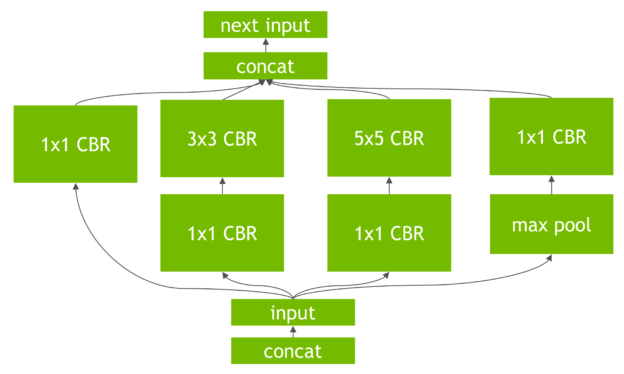
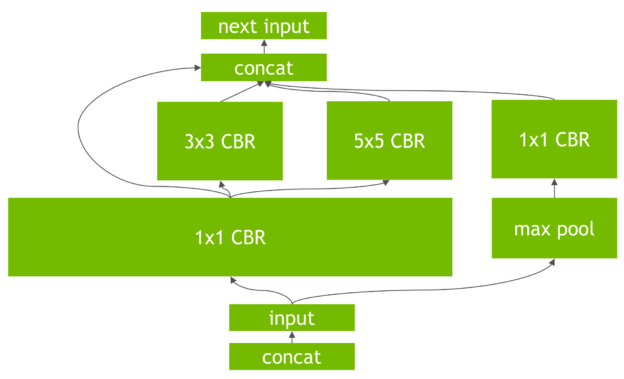
* **ONNX** (Open Neural Network Exchange)
* **UFF (Universal Framework Format )**: is a data format that describes an execution graph for a DNN (Deep Neural Network)
* **PLAN file**: is the serialized data that the runtime engine uses to execute the network

优化方法：

1. 网络优化

* 消除未使用的输出的层以避免不必要的计算；
* 在可能的情况下，convolution、bias和ReLU层被融合以形成单个层，包括垂直层融合和水平层融合。

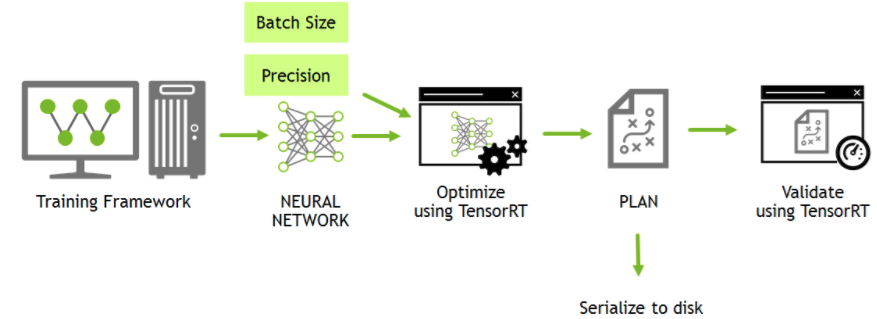


垂直层融合 水平层融合

1. 数据布局可以使用半精度FP16 和INT8 （有损）
2. 卷积的优化，使用Winograd（提升3倍） 等算法或者特定硬件方式实现；
3. cuDNN优化，根据不同的batchsize设置不同的计算模式或者GPU clock；
4. 内存优化,GPU汇编指令,提高GPU利用率(批量处理尽可能并行处理,在cuda中 使用warp对齐,提高GPU命令命中率)
5. CPU和GPU互补：CPU做一部分工作，GPU做一部分工作；

Typical development workflow:



two ways to create a TensorRT network:

1. Create from scratch using the Builder API, or
2. Import an existing NVCaffe, ONNX or Tensorflow network using the Parser API

TensorRT支持的层

Activation (ReLU,tanh,sigmoid), BatchGemm, Concatenation, Constant,

Convolution, Deconvolution, ElementWise, Flatten, FullyConnected, Gather,

LRN, MatrixMultiply, Padding, Pooling (maximum, average), Ragged SoftMax,

Reduce, RNN, RNNv2, Scale, Shuffle, SoftMax, Squeeze, TopK, Unary,

Plugin

The Plugin Layer allows you to integrate custom layer implementations that TensorRT does not natively support.

对于TensorRT 不支持的层，可以先将支持的层跑完，然后将输出作为caffe的输入，用caffe再跑，V1不支持TensorRT 和caffe同时工作，V2支持。

**Some framework models are not supported by TensorRT**. How to do?

* use the Plugin API
* modify the original network to use an equivalent supported model.

The TensorRT API allows developers to import, calibrate, generate and deploy optimized networks. Networks can be imported directly from NVCaffe, or from other frameworks via the UFF or ONNX formats. They may also be created programmatically by instantiating individual layers and setting parameters and weights directly.

A workflow is provided for those looking to implement custom layers for use in Python in tensorrt/examples/custom\_layers.

Build Phase

* elimination of layers whose outputs are not used
* fusion of convolution, bias and ReLU operations
* aggregation of operations with sufficiently similar parameters and the same source tensor (for example, the 1x1 convolutions in GoogleNet v5's inception module)
* merging of concatenation layers by directing layer outputs to the correct eventual destination

Execution Phase

* The runtime executes the optimized engine.
* The engine runs inference tasks using input and output buffers on the GPU.

安装

<https://zhuanlan.zhihu.com/p/35468450>

<https://blog.csdn.net/xll_bit/article/details/78376320>

下载（链接见第一篇blog）

TensorRT 4.0.0.3 for Ubuntu 16.04 and CUDA 8.0 tar package

注意: 版本要求cuda 8 or 9, cudnn 7， 一定要细看doc

添加路径

$sudo gedit ~/.bashrc

export TensorRT\_Root="/home/qzlin/Documents/dl/TensorRT4"

export LD\_LIBRARY\_PATH="$TensorRT\_Root/lib:$LD\_LIBRARY\_PATH"

$source ~/.bashrc

TensorRT\_Root$pip install python/tensorrt\*.whl

$ which tensorrt 测试 tensorRT是否安装成功

若出现fatal error: cuda.h: No such file or directory

注意：不能加sudo, 具体见网上解释关于安装pycuda

$pip install uff-0.1.0rc0-py2.py3-none-any.whl

$ which convert-to-uff 测试convert-to-uff是否安装成功

1. Caffe -> TensorRT

方法1: Command Line Wrapper

giexec --deploy=mnist.prototxt

--model=mnist.caffemodel

--output=prob

方法2: python接口

**方法2.1: TensorRT Lite （高度抽象层）**

the Lite engine creates the logger, TensorRT engine, runtime and context for you, then allocates the GPU memory for the engine as well.

def build\_engine(engine\_file):

MODEL\_PROTOTXT = './data/mnist/mnist.prototxt'

CAFFE\_MODEL = './data/mnist/mnist.caffemodel'

**engine = trt.lite.Engine**(framework="caffe",

deployfile=MODEL\_PROTOTXT,

modelfile=CAFFE\_MODEL,

max\_batch\_size=10,

input\_nodes={"data": (1, 28, 28)},

output\_nodes=['prob'])

**engine.save(engine\_file)**

def infer(engine\_file):

**engine = trt.lite.Engine(PLAN=engine\_file)**

for i in range(10):

img, rand\_file = generate\_data()

img = np.reshape(img, (1, 1, 28, 28))

**output = engine.infer(img)[0]**

print("Test Case: " + str(rand\_file))

print ("Prediction: " + str(np.argmax(output)))

同样支持caffe, tensorflow, and other framework(要自己构建)

可以包含输入预处理和输出后处理

def normalize(data):

for i in range(len(data)):

data[i] = 1.0 - data[i] / 255.0

return data.reshape(1,28,28)

argmax = lambda res: np.argmax(res.reshape(10))

mnist\_engine = tensorrt.lite.Engine(framework="tf", #Source framework

path=DATA\_DIR + "/mnist/lenet5\_mnist\_frozen.pb", #Model File

max\_batch\_size=10, #Max number of images to be processed at a time

input\_nodes={"in":(1,28,28)}, #Input layers

output\_nodes=["out"], #Ouput layers

preprocessors={"in":normalize}, #Preprocessing functions

postprocessors={"out":argmax}) #Postprocesssing functions

方法2.2: Using TensorRT to **Optimize Caffe Models** in Python

# 模型 -> plan

import tensorrt as trt

G\_LOGGER = trt.infer.ConsoleLogger(trt.infer.LogSeverity.ERROR)

def create\_engine(engine\_file):

MODEL\_PROTOTXT = './data/mnist/mnist.prototxt'

CAFFE\_MODEL = './data/mnist/mnist.caffemodel'

engine = trt.utils.caffe\_to\_trt\_engine(G\_LOGGER,

MODEL\_PROTOTXT,

CAFFE\_MODEL,

1,

1 << 20,

['prob'],

trt.infer.DataType.FLOAT)

trt.utils.write\_engine\_to\_file(engine\_file, engine.serialize())

engine.destroy()

# 使用plan进行inference

PyCUDA handles the CUDA operations needed to allocate memory on your GPU and to transfer data to the GPU and results back to the CPU.

NumPy is a well used tool to store and move data.

from tensorrt.parsers import caffeparser

import pycuda.driver as cuda

import pycuda.autoinit

import numpy as np

from random import randint

def inputs():

# 导入图像，(28,28) -> (28\*28, 1)

DATA = '/data/mnist/'

rand\_file = randint(0,9)

path = DATA + str(rand\_file) + '.pgm'

im = Image.open(path)

arr = np.array(im)

img = arr.ravel()

print("Test Case: " + str(rand\_file))

# 去均值

IMAGE\_MEAN = '/data/mnist/mnist\_mean.binaryproto'

INPUT\_H = 28

INPUT\_W = 28

parser = caffeparser.create\_caffe\_parser()

mean\_blob = parser.parse\_binary\_proto(IMAGE\_MEAN)

parser.destroy()

mean = mean\_blob.get\_data(INPUT\_W \*\* 2)

data = np.empty([INPUT\_W \*\* 2])

for i in range(INPUT\_W \*\* 2):

data[i] = float(img[i]) - mean[i]

mean\_blob.destroy()

return img, rand\_file

def infer\_by\_engine(context):

# convert input data to Float32

img = img.astype(np.float32)

# create output array to receive data

output = np.empty(OUTPUT\_SIZE, dtype=np.float32)

# allocate memory on the GPU with PyCUDA and

# register it with the engine.

d\_input = cuda.mem\_alloc(1 \* img.size \* img.dtype.itemsize)

d\_output = cuda.mem\_alloc(1 \* output.size \* output.dtype.itemsize)

bindings = [int(d\_input), int(d\_output)]

# create a cuda stream to run inference

stream = cuda.Stream()

cuda.memcpy\_htod\_async(d\_input, img, stream) #transfer input data to device

context.enqueue(1, bindings, stream.handle, None) #execute model

cuda.memcpy\_dtoh\_async(output, d\_output, stream) #transfer predictions back

stream.synchronize() #syncronize threads

return output

def user\_engine(engine\_file):

engine = trt.utils.load\_engine(G\_LOGGER, engine\_file)

runtime = trt.infer.create\_infer\_runtime(G\_LOGGER)

context = engine.create\_execution\_context()

img, label = inputs()

predict = infer\_by\_engine(context)

print("Test Case: " + str(label))

print ("Prediction: " + str(np.argmax(predict)))

engine.destroy()

context.destroy()

runtime.destroy()

def main():

engine\_file = "./data/mnist/new\_mnist.engine"

create\_engine(engine\_file)

user\_engine(engine\_file)

方法3： C++接口 （主要用于自定义网络）

TensorRT基本处理过程：caffe model -> GIE的model; 运行GIE引擎(数据提前copy到GPU中), 提取结果。

1. **简单例子见TensorRT/samples/sampleMNIST**

导入到netbeans,然后debug

（建议与python作对比）

自定义网络 （通过plugin实现)

功能：将最后一层全连接层name=”ip2”替换成自定义业务逻辑

1. **samplePlugin.cpp** （详细代码请参考TensorRT/samples/samplePlugin)

客户端使用自定义层

PluginFactory pluginFactory;

caffeToGIEModel("\*.prototxt", "\*.caffemodel", "prob", 1, &pluginFactory, "\*.engine");

pluginFactory.destroyPlugin();

uint8\_t img[INPUT\_H\*INPUT\_W]; // read image and substract mean

...

engine\_data, size = readFromfile(engine\_file); // read engine file

// deserialize the engine

IRuntime\* runtime = createInferRuntime(gLogger);

ICudaEngine\* engine = runtime->deserializeCudaEngine(

(const void\*)engine\_data.get(), size, &pluginFactory);

IExecutionContext \*context = engine->createExecutionContext();

// run inference

float prob[OUTPUT\_SIZE];

doInference(\*context, img, prob, 1);

void caffeToGIEModel(...){

// caffe -> network

IBuilder\* builder = createInferBuilder(gLogger);

**INetworkDefinition\* network = builder->createNetwork();**

ICaffeParser\* parser = createCaffeParser();

parser->setPluginFactoryExt(pluginFactory);

bool fp16 = builder->platformHasFastFp16();

const IBlobNameToTensor\* blobNameToTensor = parser->parse(

deployFile.c\_str(), modelFile.c\_str(), \*network,

fp16 ? DataType::kHALF : DataType::kFLOAT);

for (auto& s : outputs) **//标记某一层为输出层，每一层均可标记为输出**

network->markOutput(\*blobNameToTensor->find(s.c\_str()));

// Build the engine

builder->setMaxBatchSize(maxBatchSize);

**// MaxWorkspaceSize 设置小于需要的，则部分算法无法执行导致崩溃或结果不确定**

**// 可以通过如下方式获取最大工作空间engine->getWorkspaceSize();**

builder->setMaxWorkspaceSize(1 << 20);

builder->setHalf2Mode(fp16);

**ICudaEngine\* engine = builder->buildCudaEngine(\*network);**

// serialize the engine, then close everything down

write2file(engine->serialize(), engine\_file)

}

void doInference(IExecutionContext& context, float\* input, float\* output, int batchSize)

{

const ICudaEngine& engine = context.getEngine();

void\* buffers[2];

int inputIndex = engine.getBindingIndex(INPUT\_BLOB\_NAME),

outputIndex = engine.getBindingIndex(OUTPUT\_BLOB\_NAME);

cudaMalloc(&buffers[inputIndex], batchSize \* INPUT\_H \* INPUT\_W \* sizeof(float));

cudaMalloc(&buffers[outputIndex], batchSize \* OUTPUT\_SIZE \* sizeof(float));

cudaStream\_t stream;

cudaStreamCreate(&stream);

// DMA the input to the GPU, execute the batch asynchronously, and DMA it back:

input\_size = batchSize \* INPUT\_H \* INPUT\_W \* sizeof(float) ;

output\_size = batchSize \* OUTPUT\_SIZE\*sizeof(float);

cudaMemcpyAsync(buffers[inputIndex], input, input\_size, cudaMemcpyHostToDevice, stream);

**context.enqueue(batchSize, buffers, stream, nullptr);**

cudaMemcpyAsync(output, buffers[outputIndex], output\_size, cudaMemcpyDeviceToHost, stream);

cudaStreamSynchronize(stream);

}

void write2file(gieModelStream, engine\_file) {

std::ofstream outfile(engine\_file.c\_str(), std::ios::out | std::ios::binary);

unsigned char\* p = (unsigned char\*)gieModelStream->data();

outfile.write((char\*)p, gieModelStream->size());

}

void readFromfile(engine\_file) {

std::ifstream in\_file(engine\_file.c\_str(), std::ios::in | std::ios::binary);

std::size\_t size = fun(in\_file.tellg(), in\_file.seekg())

std::unique\_ptr<unsigned char[]> engine\_data(new unsigned char[size]);

in\_file.read((char\*)engine\_data.get(), size);

}

自定义层

class **FCPlugin**: public IPluginExt

{

public:

**FCPlugin(const Weights \*weights, int nbWeights, int nbOutputChannels)**

: mNbOutputChannels(nbOutputChannels){

// 获取权重和偏置blob

mKernelWeights = weights[0];

mBiasWeights = weights[1];

...allocate CPU memeory of mKernelWeights and mBiasWeights

...mKernelWeights.vlaues, mBiasWeights.values <- weights[0].values, weights[1].values

mNbInputChannels = int(weights[0].count / nbOutputChannels);

}

**FCPlugin(const void\* data, size\_t length)** { // create the plugin at runtime from a byte stream }

**~FCPlugin()** { //释放CPU资源 }

#----------------------required by creating the network---------------------#

int **getNbOutputs**() const override { return 1; }

Dims **getOutputDimensions**(int index, const Dims\* inputs, int nbInputDims) override

{ return DimsCHW(mNbOutputChannels, 1, 1); }

#----------------------required by builder---------------------------------#

void **configureWithFormat**(..., DataType type, ...) override { mDataType = type; }

bool **supportsFormat**(DataType type, PluginFormat format) const override

{ return (type == DataType::kFLOAT || type == DataType::kHALF) &&

format == PluginFormat::kNCHW; }

virtual size\_t **getWorkspaceSize**(int maxBatchSize) const override { return 0; }

#----------------------required by runtime----------------------------------#

int **initialize**() override {

// copy data from CPU to GPU

mDeviceKernel, mDeviceBias <- mKernelWeights, mBiasWeights

}

virtual int **enqueue**(int batchSize, const void\*const \* inputs, void\*\* outputs, void\* workspace, **cudaStream\_t stream**) override {

// 实现自定义层业务逻辑

cublasSetStream(mCublas, stream);

cudnnSetStream(mCudnn, stream);

if (mDataType == DataType::kFLOAT) { cublasSgemm(...); }

else { cublasHgemm(...); }

if (mBiasWeights.count) { cudnnAddTensor(...); }

}

virtual void **terminate**() override { //释放GPU资源 }

#----------------------required by serialization----------------------------#

virtual size\_t **getSerializationSize**() override {

return sizeof(mNbInputChannels) + sizeof(mNbOutputChannels) + sizeof(mBiasWeights.count) + sizeof(mDataType) + (mKernelWeights.count + mBiasWeights.count) \* type2size(mDataType); }

virtual void **serialize**(void\* buffer) override {

char\* d = static\_cast<char\*>(buffer), \*a = d;

write(d, mNbInputChannels);

write(d, mNbOutputChannels);

write(d, mBiasWeights.count);

write(d, mDataType);

convertAndCopyToBuffer(d, mKernelWeights);

convertAndCopyToBuffer(d, mBiasWeights);

}

private:

int mNbOutputChannels, mNbInputChannels;

Weights mKernelWeights, mBiasWeights;

void\* mDeviceKernel{nullptr};

void\* mDeviceBias{nullptr};

};

// integration for serialization

class **PluginFactory** : public nvinfer1::IPluginFactory, public nvcaffeparser1::IPluginFactoryExt

{

public:

// caffe parser plugin implementation，注意自定义层替代最后一层name=”ip2”

bool **isPlugin**(const char\* name) override { return **!strcmp(name, "ip2");** }

virtual IPlugin\* **createPlugin**(char\* layerName, const Weights\* weights, int nbWeights) override

{

assert(isPlugin(layerName) && nbWeights == 2);

const int NB\_OUTPUT\_CHANNELS = 10;

mPlugin = new FCPlugin(weights, nbWeights, NB\_OUTPUT\_CHANNELS);

return mPlugin.get();

}

// deserialization plugin implementation

IPlugin\* **createPlugin**(char\* layerName, const void\* serialData, size\_t serialLength) override

{

mPlugin = new FCPlugin(serialData, serialLength);

return mPlugin.get();

}

// the application has to destroy the plugin when it knows it's safe to do so

void **destroyPlugin**() { mPlugin.release(); }

std::unique\_ptr<FCPlugin> **mPlugin**{ nullptr };

};

方法4： TensorRT + Caffe

假定复杂网络拆解成三个网络： (如何拆解见caffe--剖析模型)

net1.prototxt, net1.caffemodel; 并优化成net1.engine (如何转换见tensorrt.lite)

net2.prototxt, net2.caffemodel; 假定很复杂无法转化为tensorrt, 直接用caffe

net3.prototxt, net3.caffemodel; 并优化成net3.engine

注意，用tensorrt.lite.Engine，不能同时优化几个网络，一个进程只能优化一个网络，否则第二个网络无法通过

级联使用engine + caffe + engine

engine\_visual = trt.lite.Engine(PLAN='step1.engine')

input\_name = 'Pooling2'

net2 = caffe.Net('net2.prototxt', 'net2.caffemodel', caffe.TEST)

engine\_ip = trt.lite.Engine(PLAN='step3.engine')

for i in range(10):

img, rand\_file = generate\_data()

img = np.reshape(img, (1, 1, 28, 28))

Pooling2 = engine\_visual.infer(img)

Pooling2 = Pooling2[0]

# net2.blobs[input\_name].reshape(\*Pooling2.shape)

net2.blobs[input\_name].data[...] = Pooling2

net2.forward()

InnerProduct1 = net2.blobs['InnerProduct1'].data

InnerProduct1 = np.reshape(InnerProduct1, (1, 1, 1, InnerProduct1.shape[-1]))

prob = engine\_ip.infer(InnerProduct1)

print("Test Case: " + str(rand\_file))

print("Prediction: " + str(np.argmax(prob)))

举例：

RefineDet

deploy.prototxt = backbone.prototxt + refinedet.prototxt

deploy.caffemodel = backbone.caffemodel + refinedet.caffemodel

**backbone.prototxt**

name: "backbone"

input: "data"

input\_shape { dim: 1 dim: 3 dim: 320 dim: 320 }

layer {

name: "conv1\_1"

type: "Convolution"

...

}

**refinedet.prototxt**

name: "refinedet"

input: "data"

input\_shape { dim: 1 dim: 3 dim: 320 dim: 320 }

input: "conv4\_3"

input\_shape { dim: 1 dim: 512 dim: 40 dim: 40 }

input: "conv5\_3"

input\_shape { dim: 1 dim: 512 dim: 20 dim: 20 }

input: "fc7"

input\_shape { dim: 1 dim: 1024 dim: 10 dim: 10 }

input: "conv6\_2"

input\_shape { dim: 1 dim: 512 dim: 5 dim: 5 }

layer {

name: "TL6\_1"

type: "Convolution"

bottom: "conv6\_2"

top: "TL6\_1"

...

