Li Deng

Artificial Intelligence in the Rising Wave of Deep Learning

The historical path and future outlook

Artificial intelligence (AI) is a branch of computer science and a technology aimed at developing the theories, methods, algorithms, and applications for simulating and extending human intelligence. Modern AI enables going from an old world—where people give computers rules to solve problems—to a new world—where people give computers problems directly and the machines learn how to solve them on their own using a set of algorithms. An algorithm is a self-contained sequence of instructions and actions to be performed by a computational machine. Starting from an initial state and initial input, the instructions describe computational steps, which, when executed, proceed through a finite number of well-defined successive states, eventually producing an output and terminating at a final ending state. AI algorithms are a rich set of algorithms used to perform AI tasks, notably those pertaining to perception and cognition that involve learning from data and experiences simulating human intelligence.

The ultimate goal of AI is to create technology that allows computational machines to function in a highly intelligent manner. This high-level goal can be broken down into a number of subareas pertaining to the human-like perceptual and cognitive capabilities that are expected to exhibit in an intelligent machine. Among the most important capabilities is that of learning, which enables automated machine algorithms to improve through experiences by themselves. The recent rise of a powerful machine-learning paradigm—deep learning—is responsible for much of the resurgence of AI over the past ten years or so (see [1], [2], and [65]). Other important cognitive capabilities include reasoning, problem solving, planning, motion (robotics) and decision making, interactions and dialogue, creativity, and knowledge representation. The latter involves how to build, represent, and exploit commonsense knowledge, which seems effortless in human behavior, and it also involves how symbolic and neural (subsymbolic) forms can be seamlessly integrated to achieve brain-like knowledge representation.

In addition to the aforementioned rich set of subareas, the wide scope of AI is also reflected in the many modalities of signal and information that humans use pervasively, whose capability is to be simulated by machines. Among the most important modalities is natural language or text. The AI processing of natural language gives machines the ability to read, comprehend, and generate the languages that humans use in our daily lives. A successful AI system with natural language capabilities would enable natural user-machine interfaces as well as the acquisition of knowledge directly from gigantic text sources created by humans in both public Internet and private enterprises. Other significant modalities include visual, speech/audio, touch/tactile, and smell/taste information, constituting another important subarea of AI called *machine perception*, which, at its core, encompasses speech recognition as well as computer vision. In developing AI methods for processing natural language and visual and speech information, AI researchers have invented many highly effective algorithms over the past three decades or so.

This article is intended for a general audience interested in the historical development, current status, popular algorithms, and future directions of AI as a technology. To limit the scope, I will survey and analyze three major rising waves of AI since the 1960s, focusing on their quite disparate paradigms, approaches, and related algorithms. In the exposition, rather than covering the extensive scope of AI outlined above, I will draw the representative and most relevant examples mainly from notable applications in machine perception. In particular, I will attempt to connect the cycles of the ups and downs of AI to their counterparts in speech recognition, which, as a core application area of AI and also as one major research discipline that our IEEE Signal Processing Society has made most significant contributions to, parallels the historical path of general AI more closely than other areas of AI. Finally, the outlook of AI will be discussed and analyzed, from both the horizontal (i.e., breakthroughs in general AI technology and algorithms)

(continued on page 173)

Digital Object Identifier 10.1109/MSP.2017.2762725
Date of publication: 9 January 2018

1053-5888/18©2018IEEE

(continued from page 180)

and the vertical (i.e., individual AI application areas) perspectives.

The first rise of AI

In the first rising wave of AI, starting in the 1960s and based on expert knowledge engineering, domain experts devised computer programs according to the knowledge about the (very narrow) application domains they have [3], [4]. The experts designed these programs using symbolic logical rules grounded on careful representation and engineering of such knowledge. These knowledge-based AI systems tend to be effective in solving narrow-domain problems by examining the "head" or most important parameters and reaching a solution about the appropriate action to take in each specific situation. These "head" parameters are identified in advance by human experts, leaving the "tail" parameters and cases untouched. Since they lack learning capability, they have difficulty in generalizing the solutions to new situations and domains.

