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:

Cross-media analysis and reasoning: advances and directions*

Yu-xin PENG[†1], Wen-wu ZHU[†‡2], Yao ZHAO[3], Chang-sheng XU[4], Qing-ming HUANG[5], Han-qing LU[4], Qing-hua ZHENG[6], Tie-jun HUANG[7], Wen GAO[7]

([1]*Institute of Computer Science and Technology, Peking University, Beijing 100871, China*)
([2]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*)
([3]*Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China*)
([4]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*)
([5]*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China*)
([6]*Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*)
([7]*School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*)
[†]E-mail: pengyuxin@pku.edu.cn; wwzhu@tsinghua.edu.cn

Received Dec. 7, 2016; Revision accepted Dec. 30, 2016; Crosschecked Jan. 1, 2017

Abstract: Cross-media analysis and reasoning is an active research area in computer science, and a promising direction for artificial intelligence. However, to the best of our knowledge, no existing work has summarized the state-of-the-art methods for cross-media analysis and reasoning or presented advances, challenges, and future directions for the field. To address these issues, we provide an overview as follows: (1) theory and model for cross-media uniform representation; (2) cross-media correlation understanding and deep mining; (3) cross-media knowledge graph construction and learning methodologies; (4) cross-media knowledge evolution and reasoning; (5) cross-media description and generation; (6) cross-media intelligent engines; and (7) cross-media intelligent applications. By presenting approaches, advances, and future directions in cross-media analysis and reasoning, our goal is not only to draw more attention to the state-of-the-art advances in the field, but also to provide technical insights by discussing the challenges and research directions in these areas.

Key words: Cross-media analysis; Cross-media reasoning; Cross-media applications
http://dx.doi.org/10.1631/FITEE.1601787 **CLC number:** TP391

1 Introduction

Along with the progress of human civilization and the development of science and technology, information acquisition, transmission, processing, and analysis have gradually changed from one form of media to multiple types of media such as text, image, video, audio, and stereo picture. Different media types on various platforms and modalities from social, cyber, and physical spaces are now mixed together to demonstrate rich natural and social properties. As a whole they represent comprehensive knowledge and reflect the behavior of individuals and groups. Consequently, a new form of information is recognized, known as cross-media information.

Over the past several decades, as the requirements for data management and utilization have increased significantly, multimedia information processing and analysis has been a research hotspot (Lew *et al.*, 2006). However, previous studies were devoted mainly to scenarios involving a single media. Research in cognitive science indicates that in the human brain, cognition of the environment is through the fusion of multiple sensory organs (McGurk and MacDonald, 1976). Although the representations of

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 61371128, U1611461, 61425025, and 61532005)
ORCID: Yu-xin PENG, http://orcid.org/0000-0001-7658-3845

different media types are heterogeneous, they may share the same semantics, and have rich latent correlations. Consider the topic of 'bird' as an example. All of the texts, images, videos, audio clips, and stereo pictures about this topic describe the same semantic concept 'bird' from complementary aspects. As a result, due to limitations in information diversity, traditional single-media analysis methods have difficulty in achieving the goal of semantic extraction from multiple modalities, and cannot deal with the analysis of cross-media data. Meanwhile, traditional reasoning methods are mainly text-based and perform reasoning under fully defined premises. They cannot deal with cross-media scenarios with sophisticated compositions, different representations, and complex correlations. Therefore, a key problem in research and application has been how to simulate the human brain's process of transforming environmental information to analytical models through vision, audition, language, and other sensory channels, and further to realize cross-media analysis and reasoning.

The topic of cross-media analysis and reasoning has attracted considerable research interest. With respect to cross-media analysis, existing studies focus mainly on modeling correlations and generating a uniform representation of two media types as in the popular correlation analysis method, called canonical correlation analysis (CCA) (Hotelling, 1936). Though there are limited studies on cross-media reasoning so far, it is an important future direction to extend traditional text-based reasoning methods to cross-media scenarios. There are also wide prospects for applications in cross-media analysis and reasoning. Effective yet efficient cross-media methods can provide more flexible and convenient ways to retrieve and manage multimedia big data. Users would like to adopt the cross-media intelligent engine for applications such as cross-media retrieval, and cross-media technology is also useful for important application scenarios, such as web content monitoring, web information trend analysis, and healthcare data fusion and reasoning. However, there still exist important challenges for cross-media intelligent applications.