}

Inference (图像大小： 每秒帧数)

# load model

engine\_visual = trt.lite.Engine(PLAN=engine\_backbone)

net2 = caffe.Net(prototxt\_refinedet, weights\_refinedet, caffe.TEST)

net2.blobs['data'].reshape(\*input\_size)

net2.blobs['conv4\_3'].reshape(1, 512, 40, 40)

net2.blobs['conv5\_3'].reshape(1, 512, 20, 20)

net2.blobs['fc7'].reshape(1, 1024, 10, 10)

net2.blobs['conv6\_2'].reshape(1, 512, 5, 5)

descriptor = []

for (i, im\_name) in enumerate(im\_names):

image\_file = 'examples/images/' + im\_name

image = caffe.io.load\_image(image\_file)

transformed\_image = transformer.preprocess('data', image)

ts = time()

output = engine\_visual.infer(transformed\_image)

net2.blobs['data'].data[...] = transformed\_image

net2.blobs['conv4\_3'].data[...] = output[0]

net2.blobs['conv5\_3'].data[...] = output[1]

net2.blobs['fc7'].data[...] = output[2]

net2.blobs['conv6\_2'].data[...] = output[3]

detections = net2.forward()['detection\_out']

te = time()

descriptor.append({

image.shape: 1.0/(te-ts)

})

1. Faster RCNN

TensorRT仅支持图像Portable PixMap (PPM)

Faster R-CNN 作了些改变

* The RPN and the ROI pooling layer is fused and replaced by a custom layer named RPROIFused .
* The reshape layer is replaced with two custom layers: ReshapeCTo2 and ReshapeCTo18 and are defined in the sample.

1. **Tensorflow -> TensorRT**
2. Unsupported Framework -> TensorRT

例子: PyTorch -> TensorRT

模型压缩

更精细模型的设计

如SqueezeNet、MobileNet等，使用更加细致、高效的模型设计，能够很大程度的减少模型尺寸，并且也具有不错的性能

模型剪枝 （有损）

将不重要的connection或者filter进行裁剪来减少模型的冗余

比如去除低阈值权重，然后稀疏编码

模型量化 （有损）

半精度浮点

为什么8bits对深度神经网络够用?

低功耗以及移动设备

八位数值的存取相对浮点数而言内存带宽降到 25%

后向传播和梯度需要浮点精度。推断时可以使用低精度

TensorFlow 自带对八位运算的生产级支持。它也能把浮点模型转换为等价的使用量化计算进行推断的图 (Graph；TensorFlow 里用来表达计算过程和内部状态的结构)

bazel -bin/tensorflow/contrib/quantization/tools/quantize\_graph \

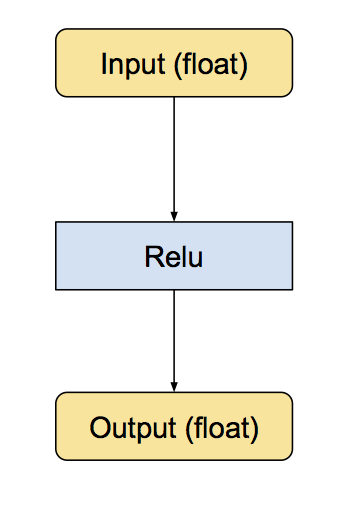
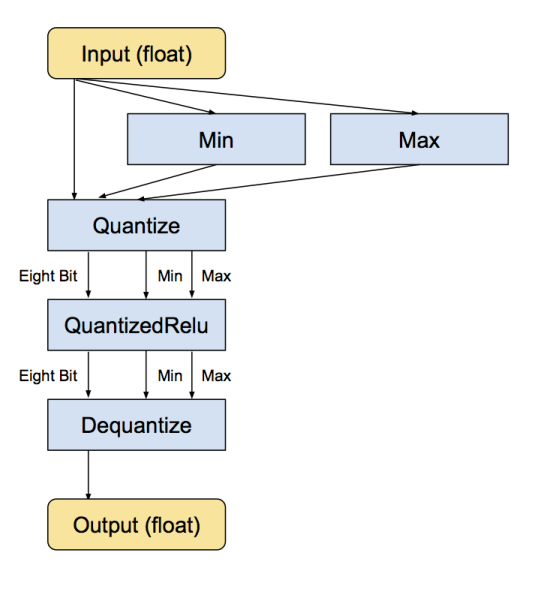
--input=/tmp/classify\_image\_graph\_def.pb \

--output\_node\_names="softmax" --output=/tmp/quantized\_graph.pb \

--mode=eightbit

量化是如何进行的？

通过把常见操作转换为等价的八位版本达到的。涉及的操作包括卷积，矩阵乘法，活化函数，池化操作，以及拼接

良好训练的神经网络必须能应对训练数据中的无关信息，这成就了神经网络对输入噪声和计算误差的强壮性

深度网络的一个魔力特性就是能够很好地应对较大的输入噪声。比如为了识别照片中的物体，网络必须忽略所有的 CCD 噪声、光照变化，以及其它与之前训练样本之间的非本质差异，而只关注重要的相似之处。这种能力意味着神经网络似乎把低精度计算视为另一种噪声来源，而在数值格式精度较低的情况下仍能给出准确结果。

通过实验也知道神经网络对噪声不敏感，所以量化噪声对结果总体来说影响不会很大

如何决定范围？如果你对八位输入进行任何算术计算，你会自然地累积得到超过八位精度能表达的数；需要把超过八位的输出缩减传递给下一个操作。对矩阵乘法而言，一个办法是根据可能的极端输入计算其输出的范围

舍入如何进行？神经网络对噪声不敏感，但如果对舍入操作不小心，会导致偏差往某个方向累积，最终影响精度

# Deep learning framework

|  |  |  |  |
| --- | --- | --- | --- |
|  | By | interface | description |
| Keras |  | Python | neural networks library  running on top of either TensorFlow or Theano, fast experimentation |
| Caffe2 | the Berkeley Vision and Learning Center (BVLC) | C, C++, Python, MATLAB, cmd | deep learning framework  made with expression, speed, and modularity in mind |
| TensorFlow | Google’s Machine Intelligence research organization | C++, Python | software library  numerical computation using data flow graphs |
| CNTK | The Microsoft Cognitive Toolkit | Python, C++, C# and cmd | unified deep-learning toolkit  train and combine popular model types across multiple GPUs and servers |
| theano |  | Python | math expression compiler  defines, optimizes, and evaluates mathematical expressions |
| Torch |  | C, C++, Lua | scientific computing framework  offers wide support for machine learning algorithms. |

Theano: a compiler for mathematical expressions in Python

TensorFlow: a python library for fast numerical computing

Keras: library addresses these concerns by providing a wrapper for both Theano and Tensorflow

scikit-learn library：general purpose machine learning framework in Python built on top

of SciPy

若内存足够大，一次将所有训练数据导入内存

model.fit(X\_train, Y\_train, batch\_size=batch\_size, nb\_epoch=12)

若CPU内存不大，无法一次性将所有训练数据导入，怎么办？

批次导入训练数据

存在多种方案：

方案一：先导入subset1, 训练12次，然后导入subset2,训练12次，...直至遍历所有训练集

方案二：先导入subset1, 训练1次，然后导入subset2, 训练1次，...遍历所有训练集，然后重复以上12次

以上是对mnist数据集的比较结果：建议采用方案二

0.989 total, 12 epoch 397s

0.986 11\*subset, 12 epoch / subset 396s

0.989 12 \* (11\*subset + 1 epoch / subset) 393s

keras 代码

from psutil import virtual\_memory

mem = virtual\_memory()

free\_memory = ratio \* float(mem.total) / 1024 \*\* 3

case\_num\_memory = f(free\_memory, case\_size)

subset\_num = ceil(all\_case\_num/subset\_num)

for j in range(epoches):

for i in range(subset\_num):

history = model.fit(train.subset(i), batch\_size=batch\_size, 1)

tensorflow代码

batch\_num = ceil(all\_case\_num/batch\_size)

for j in range(epoches):

for i in range(batch\_num):

batch = train.next\_batch(batch\_size)

sess.run([...], feed\_dict={X:batch[0], Y:batch[1]}

Caffe

<http://caffe.berkeleyvision.org/installation.html>

<https://www.cnblogs.com/go-better/p/7161006.html>

Install

cuda8, cudnn (keras有详细的说明）

$ pip install opencv-contrib-python 安装opencv

$ pkg-config --modversion opencv 获取opencv版本

sudo apt-get install libprotobuf-dev libleveldb-dev libsnappy-dev libopencv-dev libhdf5-serial-dev protobuf-compiler

sudo apt-get install --no-install-recommends libboost-all-dev

sudo apt-get install libopenblas-dev

$git clone https://github.com/BVLC/caffe.git

$cd caffe

caffe$ sudo cp Makefile.config.example Makefile.config

caffe$ sudo gedit Makefile.config

INCLUDE\_DIRS := $(PYTHON\_INCLUDE) /usr/local/include /usr/include/hdf5/serial

LIBRARY\_DIRS := $(PYTHON\_LIB) /usr/local/lib /usr/lib /usr/lib/x86\_64-linux-gnu /usr/lib/x86\_64-linux-gnu/hdf5/serial

若opencv3版本

OPENCV\_VERSION := 3

$ make all -j8

$ make test –j8

$ make runtest –j8

编译pycaffe

$ cd ~/caffe

caffe$ sudo apt-get install gfortran

$ cd ./python

caffe/python$ for req in $(cat requirements.txt); do pip install $req; done

$ cd ..

caffe$ sudo pip install -r python/requirements.txt

caffe$ make pycaffe -j8

若出现numpy找不到，需要设置Makefile.config

PYTHON\_INCLUDE := /usr/include/python2.7 \

/usr/lib/python2.7/dist-packages/numpy/core/include \

/usr/local/lib/python2.7/dist-packages/numpy/core/include

配置环境变量，以便python调用

sudo gedit ~/.bashrc

export PYTHONPATH=/home/caffe/python:$PYTHONPATH

source ~/.bashrc

tutorial

**Blob** storage and communication

a blob is an N-dimensional array stored in a C-contiguous fashion.

* batches of images

number N x channel K x height H x width W

Blob memory is row-major in layout, so the last / right most dimension changes fastest

* model parameters

For a convolution layer with 96 filters of 11 x 11 spatial dimension and 3 inputs the blob is 96 x 3 x 11 x 11.

For an inner product / fully-connected layer with 1000 output channels and 1024 input channels the parameter blob is 1000 x 1024

* derivatives for optimization

Blob stores two chunks of memories, data and diff. The former is the normal data that we pass along, and the latter is the gradient computed by the network.

要运行caffe，需要先创建一个model=layers + layer+ ...，每一屋又由许多参数组成。熟练使用caffe最重要的就是学会配置文件.prototxt的编写。

**Layer** computation and connections

A layer takes input through bottom connections and makes output through top connections.

数据层是每个模型的最底层，是模型的入口，不仅提供数据的输入，也提供数据从Blobs转换成别的格式进行保存输出。通常**数据的预处理（如减去均值, 放大缩小, 裁剪和镜像等**），也在这一层设置参数实现。数据来源可以来自高效的数据库（如LevelDB和LMDB），也可以直接来自于内存。如果不是很注重效率的话，数据也可来自磁盘的hdf5文件和图片格式文件

1. **Data Layers**

Image Data - read raw images.

Database - read data from LEVELDB or LMDB.

HDF5 Input - read HDF5 data, allows data of arbitrary dimensions.

HDF5 Output - write data as HDF5.

Input - typically used for networks that are being deployed.

Window Data - read window data file.

Memory Data - read data directly from memory.

Dummy Data - for static data and debugging.

Data layers load input and save output by converting to and from Blob to other formats

This data layer definition:

layer {

name: "mnist" #表示该层名称

**type: "Data"** # loads leveldb or lmdb storage DBs

# top或bottom: 每一层用bottom来输入数据，用top来输出数据

top: "data" # the 1st top is the data itself: the name is only convention

top: "label" # the 2nd top is the ground truth

# a rule, 指定该层属于训练或测试等

**include: { phase: TRAIN }**

data\_param { # the Data layer configuration

**source**: "examples/mnist/mnist\_train\_lmdb" # path to the DB

# type of DB: LEVELDB or LMDB (LMDB supports concurrent reads)

**backend**: LMDB

**batch\_size**: 64 # batch processing improves efficiency.

}

# 数据的预处理

transform\_param { # common data transformations

# feature scaling coefficient: this maps the [0, 255] MNIST data to [0, 1]

scale: 0.00390625

mean\_file\_size: "cifar10/mean.binaryproto" # 用一个配置文件来进行均值操作

mirror: 1 # 1/ture表示开启镜像，0/false表示关闭

crop\_size: 227 # 剪裁 227\*227图块，训练阶段随机剪裁，测试阶段从中间裁剪

}

}

数据来自于图片

layer {

**type: "ImageData"**

**image\_data\_param** {

# 一个文本文件的名字，每一行给定一个图片文件的名称和标签（label)

**source**: "examples/\_temp/file\_list.txt"

**batch\_size**: 50

# 如果设置，则将图片进行resize

**new\_height**: 256

**new\_width**: 256

}

...

}

1. **Vision Layers**

Vision layers usually take images as input and produce other images as output, other layers (with few exceptions) ignore the spatial structure of the input, effectively treating it as “one big vector” with dimension chw.

* Convolution Layer - convolves the input image with a set of learnable filters, each producing one feature map in the output image.

layer {

name: "conv1"

**type: "Convolution"**

# lr = lr\_mult\*base\_lr (学习率base\_lr在solver.prototxt设置）

param { lr\_mult: 1 } **#weight学习率系数调整**

param { lr\_mult: 2 } **#bias学习率系数调整**

convolution\_param {

**num\_output**: 20 # 卷积核（filter)的个数

**kernel\_size**: 5 # 卷积核的大小

stride: 1 # 卷积核的步长，默认为1

pad: 0 # 扩充边缘，默认为0，不扩充。 扩充的时候是左右、上下对称的，比如卷积核的大小为5\*5，那么pad设置为2，则四个边缘都扩充2个像素，即宽度和高度都扩充了4个像素,这样卷积运算之后的特征图就不会变小

# 权值初始化， 默认为“constant",值全为0， 还可以选择"xavier" / "gaussian"

weight\_filler { type: "xavier" }

# 偏置项的初始化。一般设置为"constant",值全为0

bias\_filler { type: "constant" }

}

**bottom**: "data"

**top**: "conv1"

}

注：

xavier algorithm: initialize based on the number of input/output neurons

输入：n\*c0\*w0\*h0

输出：n\*num\_output\*w1\*h1

w1=(w0+2\*pad-kernel\_size)/stride+1;

h1=(h0+2\*pad-kernel\_size)/stride+1;

一般设置stride=1, pad=(kernel\_size-1)/2, 那么宽度高度不变

* Pooling Layer - max, average, or stochastic pooling.

layer {

name: "pool1"

**type: "Pooling"**

pooling\_param {

**kernel\_size: 2** # 池化的核大小

pad: 0 # 边缘扩充, 默认为0

stride: 2# 池化的步长，默认为1。一般设为2，即不重叠

pool: MAX#池化方法，默认为MAX。MAX/AVE/STOCHASTIC

}

**bottom**: "conv1"

**top**: "pool1"

}

注：

输入：n\*c\*w0\*h0

输出：n\*c\*w1\*h1

和卷积层的区别就是其中的c保持不变

w1=(w0+2\*pad-kernel\_size)/stride+1;

h1=(h0+2\*pad-kernel\_size)/stride+1;

一般设置stride=2, 特征图池化后缩小一倍

* Spatial Pyramid Pooling (SPP)
* Crop - perform cropping transformation.
* Deconvolution Layer - transposed convolution.

1. **Activation / Neuron Layers**

* ReLU / Rectified-Linear and Leaky-ReLU - ReLU and Leaky-ReLU rectification.

layer {

name: "relu1"

**type: "ReLU"**

**bottom**: "ip1"

**top**: "ip1"

}

Since ReLU is an element-wise operation, we can do **in-place operations to save some memory**. This is achieved by simply giving the same name to the bottom and top blobs

* PReLU - parametric ReLU.
* ELU - exponential linear rectification.
* Sigmoid, TanH, Absolute Value
* Power - f(x) = (shift + scale \* x) ^ power.
* Exp - f(x) = base ^ (shift + scale \* x).
* Log - f(x) = log(x).
* BNLL - f(x) = log(1 + exp(x)).
* Threshold - performs step function at user defined threshold.
* Bias - adds a bias to a blob that can either be learned or fixed.
* Scale - scales a blob by an amount that can either be learned or fixed.

1. **Common Layers**

* Inner Product : 全连接层

layer {

name: "ip1"

**type: "InnerProduct"**

param { lr\_mult: 1 }

param { lr\_mult: 2 }

inner\_product\_param {

**num\_output:** 500

weight\_filler { type: "xavier" }

bias\_filler { type: "constant" }

}

**bottom**: "pool2"

**top**: "ip1"

}

* Dropout

layer {

name: "drop7"

type: "Dropout"

bottom: "fc7-conv"

top: "fc7-conv"

dropout\_param {

**dropout\_ratio:** 0.5

}

}

* Softmax

layers {

name: "prob"

**type: “Softmax"**

bottom: "cls3\_fc"

top: "prob"

}

* Reshape

layer {

name: "reshape"

**type: "Reshape"**

bottom: "input"

top: "output"

# 用于指定blob数据的各维的值（blob是一个四维的数据：n\*c\*w\*h）。

reshape\_param {

shape {

**dim**: 0 # 表示维度不变，即输入和输出是相同的维度。

**dim**: 0

**dim**: 14 # 将原来的维度变成14

**dim**: -1 # infer it from the other dimensions

}

}

}

注： 假设原数据为：64\*3\*28\*28， 输出数据为：64\*3\*14\*56

* Flatten: 将(n, c, h, w) -> (n, c\*h\*w, 1, 1)
* Batch Reindex,
* Split: 一输入 -> 多输出
* Concat:多输入 -> 一输出
* Slicing
* Eltwise - element-wise operations such as product or sum between two blobs.
* Filter / Mask - mask or select output using last blob.
* Parameter - enable parameters to be shared between layers.
* Reduction - reduce input blob to scalar blob using operations such as sum or mean.
* Silence - prevent top-level blobs from being printed during training.
* ArgMax
* Python - allows custom Python layers.
* Embed - for learning embeddings of one-hot encoded vector (takes index as input).
* Im2col

1. **Loss Layers**

* Softmax with Loss - computes the multinomial logistic loss of the softmax of its inputs. It’s conceptually identical to a softmax layer followed by a multinomial logistic loss layer, but provides a more numerically stable gradient.

layer {

name: "loss"

**type: "SoftmaxWithLoss"**

**bottom**: "ip2"

**bottom**: "label"

**top**: “loss”

}

to compute the loss function value, report it when backpropagation starts, and initiates the gradient with respect to ip2

* Multinomial Logistic Loss
* Infogain Loss - a generalization of MultinomialLogisticLossLayer.
* Sum-of-Squares / Euclidean - computes the sum of squares of differences of its two inputs, 12N∑Ni=1‖x1i−x2i‖22.
* Hinge / Margin - The hinge loss layer computes a one-vs-all hinge (L1) or squared hinge loss (L2).
* Sigmoid Cross-Entropy Loss - computes the cross-entropy (logistic) loss, often used for predicting targets interpreted as probabilities.
* Contrastive Loss

1. **Accuracy / Top-k layer**

scores the output as an accuracy with respect to target – it is not actually a loss and has no backward step.

layer {

name: "accuracy"

**type: "Accuracy"**

**bottom**: "ip2"

**bottom**: "label"

**top**: "accuracy"

**include { phase: TEST }** 仅在测试时，才计算accuracy

}

1. **Normalization Layers**

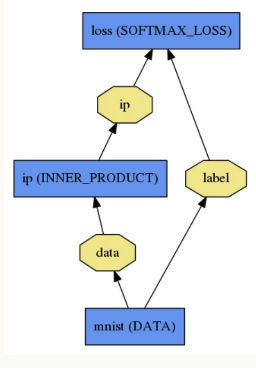
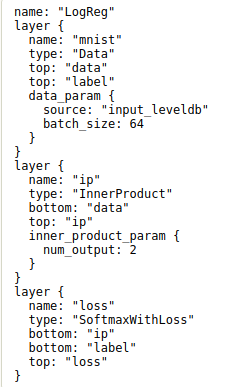
* Local Response Normalization (LRN) - performs a kind of “lateral inhibition” by normalizing over local input regions.
* Mean Variance Normalization (MVN) - performs contrast normalization / instance normalization.
* Batch Normalization - performs normalization over mini-batches.

1. **Recurrent Layers**

Recurrent, RNN, Long-Short Term Memory (LSTM)

**Net =** layers + layers + ...

The net is a set of layers connected in a computation graph – a directed acyclic graph (DAG) , A typical net begins with a data layer that loads from disk and ends with a loss layer that computes the objective for a task such as classification or reconstruction.

