

2018/1/4 上午11:40

很难有较好的泛化能力，这叫做 **covariate shift** (Shimodaira, 2000)。解决办法是**domain adaptation** (Jiang, 2008) 和迁移学习。

BN想把输入的均值方差规范化，使输入分布一致，但是仅均值、方差一样的分布就一定一样吗？但是思路是这样，而且效果好。

也可以不只关注学习系统整体，而关注它的内部，如一个**subnetwork**或 **a layer**：

损失函数为：

$$\ell = F_2(F_1(u, \Theta_1), \Theta_2)$$

subnetwork的损失函数为：

$$\ell = F_2(x, \Theta_2)$$

SGD:

$$\Theta_2 \leftarrow \Theta_2 - \frac{\alpha}{m} \sum_{i=1}^m \frac{\partial F_2(x_i, \Theta_2)}{\partial \Theta_2}$$

is exactly equivalent to that for a stand-alone network F2 with input x.

输入X的分布保持不变的话，参数就不需要 **readjust** 去补偿X分布的变化。

注意**sigmoid** 函数的特性，当输入不集中在 0 附近时（饱和时），梯度会非常小，训练会很慢。然而，输入x受前面层的参数的影响，这些参数的变化很可能导致x过早变化到**sigmoid**函数的饱和区域，收敛减慢。这种影响随着深度增加而加大。

ReLU和好的初值、更小的学习率可以解决梯度消失的问题。但是，要是我们能让输入更稳定，就不太可能 **get stuck in the saturated regime**,训练也会加快。

3.这种深度网络内部，参数不断变化导致的各层输入分布的变化 称为 **Internal Covariate Shift**.

Batch Normalization

- a.保持各层输入的均值和方差稳定，来减弱 **internal covariate shift**;
- b.也让梯度受参数及其初值的减小;
- c.它也算作正则项，减少对**Dropout**的依赖;
- d.它让卡在饱和区域的概率降低，以便可以使用 **saturating nonlinearities**

Towards Reducing Internal Covariate Shift

- 1.网络输入白化会加快收敛 (LeCun et al., 1998b; Wiesler & Ney, 2011)
- 并不是直接在每层规范化这么简单，如果模型的优化求梯度时不考虑到归一化的话，会影响模型，就是你优化半天学到了某种分布，一规范化，就把这它破坏了。
- 2.规范化与某个样本的各层输入及所有样本的各层输入都有关（对某个规范化时用到了所有样本）：

$$\hat{x} = \text{Norm}(x, \mathcal{X})$$

在反向传导时，求导需考虑下面两项：

$$\frac{\partial \text{Norm}(x, \mathcal{X})}{\partial x} \quad \text{and} \quad \frac{\partial \text{Norm}(x, \mathcal{X})}{\partial \mathcal{X}}$$

这样基于整个训练集的白化是非常耗时的，因为白化需要计算 x 的协方差矩阵及白化部分，还需计算**BP**算法中的求导。

但是基于某个或者部分样本进行规范化又会**changes the representation ability of a network**

所以本文在**minibatch**内归一化，再用可以学习的 γ 和 β 来拟合**minibatch**的统计量与整个训练集统计量之间的关系。

Normalization via Mini-Batch Statistics

- 1.有两个方面简化计算：
- a.把 x 向量中每个元素当成独立随机变量单独进行规范化，向量中各变量独立了，也没有什么协方差矩阵

dual Networks (<http://blog.csdn.net/u012816943/article/details/51702520>)

7023

论文笔记-Batch Normalization (<http://blog.csdn.net/u012816943/article/details/51691868>)

6180

论文笔记-Automatic Differentiation for Rank Matrix (<http://blog.csdn.net/u012816943/article/details/51708013>)

3746

Theano-Deep Learning Tutorials 笔记:Stacked Denoising Autoencoders (SdA) (<http://blog.csdn.net/u012816943/article/details/50514771>)

2520

Theano-Deep Learning Tutorials 笔记:LS TM Networks for Sentiment Analysis (<http://blog.csdn.net/u012816943/article/details/50513384>)

