**反向传播技术在手写邮政编码识别中的应用**

**通过提供任务域的约束条件，可以极大地增强学习网络的泛化能力。这篇论文演示了如何通过网络的架构将这些约束集成到一个反向传播网络中。该方法已成功地应用于美国邮政提供的手写邮政编码数字的识别。、从标准化的字符图像到最终分类，由单个网络学习整个识别操作。**

**1简介：**

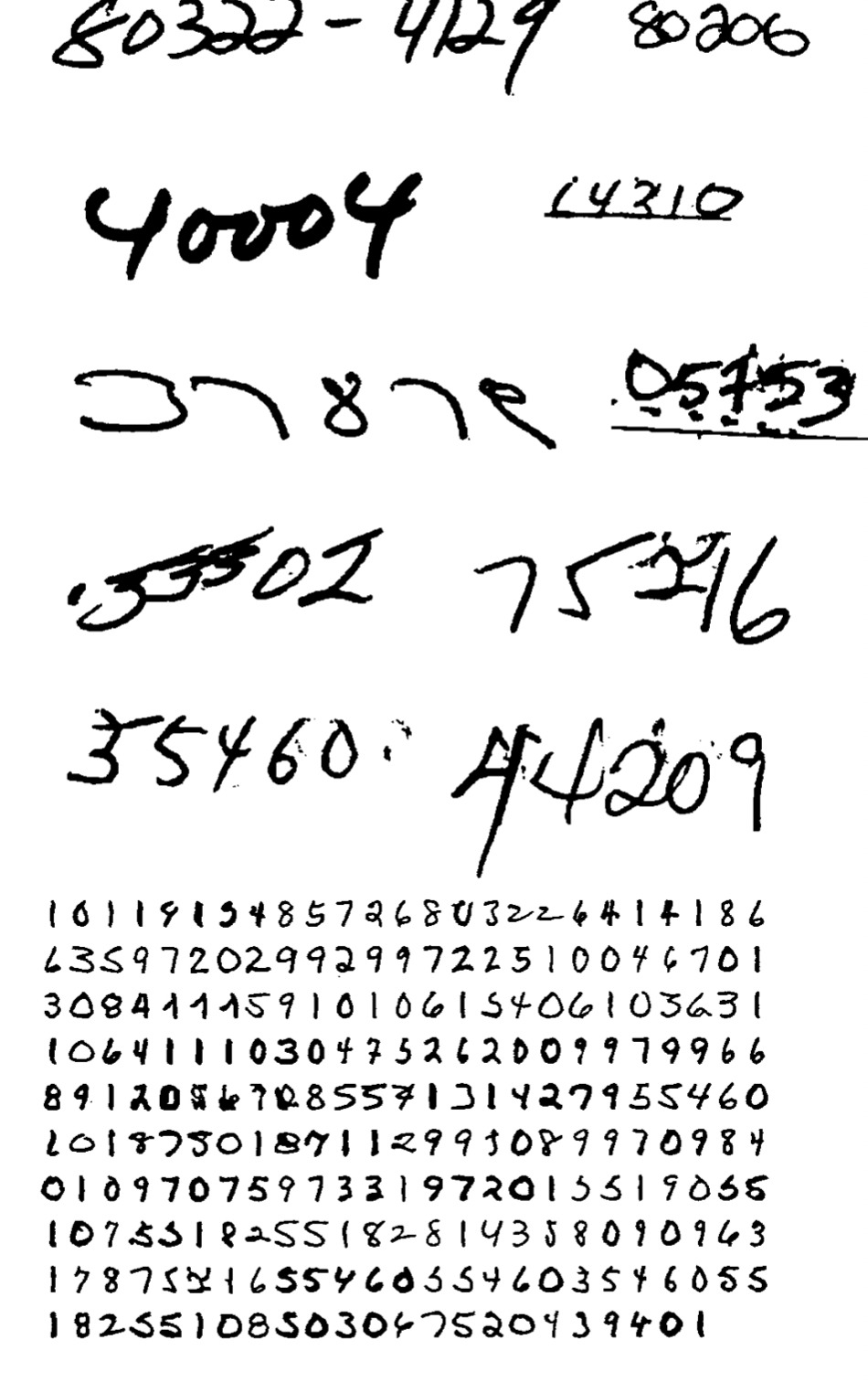
**先前关于识别简单数字图像的工作表明，通过设计一个包含了关于该任务的一定量的先验知识的网络结构，可以对复杂任务获得很好的泛化。其基本设计原则是在不过度降低网络计算能力的前提下，尽可能地减少网络中自由参数的数量。这一原则的应用增加了正确泛化的可能性，因为它产生了一个熵降低的专门网络结构，降低了Vapnik-Chervonenkis维度(Baum and Haussler 1989)。（所以通俗的理解： VC dimension就是某类模型对数据数量的包容性。VC dimension越高，就说明包容性越强。）**