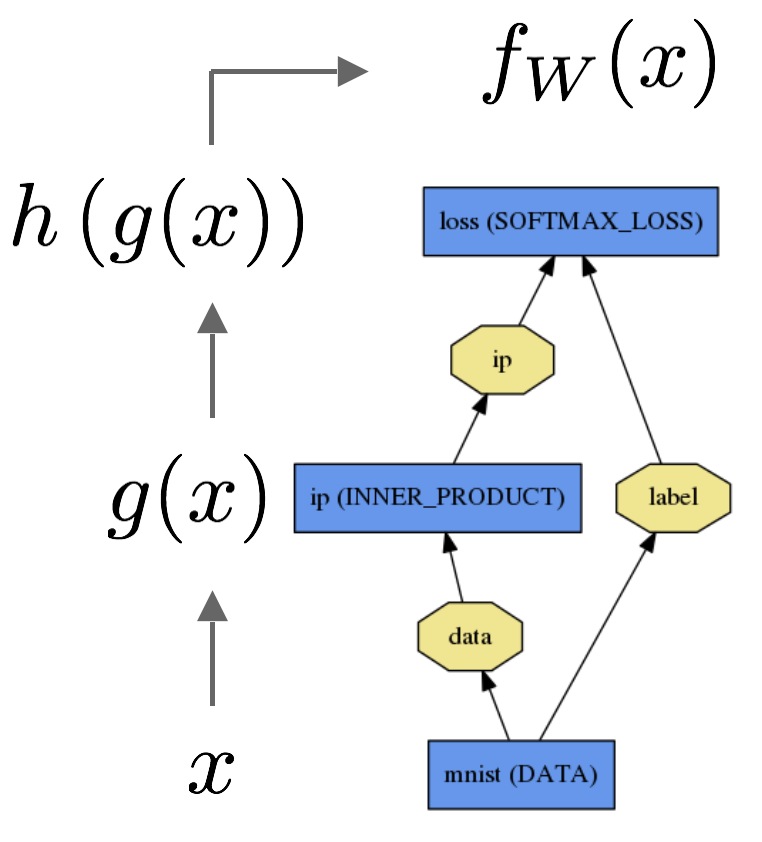
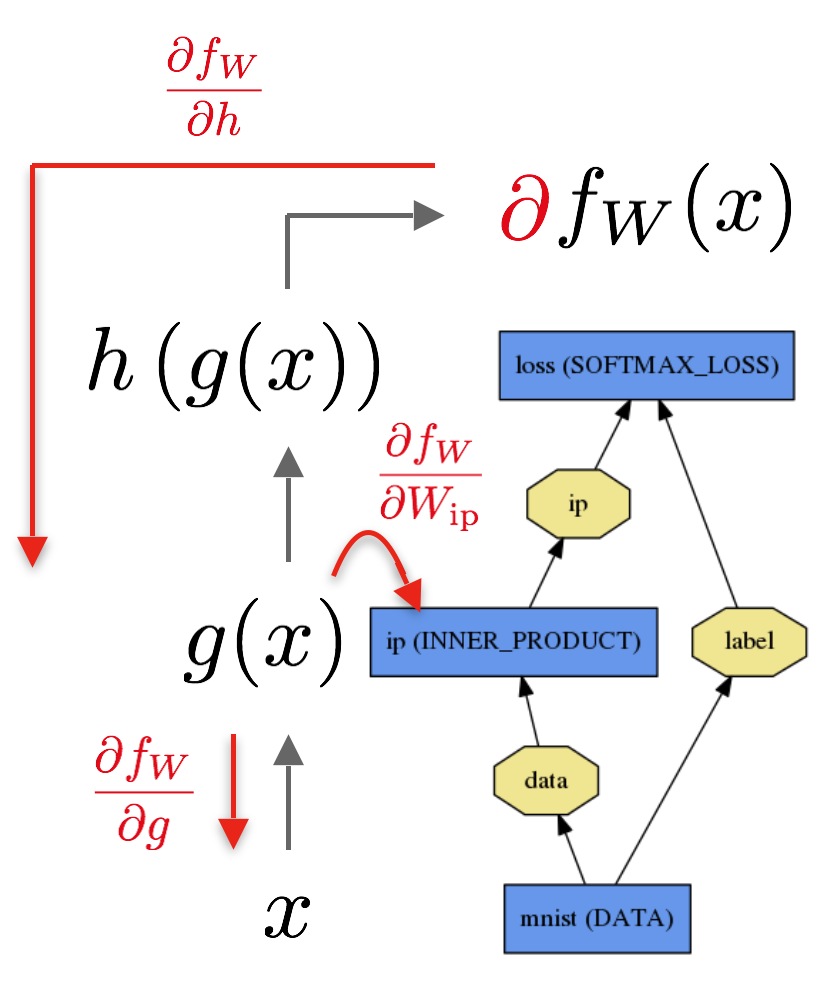
 

**Model**

models definition: **.prototxt** (in plaintext)

models learned: **.caffemodel** (serialized as binary).

**Forward / Backward**

The data x is passed through an inner product layer for g(x) then through a softmax for h(g(x)) and softmax loss to give fW(x).

The backward pass begins with the loss and computes the gradient with respect to the output ∂fW/∂h. The gradient with respect to the rest of the model is computed layer-by-layer through the chain rule. Layers with parameters, like the INNER\_PRODUCT layer, compute the gradient with respect to their parameters ∂fW/∂Wip during the backward step.

**Loss (error, cost, or objective)**

learning is driven by a loss function. By convention, Caffe layer types with the suffix Loss contribute to the loss function, but other layers are assumed to be purely used for intermediate computations. However, any layer can be used as a loss by adding a field loss\_weight: <float> to a layer definition for each top blob produced by the layer

# 添加正则化损失函数

loss := 0

for layer in layers:

for top, loss\_weight in layer.tops, layer.loss\_weights:

loss += loss\_weight \* sum(top)

在Deep Learning中，往往loss function是非凸的，没有解析解，我们需要通过优化方法来求解。solver的主要作用就是交替调用前向（forward)算法和后向（backward)算法来更新参数，从而最小化loss，实际上就是一种迭代的优化算法。

**Solver**

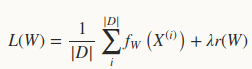
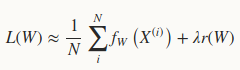
optimizes a model by

1. first calling forward to yield the output and loss,
2. then calling backward to generate the gradient of the model,
3. and then incorporating the gradient into a weight update that attempts to minimize the loss

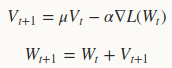
each iteration

1. calls network forward to compute the output and loss
2. calls network backward to compute the gradients
3. incorporates the gradients into parameter updates according to the solver method
4. updates the solver state according to learning rate, history, and method

损失函数: dataset D -> mini-batch N

 -> 

SGD:



learning rate α is the weight of the negative gradient.

The momentum μ is the weight of the previous update

./examples/imagenet/**alexnet\_solver.prototxt**

设置深度网络模型。每一个模型就是一个net，需要在一个专门的配置文件中对net进行配置，每个net由许多的layer所组成

注意的是：文件的路径要从caffe的根目录开始，其它的所有配置都是这样。

也可用train\_net和test\_net来对训练模型和测试模型分别设定。例如：

train\_net: "examples/hdf5\_classification/logreg\_auto\_train.prototxt"

test\_net: "examples/hdf5\_classification/logreg\_auto\_test.prototxt"

**net:** "examples/mnist/lenet\_train\_test.prototxt"

在caffe中的一次迭代iter指的是一个batch,而不是一张图片

/在测试的时候，需要迭代的次数; test\_iter = 测试样本总数/test\_batch\_size

**test\_iter**: 100

//经过多少次batch\_size的训练，然后进行一次测试

一般遍历一次全部训练数据(即一个epoch), 然后测试一次，则test\_interval = 训练样本总数/ train\_batch\_size

**test\_interval**: 500

**max\_iter**: 350000 # 350K次batch\_size训练， 建议epoch \* 训练集大小/batch\_size

iter\_size： X

每次循环都会以batch\_size大小计算梯度和loss，最后再取iter\_size次的平均。可以看成iter\_size\*batch\_size次更新一次参数;即处理batchsize\*itersize张图片后，才进行梯度下降

如何更新学习率？

lr\_policy可以设置为下面这些值，相应的学习率的计算为：

fixed #保持base\_lr不变.

step　 #step + stepsize + gamma

#返回 base\_lr \* gamma ^ (floor(iter / stepsize)),iter当前迭代次数

exp 　　 #返回base\_lr \* gamma ^ iter， iter为当前迭代次数

inv　　 #inv + power

#返回base\_lr \* (1 + gamma \* iter) ^ (- power)

multistep #multistep + stepvalue

#step是均匀等间隔变化，而multistep则是根据stepvalue值变化

poly 　 #学习率进行多项式误差, 返回 base\_lr (1 - iter/max\_iter) ^ (power)

sigmoid #学习率进行sigmod衰减

#返回 base\_lr ( 1/(1 + exp(-gamma \* (iter - stepsize))))

**base\_lr**: 0.01 # begin training at a learning rate of 0.01 = 1e-2

**lr\_policy**: "step" # learning rate policy: drop the learning rate

**gamma**: 0.1 # drop the learning rate by a factor of 10

**stepsize**: 100000 # drop the learning rate every 100K iterations

**momentum**: 0.9

type: SGD # 默认值就是SGD， AdaDelta/AdaGrad/Adam/Nesterov/RMSProp

**weight\_decay**: 0.0005 # 权重衰减项

**display**: 100 #每训练100次，在屏幕上显示一次。如果设置为0，则不显示

快照。将训练出来的model和solver状态进行保存，snapshot用于设置训练多少次后进行保存，默认为0，不保存。snapshot\_prefix设置保存路径

**snapshot**: 5000

**snapshot\_prefix**: "examples/mnist/lenet"

#保存的类型HDF5/BINARYPROTO ，默认为BINARYPROTO

snapshot\_format: "BINARYPROTO"

snapshot\_diff: false 是否保存梯度值，默认为false

solver\_mode: CPU CPU/GPU, 默认为GPU

If learning diverges (e.g., you start to see very large or NaN or inf loss values or outputs), try dropping the base\_lr (e.g., base\_lr: 0.001) and re-training, repeating this until you find a base\_lr value that works.

感觉训练时噪音太大时:

调大batch size到喂饱硬件 (显卡容易爆）

带learning rate下降的sgd，开始时依赖batch带来的噪音快速下降，接下来使用较低的learning rate消除这些噪音寻求稳定收敛

大batch size相当于小lr，反之亦然。可以根据收敛速度进行一定选择。

另外如果用了batchnorm，batch size别太小（大于64？）

如果不用batchnorm，可以考虑用小batch size甚至1 来得到最优的结果

可视化.prototxt

可视化编辑器

网址：<http://ethereon.github.io/netscope/#/editor>

内容copy到左边编辑框，然后shift+enter

Mouse指者对应层，可以显示对应层的信息

.prototxt -> 图片

$sudo apt-get install graphviz

$sudo pip install pydot

# **sudo python python/draw\_net.py** \*.prototxt \*.png --rankdir=BT

第一个参数：网络模型的prototxt文件

第二个参数：保存的图片路径及名字

第二个参数：--rankdir=x , x 有四种选项，分别是LR, RL, TB, BT 。用来表示网络的方向，分别是从左到右，从右到左，从上到小，从下到上。默认为ＬＲ。

**Interfaces--Command Line (train, test, time)**

$ sudo gedit ~/.bashrc

export PATH="/home/qzlin/Documents/dl/caffe/build/tools:$PATH"

$ **caffe train**

#模型的配置文件

**-solver** examples/finetuning/solver**.prototxt**

#预先训练好的权重来fine-tuning模型，不能和-snapshot同时使用

**-weights** models/bvlc\_reference\_caffenet/ bvlc**.caffemodel**

# 指定用哪一块gpu运行，或-gpu all使用所有gpu运行

**-gpu all**

从snapshot中恢复训练。可以在solver配置文件设置快照，保存solverstate

**-snapshot** model\_iter\_1000**.solverstate**

$ **caffe test**

**-model** examples/mnist/lenet\_train\_test.prototxt /

**-weights** examples/mnist/lenet\_iter\_10000.caffemodel /

-gpu 0,1 -iterations 100

$ **caffe time**

-model examples/mnist/lenet\_train\_test.prototxt /

-weights examples/mnist/lenet\_iter\_10000.caffemodel /

-gpu 0 -iterations 10

**训练和测试自己的图片**

1. 图像数据 -> db（leveldb/lmdb)

# sudo vi examples/images/create\_filelist.sh

# sudo vi examples/images/create\_lmdb.sh

1. 计算均值并保存

图片减去均值再训练，会提高训练速度和精度

# sudo build/tools/compute\_image\_mean examples/myfile/img\_train\_lmdb examples/myfile/mean.binaryproto

1. 创建模型并编写配置文件

# sudo cp models/bvlc\_reference\_caffenet/solver.prototxt examples/myfile/

# sudo cp models/bvlc\_reference\_caffenet/train\_val.prototxt examples/myfile/

1. 训练

# sudo build/tools/caffe train -solver examples/myfile/solver.prototxt

example:

下载mnist数据, 在 data/mnist/目录下有四个文件

train-images-idx3-ubyte: 训练集样本 (9912422 bytes)

train-labels-idx1-ubyte: 训练集对应标注 (28881 bytes)

t10k-images-idx3-ubyte: 测试集图片 (1648877 bytes)

t10k-labels-idx1-ubyte: 测试集对应标注 (4542 bytes)

caffe$ sudo sh data/mnist/get\_mnist.sh

解析手写体，从而获取图像，代码见

<https://www.jianshu.com/p/84f72791806f>

转换成LMDB数据

在 examples/mnist/目录下，生成两个文件夹

mnist\_train\_lmdb

data.mdb

lock.mdb

mnist\_test\_lmdb

caffe$ sudo sh examples/mnist/create\_mnist.sh

训练

$ sudo time sh examples/mnist/train\_lenet.sh

or

$ caffe train -solver examples/mnist/lenet\_solver.prototxt

若没有权限访问mnist

$sudo chmod -R 777 .

**Interfaces--Python**

安装

$ make pycaffe

$ sudo gedit ~/.bashrc

export PYTHONPATH=/path/to/caffe/python:$PYTHONPATH

**caffe.Net:** is the central interface for loading, configuring, and running models. **caffe.Classifier and caffe.Detector:** provide convenience interfaces for common tasks.

**caffe.SGDSolver:** exposes the solving interface.

**caffe.io:** handles input / output with preprocessing and protocol buffers.

**caffe.draw:** visualizes network architectures.

Caffe blobs are exposed as numpy ndarrays for ease-of-use and efficiency.

from caffe import layers, params

import caffe

**layers**里面包含了Caffe内置的层（比如卷积，ReLU等），

**params**包含了各种枚举值

net=caffe.NetSpec() //获取Caffe的一个Net

...填充这个net

...最后输出到文件即prototxt

1. **用代码来生成train.prototxt, test.prototxt, solver.prototxt文件**

**deploy文件最好手动生成**

# 添加caffe/python路径

import sys

sys.path.insert(0, "python")

import numpy as np

from caffe import layers as L, params as P

import caffe

data\_dir = 'examples/mnist'

file\_dir = 'examples/qzlin/mnist'

def create\_net(lmdb, batch\_size, include\_acc=False):

data, label = **L.Data**(source=lmdb, backend=P.Data.LMDB, batch\_size=batch\_size, ntop=2, transform\_param=dict(scale=0.00390625))

conv1 = **L.Convolution**(data, kernel\_size=5, stride=1, num\_output=16, pad=2, weight\_filler=dict(type='xavier'))

relu1 = **L.ReLU**(conv1, in\_place=True)

pool1 = **L.Pooling**(relu1, pool=P.Pooling.MAX, kernel\_size=3, stride=2)

conv2 = L.Convolution(pool1, kernel\_size=3, stride=1, num\_output=32, pad=1, weight\_filler=dict(type='xavier'))

relu2 = L.ReLU(conv2, in\_place=True)

pool2 = L.Pooling(relu2, pool=P.Pooling.MAX, kernel\_size=3, stride=2)

fc3 = **L.InnerProduct**(pool2, num\_output=1024, weight\_filler=dict(type='xavier'))

relu3 = L.ReLU(fc3, in\_place=True)

drop3 = **L.Dropout**(relu3, in\_place=True)

fc4 = L.InnerProduct(drop3, num\_output=10, weight\_filler=dict(type='xavier'))

loss = **L.SoftmaxWithLoss**(fc4, label)

if include\_acc:

acc = **L.Accuracy**(fc4, label)

return **caffe.to\_proto**(loss, acc)

else:

return caffe.to\_proto(loss)

def write\_net(train\_proto=file\_dir+'/train.prototxt',

test\_proto=file\_dir+'/test.prototxt',

train\_lmdb=data\_dir+'/mnist\_train\_lmdb',

test\_lmdb=data\_dir+'/mnist\_test\_lmdb'):

with open(train\_proto, 'w') as f:

f.write(str(create\_net(train\_lmdb, batch\_size=64)))

with open(test\_proto, 'w') as f:

f.write(str(create\_net(test\_lmdb, batch\_size=100, include\_acc=True)))

def create\_solver(train\_proto, test\_proto):

s = **caffe.proto.caffe\_pb2.SolverParameter**()

s.train\_net = train\_proto

s.test\_net.append(test\_proto)

# 每遍历完整个训练集,测试一次

s.test\_interval = int(np.ceil(50000.0/64))

# 20为等价于keras的epoches

s.max\_iter = s.test\_interval \* 20

# 测试集迭代次数 = 测试集/batch\_size

s.test\_iter.append(int(10000.0/100))

s.base\_lr = 0.01

s.momentum = 0.9

s.weight\_decay = 0.0005

s.lr\_policy = 'inv'

s.gamma = 0.0001

s.power = 0.75

s.display = s.test\_interval

s.snapshot = s.max\_iter/2

s.snapshot\_prefix = file\_dir+'shapshot'

s.solver\_mode = caffe.caffe\_pb2.SolverParameter.GPU

return s

def write\_solver(train\_proto=file\_dir+'/train.prototxt',

test\_proto=file\_dir+'/test.prototxt',

solver\_file=file\_dir+'/solver.prototxt'):

s = create\_solver(train\_proto, test\_proto)

with open(solver\_file, 'w') as f:

f.write(str(s))

deploy.prototxt

<https://github.com/BVLC/caffe/wiki/Using-a-Trained-Network:-Deploy>

1. Remove the data layer that was used for training
2. Remove any layer that is dependent upon data labels.
3. Set the network up to accept data.
4. Have the network output the result.

和test.prototxt文件差不多，只是头尾不相同。

deploy文件

第一层数据输入层只有维度信息

没有最后的Accuracy层，换成输出层(Softmax概率层)

name: "Mnist"

layer {

name: "data"

type: "Input"

top: "Data1"

input\_param { shape: { dim: 1 dim: 1 dim: 28 dim: 28 } }

}

**input: "Data1"**

**input\_dim: 1**

**input\_dim: 1**

**input\_dim: 28**

**input\_dim: 28**

输入层，以上左右框演示，任选一个

...

test.prototxt去除头尾层

...

layer {

name: "prob"

type: "Softmax"

bottom: "InnerProduct2"

top: "prob"

}

1. **剖析模型**

<https://github.com/BVLC/caffe/blob/master/examples/net_surgery.ipynb>

训练好的模型:

**.caffemodel**

这个文件里面存放的就是各层的参数，即net.params，里面没有数据(net.blobs)

**.solverstate**文件，这个和caffemodel差不多，但它多了一些数据，如模型名称、当前迭代次数; solverstate是用来恢复训练的，防止意外终止而保存的快照

导入模型

net = caffe.Net(deploy, caffemodel, caffe.TEST)

# 获取模型权重和偏置等参数

for layer\_name, param in net.params.items():

print layer\_name + '\t' + str(param[0].data.shape), str(param[1].data.shape)

根据blobs名称获取输出数据

net.blobs['data'].data #取出数据层blob 1\*c\*h\*w

net.blobs['conv'].data #取出conv层输出 1\*16\*h\*w

根据层名称获取层权重和偏置

net.params['conv'][0].data #取出conv权重 (16, 3, 5, 5)

#[‘conv’][1].data为偏置

# 获取各层**blobs**

net.blobs[input\_name].data[...] = input //首先输入数据

net.forward() //运行网络前向计算

for layer\_name, blob in net.blobs.items(): //获取各个层blobs

print layer\_name + '\t' + str(blob.data.shape)

可视化卷积核和特征图，见链接

修改模型

训练后的prototxt and caffemodel， 层参数是一一对应的

创建网络时，prototxt层参数可以比caffemodel少，比如剪裁版prototxt, 那么创建的网络只导入prototxt层参数，然后保存该网络，只获取prototxt对应的caffemodel

例子：deploy.txt + deploy.caffemodel ->

net1.prototxt, net1.caffemodel

net2.prototxt, net2.caffemodel

net3.prototxt, net3.caffemodel

deploy.prototxt

net3.prototxt

net2.prototxt

net1.prototxt

name: "Mnist"

input: "data"

input\_shape { dim: 1 dim: 1 dim: 28 dim: 28}

layer {

name: "Convolution1"

type: "Convolution"

bottom: "Data1"

...

}

...

layer {

name: "Pooling2"

type: "Pooling"

bottom: "Convolution2"

top: "Pooling2"

...

}

...

layer {

name: "InnerProduct1"

type: "InnerProduct"

bottom: "Pooling2"

top: "InnerProduct1"

...

}

layer {

name: "InnerProduct2"

type: "InnerProduct"

bottom: "InnerProduct1"

top: "InnerProduct2"

...

}

layer {

name: "prob"

type: "Softmax"

bottom: "InnerProduct2"

top: "prob"

}

name: "net1"

input: "data"

input\_shape { dim: 1 dim: 1 dim: 28 dim: 28}

layer {

name: "Convolution1"

type: "Convolution"

bottom: "Data1"

...

}

...

layer {

name: "Pooling2"

type: "Pooling"

bottom: "Convolution2"

top: "Pooling2"

...

}

name: "net2"

input: "Pooling2"

input\_shape: { dim: 1 dim: 32 dim: 7 dim: 7 }

layer {

name: "InnerProduct1"

type: "InnerProduct"

bottom: "Pooling2"

...

}

...

layer {

name: "Dropout1"

type: "Dropout"

bottom: "InnerProduct1"

top: "InnerProduct1"

}

name: "net3"

input: "InnerProduct1"

input\_shape: { dim: 1 dim: 1dim: 1 dim: 1024 }

layer {

name: "InnerProduct2"

type: "InnerProduct"

bottom: "InnerProduct1"

top: "InnerProduct2"

...

