Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Challenges and opportunities: from big data to knowledge in AI 2.0

Yue-ting ZHUANG, Fei WU‡, Chun CHEN, Yun-he PAN

(*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

E-mail: yzhuang@zju.edu.cn; wufei@zju.edu.cn; chenc@zju.edu.cn; panyh@cae.cn

Received Dec. 31, 2016; Revision accepted Jan. 9, 2017; Crosschecked Jan. 11, 2017

Abstract: In this paper, we review recent emerging theoretical and technological advances of artificial intelligence (AI) in the big data settings. We conclude that integrating data-driven machine learning with human knowledge (common priors or implicit intuitions) can effectively lead to explainable, robust, and general AI, as follows: from shallow computation to deep neural reasoning; from merely data-driven model to data-driven with structured logic rules models; from task-oriented (domain-specific) intelligence (adherence to explicit instructions) to artificial general intelligence in a general context (the capability to learn from experience). Motivated by such endeavors, the next generation of AI, namely AI 2.0, is positioned to reinvent computing itself, to transform big data into structured knowledge, and to enable better decision-making for our society.

Key words: Deep reasoning; Knowledge base population; Artificial general intelligence; Big data; Cross media
http://dx.doi.org/10.1631/FITEE.1601883 **CLC number:** TP391.4

1 Introduction

The idea of inanimate objects coming to life as intelligent beings has been around for a long time. Modern artificial intelligence (AI) was formally founded in 1956 at a workshop at Dartmouth College, where the term 'artificial intelligence' was coined (McCarthy *et al.*, 2006). After advancing over the last few decades, AI has become one of the most profound undertakings in science, and one that will affect every aspect of human life.

We have witnessed the remarkable successes brought about by AI, such as machine translation, speech recognition, image classification, and information retrieval. However, there are clear differences between human-intelligence and machine-intelligence. For example, although people and computers can both play chess, it is far from clear whether they do it the same way. To be successful in realistic environments, existing AIs should identify and implement effective actions, given the fact of inescapable incompleteness in their knowledge about the world. That is, the next-generation AI with regard to big data is an explainable, robust, and general AI: it can perform deep neural reasoning, instead of brute-force shallow computation (e.g., search); it is capable of harnessing data-driven models with structured logical rules; it can learn from experience (Fig. 1).

We review the recent advances of AI in terms of AI platforms, natural language processing (NLP), multimedia, computer vision, knowledge base population and visualization, as well as the challenges and opportunities of future AI.

2 Artificial intelligence platforms

The development of the entire AI field may be considered in three stages as follows: (1) From the 1940s to the 1970s, researchers focused on studying traditional AI problems such as reasoning with

‡ Corresponding author
ORCID: Fei WU, http://orcid.org/0000-0003-2139-8807

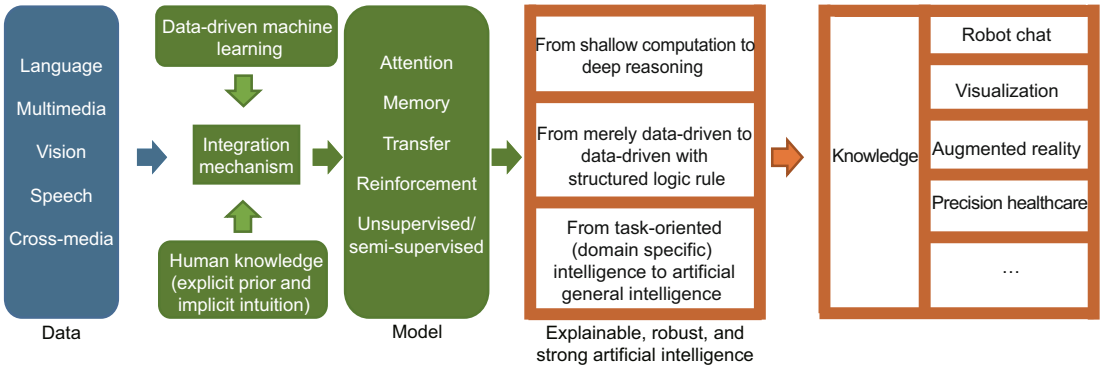

Fig. 1 Flowchart: from data to knowledge

methodologies, which were based mainly on logic and heuristic algorithms. (2) After the 1970s, AI problems were split into many other new fields, such as natural language processing, multimedia and computer vision, and statistical machine learning, which turned out to be workable on these problems after the 1990s. (3) In recent years, the third stage, it is looking like the key to the AI world is deep learning. At different stages, engineers and scientists try to build different AI frameworks and apply them to specific problems. In this paper, we review three major types of AI framework corresponding to the three different stages of AI field, and how these each contribute to and benefit the design of a new framework for the AI 2.0 age.

2.1 Traditional artificial intelligence frameworks

During the first stage, AI researchers worked on traditional AI problems like game playing, knowledge representation and reasoning, and expert systems. The frameworks developed are task-specific, in other words, suitable only for one specific problem. The methodologies are mostly based on heuristic algorithms. For instance, DeepBlue, the first chess-playing system to beat a world chess champion, uses heuristic search (e.g., alpha-beta pruning) to find an optimized play during the game. However, DeepBlue could hardly handle large-scale data (e.g., the game of Go) due to the high complexity of its heuristic algorithm and its inability to be applied to any scenarios other than chess.

