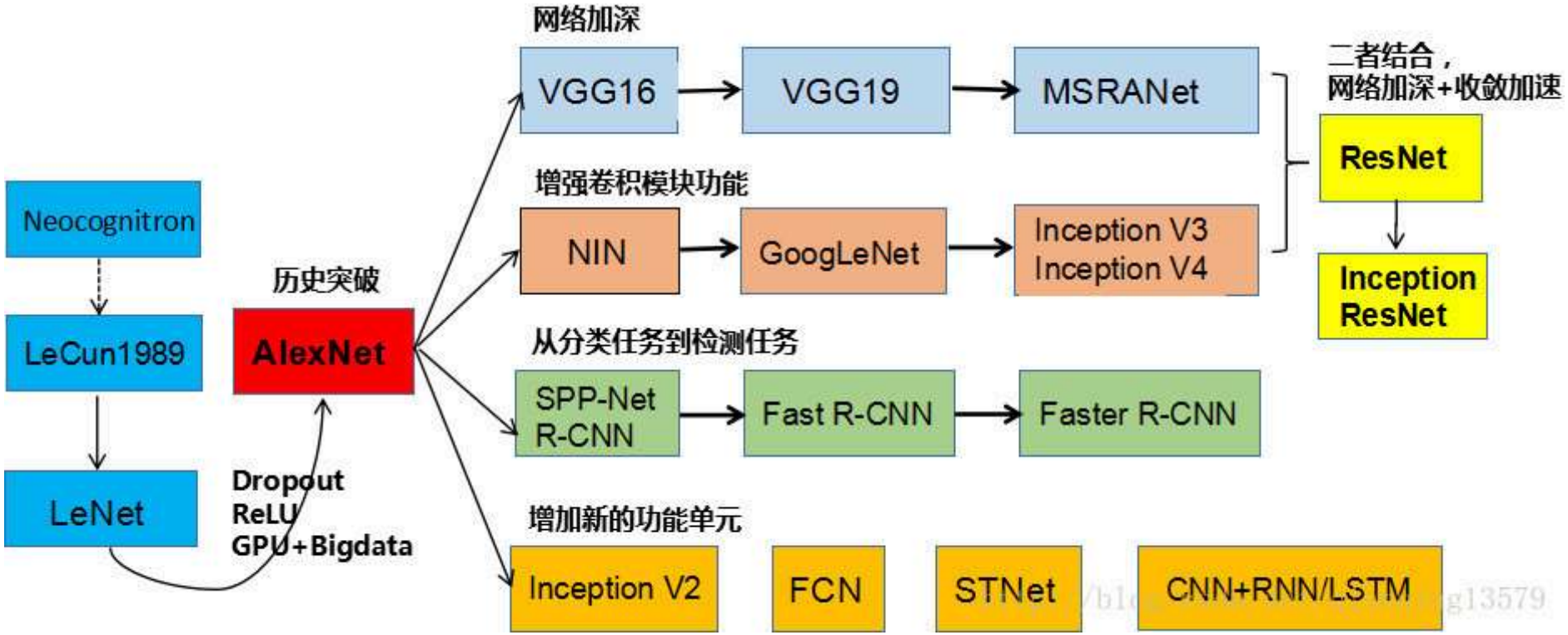
CNN论文阅读（一） LeNet：Gradient-based learning applied to document recognition

1、CNN结构演化历史图



CNN经典论文学习第一篇，卷积神经网络开山鼻祖，经典的手写体识别论文——LeNet:《Gradient-Based Learning Applied to Document Recognition》，作者包括深度学习三大巨头之一Yann Lecun，花书《深度学习》作者之一Yoshua Bengio。

原文篇幅很长，选择记录其中最重要的介绍CNN网络结构的第二章的A和B部分。

2、 用于字符识别的卷积神经网络

使用梯度下降法的多层网络可以从大量的数据中学习复杂的，高纬，非线性的映射，这使得他们成为图像识别任务的首选。在传统的模式识别的模型中，手工设计的特征提取器从图像中提取相关特征清除不相关的信息。分类器可以将这些特征进行分类。全连接的多层网络可以作为分类器。一

个更有意思的模式就是尽量依赖特征提取器本身进行学习。对于字符识别，可以将图像作为行向量作为输入输入到网络中。虽然这些任务(比如字符识别)可以使用传统的前向全连接网络完成。但是还存在一些问题。