}

layer {

name: "prob"

type: "Softmax"

bottom: "InnerProduct2"

top: "prob"

}

CAFFE\_MODEL = 'examples/qzlin/mnist/snapshot\_iter\_15640.caffemodel'

def create\_caffemodel\_by\_prototxt(keyname):

deploy = 'mnist/{}.prototxt'.format(keyname)

net = caffe.Net(deploy, CAFFE\_MODEL, caffe.TEST)

net.save('mnist/{}.caffemodel'.format(keyname))

for i in range(1, 4):

create\_caffemodel\_by\_prototxt("step{}".format(i))

级联使用net1.caffemodel, net2.caffemodel, net3.caffemodel

im = caffe.io.load\_image('mnist/0.png', False)

input = im[np.newaxis, :, :, :].transpose(0, 3, 1, 2) # hwc -> 1chw

net1 = dig\_net(input, input\_name='Data1',

deploy='mnist/step1.prototxt',

caffemodel='mnist/step1.caffemodel')

Pooling2 = net1.blobs['Pooling2'].data

net2 = dig\_net(Pooling2, input\_name='Pooling2', deploy, caffemodel)

InnerProduct1 = net2.blobs['InnerProduct1'].data

net3 = dig\_net(InnerProduct1, input\_name='InnerProduct1', deploy, caffemodel)

prob = net3.blobs['prob'].data[0]

1. **训练--绘制loss and accuracy曲线**

准备数据集

caffe\_root$ ./data/cifar10/get\_cifar10.sh 下载数据集

caffe\_root$ ./examples/cifar10/create\_cifar10.sh转化为lmdb格式(含均值文件)

def loss\_accuracy():

# 注意：数据路径,模型,优化等都包含在内

# 1. 无论模型中Slover类型是什么统一设置为SGD

solver = caffe.SGDSolver('examples/cifar10/cifar10\_quick\_solver.prototxt')

# 2. 根据solver的prototxt中solver\_type读取，默认为SGD

solver = caffe.get\_solver('/home/xxx/data/solver.prototxt')

#方式1: 训练模型

# solver.solve()

# 方式2: 迭代训练模型

niter = 4000

test\_interval = 200

train\_loss = np.zeros(niter)

test\_acc = np.zeros(int(np.ceil(niter/test\_interval)))

for it in range(niter):

#进行一次前向传播一次反向传播并根据梯度更新参数

#solver.net.forward() # 前向传播

#solver.net.backward() # 反向传播,计算梯度

solver.step(1)

train\_loss[it] = solver.net.blobs['loss'].data

solver.test\_nets[0].forward(start='conv1')

if it % test\_interval == 0:

acc = solver.test\_nets[0].blobs['accuracy'].data

test\_acc[it // test\_interval] = acc

保存模型

solver.net.save('mymodel.caffemodel')

1. **测试**

def evaluate():

#加载Model

net\_file = '.../deploy.prototxt'

caffe\_model = '....caffemodel'

net = caffe.Net(net\_file, caffe\_model, caffe.TEST)

# 图像预处理

# 设定图片的shape格式为网络data层格式

transformer = caffe.io.Transformer({'data': net.blobs['data'].data.shape})

#HWC -> CHW

transformer.set\_transpose('data', (2,0,1))

#去均值

transformer.set\_mean('data', np.load('\*\_mean.npy').mean(1).mean(1))

# rescale -> [0, 255]

transformer.set\_raw\_scale('data', 255)

# RGB -> BGR

transformer.set\_channel\_swap('data', (2,1,0))

# 加载图片

im\_test = caffe.io.load\_image(caffe\_root+'examples/images/cat.jpg')

# 执行预处理

net.blobs['data'].data[...] = transformer.preprocess('data',im\_test)

# 执行测试

out = net.forward()

# 排序softmax(即类别概率),显示对应label

imagenet\_labels\_filename = caffe\_root + 'data/ilsvrc12/synset\_words.txt'

labels = np.loadtxt(imagenet\_labels\_filename, str, delimiter='\t')

top\_k = net.blobs['prob'].data[0].flatten().argsort()[-1:-6:-1]

for i in np.arange(top\_k.size):

print top\_k[i], labels[top\_k[i]]

**Interfaces--C++**

Caffe源代码

data 用于存放下载的训练数据

docs 帮助文档

example 一些代码样例

matlab MATLAB接口文件

python Python接口文件

model 一些配置好的模型参数

scripts 一些文档和数据用到的脚本

tools 保存的源码是用于生成二进制处理程序的，训练时直接调用这些二进制文件

include Caffe的实现代码的头文件

src 实现Caffe的源文件

util 数据转换时用的一些代码

proto 数据存储格式Protobuf

layers

caffe核心代码

blob[.cpp .h] 基本的数据结构Blob类。

common[.cpp .h] 定义Caffe类

internal\_thread[.cpp .h] 使用boost::thread线程库

net[.cpp .h] 网络结构类Net

solver[.cpp .h] 优化方法类Solver

data\_transformer[.cpp .h] 输入数据的基本操作类DataTransformer

syncedmem[.cpp .h] 分配内存和释放内存类CaffeMallocHost，用于同步GPU，CPU数据

layer\_factory.cpp layer.h 层类Layer

caffemodel 解析

<https://blog.csdn.net/zr459927180/article/details/50904938>

caffe.proto

caffe中，数据的读取、运算、存储都是采用Google Protocol Buffer (PB), PB是一种轻便、高效的结构化数据存储格式，可以用于结构化数据串行化，很适合做数据存储或 RPC 数据交换格式。它可用于通讯协议、数据存储等领域的语言无关、平台无关、可扩展的序列化结构数据格式。是一种效率和兼容性都很优秀的二进制数据传输格式，目前提供了 C++、Java、Python 三种语言的 API。Caffe采用的是C++和Python的API。caffemodel存储的数据也就是网络参数net\_param的PB

helloworld.proto

package lm

message helloworld

{

required int32 id = 1;

repeated string str = 2;

optional int32 opt = 3;

}

proto文件类似C++语言数据定义, Message等价于class

修饰符

required表明是必须包含该成员

optional表明可选成员

repeated等价于vector数组

-> PB编译器(protoc) -> c++

lm.helloworld.pb.h

lm.helloworld.pb.cc

#include "lm.helloworld.pb.h"

lm::helloworld msg1;

msg1.set\_id(101);

msg1.set\_str(“hello”);

fstream output("./log", ios::out | ios::trunc | ios::binary);

msg1.SerializeToOstream(&output)

fstream input("./log", ios::in | ios::binary);

msg1.ParseFromIstream(&input)

caffe 调试

<http://suanfazu.com/t/eclipse-caffe/13450>

<https://blog.csdn.net/mounty_fsc/article/details/51089864>

$nsight 打开 IDE

File -> New -> Project -> C/C++ -> Makefile Project with Existing Code

工程配置：

Project name： caffe

Existing code location：/home/qzlin/Documents/dl/caffe

Language：C and C++

Toolchain：Linux GCC

Run -> Debug Configurations -> C/C++ Application -> New

Name: RefineDet Defulat

Main: /home/qzlin/Documents/dl/RefineDet/build/tools/caffe

Arguments: train -solver=.../solver.prototxt

working directory: ...

编译： Project -> Projecties -> build command “make” 改为”make -j”

Project -> clean, Project -> Build All

注意：将Makefile.config注释掉 DEBUG :=1 这一行来打开调试信息并关闭优化(gcc flag -g -O0)

Caffe的整体流程图

<https://www.cnblogs.com/liuzhongfeng/p/7289956.html>

tools/caffe.cpp

void main(输入参数train -solover=solover.prototxt) {

GetBrewFunction("train")获取train()

}

int train() {

//从solver.prototxt读取solver\_param

SolverParameter solver\_param = ReadSolverParamsFromTextFileOrDie();

//从参数创建solve

Solver solver = CreateSolver(solver\_param);

// 从snapshot恢复模型

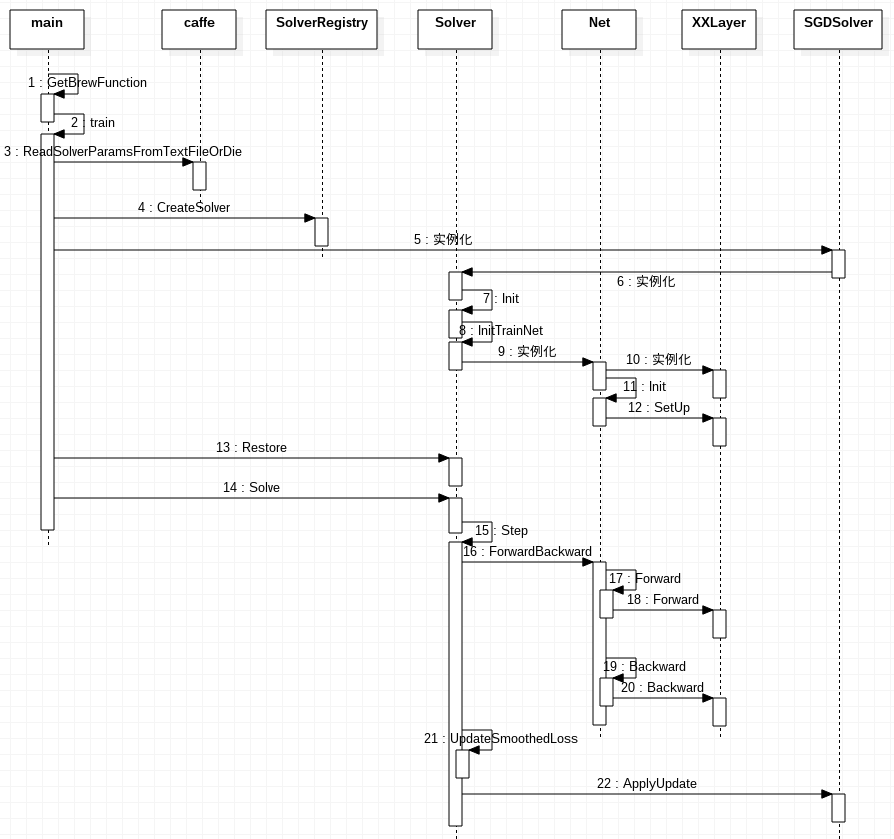
solver->Restore(snapshot);

//若采用finetuning，则拷贝weight到指定模型

CopyLayers(solver.get(), FLAGS\_weights);

solver->Solve();//开始训练网络

}



初始化总体流程

1. 新建一个Solver对象
2. 在Solver的构造函数中新建Net类实例
3. 在Net类的构造函数中新建各个layer实例

设置每个Blob

void Solver::InitTrainNet() {

// 获取NetParameter from SolverParameter

NetParameter net\_param;

net\_param.CopyFrom(param\_.net\_param());

ReadNetParamsFromTextFileOrDie(param\_.net(), &net\_param);

// 实例化net

net =new Net(net\_param);

}

void Net::Init(const NetParameter& in\_param) {

FilterNet(in\_param, &filtered\_param); //过滤校验参数FilterNet

InsertSplits(filtered\_param, &param);//插入Splits层

...// 构建网络中输入输出存储结构

for (int layer\_id = 0; layer\_id < param.layer\_size(); ++layer\_id) {

//创建层, 然后调用REGISTER\_LAYER\_CLASS(AnnotatedData);从而实例化AnnotatedDataLayer

layers.push\_back(LayerRegistry::CreateLayer(layer\_param));

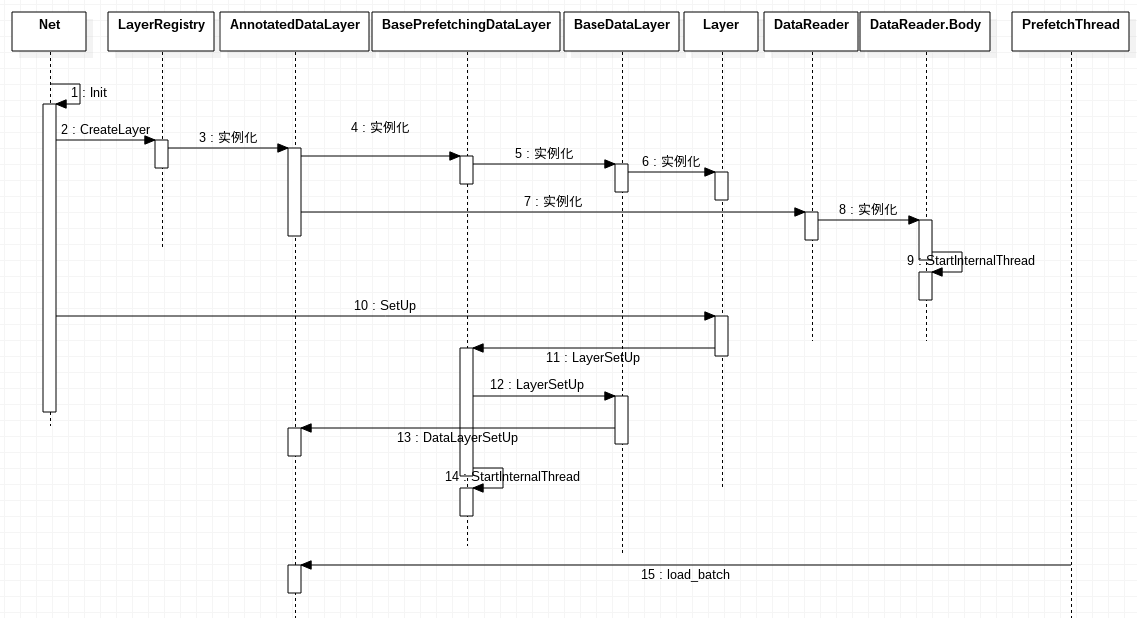
........//创建相关blob

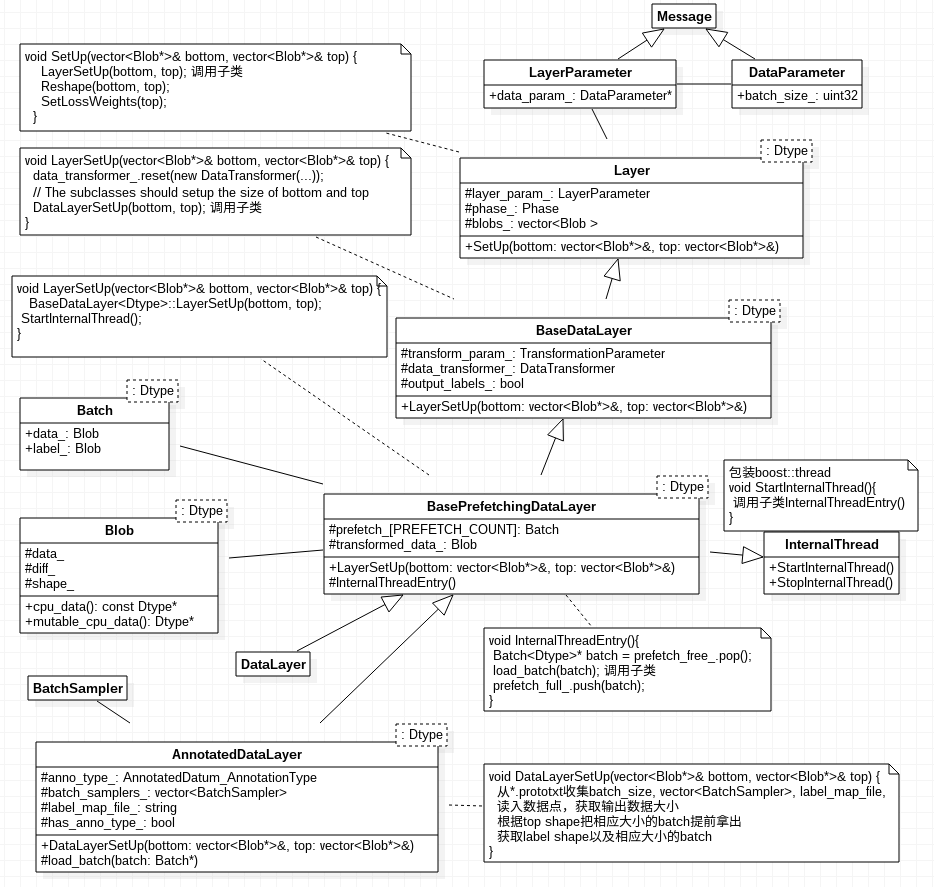
//执行SetUp()

layers\_[layer\_id]->SetUp(bottoms[layer\_id], tops[layer\_id]);

}

}





训练

void Solver::Step(int iters) {

while (iter\_ < stop\_iter) {

net\_->ClearParamDiffs(); // 将net\_中的Bolb梯度参数置为零

// accumulate the loss and gradient

Dtype loss = 0;

for (int i = 0; i < param\_.iter\_size(); ++i) {

loss += net\_->ForwardBackward(); // 正向传导和反向传导，并计算loss

}

loss /= param\_.iter\_size();

// 为了输出结果平滑，将临近的average\_loss个loss数值进行平均，存储在成员变量smoothed\_loss\_

UpdateSmoothedLoss(loss, start\_iter, average\_loss);

ApplyUpdate(); // BP算法更新权重

++iter\_;

}

}

void SGDSolver::ApplyUpdate() {

Dtype rate = GetLearningRate(); // 获取当前学习速率

// 在计算当前梯度的时候，如果该值超过了阈值clip\_gradients，则将梯度直接设置为该阈值，

//此处阈值设为-1，即不起作用

ClipGradients();

// 逐层更新网络中的可学习层

for (int id = 0; id < this->net\_->learnable\_params().size(); ++id) {

Normalize(id); // 归一化

Regularize(id); // L2范数正则化添加衰减权重

ComputeUpdateValue(id, rate); // 随机梯度下降法计算更新值

}

this->net\_->Update(); // 更新权重

}

Blobs

(N, C, H, W), 同时保存data and diff(梯度）

shared\_ptr<SyncedMemory> data\_ //数据

shared\_ptr<SyncedMemory> diff\_ //梯度

//重新修改Blob的形状(4维)，并根据形状来申请动态内存存储数据和梯度。

void Reshape(const int num, const int channels, const int height, const int width)

Layers

vector<Dtype> loss\_ ; //每一层都会有一个loss值，但只有LossLayer才会产生非0的loss

vector<shared\_ptr<Blob<Dtype> > > blobs\_ ; //Layer所学习的参数，包括权值和偏差

//通过bottom Blob和LayerParameter(从prototxt读入)确定Layer的学习参数的形状

void LayerSetUp(const vector<Blob<Dtype>\*>& bottom, const vector<Blob<Dtype>\*>& top)

//通过bottom Blob对象的形状以及Layer的学习参数的形状来确定top Blob对象的形状

void Reshape(const vector<Blob<Dtype>\*>& bottom, const vector<Blob<Dtype>\*>& top)

// Layer内部数据正向传播，从bottom到top方向

void Forward(const vector<Blob<Dtype>\*> &bottom, vector<Blob<Dtype>\*> \*top) = 0

// Layer内部梯度反向传播，从top到bottom方向

void Backward(const vector<Blob<Dtype>\*> &top,

const vector<bool> &propagate\_down,

vector<Blob<Dtype>\*> \*bottom) = 0

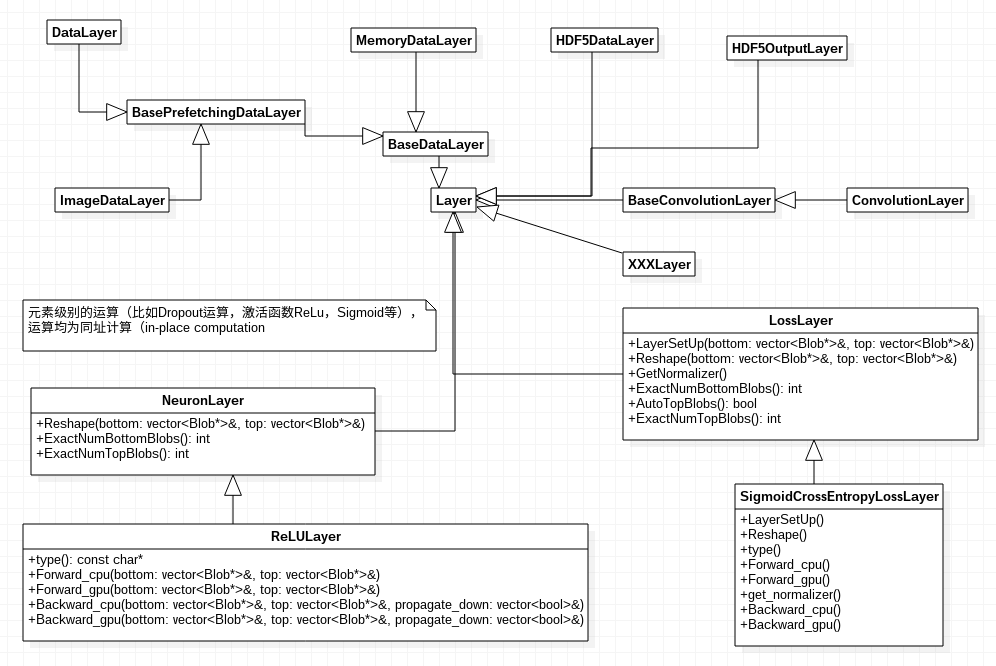
1.Data Layers

2.Vision Layers

3.Neuron Layers: 逐点运算层(如Activation)继承于Neuron Layer

4.Loss Layers

5.Common Layers (InnerProductLayer, SoftmaxLayer, Accuracy / Top-k layer , NormalizeLayer, RecurrentLayer, ...)直接继承于Layer



Net

用容器的形式将多个Layer有序地放在一起，主要功能是对逐层Layer进行初始化，以及提供Update( )的接口（更新网络参数），本身不能对参数进行有效地学习过程

vector<shared\_ptr<Layer<Dtype> > > layers\_; //构成该net的layers

vector<vector<Blob<Dtype>\*> > bottom\_vecs\_; //每一层layer中的bottom Blobs

vector<vector<Blob<Dtype>\*> > top\_vecs\_; //每一层layer中的top Blobs

vector<shared\_ptr<Blob<Dtype> > > params\_; //整个net中的learnable parameter

//根据NetParameter进行net初始化,简单的来说就是先把网络中所有层的bottom Blobs&top Blobs（无重复）实例化，并从输入层开始，逐层地进行Setup的工作，从而完成了整个网络的搭建，为后面的数据前后传输打下基础

void Init(const NetParameter& param);

//对整个网络的前向和方向传导，各调用一次就可以计算出网络的loss

vector<Blob\*>& Forward(const vector<Blob<Dtype>\* > & bottom,Dtype\* loss = NULL)

void Net<Dtype>::Backward()

Solver

Solver类中包含一个Net的指针，主要是实现了训练模型参数所采用的优化算法，根据优化算法的不同会派生不同的类，而基于这些子类就可以对网络进行正常的训练过程

shared\_ptr<Net<Dtype> > net\_; //net对象

//对已初始化后的网络进行固定次数的训练迭代过程

void Step(int iters)

// 不同的模型训练方法通过重载函数ComputeUpdateValue( )实现计算update参数的核心功能

ComputeUpdateValue();

net\_->Update();

自定义层

include/caffe/layers/your\_layer.hpp

src/caffe/layers/your\_layer.cpp

src/caffe/layers/your\_layer.cu (若实现Forward\_gpu,Backward\_gpu)

your\_layer继承common\_layer.hpp, data\_layer.hpp, loss\_layer.hpp, neuron\_layer.hpp, vision\_layer.hpp

// 为了在写net.prototxt时，layer{type："YourLayerName"}有所对应

virtual inline const char\* type() const { return "YourLayerName"; }

// 限制bottom和top的blob个数

virtual inline int ExactNumBottomBlobs() const { return 0; }

virtual inline int MinBottomBlobs() const { return -1; }

virtual inline int MaxBottomBlobs() const { return -1; }

virtual inline int ExactNumTopBlobs() const { return 1; }

virtual inline int MinTopBlobs() const { return -1; }

virtual inline int MaxTopBlobs() const { return -1; }

virtual inline bool EqualNumBottomTopBlobs() const { return false; }

//If true, Net::Init will create enough "anonymous" top blobs

virtual inline bool AutoTopBlobs() const { return false; }

// Layer初始化: 读取layer的参数，权重进行初始化等

virtual void LayerSetUp(vector<Blob\*>& bottom, vector<Blob\*>& top);

// layer初始化: 根据bottom的shape，修改top的shape

virtual void Reshape(vector<Blob\*>& bottom, vector<Blob\*>& top) = 0;

// 前向传播计算loss和top，反向传播计算diff(梯度)

Dtype Forward\_cpu(vector<Blob\*>& bottom, vector<Blob\*>& top);

void Backward\_cpu(...);

在.cpp文件末尾注册Layer，便于运行时统一创建

INSTANTIATE\_CLASS(XXXLayer);

REGISTER\_LAYER\_CLASS(XXX);

在src/caffe/proto/caffe.proto添加your\_layer的message

如果想要在net.prototxt中设置你的layer的参数的话，你需要在caffe.proto中定义

message AnnotatedDataParameter {

...