The typical approach during this wave is exemplified by the expert system, a computer system that emulates the decision-making ability of a human expert. Such systems are designed to solve complex problems by reasoning about knowledge [3]. The first expert systems were created in the 1970s and then proliferated in the 1980s. The main "algorithm" used was the inference rules in the form of "if-then-else" [5].

The main strength of these first-generation AI systems is its transparency and interpretability in their (limited) capability in performing logical reasoning. They use handcrafted expert knowledge that is often effective in narrowly defined problems, although the reasoning cannot handle uncertainty that is ubiquitous in practical applications. Due to this strength, the first-generation AI systems are still in use today. Examples are narrow-domain dialogue systems and chatbots, chess-playing programs, traffic light controllers, optimization software for logistics of good deliveries, etc.

The early research and system design of speech recognition, a long-standing AI challenge in machine perception, were based on the AI paradigm of expert knowledge engineering during the first rising wave of AI. During the 1970s and early 1980s, the expert-system approach to speech recognition was quite popular, driven largely by Carnegie Mellon University and Massachusetts Institute of Technology speech researchers; e.g., the spectrogram reading method as elaborated in [6]. However, the lack of general abilities to learn algorithmically from data and to handle uncertainty in reasoning in the knowledge-based approach was acutely recognized by researchers, along with the second rise of AI in the 1980s. The author has been part of the speech research community since the late 1980s and contributed to the transition from knowledge-based speech recognition to a data-driven one powered by statistical machine-learning methods [7]–[13].

The second rise of AI

The second rising wave of AI for speech arrived in the 1980s (and somewhat later for other AI areas) after clear evidence that learning and perception capabilities are crucial for complex AI systems but missing in knowledge-based expert systems. This is not just for speech recognition but also for vision and other AI systems. For example, when the Defense Advanced Research Projects Agency opened its first Grand Challenge for autonomous driving, most vehicles then relied on the knowledge-based paradigm. Much like speech recognition, the autonomous driving and vision researchers quickly realized the limitation of the first-generation AI paradigm due to the need for automatic learning equipped with uncertainty handling and generalization capabilities.

This second-generation AI paradigm was based on machine learning, which we now call *shallow* due to the lack of abstractions constructed by many-layer or "deep" representations of data which would come in the third rise of AI. In such shallow machine learning, engineers do not need to be concerned with constructing precise and exact rules as required for the first-generation AI systems. Rather, they focus on statistical models [14], [15] or simple neural networks [16] as an underlying engine and then automatically learn or "tune" the parameters of the engine using the training data to make them handle uncertainty and generalize well from one condition to another and from one domain to another. The key algorithms and methods for machine learning include expectation-maximization (EM), Bayesian networks, support vector machines, decision trees, and, for neural networks, the backpropagation algorithm.

Generally, the machine-learning-based AI systems perform much better than the earlier, knowledge-based counterparts. Successful examples include almost all AI tasks in machine perception—speech recognition [10], [17]–[20], face recognition [21], visual object recognition [22], handwriting recognition [23], and machine translation [24].

In speech recognition, for more than 20 years from the 1980s to 2010, the paradigm was completely switched to and dominated by the (shallow) machine-learning paradigm using a statistical generative model called the *hidden Markov model* (*HMM*) integrated with Gaussian mixture models, along with various versions of its generalization [25]–[28]. The main algorithms and methods include the Viterbi algorithm, Baum–Welch algorithm (which is a special case of EM when applied to the HMM), and extended Baum–Welch (which includes learning algorithms for maximizing mutual information, minimizing classification errors, and minimizing phone errors) [29].

Among many versions of the generalized HMMs were statistical and neural-net-based hidden dynamic models (see [30]–[32] and [66]). The former adopted EM and extended Kalman filter algorithms for learning model parameters [33], [34] and the latter used backpropagation [35]. Both of them made extensive use of multiple latent layers of representations for the generative process of speech waveforms following the long-standing framework of analysis by synthesis in human speech perception. More significantly, inverting this deep generative process to its counterpart of an end-to-end discriminative process gave rise to the very first industrial success of deep learning [36]–[39], which is the foundation of the third rising wave of AI that will be elaborated on next.

