试论训练递归神经网络的di伍culty

Razvan Pascanu pascanur@iro .umontreal .ca
蒙特利尔大学

托马斯米科洛夫 t.mikolov@gmail.com
布尔诺大学

约书亚班吉奥 约书亚。bengio@umontreal.ca
蒙特利尔大学

摘要

在正确训练递归神经网络中有两个众所周知的问题，即Bengio等人详细介绍的消失梯度问题和爆炸梯度问题。(1994).在本文中，我们试图通过从分析、几何和动力系统的角度探讨这些问题来提高对潜在问题的理解。我们的分析被用来证明一个简单而有效的解决方案。本文提出了一种处理爆炸梯度的梯度范数剪切策略和处理梯度消失问题的软约束方法。我们在实验部分验证了我们的假设和提出的解决方案。

1．介绍

一个递归神经网络（RNN），e。g.图1，是80年代提出的一种神经网络模型。，1986年；Elman，1990；Werbos，1988)为建模时间序列。该网络的结构类似于一个标准的多层感知器，有区别的是，我们允许与时间延迟相关联的隐藏单元之间的连接。通过这些连接，该模型可以保留关于过去输入的信息，使其能够发现数据中可能彼此相距遥远的事件之间的时间相关性（这是正确学习时间序列的一个关键属性）。

虽然原则上循环网络是一个简单而强大的模型，但在实践中，不幸的是很难进行正确的训练。为什么这个模型如此笨拙的主要原因之一是梯度的消失

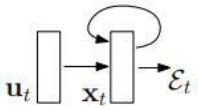

图1。一个递归神经网络的示意图。隐藏层中的循环连接允许信息从一个输入持续保存到另一个输入。

以及Bengio等人所描述的爆炸性梯度问题。(1994).

1．1．培训循环网络

一个输入值为u的通用递归神经网络t和状态xt对于时间步长t，由式(1)给出。在本文的理论部分，我们有时会利用方程（11）给出的特殊参数化[1]为了提供更精确的条件和直觉的日常用例。

(1)

(2)

该模型的参数由循环权值矩阵W给出rec，偏差b和输入权重矩阵w在，收集为一般情况。$\mathbf{x}_0$由用户提供，设置为零或学习，a是一个元素级函数（通常是tanh或s型）。成本E衡量网络在某些给定任务上的性能，它可以分解为每个步骤的单独成本对 $1 \leqslant t \leqslant T$ E_t，其中$E_t = L(x_t)$.

一种可以用于计算必要的梯度的方法是通过时间的反向传播（BPTT），其中递归模型被表示为

[1]这个公式等价于已知的方程$x_t = a(W_{rec}\mathbf{x}_t$ 更广泛的+ 一 1+ W在$\mathbf{u}_t$是为了方便而选择的。 b)，和它

在已展开的模型上应用了一个深度的多层模型（具有无限的层数）和反向传播(见图。2).

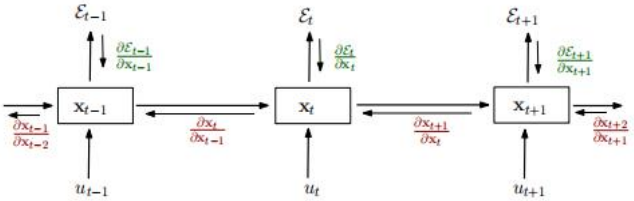

图2。通过为每个时间步长创建一个模型的副本，及时展开递归神经网络。我们用x表示t网络在t时刻的隐藏状态，由ut时间t和时间E的网络输入t从t时刻的输出中得到的误差。

在这一点上，我们将偏离经典的BPTT方程，并重写梯度（见方程(3)，(4)和(5)），以更好地突出爆炸的梯度问题。这些方程是通过写出梯度以乘积之和的形式得到的。

(3)

(4)

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{5}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right)$$

。θ

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1}))$$

;矩阵的形式，其中diag将一个向量转换为一个对角矩阵，和a\以元素级的方式计算a的导数。

$\frac{\partial \mathcal{E}_t}{\partial \theta}$请注意，方程(3)中的每个项都具有相同的形式，这些单独的项的行为决定了和的行为。从今以后，我们将关注这样一个通用术语，在没有混淆的情况下，将它简单地称为梯度。

$\frac{\partial \mathcal{E}_t}{\partial \theta}$任何梯度分量也是一个和（见方程(4)），我们将其术语称为时间贡献或时间分量。我们可以看到每一个这样的时间贡献$\frac{@E_t}{@\mathbf{x}_t} \frac{@\mathbf{x}_t}{@\mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta}$测量如何θ在步骤>k影响了步骤t > k的成本。因素$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}$（公式(5)）将误差“时间”从步骤t返回回步骤k。我们将进一步松散地区分长期和短期贡献，其中长期指的是短期和短期对其他一切的贡献。

2．爆炸和消失的梯度

如Bengio等人介绍。（1994），爆炸性梯度问题是指在训练过程中梯度常数的大幅增加。这类事件是由长期成分的爆炸引起的，长期成分可以比短期成分呈指数级增长。消失梯度问题指的是相反的行为，当长期成分以指数快到范数0，使得模型不可能学习时间遥远事件之间的相关性。

.1．2机械师

为了理解这一现象，我们需要观察每个时间分量的形式，特别是以t - k雅可比矩阵乘积形式存在的矩阵因子（见方程(5)）。$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}$同样地，t - k实数的乘积可以收缩到零或爆炸到永恒性，这个矩阵的乘积也可以（沿着某个方向v）。

在接下来的内容中，我们将尝试将这些直觉形式化(扩展了在Bengio等人的论文中所做的类似推导。（1994年），其中只考虑了一个隐藏单元的情况)。

如果我们考虑该模型的一个线性版本(即。对方程（11）)中的恒等函数设a，我们可以使用幂迭代方法正式分析雅可比矩阵的乘积，得到梯度爆炸或消失的严格条件（这些条件的详细推导见补充材料）。ffi它适合于最大的特征值λ1的循环权重矩阵应小于1，以使长期分量消失(如t! 1)和它必须大于1的梯度爆炸。

我们可以将这些结果推广到非线性函数a，其中a的绝对值\(x)是有界的（比如用一个值V 2 R），因此|diag(a\(xk))| ≤ V .