**在这篇论文中，我们应用反向传播算法(Rumelhart et al. 1986)到一个现实世界的问题来：识别从美国邮件中提取的手写数字。与我们小组之前在这个问题上的研究结果(Denker et al. 1989)不同的是，学习网络是直接输入图像而不是特征向量，这证明了反向传播网络有处理大量低级信息的能力。**

**2邮政编码：**

**2.1数据集**

**用于训练和测试网络的数据库由9298个分段数字组成，这些数字是流经纽约布法罗邮局的美国邮件上手写邮政编码的数字。图1显示了这些图像的示例。这些数字是由不同的人书写的，使用不同的大小、书写风格和工具，书写的谨慎程度也大不相同;7291例用于训练网络，2007例用于测试泛化性能。该数据库的一个重要特征是，训练集和测试集都包含许多不明确、不可分类甚至错误分类的示例。**



**2.2预处理**

**邮政服务承包商(Wang and Srihari 1988)负责在信封上定位邮政编码，并将每个数字与相邻的数字区分开来，这是一项非常艰巨的任务。此时，数字图像的大小会发生变化，但通常在40 × 60像素左右。然后应用线性变换使图像符合16 × 16像素图像。这种变换保留了字符的长宽比，并在图像中的无关标记被移除后执行。由于线性变换，得到的图像不是二值化的，而是有多个灰度级别，因为原始图像中可变数量的像素可以落在目标图像中的给定像素中。每个图像的灰度级别被缩放并转换为-1到1的范围。**

**3网络设计**

**3.1输入和输出**

**识别的其余部分完全由多层网络来完成。网络中的所有连接都是自适应的，尽管受到严重的限制，并且使用反向传播进行训练。这与早期的研究相反，最初的几层连接是手工选择的常数，在一个神经网络芯片上实现。网络的输入是一个16 × 16的归一化图像。输出由10个单元组成(每个类一个单元)，并使用位置编码。**

**3.2特征映射和权重共享**

**视觉模式识别的经典研究已经证明了提取局部特征并将其结合形成高阶特征的优点。这样的知识可以很容易地应用到网络中，迫使隐藏的单位只结合局部的信息来源。一个物体的不同特征可以出现在输入图像的不同位置。因此，拥有一组特征检测器似乎是明智的，它可以检测到输入平面上任何地方的特定特征实例。由于特征的精确定位与分类无关，在这个过程中我们可以承受丢失一些位置信息的代价。然而，近似位置信息必须保留，以便下一层能检测到更高阶、更复杂的特征。**

**使用“权重共享”技术，可以很容易地检测到输入端任意位置的特定特征。权重共享由Rumelhart等人(1986)为所谓的T-C问题提出，它包括由单个参数(权重)控制的几个连接。它可以被解释为在各连接之间施加相等约束。这种技术可以用很少的计算开销来实现。**

**权值共享不仅大大减少了网络中空闲参数的数量，而且可以表达任务的几何和拓扑结构。在我们的例子中，第一个隐藏层是由几个我们称之为特征映射的平面组成的。平面上的所有单元共享同一组权重，从而在不同的位置检测到相同的特征。由于特征的确切位置并不重要，所以特征映射不需要像输入那样有那么多的单元。**