The third (current) rise of AI

While the second generation of AI systems performed a lot better than the previous generation, they were far from human-level performance. With a few exceptions, the shallow machine-learning models often did not have the sufficiently large capacity to absorb the huge amounts of training data. Further, the learning algorithms, methods, and computing infrastructures were not powerful enough. All of this changed about one decade ago, bringing about the third rise of AI, propelled by the new paradigm of deep-structured machine learning or deep learning**.**

In traditional machine-learning approaches, features are designed by humans and feature engineering is a bottleneck requiring significant human expertise. Concurrently, the models lack the representation power and, hence, the ability to form levels of decomposable abstractions that automatically disentangle complex factors in shaping the observed data. Deep learning breaks away the aforementioned difficulties by the use of a deep, layered model structure, often in the form of neural networks, and the associated end-to-end learning algorithms. The advances in deep learning are one major driving force behind the current AI inflection point and the resurgence of neural networks.

Speech recognition is the first real-world AI application impacted strongly by deep learning. Industrial applications of deep learning to large-scale speech recognition started around 2010. In late 2009 and also 2010, I invited my academic collaborator Prof. Geoffrey Hinton, of the University of Toronto, and later his students to work with me and my colleagues at Microsoft Research in Redmond to apply deep learning to speech recognition. We co-organized the 2009 Neural Information Processing Systems Workshop on Deep Learning for Speech Recognition and Related Applications. The workshop was motivated by the limitations of deep generative models of speech, and the possibility that the big-compute, big-data era warrants a serious exploration of deep neural nets (DNNs) [40]. It was believed then that pretraining DNNs using generative models of deep belief nets based on the contrastive-divergence learning algorithm would overcome the main difficulties of neural nets encountered in the 1990s. However, early into this research at Microsoft, it was discovered that without contrastive-divergence pretraining, but with the use of large amounts of training data together with the DNNs designed with corresponding large, context-dependent output layers, dramatically lower recognition errors were possible than then state-of-the-art (shallow) machine-learning systems based on the second-generation AI paradigm and algorithms. This finding was quickly verified by several other major speech recognition research groups. Further, the nature of recognition errors produced by the two types of systems were found to be characteristically different, offering technical insights into how to integrate deep learning into the existing highly efficient, run-time speech decoding system deployed by major players in the speech recognition industry today [38], [41]–[43]. More recent advances in deep learning for speech recognition can be found in [44]–[46]. The backpropagation algorithm is uniformly used in all sorts of deep neural networks in all current speech recognition systems.

Large-scale speech recognition is the first and most convincing successful case of deep learning in recent history, embraced by both industry and academia across the board. Since 2010, the two major conferences on signal processing and speech recognition—the IEEE International Conference on Acoustics, Speech, and Signal Processing, and Interspeech—have seen a huge increase year by year in the number of accepted papers in their respective annual conferences on the topic of deep learning for speech recognition. More importantly, all major commercial speech recognition systems (e.g., Microsoft Cortana, Xbox, Skype Translator, Amazon Alexa, Google Now, Apple Siri, Baidu, and iFlyTek voice search, and a range of Nuance speech products, etc.) are all based on deep-learning methods. The most cited article per year in speech recognition's history was published in 2012 in *IEEE Signal Processing Magazine* [38].

Quickly following the striking success of speech recognition in 2010 heralding the arrival of the third AI wave, two other important AI application areas—computer vision [47], [48] and machine translation [49]—were completely taken over by the similar deep-learning paradigm. In addition, a large number of other real-world applications have been made successful due to deep learning, including

- image captioning [50]–[53]
- visual question answering [54]
- web search [55]
- natural language processing [67]
- drug discovery and toxicology
- customer relationship management
- recommendation systems
- medical informatics
- Internet advertisements
- medical image analysis
- robotics
- self-driving vehicles
- board games (e.g. AlphaGo, Poker, and DOTA2), etc.

Setting aside their huge empirical successes, models of neural-network-based deep learning are often simpler and easier to design than the traditional machine-learning models developed in second-generation AI. In many applications, deep learning is performed simultaneously for all parts of the model, from feature extraction all the way to prediction, in an end-to-end manner. Another factor contributing to the simplicity of neural network models is that the same model building blocks (i.e., the different types of layers) are generally used in many different applications. Using the same building blocks for a large variety of tasks in a uniformed manner makes the adaptation of models reused for one task or data to another task or data relatively easy. In addition, software toolkits have been developed to allow faster and more efficient implementation of these models. For these reasons, deep neural networks are today a prominent method of choice for a wide variety of machine-learning and AI tasks over large data sets.