我们首先证明它对λ是合理的ffi1<$\frac{1}{}$，其中λ1是循环权值矩阵W的最大特征值的绝对值rec，导致消失梯度问题的发生。注意，我们假设由方程（11）给出的参数化。雅各布人

矩阵由W给出$\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k}^T$rec二氧化碳\(xk)).2-

这个雅可比矩阵的范数的有界是

这两个矩阵的范数（见等式(6)）。根据我们的假设，这意味着它小于1。

$$\forall k, \left\| \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\| \leq \left\| \mathbf{W}_{rec}^T \right\| \left\| diag(\sigma'(\mathbf{x}_k)) \right\| < \frac{1}{\gamma}\gamma < 1$$

η设2R是这样的，Ak，≤<1。$\left\|\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k}\right\|\eta$这个的存在性由方程(6)给出。$\eta$通过对i的归纳法，我们可以证明它

$$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}\left(\prod_{i=k}^{t-1}\frac{\partial \mathbf{x}_{i+1}}{\partial \mathbf{x}_i}\right) \leq \eta^{t-k}\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \qquad (7)$$

η作为<1，它可以得出，根据公式(7)，长期贡献（其中tk很大）进入—
0与tk呈指数级快。—样

通过反证明，得到了爆炸的必要条件，即最大特征值λ1$\frac{1}{\gamma}$大于（否则长期的成分将会消失，而不是爆炸）。

对于tanh，我们有V = 1，而对于s型型，我们有V=$^1/_4$.

2．2．绘制与动态图形的相似性

系统

我们可以通过使用动态系统的视角来提高我们对爆炸梯度和消失梯度问题的理解，就像之前在Doya（1993）中所做的那样；Bengio等人。(1993).

我们建议阅读Strogatz（1994）对动力系统理论的正式和详细的处理。对于任何参数分配，取决于初始状态xθ0，状态xt在映射F的重复应用下，自治动态系统的收敛收敛到几种可能的不同吸引子状态中的一种。点吸引子，尽管存在其他类型的吸引子)。该模型也可以在混沌状态下运行，在这种情况下，下面的一些观察结果可能不成立，但在这里没有深入讨论。吸引子描述了模型的渐近行为。状态空间被划分为吸引盆地，每个吸引子一个。如果模型在一个吸引盆地中启动，则随着t的增长，模型将收敛到相应的吸引子。

θ动力系统理论告诉我们，随着变化的缓慢，渐近行为几乎在任何地方都平滑地变化，除了某些发生剧烈变化的关键点（新的渐近行为在拓扑上不再与旧的等价）。这些点被称为分岔边界，它们是由出现、消失或改变形状的吸引子引起的。

（Doya，1993）假设这种分叉交叉可能会导致梯度爆炸。我们想将这一观察结果扩展到梯度爆炸的一个合理条件，因此，我们将重复使用（Doya，1993）中的单隐藏单元模型（和图）(见图。ffi3).

x轴包含参数b，y轴包含渐近状态x1.粗体线跟随着点吸引子x的移动1，随着b的变化。在b1我们有一个分岔边界，其中一个新的吸引子出现（当b从1减小），而在b2我们还有另一个方法，它会导致这两个吸引子之一的消失。在间隔（b1，b2)我们处于一个丰富的状态，其中有两个吸引子，它们之间的边界位置的变化，当我们变化b时，用虚线追踪出来。向量ield（灰色虚线箭头）描述了在该区域初始化网络时状态x的演化。

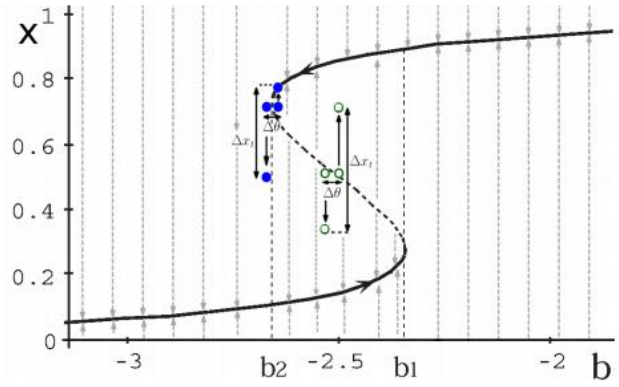

图3。单个隐藏单元RNN的分叉图（混合循环权重为5.0，可调偏差b；Doya（1993）中介绍的例子）。请参见文本。

我们证明了有两种类型的事件可能导致x的大变化t，有t ！1.一个是跨越吸引盆地之间的边界（用未填充的圆表示），而另一个是跨越分叉边界（未填充的圆）。对于大的t，xΔt即使b的变化很小（因为b的变化也会被不同的吸引子吸引），这也会导致很大的梯度。

然而，要使梯度爆炸，既没有必要也不可能跨越一个分岔，因为分岔是不可能在局部没有缺陷的全局事件。ffi学习在参数状态空间中追踪出一条路径。如果我们在一个分岔边界，但模型的状态是这样的，它在一个吸引子的吸引盆地，当分叉交叉时不会改变形状或消失，那么这个分岔将不会影响学习。

跨越吸引盆地之间的边界是一个局部事件，而梯度的爆炸是必然的。ff如果我们假设跨越到一个新兴吸引子或从一个消失（由于分支）符合穿越一些吸引子之间的边界，我们可以制定一个可靠的条件梯度爆炸封装的观察Doya（1993），扩展也正常跨越不同盆地之间的边界的吸引力。ff注意，在igure中，只有两个具有分叉的b值，但有一个整个范围的值可以有一个边界交叉。

先前分析的另一个局限性是，它们只考虑自治系统，并假设观察结果适用于输入驱动的模型。在(Bengio等。通过假设输入是有界噪声来处理它。这种方法的缺点是，它限制了人们对输入的推理方式。在实践中，输入应该驱动动力系统，能够使模型处于某种吸引子状态，或者在某些触发模式出现时，将其踢出吸引盆地。