2.2 Statistical machine learning frameworks

Traditional AI frameworks suffer from the requirement of designing precise rules, which cost human resources. After the 1970s, researchers paid more attention to extracting the 'rule' automatically from observed data, using machine learning technologies like Perceptron, the first successful model designed in 1969. From then on, AI research has made machine learning its focus. Many statistical learning algorithms have been designed and published, such as support vector machine, Bayesian networks, and conditional random fields. However, there is no single model that works best on all given problems. Determining a suitable model for a given problem is still more an art than a science. Researchers therefore tried to uniform these models into a framework. For instance, GraphLab (Low *et al.*, 2014), an open source project started by Carlos Guestrin of Carnegie Mellon University in 2009, provides great features including multicore and distributed application programming interface (API). More specifically, GraphLab contains topic models, graph analytical algorithms, graphical models, clustering algorithms, collaborative filtering algorithms, etc. Frameworks like GraphLab make it easier to apply statistical machine learning models to specific AI problems. However, it requires a large amount of data to drive statistical machine learning algorithms. Besides, it turns out that supervised algorithms perform better than unsupervised ones at many tasks. How to learn from a small set of observed data without label information, just like the way human beings learn things, is still an open direction for the entire AI field.

2.3 Deep learning frameworks

Statistical machine learning methodologies require precise domain knowledge from scientists and engineers to design the features. This limits the

application of statistical machine learning in large-scale data. For instance, imagine that we were expected to design a feature engineering based version of AlphaGo, what kind of features should we propose to extract from the game of Go? One of the best experts to give guidance on feature extraction is Lee Se-dol, who might use some of his gaming experience and feed it into a machine learning model. However, by this method, AlphaGo could hardly beat Lee Se-dol as his 'student', since human-designed features are in most cases considering the large-scale data limitedly. Instead, deep learning has the ability to acquire feature hierarchies from data automatically. Then, these features are encoded within multiple non-linear neural networks.

In recent years, several deep learning frameworks have been released. Table 1 illustrates the comparisons of the major aspects of several mainstream frameworks. For example, Google developed TensorFlow (Abadi *et al.*, 2016), a deep learning framework that uses data flow graphs, where nodes represent a computation and edges indicate the flow of information from one node to another. TensorFlow also has the advantage of being able to do partial subgraph computation, which facilitates the distributed training.

Other state-of-the-art deep learning frameworks include: Caffe (Jia *et al.*, 2014), which was developed at the Berkeley Vision and Learning Center and is used widely by the Facebook deep learning research team led by Yann LeCun; Torch (Collobert *et al.*, 2002), which was originally developed at NYU and then featured a large number of community-contributed packages; Theano (Bergstra *et al.*, 2010), which was developed at the University of Montreal, to perform symbolic differentiation or integration on complicated non-linear functions, and has been competitive on execution speed with Torch; and Neon, which was open-sourced by Nervana Systems, and has recently been ranked as the fastest framework across several performance categories.

The aforementioned frameworks offer the solution to the acquirement of software and hardware to apply neural networks to specific problems across a wide range of disciplines. However, the major limitation of deep learning still remains unsolved: the results of the computations are unexplained. More specifically, the features acquired by deep learning are just continuous values with less semantic information. We have no idea about the reason why AlphaGo rated a specific move as a good play. Thus, human players could hardly learn from AlphaGo to improve their skills.

To conclude this section, we review three types of AI frameworks with their major advantages and disadvantages. We see that none of them could achieve the fundamental goal of AI, which is the claim that human intelligence "can be so precisely described that a machine can be made to simulate it". Thus, in the age of AI 2.0, we are expecting a new framework, which is explainable at the human logical level, computationally powerful to handle large-scale data, and also workable on a small set of labeling data.

3 Artificial intelligence for big data

3.1 Natural language processing

NLP is a field of computer science, AI, and computational linguistics, concerning the interactions between computers and human natural languages. It aims to discover effective theories and methods that enable humans to communicate better with computers using natural language. The 'natural language' in this research field is the language that we commonly use in our daily life. Thus, NLP is closely related to the study of linguistics, but there are some important differences between them. NLP is not a general study of natural language. NLP focuses on developing a computer system which can actualize natural language communication.

3.1.1 Traditional natural language processing

Harris (1954) proposed the bag-of-words (BOW) model to represent human language. The BOW model is a simplified representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as a bag (multi-set) of its words, ignoring grammar and even the syntactic order but keeping multiplicity. In practice, the BOW model is used mainly as a tool of feature generation. After transforming the text into a 'bag of words', we can calculate various measures to characterize the text. The BOW model now has many successful applications, e.g., document classification, feature extraction, and spam filtering.

Table 1 Comparison of several deep learning frameworks

Framework	Language	Multi-GPU	Speed*	Applicability
TensorFlow	Python and C++	Yes	⋆	General
Torch	Lua	Yes	⋆⋆	General
Caffe	C++	Yes	⋆	Image
Theano	Python	No	⋆⋆	General
Neon	Python	Yes	⋆⋆⋆	General

*Speed is based on benchmarks published at ConvNet-benchmarks on GitHub. GPU: graphics processing unit

Part-of-speech tagging is one of the achievements in the traditional NLP field, used by a large number of language processing systems for preprocessing. The tagger assigns a (unique or ambiguous) part-of-speech tag to each token in the input, and passes its output to the next processing level, i.e., usually a parser. Furthermore, there is a large interest in part-of-speech tagging for corpus annotation projects, which create valuable linguistic resources by a combination of automatic processing and human correction. One promising solution for this task uses statistical approaches, e.g., the maximum entropy framework and Markov models. Stochastic taggers have obtained a high degree of accuracy without performing any syntactic analysis on the input. Later, Brill (1992) presented a simple rule-based part-of-speech tagger, which has many advantages over stochastic taggers as follows: a vast reduction in the stored information required, the perspicuity of a small set of meaningful rules as opposed to the large tables of statistics needed for stochastic taggers, ease of finding and implementing improvements to the tagger, and better portability from one tag set or corpus genre to another.