}

message LayerParameter {

optional AnnotatedDataParameter annotated\_data\_param = 200;

}

CAFFE\_ROOT$ make clean

CAFFE\_ROOT$ make -j

**举例：**

增加对应cpp、h、cu的声明和实现,修改caffe.proto文件，编译caffe库即可

1. **$CAFFE\_ROOT/include/caffe/layers/AllPassLayer.hpp**

#ifndef CAFFE\_ALL\_PASS\_LAYER\_HPP\_

#define CAFFE\_ALL\_PASS\_LAYER\_HPP\_

#include <vector>

#include "caffe/blob.hpp"

#include "caffe/layer.hpp"

#include "caffe/proto/caffe.pb.h"

#include "caffe/layers/neuron\_layer.hpp"

namespace caffe {

template <typename Dtype>

class AllPassLayer : public NeuronLayer<Dtype> {

public:

explicit AllPassLayer(const LayerParameter& param)

: NeuronLayer<Dtype>(param) {}

virtual inline const char\* type() const { return "AllPass"; }

protected:

virtual void Forward\_cpu(const vector<Blob<Dtype>\*>& bottom,

const vector<Blob<Dtype>\*>& top);

virtual void Forward\_gpu(const vector<Blob<Dtype>\*>& bottom,

const vector<Blob<Dtype>\*>& top);

virtual void Backward\_cpu(const vector<Blob<Dtype>\*>& top,

const vector<bool>& propagate\_down, const vector<Blob<Dtype>\*>& bottom);

virtual void Backward\_gpu(const vector<Blob<Dtype>\*>& top,

const vector<bool>& propagate\_down, const vector<Blob<Dtype>\*>& bottom);

};

} // namespace caffe

#endif // CAFFE\_ALL\_PASS\_LAYER\_HPP\_

1. **$CAFFE\_ROOT/src/caffe/layers/AllPassLayer.cpp**

#include <algorithm>

#include <vector>

#include "caffe/layers/all\_pass\_layer.hpp"

#include <iostream>

using namespace std;

#define DEBUG\_AP(str) cout<<str<<endl

namespace caffe {

template <typename Dtype>

void AllPassLayer<Dtype>::Forward\_cpu(const vector<Blob<Dtype>\*>& bottom,

const vector<Blob<Dtype>\*>& top) {

const Dtype\* bottom\_data = bottom[0]->cpu\_data();

Dtype\* top\_data = top[0]->mutable\_cpu\_data();

const int count = bottom[0]->count();

for (int i = 0; i < count; ++i) {

top\_data[i] = bottom\_data[i];

}

DEBUG\_AP("Here is All Pass Layer, forwarding.");

DEBUG\_AP(this->layer\_param\_.all\_pass\_param().key());

}

template <typename Dtype>

void AllPassLayer<Dtype>::Backward\_cpu(const vector<Blob<Dtype>\*>& top,

const vector<bool>& propagate\_down,

const vector<Blob<Dtype>\*>& bottom) {

if (propagate\_down[0]) {

const Dtype\* bottom\_data = bottom[0]->cpu\_data();

const Dtype\* top\_diff = top[0]->cpu\_diff();

Dtype\* bottom\_diff = bottom[0]->mutable\_cpu\_diff();

const int count = bottom[0]->count();

for (int i = 0; i < count; ++i) {

bottom\_diff[i] = top\_diff[i];

}

}

DEBUG\_AP("Here is All Pass Layer, backwarding.");

DEBUG\_AP(this->layer\_param\_.all\_pass\_param().key());

}

#ifdef CPU\_ONLY

STUB\_GPU(AllPassLayer);

#endif

INSTANTIATE\_CLASS(AllPassLayer);

REGISTER\_LAYER\_CLASS(AllPass);

} // namespace caffe

1. **$CAFFE\_ROOT/src/caffe/proto/caffe.proto**

message AllPassParameter {

optional float key = 1 [default = 0];

}

message LayerParameter {

...

optional AllPassParameter all\_pass\_param = 155;

}

1. **$CAFFE\_ROOT$ make clean && make all**

使用于deploy.prototxt

name: "AllPassTest"

layer {

name: "data"

type: "Input"

top: "data"

input\_param { shape: { dim: 10 dim: 3 dim: 227 dim: 227 } }

}

layer {

name: "ap"

type: "AllPass" #类名去掉 Layer 后的名称。

bottom: "data"

top: "conv1"

all\_pass\_param { key: 12.88 }

}

测试

$CAFFE\_ROOT$ ./build/tools/caffe.bin time -model deploy.prototxt

Caffe2

Install

1. Install Dependencies

sudo apt-get update

sudo apt-get install -y --no-install-recommends \

build-essential \

cmake \

git \

libgoogle-glog-dev \

libgtest-dev \

libiomp-dev \

libleveldb-dev \

liblmdb-dev \

libopencv-dev \

libopenmpi-dev \

libsnappy-dev \

libprotobuf-dev \

openmpi-bin \

openmpi-doc \

protobuf-compiler \

python-dev \

python-pip

sudo pip install \

future \

numpy \

protobuf

# for Ubuntu 16.04

sudo apt-get install -y --no-install-recommends libgflags-dev

1. Clone & Build

git clone --recursive https://github.com/caffe2/caffe2.git && cd caffe2

# This will build Caffe2 in an isolated directory so that Caffe2 source is

# unaffected

mkdir build && cd build

# This configures the build and finds which libraries it will include in the

# Caffe2 installation. The output of this command is very helpful in debugging

cmake ..

# This actually builds and installs Caffe2 from makefiles generated from the

# above configuration step

sudo make install

Note:

If download from github slowly, add IP of github into hosts (view gitcmd.txt)

If load github without complete, check and load submodule again

Caffe2$git submodule update --init --recursive

(refer to: <http://blog.csdn.net/u013553529/article/details/78307072)>

1. Environment Variables

$sudo gedit ~/.bashrc

echo $PYTHONPATH

# export PYTHONPATH=/usr/local:$PYTHONPATH

# export PYTHONPATH=$PYTHONPATH:/home/ubuntu/caffe2/build

echo $LD\_LIBRARY\_PATH

# export LD\_LIBRARY\_PATH=/usr/local/lib:$LD\_LIBRARY\_PATH

1. Test the Caffe2 Installation

cd ~ && python -c 'from caffe2.python import core' 2>/dev/null && echo "Success" || echo "Failure"

Install with GPU Support

same as keras

then test GPU works

python caffe2/python/operator\_test/relu\_op\_test.py

Note: ImportError: No module named hypothesis ##缺少hypothesis模块

$ pip install hypothesis

Tutorial

the primary idea of Caffe2 API: use Python to conveniently compose nets to train your model, pass those nets to C++ code as serialized protobuffers, and then let the C++ code run the nets with full performance.

caffe基本概念：

blob：存储数据和权值

layer：输入数据blob 形式，输出数据blob形式，层定义了计算

net：由多个layers组成，构成整体的网络

solver：定义了训练规则

caffe2基本概念

**Blob**：存储数据 （Data in Caffe2 is organized as blob, in Python they are translated to numpy arrays）

**Operator**：输入blob，输出blob，定义了计算规则

**Net**：由多个operator组合实现（A net is a graph of operators and each operator takes a set of input blobs and produces one or more output blobs）

**Workspace**：caffe中没有，可以理解成变量的空间，便于管理网络和变量（A Workspace stores all the blobs）

workspace是工作空间(可以把workspace理解成matlab中变量存储区)，在worspace中，可以存储网络结构类Net和数据存储类Blob.

from caffe2.python import core, workspace

workspace.Blobs()

workspace.HasBlob("X")

workspace.FeedBlob("X", X)

workspace.FetchBlob("X")

用多个名字定义多个workspace，并且可以切换工作空间。我们可以使用currentworkspace()在访问当前工作空间，使用switchworkspace(name)来切换工作空间

workspace.CurrentWorkspace()

workspace.SwitchWorkspace("another\_workspace", True)

**输入数据、权值、输出数据都存储在Blob中**

for name in workspace.Blobs():

print("{}:\n{}".format(name, workspace.FetchBlob(name)))

**Operator类用来定义来数据如何计算**

可以使用core.CreateOperator来直接创造

也可以使用core.Net来访问创建operator

还可以使用modelHelper来访问创建operators

**Net类由operator构成的计算图**

net = core.Net("my\_first\_net") #创建网络

X = net.GaussianFill([], ["X"], mean=0.0, std=1.0, shape=[2, 3], run\_once=0)

W = net.GaussianFill([], ["W"], mean=0.0, std=1.0, shape=[5, 3], run\_once=0)

b = net.ConstantFill([], ["b"], shape=[5,], value=1.0, run\_once=0)

Y = net.FC([X, W, b], ["Y"])

运行net

方法1:使用workspace.RunNetOnce,初始化网络，运行网络，然后销毁网络。

workspace.ResetWorkspace()

workspace.RunNetOnce(net)

方法2:先使用workspace.CreateNet初始化网络，然后使用workspace.RunNet来运行网络

workspace.ResetWorkspace()

workspace.CreateNet(net)

workspace.RunNet(net.Proto().name)

caffe2中一个模型通常包含一个初始化网络，一个训练网络

**RunNetOnce用来运行生成权值和数据的网络，常用于初始化，这样的网络一次生成完，权值输出或数据就存在当前的workspace中，网络本身就没有存在的必要了，就直接销毁**

RunNet可以用来重复训练网络，一开始使用CreateNet，不断迭代调用RunNet就可以不断运行网络更新参数了

加载预训练模型, 保存到caffe2/build/caffe2/python/models

python -m caffe2.python.models.download -i squeezenet

from caffe2.python.models import squeezenet

print(squeezenet.init\_net.name)

print(squeezenet.predict\_net.name)

注意：-i表示install a model, which allows later import the model directly in Python

若没有-i, 表示直接下载model，含init\_net.pb and predict\_net.pb文件

读入protobuf文件

with open("init\_net.pb") as f:

init\_net = f.read()

with open("predict\_net.pb") as f:

predict\_net = f.read()

p = workspace.Predictor(init\_net, predict\_net)

results = p.run([img])

**ModelHelper** class should only contain network definition and parameter information.

The **brew** module will have the functions to build network and initialize parameters.

方法1： Building your model/network using merely single operators

model = model\_helper.ModelHelper(name="train")

# initialize your weight

weight = model.param\_init\_net.XavierFill([], 'w', shape=[dim\_out, dim\_in])

# initialize your bias

bias = model.param\_init\_net.ConstantFill([], 'b', shape=[dim\_out, ])

# finally building FC

model.net.FC([blob\_in, weights, bias], blob\_out)

方法2：python/helpers/fc.py

fcLayer = fc(model, blob\_in, blob\_out, \*\*kwargs) # returns a blob reference

方法3（推荐）：from caffe2.python import brew

brew.fc(model, blob\_in, blob\_out, ...)

from caffe2.python import brew

def AddLeNetModel(model, data):

conv1 = brew.conv(model, data, 'conv1', 1, 20, 5)

pool1 = brew.max\_pool(model, conv1, 'pool1', kernel=2, stride=2)

conv2 = brew.conv(model, pool1, 'conv2', 20, 50, 5)

pool2 = brew.max\_pool(model, conv2, 'pool2', kernel=2, stride=2)

fc3 = brew.fc(model, pool2, 'fc3', 50 \* 4 \* 4, 500)

fc3 = brew.relu(model, fc3, fc3)

pred = brew.fc(model, fc3, 'pred', 500, 10)

softmax = brew.softmax(model, pred, 'softmax')

**arg\_scope** is a syntax sugar for you to set default helper function argument values within its context

with brew.arg\_scope([brew.conv], weight\_init=('XavierFill', {})):

brew.conv(model, ...) # no weight\_init needed here!

brew.conv(model, ...)

Caffe Model Files

.caffemodel and .pb: these are the models; they’re binary and usually large files

**.caffemodel**: from original Caffe

**.pb**: from Caffe2 and generally have init and predict together

**.pbtxt**: human-readable form of the Caffe2 pb file

**deploy.prototxt**: describes the network architecture for deployment (and not training) time

**solver.prototxt**: describes the variables used during training, including learning rates, regularization, etc.

**train\_val.prototxt**: describes the network architecture for training (and validation) time

Caffe2 DB

caffe2 使用二值 DB 存储模型训练的数据

假定数据集

(train\_features, train\_labels), (test\_features, test\_labels)

将numpy数据写入db中

def write\_db(db\_type, db\_name, features, labels):

db = core.C.create\_db(db\_type, db\_name, core.C.Mode.write)

transaction = db.new\_transaction()

for i in range(features.shape[0]):

feature\_and\_label = caffe2\_pb2.TensorProtos()

feature\_and\_label.protos.extend([

utils.NumpyArrayToCaffe2Tensor(features[i]),

utils.NumpyArrayToCaffe2Tensor(labels[i])])

transaction.put('train\_%03d'.format(i), feature\_and\_label.SerializeToString())

del transaction

del db

write\_db("minidb", "iris\_train.minidb", train\_features, train\_labels)

write\_db("minidb", "iris\_test.minidb", test\_features, test\_labels)

如何读取db数据

通过TensorProtosDBInput加载

def read\_db(db\_type, db):

net\_proto = core.Net("example\_reader")

dbreader = net\_proto.CreateDB([], "dbreader", db=db, db\_type=db\_type)

net\_proto.TensorProtosDBInput([dbreader], ["X", "Y"], batch\_size=16)

print(str(net\_proto.Proto()))

workspace.CreateNet(net\_proto)

# Let's run it to get batches of features.

for i in range(2):

workspace.RunNet(net\_proto.Proto().name)

print("The {}th batch of feature is:".format(i))

print(workspace.FetchBlob("X"))

print("The {}th batch of label is:".format(i))

print(workspace.FetchBlob("Y"))

read\_db("minidb", "iris\_train.minidb")

lmdb文件创建和读写见caffe2/python/examples/lmdb\_create\_example.py

结果会存成

mnist-train-nchw-lmdb

data.mdb

lock.mdb

如何读取?

# load the data

data\_uint8, label = model.TensorProtosDBInput(

[], ["data\_uint8", "label"], batch\_size=batch\_size,

db=db, db\_type=db\_type)

peek\_data(model)

def peek\_data(model):

workspace.RunNetOnce(model.param\_init\_net)

workspace.CreateNet(model.net)

workspace.RunNet(model.net.Proto().name)

data\_uint8 = workspace.FetchBlob('data\_uint8')

实际上，打包成caffe2 db就是为了直接作为模型的输入

1\*28\*28 -> 20\*24\*24 -> 20\*12\*12 -> 100\*8\*8 -> 100\*4\*4 -> 500 -> 10

def AddLeNetModel(model, data):

conv1 = brew.conv(model, data, 'c1', dim\_in=1, dim\_out=20, kernel=5)

pool1 = brew.max\_pool(model, conv1, 'p1', kernel=2, stride=2)

conv2 = brew.conv(model, pool1, 'c2', dim\_in=20, dim\_out=100, kernel=5)

pool2 = brew.max\_pool(model, conv2, 'p2', kernel=2, stride=2)

fc3 = brew.fc(model, pool2, 'fc3', dim\_in=100 \* 4 \* 4, dim\_out=500)

relu = brew.relu(model, fc3, fc3)

pred = brew.fc(model, relu, 'pred', 500, 10)

softmax = brew.softmax(model, pred, 'softmax')

return softmax

def AddTrainingOperators(model, softmax, label):

brew.accuracy(model, [softmax, label], "accuracy")

# compute the expected loss

xent = model.LabelCrossEntropy([softmax, label], 'xent')

loss = model.AveragedLoss(xent, "loss")

model.AddGradientOperators([loss])

# stochastic gradient descent

ITER = brew.iter(model, "iter")

LR = model.LearningRate(ITER, "LR", base\_lr=-0.1, policy="step", stepsize=1, gamma=0.999 )

ONE = model.param\_init\_net.ConstantFill([], "ONE", shape=[1], value=1.0)

# for each parameter, we do the gradient updates.

for param in model.params:

# Note how we get the gradient of each parameter - ModelHelper keeps track of that.

param\_grad = model.param\_to\_grad[param]

# The update is a simple weighted sum: param = param + param\_grad \* LR

model.WeightedSum([param, ONE, param\_grad, LR], param)

def AddBookkeepingOperators(model):

model.Print('accuracy', [], to\_file=1)

model.Print('loss', [], to\_file=1)

for param in model.params:

model.Summarize(param, [], to\_file=1)

model.Summarize(model.param\_to\_grad[param], [], to\_file=1)

训练

#建模型，注意输入是db输入

model = ModelHelper(name="mnist\_train", arg\_scope={"order": "NCHW"})

X, y = model.TensorProtosDBInput([], ["X", "y"], batch\_size=64,

db='mnist-train-nchw-lmdb', db\_type='lmdb')

softmax = AddLeNetModel(train\_model, X)

#训练方案: loss, 每个参数进行梯度更新

AddTrainingOperators(train\_model, softmax, label)

# 训练过程中的内部信息输出，便于诊断

AddBookkeepingOperators(train\_model)

# 初始化参数

workspace.RunNetOnce(train\_model.param\_init\_net)

# 迭代训练

total\_iters = 200

accuracy, loss = np.zeros(total\_iters), np.zeros(total\_iters)

workspace.CreateNet(train\_model.net, overwrite=True)

for i in range(total\_iters):

workspace.RunNet(train\_model.net)

accuracy[i] = workspace.FetchBlob('accuracy')

loss[i] = workspace.FetchBlob('loss')

可以对训练过程的accuracy and loss可视化

测试

对于测试模型，需要data, model, and accuracy, 注意模型千万不要初始化，即init\_params=False, 模型参数使用workspace里训练模型的同名参数

test\_model = ModelHelper(name="mnist\_test", arg\_scope=arg\_scope, **init\_params=False)**

data, label = **AddInput**(test\_model, batch\_size=100, db='mnist-test-nchw-lmdb', db\_type='lmdb')

注意：测试模型共享训练模型，通过workspace里的同名参数

softmax = **AddLeNetModel**(test\_model, data)

**AddAccuracy**(test\_model, softmax, label)

# run a test pass on the test net

workspace.RunNetOnce(test\_model.param\_init\_net)

workspace.CreateNet(test\_model.net, overwrite=True)

test\_accuracy = np.zeros(100)

for i in range(100):

workspace.RunNet(test\_model.net.Proto().name)

test\_accuracy[i] = workspace.FetchBlob('accuracy')

# deploy模型，仅需要模型部分

deploy\_model = ModelHelper(name="mnist\_deploy", arg\_scope=arg\_scope, **init\_params=False**)

**AddLeNetModel**(deploy\_model, "data")

# 注意输入输出需要指定

pe\_meta = **pe.PredictorExportMeta**(

predict\_net=deploy\_model.net.Proto(),

parameters=[str(b) for b in deploy\_model.params],

inputs=["data"],outputs=["softmax"],

)

# 保存模型

**pe.save\_to\_db**("minidb", "mnist\_model.minidb", pe\_meta)

# load the predict net

predict\_net = **pe.prepare\_prediction\_net**("mnist\_model.minidb", "minidb")

# 导入输入数据，进行预测

workspace.FeedBlob("data", blob)

# predict

**workspace.RunNetOnce(predict\_net)**

softmax = workspace.FetchBlob("softmax")

## Keras

<https://keras-cn.readthedocs.io/en/latest/> Keras的核心数据结构是“模型”, Keras的底层库使用Theano或TensorFlow(“符号主义”的库)

与传统的Python代码区别？

**符号主义的计算首先定义各种变量，然后建立“计算图”，计算图规定了各个变量之间的计算关系。建立好的计算图需要编译已确定其内部细节，然而，此时的计算图还是一个“空壳子”，里面没有任何实际的数据，只有当你把需要运算的输入放进去后，才能在整个模型中形成数据流，从而形成输出值**

深度学习的优化算法，说白了就是梯度下降。每次的参数更新有两种方式：

* Batch gradient descent（批梯度下降）：遍历全部数据集算一次损失函数，然后算函数对各个参数的梯度，更新梯度。缺点：计算量开销大，计算速度慢，不支持在线学习
* stochastic gradient descent（随机梯度下降）：速度快，但收敛性不好，可能在最优点附近晃来晃去，hit不到最优点。两次参数的更新也有可能互相抵消掉，造成目标函数震荡的比较剧烈
* mini-batch gradient decent（小批的梯度下降）：数据分为若干批，按批来更新参数，批中的一组数据共同决定了本次梯度的方向，下降起来就不容易跑偏，减少了随机性

张量可以看作是向量、矩阵的自然推广，我们用张量来表示广泛的数据类型

0阶张量，即标量，也就是一个数

1阶张量，也就是一个向量

2阶张量，也就是一个矩阵

3阶张量，一个立方体

张量的阶数有时候也称为维度，或者轴

Ubuntu 16.04 LTS是Nvidia官方以及绝大多数深度学习框架默认开发环境

### 安装

安装tensorflow-gpu

# 安装CUDA开发环境: nvidia driver, cuda toolkit, cuDNN

<https://docs.nvidia.com/cuda/cuda-installation-guide-linux/index.html>

<https://www.tensorflow.org/install/install_linux>

<http://www.python36.com/install-tensorflow-using-official-pip-pacakage/>

若已装有最新的版本，比如cuda9.1,需要remove，否则装不上

sudo apt-get purge nvidia\*

sudo apt-get auto-remove

重启机器

注意：cuda 9.1需要R390 nvidia driver

注意：cuda 9.0需要387 nvidia driver

$ cat /proc/driver/nvidia/version 查看nvidia版本号

sudo add-apt-repository ppa:graphics-drivers/ppa && sudo apt update

sudo apt-get install nvidia-390\*

or sudo apt-get install nvidia-387

Google: cuda tookit 9.0

CUDA Toolkit9.0

sudo dpkg -i cuda-repo-ubuntu1604-9-0-local\_9.0.176-1\_amd64.deb

sudo apt-key add /var/cuda-repo-9-0-local/7fa2af80.pub

sudo apt-get update

sudo apt-get install cuda-9.0

sudo dpkg -i libcudnn7\_7.0.5.15-1+cuda9.0\_amd64.deb

$ sudo gedit ~/.bashrc

export CUDA\_HOME="/usr/local/cuda-9.0"

export PATH="/usr/local/cuda-9.0/bin:$PATH"

export LD\_LIBRARY\_PATH="/usr/local/cuda-9.0/lib64:$LD\_LIBRARY\_PATH"