我们建议通过将输入折叠到地图中，将我们的分析扩展到输入驱动模型。我们考虑映射F的族t，其中我们应用了一个不同的Ft在每一步。直观地说，为了使梯度爆炸，我们需要与之前相同的行为，其中（至少在某个方向上）映射F1，..，Ft同意并改变方向。图4描述了这种行为。

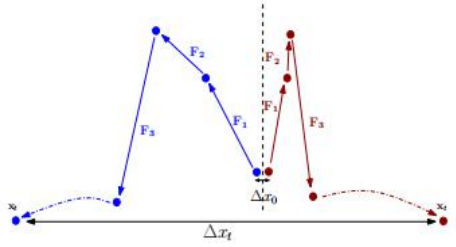

图4。这个图说明了x中的变化t， xΔt，对于一个小的x可以很大Δ0.蓝色和红色（左和右）的轨迹是由相同的地图F生成的1，F2，...对于两个不同的初始状态。

对于方程（11）所提供的特殊参数化，我们可以通过分解映射F来进一步进行类比t$\bar{F}$变成一个混合的地图和一个时变的Ut。 F（x）= Wreca（x）+ b对应于一个无输入的递归网络，而Ut(x) = x + W在ut描述了输入的效果。如图所示。5.自Ut随着时间的变化，它不能用标准的动力系统工具进行分析，但可以。$\bar{F}$这意味着，当一个景点的盆地之间的边界被跨越时 $\bar{F}$，国家将走向一个不同的吸引子，

对于大的t可能导致(除非输入映射Ut反对x的大差异吗t$\bar{F}$.因此，研究的渐近行为可以提供关于此类事件可能发生的地点的有用信息。

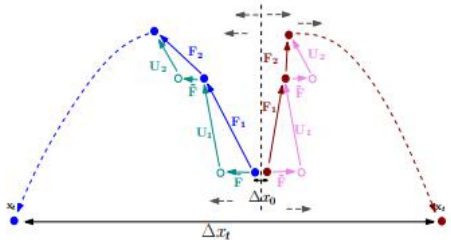

图5。说明了人们如何分解地图F1，F..t变成一个常数的地图和地图U$\bar{F}$1，..，Ut.垂直虚线表示吸引盆地之间的边界，直虚线箭头表示边界两侧地图的方向。$\bar{F}$这张图是对图的一个扩展。4.

从动力学系统的角度来看，关于消失的梯度的一个有趣的观察结果如下。如果因子趋于零（对于t-k很大），这意味着x$\frac{\partial x_t}{\partial x_k}$t不依赖于xk(如果我们改变xk通过一些，xΔt保持不变）。这就转化为x处的模型t接近收敛到某个吸引子(它将从x附近的任何地方到达k).

. 3 . 2几何解释

让我们考虑一个简单的单隐单元模型（方程(8)），其中我们提供了一个初始状态x0并训练模型在50步后得到一个特定的目标值。请注意，为了简单起见，我们假设没有输入。

$$x_t = w a\ (x_t\ —\ 1)\ +\ b \tag{8}$$

图6表示误差面E50= (a(x50) — 0.7)2，其中x0= .5和a是s型函数。

我们可以更容易地分析这个模型的行为，通过进一步将它简化为线性的（a然后是恒等函数），使用b = 0。xt= x0wt由此是= tx$\frac{\partial x_t}{\partial w}$0wt — 1和$\frac{\partial^2 x_t}{\partial w^2}$= t(t — 1)x0wt —2，这意味着当一阶导数爆炸时，二阶导数也会爆炸。在一般情况下，当梯度爆炸时，它们会沿着某些方向v。这说明，在这种情况下，存在一个向量v，v Cα $\frac{\partial \varepsilon_t}{\partial \theta} \geq$t，其中C，α 2 R和α > 1。对于线性情况（a是恒等函数），v是W的最大特征值对应的特征向量rec.如果这个界限是紧的，我们假设一般当梯度

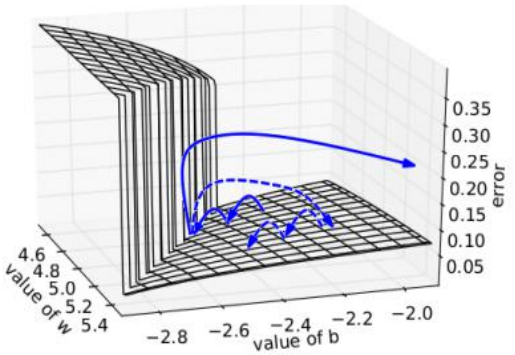

图6。我们绘制了单个隐单元循环网络的误差面，突出了高曲率壁的存在。实线描述了梯度下降可能遵循的标准轨迹。使用虚线箭头，图表显示了当梯度的范数超过一个阈值时，如果梯度被重新缩放到一个混合的大小会发生什么。

沿着v的曲率也会爆炸，导致误差面的一个壁，如图中所示。6.

如果这成立，那么它给了我们一个简单的爆炸梯度问题的解决方案。6.

如果曲率的梯度和前导特征向量都与爆炸方向v对齐，则误差面有一个垂直于v的陡峭壁（因此也垂直于梯度）。这意味着，当随机梯度下降（SGD）到达墙壁并进行梯度下降步骤时，它将被迫以垂直于陡峭墙壁的方式移动穿过山谷，可能离开山谷并破坏学习过程。

图中的虚线箭头。6对应于忽略这个大步骤的规范，确保模型保持靠近墙壁。关键的见解是，当梯度爆炸时所采取的所有步骤都与v对齐，而忽略了其他下降方向(i。e.该模型垂直于墙移动）。因此，在墙上，沿梯度方向的小标准步只是把我们推回到墙之外更平滑的低曲率区域，而规则的梯度步会把我们带得非常远，从而减缓或阻止进一步的训练。相反，通过一个有界的步骤，我们回到了在墙附近的平滑区域，在那里SGD可以自由地探索其他下降方向。

在这种情况下，对经典的高曲率谷的重要补充是，我们假设谷是宽的，因为我们在墙周围有一个很大的区域，如果我们着陆，我们可以依靠一阶方法向局部极小值移动。这就是为什么仅仅裁剪梯度可能是合理的，而不需要使用二阶方法。注意，这就是