An outlook for upcoming AI technology

Despite the spectacular successes of deep learning during the third rising wave of AI, there remain huge challenges. The current deep-learning methods lack interpretability, in contrast to the knowledge-based AI paradigm established during the first rise. In a number of applications, deep-learning methods prove to give the recognition accuracy close to or exceeding humans, but they require considerably more training data, power consumption, and computing resources than humans. Also, the accuracy results are statistically impressive but often unreliable on the individual basis. Further, most of the current deep-learning models have no reasoning and explaining capabilities, making them vulnerable to disastrous failures or attacks without the ability to foresee and thus prevent them.

To overcome the aforementioned challenges and to achieve ultimate successes of general AI, both fundamental and applied research are needed. The next new wave of AI will not come until new paradigmatic, algorithmic, and hardware breakthroughs are brought about. In this section, I will discuss future directions of AI from this "horizontal" perspective in terms of the advancement of general AI methodology. The next section will be devoted individual vertical AI application areas.

One potential breakthrough in AI research is in developing interpretable deep-learning models that can be learned and applied with no black box in it. The success in this endeavor will create new AI algorithms and methods that can overcome the current limitation of AI systems in their lack of the ability to explain the actions, decision, and prediction outcomes to human users while promising to perceive, learn, decide, and act on their own. The new type of AI systems will desirably allow the users to understand and thus trust the AI system's outputs and to foresee and predict future behaviors of the systems. To this end, neural networks and symbolic systems need to be usefully integrated, enabling the AI systems themselves to construct models that will explain how the world works. That is, they will discover by themselves the underlying causes or logical rules that shape their prediction and decision-making processes understandable to human users in symbolic and natural language forms. Initial work in this direction makes use of an integrated neural-symbolic representation called *tensor-product neural memory cells*, which can be decoded back to symbolic form without loss of information after extensive learning in the neural-tensor domain [56], [57], [68]. Active research is ongoing in this direction. When successful, the AI system built on such tensor-product representations will learn to read and understand massive natural language documents. Then it will be able not only to answer questions sensibly but also to truly understand what it reads to the extent that it can convey such understanding to human users in providing what steps it takes to reach the answer. These steps may be in the form of logical reasoning expressed in natural language that is easily understood by the human users of this type of machine reading comprehension AI systems.

Another potential breakthrough in AI research is in new algorithms for reinforcement and unsupervised deep learning that make use weak or even no teaching signals paired to inputs to guide the learning. Effective reinforcement-learning algorithms would allow the AI systems to learn via interactions with possibly adversarial environments and with themselves. Hierarchical rewards can be beneficially modeled in reinforcement learning. The most challenging problem, however, is unsupervised learning, for which no satisfactory learning algorithms have been devised so far in practical applications. The development of unsupervised learning algorithms is significantly behind that of supervised and reinforcement deep learning where backpropagation and Q-learning algorithms have been reasonably mature. The most recent development in unsupervised learning takes the approach of exploiting sequential output structure and advanced optimization methods to alleviate the need for using labels in training prediction systems [58], [59].

Future advances in unsupervised learning are promising by exploiting new sources of learning signals including the structure of input data and the mapping relationships from input to output and vice versa. Exploiting the relationship from output to input is closely connected to building conditional generative models. To this end, the recent popular topic in deep learning—generative adversarial networks [60]—is a highly promising direction where the long-standing concept of analysis-by-synthesis in pattern recognition and machine learning (especial for speech recognition) is likely to return to the spotlight in the near future.

A closely related topic and research direction is multimodal deep learning with cross-domain information as low-cost supervision. Standard speech recognition, image recognition, and text-classification methods make use of supervision labels within each of the speech, image, and text modalities separately. This is far from how children learn to recognize speech and images and classify text. For example, children often get the distant "supervision" signal for speech sounds by an adult pointing to an image scene or text or handwriting that is associated with the speech sounds. Similarly, for children to learn image categories, speech sounds or text can be used as the supervision signal. This has motivated a computational model [61], which is aimed to effectively leverage multimodal data to improve engineering systems for multimodal processing using a similarity measure defined in the same semantic space that speech, image, and text are all mapped into via DNNs trained using maximum mutual information across different modalities.