首先，图像是非常大的，由很多像素组成。具有100个隐藏单元的全连接网络包含成千上万的权重，这么多参数提高了系统的消耗和内存占用，因此需要更大的训练集。但是没有结构的网络的主要缺点是，多于图像或者音频这些应用来说，不具备平移，形变扭曲的不变性。在输入到固定大小输入的网络钱，字符图像的大小必须归一化，并且放在输入的中间，不幸的是，没有哪种预处理能够达到如此完美：由于手写体以字符为归一化单位，会导致每个字符的大小，倾斜，位置存在变化，再加上书写风格的差异，将会导致特征位置的变化，原则上，足够大小的全连接网络可以对这些变化鲁棒，但是，要达到这种目的需要更多的在输入图像不同位置的神经元，这样可以检测到不同的特征，不论他们出现在图像的什么位置。学习这些权值参数需要大量的训练样本去覆盖可能的样本空间，在下面描述的卷积神经网络中，位移不变性(shift invariance 通过权值共享实现。

第2点，全连接的网络的另一个缺点就是完全忽略了输入的拓扑结构。在不影响训练的结果的情况下，输入图像可以是任意的顺序。然而，图像具有很强的二维局部结构：空间相邻的像素具有高度相关性。局部相关性对于提取局部特征来说具有巨大优势，因为相邻像素的权值可以分成几类。CNN通过将隐藏结点的感受野限制在局部来提取特征。

A 卷积网络

CNN通过局部感受野(local receptive fields)，权值共享(shared weights)，下采样(sub-sampling)实现位移，缩放，和形变的不变性(shift,scale,distortion invariance)。一个典型的用于字符识别的网络结构如图2所示，该网络结构称为LeNet-5。输入层输入大小归一化并且字符位于中间的字符图像。每一层的每个神经元(each unit)接受上一层中一组局部领域的神经元的输入(就是局部感受野)。将多个神经元连接为局部感受野的思想可以追溯到60年代的感知机，与Hubel and Wiesel’s在猫的视觉系统中发现的局部感受和方向选择的神经元几乎是同步的（神经网络和神经科学关系密切）。局部感受野在视觉学习神经模型中使用很多次了，使用局部感受野，神经元能够提取边缘，角点等视觉特征，这些特征在下一层中进行结合形成更高层的特征，之前提到，形变和位移会导致显著特征位置的变化，此外图像局部的特征检测器也可以用于整个图像，基于这个特性，我们可以将局部感受野位于图像不同位置的一组神经元设置为相同的权值(这就是权值共享)。每一层中所有的神经元形成一个平面，这个平面中所有神经元共享权值。神经元(unit)的所有输出构成特征图，特征图中所有单元在图像的不同位置执行相同的操作，这样他们可以在输入图像的不同位置检测到同样的特征，一个完整的卷积层由多个特征图组成(使用不同的权值向量)，这样每个位置可以提取多种特征。一个具体的示例就是图2 LeNet-5中的第一层，第一层隐藏层中的所有单元形成6个平面，每个是一个特征图。一个特征图中的一个单元对应有25个输入，这25个输入连接到输入层的5x5区域，这个区域就是局部感受野。每个单元有25个输入，因此有25个可训练的参数加上一个偏置。由于特征图中相邻单元以前一层中连续的单元为中心，所以相邻单元的局部感受野是重叠的。比如，LeNet-5中，水平方向连续的单元的感受野存在5行4列的重叠，之前提到过，一个特征图中所有单元共享25个权值和一个偏置，所以他们在输入图像的不同位置检测相同的特征，每一层的其他特征图使用不同的一组权值和偏置，提取不同类型的局部特征。LeNet中，每个输入位置会提取6个不同的特征。特征图的一种实现方式就是使用一个带有感受野的单元，扫面整个图像，并且将每个对应的位置的状态保持在特征图中，这种操作等价于卷积，后面加入一个偏置和一个函数，因此，取名为卷积网络，卷积核就是连接的权重。卷积层的核就是特征图中所有单元使用的一组连接权重。卷积层的一个重要特性是如果输入图像发生了位移，特征图会发生相应的位移，否则特征图保持不变。这个特性是CNN对位移和形变保持鲁棒的基础。