Cross-media analysis and reasoning has been an active research area in computer science, and an important future direction in artificial intelligence. As discussed in Pan (2016), cross-media intelligence plays the role of a cornerstone in artificial intelligence, through which the machines can recognize the external environment. Although considerable improvement has been made in the research of cross-media analysis and reasoning (Rasiwasia *et al.*, 2010; Yang *et al.*, 2012; Peng *et al.*, 2016a; 2016b), there remain some important challenges and unclear points in future research directions. In this paper, we give a comprehensive overview of not only the advances achieved by existing studies, but also future directions for cross-media analysis and reasoning. The aim is to attract more researchers to the research field in cross-media analysis and reasoning, and thus we provide insights by discussing challenges and research directions, to facilitate new studies and applications on this new and exciting research topic.

2 Cross-media analysis and reasoning

The advances and directions in cross-media analysis and reasoning can be summarized as seven parts: (1) theory and model for cross-media uniform representation; (2) cross-media correlation understanding and deep mining; (3) cross-media knowledge graph construction and learning methodologies; (4) cross-media knowledge evolution and reasoning; (5) cross-media description and generation; (6) cross-media intelligent engines; (7) cross-media intelligent applications. In this section, we will provide descriptions of these seven parts, so as to present a comprehensive overview of cross-media analysis and reasoning.

2.1 Theory and model for cross-media uniform representation

Cross-media data naturally carries different kinds of information, which needs to be integrated to obtain comprehensive results in real-world applications. A fundamental research problem is how to learn uniform representation for cross-media data. Generally, this approach tries to build a commonly shared space where similarities between heterogeneous data objects can be computed directly using common distance metrics like Euclidean and cosine distances after mapping data into this space (Fig. 1). In this way, the heterogeneous gap among data from different modalities is reduced. To this end, two issues should be addressed: (1) how to build the shared space; (2) how to project data into it. To deal with these issues, learning schemes based on different models have been proposed recently.

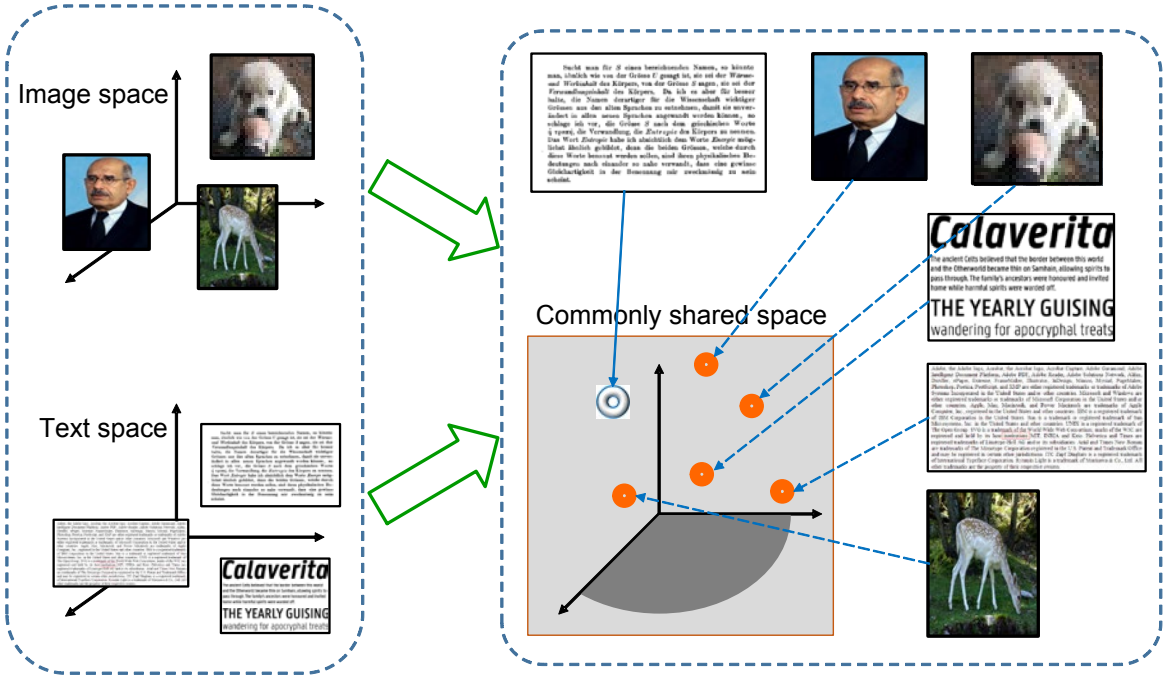

Fig. 1 An example of uniform representation methods for multimodal data (considering the images and texts as examples)