2448


内容举报


返回顶部

了。这种规范化在各变量相关的情况下依然能加速收敛, (LeCun et al., 1998b),此外, 如果看成向量中变量的联合概率, 需要计算协方差矩阵, 如果变量个数大于minibatch中样本数, 协方差矩阵不可逆!!

b.在每个mini-batch中计算得到mini-batch mean和variance来替代整体训练集的mean和variance. Algorithm 1.

simply normalizing each input of a layer may change what the layer can represent.normalizing the inputs of a sigmoid would constrain them to the linear regime of the nonlinearity

为了解决这个问题, we make sure that the transformation inserted in the network can represent the identity transform.也就是用可以学习的 γ 和 β 去拟合出与原先等价的变换。

we introduce, for each activation $x^{(k)}$, a pair of parameters $\gamma^{(k)}, \beta^{(k)}$, which scale and shift the normalized value:

$$y^{(k)} = \gamma^{(k)} \hat{x}^{(k)} + \beta^{(k)}.$$

These parameters are learned along with the original model parameters, and restore the representation power of the network. Indeed, by setting $\gamma^{(k)} = \sqrt{\text{Var}[x^{(k)}]}$ and $\beta^{(k)} = \text{E}[x^{(k)}]$, we could recover the original activations, if that were the optimal thing to do.

采用 normalize via mini-batch statistics, the statistics used for normalization can fully participate in the gradient backpropagation

Batch Normalizing Transform:

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1 \dots x_m\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{ mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{ mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{ normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{ scale and shift}$$

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

只要minibatch中的样本采样与同一分布, 规范化后的输入 x 期望为0, 方差为1, 把规范后的 x 进行线性变换得到 y 作为后续层的输入, 可以发现 后续层的输入具有固定的均值和方差的。尽管 规范化后的 x 的联合分布在训练过程中会改变(源于第一个简化, 本文的规范化是把 x 向量中各个变量当作独立的, 单独规范化的, 所以他们的联合分布并不稳定, 只是单独是稳定的), 但还是可以使训练加速。

2.优化中也需要对 BN 变换的两个参数进行优化, 链式法则求导就可以了:



内容举报



返回顶部



$$\frac{\partial \ell}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial y_i} \cdot \gamma$$

$$\frac{\partial \ell}{\partial \sigma_B^2} = \sum_{i=1}^m \frac{\partial \ell}{\partial \hat{x}_i} \cdot (x_i - \mu_B) \cdot \frac{-1}{2} (\sigma_B^2 + \epsilon)^{-3/2}$$

$$\frac{\partial \ell}{\partial \mu_B} = \left(\sum_{i=1}^m \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \frac{\sum_{i=1}^m -2(x_i - \mu_B)}{m}$$

$$\frac{\partial \ell}{\partial \hat{x}_i} = \frac{\partial \ell}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \frac{2(x_i - \mu_B)}{m} + \frac{\partial \ell}{\partial \mu_B} \cdot \frac{1}{m}$$

$$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^m \frac{\partial \ell}{\partial y_i} \cdot \hat{x}_i$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^m \frac{\partial \ell}{\partial y_i}$$

BN变换是可微的，通过BN变换，可以减弱输入分布的 internal covariate shift，并且学习到这个线性变换让 BN变换 与网络本来的变换 等价，preserves the network capacity

Training and Inference with Batch-Normalized Networks

- 1.使用BN，把网络中各层输入 x 变为 $\text{BN}(x)$ 即可。可以使用SGD及其各种变种训练。
- 2.训练时候在minibatch内规范化非常高效，但是推断时就不需要而且不应该这样了。推断是我们希望输出只依赖于输入，所以规范化中的期望、方差用全部数据计算：

$$E_B[\sigma_B^2]$$

就是各minibatch的方差求平均，minibatch数量为 m 。

注意：推断时，均值和方差是固定的，那么规范化这步线性变换可以和 γ 、 β 这步线性变换 合成 一个线性变换。训练BN网络步骤如下：



内容举报



返回顶部



Input: Network N with trainable parameters Θ ;
subset of activations $\{x^{(k)}\}_{k=1}^K$