A further future direction for fruitful AI research is the paradigm of learning-to-learn or metalearning, that is, how to design an AI system that improves or automatically discovers a learning algorithm, such as a complex optimization algorithm. The study of this paradigm started in 2001 [62], but it was not until around 2015 when the deep-learning methodology became reasonably matured that stronger evidence of the potential impact of learning-to-learning has become apparent. If successful, the development of algorithms for solving most computer science problems and even the programming itself may be reformulated as a deep learning problem and be solved by a uniform infrastructure designed for deep learning today. Learning-to-learn is a powerful emerging AI paradigm and is a fertile research direction expected to impact real-world AI applications which I will discuss next.

Vertical applications

The third rise of AI over the past several years has been creating not only technology breakthroughs, but it also promises to create business model disruptions in a wide range of vertical applications. The current and future application areas that may be transformed by AI include financial services, health care, transportation (e.g., self-driving vehicles), agriculture, retail, energy, logistics, paralegal, education, human resource management, sales and marketing (e.g., customer relationship management), customer service automation, and more. Here a few of them are selected to elaborate.

Learning-to-learn is a powerful emerging AI paradigm and is a fertile research direction expected to impact real-world AI applications.

In financial services, banks have long been using AI systems to detect claims of anomaly and flag them for human investigation. AI technology can also help reduce fraud and crime by monitoring behavioral patterns of users for unusual changes. Banks also use AI systems to organize operations, maintain bookkeeping, invest in stocks, and manage properties, reacting to changes overnight or when business is not taking place. With limited reasoning capabilities, AI technology can further scan through millions of e-mails, instant messages, and texts a day on Wall Street to map out ordinary behavioral patterns among the employees of its banks. For example, AI systems can screen whom traders ordinarily communicate with and what kinds of information they typically send. When behavior is detected to be abnormal or falls into a recognizable pattern of wrongdoing, AI systems can alert the compliance staff of its clients to trigger an investigation.

In health care, AI systems can assist doctors. For example, Microsoft has been reported to develop AI methods to help doctors find the right treatments for cancer. There is a great amount of research and drugs developed relating to cancer. The very large number of medicines and vaccines to treat cancer makes it difficult for doctors to choose the right drugs for the patients. The goal may be to distill information from a large quantity of medical papers relevant to cancer treatment and help predict which combinations of drugs will be most effective for each patient. Besides drug selection and discovery, AI also can help improve hospital efficiency.

Over the past two years or so, one particular AI technology, called *bots*, has made significant advances. Bots promise to become the future of the user interface, where the user experience evolves from click based to conversational (text or voice). Interactions likewise shift from app oriented to messaging oriented [63]. After bots technology is widely adopted, users no longer need to open different apps to book travel, shop for clothes, and engage customer services. Rather, a user could engage directly in conversation with the bots via a messenger. Bots technology promises to facilitate businesses in e-commerce, customer support, and employee workflows and productivity.

A special type of AI bot is the digital personal assistant, which is personalized and equipped with the abilities to complete or automate simple tasks based on voice commands or text inputs. It is also equipped with forecasting and inference capabilities of recommendation engines. Popular examples are Microsoft's Cortana, Apple's Siri, Amazon's Alexa, and Google Assistant. The AI technology underlying these systems is expected to improve at a fast pace in the future.

In summary, this article exposes and analyzes the rise of technologies that are enabling AI and deep learning to become a consistent way of learning from and exploiting a wide range of signals and information—from speech and image to bots and other applications. An outlook of future AI technology development is provided from a personal perspective based on the past and ongoing research experience.

Acknowledgment

This manuscript was submitted on 23 April 2017 when the author was previously with Microsoft Research, Redmond, Washington. The author thanks associate editor, Prof. Rodrigo Guido, who kindly invited and assisted him in writing this survey article.