**3.3网络架构**

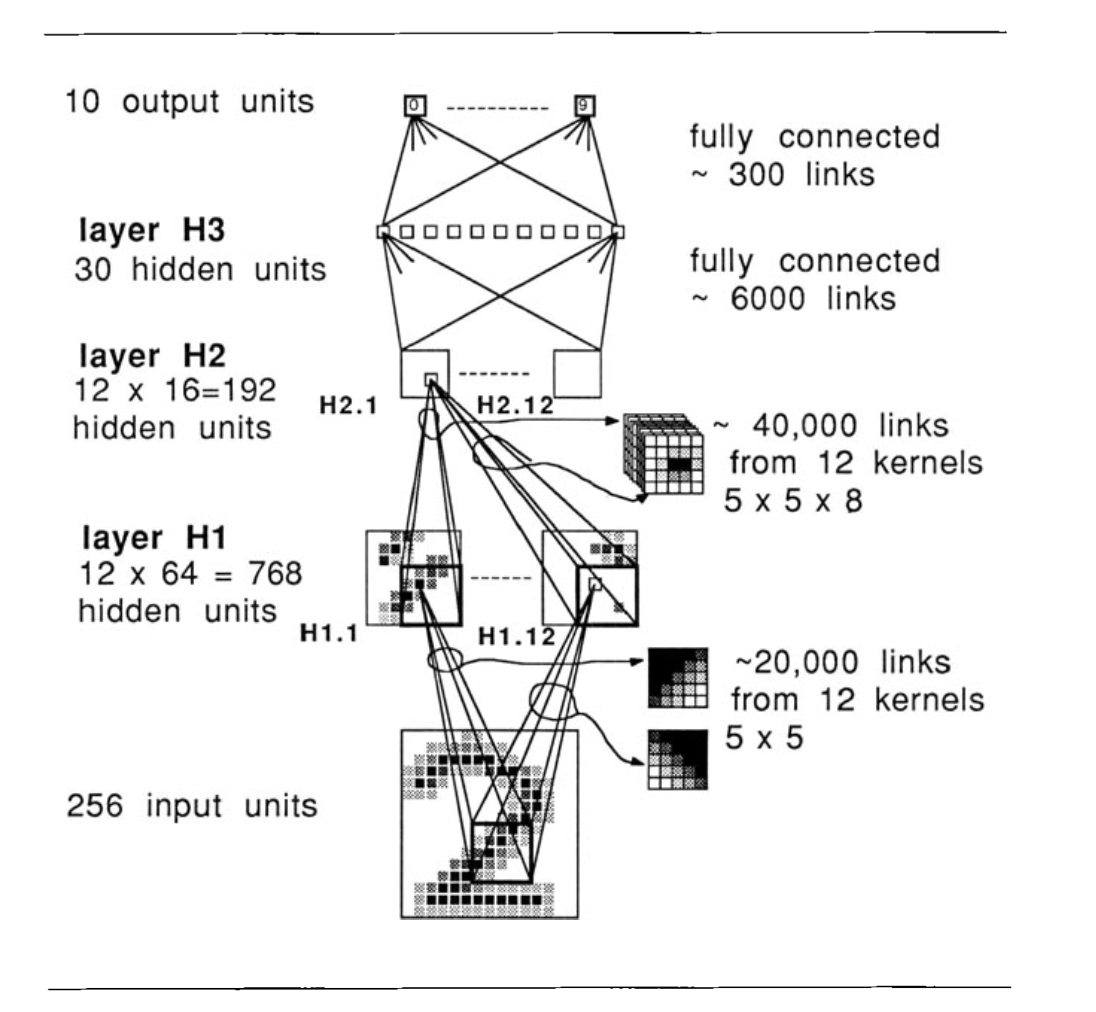
**网络如图2所示。它的架构是对LeCun(1989)提出的架构的直接扩展。网络有三个隐藏层，分别命名为H1、H2和H3。进入H1和H2的连接是本地的，并且受到很大的限制。**