一旦计算出feature map,那么精确的位置就变得不重要了，相对于其他特征的大概位置是才是相关的。比如，我们知道左上方区域有一个水平线段的一个端点，右上方有一个角，下方垂直线段有一个端点，我们就知道这个数字是7。这些特征的精确位置不仅对识别没有帮助，反而不利于识别，因为对于不同的手写体字符，位置会经常变动。在特征图中降低特征位置的精度的方式是降低特征图的空间分辨率，这个可以通过下采样层达到，下采样层通过求局部平均降低特征图的分辨率，并且降低了输出对平移和形变的敏感度。LeNet-5中的第二个隐藏层就是下采样层。这个层包含了6个特征图，与前一层的6个特征图对应。每个神经元的感受野是2x2,每个神经元计算四个输入的平均，然后乘以一个系数，最后加上一个偏执，最后将值传递给一个sigmoid函数。相邻的神经元的感受野没有重叠。因此，下采样层的特征图的行和列是前一层特征图的一半。系数和偏置影响了sigmoid函数的效果。如果系数比较小，下采样层相当于对输入做了模糊操作。如果系数较大，根据偏置的值下采样层可以看成是“或”或者“与”操作。卷积层和下采样层是交替出现的，这种形式形成一个金字塔：每一层，特征图的分辨率逐渐减低，而特征图的数量逐渐增加。LeNet-5中第三个隐藏层(C3层)的每个神经元的输入可以来自前一层(S2)的多个特征图。卷积和下采样的结合的灵感来源于Hubel and Wiesel’s”简单”和”复杂”细胞的概念,虽然那个时候没有像反向传播的全局监督学习过程。下采样以及多个特征结合可以大大提高网络对几何变换的不变性。

由于所有的权值都是通过反向传播学习的，卷积网络可以看成是一个特征提取器。权值共享技术对降低参数的数量有重要的影响，同时权值共享技术减小了测试误差和训练误差之间的差距。LeNet-5包含了340908个连接，但是由于权值共享只包含了60000个可训练的参数。卷积神经网络以及被应用在多个领域，包括手写体识别，打印字符识别，在线手写体提识别，以及人脸识别。在单个时间维度上权值共享的卷积神经网络被称为延时神经网络(TDNNs),TDNNs已经被用在场景识别(没有下采样)[40],语音识别（没有下采样），独立的手写体字符识别[44]以及手势验证[45]。