Author

Li Deng (l.deng@ieee.org) received his Ph.D. degree from the University of Wisconsin–Madison. He was a tenured professor from 1989 to 1999 at the University of Waterloo, Ontario, Canada, and then joined Microsoft Research, Redmond, Washington, where he was the chief scientist of artificial intelligence (AI) and partner research manager. He recently joined Citadel as its chief AI officer. He is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association. He was the editor-in-chief of *IEEE Signal Processing Magazine* (2009–2011) and *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2012–2014), for which he received the IEEE Signal Processing Society (SPS) Meritorious Service Award. He received the 2015 IEEE SPS Technical Achievement Award for "outstanding contributions to automatic speech recognition and deep learning" and numerous best paper and scientific awards for the contributions to AI, machine learning, multimedia signal processing, speech and human language technology, and their industrial applications.

References

[1] L. Deng and D. Yu, *Deep Learning: Methods and Applications*. Boston: NOW Publishers, 2014.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[3] N. Nilsson, *Principles of Artificial Intelligence*. New York: Springer, 1982.

[4] P. Winston, *Artificial Intelligence*. New York: Springer, 1993.

[5] P. Jackson, *Introduction to Expert Systems*. Reading, MA: Addison-Wesley, 1998.

[6] V. Zue, "The use of speech knowledge in automatic speech recognition," *Proc. IEEE*, vol. 73, no. 11, Nov. 1985.

[7] L. Deng, M. Lennig, V. Gupta, and P. Mermelstein, "Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 1988.

[8] L. Deng, M. Lennig, and P. Mermelstein, "Use of vowel duration information in a large vocabulary word

recognizer," *J. Acoustic. Soc. Amer.*, vol. 86, no. 2, pp. 540–548, 1989.

[9] L. Deng, P. Kenny, M. Lennig, V. Gupta, and P. Mermelstein, "A locus model of coarticulation in an HMM speech recognizer," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 1989, pp. 97–100.

[10] L. Deng, M. Lennig, V. Gupta, and P. Mermelstein, "Modeling microsegments of stop consonants in a hidden Markov model based word recognizer," *J. Acoust. Soc. Amer.*, vol. 87, no. 6, pp. 2738–2747, 1990.

[11] L. Deng, P. Kenny, M. Lennig, and P. Mermelstein, "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 2, pp. 265–271, 1990.

[12] L. Deng, V. Gupta, M. Lennig, P. Kenny, and P. Mermelstein, "Acoustic recognition component of an 86000-word speech recognizer," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 1990, pp. 741–744.

[13] L. Deng and K. Erler, "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units," *J. Acoust. Soc. Amer.*, vol. 92, pp. 3058–3067, 1992.

[14] C. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer, 2006.

[15] K. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[16] C. Bishop, *Neural Networks for Pattern Recognition.* Oxford, U.K.: Oxford Univ. Press, 1995.

[17] F. Jelinek, *Statistical Models for Speech Recognition*. Cambridge, MA: MIT Press, 1998.

[18] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[19] R. Chengalvarayan and L. Deng, "HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 243–256, May 1997.

[20] L. Deng, D. Yu, and A. Acero, "Structured speech modeling," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 5, pp. 1492–1504, 2006.

[21] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.

[22] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2005, pp. 524–531.

[23] R. Plamondon and S. Srihari, "Online and off-line handwriting recognition: A comprehensive survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 63–84, Jan. 2000.

[24] F. Och, "Minimum error rate training in statistical machine translation," in *Proc. Assoc. Computational Linguistics Conf.*, 2003, pp. 160–167.

[25] J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 75–80, 2009.

[26] J. M. Baker, L. Deng, S. Khudanpur, C.-H. Lee, J. R. Glass, N. Morgan, and D. O'Shaughnessy, "Updated MINDS report on speech recognition and understanding," *IEEE Signal Process. Mag.*, vol. 26, no. 4, pp. 78–85, July 2009.

[27] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing.* Englewood Cliffs, NJ: Prentice Hall, 2000.

[28] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach.* New York: Marcel Dekker, 2003.

[29] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition: A unifying review for optimization-oriented speech recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, 2008.

[30] J. Bridle, L. Deng, J. Picone, H. B. Richards, J. Ma, T. Kamm, M. Shuster, S. Pike, and R. Roland, "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," Final Rep. for 1998 Workshop on Language Engineering, CLSP, Johns Hopkins, Baltimore, MD, 1998.