即使梯度的增长率与曲率的增长率不相同（在这种情况下，由于梯度和曲率之间的比率仍然会爆炸，二阶方法就会失效）。

我们的假设也有助于理解与无黑森方法相比，其他二阶方法最近的成功。没有黑森和其他第二之间有两个关键的区别
排序算法。首先，它使用了完整的黑森矩阵，因此可以处理不一定是轴对齐的爆炸方向。其次，它在每个更新步骤之前计算黑森矩阵的新估计，并可以考虑曲率的突变（如我们的假设提出的），而大多数其他方法使用平滑假设，i。e.，在许多步骤中平均二阶信号。

3．处理爆炸和消失的梯度

.1．3以前的解决方案

对循环权值使用L1或L2惩罚可以帮助解决爆炸梯度。假设参数初始化的值较小，则其光谱半径为W_{rec}可能小于1，由此可以得出梯度不会爆炸（见第2.1节中发现的必要条件）。正则化项可以确保在训练过程中，光谱半径永远不会超过1。这种方法将模型限制在一个简单的范围内（在原点有一个单点吸引子），其中插入到模型中的任何信息都必须在时间上以指数速度消失。在这种情况下，我们不能训练一个发电机网络，也不能表现出长期的记忆痕迹。

Doya（1993）提出对模型进行预编程（在正确的制度下初始化模型）或使用教师强制使用。第一个建议假设，如果模型从一开始就表现出与目标所要求的相同的渐近行为，那么就不需要跨越一个分岔边界。缺点是，人们不可能总是知道所需的渐近行为，而且，即使这些信息是已知的，在这个特殊的情况下初始化一个模型也不是微不足道的。我们还应该注意到，这种初始化并不能阻止跨越吸引盆地之间的边界，如图所示，即使没有跨越分岔边界，也可能发生。

教师强迫是一个更有趣的方法，但又不是一个很容易理解的解决方案。它可以被看作是在正确的情况下初始化模型的一种方式

区域的空间。研究表明，在实践中，它可以减少梯度爆炸的机会，甚至允许训练生成器模型或使用无限内存的模型（Pascanu和Jaeger，2011；多雅和吉泽，1991）。一个重要的缺点是，它需要在每个时间步长中都确定一个目标。

在霍克雷特和施米德胡伯（1997）；Graves et al。（2009）提出了一个解决消失梯度问题的方法，其中模型的结构为
变化的具体地说，它引入了一组特殊的单元，称为LSTM单元，它们是线性的，与自身有一个循环连接，该单元相加为1。进入单元和来自单元的低信息由输入和输出门保护（它们的行为被学习）。这个基本原理有几个变体构造这个解决方案并没有明确地解决爆炸性梯度的问题。

苏茨克弗等人。（2011）使用无海森优化器与结构阻尼相结合，这是海森的一种特殊的阻尼策略。这种方法似乎能很好地处理消失的梯度，尽管仍然缺少更详细的分析。据推测，这种方法之所以有效，是因为在高维空间中，长期分量与短期分量正交的可能性很高。这将允许黑森人独立地重新调整这些分量。在实践中，我们不能保证这个财产是否成立。如第2.3节所述，这种方法也能够处理爆炸梯度。结构阻尼是一种增强，当参数变化较小时，迫使状态变化很小。$\Delta\theta$这就需要雅可比矩阵$\frac{\partial \mathbf{x}_{t}}{\partial \theta}$为了有一个小的规范，从而进一步帮助解决爆炸性的梯度问题。当在长序列上训练循环神经模型时，它很有帮助，这表明虽然曲率可能与梯度同时爆炸，但它可能不会以相同的速度增长，因此不适合处理爆炸梯度。ffi

""回波状态网络（卢科塞维丘斯和Jaeger，2009）通过不学习循环权值和输入权值来避免爆炸和消失的梯度问题。它们是从手工制作的分布图中取样的。由于通常循环权值的最大特征值小于1，因此输入模型的信息必须以指数速度消失。这意味着这些模型不能轻易地处理长期的依赖关系，即使其原因与消失的梯度问题略有不同。对经典模型的扩展是由泄漏积分单元表示的(Jaeger等。，2007)

$\mathbf{x}_{k}=\alpha x_{k}-1+(1-\alpha)a(W_{rec}\mathbf{x}_{k}-1+W在\mathbf{u}_{k}+b)$.

虽然这些单元可以用于解决霍克雷特和施米德胡伯（1997）提出的学习长期依赖的标准基准（见（Jaeger，2012））），但它们更适合处理低频信息，因为它们充当低通滤波器。因为大部分的重量都是随机分布的
目前还不清楚人们需要什么样的模型来解决复杂的现实世界的任务。

我们将特别注意到托马斯·米科洛夫在他的博士论文（Mikolov，2012）中提出的方法(并在语言建模的最新结果中隐含地使用(Mikolov等人。，2011))。它涉及到按元素方式剪切梯度的时间分量（当它超过绝对值的混合阈值时，剪切一个项）。剪切在实践中已经被证明做得很好，它构成了我们的方法的支柱。

3.2.缩放梯度

正如在第2.3节中所建议的，处理梯度范数突然增加的一个简单机制是当它们超过一个阈值时重新调整它们（参见算法1）。

算法1的伪代码，用于范数剪辑梯度，无论何时它们爆炸

$\hat{\mathbf{g}}\frac{\partial \mathcal{E}}{\partial \theta}$
$\hat{\mathbf{g}}\geq$如果||阈值
$\hat{\mathbf{g}}\frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
如果结束

这个算法非常类似于托马斯·米科洛夫提出的，我们只是偏离最初的提议，试图提供一个更好的理论基础（确保我们总是在下降方向对当前小批），尽管在实践中两个变体行为相似。

ffi所提出的剪切易于实现和计算效果显著，但它确实引入了一个额外的超参数，即阈值。设置这个阈值的一个很好的启发式方法是查看在通常的大量更新中的平均范数的统计数据。ffi在我们的实验中，我们注意到，对于给定的任务和模型大小，训练对这个超参数不是很敏感，即使在相当小的阈值下，算法也表现得很好。