$ source ~/.bashrc

Old version

http://shomy.top/2016/12/29/gpu-tensorflow-install/

$ sudo dpkg -i cuda-repo-ubuntu1604-8-0-local\_8.0.44-1\_amd64.deb

$ sudo apt-key add /var/cuda-repo-8-0-local/7fa2af80.pub

$ sudo apt update

$ sudo apt install cuda

下载并解压cudnn v7版本：cudnn-8.0-linux-x64-v7.tgz

include -> /usr/local/cuda/include

lib64 -> /usr/local/cuda/lib64

# 将CUDA路径添加至环境变量

$ sudo gedit /etc/bash.bashrc

在bash.bashrc文件中添加：

export CUDA\_HOME=/usr/local/cuda-8.0

export PATH=/usr/local/cuda-8.0/bin${PATH:+:${PATH}}

export LD\_LIBRARY\_PATH=/usr/local/cuda-8.0/lib64${LD\_LIBRARY\_PATH:+:${LD\_LIBRARY\_PATH}}

$ source /etc/bash.bashrc

在.bashrc中添加如上相同内容

$ sudo gedit ~/.bashrc

$ source ~/.bashrc

$ nvcc -V 测试nVidia cuda版本号

/usr/local/cuda/samples$sudo make all –j8 测试是否成功，运行某个sample

若出现nvidia driver 不匹配，要安装对应nvidia-driver

安装Keras

# 系统升级

$ sudo apt update

$ sudo apt upgrade

# 安装python基础开发包

$ sudo apt install -y python-dev python-pip python-nose gcc g++ git gfortran

# 安装运算加速库

$ sudo apt install -y libopenblas-dev liblapack-dev libatlas-base-dev

#安装依赖包

$ sudo pip install pip setuptools wheel

$ sudo pip install numpy scipy matplotlib scikit-learn scikit-image

$ sudo pip install h5py pyyaml

$ sudo pip install theano

$ sudo pip install tensorflow-gpu

#安装keras

$ sudo pip install keras

(note: 或者下载源代码keras$ sudo python setup.py install)

Keras默认使用TensorFlow作为后端来进行张量操作,

若有tensorflow-gpu，直接使用gpu

**Note: tensorflow-gpu对cuda版本 and cuCNN版本有要求**

若cuda-8.0 and cndnn5.1,则tensorflow-gpu-1.2

若装错了

$sudo pip uninstall tensorflow-gpu

$sudo pip install -U--pre tensorflow-gpu-1.2

$ python 验证

>>> import tensorflow

>>> import keras

Keras环境设置

修改默认keras后端: gedit ~/.keras/keras.json

若配置theano文件: gedit ~/.theanorc

[global]

openmp=False

device = gpu

floatX = float32

allow\_input\_downcast=True

[lib]

cnmem = 0.8

[blas]

ldflags= -lopenblas

[nvcc]

fastmath = True

加速测试

keras/examples/$ python mnist\_mlp.py

注意:

下载的数据集统一放于$HOME/.keras/datasets/

下载的模型统一放于$HOME/.keras/models/

**重点概念**

张量

层

**多输入和多输出模型**



main\_input = Input(shape=(100,), dtype='int32', name='main\_input')

x = Embedding(output\_dim=512, ...)(main\_input)

lstm\_out = LSTM(32)(x)

auxiliary\_output = Dense(1, 'sigmoid', name='aux\_output')(lstm\_out)

auxiliary\_input = Input(shape=(5,), name='aux\_input')

x = keras.layers.concatenate([lstm\_out, auxiliary\_input])

for i in range(3)

x = Dense(64, activation='relu')(x)

x = Dense(64, activation='relu')(x)

x = Dense(64, activation='relu')(x)

main\_output = Dense(1, activation='sigmoid', name='main\_output')(x)

model = Model(inputs=[main\_input, auxiliary\_input],

**outputs=[main\_output, auxiliary\_output]**)

model.compile(optimizer='rmsprop',

**loss={'main\_output':** 'binary\_crossentropy',

**'aux\_output':** 'binary\_crossentropy'},

loss\_weights={'main\_output': 1.,

'aux\_output': 0.2})

model.fit({'main\_input': headline\_data, 'aux\_input': additional\_data},

**{'main\_output': labels, 'aux\_output': labels}**,

epochs=50, batch\_size=32)

**定制化层**

若无状态操作，建议用Lambda层

*Lambda(function, output\_shape=None, mask=None, arguments=None)*

本函数用以对上一层的输出施以任何Theano/TensorFlow表达式

# add a x -> x^2 layer

model.add(Lambda(lambda x: x \*\* 2))

若有状态操作（如含训练权重），则定制化层

from keras import backend as K

from keras.engine.topology import Layer

class MyLayer(Layer):

def \_\_init\_\_(self, output\_dim, \*\*kwargs):

self.output\_dim = output\_dim

super(MyLayer, self).\_\_init\_\_(\*\*kwargs)

def build(self, input\_shape): # 定制权重

self.kernel = ...

super(MyLayer, self).build(input\_shape)

def call(self, x): # 实现层的业务逻辑

return K.dot(x, self.kernel)

# 若层改变输出层shape,需要指明

def compute\_output\_shape(self, input\_shape):

return (input\_shape[0], self.output\_dim)

### 概念

**Keras模型= Sequential模型 / 函数式模型Model**

model.**summary**()：打印出模型概况

config = model.get\_config():返回包含模型配置信息的Python字典。

model = Model.from\_config(config): 模型也可以从它的config信息中重构回去

model.get\_layer()：依据层名或下标获得层对象

model.**get\_weights**()：返回模型权重张量的列表，类型为numpy array

model.**set\_weights**()：从numpy array里将权重载入给模型，要求数组具有与model.get\_weights()相同的形状。

json\_string =model.to\_json：返回代表模型的JSON字符串，仅包含网络结构，不包含权值。

model = model\_from\_json(json\_string): 可以从JSON字符串中重构原模型

model.save\_weights(filepath)：将模型权重保存到指定路径，文件类型是HDF5（后缀是.h5）

model.load\_weights(filepath, by\_name=False)：从HDF5文件中加载权重到当前模型中, 默认情况下模型的结构将保持不变。如果想将权重载入不同的模型（有些层相同）中，则设置by\_name=True，只有名字匹配的层才会载入权重

Sequential模型

model.layers: 是添加到模型上的层的list

add(self, layer): 向模型中添加一个层, Layer对象

**compile**(self, optimizer, loss, metrics=None, sample\_weight\_mode=None)

**fit**(...)

**predict**(…)

evaluate(...)

predict\_classes(self, x, batch\_size=32, verbose=1)

predict\_proba(self, x, batch\_size=32, verbose=1)

**train\_on\_batch**(self, x, y, class\_weight=None, sample\_weight=None)

**predict\_on\_batch**(self, x)

**fit\_generator**(...)

evaluate\_generator(...)

predict\_generator(…)

函数式模型Model

广义的拥有输入和输出的模型，用Model来初始化一个函数式模型

from keras.models import Model

model = Model(inputs=[a1, a2], outputs=[b1, b3, b3])

**model.layers：组成模型图的各个层**

**model.inputs：模型的输入张量列表**

**model.outputs：模型的输出张量列表**

其余api同Sequential模型

**输入是张量，输出也是张量的一个框架就是一个模型，通过Model定义。**

# This creates a model that includes the Input layer and three Dense layers

model = Model(inputs=inputs, outputs=predictions)

model.compile(optimizer='rmsprop',loss='categorical\_crossentropy', metrics=['accuracy'])

model.fit(data, labels) # starts training

**重用已经训练好的模型**

把模型当作一个层一样，通过提供一个tensor来调用它。注意当你调用一个模型时，你不仅仅重用了它的结构，也重用了它的权重。

x = Input(shape=(784,))

# This works, and returns the 10-way softmax we defined above.

y = model(x)

**预训练权重的Keras模型**

模型的预训练权重将下载到~/.keras/models/并在载入模型时自动载入,这些模型可以用来进行预测、特征提取和finetune

ResNet50模型: from keras.applications.resnet50 import ResNet50

VGG16模型: from keras.applications.vgg16 import VGG16

model = VGG16(weights='imagenet', include\_top=False)

img = image.load\_img(img\_path, target\_size=(224, 224))

x = image.img\_to\_array(img)

x = preprocess\_input(x)

features = model.predict(x)

VGG19模型: from keras.applications.vgg19 import VGG19

InceptionV3模型: from keras.applications.inception\_v3 import InceptionV3

Xception模型: from keras.applications.xception import Xception

模型可视化 (依赖 pydot-ng 和 graphviz)

from keras.utils import vis\_utils, plot\_model

plot\_model(model, to\_file='model.png')

在ipython中展示图片

from IPython.display import SVG

from keras.utils.visualize\_util import model\_to\_dot

SVG(model\_to\_dot(model).create(prog='dot', format='svg'))

**多输入和多输出模型**



main\_input = Input(shape=(100,), dtype='int32', name='main\_input')

x = Embedding(output\_dim=512, ...)(main\_input)

lstm\_out = LSTM(32)(x)

auxiliary\_output = Dense(1, 'sigmoid', name='aux\_output')(lstm\_out)

auxiliary\_input = Input(shape=(5,), name='aux\_input')

x = keras.layers.concatenate([lstm\_out, auxiliary\_input])

for i in range(3)

x = Dense(64, activation='relu')(x)

x = Dense(64, activation='relu')(x)

x = Dense(64, activation='relu')(x)

main\_output = Dense(1, activation='sigmoid', name='main\_output')(x)

model = Model(inputs=[main\_input, auxiliary\_input],

**outputs=[main\_output, auxiliary\_output]**)

model.compile(optimizer='rmsprop',

**loss={'main\_output':** 'binary\_crossentropy',

**'aux\_output':** 'binary\_crossentropy'},

loss\_weights={'main\_output': 1.,

'aux\_output': 0.2})

model.fit({'main\_input': headline\_data, 'aux\_input': additional\_data},

**{'main\_output': labels, 'aux\_output': labels}**,

epochs=50, batch\_size=32)

**张量Tensor**

from keras.layers import Input, Dense

from keras.models import Model

# This returns a **tensor**

inputs = Input(shape=(784,))

**层对象接受张量为参数，返回一个张量。**

# a layer instance is callable on a tensor, and returns a tensor

x = Dense(64, activation='relu')(inputs)

x = Dense(64, activation='relu')(x)

predictions = Dense(10, activation='softmax')(x)

**层Layer**

layer = Dense(32)

reconstructed\_layer = Dense.from\_config(config)层也可以借由配置信息重构

layer.get\_weights()：返回层的权重（numpy array）

layer.set\_weights(weights)：从numpy array中将权重加载到该层中，要求numpy array的形状与\* layer.get\_weights()的形状相同

config = layer.get\_config()：返回当前层配置信息的字典

获取层输入输出信息

非共享层

layer.input 输入张量

layer.output 输出张量

layer.input\_shape 输入数据的形状

layer.output\_shape 输出数据的形状

层有多个计算节点

layer.**get\_input\_at**(node\_index)

layer.**get\_output\_at**(node\_index)

layer.**get\_input\_shape\_at**(node\_index)

layer.**get\_output\_shape\_at**(node\_index)

**共享层**

把一个相同的Conv2D应用于一个大小为(3,32,32)的数据，然后又将其应用于一个(3,64,64)的数据，那么此时该层就具有了多个输入和输出的shape

a = Input(shape=(3, 32, 32))

b = Input(shape=(3, 64, 64))

conv = Conv2D(16, (3, 3), padding='same')

conved\_a = conv(a)

# Only one input so far, the following will work:

assert conv.input\_shape == (None, 3, 32, 32)

conved\_b = conv(b)

# now the `.input\_shape` property wouldn't work, but this does:

assert conv.get\_input\_shape\_at(0) == (None, 3, 32, 32)

assert conv.get\_input\_shape\_at(1) == (None, 3, 64, 64)

#### 常用层

* **常用的网络层**

**from keras.layers** import Dense, Dropout, Flatten, Reshape, Permute, RepeatVector, Lambda, ActivityRegularizer, Masking, GaussianNoise, GaussianDropout, AlphaDropout, BatchNormalization,TimeDistributed, Bidirectional

*Dense(units, activation, use\_bias, kernel\_initializer, bias\_initializer, kernel\_regularizer, bias\_regularizer, activity\_regularizer, kernel\_constraint, bias\_constraint)*

kernel\_initializer：权值初始化方法，为预定义初始化方法名的字符串，或用于初始化权重的初始化器。

bias\_initializer：权值初始化方法，为预定义初始化方法名的字符串，或用于初始化权重的初始化器。

kernel\_regularizer：施加在权重上的正则项，为Regularizer对象

bias\_regularizer：施加在偏置向量上的正则项，为Regularizer对象

activity\_regularizer：施加在输出上的正则项，为Regularizer对象

kernel\_constraints：施加在权重上的约束项，为Constraints对象

bias\_constraints：施加在偏置上的约束项，为Constraints对象

*Dropout(rate, noise\_shape=None, seed=None)*

Dropout将在训练过程中每次更新参数时随机断开一定百分比（rate）的输入神经元，Dropout层用于防止过拟合。

*Flatten()*

Flatten层用来将输入“压平”，即把多维的输入一维化

*Reshape(target\_shape)*

Reshape层用来将输入shape转换为特定的shape

# as first layer in a Sequential model

model = Sequential()

model.add(Reshape((3, 4), input\_shape=(12,)))

# now: model.output\_shape == (None, 3, 4)

# note: `None` is the batch dimension

# as intermediate layer in a Sequential model

model.add(Reshape((6, 2)))

# now: model.output\_shape == (None, 6, 2)

# also supports shape inference using `-1` as dimension

model.add(Reshape((-1, 2, 2)))

# now: model.output\_shape == (None, 3, 2, 2)

*Permute(dims)*

**Permute层将输入的维度按照给定模式进行重排**，例如，当需要将RNN和CNN网络连接时，可能会用到该层。

model = Sequential()

model.add(Permute((2, 1), input\_shape=(10, 64)))

# now: model.output\_shape == (None, 64, 10)

# note: `None` is the batch dimension

*RepeatVector(n)*

RepeatVector层将输入重复n次

model = Sequential()

model.add(Dense(32, input\_dim=32))

# now: model.output\_shape == (None, 32)

# note: `None` is the batch dimension

model.add(RepeatVector(3))

# now: model.output\_shape == (None, 3, 32)

*Lambda(function, output\_shape=None, mask=None, arguments=None)*

本函数用以对上一层的输出施以任何Theano/TensorFlow表达式

# add a x -> x^2 layer

model.add(Lambda(lambda x: x \*\* 2))

*ActivityRegularization(l1=0.0, l2=0.0)*

基于其激活值更新损失函数值,

l1：1范数正则因子（正浮点数）

l2：2范数正则因子（正浮点数）

*BatchNormalization*

**该层在每个batch上将前一层的激活值重新规范化，即使得其输出数据的均值接近0，其标准差接近1**

（1）加速收敛

（2）控制过拟合，可以少用或不用Dropout和正则

（3）降低网络对初始化权重不敏感

（4）允许使用较大的学习率

*GaussianNoise, GaussianDropout, AlphaDropout*

GaussianNoise(stddev)

为数据施加0均值，标准差为stddev的加性高斯噪声。该层在克服过拟合时比较有用，你可以将它看作是**随机的数据提升**。高斯噪声是需要对输入数据进行破坏时的自然选择。

一个使用噪声层的典型案例是构建去噪自动编码器，即Denoising AutoEncoder（DAE）。该编码器试图从加噪的输入中重构无噪信号，以学习到原始信号的鲁棒性表示

GaussianDropout(rate)

为层的输入施加以1为均值，标准差为sqrt(rate/(1-rate)的乘性高斯噪声

from keras.layers import *TimeDistributed, Bidirectional*

* **卷积层，反卷积，Cropping，UpSampling，ZeroPadding，Pooling**

**from keras.layers** import Conv1D, Conv2D, SeparableConv2D, Conv2DTranspose, Conv3D, Cropping1D, Cropping2D, Cropping3D, UpSampling1D, UpSampling2D, UpSampling3D, ZeroPadding1D, ZeroPadding2D, ZeroPadding3D

[http://deeplearning.net/software/theano\_versions/dev/tutorial/conv\_arithmetic.html#transposed-convolution-arithmetic](http://deeplearning.net/software/theano_versions/dev/tutorial/conv_arithmetic.html" \l "transposed-convolution-arithmetic)

一维卷积层（即时域卷积），用以在一维输入信号上进行邻域滤波

*Conv1D(...)*

二维卷积层，即对图像的空域卷积。该层对二维输入进行滑动窗卷积

*Conv2D*(filters, kernel\_size, strides=(1, 1), padding='valid', data\_format=None, dilation\_rate=(1, 1), activation=None, use\_bias=True, kernel\_initializer='glorot\_uniform', bias\_initializer='zeros', kernel\_regularizer=None, bias\_regularizer=None, activity\_regularizer=None, kernel\_constraint=None, bias\_constraint=None)

**filters：卷积核的数目**（即输出的维度）

**kernel\_size：卷积核的宽度和长度**。如为单个整数，则表示在各个空间维度的相同长度。

**strides：卷积的步长**。如为单个整数，则表示在各个空间维度的相同步长。

**padding：补0策略**，为“valid”, “same” 。“valid”代表只进行有效的卷积，“same”代表保留边界处的卷积结果

**activation：激活函数**

dilation\_rate：指定dilated convolution中的膨胀比例。

data\_format：字符串，“channels\_first”或“channels\_last”之一，代表图像的通道维的位置。该参数是Keras 1.x中的image\_dim\_ordering，**“channels\_last”对应原本的“tf”，“channels\_first”对应原本的“th”。**

use\_bias:布尔值，是否使用偏置项

kernel\_initializer：权值初始化方法，为预定义初始化方法名的字符串，或用于初始化权重的初始化器

bias\_initializer：权值初始化方法，为预定义初始化方法名的字符串，或用于初始化权重的初始化器

**kernel\_regularizer**：施加在权重上的正则项，为Regularizer对象

**bias\_regularizer**：施加在偏置向量上的正则项，为Regularizer对象

**activity\_regularizer**：施加在输出上的正则项，为Regularizer对象

kernel\_constraints：施加在权重上的约束项，为Constraints对象

bias\_constraints：施加在偏置上的约束项，为Constraints对象

输入shape

‘channels\_first’模式下，输入形如（samples, channels，rows，cols）的4D张量

**‘channels\_last’模式下，输入形如（samples，rows，cols，channels）的4D张量**

输出shape

‘channels\_first’模式下，为形如（samples，nb\_filter, new\_rows, new\_cols）的4D张量

**‘channels\_last’模式下，为形如（samples，new\_rows, new\_cols，nb\_filter）的4D张量**

输出的行列数可能会因为填充方法而改变

可分离卷积: SeparableConv2D(...)

**反卷积: Conv2DTranspose**(...)

*Conv3D*(...)

三维卷积对三维的输入进行滑动窗卷积，例如input\_shape = (3,10,128,128)代表对10帧128\*128的彩色RGB图像进行卷积

*Cropping1D*(cropping=(1, 1))

在时间轴（axis1）上对1D输入（即时间序列）进行裁剪

*Cropping2D*(cropping=((0, 0), (0, 0)), data\_format=None)

对2D输入（图像）进行裁剪，将在空域维度，即宽和高的方向上裁剪

cropping：长为2的整数tuple，分别为宽和高方向上头部与尾部需要裁剪掉的元素数

*Cropping3D*(...)

*UpSampling1D*(size=2): 在时间轴上，将每个时间步重复length次

*UpSampling2D*(size=(2, 2), data\_format=None)

**将数据的行和列分别重复size[0]和size[1]次**

*UpSampling3D*(size=(2, 2, 2), data\_format=None)

*ZeroPadding1D*(padding=1):对1D输入的首尾端（如时域序列）填充0

*ZeroPadding2D*(padding=(1, 1), data\_format=None)

对2D输入（如图片）的边界填充0，以控制卷积以后特征图的大小

**from keras.layers** import *MaxPooling1D, MaxPooling2D, MaxPooling3D, AveragePooling1D, AveragePooling2D, AveragePooling3D, GlobalMaxPooling1D, GlobalMaxPooling2D, GlobalAveragePooling2D*

* **局部连接层(locally-connected)，循环层(Recurrent),嵌入层(Embedding)**

from keras.layers import *LocallyConnected1D, LocallyConnected2D*

LocallyConnected1D层与Conv1D工作方式类似，唯一的区别是不进行权值共享。即施加在不同输入位置的滤波器是不一样的。

# apply a 3x3 unshared weights convolution with 64 output filters on a 32x32 image

# with `data\_format="channels\_last"`:

model = Sequential()

model.add(LocallyConnected2D(64, (3, 3), input\_shape=(32, 32, 3)))

# now model.output\_shape == (None, 30, 30, 64)

# notice that this layer will consume (30\*30)\*(3\*3\*3\*64) + (30\*30)\*64 parameters

from keras.layers.recurrent import *RNN, SimpleRNN, GRU, LSTM, ConvLSTM2D, SimpleRNNCell, GRUCell, LSTMCell, CuDNNGRU, CuDNNLSTM*

SimpleRNN(...): 全连接RNN网络

GRU(...): 门限循环单元

LSTM(...): 长短期记忆模型

* **嵌入层(Embedding), 融合层(Merge),激活层(Activation), 规范层(BatchNormalization), 噪声层(Noise),包装器(Wrapper)**

from keras.layers import *Embedding*

from keras.layers import *（类）Add, Substract, Multiply, Average, Maximum, Concatenate****,*** *Dot, (函数式接口）add, substract, multiply, average, maximum, concatenate, dot*

Merge层提供了一系列用于融合两个层或两个张量的层对象和方法。以大写首字母开头的是Layer类，以小写字母开头的是张量的函数。小写字母开头的张量函数在内部实际上是调用了大写字母开头的层。

Add(): 将Layer that adds a list of inputs.