3.1.2 Natural language processing with deep learning

The earlier approaches in NLP research were based mainly on hand-crafted rules or features, but ignored a lot of information during data processing. The rise of deep learning makes it possible to extract more information from a large scale of raw data.

One outstanding result lies in the latest proposed model for representing words in corpus, namely Word vector (Mikolov *et al.*, 2013). Word2vec is a group of related models used to produce word embedding. These models are shallow two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes a large corpus of text as its input, and produces a high-dimensional space (typically of several hundred dimensions), with each unique word in the corpus corresponding to one vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. Word2vec can use either of the two model architectures, continuous bag-of-words (CBOW) or continuous skip-gram, to produce a distributed representation of words. By CBOW, the model predicts the current word by using a window of surrounding context words. The order of the context words does not affect the prediction (bag-of-words assumption). By the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weights nearby context words more heavily than more distant context words. According to the authors' note, CBOW is faster while skip-gram is slower, but the latter does a better job for infrequent words.

Based on the power of Word vector, convolutional neural networks (CNNs) are leveraged in NLP like semantic parsing, information retrieval, sentence modeling, and other traditional NLP tasks. These methods simply use the CNN models that were originally invented for computer vision. However, several disadvantages emerged when directly applying CNN to NLP tasks such as sentence modeling, as the boundary information of a sentence is more important than that of an image. To overcome this defect, Kalchbrenner *et al.* (2014) introduced the 'wide convolution' method into CNN. Moreover, the model using k-max pooling over a linear sequence of values returns the subsequence of k maximum values in the sequence, instead of the single maximum value. Secondly, the pooling parameter k can be dynamically chosen by making k a function of other aspects of the network or the input.

Another successful application of deep learning in NLP is machine translation. However, CNN cannot be used to map sequences to sequences, although it works well whenever large labeled training sets are available. Sutskever *et al.* (2014) used a general end-to-end approach to sequence learning, which makes minimal assumptions on the sequence structure. This approach uses a multi-layered long short-term memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another hierarchical LSTM, to decode the target sequence from the vector. The model showed its effectiveness even on long sentence translation, which indicated that LSTM does not have any difficulty handling long sentences. Meanwhile, LSTM learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. The obtained experimental results are close to the state-of-the-art performance of the conventional phrase-based machine translation system on an English-to-French translation task.

STM also shows certain promise in dialogue generation, but this method tends to be short-sighted since it predicts utterances one at a time, while ignoring the influence on future outcomes. Modeling the future direction of a dialogue is crucial to the generation of coherent and interesting dialogues. Reinforcement learning is therefore employed in dialogue generation. Li *et al.* (2016) integrated reinforcement learning with deep learning to model future reward in chatbot dialogue. The model simulates dialogues between two virtual agents, using policy gradient methods to reward sequences that display three useful conversational properties: informativity, coherence, and ease of answering (related to the forward-looking function).

3.1.3 Attention and memory in natural language processing

Recently, the rise of attention mechanisms in computer vision (CV) has inspired some work in NLP. Neural machine translation is a typical example, as the performance is greatly boosted when taking the attention mechanism into consideration. The models previously proposed for neural machine translation often belong to a family of encoder-decoders; i.e., they encode a source sentence into a fixed-length vector, from which a decoder generates a translation. Bahdanau *et al.* (2014) conjectured that the use of a fixed-length vector was a bottleneck preventing improvement of the performance of this basic encoder-decoder architecture. They proposed to circumvent this by allowing a model to automatically search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, they achieved a translation performance comparable to those of the state-of-the-art methods.

The aforementioned applications in NLP have already been the benchmarks in the corresponding research field of NLP. However, one drawback of the current methods is that the model lacks reasoning ability. To overcome this shortcoming, Weston *et al.* (2014) described a new class of learning models called memory networks. Memory networks reason with inference components combined with a long-term memory component, in which they learn how to use these mechanisms jointly. The long-term memory can be read and written to, with the goal of using it for prediction. The authors investigated their approach in the context of question answering (QA), where the long-term memory effectively acts as a (dynamic) knowledge base, and the output is a textual response, demonstrating the reasoning power of such models by chaining multiple supporting sentences to answer questions that require understanding the intention of the verbs.

3.2 Multimedia

Conventional multimedia computing is built on top of hand-crafted features, e.g., scale-invariant feature transform (SIFT), bag of visual words, bag of latent topics, and sparse representation (Wu *et al.*, 2014; 2015). However, hand-crafted features are in general restrictive in capturing complex semantics in multimedia (image, audio, speech, and text). In recent years, we have witnessed the unprecedented advance in deep learning in various multimedia applications, such as image and video classification, content-based multimedia retrieval, and cross-media analysis (Cho *et al.*, 2015).

Unlike the conventional machine learning methods that often use 'shallow' architectures, deep learning mimics the human brain, which is organized in a deep architecture and processes information through multiple stages of transformation and

representation. By exploring deep architectures to employ the feature learning at multiple levels of abstractions from raw data, deep learning approaches can learn complex nonlinear functions via multiple layers, and transform the input data (whether images, speech, or text) into some output. This is useful for making decisions (e.g., whether an attentional object in a camera image, or inferring context from text), without relying on human-crafted features using domain-specific knowledge.