该算法也可以被认为是基于梯度的范数来适应学习速率。与其他学习速率自适应策略相比，后者侧重于通过收集梯度上的统计数据来提高收敛性(例如在

Duchi等。（2011），或Moreira和Fiesler（1995）的概述），我们依赖于瞬时梯度。这意味着我们可以处理规范中非常突然的变化，而其他方法将不能这样做。

．3．3消失梯度正则化

我们选择使用一个正则化项来解决消失的梯度问题，它表示对参数值的偏好，这样反向传播的梯度既不会增加也不会减少太多。我们的直觉是增加了规范 $\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_k}$ 表示t时刻的误差对所有输入u更敏感t，..，uk $\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}$（是一个因素）。 $\frac{\partial \mathcal{E}_t}{\partial \mathbf{u}_k}$ 在实践中，其中一些输入将与t时刻的预测无关，并且会表现得像网络需要学习忽略的噪声。网络不能学会忽略这些不相关的输入，除非有一个错误信号。 $\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}$ 这两个问题不能并行解决，似乎自然期望我们需要迫使网络增加规范的代价更大的错误（由不相关的输入条目），然后等待它学习忽略这些无关的输入条目。这表明，在遵循误差E的下降方向时（例如，二阶方法会尝试做什么），因此我们需要通过正则化项来执行它。 $\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}$

我们在下面提出的正则化器更倾向于使误差信号在返回时间时保持范数的解决方案：

$$\Omega = \sum_k \Omega_k = \sum_k \left(\frac{\left\| \frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}} \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\|}{\left\| \frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}} \right\|} - 1 \right)^2 \tag{9}$$

为了在计算上很显著，我们只使用关于W的“直接”偏导数ffirec(我们认为xk $\frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}}$ 对于W是恒定的rec当计算中的导数时k)，如等式（10）所示。注意，我们使用了方程（11）的参数化。这可以简单地完成，因为我们从BPTT中得到的值。 $\frac{\partial \mathcal{E}}{\partial \mathbf{x}_k}$ 我们使用Theano来计算这些梯度(伯格斯特拉等人。，2010年；Bastien等人。，2012).

$$\begin{aligned} \frac{\partial^+ \Omega}{\partial \mathbf{W}_{rec}} &= \sum_k \frac{\partial^+ \Omega_k}{\partial \mathbf{W}_{rec}} \\ &= \sum_k \frac{\partial^+ \left(\frac{\left\| \frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_k)) \right\|}{\left\| \frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}} \right\|} - 1 \right)^2}{\partial \mathbf{W}_{rec}} \end{aligned} \tag{0}$$

请注意，我们的正则化项只强制使用Ja-
可比矩阵，以保持相关的范数 $\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k}$

误差的方向，而不是对任何方向(即。 $\frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}}$ 我们并不强制要求所有的特征值都接近于1)。
第二个观察结果是，我们使用的是一个软约束，因此我们不能保证误差信号的范数被保留下来。如果这些雅可比矩阵发生范数爆炸（随着tk增加），那么这可能导致导致梯度的问题，我们需要处理它，如3.2节所述。这可以从
动力系统的角度来看：防止梯度消失意味着我们推动模型等远离吸引子（这样它不收敛，梯度消失的情况）和更接近盆地之间的边界，使它更有可能的梯度爆炸。

4．实验和结果

．1．4.病理综合问题

正如在马滕斯和萨茨基弗（2011）中所做的那样，我们解决了霍克雷特和施米德胡伯（1997）提出的病理问题，这些问题需要学习长期的相关性。我们建议读者参考这篇原始论文对任务的详细描述，并参考补充材料来对实验设置的完整描述。

4 ．1 ．1 ．时间顺序问题

我们将时间顺序问题作为典型的病理问题，并将我们的结果扩展到其他提出的任务。输入是一长串离散的符号。在两个时间点（在序列的开始和中间），发出{A，B}中的一个符号。该任务包括对序列末尾的顺序（AA、AB、BA、BB）进行分类。

图7显示了标准SGD、SGD-c（带出裁剪策略的SGD增强）和SGDCR（带出裁剪策略和正则化项的SGD）的成功率。请注意，对于长度超过20的序列，消失梯度问题确保了SGD和SGD-C算法都不能解决该任务。x轴是基于对数尺度的。

这项任务提供了经验证据，表明爆炸性的梯度与需要长时间记忆痕迹的任务有关。我们知道，最初的模型运行在单吸引子制度(i。e. λ1< 1)，其中内存量由λ控制1.更多的内存意味着更大的光谱半径，当这个值超过某个阈值时，模型进入了梯度可能爆炸的丰富状态。我们在图中看到。7，只要消失的梯度问题

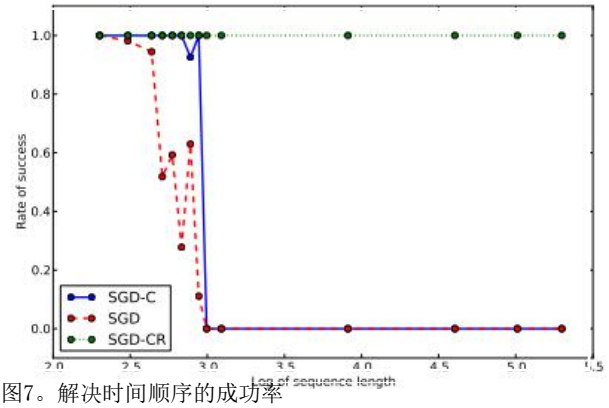

图7。解决时间顺序的成功率
问题与序列长度的日志之间。请参见文本。

lem不会成为一个问题，解决爆炸性梯度问题确保了更好的成功率。

当结合剪切和第3.3节中提出的正则化项时，我们称这种算法为SGD-CR。SGD-CR对高达200步的序列（马滕斯和Sutskever（2011）中使用的最大长度）的成功率为100%。此外，我们可以训练一个单一的模型来处理任何长度为50到200的序列（通过为不同的SGD步骤提供不同长度的序列）。