Output: Batch-normalized network for inference, $N_{\text{BN}}^{\text{inf}}$

```

1:  $N_{\text{BN}}^{\text{tr}} \leftarrow N$  // Training BN network
2: for  $k = 1 \dots K$  do
3:   Add transformation  $y^{(k)} = \text{BN}_{\gamma^{(k)}, \beta^{(k)}}(x^{(k)})$  to
4:    $N_{\text{BN}}^{\text{tr}}$  (Alg. 1)
5:   Modify each layer in  $N_{\text{BN}}^{\text{tr}}$  with input  $x^{(k)}$  to take
6:    $y^{(k)}$  instead
7: end for
8: Train  $N_{\text{BN}}^{\text{tr}}$  to optimize the parameters  $\Theta \cup$ 
9:    $\{\gamma^{(k)}, \beta^{(k)}\}_{k=1}^K$ 
10:  $N_{\text{BN}}^{\text{inf}} \leftarrow N_{\text{BN}}^{\text{tr}}$  // Inference BN network with frozen
    // parameters
11: for  $k = 1 \dots K$  do
12:   // For clarity,  $x \equiv x^{(k)}, \gamma \equiv \gamma^{(k)}, \mu_{\mathcal{B}} \equiv \mu_{\mathcal{B}}^{(k)}$ , etc.
13:   Process multiple training mini-batches  $\mathcal{B}$ , each of
    size  $m$ , and average over them:
    
$$E[x] \leftarrow E_{\mathcal{B}}[\mu_{\mathcal{B}}]$$

    
$$\text{Var}[x] \leftarrow \frac{m}{m-1} E_{\mathcal{B}}[\sigma_{\mathcal{B}}^2]$$

14:   In  $N_{\text{BN}}^{\text{inf}}$ , replace the transform  $y = \text{BN}_{\gamma, \beta}(x)$  with
    
$$y = \frac{\gamma}{\sqrt{\text{Var}[x] + \epsilon}} \cdot x + \left( \beta - \frac{\gamma E[x]}{\sqrt{\text{Var}[x] + \epsilon}} \right)$$

15: end for
```

Algorithm 2: Training a Batch-Normalized Network

Batch-Normalized Convolutional Networks

1. BN 可以用于任意层的 activations，但是把 BN 加在 $Wu + b$ 之后，非线性激活函数之前更好！

因为前层的 activations（这层的输入 u ）是非线性输出的，其分布很可能在训练中变化；而 $Wu + b$ 更可能有 a symmetric, non-sparse distribution, that is “more Gaussian”

(Hyvärinen & Oja, 2000); 规范化它更有可能得到稳定的 activations 分布。

2. 注意 b 可以不管，因为减均值时 b 会被消掉， b 的作用其实被 β 代替了，所以：

$$z = g(\text{BN}(Wu))$$

BN 是对 $x = Wu$ 的每一维单独规范化

3. 对卷积层，规范化也该保持卷积特性，即 相同 feature map，不同 location 的元素 用相同方式规范化：

a mini-batch of size m and feature maps of size $p \times q$, $m \times p \times q$ 个元素一起规范化！每个 feature map 有一对 $\gamma \beta$ 。

Batch Normalization enables higher learning rates

1. 太大的学习率可能导致 梯度爆炸或消失以及卡在局部极值，BN 可以防止参数小变换被逐层放大，通过修改 γ 、 β 可以优化 activations 的变化。

一般来说，大学习率增加参数的 scale，在 BP 中放大了梯度，导致模型爆炸。然而使用了 BN，每层的 BP 不受其参数影响：



内容举报



返回顶部



$$\frac{\partial \text{BN}((aW)u)}{\partial u} = \frac{\partial \text{BN}(Wu)}{\partial u}$$

$$\frac{\partial \text{BN}((aW)u)}{\partial (aW)} = \frac{1}{a} \cdot \frac{\partial \text{BN}(Wu)}{\partial W}$$

The scale does not affect the layer Jacobian nor, consequently, the gradient propagation.