B LeNet-5

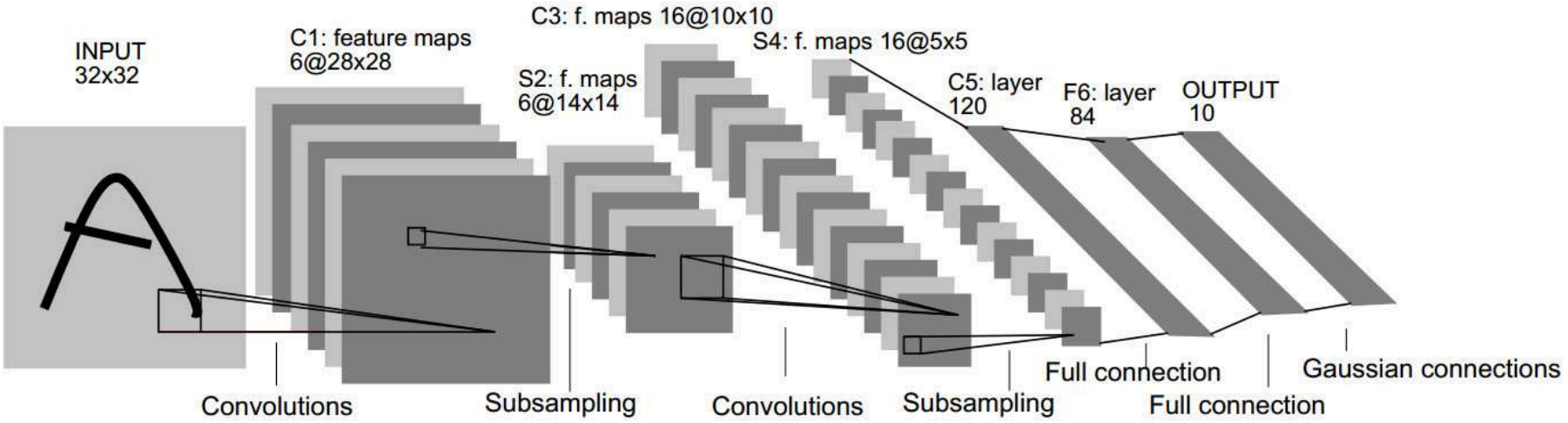


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

LeNet-5共有7层，不包含输入，每层都包含可训练参数（连接权重）。输入图像为32*32大小。这要比Mnist数据库（一个公认的手写数据库）中最大的字母还大（28*28）。这样做的原因是希望潜在的明显特征如笔画端点或角点能够出现在最高层特征监测器感受野的中心。在LeNet-5中，最后一层卷积层的感受野的中心在32x32的输入图像中形成了一个20x20的区域，输入像素值被归一化了，这样背景(白色)对应-0.1，前景(黑色)对应1.175.这使得输入的均值约等于0，方差约等于1，这样能够加速学习[46]。

下文中，卷积层标识为Cx,下采样层标识为Sx,全连接层标识为Fx,x标识层的索引。

C1层是一个卷积层，由6个特征图Feature Map构成。特征图中每个神经元与输入中5*5的邻域相连。特征图的大小为28*28，这样能防止输入的连接掉到边界之外。C1有156个可训练参数（每个滤波器5*5=25个unit参数和一个bias参数，一共6个滤波器，共(5*5+1)*6=156个参数），共122,304个连接（26*28*28*6，每个神经元对应26个连接，每个feature map有28*28个unit，一共有6个feature map）。（25个输入和1个偏置共26个连接，得到输出特征图里一个像素。）

S2层是一个下采样层，有6个14*14的特征图。特征图中的每个单元与C1中相对应特征图的2*2邻域相连接。S2层每个单元的4个输入相加，乘以一个可训练参数，再加上一个可训练偏置。结果通过sigmoid函数计算。可训练系数和偏置控制着sigmoid函数的非线性程度。如果系数比较小，那么运算近似于线性运算，下采样相当于模糊图像。如果系数比较大，根据偏置的大小下采样可以被看成是有噪声的“或”运算或者有噪声的“与”运算。每个单元的2*2感受野并不重叠，因此S2中每个特征图的行列分别是C1中特征图的一半。S2层有12个（池化层没有要学习的参数）可训练参数(每个feature map有一个系数和偏置)和5880（5*14*14*6）个连接。

C3是一个有16个特征图的卷积层。C3层的卷积核大小为5*5，每个特征图中的每个单元与S2中的多个特征图相连，表1显示了C3中每个特征图与S2中哪些特征图相连。

那为什么不把S2中的每个特征图连接到每个C3的特征图呢？原因有2点。

第一，不完全的连接机制将连接的数量保持在合理的范围内。

第二，也是更加重要的，其破坏了网络的对称性。不完全连接能够保证C3中不同特征图提取不同的特征（希望是互补的），因为他们的输入不同。

表1中展示了一个合理的连接方式：C3的前6个特征图以S2中3个相邻的特征图为输入。接下来6个特征图以S2中4个相邻特征图为输入，下面的3个特征图以不相邻的4个特征图为输入。最后一个特征图以S2中所有特征图为输入。这样C3层有1516个可训练参数((25*3+1)*6+(25*4+1)*9+(25*6+1))和151600个（C3层特征图大小10*10）连接。

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

TABLE I

EACH COLUMN INDICATES WHICH FEATURE MAP IN S2 ARE COMBINED BY THE UNITS IN A PARTICULAR FEATURE MAP OF C3.