**H1由12组64个单位组成，按12个独立的8 x 8特征图排列。这12张特征图被命名为H1.1,…,H1.12。特征图中的每个单元由输入平面上的一个5乘5的邻域进行输入。对于层H1中相隔一个单位的单元，它们的接收域（在输入层中）相隔两个像素。 因此，输入图像被欠采样，并且一些位置信息被消除。 从层H1到H2发生了类似的二对一欠采样。 其动机是需要高分辨率来检测特征的存在，而无需以同样高的精度确定其确切位置。**

**我们还知道，在图像的一个地方很重要的特征可能在其他地方也很重要。因此，给定特征映射中每个单元上的对应连接被约束为具有相同的权值。换句话说，H1.1中的64个单元。使用相同的一组25个权重。每个单元对图像的相应部分执行相同的操作。因此，由特征映射执行的函数可以被解释为一个5 × 5核的非线性下采样卷积。**

**当然，另一个映射（例如H1.4）中的单位共享另一组25个权重。 单位不共享其偏置（阈值）。 每个单元因此具有25条输入连接和一个偏置。 延伸超出输入平面边界的连接从虚拟背景平面获取其输入，该虚拟背景平面的状态等于恒定的预定背景水平（在我们的情况下为-1）。 因此，层H1包含768个单元（8 \*8\*12、19,968个连接（768 \*26），但由于许多连接共享相同的权重，因此仅包含1068个自由参数（768个偏差加上25\*12个特征核）。**

**图层H2也由12个特征图组成。每个特征图包含16个单元，排列在一个4乘4的平面上。和之前一样，这些特征图被指定为H2.1, H2.2，…，H2.12。H1和H2之间的连接模式与输入和H1之间的连接模式非常相似，但稍微复杂一些，因为H1有多个二维映射。H2中的每个单元都结合了来自H1中12个不同特征映射图中的8个的局部信息。它的感受区由8个5乘5的相邻单元组成，围绕在8张地图中相同位置的单位。因此，H2中的一个单位有200个输入、200个权重和一个偏差。同样，给定特征图中的所有单位都有相同的权重向量。H1中的8个映射是根据一种方案选择的，这里不作描述。对于脱离边界的连接的处理跟H1中相同。总而言之，层H2包含192个单元(12 × 4 × 4)，层H1和层H2之间总共有38592个连接(192× 201)。所有这些连接仅由2592个自由参数控制(12个特征映射乘以200个权重加上192个偏差)。**



**4实验环境：**

**所有的模拟都是使用运行在SUN-4/260上的反向传播模拟器SN进行的。**

**在每个节点上使用的非线性函数是一个缩放的双曲正切函数。这种对称函数被认为可以产生更快的收敛，尽管如果一些权值太小，学习可能会非常慢。在S形的准线性范围内选择输出单位的目标值。 这样可以防止权重无限期增长，并且可以防止输出单元在S型曲线的平坦位置工作。输出成本函数是均方误差。**

**训练前，用随机值初始化权值，使其均匀分布在-2.4/F和2.4/F之间，其中F为连接所属单元的输入数(fan-in)。这种技术倾向于将总输入保持在Sigmoid的操作范围内。**

**在每次学习实验中，这些模式都以恒定的顺序重复呈现。权重的更新是根据所谓的随机梯度或“在线”程序(在单个模式的每次呈现后更新)，而不是“真正的”梯度过程(在更新权重之前对整个训练集进行平均)。从实证研究(有理论论据支持)发现，随机梯度比真梯度收敛得快得多，特别是在大型冗余数据库上。它还能找到更可靠的解决方案。**

