# 1

# Introduction

## 1.1 Definitions and background

Since 2006, deep structured learning, or more commonly called deep learning or hierarchical learning, has emerged as a new area of machine learning research [20, 163]. During the past several years, the techniques developed from deep learning research have already been impacting a wide range of signal and information processing work within the traditional and the new, widened scopes including key aspects of machine learning and artificial intelligence; see overview articles in [7, 20, 24, 77, 94, 161, 412], and also the media coverage of this progress in [6, 237]. A series of workshops, tutorials, and special issues or conference special sessions in recent years have been devoted exclusively to deep learning and its applications to various signal and information processing areas. These include:

- 2008 NIPS Deep Learning Workshop;
- 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications;
- 2009 ICML Workshop on Learning Feature Hierarchies;

- 2011 ICML Workshop on Learning Architectures, Representations, and Optimization for Speech and Visual Information Processing;
- 2012 ICASSP Tutorial on Deep Learning for Signal and Information Processing;
- 2012 ICML Workshop on Representation Learning;
- 2012 Special Section on Deep Learning for Speech and Language Processing in IEEE Transactions on Audio, Speech, and Language Processing (T-ASLP, January);
- 2010, 2011, and 2012 NIPS Workshops on Deep Learning and Unsupervised Feature Learning;
- 2013 NIPS Workshops on Deep Learning and on Output Representation Learning;
- 2013 Special Issue on Learning Deep Architectures in IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI, September).
- 2013 International Conference on Learning Representations;
- 2013 ICML Workshop on Representation Learning Challenges;
- 2013 ICML Workshop on Deep Learning for Audio, Speech, and Language Processing;
- 2013 ICASSP Special Session on New Types of Deep Neural Network Learning for Speech Recognition and Related Applications.

The authors have been actively involved in deep learning research and in organizing or providing several of the above events, tutorials, and editorials. In particular, they gave tutorials and invited lectures on this topic at various places. Part of this monograph is based on their tutorials and lecture material.

Before embarking on describing details of deep learning, let's provide necessary definitions. Deep learning has various closely related definitions or high-level descriptions:

- ***Definition 1***: A class of machine learning techniques that exploit many layers of non-linear information processing for

supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification.

- ***Definition 2***: "A sub-field within machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture. Most of these models are based on unsupervised learning of representations." (Wikipedia on "Deep Learning" around March 2012.)
- ***Definition 3***: "A sub-field of machine learning that is based on learning several levels of representations, corresponding to a hierarchy of features or factors or concepts, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts. Deep learning is part of a broader family of machine learning methods based on learning representations. An observation (e.g., an image) can be represented in many ways (e.g., a vector of pixels), but some representations make it easier to learn tasks of interest (e.g., is this the image of a human face?) from examples, and research in this area attempts to define what makes better representations and how to learn them." (Wikipedia on "Deep Learning" around February 2013.)
- ***Definition 4***: "Deep learning is a set of algorithms in machine learning that attempt to learn in multiple levels, corresponding to different levels of abstraction. It typically uses artificial neural networks. The levels in these learned statistical models correspond to distinct levels of concepts, where higher-level concepts are defined from lower-level ones, and the same lower-level concepts can help to define many higher-level concepts." See Wikipedia http://en.wikipedia.org/wiki/Deep_learning on "Deep Learning" as of this most recent update in October 2013.
- ***Definition 5***: "Deep Learning is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals: Artificial

> Intelligence. Deep Learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text." See https://github.com/lisa-lab/DeepLearningTutorials

Note that the deep learning that we discuss in this monograph is about learning with deep architectures for signal and information processing. It is not about deep understanding of the signal or information, although in many cases they may be related. It should also be distinguished from the overloaded term in educational psychology: "Deep learning describes an approach to learning that is characterized by active engagement, intrinsic motivation, and a personal search for meaning." http://www.blackwellreference.com/public/tocnode?id=g9781405161251_chunk_g97814051612516_ss1-1

Common among the various high-level descriptions of deep learning above are two key aspects: (1) models consisting of multiple layers or stages of nonlinear information processing; and (2) methods for supervised or unsupervised learning of feature representation at successively higher, more abstract layers. Deep learning is in the intersections among the research areas of neural networks, artificial intelligence, graphical modeling, optimization, pattern recognition, and signal processing. Three important reasons for the popularity of deep learning today are the drastically increased chip processing abilities (e.g., general-purpose graphical processing units or GPGPUs), the significantly increased size of data used for training, and the recent advances in machine learning and signal/information processing research. These advances have enabled the deep learning methods to effectively exploit complex, compositional nonlinear functions, to learn distributed and hierarchical feature representations, and to make effective use of both labeled and unlabeled data.