（表1中第1列表示C3的第0个特征图，与S2中的第0,1,2个特征图连接）

S4层是一个下采样层，由16个5*5大小的特征图构成。特征图中的每个单元与C3中相应特征图的2*2邻域相连接，跟C1和S2之间的连接一样。S4层有32个可训练参数（每个特征图1个系数和一个偏置）和2000个连接(5*5*5*16,对于S4的每个unit,对应感受野4个参数，加上一个偏置)。

C5层是一个卷积层，有120个特征图。每个单元与S4层的全部16个特征图的5*5领域相连。由于S4层特征图的大小也为5*5（同滤波器一样），故C5特征图的大小为1*1：这构成了S4和C5之间的全连接。之所以仍将C5标示为卷积层而非全连接层，是因为如果LeNet-5的输入变大，而其他的保持不变，那么此时特征图的维数就会比1*1大。C5层有48120个可训练连接((5*5*16+1)*120)。

F6层有84个单元（之所以选这个数字的原因来自于输出层的设计，下面会有说明），与C5层全相连。有10164个可训练参数。

如同经典神经网络，F6层计算输入向量和权重向量之间的点积，再加上一个偏置。神经元i的加权和表示为aiai,然后将其传递给sigmoid函数产生单元i的一个状态，表示为xi,

xi=f(ai)

Sigmoid函数是一个双曲线正切函数：

f(a)=Atanh(Sa)

A表示函数的振幅，S决定了斜率，这个函数是一个奇函数，水平渐近线为+A，-A。常量A通常取1.7159。选择该函数的原因见附录A。

最后，输出层(其实就是softmax loss)由欧式径向基函数（Euclidean Radial Basis Function,RBF）单元组成，每类一个单元，每个单元有84个输入，每个RBF单元yiyi的输出按照如下方式计算：

$$yi=\sum(xj-\omega ij)2$$

$$yi=\sum(xj-\omega ij)2$$

换句话说，每个输出RBF单元计算输入向量和参数向量之间的欧式距离。输入离参数向量越远，RBF输出的越大。一个RBF输出可以被理解为衡量输入模式和与RBF相关联类的一个模型的匹配程度的惩罚项。用概率术语来说，RBF输出可以被理解为F6层配置空间的高斯分布的负的log似然(log-likelihood)。给定一个输入模式，损失函数应能使得F6的配置与RBF参数向量（即模式的期望分类）足够接近。这些单元的参数是人工选取并保持固定的（至少初始时候如此）。这些参数向量的成分被设为-1或1。虽然这些参数可以以-1和1等概率的方式任选，或者构成一个纠错码，但是被设计成一个相应字符类的7*12大小（即84）的格式化图片。这种表示对识别单独的数字不是很有用，但是对识别可打印ASCII集中的字符串很有用。基本原理就是字符是相似的，容易混淆，比如大小的O,小写的o和数字0或者小写的l与数字1，方括号和大写的I，会有相似的输出编码。如果一个系统与一个能够纠正此混淆的语言处理器相结合，这个就非常有用了。由于容易混淆的类别的编码是相似的，有歧义的字符的RBF输出是相似的，这个语言处理器就能够选择出合适的解释。图3给出了所有ASCII字符集的输出编码。

Fig. 3. Initial parameters of the output RBFs for recognizing the full ASCII set.

使用这种分布编码而非更常用的“1 of N”编码(又叫位置编码或者细胞编码)用于产生输出的另一个原因是，当类别比较大的时候，非分布编码的效果比较差。原因是大多数时间非分布编码的输出必须是关闭状态。这使得用sigmoid单元很难实现。另一个原因是分类器不仅用于识别字母，也用于拒绝非字母。使用分布编码的RBF更适合该目的，因为与sigmoid不同，他们在输入空间的较好得限制区域内兴奋，而非典型模式更容易落到外边。

RBF参数向量起着F6层目标向量的角色。需要指出这些向量的成分是+1或-1，这正好在F6 sigmoid的范围内，因此可以防止sigmoid函数饱和。实际上，+1和-1是sigmoid函数的最大曲率的点。这使得F6单元运行在最大非线性范围内。必须避免sigmoid函数的饱和，因为这将会导致损失函数较慢的收敛和病态问题。

3、重要的点

神经元是一个包含完整输入和输出完整过程的计算模型。

一个神经元对应一组权值（卷积神经网络中的卷积核+偏置，全连接网络中的权重w和偏置），执行一次计算 $y=f(\sum \omega ixi+b)$(CNN中的卷积计算,全连接网络中的权重计算)，产生一个输出（CNN中特征图的一个像素，全连接网络中的下一层的一个神经元）的过程。

————————————————————

posted on 2020-04-20 22:44 枫勤雪 阅读(2934) 评论(0) 编辑 收藏 举报