有趣的是，训练后的模型可以推广到新序列，可以是训练期间看到的序列的两倍。

4 .1 ..2 其他病理任务

SGD-CR也能够解决（100%成功的下面列出的长度，除了一个任务）其他病理问题（1997），即添加问题，乘法问题，3位时间顺序问题，随机排列问题和无噪声记忆问题在两个变量（当模式需要记忆5位长度和包含超过20位的信息；参见马滕斯和苏斯克弗（2011））。对于前4个问题，我们使用一个长度高达200的单一模型，而对于无噪声记忆，我们对每个序列长度（50、100、150和200）使用一个不同的模型。在8条路径中只有一条成功的最困难的问题是随机排列问题。在所有情况下，我们观察到成功地泛化到比训练序列更长的序列。在大多数情况下，这些结果在成功率方面优于马滕斯和苏茨克弗（2011），他们处理的序列比霍克雷特和施米德胡伯（1997）更长，与（Jaeger，2012）相比，它们可以推广到更长的序列。

表1。每个时间步长负对数似然复调音乐预测结果。降低是更好的。

数据集	数据折叠	SGD	SGD+C	SGD+CR
钢琴。德诺丁汉	列车	6.87	6.81	7.01
	测试	7.56	7.53	7.46
Muse数据	列车	3.67	3.21	3.24
	试验	3.80	3.48	3.46
		8.25	6.54	6.51
		7.11	7.00	6.99

表2。关于下一个熵的结果（位 字符预测 任务在/字符）

数据集	数据折叠	SGD	SGD+C	SGD+CR
1步	火车	1.46	1.34	1.36
	试验	1.50	1.42	1.41
5步	火车	N/A	3.76	3.70
	试验	N/A	3.89	3.74

.2.4自然问题

我们解决了复调音乐预测的任务，使用数据集钢琴-米迪。布朗格-莱万多夫斯基等人描述的博物馆数据。（2012年）和在宾夕法尼亚大学树库数据集上的字符级别上的语言建模(Mikolov等人。，2012).我们还探索了该任务的一个修改版本，其中我们要求模型预测未来的第5个字符（而不是下一个）。我们的假设是，要解决这个修改后的任务，长期相关性比短期相关性更重要，因此我们的正则化术语应该更有帮助。

表1中报告的训练和测试分数是每个时间步长的平均负对数似然值。除了正则化因子和剪切阈值外，我们在三次运行中裁剪了超参数。

SGD-CR在所有复调音乐预测任务上的改进，除了博物馆数据，我们得到了与现有的完全相同的性能(Bengio等人。，它使用了一个不同的架构。表2包含了语言建模的结果（每个字母的位）。

这些结果表明，剪切梯度解决了一个优化问题，而不是作为一个正则化器，因为训练误差和测试误差都普遍提高。宾夕法尼亚树库的研究结果达到了米科洛夫等人取得的最新水平。（2012），他使用了与我们类似的不同的剪切算法，从而提供了两者行为相似的证据。正则化模型的性能与无黑森训练模型一样好。

通过使用所提出的正则化项，我们甚至能够改进测试错误，即使是在任务上没有

以长期贡献为主的。

5．总结和结论

我们提供了不同的视角，通过这些视角，我们可以更多地了解爆炸和消失的梯度问题。为了处理爆炸梯度的问题，我们提出了一个解决方案，包括在爆炸梯度太大时裁剪爆炸梯度的范数。该算法的动机是假设当梯度爆炸时，曲率和高阶导数也会爆炸，我们在误差面面临一个特殊的模式，即具有单一陡峭壁的山谷。为了处理消失梯度问题，我们使用了一个正则化项，它迫使误差信号在随时间移动时不消失。这个正则化项迫使雅可比矩阵只在相关方向上保持范数。$\frac{\partial x_t}{\partial x_{t-1}}$在实践中，我们证明了这些解决方案提高了在所考虑的病理合成数据集以及复调音乐预测和语言建模上的性能。

致谢

我们也要感谢西奥诺开发团队（特别是弗雷德里克·巴斯蒂安，帕斯卡尔·兰布林和詹姆斯·伯格斯特拉）的帮助。

我们感谢NSERC，FQRNT，CIFAR，RQCHP和计算加拿大为他们提供的资源。

参考文献

巴斯蒂安、兰布林、帕斯卡纳、伯格斯特拉、古德费罗、伯杰隆、布沙尔、N.和本吉奥。(2012).新的功能和速度的改进。适用于深度学习和无监督特征学习NIPS 2012研讨会。

本吉奥，弗拉斯科尼，和西马德。(1993).循环网络中长期依赖关系的学习问题。第1183-1195页，旧金山。IEEE出版社。特邀报告

本吉奥，Y，西马德，P。和弗拉斯科尼，P。(1994).通过梯度下降学习长期依赖关系是一种的。ffi《IEEE《神经网络学报》，5(2)，157-166。

本吉奥，布兰格-莱万多夫斯基和帕斯卡努。(2012).优化循环网络的进展。技术报告，arXiv：1212.0901，U。蒙特利尔

伯格斯特拉、布鲁勒、巴斯丁、兰布林、帕斯卡努、德斯贾丁、图里安、沃德法利、D.和本吉奥。(2010).Theano：CPU和GPU数学表达式编译器。在科学计算会议的Python论文集（SciPy）中。口头陈述。

布兰格-莱万多夫斯基，N.，本吉奥，Y.，和文森特，P。(2012).高维序列的时间依赖建模：在复调音乐生成和转录中的应用。在二十九届机器学习国际会议（ICML‘12）的会议记录中。ACM。

Doya，K.(1993).梯度下降学习中递归神经网络的分岔。IEEE神经网络学报，1,75-80。

Doya，K.和吉泽，S。(1991).神经和物理振荡器的自适应同步。在J。E.穆迪。J.汉森和R。李普曼，编辑，NIPS，第109-116页。摩根考夫曼。

杜奇，J。C.，哈赞和辛格。(2011).在线学习和随机优化的自适应次梯度方法。《机器学习研究杂志》，12,2121-2159。

埃尔曼，J。(1990).及时找到结构。认知科学，14(2)，179-211。

格雷夫斯，利维基，M.，费尔南德斯，贝尔托拉米，邦克，H.和施米德胡贝尔，J。(2009).一种新型的无约束手写识别的连接主义系统。IEEE《模式分析与机器智能学报》，31(5)，855-868。