Multiply(): 逐元素积的张量

Average(): 逐元素均值

Maximum(): 逐元素最大值

Concatenate(axis=-1): 按照给定轴相接构成的向量。

Dot(axes, normalize=False):两张量乘积

例如，如果两个张量a和b的shape都为（batch\_size, n），则输出为形如（batch\_size,1）的张量，结果张量每个batch的数据都是a[i,:]和b[i,:]的矩阵（向量）点积。

normalize: 布尔值，是否沿执行成绩的轴做L2规范化，如果设为True，那么乘积的输出是两个样本的余弦相似性

Add()([X1, X2]) 类使用

add([X1, X2]) 函数使用

* **激活层**

from keras.layers import *Activation*

预定义激活函数

**softmax：对输入数据的最后一维进行softmax**

elu

softplus

softsign

**relu**

tanh

**sigmoid**

hard\_sigmoid

linear

可以通过传递一个逐元素运算的Theano/TensorFlow函数来作为激活函数

from keras import backend as K

def tanh(x):

return K.tanh(x)

model.add(Dense(64, activation=tanh))

model.add(Activation(tanh)

高级激活函数

from keras.layers import *LeakyReLU, PReLU, ELU, ThresholdedReLU, Softmax*

当不激活时，LeakyReLU仍然会有非零输出值，从而获得一个小梯度，避免ReLU可能出现的神经元“死亡”现象

* **定制化层**

若无状态操作，建议用Lambda层

*Lambda(function, output\_shape=None, mask=None, arguments=None)*

本函数用以对上一层的输出施以任何Theano/TensorFlow表达式

# add a x -> x^2 layer

model.add(Lambda(lambda x: x \*\* 2))

若有状态操作（如含训练权重），则定制化层

from keras import backend as K

from keras.engine.topology import Layer

class MyLayer(Layer):

def \_\_init\_\_(self, output\_dim, \*\*kwargs):

self.output\_dim = output\_dim

super(MyLayer, self).\_\_init\_\_(\*\*kwargs)

def build(self, input\_shape): # 定制权重

self.kernel = ...

super(MyLayer, self).build(input\_shape)

def call(self, x): # 实现层的业务逻辑

return K.dot(x, self.kernel)

# 若层改变输出层shape,需要指明

def compute\_output\_shape(self, input\_shape):

return (input\_shape[0], self.output\_dim)

#### 数据预处理

**from keras.preprocessing.image import ImageDataGenerator**

ImageDataGenerator(featurewise\_center=False,

samplewise\_center=False,

featurewise\_std\_normalization=False,

samplewise\_std\_normalization=False,

zca\_whitening=False,

rotation\_range=0.,

width\_shift\_range=0.,

height\_shift\_range=0.,

shear\_range=0.,

zoom\_range=0.,

channel\_shift\_range=0.,

fill\_mode='nearest',

cval=0.,

horizontal\_flip=False,

vertical\_flip=False,

rescale=None,

**preprocessing\_function**=None,

data\_format=K.image\_data\_format())

用以生成一个batch的图像数据，支持实时数据提升。训练时该函数会无限生成数据，直到达到规定的epoch次数为止。

featurewise\_center：布尔值，**使输入数据集去中心化**（均值为0）, 按feature执行

featurewise\_std\_normalization：布尔值，将输入除以数据集的标准差以完成标准化, 按feature执行

rotation\_range：整数，数据提升时图片随机转动的角度

width\_shift\_range：浮点数，图片宽度的某个比例，数据提升时图片水平偏移的幅度

height\_shift\_range：浮点数，图片高度的某个比例，数据提升时图片竖直偏移的幅度

shear\_range：浮点数，剪切强度（逆时针方向的剪切变换角度）

zoom\_range：浮点数或形如[lower,upper]的列表，随机缩放的幅度，若为浮点数，则相当于[lower,upper] = [1 - zoom\_range, 1+zoom\_range]

fit(x, augment=False, rounds=1)：计算依赖于数据的变换所需要的统计信息(均值方差等),只有使用featurewise\_center，featurewise\_std\_normalization或zca\_whitening时需要此函数。

flow(self, X, y, batch\_size=32, shuffle=True, seed=None, ...)：接收numpy数组和标签为参数,生成经过数据提升或标准化后的batch数据,并在一个无限循环中不断的返回batch数据

(x\_train, y\_train), (x\_test, y\_test) = cifar10.load\_data()

y\_train = np\_utils.to\_categorical(y\_train, num\_classes)

y\_test = np\_utils.to\_categorical(y\_test, num\_classes)

**datagen = ImageDataGenerator**(

featurewise\_center=True,

featurewise\_std\_normalization=True,

rotation\_range=20,

width\_shift\_range=0.2,

height\_shift\_range=0.2,

horizontal\_flip=True)

# compute quantities required for featurewise normalization

# (std, mean, and principal components if ZCA whitening is applied)

datagen.fit(x\_train)

# fits the model on batches with real-time data augmentation:

**model.fit\_generator(datagen.flow**(x\_train, y\_train, batch\_size=32),

steps\_per\_epoch=len(x\_train), epochs=epochs)

同时变换图像和mask

# we create two instances with the same arguments

data\_gen\_args = dict(featurewise\_center=True,

featurewise\_std\_normalization=True,

rotation\_range=90.,

width\_shift\_range=0.1,

height\_shift\_range=0.1,

zoom\_range=0.2)

image\_datagen = ImageDataGenerator(\*\*data\_gen\_args)

mask\_datagen = ImageDataGenerator(\*\*data\_gen\_args)

# Provide the same seed and keyword arguments to the fit and flow methods

seed = 1

image\_datagen.fit(images, augment=True, seed=seed)

mask\_datagen.fit(masks, augment=True, seed=seed)

image\_generator = image\_datagen.flow\_from\_directory(

'data/images',

class\_mode=None,

seed=seed)

mask\_generator = mask\_datagen.flow\_from\_directory(

'data/masks',

class\_mode=None,

seed=seed)

# combine generators into one which yields image and masks

train\_generator = zip(image\_generator, mask\_generator)

model.fit\_generator(train\_generator, steps\_per\_epoch=2000, epochs=50)

#### 性能评估

**from keras import losses**

真实的优化目标函数是在各个数据点得到的损失函数值之和的均值

可用的目标函数:

mean\_squared\_error或mse

mean\_absolute\_error或mae

mean\_absolute\_percentage\_error或mape

mean\_squared\_logarithmic\_error或msle

squared\_hinge

hinge

**binary\_crossentropy**（亦称作对数损失，logloss）

logcosh

**categorical\_crossentropy：亦称作多类的对数损失，注意使用该目标函数时，需要将标签转化为形如(nb\_samples, nb\_classes)的二值序列**

sparse\_categorical\_crossentrop：如上，但接受稀疏标签

kullback\_leibler\_divergence:从预测值概率分布Q到真值概率分布P的信息增益,用以度量两个分布的差异.

poisson：即(predictions - targets \* log(predictions))的均值

cosine\_proximity：即预测值与真实标签的余弦距离平均值的相反数

**from keras import optimizers**

optimizers.**SGD**(lr=0.01, momentum=0.0, decay=0.0, nesterov=False)

随机梯度下降法，支持动量参数，支持学习衰减率，支持Nesterov动量

optimizers.**RMSprop**(lr=0.001, rho=0.9, epsilon=1e-06)

除学习率可调整外，建议保持优化器的其他默认参数不变,递归神经网络时的一个良好选择

optimizers.Adagrad(lr=0.01, epsilon=1e-06): 建议保持优化器的默认参数不变

Adadelta(lr=1.0, rho=0.95, epsilon=1e-06): 建议保持优化器的默认参数不变

optimizers.**Adam**(lr=0.001, beta\_1=0.9, beta\_2=0.999, epsilon=1e-08):该优化器的默认值来源于参考文献

Nadam(lr=0.002, beta\_1=0.9, beta\_2=0.999, epsilon=1e-08, schedule\_decay=0.004)

**Adam本质上像是带有动量项的RMSprop，Nadam就是带有Nesterov 动量的Adam RMSprop**

**from keras import metrics**

metrics=[metrics.mae, metrics.categorical\_accuracy]

categorical\_accuracy:对多分类问题,计算所有预测值上的平均正确率

**权重初始化**

**from keras import initializers**

不同的层可能使用不同的关键字来传递初始化方法

Dense(64, kernel\_initializer='random\_uniform', bias\_initializer='zeros')

kernel\_initializer=initializers.random\_normal(stddev=0.01)

kernel\_initializer='random\_normal'

预定义初始化方法

initializers.Zeros()

initializers.Ones()

Constant(value=0)

RandomNormal(mean=0.0, stddev=0.05, seed=None))

RandomUniform(minval=-0.05, maxval=0.05, seed=None)

TruncatedNormal(mean=0.0, stddev=0.05, seed=None):截尾高斯分布初始化，该初始化方法与RandomNormal类似，但位于均值两个标准差以外的数据将会被丢弃并重新生成，形成截尾分布。该分布是神经网络权重和滤波器的推荐初始化方法。

VarianceScaling(scale=1.0, mode='fan\_in', distribution='normal', seed=None):自适应目标张量的shape。

Orthogonal(gain=1.0, seed=None):用随机正交矩阵初始化

Identity(gain=1.0)

lecun\_uniform(seed=None): LeCun均匀分布初始化方法

glorot\_normal(seed=None):Glorot正态分布初始化方法，也称作Xavier正态分布初始化

he\_normal(seed=None):He正态分布初始化方法，也称作Xavier正态分布初始化

he\_uniform(seed=None)

自定义初始化器

from keras import backend as K

def my\_init(shape, dtype=None):

return K.random\_normal(shape, dtype=dtype)

model.add(Dense(64, init=my\_init))

#### 正则项，约束项

**正则项在优化过程中层的参数或层的激活值添加惩罚项，这些惩罚项将与损失函数一起作为网络的最终优化目标**

惩罚项基于层进行惩罚，目前惩罚项的接口与层有关，但Dense, Conv1D, Conv2D, Conv3D具有共同的接口。

kernel\_regularizer：施加在权重上的正则项，为keras.regularizer.Regularizer对象

bias\_regularizer：施加在偏置向量上的正则项，为keras.regularizer.Regularizer对象

activity\_regularizer：施加在输出上的正则项，为keras.regularizer.Regularizer对象

**from keras import regularizers**

model.add(Dense(64, input\_dim=64, kernel\_regularizer=regularizers.l2(0.01),

activity\_regularizer=regularizers.l1(0.01)))

开发新的正则项

任何以权重矩阵作为输入并返回单个数值的函数均可以作为正则项

from keras import backend as K

def l1\_reg(weight\_matrix):

return 0.01 \* K.sum(K.abs(weight\_matrix))

Dense(64, input\_dim=64, kernel\_regularizer=l1\_reg)

约束项

在优化过程中为网络的参数施加约束

惩罚项基于层进行惩罚，目前惩罚项的接口与层有关，但Dense, Conv1D, Conv2D, Conv3D具有共同的接口。

kernel\_constraint：对主权重矩阵进行约束

bias\_constraint：对偏置向量进行约束

from keras.constraints import maxnorm

Dense(64, kernel\_constraint=max\_norm(2.))

预定义约束项

max\_norm(m=2)：最大模约束

non\_neg()：非负性约束

unit\_norm()：单位范数约束, 强制矩阵沿最后一个轴拥有单位范数

#### 回调函数Callbacks

**回调函数是一组在训练的特定阶段被调用的函数集**，你可以使用回调函数来观察训练过程中网络内部的状态和统计信息。

回调函数以字典logs为参数，该字典包含了一系列与当前batch或epoch相关的信息。

在每个epoch的结尾处（on\_epoch\_end），logs将包含训练的正确率和误差，acc和loss，如果指定了验证集，还会包含验证集正确率和误差val\_acc和val\_loss，val\_acc还额外需要在.compile中启用metrics=['accuracy']。

在每个batch的开始处（on\_batch\_begin）：logs包含size，即当前batch的样本数

在每个batch的结尾处（on\_batch\_end）：logs包含loss，若启用accuracy则还包含acc

**from keras import callbacks**

callbacks.BaseLogger(): 对每个epoch累加metrics指定的监视指标的epoch平均值, 在Keras模型中会被自动调用

callbacks.History(): 在Keras模型上会被自动调用，History对象即为fit方法的返回值

callbacks.ProgbarLogger(): 将metrics指定的监视指标输出到标准输出上

**callbacks.ModelCheckpoint(filepath ...): 在每个epoch后保存模型到filepath**

**callbacks.EarlyStopping(...): 当监测值不再改善时，该回调函数将中止训练**

callbacks.RemoteMonitor(root='http://localhost:9000'):向服务器发送事件流，该回调函数需要requests库

callbacks.TensorBoard(log\_dir='./logs', ...): 可视化的展示器

**callbacks.ReduceLROnPlateau(monitor='val\_loss', factor=0.1,...):当评价指标不在提升时，减少学习率**

callbacks.CSVLogger(filename, separator=',', append=False):将epoch的训练结果保存在csv文件中

编写自己的回调函数

class LossHistory(keras.callbacks.Callback):

def on\_train\_begin(self, logs={}):

self.losses = []

def on\_batch\_end(self, batch, logs={}):

self.losses.append(logs.get('loss'))

history = LossHistory()

model.fit(X\_train, Y\_train, batch\_size=128, epochs=20, verbose=0, callbacks=[history])

print history.losses

keras.utils提供的使得方法

x\_data = keras.utils.io\_utils.HDF5Matrix('input/file.hdf5', 'data')

model.predict(x\_data)

keras.utils.to\_categorical(y, num\_classes=None)

keras.utils.normalize(x, axis=-1, order=2)

x：待规范化的数据

axis: 规范化的轴

order：规范化方法，如2为L2范数

#### Keras后端

Keras是一个模型级的库，提供了快速构建深度学习网络的模块。Keras并不处理如张量乘法、卷积等底层操作。这些操作依赖于某种特定的、优化良好的张量操作库。**Keras依赖于处理张量的库就称为“后端引擎”**。Keras提供了两种后端引擎Theano/Tensorflow

Keras的配置文件$HOME/.keras/keras.json

文件的默认配置如下：

{

"image\_data\_format": "channels\_last",

"epsilon": 1e-07,

"floatx": "float32",

"backend": "tensorflow"

}

也可以通过定义环境变量KERAS\_BACKEND来覆盖上面配置文件中定义的后端

使用抽象的Keras后端来编写代码

from keras import backend as K

**定义tensor**

#tf.placeholder() ，T.matrix()，T.tensor3()

input = K.placeholder(shape=(2, 4, 5))

**#共享变量（shared），等价于tf.variable()**或 theano.shared()

val = np.random.random((3, 4, 5))

var = K.variable(value=val)

var = K.zeros(shape=(3, 4, 5)) # all-zeros variable:

var = K.ones(shape=(3, 4, 5)) # all-ones

K.eye(size, dtype='float32', name=None)

**大多数你需要的张量操作都可以通过统一的Keras后端接口完成**，而不关心具体执行这些操作的是Theano还是TensorFlow

a = b + c \* K.abs(d)

c = K.dot(a, K.transpose(b))

a = K.sum(b, axis=2)

a = K.softmax(b)

eps = K.epsilon()

K.set\_epsilon(1e-05)

K.floatx()

K.set\_floatx('float16')

data\_format = K.image\_data\_format() #‘channels\_last’或‘channels\_first’

K.set\_image\_data\_format(data\_format)

K.is\_keras\_tensor(input)

zeros\_like, ones\_like, random\_uniform\_variable, max, min, sum, prod, cumsum, cumprod, var, std, mean, any, all, argmax, square, abs, sqrt, exp, log, logsumexp, round, sign, pow, clip, equal, sin, cos, concatenate, reshape, permute\_dimensions, resize\_images, resize\_volumes, repeat\_elements, repeat, arange, tile, batch\_flatten, expand\_dims, squeeze,random\_normal, random\_uniform,random\_binomial,truncated\_normall

normalize\_batch\_in\_training(): 对一个batch数据先计算其均值和方差，然后再进行batch\_normalization

batch\_normalization(): output = (x-mean)/(sqrt(var)+epsilon)\*gamma+beta

one\_hot(indices, nb\_classes): 输出为(n+1)维的one-hot编码

get\_value(x): 以Numpy array的形式返回张量的值

gradients(loss, variables): 返回loss函数关于variables的梯度，variables为张量变量的列表

relu(x, alpha=0.0, max\_value=None)

elu(x, alpha=1.0)

softmax(x)

softplus(x)

softsign(x)

sigmoid(x)

hard\_sigmoid(x)

tanh(x)

categorical\_crossentropy(output, target, from\_logits=False): 计算输出张量和目标张量的Categorical crossentropy（类别交叉熵）

l2\_normalize(x, axis)

dropout(x, level, seed=None)

conv1d(...)

conv2d(...)

deconv2d(...)

conv3d(...)

pool2d(...)

pool3d(...)

bias\_add(x, bias, data\_format=None)

**Key points:**

* 数据集

A batch generally approximates the distribution of the input data better than a single input. The larger the batch, the better the approximation

大数据集，内存放不下，如何处理？

model.train\_on\_batch(x, y)

model.test\_on\_batch(x, y)

model.fit\_generator(data\_generator, steps\_per\_epoch, epochs)

小数据集，如何解决overfitting?

blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html

模型太大，不仅学到共性特征，也学到训练集独有的特征，可以添加regular, dropout等解决

Data augmentation，使model训练时见到没有同样的case

Transfer 学习， 用相近的更大规模训练，然后freeze 低层，只对top层进行训练

开发阶段，如何避免性能改变是由于随机改变，还是模型提升？

keras.io/getting-started/faq/

import os

os.environ['PYTHONHASHSEED'] = '0'

np.random.seed(42)

rn.seed(12345)

* 模型 f(input\_shape, n\_class)

Model 需要指定输入维数（不需要包含batch\_size）, 由input tensor and output tensor定义

可以将model视为layers，输入tensor, 输出tensor

input\_shape = (w, h, channel)

model = Model(Input(input\_shape), output)

Y\_predict = model(X\_train)

注意，同样地，所有层，不考虑batch\_size

比如

h1 = Dense(32, input\_shape=(16, )) 输入(\*, 16), 输出(\*, 32)

c1 = Conv2D(64, 3, 3)(pool1) 若pool1(\*, 32, 32, 3), 输出(\*, 32, 32, 64)

f1 = Flatten()(c1) 输出(\*, 32\*32\*64)

r1 = Reshape((32, 32, 64))(f1) 输出(\*, 32, 32, 64)

Permute((1, 3))(r1) 输出(\*, 64, 32, 32)

RepeatVector(3)(h1) 输出(\*, 3, 32)

自定义层

方法一：若无状态操作，采用layers.core.Lambda

若输出维数不变：model.add(Lambda x: x\*\*2)

若输出维数改变：

def antirectifier(x):

pos = K.relu(x)

neg = K.relu(-x)

return K.concatenate([pos, neg], axis=1)

def output\_shape(input\_shape):

shape = list(input\_shape)

shape[-1] \*= 2

return tuple(shape)

model.add(Lambda(antirectifier, output\_shape=output\_shape))

方法二：若有状态操作，如含权重，需要继承Layer

**from keras import backend as K**

**from keras.engine.topology import Layer**

import numpy as np

#阅读源代码

class Conv2D(**Layer**):

def \_\_init\_\_(self, kernel\_size=(3,3), filters=32, \*\*kwargs):

self.kernel\_size = kernel\_size

self.filters = filters

super(MyLayer, self).\_\_init\_\_(\*\*kwargs)

#**定义权重**的方法 # Create a trainable weight variable for this layer.

**def build(self, input\_shape):**

channels\_axis = -1 if self.data\_format==’channels\_last’ else 1

input\_dim = input\_shape[channels\_axis]

kernel\_shape = self.kernel\_size + (input\_dim, self.filters)

self.kernel = self.add\_weight(shape=kernel\_shape, initializer='uniform')

super(MyLayer, self).build(input\_shape)

**#定义层功能的方法**

**def call(self, inputs):**

return K.con2d(inputs, self.kernel)

#**指定shape**变化的方法

**def compute\_output\_shape(self, input\_shape):**

return (input\_shape[0], input\_shape[1], self.filters )

保存/导入模型Saving/loading whole models (architecture + weights + optimizer state)

from keras.models import load\_model

model.save('my\_model.h5') # creates a HDF5 file 'my\_model.h5'

del model # deletes the existing model

# returns a compiled model

# identical to the previous one

model = load\_model('my\_model.h5')

* 训练

若指标没有随训练下降，停止学习

from keras.callbacks import EarlyStopping

early\_stopping = EarlyStopping(monitor='val\_loss', patience=2)

model.fit(x, y, validation\_split=0.2, callbacks=[early\_stopping])

* 正则化：

mnist 60000训练集，加了reg, acc从99.11 → 99.12

Conv2D(32, (5, 5), activation='relu',

kernel\_regularizer=l2(weight\_decay),

bias\_regularizer=l2(weight\_decay),

activity\_regularizer=l2(weight\_decay))

* 不平衡类：

mnist 60000训练集，加了class\_weight, acc从99.12 → 99.19

class\_weight = {0 : 1., 1: 50., 2: 2.} 对应类权重

model.fit(X\_train, Y\_train， class\_weight = class\_weight)

计算类权重，越少类权重越大

from sklearn.utils.class\_weight import compute\_class\_weight

class\_weight = compute\_class\_weight('balanced', np.unique(y\_train), y\_train)

model.fit(X\_train, y\_train, class\_weight=class\_weight)

熵作为类权重

log(1 / yi) 其中yi ＝ count / (mu\*total), mu = 0.15

# labels\_dict : {ind\_label: count\_label}

# mu : parameter to tune

def create\_class\_weight(labels\_dict, mu=0.15):

total = np.sum(labels\_dict.values())

class\_weight = {}

for key in labels\_dict.keys():

score = math.log(mu\*total/float(labels\_dict[key]))

class\_weight[key] = score if score > 1.0 else 1.0

return class\_weight

* 损失函数 f(target, output)

**input have shape [batch\_size, w, h, 1]**

**output have shape [batch\_size, 1]**

target = K.reshape(target, [-1, w\*h\*1])

output = K.reshape(output, [-1, w\*h\*1])

dice\_coef

intersection = K.sum(y\_true \* y\_pred, axis=-1)

return (2. \* intersection + smooth) /

(K.sum(y\_true, axis=-1) + K.sum(y\_pred, axis=-1) + smooth)

binary\_cross\_entropy

pos\_weight = 2

loss = pos\_weight \* y\_true \* K.log(y\_pred+K.epsilon()) + (1-y\_true)\*K.log(1-y\_pred+K.epsilon())

return -K.mean(loss, axis=-1)

keras的损失函数binary\_crossentropy支持[batch\_size, w, h, 1], 但最后一层需为sigmoid激活函数

model.compile(loss=binary\_crossentropy, metrics=[dice\_coef])

但是class\_weight不支持3D以上, 自定义可以实现此功能

输出为 [batch\_size, 1], 每个元素表示case的loss, keras会求平均

若输出为标量，则表示batch\_size case的平均loss

dice\_coef

intersection = K.sum(y\_true \* y\_pred)

return 2. \* intersection / (K.sum(y\_true) + K.sum(y\_pred) )

categorical\_crossentropy

若logits and y have same shape (32, 512, 512, 10)

则class\_weights have shape (1, 10)

flat\_logits = tf.reshape(logits, [-1, n\_class]) (32\*512\*512, 10)

flat\_labels = tf.reshape(y, [-1, n\_class])

class\_weights = tf.constant(np.array(class\_weights, dtype=np.float32))

weight\_map = tf.multiply(flat\_labels, class\_weights) (32\*512\*512, 10)

weight\_map = tf.reduce\_sum(weight\_map, axis=1) (32\*512\*512, 1)

(32\*512\*512, 1) ← (32, 512, 512, 10) , (32, 512, 512, 10)

loss\_map = tf.nn.softmax\_cross\_entropy\_with\_logits(flat\_logits, flat\_labels)

weighted\_loss = tf.multiply(loss\_map, weight\_map) (32\*512\*512, 1)

loss = tf.reduce\_mean(weighted\_loss) scalar

* Metrics f(y\_true, y\_pred)

**#(y\_true, y\_pred) as arguments and return a single tensor value.**

import keras.backend as K

def mean\_pred(y\_true, y\_pred):

return K.mean(y\_pred)

model.compile(optimizer='rmsprop',

loss='binary\_crossentropy',

metrics=['accuracy', mean\_pred])

### workflow

Load Data.