[31] L. Deng, *Dynamic Speech Models: Theory, Algorithm, and Application*. San Rafael, CA: Morgan & Claypool, 2006.

[32] L. Deng and D. Yu, "Use of differential cepstra as acoustic features in hidden trajectory modeling for phonetic recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 2007.

[33] L. Deng, "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*. New York: Springer-Verlag, 2003.

[34] J. Ma and L. Deng, "Target-directed mixture dynamic models for spontaneous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 47–58, 2004.

[35] J. Picone, P. S. Pike, R. Regan, T. Kamm, J. bridle, L. Deng, Z. Ma, H. Richards, and M. Schuster, "Initial evaluation of hidden dynamic models on conversational speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1999.

[36] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, "Binary coding of speech spectrograms using a deep autoencoder," in *Proc. IEEE Int. Conf. Speech Communication Assoc.*, 2010.

[37] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 2013.

[38] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, pp. 82–97, Nov. 2012.

[39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[40] A. Mohamed, G. Dahl, and G. Hinton. "Acoustic modeling using deep belief networks," in *Proc. Conf. Neural Information Processing Systems Workshop on Speech Recognition*, 2009.

[41] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach.* New York: Springer, 2015.

[42] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 20, pp. 30–42, Jan. 2012.

[43] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 22, pp. 1533–1545, July 2014.

[44] D. Amodei, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 173–182.

[45] W. Xiong, et al., "Achieving human parity in conversational speech recognition," in *Proc. IEEE Int. Conf. Speech Communication Assoc.*, 2016.

[46] G. Saon, et al., "English conversational telephone speech recognition by humans and machines," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 2017.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Conf. Neural Information Processing Systems*, 2012.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016.

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Machine Learning*, 2015.

[50] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2015.

[51] J. Devlin, et al., "Language models for image captioning: The quirks and what works," in *Proc. Assoc. Computational Linguistics Conf.*, 2015.

[52] H. Fang, et al., "From captions to visual concepts and back," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2015.

[53] Z. Gan, et al., "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2017.

[54] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Computer Vision Pattern Recognition,* 2016.

[55] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. Conf. Information and Knowledge Management*, 2013.

[56] P. Smolensky, M. Lee, X. He, W.-t. Yih, J. Gao, and L. Deng, "Basic reasoning with tensor product representations," arXiv preprint, Jan. 2016.

[57] M. Lee, X. He, W.-t. Yih, J. Gao, L. Deng, and P. Smolensky, "Reasoning in vector space: An exploratory study of question answering," in *Proc. Int. Conf. Learning Representations*, 2016.

[58] R. Stewart and S. Ermon, "Label-free supervision of neural networks with physics and domain knowledge," in *Proc. Assoc. Advancement Artificial Intelligence Conf.*, 2017.

[59] Y. Liu, J. Chen, and L. Deng, "Unsupervised sequence classification using sequential output statistics," in *Proc. Conf. Neural Information Processing Systems*, 2017.

[60] I. Goodfellow, et al., "Generative adversarial nets," in *Proc. Conf. Neural Information Processing Systems*, 2014.

[61] L. Deng. "Cross-modality distant supervised learning for speech, text, and image classification," in *Proc. Conf. Neural Information Processing Systems Workshop Multimodal Machine Learning*, 2015.

[62] S. Hochreiter, A. S. Younger, and P. R. Conwell, "Learning to learn using gradient descent," in *Proc. Int. Conf. Artificial Neural Networks*, 2001, pp. 87–94.

[63] L. Deng, *How Deep Reinforcement Learning Can Help Chatbot*, Venturebeat, Aug. 2016. [Online]. Available: https://venturebeat.com/2016/08/01/how-deep-reinforcement-learning-can-help-chatbots/

[64] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. London: Pearson, 2009.

[65] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets, *Neural Computation*, vol. 18, 2006, pp. 1527–1554.

[66] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Communication*, vol. 24, no. 4, pp. 299–323, 1998.

[67] L. Deng and Y. Liu, Eds, *Deep Learning in Natural Language Processing*. Beijing: Springer, 2018.

[68] H. Palangi, P. Smolensky, X. He, and L. Deng. "Question-answering with grammatically-interpretable representations," in *Proc. AAAI Conf. Artificial Intelligence*, 2018, to be published.