In the multimedia field, the heterogeneity-gap among multimodal data is a fundamental barrier. Nowadays, more and more applications tend to involve multimodal data, where information inherently consists of data with different modalities, such as a Web image with loosely related narrative text descriptions, or a news article with paired texts and images. Data in different modalities can be used to discover the underlying latent correlation between data objects in single or multiple modalities. Much work has been done to map the data from one modality to another. In general, there are mainly three kinds of approaches to boost multimodal embedding, i.e., cross-media embedding for cross-media retrieval, such as Zhuang *et al.* (2016). The first one is canonical correlation analysis (CCA) and its variants, which map the multimodal data into a common space, such that the distance between two similar samples is minimized. The second kind of approach is latent Dirichlet allocation (LDA) and its extensions. The LDA-based approaches attempt to model correlations among multimodal data in terms of a latent semantic level. Motivated by the recent remarkable advance of deep learning, several deep architectures of the third kind of approach have been claimed to learn the joint multimodal representation. For example, Karpathy *et al.* (2014) broke down images into objects and sentences into fragments, and then evaluated their alignments in a latent common space.

Deep neural networks take a feed-forward hierarchical approach to learn features, and do classification simultaneously via a backpropagation error strategy. From a cognitive science perspective, human knowledge can be used to enhance or suppress the middle relevant or irrelevant neurons since the features learned at hidden layers are not always transparent in their delivering cues.

It is well known that semantics in multimedia is governed by concepts via common sense knowledge (e.g., bird is a hypernym of both penguin and canary). Although deep learning has achieved remarkable progress in its renaissance, learning models merely trained by an amount of data without recognizing the inherent knowledge (e.g., the latent structure, the uncertainty, and implicit priors) in multimedia may not be the most efficient strategy (Neal, 2012). Given the existence and availability of rich knowledge stored in different forms, such as WordNet and ImageNet, or even implicit knowledge extracted from click-through logs from search engines and social media (e.g., the individual's preference), we believe that deep learning frameworks can benefit substantially from leveraging such knowledge to further advance the state-of-the-art of various multimedia computing tasks. For example, deeply-supervised nets (DSNs) (Lee *et al.*, 2015) introduced direct supervision to the hidden layers, rather than the standard approach of providing supervision at only the output layer, and propagating this supervision back to earlier layers.

As a result, how to appropriately integrate the power of data-driven deep learning and the subtlety of knowledge-guided frameworks together is one attractive direction in the future multimedia computing (Pan, 2016).

3.3 Computer vision

Computer vision aims at simulating human perception capability with the power of computational modeling of the visual domain. Therefore, it theoretically lies in an extremely interdisciplinary field that effectively involves the comprehensive factors derived from multiple research areas such as image processing and statistical learning. In principle, the core of computer vision is how to set up an effective visual computing pipeline that is able to carry out a variety of visual tasks ranging from low level (e.g., image feature analysis) to high level (e.g., visual recognition), resulting in the realization of image understanding in a human-like way.

One of the most important problems in computer vision is how to model the semantic structural properties of an image space. In this area, the pioneering work by David Marr seeks to establish a computational theory of vision (Kitcher, 1988), which aims to form an effective symbolic description of the

image world from 2D to 3D. This vision theory is based on a structured computational pipeline for object recognition, which follows the stages of primal sketch, 2.5D sketch, and 3D model. In theory, these stages are supposed to cope with a set of challenging 3D geometry problems (e.g., feature representation, camera calibration, and 3D reconstruction), which are heuristic and highly dependent on expertise and experience. As a result, such a visual computing pipeline is often restricted in many practical scenarios, such as the drastic appearance changes and complicated environmental conditions.

Motivated by the above observations, which attempt to automatically build the feature representation machine in terms of supervised deep learning over massive visual data, a data-driven learning strategy is introduced to the visual computing process. The benefit of such a strategy is in taking an end-to-end learning architecture (Gordo *et al.*, 2016), which is formulated as deep neural networks that can learn informative features across multiple levels (e.g., edge, object shape, and object attention). Such a deep learning architecture is optimized in a GPU-driven parallel computing framework. In this way, the process of feature representation is governed by deep learning machines, which adaptively produce a set of visual task-specific features, instead of heuristically hand-crafted features, e.g., local binary pattern (LBP), histogram of oriented gradient (HOG), and Gist. Following this trend, a large body of deep learning approaches have been proposed to improve the power of feature representation, such as AlexNet (Krizhevsky *et al.*, 2012), VGG (Simonyan and Zisserman, 2014), GoogleNet (Szegedy *et al.*, 2015), and ResNet (He *et al.*, 2015). These approaches focus on designing more feasible network architectures to model the visual computing process from sparse connections to dense links. For the sake of video applications, the aforementioned learning machines are further extended to introduce spatio-temporal memory constraints into the deep learning process, e.g., recurrent neural network (RNN), LSTM (Liu *et al.*, 2016), and gated recurrent units (GRUs), which leads to the higher-order feature representation, taking into account the long-range dependency among visual elements in either the spatial or temporal dimensions.

More recently, many vision researchers have devoted their efforts to exploiting unsupervised learning for the ease of image sample annotation in deep learning. One of the most representative approaches is based on generative adversarial networks, which take an adversarial learning strategy to effectively balance the learning power of a discriminator and a generator network until convergence. As a result, they use the generator network to generate the synthesized samples, and let the discriminator network distinguish the differences between real samples and synthesized ones. By solving a joint min-max optimization problem iteratively, both of the networks are mutually reinforced until the discriminator network cannot distinguish the differences between the synthesized and real samples. In this case, the learned generator network can serve as the sample generator that can produce a set of high-quality training samples to improve the quality of deep learning.