HochreiterS。和施密德胡伯，J。(1997).长期记忆。神经计算法，9(8)，1735-1780年。

杰格，H。(2012).回波状态网络中的长短期记忆：模拟研究的细节。技术报告，不来梅雅各布斯大学。

杰格，H。卢科塞维丘斯，西维特，大学。(2007).带有泄漏积分器神经元的回波状态网络的优化与应用。神经网络，20(3)，335-352。

Lukosevicius，M.“JaegerH。(2009).递归神经网络训练的储层计算方法。《计算机科学评论》，3(3)，127-149。

马滕斯，J。和苏特斯克弗，我。(2011).使用无海森优化技术学习递归神经网络。在程序中。ICML’2011.ACM。

米科洛夫，T。(2012).基于神经网络的统计语言模型。菲律宾D.论文，布尔诺理工大学。

米科洛夫，T.，多拉斯，A.，孔布林克，S.，伯吉特，L.，和塞诺基，J。(2011).经验的评估和先进的语言建模技术的结合。在程序中。第12届国际演讲传播协会年会（2011年演讲间会议）。

米科洛夫，T.，苏茨克弗，我，德拉斯，A.，勒，科姆布林克，S.和塞诺基，J。.-(2012).用神经网络进行子词语言建模。预先印好的http://www.它vutbr.cz/imikolov/rnnlm/char.pdf).

莫雷拉，M。和费斯勒，E。(1995).具有自适应学习率和动量项的神经网络。1995年，瑞士马尔蒂尼，IDIAP。

Pascanu, R.JaegerH。(2011).一种工作记忆的神经动力学模型。神经网络。，24，199 - 207.

Rumelhart, D.E.，辛顿，G。E.，和威廉姆斯，R。J.(1986).通过反向传播错误来学习表示法。自然，323（6088），533-536。

斯特罗加茨。(1994).非线性动力学和混沌：应用于物理、生物、化学和工程（非线性的研究）。研究在非线性。珀尔修斯图书集团，1版。

萨茨克弗，我，马滕斯和辛顿。(2011).用递归神经网络生成文本。指导中尉Getoor和T。谢弗，编辑，第28届机器学习国际会议的会议记录（ICML11），ICML '11，第1017-1024页，纽约，纽约，美国。ACM。

沃博斯，P。J.(1988).反向传播的推广，并应用于循环天然气市场模型。神经网络，1(4)，339-356。

爆炸和消失梯度问题的分析分析

$$\mathbf{x}_t = W_{rec}a(x_t - 1) + W在\mathbf{u}_t + b \quad (11)$$

让我们考虑一下术语$g_k^T = \frac{\partial E_t}{\partial \mathbf{x}_t}\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}\frac{\partial^+ \mathbf{x}_k}{\partial \theta}$为
方程（11）中的参数化的线性版本。设a为恒等函数），假设t为初等，1==。我们有：

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \left(\mathbf{W}_{rec}^T\right)^l \quad (12)$$

通过基于一般幂迭代方法的证明，可以证明在一定条件下，
$\frac{\partial E_+}{\partial \mathbf{x}_-}(\mathbf{W}_{rec}^T)1$呈指数级增长。

证明让W_{rec}具有特征值$\lambda 1, .., \lambda n$与$|\lambda 1| > |\lambda 2..| > > |\lambda n|$和相应的特征向量q1，q2，..，qn它形成了一个向量基。我们现在可以把行向量写入到这个基中：$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}$

$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} = 对 \sum_{i=1}^{N} c_i \mathbf{q}_i^T$

如果j是这样的，那么$c_j \neq 0$和任何$j\backslash < j, c_j\backslash = 0$，使用$\mathbf{q}_i^T \mathbf{W} \lambda \mathbf{q}_{rec}^{T\ l} = {}_i^T{}_i^T$我们有这个

$$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = c_j\lambda_j^l\mathbf{q}_j^T + \lambda_j^l \sum_{i=j+1}^{n} c_i \frac{\lambda_i^l}{\lambda_j^l}\mathbf{q}_i^T \approx c_j\lambda_j^l\mathbf{q}_j^T \quad (13)$$

我们使用的事实是$'\lambda i/\lambda j'<1$表示i > j，这意味着
$\lim l!1 '\lambda i/\lambda j'^l = 0$.如果$|\lambda j\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k}| > 1$，它随l呈指数快速增长，而且确实如此
所以沿着方向qj.
样

证明假定W_{rec}为了简单起见，是否可以对角化，尽管使用的是约当范型W_{rec}我们可以通过不仅考虑最大特征值的特征向量，还可以考虑由具有相同（最大）特征值的特征向量所张成的整个子空间来扩展这个证明。

这一结果为梯度的增长提供了一个必要的条件，即W的光谱半径（最大特征值的绝对值）rec必须大于1。

如果qj不在的空空间中$\frac{\partial^+ \mathbf{x}_k}{\partial \theta}$整个时间
分量随l呈指数增长。这种方法可以很容易地扩展到整个梯度。如果我们用W的特征分解来重写它，我们得到：

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{j=1}^{n}\left(\sum_{i=k}^{t} c_j\lambda_j^{t-k}\mathbf{q}_j^T\frac{\partial^+ \mathbf{x}_k}{\partial \theta}\right) \quad (14)$$

我们现在可以选择j和k，使$c_j q_j^T \frac{\partial^+ x_k}{\partial \theta}$没有0规范，而最大化$|\lambda_j|$.如果对于所选的j，它保持该$|\lambda_j|>$是1，然后是$\lambda_j^{t-k} c_j q_j^T \frac{\partial^+ x_k}{\partial \theta}$遗嘱主体
确定和，因为这个项随t以指数快速增长到初始值，所以和也会发生同样的情况。