**所有的实验都是用牛顿算法的一个特殊版本来完成的——使用了Hessian矩阵的正对角线近似。该算法不被认为会带来学习速度的巨大提高，但它可以可靠地收敛，不需要大量的参数调整。**

**5结果**

**每次通过训练集之后，在训练集和测试集上都测量性能。该网络使用训练集训练了23次(167,693个模式演示)。**

**在这23次训练之后，模式和输出单元上的平均误差在训练集上是wpsoffice，在测试集上是 wpsoffice。训练集错误数为0.14%(10个错误)，测试集错误数为5.0%(102个错误)。从图3中可以看出，该收敛速度非常快，表明反向传播可以在较合理的时间内训练较大的数据集。这部分归因于实际数据的高度冗余。**

**在现实的应用程序中，用户通常感兴趣的是达到给定精度所需的拒绝数，而不是原始的错误率。我们测量了必须被拒绝的测试模式的百分比，以便在剩余的测试图片上获得1%的错误率。我们的主要拒绝标准是，两个最活跃的单元的活动水平之间的差异应超过给定的阈值。**

**在其余（未拒绝）测试图片上，对于1％的分类错误，拒绝的百分比为12.1％。应该强调的是，拒绝阈值是使用测试集上的性能指标获得的。**

**通过网络合成的某些神经元可以解释为特征检测器，与发现于生物视觉系统中的，或设计为先前的人工字符识别器的检测器非常相似，例如空间导数估计器或偏心/环绕型特征 探测器。**

**大多数错误分类是由于错误地将图像分割成单个字符。分割是一个非常困难的问题，尤其是当字符大量重叠时。其他错误是由于训练集中的图案太过模凌两可，低分辨率效果或写作风格所致。**

**其他特征图较少的网络也尝试过，但结果更差。也尝试了各种全连接、无约束网络，但泛化性能相当差。例如，一个隐含层为40个单元(共10690个连接)的全连接网络给出了如下结果:训练集的错误分类率为1.6%，测试集的错误分类率为8.1%，其余测试模式的错误率为1%，拒绝率为19.4%。完整的比较研究将在另一篇文章中描述。**

**5.1与其他网络的比较**

**在我们以前的系统中，处理的前几个阶段涉及到卷积，其中系数是手工设计的。在本系统中，网络的前两层被约束为卷积，但是系统会自动学习组成内核的系数。这种“受限反向传播”是本系统成功的关键：它不仅建立了平移不变性，而且极大地降低了熵，Vapnik-Chervonenkis维数和自由参数的数量，从而成比例地减少了为达到给定泛化水平性能所需的数据训练集。现在的系统比以前的系统稍微好一些。这是值得注意的，因为有关这个问题的具体信息在网络中构建得非常少。此外，通过设计更多连接和更少自由参数的专门化体系结构，新方法似乎有更大的改进潜力。**

**Waibel在1989年描述了一个大型网络（但与我们的网络相比仍然很小），该网络具有约18,000个连接和1800个自由参数，在一个语音识别任务上训练。 由于培训时间过长（在Alliant小型超级计算机上为18天），因此他建议从较小的，经过单独培训的网络中构建网络。 我们不需要这样的模块化构建过程，因为在Sun工作站上，我们的培训时间只要3天，而且无论如何，尚不清楚如何将我们的问题划分为可单独培训的子问题。**

**5.2数据信号处理（DSP）的实现**

**在识别过程中，几乎所有的计算时间都花在了乘法累加运算上，这是专门为数字信号处理器(DSP)设计的任务。我们使用了一个现成的板，它包含256k字节的本地内存和一个AT&T DSP- 32c通用DSP，在32位浮点数(25 MFLOPS)上的最高性能为每秒1250万次乘法运算。DSP作为协处理器运行;主机是一台个人计算机(PC)，它也包含一个连接到摄像机的视频采集板。**

**个人计算机使用自适应阈值技术将图像数字化并二值化。阈值图像然后被扫描，并给切每个连接的部分(或段)被隔离。太小或太大的部分被丢弃;剩下的部分送到DSP进行归一化和识别。PC机将单个数字的可变大小像素图表示给DSP，由DSP进行归一化和分类。**