Some vision learning methodologies (Shojaee and Baghshah, 2016; Rezende *et al.*, 2016) have started to focus on how to recognize the unseen object classes, without training data, on the basis of seen-class samples. Therefore, zero/one shot learning emerges to model the following two aspects: (1) modeling the semantic interactions between the image feature space and the label class space; (2) capturing the domain distribution connections between seen-class and unseen-class data. Many vision works rely on deep reinforcement learning, which aims to set up an interactive learning mechanism between learning agents and environments by adaptively learning reward and policy functions.

Of course, the ultimate goal of visual computing is to truly understand the semantics of visual data, and provide knowledge service to human beings. To that end, some recent vision works have made an attempt to incorporate prior knowledge into the process of visual computing, resulting in knowledge-guided visual computation. As a result, a number of knowledge representation techniques (Shijia *et al.*, 2016) (e.g., knowledge graphs) were introduced to improve the semantic understanding power of visual computing (e.g., knowledge-guided image captioning).

4 From data to knowledge

The rapid development of information technology has generated massive amounts of data globally,

80% of which is unstructured. The idea of converting unstructured data into formalized knowledge is so appealing that its practices can be traced back to classic AI research in the 1980s (Russell *et al.*, 2003). The major challenge of this task is to build up knowledge bases (or knowledge graphs) that can enable and contribute to intelligent applications, such as semantic search, question answering, or even reasoning and large-scale machine reading.

The construction of knowledge bases is a long-term studied area, and there are various ways of building one. The Cyc (Sarjant *et al.*, 2009) is one of the earliest attempts to create a universal knowledge base. The goal of Cyc is to enable AI applications to perform human-like reasoning by assembling a comprehensive knowledge base of everyday common sense knowledge. Cyc is fully created and refined by human exports through curation. However, after more than 10 years of hard endeavor including more than 900 person years, Cyc is still far from its ultimate goal, because the creation of a universal knowledge base by pure curation is infeasible and unsustainable.

Instead of relying on a small group of experts, modern knowledge bases are usually constructed in different ways. One way is to use the idea of crowdsourcing to share the workload, i.e., using a crowd of workers from the Internet to collaboratively edit a knowledge base. Typical examples of this kind are Freebase (Bollacker *et al.*, 2008) and Wikidata (Vrandečić and Krötzsch, 2014). Another way is to use automatic knowledge extraction methods to harvest knowledge from semi-structured data. For example, DBpedia (Auer *et al.*, 2007) is a knowledge base that is extracted from semi-structured data in Wikipedia. The main source for DBpedia is the key-value pairs in the Wikipedia infoboxes. The tasks needed for creating automatic tools to map infoboxes to the DBpedia ontology and corresponding properties are released and distributed in a crowd-sourced way.

The aforementioned two approaches have achieved great success in constructing general knowledge bases; however, extending or refining an existing knowledge base, or creating a domain specific knowledge base from scratch using unstructured text data sources, is still challenging, especially when confronted with large-scale unstructured data. A proposed solution to this challenge is to construct data-driven knowledge bases by using machine learning approaches. Its process can usually be divided into three sequential steps. The first step is mention detection, which aims to recognize mentions (proper noun phrases) from the text by their surface names, locate their span in the text, and classify them into a predefined set of categories (typing). The task of mention detection can be treated as a classification or sequence-labeling problem. The major challenge of this task is that an entity mention may have many surface names, and can be associated with multiple types. Identified mentions of the first step can be highly ambiguous, and therefore the second step, entity linking (EL), tends to link an identified mention to a specific entity in a predetermined knowledge base. The main information used to address such inherent ambiguity is the candidate entity popularity, context and topic similarities, as well as other mentions in the context. Learning methods such as regression and ranking are often applied in this scenario. As a disambiguation tool, EL can improve the performance of a retrieval system. The mentions that cannot link to the knowledge base will be clustered for further investigation. Therefore, EL can also be used to refine a knowledge base. Finally, the third step of this process is relation extraction (RE). RE extracts properties of linked or clustered mentions, and finds their relations from plain text documents. The subtask of property extraction is also called slot-filling, which aims to fill in values for specific slots (attributes) with reference to specific entities. Since it is not feasible to annotate large-scale documents with sentences that contain candidate entity-value pairs, distant supervision assumption, as well as multi-instance, multi-label learning is often applied to this task. Another subtask in this step, which aims to extract the relation between entities, is called open information extraction (IE). In open IE, relations are represented as subject–relation–object triples. Schemas for these relations do not need to be specified in advance. NLP tools such as chunker and dependency parser are used to provide features to segment sentences into relation arguments and relation names.