而且：大权重会导致更小的梯度，所以BN可以稳定参数的增长。

2. BN还可以使 layer Jacobians 的奇异值接近 1 .这更利于训练 (Saxe et al.,2013).

论文中有在高斯、独立且变换为线性等条件下，可以推出来，但是说实话假设有点太苛刻，有点强行解释的味道，论文也提出更普适的结论需后续研究。

Batch Normalization regularizes the model

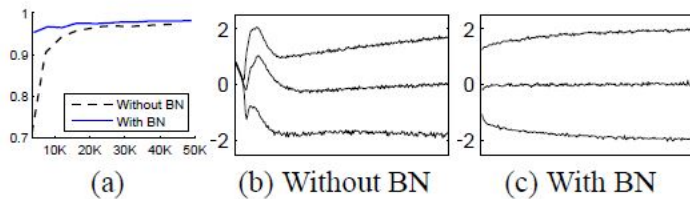
1.使用BN后，训练时对于单个样本与整个minibatch综合考虑了， training network no longer producing deterministic values for a given training example

这有利于提升网络的泛化能力，可以代替 Dropout

Experiments

Accuracies over time

在mnist上用3个隐层，每层100个的神经元的网络进行实验，初值为高斯，sigmoid函数，迭代50000次，minibatch为60个样本，损失为交叉熵。



(a) 可以看到，加BN测试精度更高，而且最开始就达到了较高的精度；
(b,c) 可以明显看到加BN后分布更加稳定。图中三条线为{15, 50, 85}分位数。
后面的实验就看论文了。

版权声明：本文为博主原创文章，未经博主允许不得转载。



- chuchus (/chuchus) 2017-11-15 09:52 2楼

博主的

回复
-
- seashell_9 (/seashell_9) 2017-03-29 21:10 1楼


博主你好，想问一下用不用每一个conv层后面都加BN？还是说前面几个conv层加BN就行？

回复

Batch Normalization论文翻译——中文版

(http://blog.csdn.net/Quincutial/article/details/78124629)


Batch Normalization 导读

 malefactor 2016年05月24日 19:08 39220

Batch Normalization作为最近一年来深度学习的重要成果，已经被证明其有效性和重要性。本文对原始论文进行导读，帮助读者更好地理解BatchNorm。...

(http://blog.csdn.net/malefactor/article/details/51476961)


关于Batch Normalization的另一种理解

 Alchipmunk 2017年01月11日 16:00 3163

Batch Norm可谓深度学习中非常重要的技术，不仅可以使训练更深的网络变容易，加速收敛，还有一定正则化的效果，可以防止模型过拟合。在很多基于CNN的分类任务中，被大量使用。但我最近在图像超分辨...

(http://blog.csdn.net/Alchipmunk/article/details/54234646)


深度学习（二十九）Batch Normalization 学习笔记

 hjimce 2016年03月12日 17:00 68802

近年来深度学习捷报连连，声名鹊起，随机梯度下架成了训练深度网络的主流方法。尽管随机梯度下降法，对于训练深度网络，简单高效，但是它有个毛病，就是需要我们人为的去选择参数，比如学习率、参数初始化等，这些...

(http://blog.csdn.net/hjimce/article/details/50866313)

为什么会出现Batch Normalization层

 NNNNNNNNNNNNY 2017年04月21日 16:36 2223

训练模型时的收敛速度问题众所周知，模型训练需要使用高性能的GPU，还要花费大量的训练时间。除了数据量大及模型复杂等硬性因素外，数据分布的不断变化使得我们必须使用较小的学习率、较好的权重初值和不容易饱和...

(http://blog.csdn.net/NNNNNNNNNNNNNY/article/details/70331796)


Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

 wspba 2017年03月23日 14:31 1279

虽然GoogLeNet在ILSVRC2014上斩获了各种大奖，但是GoogLe的各位大牛依然没有闲着，他们马上又抛出了一个新的问题：Internal Covariate Shift，也就是本文要研究的...

(http://blog.csdn.net/wspba/article/details/65440004)

论文笔记——《Batch Normalization Accelerating Deep Network Training by Reducing Internal Covariate Shift》

 jzrita 2017年05月26日 11:41 284

今年过年之前，MSRA和Google相继在Imagenet图像识别数据集上报告他们的效果超越了人类水平，下面将分两期介绍两者的算法细节。这次先讲Google的这篇《Batch Normalization Accelerating Deep Network Training by Reducing Internal Covariate Shift》...

(http://blog.csdn.net/jzrita/article/details/72765114)


[深度学习论文笔记][Weight Initialization] Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

 Hao_Zhang_Vision 2016年09月20日 13:54 706

Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03193. 2015.