Active researchers in this area include those at University of Toronto, New York University, University of Montreal, Stanford University, Microsoft Research (since 2009), Google (since about 2011), IBM Research (since about 2011), Baidu (since 2012), Facebook (since 2013), UC-Berkeley, UC-Irvine, IDIAP, IDSIA, University College London, University of Michigan, Massachusetts Institute of

Technology, University of Washington, and numerous other places; see http://deeplearning.net/deep-learning-research-groups-and-labs/ for a more detailed list. These researchers have demonstrated empirical successes of deep learning in diverse applications of computer vision, phonetic recognition, voice search, conversational speech recognition, speech and image feature coding, semantic utterance classification, natural language understanding, hand-writing recognition, audio processing, information retrieval, robotics, and even in the analysis of molecules that may lead to discovery of new drugs as reported recently by [237].

In addition to the reference list provided at the end of this monograph, which may be outdated not long after the publication of this monograph, there are a number of excellent and frequently updated reading lists, tutorials, software, and video lectures online at:

- http://deeplearning.net/reading-list/
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Recommended_Readings
- http://www.cs.toronto.edu/~hinton/
- http://deeplearning.net/tutorial/
- http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

## 1.2 Organization of this monograph

The rest of the monograph is organized as follows:

In Section 2, we provide a brief historical account of deep learning, mainly from the perspective of how speech recognition technology has been hugely impacted by deep learning, and how the revolution got started and has gained and sustained immense momentum.

In Section 3, a three-way categorization scheme for a majority of the work in deep learning is developed. They include unsupervised, supervised, and hybrid deep learning networks, where in the latter category unsupervised learning (or pre-training) is exploited to assist the subsequent stage of supervised learning when the final tasks pertain to classification. The supervised and hybrid deep networks often have the

same type of architectures or the structures in the deep networks, but the unsupervised deep networks tend to have different architectures from the others.

Sections 4–6 are devoted, respectively, to three popular types of deep architectures, one from each of the classes in the three-way categorization scheme reviewed in Section 3. In Section 4, we discuss in detail deep autoencoders as a prominent example of the unsupervised deep learning networks. No class labels are used in the learning, although supervised learning methods such as back-propagation are cleverly exploited when the input signal itself, instead of any label information of interest to possible classification tasks, is treated as the "supervision" signal.

In Section 5, as a major example in the hybrid deep network category, we present in detail the deep neural networks with unsupervised and largely generative pre-training to boost the effectiveness of supervised training. This benefit is found critical when the training data are limited and no other appropriate regularization approaches (i.e., dropout) are exploited. The particular pre-training method based on restricted Boltzmann machines and the related deep belief networks described in this section has been historically significant as it ignited the intense interest in the early applications of deep learning to speech recognition and other information processing tasks. In addition to this retrospective review, subsequent development and different paths from the more recent perspective are discussed.

In Section 6, the basic deep stacking networks and their several extensions are discussed in detail, which exemplify the discriminative, supervised deep learning networks in the three-way classification scheme. This group of deep networks operate in many ways that are distinct from the deep neural networks. Most notably, they use target labels in constructing *each* of many layers or modules in the overall deep networks. Assumptions made about part of the networks, such as linear output units in each of the modules, simplify the learning algorithms and enable a much wider variety of network architectures to be constructed and learned than the networks discussed in Sections 4 and 5.

In Sections 7–11, we select a set of typical and successful applications of deep learning in diverse areas of signal and information processing. In Section 7, we review the applications of deep learning to speech recognition, speech synthesis, and audio processing. Subsections surrounding the main subject of speech recognition are created based on several prominent themes on the topic in the literature.

In Section 8, we present recent results of applying deep learning to language modeling and natural language processing, where we highlight the key recent development in embedding symbolic entities such as words into low-dimensional, continuous-valued vectors.

Section 9 is devoted to selected applications of deep learning to information retrieval including web search.

In Section 10, we cover selected applications of deep learning to image object recognition in computer vision. The section is divided to two main classes of deep learning approaches: (1) unsupervised feature learning, and (2) supervised learning for end-to-end and joint feature learning and classification.

Selected applications to multi-modal processing and multi-task learning are reviewed in Section 11, divided into three categories according to the nature of the multi-modal data as inputs to the deep learning systems. For single-modality data of speech, text, or image, a number of recent multi-task learning studies based on deep learning methods are reviewed in the literature.

Finally, conclusions are given in Section 12 to summarize the monograph and to discuss future challenges and directions.

This short monograph contains the material expanded from two tutorials that the authors gave, one at APSIPA in October 2011 and the other at ICASSP in March 2012. Substantial updates have been made based on the literature up to January 2014 (including the materials presented at NIPS-2013 and at IEEE-ASRU-2013 both held in December of 2013), focusing on practical aspects in the fast development of deep learning research and technology during the interim years.