Data-driven knowledge base construction has received considerable attention recently. The U.S. National Institute of Standards and Technology (NIST) organizes the Annual Text Analysis Conference Knowledge Base Population (KBP) Workshop to promote research in automated data-driven

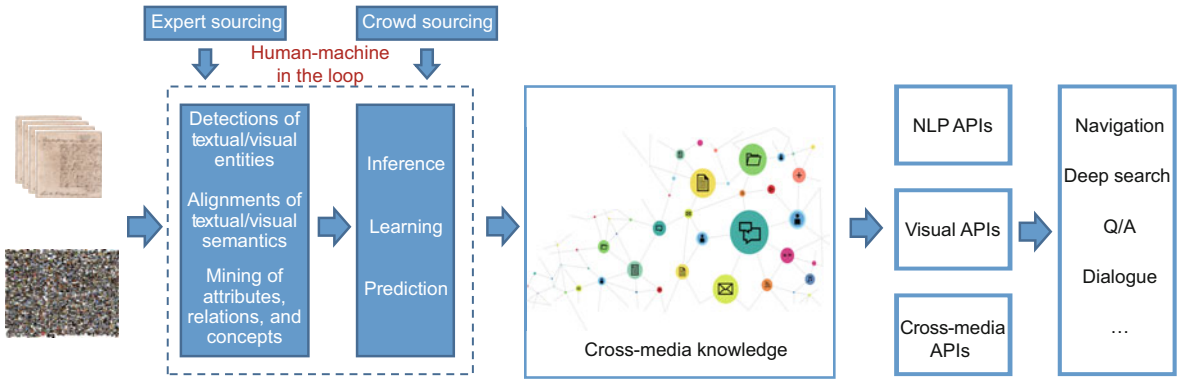

Fig. 2 Knowledge computation engine in KS-Studio

NLP: natural language processing; API: application programming interface

systems that discover information about the entities as found in a large unstructured text corpus, and incorporate this information to enrich a knowledge base. In other domains, such as biology, a series of BioCreAtIvE (a critical assessment of text mining methods in molecular biology) workshops have been organized with similar challenges such as biological entity extraction and gene name normalization (linking).

However, knowledge bases constructed by data-driven methods also suffer from several problems such as inaccurate entity recognition and unreliable property/relation discovery due to insufficient training data. One possible solution to these problems is to refine the knowledge that is learned by machines by human corrections. One of the earliest approaches that employs such human-machine collaboration is the never ending language learning (NELL) project (Carlson *et al.*, 2010). The project firstly extracts relation assertions by learning algorithms, and then sends them to anonymous online reviewers for correction. Finally, the corrected results are fed back to the learning algorithms as training examples. This process can be carried out iteratively to gradually improve the system performance. This kind of human-machine interaction can be viewed as a combination of crowdsourcing and data-driven methods. We believe that such human-machine collaboration has a great potential to improve the method of knowledge collection from unstructured data.

Fig. 2 illustrates the architecture of KS-Studio, a knowledge computation engine developed at Zhejiang University. KS-Studio is a knowledge computing engine consisting of a set of APIs and tools that convert unstructured data into structured one, covering the whole 'from data to knowledge' process. For example, it includes entity discovery, entity linking, slot filling, event extraction, image recognition, and cross-media analysis. KS-Studio can provide knowledge services with the help of knowledge deep understanding.

5 Visual analysis and visualization

The design of visualization and visual analysis has gone through three stages: visualization (visual representation), interactive visualization, and visual reasoning. The progress implicitly involves more and more human intelligence in the data process.

In the information visualization field, the pioneering visualization research is in statistical charts, which have been applied to various fields: sociology, economics, geography, medical sciences, astronomy, etc. To transform data into an effective statistical chart (visualization) is nontrivial. That is probably the reason why the visualization enterprise Tableau became so successful more than a hundred years after the statistical charts were first invented. The 'show me' feature (Mackinlay *et al.*, 2007) automatically suggests a proper visual representation and a color scheme, according to the underlying data. In the scientific data visualization field, visualization is generally about the design of the transfer function. The most representative scientific data visualization work is the visualization toolkit (VTK) (Schroeder *et al.*, 2004), which supports visualization algorithms including scalar, vector, tensor, texture, volumetric methods, as well as the advanced modeling techniques such as implicit modeling, polygon reduction, mesh smoothing, cutting, contouring, and Delaunay

triangulation. Visualization in the sense of visual representation implies the design of visual channels (color, shape, size, etc.) that facilitate human perception and cognition of the latent patterns in data.

With the increase of information collection and social connections, it became clear that visual representations alone could not satisfy the dynamic requirements of information understanding and communication. Interaction techniques such as focus + context (Baudisch *et al.*, 2002), overview + details (Shneiderman, 1996), and saliency-aware navigation (Kim and Varshney, 2006) were proposed to provide multiple perspectives of data patterns according to the users' demand. In the exploration of large-scale landscape images (gigapixels), saliency-guided navigation (Ip and Varshney, 2011) seems to be a reasonable solution to rapid identification of regions of interest. Although users are able to express their data requirements with interactive visualization, they are not able to contribute their domain knowledge to visualization for better data understanding. In other words, human intelligence was not part of the data fed into the visualization process.

There are two reasons for adopting visualization in data analysis: data are too complex to interpret, and data in their raw form still require further processing. Visualization (interactively) addresses the former challenge, while visual analysis addresses the latter. Different from interactive visualization, human intelligence integrated in a visual analysis process is not limited to navigation or parameter adjustment. More importantly, it also includes logical reasoning skills and domain knowledge that constitute an iterative knowledge generation model (Sacha *et al.*, 2014) for more profound findings. Palantir is one representative of a visual analytical tool for knowledge generation. It integrates human domain knowledge to build connections between data from various sources, from which analysts are able to validate obsolete knowledge and generate new knowledge.