实验装置

注意，所有超参数都是基于使用网格搜索在验证集上的性能选择的。

病理合成任务

下面的所有任务都使用了类似的成功标准（借用马滕斯和萨斯克弗（2011）），即模型在10000个测试样本上的误差不超过1%。在所有的情况下，离散符号都用一个热编码来描述，在回归的情况下，对给定的se-进行预测如果误差小于0.04，则认为定量为成功。

附加问题

输入由一系列随机数组成，其中两个随机位置（一个在开始，一个在序列的中间）被标记。该模型需要在看到整个序列后，预测这两个随机数的总和。对于每个生成的序列，我们采样长度为$T\backslash\frac{11}{10}$虽然为了清晰起见，我们把T称为论文中序列的长度。第一个位置从[1，]中采样，而第二个位置从[，]中采样。$\frac{T'}{10}\frac{T'}{10}\frac{T'}{2}$这些位置i，j在哪里被标记在不同的输入通道中，除了两个采样位置在哪里都是0。该模型需要预测在采样位置i，j处发现的随机数之和除以2。

为了解决这个问题，我们使用了一个50个隐藏单元模型，具有一个tanh激活函数。.01学习速率设置为，正则化项前的因子α为0.5。我们在梯度的范数上使用裁剪阈值为6的裁剪。权值初始化从一个均值为0的正态分布和标准推导。1.
该模型是在长度T在50到200之间的序列上进行训练的。我们成功地获得了解决这个任务的100%的成功率，这优于马滕斯和萨斯基弗（2011）（使用黑森自由）的结果，我们看到随着序列的长度接近200，成功率下降

步骤霍克雷特和施米德胡伯（1997）只考虑了最多100步的序列。Jaeger（2012）也以100%的成功率解决了这个任务，尽管该解决方案似乎不能很好地推广，因为它依赖于非常大的输出权值，对于esn来说，这通常是不稳定的标志。我们使用一个单一的模型来处理所有长度的序列（50、100、150 200），训练后的模型可以推广到可以增加400步的新序列（而误差仍然低于1%）。

乘法问题

这个任务类似于上面的问题，只是预测值是随机数的乘积，而不是总和。我们使用了与前面的情况相同的超参数，并得到了非常相似的结果。

时间顺序问题

对于时间顺序，序列的长度是T，我们有两个符号{aB}和4个干扰符号{cdef}。序列条目除了来自两个随机位置，第一个位置来自[，]，而第二个位置来自[，]。$\frac{T}{10}\frac{2T}{10}\frac{4T}{10}\frac{5T}{10}$任务是预测非干扰物符号被提供的顺序，i。e.{AAABBABB}。

.001我们使用一个50隐藏单元模型，学习速率为α，正则化共，设置为2。ffi剪切梯度范数的阈值值留为6。至于另外两个任务，我们在训练一个单一模型来处理50到200步之间的序列时，我们有100%的成功率。由于成功率，这方面优于以前的技术水平，但单一模型也可以推广到更长的序列（最多400步）。

3位时间顺序问题

与前一个类似，除了我们有3个随机位置，首先从[，]，第二个从[，]，最后从[，]。$\frac{T}{10}\frac{2T}{10}\frac{3T}{10}\frac{4T}{10}\frac{6T}{10}\frac{7T}{10}$

我们使用了与上面类似的超参数，但是我们将隐藏层的大小增加到100个隐藏单元。与之前一样，我们在训练一个能够泛化到新序列长度的单一模型时，表现超过了最先进的水平。

随机排列问题

在这种情况下，我们有一个包含100个符号的字典。除了从{1,2}采样的具有相同值的第一个和最后一个位置外，其他条目都运行-

多姆利选自[3,100]网站。任务是做下一个符号预测，尽管唯一可预测的符号是最后一个。

.001我们使用一个100个隐藏单元，学习速率为α，正则化共，设置为1。ffi切割阈值留为6。ffi这项任务更需要学习，8个实验中只有1个成功了。与之前一样，我们使用单一模型来处理T的多个值（从50到200个单位）。

无噪声记忆问题

对于无噪声记忆，我们得到了一个长度为5的二进制模式，然后是常值的T步。在这些T步骤之后，模型需要生成最初看到的模式。我们还考虑了马滕斯和萨斯克弗（2011）对这个问题的扩展，其中模式的长度为10，符号集的基数为5而不是2。

我们在这些任务上管理了100%的成功率，尽管我们为所考虑的5个序列长度（50、100、150、200）训练了一个不同的模型。

自然任务

复调音乐预测

我们训练我们的模型，一个s型单位RNN，在200步的序列上。ffi在所有情况下，共切阈值都是相同的，即8（注意，在计算梯度时，必须取序列长度的平均值）。

在Piano-midi.de数据集的情况下，我们使用300个隐藏单元和初始学习率为1.0（每次在一个时期内的误差增加而不是减少时，学习率就会减半）。对于正则化模型，我们使用正则化系数α的初始值为0.5，其中α遵循一个1/t的时间表，i。e.$\alpha_t=\frac{1}{2t}$（其中t测量时代的数量）。

对于诺丁汉数据集，我们使用了完全相同的设置。对于博物馆数据，我们将隐藏层增加到400个隐藏单元。学习率也下降到0.5。对于正则化模型，α的初始值为0.1，而$\alpha_t=\ .\frac{1}{t}$

我们观察到，对于自然任务，使用一个减少正规化术语的时间表似乎很有用。我们假设正则化项迫使模型关注长期相关性，而代价是短期相关性，所以有这个衰减因子可能是有用的
该模型可以更好地利用短期信息。

语言建模

对于语言建模任务，我们使用了一个没有偏差的500个s型隐藏单元模型(Mikolov等人。，2012).该模型通过200步的序列进行训练，其中隐藏状态从一个步骤延续到下一个步骤。

我们对所有实验使用45的阈值切割（尽管我们取序列长度的成本之和）。对于下一个字符预测，当我们使用没有正则化项的剪切时学习率为0.01，当我们添加正则化项时学习率为0.05，当我们不使用剪切时学习率为0.001。在预测未来的第5个字符时，我们使用0.05的学习率和正则化项和0。1没有它。

下一个字符预测的正则化因子α被设置为并保持不变，而对于修改后的任务，我们使用一个调度的初始值为0.05。.01$\frac{1}{t}$