(http://blog.csdn.net/Hao_Zhang_Vision/article/details/52595249)

深度学习 Batch Normalization 论文笔记

 Cyiano 2017年07月10日 16:19 1101

深度学习 Batch Normalization 论文笔记标题: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

 内容举报

 返回顶部

(http://blog.csdn.net/Cyiano/article/details/74928415)

论文笔记——Batch Normalization


 jzrita 2017年05月26日 14:56 175

Batch Normalization 学习笔记原文地址: http://blog.csdn.net/hjimce/article/details/50866313 作者: hjimce 一、背...

(http://blog.csdn.net/jzrita/article/details/72770076)

Batch Normalization 论文翻译——中英文对照


Batch Normalization 论文翻译——中英文对照

 Quincuntial 2017年09月28日 15:59 1556

(http://blog.csdn.net/Quincuntial/article/details/78124582)


【深度学习】论文导读: google的批正则方法 (Batch Normalization: Accelerat...

google 2015年的论文, 首次提出批正则方法, 优化深度神经网络的学习摘要在深度网络的训练中, 每一层网络的输入都会因为前一层网络参数的变化导致其分布发生改变, 这就要求我们必须使用一个很小的学...

 mao_xiao_feng 2016年11月13日 17:07 829

(http://blog.csdn.net/mao_xiao_feng/article/details/53150037)

Batch Normalization 学习笔记 (一)

 qq_30478885 2017年12月11日 21:19 40

原文地址: http://blog.csdn.net/hjimce/article/details/50866313 作者: hjimce 一、背景意义本篇博文主要讲解2015年深度学习领域, 非常...

(http://blog.csdn.net/qq_30478885/article/details/78776893)

深度学习 (二十九) Batch Normalization 学习笔记

Batch Normalization 学习笔记原文地址: http://blog.csdn.net/hjimce/article/details/50866313 作者: hjimce 一、背...

(http://blog.csdn.net/haoji007/article/details/52788634)

笔记: batch normalization: accelerating deep network training by reducing in...


概述: 机器学习系统中输入分布不均称为covariate shift。对于复杂网络系统, 某层输入分布不稳定称为internal covariate shift。本文提出了batch normaliza...

 a1154761720 2016年03月06日 11:37 1778

(http://blog.csdn.net/a1154761720/article/details/50812607)


《Batch Normalization Accelerating Deep Network Training by Reducing Intern...

本文转自: http://www.aichengxu.com/view/1422042 今年过年之前, MSRA和Google相继在Imagenet图像识别数据集上报告他们的效果超越了人类水平, 下面...

 xiaoyanghijk 2016年08月03日 12:01 153

(http://blog.csdn.net/xiaoyanghijk/article/details/52102302)

Batch Normalization 学习笔记

 zy3381 2016年04月08日 14:04 1012

Batch Normalization 学习笔记原文地址: http://blog.csdn.net/hjimce/article/details/50866313 作者: hjimce 一、...

(http://blog.csdn.net/zy3381/article/details/51096165)

加入CS231n, 第4集 Data Preprocessing, Weights Initialization 与 Batch Normalization

登录

注册

×





内容举报





返回顶部

Data Preprocessing, Weights Initialization与Batch NormalizationData Preprocessing Weights Initializat...

 u012767526 2016年05月14日 14:57  4820



(http://blog.csdn.net/u012767526/article/details/51405701)

深度学习（二十九）Batch Normalization 学习笔记

近年深度学习捷报连连，声名鹊起，随机梯度下架成了训练深度网  hjimce 2016年03月12日 17:00  68802
络的主流方法。尽管随机梯度下降法，将对于训练深度网络，简单高
效，但是它有个毛病，就是需要我们人为的去选择参数，比如学习率、参数初始化等，这些...

(http://blog.csdn.net/hjimce/article/details/50866313)

深度学习（二十九）Batch Normalization 学习笔记

Batch Normalization 学习笔记原文地址: http://blog.csdn.net/hjimce/article/  jzrita 2017年05月23日 16:53  215
details/50866313 作者: hjimce 一、背...

(http://blog.csdn.net/jzrita/article/details/72650826)