Though visualization has been fully involved in every component of big data, we are still facing many challenges. First of all, in the era of big data, visualization is often faced with TB- or even PB-level data sets, which brings great computing efficiency and usability challenges to data cleaning, data statistics, feature extraction, and data display. Secondly, many existing visualization approaches perform post-hoc analysis. The expectation is to integrate visualization into real-time analysis systems, evolving from the analysis of static historical data to the analysis of dynamic streaming data, as well as forecasts of the future. Thirdly, many existing visualization applications are designed for only one type or class of specific data. General visualization methods and platforms for all kinds of cross-media data will need to be developed to standardize visualization, which is crucial for the future development, popularization, and application of visualization tools. Finally, in the future, visualization will be able to connect physical space with cyber space using large shared displays (Marrinan *et al.*, 2014), such as virtual reality (VR), augmented reality (AR), wearable, and other technologies, making visual interactions more natural.

6 Prospective trends and conclusions

We human beings learn by way of concrete examples, different forms of general knowledge, and rich experiences in the physical world. As pointed out in Pan (2016), with the rapid partial overlapping of cyberspace with physical space and human society (CPH), AI has been profoundly changed. Here we describe in detail some of the emerging trends from data to knowledge depicted in Fig. 1 as follows:

The effective integration of rule-based symbolic reasoning and data-driven learning (i.e., connectionist learning). An appropriate integration is desirable for enhancing the ability to explain intelligent actions, e.g., prediction and classification (Hu *et al.*, 2016).

Cross-media inference and reasoning. Data with multi-modalities from multi-domains can improve the robustness and reliability of inference and reasoning. In general, data with different modalities have different discriminative powers to encode their particular semantics. As a result, how to employ the intrinsic interaction between cross-media during inference and reasoning is a fundamental challenge to sense our real world.

Creative ability via artificial intelligence. In recent years, we have witnessed an explosion of AI generated arts such as pop music and painted photos. Whether next generation AI will have creative ability in some domain-specific field is an amazing research direction.

In this paper, we survey some recent advances in terms of AI platforms, NLP, multimedia, computer vision, knowledge base population, and visualization. We believe that an appropriate integration of data-driven machine learning approaches (bottom-up) with knowledge-guided methods (top-down) will open a new door for the future of AI.

Acknowledgements

The authors would like to thank the following contributors from the College of Computer Science and Technology, Zhejiang University: Wei CHEN, Xi LI, Si-liang TANG, Zhou ZHAO, Yang YANG, and Zi-cheng LIAO. Special thanks to Zhong-fei (Mark) ZHANG and Ya-hong HAN.

References

Abadi, M., Agarwal, A., Barham, P., *et al.*, 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. ePrint Archive, arXiv:1603.04467.

Auer, S., Bizer, C., Kobilarov, B., *et al.*, 2007. DBpedia: a nucleus for a web of open data. Proc. 6th Int. Semantic Web Conf. & 2nd Asian Semantic Web Conf., p.722-735. http://dx.doi.org/10.1007/978-3-540-76298-0_52

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. ePrint Archive, arXiv:1409.0473.

Baudisch, P., Good, N., Bellotti, V., *et al.*, 2002. Keeping things in context: a comparative evaluation of focus plus context screens, overviews, and zooming. Proc. SIGCHI Conf. on Human Factors in Computing Systems, p.259-266. http://dx.doi.org/10.1145/503376.503423

Bergstra, J., Breuleux, O., Bastien, F., *et al.*, 2010. Theano: a CPU and GPU math compiler in Python. Proc. 9th Python in Science Conf., p.1-7.

Bollacker, K., Evans, C., Paritosh, P., *et al.*, 2008. Freebase: a collaboratively created graph database for structuring human knowledge. Proc. ACM SIGMOD Int. Conf. Management of Data, p.1247-1250. http://dx.doi.org/10.1145/1376616.1376746

Brill, E., 1992. A simple rule-based part of speech tagger. Proc. Workshop on Speech and Natural Language, p.112-116. http://dx.doi.org/10.3115/1075527.1075553

Carlson, A., Betteridge, J., Kisiel, B., *et al.*, 2010. Toward an architecture for never-ending language learning. Proc. 24th AAAI Conf. on Artificial Intelligence, p.3-11.

Cho, K., Courville, A., Bengio, Y., 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. Multim.*, **17**(11):1875-1886. http://dx.doi.org/10.1109/TMM.2015.2477044

Collobert, R., Bengio, S., Mariéthoz, J., 2002. Torch: a Modular Machine Learning Software Library. IDIAP Research Report No. IDIAP-RR 02-46, Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland.

Gordo, A., Almazan, J., Revaud, J., *et al.*, 2016. End-to-end learning of deep visual representations for image retrieval. ePrint Archive, arXiv:1610.07940.

Harris, Z.S., 1954. Distributional structure. *In:* Hiz, H. (Ed.), Formal Linguistics Series. Springer Netherlands, Houten, Netherlands. http://dx.doi.org/10.1007/978-94-017-6059-1_36

He, K.M., Zhang, X.Y., Ren, S.Q., *et al.*, 2015. Deep residual learning for image recognition. ePrint Archive, arXiv:1512.03385.

Hu, Z.T., Ma, X.Z., Liu, Z.Z., *et al.*, 2016. Harnessing deep neural networks with logic rules. ePrint Archive, arXiv:1603.06318.

Ip, C.Y., Varshney, A., 2011. Saliency-assisted navigation of very large landscape images. *IEEE Trans. Visual. Comput. Graph.*, **17**(12):1737-1746. http://dx.doi.org/10.1109/TVCG.2011.231

Jia, Y.Q., Shelhamer, E., Donahue, J., *et al.*, 2014. Caffe: convolutional architecture for fast feature embedding. Proc. 22nd ACM Int. Conf. on Multimedia, p.675-678. http://dx.doi.org/10.1145/2647868.2654889

Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences. ePrint Archive, arXiv:1404.2188.