**包括图像采集在内的数字识别器的总吞吐量为每秒10到12个分类，主要受到归一化步骤的限制。对于标准化数字，DST每秒执行30个以上的分类。**

**6总结**

**我们已经成功地将反向传播学习应用到一个大型的现实任务中。我们的结果在数字识别方面似乎处于最先进的水平。我们的网络是在一个低层次表示的数据上训练的，它有最少的预处理(相对于复杂的特征提取)。网络有许多连接，但相对较少的自由参数。设计了网络结构和权值约束，将任务的几何知识纳入系统。由于数据的冗余性质和网络上的约束，考虑到训练集的大小，学习时间相对较短。缩放特性远远好于仅根据反向传播对较小的人为问题进行的推断得出的结果。**

**通过反向传播学习获得的最终连接和权重网络很容易在商业数字信号处理硬件上实现。 从相机到分类图像的吞吐率都超过了每秒10位。**

**这项工作指出了设计灵活的“网络设计”软件工具的必要性，这些工具可以简化复杂的专用网络体系结构的设计。**

**7致谢**

**感谢美国邮政服务公司及其承包商为我们提供的数据库。 神经网络模拟器SN是LCon-Yves Bottou和Yann LeCun合作的结果。**

**8参考文献**

**Baum, E. B., and Haussler, D. 1989. What size net gives valid generaliztion? Neural Comp. 1,151-160.**

**Becker, S., and LeCun, Y. 1988. Improving the Convergence of Back-Propagation Learning With Second-Order Methods. Tech. Rep. CRG-TR-88-5, University of Toronto Connectionist Research Group.**

**Bottou, L.-Y., and LeCun, Y. 1988. Sn: A simulator for connectionist models. In Proceedings of NeuroNimes 88, Nimes, France.**

**Denker, J., Schwartz, D., Wittner, B., Solla, S. A., Howard, R., Jackel, L., and Hopfield, J. 1987. Large automatic learning, rule extraction and generaliza- tion. Complex Syst. 1,877-922.**

**Denker, J. S., Gardner, W. R., Graf, H. P., Henderson, D., Howard, R. E., Hub- bard, W., Jackel, L. D., Baird, H. s.,and Guyon, I. 1989. Neural network recognizer for hand-written zip code digits. In D. Touretzky, ed., Advances in Neural Information Processing Systems, pp. 323-331. Morgan Kaufmann, San Mateo, CA.**

**Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Bid. Cybemet. 36,193-202.**

**Hubel, D. H., and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. of Physiol. 160,106-154.**

**LeCun, Y. 1987. ModPles connexionnistesde l'apprentissage. Ph.D. thesis, Uni- versit6 Pierre et Marie Curie, Paris, France.**

**LeCun, Y. 1989. Generalization and network design strategies. In Connectionism in Perspective, R. Pfeifer, Z. Schreter, F. Fogelman, and L. Steels, eds. North- Holland, Amsterdam.**

**Mozer, M. C. 1987. Early parallel processing in reading: A connectionist ap- proach. In Attention and Performance, XII: The Psychology of Reading, M. Colt- heart, ed., Vol. XII, pp. 83-104. Erlbaum, Hillsdale, NY.**

**Patarnello, S., and Carnevali, P. 1987. Learning networks of neurons with boolean logic. Europhys. Lett. 4(4), 503-508.**

**Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In Parallel Distributed Processing: Ex- plorations in the Microstructure of Cognition, D. E. Rumelhart and J. L. Mc- Clelland, eds., Vol. I, pp. 318-362. Bradford Books, Cambridge, MA.**

**Tishby, N., Levin, E., and Solla, S. A. 1989. Consistent inference of probabilities in layered networks: Predictions and generalization. In Proceedings of the International Joint Conference on Neural Networks, WashingtonDC.**

**Waibel, A. 1989. Consonant recognition by modular construction of large phone- mic time-delay neural networks. In Advances in Neural Information Processing Systems, D. Touretzky, ed., pp. 215-223. Morgan Kauhann, San Mateo, CA.**

**Wang, C. H., and Srihari, S. N. 1988. A framework for object recognition in a visually complex environment and its application to locating address blocks on mail pieces. Int. I. Computer Vision 2, 125.**