第一章 绪论

> 一个人在不接触对方的情况下，通过一种特殊的方式，和对方进行一系列的问答。如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那么就可以认为这个计算机是智能的。
>
> — Alan Turing [1950]，《机器能思维吗?》

让机器具备智能是人们长期追求的目标，但是关于智能的定义也十分模糊。Alan Turing 在 1950 年提出了著名的图灵测试：“一个人在不接触对方的情况下，通过一种特殊的方式，和对方进行一系列的问答。如果在相当长时间内，他无法根据这些问题判断对方是人还是计算机，那么就可以认为这个计算机是智能的”。

要通过真正地通过图灵测试，计算机必须具备理解语言、学习、记忆、推理、决策等能力。这也延伸出很多不同的学科，比如机器感知（计算机视觉、自然语言处理），学习（模式识别、机器学习、增强学习），记忆（知识表示）、决策（规划、数据挖掘）等。所有这些分支学科都可以看成是**人工智能**（Artificial Intelligence，AI）的研究范畴。其中，**机器学习**（Machine Learning，ML）因其在很多领域的出色表现逐渐成为热门学科。机器学习的主要目的是设计和分析一些**学习算法**，让计算机从数据中获得一些决策函数，从而可以帮助人们解决一些特定任务，提高效率。对于人工智能来说，机器学习从一开始就是一个重要的研究方向，并涉及了概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科。

人工神经网络（Artificial Neural Network，ANN），也简称**神经网络**，是众多机器学习算法中比较接近生物神经网络特性的数学模型。人工神经网络通过

模拟生物神经网络（大脑）的结构和功能，由大量的节点（或称“神经元”，或“单元”）和之间相互联接构成，可以用来对数据之间的复杂关系进行建模。

Rosenblatt [1958] 最早提出可以模拟人类感知能力的数学模型，并称之为**感知器**（Perceptron），并提出了一种接近于人类学习过程（迭代、试错）的学习算法。但感知器因其结构过于简单，不能解决简单的异或（XOR）等线性不可分问题，造成了人工神经领域发展的长年停滞及低潮。直到 1980 年以后，Geoffrey Hinton、Yann LeCun 等人将**反向传播算法**（Backpropagation，BP）引入到多层感知器 [Williams and Hinton, 1986]，人工神经网络才又重新引起人们的注意，并开始成为新的研究热点。但是，2000 年以后，因为当时计算机的计算能力不足以支持训练大规模的神经网络，并且随着支持向量机（Support Vector Machines，SVM）等方法的兴起，人工神经网络又一次陷入低潮。

直到 2006 年，Hinton and Salakhutdinov [2006] 发现多层前馈神经网络可以先通过逐层预训练，再用反向传播算法进行精调的方式进行有效学习。并且近年来计算机计算能力的提高（大规模并行计算，GPU），计算机已经可以训练大规模的人工神经网络。随着深度的人工神经网络在语音识别 [Hinton et al., 2012] 和图像分类 [Krizhevsky et al., 2012] 等任务上的巨大成功，越来越多的人开始关注这一个“崭新”的研究领域：深度学习。目前，深度学习技术在学术界和工业界取得了广泛的成功，并逐渐受到了高度重视。

深度学习（Deep Learning，DL）是从机器学习中的人工神经网络发展出来的新领域。早期所谓的“深度”是指超过一层的神经网络。但随着深度学习的快速发展，其内涵已经超出了传统的多层神经网络，甚至机器学习的范畴，逐渐朝着人工智能的方向快速发展。

深度学习的一个内在的特性是可以自动学习特征（数据的表示）。在传统机器学习中，除了学习算法外，特征也是影响最终学习效果的重要因素，甚至在很多的任务上比算法重要。因此，要提高开发一个实际系统，人们往往需要去花费大量的精力去尝试设计不同的特征以及特征组合，来提高最终的系统能力，这就是所谓的“特征工程”问题。因此，人们也越来越关心如何从数据中自动学习有效的特征，也叫作“表示学习”（Representation Learning）。深度学习技术在一定程度上可以看作是一个表示学习技术，通过多层的非线性转换，把原始数据变成为更高层次、更抽象的表示。这些表示可以替代人工设计的特征。

本书主要介绍人工神经网络与深度学习中的基础知识、主要模型（卷积神经网络、递归神经网络等）以及在计算机视觉、自然语言处理等领域的应用。

总结和深入阅读

若希望全面了解人工神经网络和深度学习的知识，可以参考如下材料：

1. Ian Goodfellow, Aaron Courville, and Yoshua Bengio. Deep learning. Book in preparation for MIT Press, 2015. URL http://goodfeli.github.io/dlbook/
2. Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009

另外，网站也给出很好的教程，比如http://deeplearning.net/等。

参考文献

James A Anderson and Edward Rosenfeld. *Talking nets: An oral history of neural networks.* MiT Press, 2000.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994.

Yoshua Bengio, Jean-Sébastien Senécal, et al. Quick training of probabilistic neural nets by importance sampling. In *AISTATS Conference*, 2003.

Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8624–8628. IEEE, 2013.

C.M. Bishop. *Pattern recognition and machine learning.* Springer New York., 2006.

P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, and J.C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–479, 1992.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002.

Hal Daumé III. A course in machine learning. http://ciml.info/. [Online].

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, 2nd edition, 2001. ISBN 0471056693.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in Neural Information Processing Systems*, pages 472–478, 2001.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

Ian Goodfellow, Aaron Courville, and Yoshua Bengio. Deep learning. Book in preparation for MIT Press, 2015. URL http://goodfeli.github.io/dlbook/.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

M.I. Jordan. *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 456–464. Association for Computational Linguistics, 2010.

Marvin Minsky and Papert Seymour. Perceptrons. 1969.

Marvin L Minsky and Seymour A Papert. *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*. MIT press Boston, MA:, 1987.

T.M. Mitchell. *Machine learning*. Burr Ridge, IL: McGraw Hill, 1997.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*, 2012.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

Albert BJ Novikoff. On convergence proofs for perceptrons. Technical report, DTIC Document, 1963.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(3):328–339, 1989.

Paul Werbos. Beyond regression: New tools for prediction and analysis in the behavioral sciences. 1974.

Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

DE Rumelhart GE Hinton RJ Williams and GE Hinton. Learning representations by back-propagating errors. *Nature*, pages 323–533, 1986.

Matthew D Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.