Karpathy, A., Joulin, A., Li, F.F.F., 2014. Deep fragment embeddings for bidirectional image sentence mapping. Proc. Advances in Neural Information Processing Systems, p.1889-1897.

Kim, Y.M., Varshney, A., 2006. Saliency-guided enhancement for volume visualization. *IEEE Trans. Visual. Comput. Graph.*, **12**(5):925-932. http://dx.doi.org/10.1109/TVCG.2006.174

Kitcher, P., 1988. Marr's computational theory of vision. *Philos. Sci.*, **55**(1):1-24.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. 26th Annual Conf. on Neural Information Processing Systems, p.1097-1105.

Lee, C.Y., Xie, S., Gallagher, P., *et al.*, 2015. Deeply-supervised nets. Artificial Intelligence and Statistics Conf., p.562-570.

Li, J.W., Monroe, W., Ritter, A., *et al.*, 2016. Deep reinforcement learning for dialogue generation. ePrint Archive, arXiv:1606.01541.

Liu, Y., Sun, C.J., Lin, L., *et al.*, 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. ePrint Archive, arXiv:1605.09090.

Low, Y.C., Gonzalez, J.E., Kyrola, A., *et al.*, 2014. GraphLab: a new framework for parallel machine learning. ePrint Archive, arXiv:1408.2041.

Mackinlay, J., Hanrahan, P., Stolte, C., 2007. Show me: automatic presentation for visual analysis. *IEEE Trans. Visual. Comput. Graph.*, **13**(6):1137-1144. http://dx.doi.org/10.1109/TVCG.2007.70594

Marrinan, T., Aurisano, J., Nishimoto, A., *et al.*, 2014. SAGE2: a new approach for data intensive collaboration using scalable resolution shared displays. Int. Conf. on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), p.177-186. http://dx.doi.org/10.4108/icst.collaboratecom.2014.257337

McCarthy, J., Minsky, M.L., Rochester, N., *et al.*, 2006. A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Mag.*, **27**(4):12-14.

Mikolov, T., Chen, K., Corrado, G., *et al.*, 2013. Efficient estimation of word representations in vector space. ePrint Archive, arXiv:1301.3781.

Neal, R.M., 2012. Bayesian Learning for Neural Networks. Springer Science & Business Media, Berlin, Germany.

Pan, Y.H., 2016. Heading toward artificial intelligence 2.0. *Engineering*, **2**(4):409-413. http://dx.doi.org/10.1016/J.ENG.2016.04.018

Rezende, D.J. Mohamed, S., Danihelka, I., *et al.*, 2016. One-shot generalization in deep generative models. ePrint Archive, arXiv:1603.05106.

Russell, S.J., Norvig, P., Canny, J., *et al.*, 2003. Artificial Intelligence: a Modern Approach. Prentice Hall, Upper Saddle River, USA.

Sacha, D., Stoffel, A., Stoffel, F., *et al.*, 2014. Knowledge generation model for visual analytics. *IEEE Trans. Visual. Comput. Graph.*, **20**(12):1604-1613. http://dx.doi.org/10.1109/TVCG.2014.2346481

Sarjant, S., Legg, C., Robinson, M., *et al.*, 2009. All you can eat ontology-building: feeding Wikipedia to Cyc. Proc. Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology, p.341-348. http://dx.doi.org/10.1109/WI-IAT.2009.60

Schroeder, W.J., Lorensen, B., Martin, K., 2004. The Visualization Toolkit: an Object-Oriented Approach to 3D Graphics. Kitware, New York, USA.

Shijia, E., Jia, S.B., Yang, X., *et al.*, 2016. Knowledge graph embedding for link prediction and triplet classification. China Conf. on Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data, p.228-232. http://dx.doi.org/10.1007/978-981-10-3168-7_23

Shneiderman, B., 1996. The eyes have it: a task by data type taxonomy for information visualizations. Proc. IEEE Symp. on Visual Languages, p.336-343. http://dx.doi.org/10.1109/VL.1996.545307

Shojaee, S.M., Baghshah, M.S., 2016. Semi-supervised zero-shot learning by a clustering-based approach. ePrint Archive, arXiv:1605.09016.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. ePrint Archive, arXiv:1409.1556.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Conf. on Neural Information Processing Systems, p.3104-3112.

Szegedy, C., Liu, W., Jia, Y.Q., *et al.*, 2015. Going deeper with convolutions. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-9.

Vrandečić, D., Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, **57**(10):78-85. http://dx.doi.org/10.1145/2629489

Weston, J., Chopra, S., Bordes, A., 2014. Memory networks. ePrint Archive, arXiv:1410.3916.

Wu, F., Yu, Z., Yang, Y., *et al.*, 2014. Sparse multi-modal hashing. *IEEE Trans. Multim.*, **16**(2):427-439. http://dx.doi.org/10.1109/TMM.2013.2291214

Wu, F., Jiang, X.Y., Li, X., *et al.*, 2015. Cross-modal learning to rank via latent joint representation. *IEEE Trans. Imag. Process.*, **24**(5):1497-1509. http://dx.doi.org/10.1109/TIP.2015.2403240

Zhuang, Y.T., Song, J., Wu, F., *et al.*, 2016. Multi-modal deep embedding via hierarchical grounded compositional semantics. *IEEE Trans. Circ. Syst. Video Technol.* http://dx.doi.org/10.1109/TCSVT.2016.2606648