Define Model.

Compile Model.

Fit Model.

Evaluate Model.

Tie It All Together.

#### Image denoising

### 数据预处理

图像以较小的边缩放至[256, 480], 这样便于scale augmentation,然后从中随机裁出224\*224

水平翻转

标准颜色扩充

减均值

优化

Initialization

Initialization under ReLU activation --kaiming he

## tensorflow

TensorFlow + TensorBoard + TensorFlow Serving

TensorFlow =

TensorFlow Core (low level API) +

tf.contrib.learn (high-level API)

设计目标：在多台计算机以及单机多CPU,单机多GPU环境中具有良好的可伸缩性

定义计算图 ＋ 运行计算图（在数据上）

**计算图本质是一组链接在一起的函数**，每个函数都会将其输出传递给0,1,多个位于这个级联链上的其他函数

计算图（数据流图） ＝ 节点（对数据所做的运算或输入操作）＋边（tensor)

A computational graph is a series of TensorFlow operations arranged into a graph of nodes

To actually evaluate the nodes, we must run the computational graph within a session. A session encapsulates the control and state of the TensorFlow runtime.

一个 TensorFlow 图描述了计算的过程. 为了进行计算, 图必须在 会话 里被启动. 会话 将图的 op 分发到诸如 CPU 或 GPU 之类的 设备 上, 同时提供执行 op 的方法. 这些方法执行后, 将产生的 tensor 返回

当TensorFlow库被加载时，它会自动创建一个Graph对象，并将其作为默认的数据流图。因此在Graph.as\_default()上下文管理器之外定义的任何Op, Tensor对象都会自动放置在默认的数据流图中。

在大多数TensorFlow程序中，只使用默认数据流图就足够了。然而，如果需要定义多个相互之间不存在依赖关系的模型，则创建多个Graph对象十分有用。

g1 = tf.get\_default\_graph()

g2 = tf.Graph()

with g1.as\_default():

定义g1的Op 和张量等

with g2.as\_default():

定义g2的Op和张量等

Tensor

一个 op 获得 0 个或多个 Tensor, 执行计算, 产生 0 个或多个 Tensor. 每个 Tensor 是一个类型化的多维数组

**TensorFlow Op可接收标准Python数据类型，如整数或字符串，并将它们自动转化为张量，返回的 tensor 是 numpy ndarray 对象**

**经典做法是借助Numpy数组手工定义Tensor对象**,TensorFlow的数据类型是基于Numpy的数据类型的。实际上，语句np.int32==tf.int32。任何NumPy数组都可以传递给TensorFlow Op

t\_0 = np.array(50, dtype=np.int32) 0阶张量 即标量

t\_1 = np.array([1., 2.], dtype=np.float32) 1阶张量即向量

t\_2 = np.array([ 2阶张量即矩阵

[1, 2, 3],

[4, 5, 6]

], dtype=np.int32)

除了能够将张量的每一维指定为固定长度，也可以将None作为某一维的值，使该疑张量具有可变长度。

s\_0 = () 0阶张量的形状

s\_1 = (3) 1阶张量的形状

s\_2 = (3, 2)

**s\_2\_flex = (None, 2) 行数任意，列数为3的矩阵形状**

s\_any = None 形状任意的张量

shape = tf.shape(tensor1)

sess.run(shape) 求张量tensor1的shape

节点Op

a = np.array([2, 3], dtype=np.int32) tensor

b = tf.constant([5, 3], name=”input\_b”) 节点，指向常量操作输出Tensor句柄

**c = tf.add(a, b) 变量c为指向该Op输出的Tensor对象的句柄**

实际上，存在对应的运算符重载，张量逐元素运算+-\*/, //, %, \*\*, < <= > >= & | ^

c = a + b 运算符重载

如何确保计算机中对必要的节点执行运算，而无需手工指定？

答案是：利用节点之间的依赖关系

可对数据流中的任意Op使用run()函数。当将某个Op传入sess.run()时，本质上是通过TensorFlow“这里有一个节点，我希望得到它的输出，请执行所有必要的运算来求取这个节点的输出”

a = tf.add(2, 3)

b = tf.mul(a, 3)

with tf.Session() as sess: 等价于tf.Session(graph=tf.get\_default\_graph())

sess.run(tf.initialize\_all\_variables())

sess.run(fetches=[b], feed\_dct={a: 15})

fetches参数接收任意的数据流图元素(Op or Tensor对象），后者指定了用户希望执行的对象。若请求的是Tensor对象，则run()输出的是Numpy数组； 若请求的是Op，则输出为None

feed\_dict用于覆盖数据流图中的Tensor对象值，**字典键为Tensor对象的句柄**

sess.run([b], feed\_dct={a: 15})

例子中取张量b， TensorFlow便会得到通知，Session对象应当找到为计算b的值所需的全部节点，顺序执行这些节点，然后b的值输出

sess.run(tf.initialize\_all\_variables()) 执行初始化Variable对象所需的计算，返回值为None

summary.FileWriter 对象保存来自数据流图的数据和概括流计量，第一个参数是数据流图的描述在磁盘中的存放路径。第二个参数是要追踪的数据流图

writer = tf.summary.FileWriter(‘./my\_graph’, sess.graph)

一量执行完以上代码，便可启动TensorBoard

$ tensorboard –-logdir=”.my\_graph”

localhost:6006会显示利用Writer对象要求TensorFlow所保存的信息。

利用占位节点添加输入

a = tf.placeholder(tf.int32, shape=[2])

b = tf.reduce\_sum(a)

sess.run(b, feed\_dict={a: np.array([5, 3], dtype=np.int32)})

变量

Tensor对象和Op对象都是不可变的，如何维护图执行过程中的状态信息？

答案是：变量(Variables for more details),包含了在对Session.run()多次调用中可持久化的可变张量值

通常会将一个统计模型中的参数表示为一组变量. 例如, 你可以将一个神经网络的权重作为某个变量存储在一个 tensor 中. 在训练过程中, 通过重复运行训练图, 更新这个 tensor.

weight = tf.Variable(3)

Variables对象通常需要初值，TensorFlow提供辅助Op

weights = tf.Varaible(tf.zeros([2, 2]))

tf.ones([6])

tf.random\_uniform([3, 3, 3], minval=0, maxval=10)

tf.random\_normal([3, 3, 3], mean=0.0, stddev=2.0)

#创建[mean-2std, mean+2std]

tf.truncated\_normal([2, 2], mean=5.0, stddev=1.0)

为使用Variable对象，需要采取一些额外的步骤，必须在一个Session对象内对Variable对象进行初始化。这样会使Session对象开始追踪这个Variable对象的值的变化

sess.run(tf.initialize\_all\_variables())

修改Variable

weights\_update = weights.assign(weights \* 2)

sess.run(weights) 因为还未运行操作weights.assign，所以weights未改变

sess.run(weights\_update)

运行操作weights.assign, weights and weights\_update均改变

由于不同Session对象会各自独立地维护Variable对象的值，因此每个Session对象都拥有自己的，在Graph对象中定义的Variable对象的当前值

若将所有Variable对象的值重置为初始值，则只需再次调用sess.run(tf.initialize\_all\_variables())

weights变量一般作为trainable参数，Optimizer会自动修改Variable对象的值。若该权重不参与训练，可以设置参数trainable=False

weight\_no = tf.Variable(0, trainable=False)

在TensorFlow中，真正的循环依赖关系是无法表示的，这并非坏事。在实际使用中，完全可通过对数据流图进行有限次的复制，然后将它们并排放置，并将代表相邻迭代轮次的副本的输出与输入串接。该过程通常被称为数据流图的“展开”unrolling

保存训练检查点：周期性地保存所有变量，创建检查点checkpoint文件，并在必要时从最近的检查点恢复训练。默认情况下，Saver对象 只会保留最近的5个文件

tf.to\_float(weights)

tf.cast(weights, “float”)

tf.reshape(X, [None, 28, 28, 1])

tf.range(-1, 3)

张量逐元素运算+-\*/, //, %, \*\*, < <= > >= & | ^

tf.matmul(X, W) 矩阵乘

tf.transpose(W)

将几个行向量的特征 格式化为矩阵：行（样本），列（特征）

features = tf.transpose(tf.pack([sepal\_length, sepal\_width]))

将类名称转换为从0开始计的类别索引

label = tf.to\_int32(tf.argmax(tf.to\_int32(tf.pack([

tf.equal(label\_name, [“Iris-setosa”]),

tf.equal(label\_name, [“Iris-versicolor”]),

tf.equal(label\_name, [“Iris-virginica”]) ])), 0)

回归: 总平方差(回归损失函数)

loss = tf.reduce\_sum(tf.squared\_difference(Y\_predicted, Y))

分类问题

字符串特征转数值型特征？

为什么不能为每个可能的取值分配一个数值？如用1代表一等船票，2 and 3分别代表二，三等船票，因为这种方式会为这些取值强加一种实际并不存在的线性关系。我们不能说“三等票是一等票的3倍“

正确的做法是将每个属性特征扩展为N维的布尔型特征，每个可能的取值对应一维。若具备该属性，则相应的维度上取值为1.这样就可以使模型独立地学习到每个可能的取值的重要性。

对于只可能取两种值的属性，如果性别，用单个变量来表示已经足够，这是因为可表达这些值之间的线性关系。如令female = 1, male = 0, 则male = 1 – female, 因此单个权值便可学习同时表示两种状态

两类问题：交叉熵

tf.sigmoid(f(X, W, b) sigmoid （概率值）

loss = sum(yi\*log(y\_predictedi )+ (1-yi)log(1-y\_predictedi)) i为case

loss = tf.reduce\_mean(

tf.nn.sigmoid\_cross\_entropy\_with\_logits(Y\_predicted, Y))

注：

交叉熵与香农熵区别：

香农熵 H = -sum(yi\*log(yi ))

物理意义：已知字符串中每个字符出现的概率 yi，采用最优编码方案，则字符串编码每个字符的平均位数

交叉熵 H = -sum(yi\*log(y\_predictedi ))

实际中，对符号编码时，得到的概率是y\_predictedi 而非真实概率 yi，则每个符号所需的编码长度会更大，交叉熵允许用户以次优编码方案对字符串编码。

当建模得到的概率是y\_predictedi ＝＝真实概率 yi时，交叉熵取得最小值。交叉熵度量的是模型分布与真实分布的吻合情况。若交叉熵越接近于熵，则得到的概率是y\_predictedi 越逼近真实概率yi

多类问题：

tf.nn.softmax(f(X, W, b)) 概率值

单样本的所有类别概率和为1, 若<1, 则意味着存在一些隐藏的类别； 若>1,则说明每个样本可能同时属于多个类别

loss = -sum(sum(yic\*log(y\_predictedic ))) i为case, c为类别

若每个样本只对应单个类别

loss = tf.reduce\_mean(

tf.nn.spare\_softmax\_cross\_entropy\_with\_logits(Y\_predicted, Y))

注意：类别可以为0, 1, 2, 3, ...

若每个样本含多个类别，即类别0的概率，类别1的概率，。。。

loss = tf.reduce\_mean(

tf.nn.softmax\_cross\_entropy\_with\_logits(Y\_predicted, Y))

注意：类别只能为one-hot编码

例子：手写体

softmax regression

softmax模型可以用来给不同的对象分配概率。我们训练更加精细的模型时，最后一步也需要用softmax来分配概率。

y = softmax(xW + b)

损失函数：交叉熵cross-entropy

Hy’(y) = - sum (yi’log(yi)) 度量真实分布yi’与预测分布 yi的差异性信息

class CNN(object):

def \_\_init\_\_(self):

self.keep\_prob = 0.5

self.y\_class = 10

self.define\_graph()

self.sess = tf.Session()

self.sess.run(tf.global\_variables\_initializer())

# 保存训练检查点

self.saver = tf.train.Saver()

self.initial\_epoches = 0

ckpt = tf.train.get\_checkpoint\_state(os.path.dirname(\_\_file\_\_))

if ckpt and ckpt.model\_checkpoint\_path:

self.saver.restore(self.sess, ckpt.model\_checkpoint\_path)

self.initial\_epoches = int(ckpt.model\_checkpoint\_path.rsplit('-', 1)[1])

**def inference(self, X):**

...

return Y\_predicted

@staticmethod

**def loss(Y, Y\_predicted):**

return -tf.reduce\_sum(Y \* tf.log(Y\_predicted))

tf.argmax 给出某个tensor对象在某一维上的其数据最大值所在的索引值

@staticmethod

**def acc(Y, Y\_predicted):**

correct\_prediction = tf.equal(tf.argmax(Y\_predicted, 1), tf.argmax(Y, 1))

return tf.reduce\_mean(tf.cast(correct\_prediction, "float"))

因为TensorFlow拥有一张描述你各个计算单元的图，它可以自动地使用反向传播算法(backpropagation algorithm)来有效地确定你的变量是如何影响你想要最小化的那个成本值的。然后，TensorFlow会用你选择的优化算法来不断地修改变量以降低成本

用梯度下降算法训练你的模型，微调你的变量，不断减少成本

为什么梯度下降法需要学习率？即Wi+1 = Wi – lr\*gradient(loss(Wi))

因为梯度向量的长度实际上是一个在“损失函数单元”中而非“权值单元”中度量的量，因此需要对梯度进行缩放，使其能够与权值相加

反向传播算法：据链式法则计算梯度，每一层的导数＝后一层的导数\*前一层输出

@staticmethod

**def optimize(total\_loss):**

return tf.train.AdamOptimizer(1e-4).minimize(total\_loss)

**def define\_graph(self, ):**

self.X = tf.placeholder(tf.float32, [None, 28, 28, 1])

self.Y = tf.placeholder(tf.float32, [None, 10])

Y\_predicted = self.inference(self.X)

self.total\_loss = self.loss(self.Y, Y\_predicted)

self.optimizer = self.optimize(self.total\_loss)

self.accuracy = self.acc(self.Y, Y\_predicted)

该循环的每个步骤中，我们都会随机抓取训练数据中的128个批处理数据点，使用一小部分的随机数据来进行训练被称为随机训练（stochastic training），在理想情况下，我们希望用我们所有的数据来进行每一步的训练，因为这能给我们更好的训练结果，但显然这需要很大的计算开销。所以，每一次训练我们可以使用不同的数据子集，这样做既可以减少计算开销，又可以最大化地学习到数据集的总体特性。

**def train(self, ds\_train, batch\_size=128, epoches=5):**

total\_num = ds\_train.num\_examples

epoch = int(math.ceil(total\_num/float(batch\_size)))

for i in range(self.initial\_epoches, epoches):

print('Epoch {0}/{1}'.format(i+1, epoches))

for j in range(epoch):

**batch = ds\_train.next\_batch(batch\_size)**

\_, loss\_value, acc\_value = self.sess.run([self.optimizer, self.total\_loss, self.accuracy], feed\_dict={self.X: batch[0].reshape([-1, 28, 28, 1]), self.Y: batch[1].reshape([-1, 10])})

print('{0}/{1} [........] - loss: {2} - acc: {3}'.format(

(j+1)\*batch\_size, total\_num, loss\_value, acc\_value))

if epoches % 2 == 0:

self.saver(self.sess, 'mnist-model', global\_step=epoches)

**self.saver.save(self.sess, 'mnist-model', global\_step=epoches)**

return self

def test(self, ds\_test, batch\_size=32):

loop = int(math.ceil(ds\_test.images.shape[0] / float(batch\_size)))

accs = []

for i in range(loop):

batch = ds\_test.next\_batch(batch\_size)

accs.append(self.sess.run(self.accuracy, feed\_dict={

self.X: batch[0].reshape([-1, 28, 28, 1]),

self.Y: batch[1].reshape([-1, 10])}))

accuracy\_mean = sum(accs) / len(accs)

print('test accuracy %g' % accuracy\_mean)

多层卷积网络

权重初始化

为了创建这个模型，我们需要创建大量的权重和偏置项。这个模型中的权重在初始化时应该加入少量的噪声来打破对称性以及避免0梯度。由于我们使用的是ReLU神经元，因此比较好的做法是用一个较小的正数来初始化偏置项，以避免神经元节点输出恒为0的问题（dead neurons）

with tf.name\_scope("conv1"):

W\_conv1 = tf.Variable(**tf.truncated\_normal**([5, 5, 1, 32], stddev=0.1))

b\_conv1 = **tf.Variable**(tf.constant(0.1, shape=[32]), name="bias")

c1 = **tf.nn.conv2d**(X, W\_conv1, strides=[1, 1, 1, 1], padding='SAME')

h\_conv1 = **tf.nn.relu**( c1 + b\_conv1)

h\_pool1 = **tf.nn.max\_pool**(h\_conv1, ksize=[1, 2, 2, 1], strides=[1, 2, 2, 1], padding='SAME')

strides参数的格式与输入向量相同，即(image\_batch\_size\_stride, image\_height\_stride, image\_width\_stride, image\_channels\_stride)

激活函数

函数单调 （便于梯度下降法寻找局部极值点成为可能）

函数可微分（从而能够计算梯度）

tf.nn.relu, tf.sigmoid, tf.tanh,

归一化层

一般对tf.nn.relu的输出进行归一化

tf.nn.local\_response\_normalization

池化层

tf.nn.max\_pool, tf.nn.avg\_pool

为了减少过拟合，常在输出层之前加入dropout

TensorFlow的tf.nn.dropout操作除了可以屏蔽神经元的输出外，还会自动处理神经元输出值的scale。所以用dropout的时候可以不用考虑scale。

keep\_prob = tf.placeholder("float")

h\_fc1\_drop = **tf.nn.dropout**(h\_fc1, keep\_prob)

输出层

y\_conv = **tf.nn.softmax**(tf.matmul(h\_fc1\_drop, W\_fc2) + b\_fc2)

获取层张量的shape

sess.run(h\_pool1.get\_shape())

高级层

加载图像

建议在训练之前对图像进行预处理并将预处理结果保存下来。每次加载图像时才对其进行处理是不推荐的做法

当输入为一幅图像时，不应使用tf.to\_float, 而应使用tf.image.convert\_image\_dtype,该方法将以恰当的方式调整各分量以表示顔色值。

预处理包括对图像裁剪，缩放，灰度调整，翻转，扭曲处理等等，以使输入给网络的训练信息多样化。虽然这个步骤会进一步增加处理时间，但却有助于缓解过拟合现象

tf.image.rag\_to\_grayscale

tf.image.convert\_image\_dtype tf.uint8 → tf.float，且归一化

tf.image.central\_crop

tf.image.crop\_to\_bounding\_box

tf.image.pad\_to\_bounding\_box

tf.image.resize\_image\_with\_crop\_or\_pad

tf.image.flip\_left\_right

tf.image.flip\_up\_down

tf.image.adjust\_brightness

tf.image.adjust\_contrast

tf.image.adjust\_hue

tf.image.adjust\_saturation

tf.image.rgb\_to\_hsv

含权值初始化，偏置初始化，激活函数

tf.contrib.layers.convolution2d

含权值初始化，偏置初始化

tf.contrib.layers.fully\_connected

深度值的增加，减少使用该网络所需的计算量

from **tf.contrib.layers** import convolution2d, max\_pool2d, fully\_connected

def inference(self, X):

with tf.name\_scope("conv1"):

conv1 = convolution2d(X, 32, (5, 5))

pool1 = max\_pool2d(conv1, (2, 2))

with tf.name\_scope("conv2"):

conv2 = convolution2d(pool1, 64, (5, 5))

pool2 = max\_pool2d(conv2, (2, 2))

with tf.name\_scope("fc1"):

pool2\_flat = tf.reshape(pool2, [-1, 7 \* 7 \* 64])

fc1 = fully\_connected(pool2\_flat, 1024)

with tf.name\_scope("fc2"):

fc1\_drop = tf.nn.dropout(fc1, self.keep\_prob)

y\_conv =fully\_connected(fc1\_drop, self.y\_class, activation\_fn=tf.nn.softmax)

return y\_conv

官方站点：

tensorflow.org/get\_started/

import tensorflow as tf

import numpy as np

features = [tf.contrib.layers.real\_valued\_column("x", dimension=1)]

estimator = tf.contrib.learn.LinearRegressor(feature\_columns=features)

x\_train = np.array([1., 2., 3., 4.])

y\_train = np.array([0., -1., -2., -3.])

x\_eval = np.array([2., 5., 8., 1.])

y\_eval = np.array([-1.01, -4.1, -7, 0.])

input\_fn = tf.contrib.learn.io.numpy\_input\_fn({"x":x\_train}, y\_train, batch\_size=4, num\_epochs=1000)

eval\_input\_fn = tf.contrib.learn.io.numpy\_input\_fn({"x":x\_eval}, y\_eval, batch\_size=4, num\_epochs=1000)

estimator.fit(input\_fn=input\_fn, steps=1000)

print("train loss: %r"% estimator.evaluate(input\_fn=input\_fn))

print("eval loss: %r"% estimator.evaluate(input\_fn=eval\_input\_fn))

Pycharm tensorflow ImportError but works fine with Terminal?

open PyCharm from the command line and everything works now