To the best of our knowledge, the first well-known cross-media model is based on CCA (Rasiwasia *et al.*, 2010). It learns a commonly shared space by maximizing the correlation between pairwise co-occurring heterogeneous data and performs projection by linear functions. Although the scheme is simple, it has inspired subsequent studies. CCA has many variants (Andrew *et al.*, 2013; Gong *et al.*, 2014; Rasiwasia *et al.*, 2014). For example, Andrew *et al.* (2013) extended this method using a deep learning technique to learn the correlations more comprehensively than those using CCA and kernel CCA. These methods can, for the most part, model only the correlations of two media types. To overcome this limitation, researchers have also attempted to develop datasets and methods for scenarios with more media types. For example, the newly constructed XMedia dataset (http://www.icst.pku.edu.cn/mipl/XMedia) is the first dataset containing five media types (text, image, video, audio, and 3D model), and methods such as those proposed by Zhai *et al.* (2014) and Peng *et al.* (2016b) can jointly model the correlations and semantic information in a unified framework with graph regularization for the five media types on the XMedia dataset. Yang *et al.* (2008) introduced another model called the multimedia document (MMD) to represent data, where each MMD is a set of media objects of different modalities but carrying the same semantics. The distances between MMDs are related to each modality, and in this way we can perform cross-media retrieval. Daras *et al.* (2012) employed a radial basis function (RBF) network to address the problem of missing modalities. However, the main problem with the MMD is that it only handles data from different modalities together, which is not flexible in many applications. Most cross-media representation learning models still belong to subspace learning techniques.

The topic model is another frequently used technique in cross-media uniform representation learning tasks, assuming that heterogeneous data containing the same semantics shares some latent topics. For example, Roller and Schulte im Walde (2013) integrated visual features into latent Dirichlet allocation (LDA) and proposed a multimodal LDA model to learn representations for textual and visual data. Wang Y *et al.* (2014) proposed a scheme called the multimodal mutual topic reinforce model (M^3R), which seeks to discover mutually consistent semantic topics via appropriate interactions between model factors. These schemes represent data as topic distributions, and similarities are measured by the

likelihood of observed data in terms of latent topics. Metric learning is usually performed if we know which data pairs are similar and which are dissimilar from heterogeneous modalities. An appropriate distance metric is designed to measure heterogeneous similarity, and learned using the given labeled data pairs to achieve the best performance. When the learned metric is decomposed into modality-specific projection functions (Wu *et al.*, 2010), data can be explicitly projected into a uniform representation as CCA does. Apart from the above-mentioned models, Mao *et al.* (2013) proposed a manifold-based model called parallel field alignment retrieval (PFAR), which considers cross-media retrieval as a manifold alignment problem using parallel fields.

In recent years, since deep learning has shown superiority in image classification (Krizhevsky *et al.*, 2012) and image content representation (Babenko *et al.*, 2014), it has also been widely used in cross-media research to learn uniform representations. Ngiam *et al.* (2011) proposed an autoencoder model to learn uniform representations for speech audios coupled with videos of the lip movements. Srivastava and Salakhutdinov (2012) introduced a deep restricted Boltzmann machine to learn joint representations for multimodal data. Andrew *et al.* (2013) proposed a deep CCA method which is a deep extension of the traditional CCA method. Socher *et al.* (2014) introduced dependency tree recursive neural networks (DT-RNNs), employing dependency trees to embed sentences into a vector space in order to retrieve images described by those sentences. Feng *et al.* (2014) and Wang W *et al.* (2014) applied autoencoders to perform cross-modality retrieval. More recently, Wang *et al.* (2015) proposed a multimodal deep learning scheme to learn accurate and compact multimodal representations for multimodal data (Fig. 2). This method facilitates efficient similarity search and other related applications on multimodal data. Zhang *et al.* (2014a) presented an attribute discovery approach, named the independent component multimodal autoencoder (ICMAE), which can learn shared high-level representation to identify attributes from a set of image and text pairs. Zhang *et al.* (2016) further proposed to learn image-text uniform representation from web social multimedia content, which is noisy, sparse, and diverse under weak supervision. Wei *et al.* (2017) proposed a deep semantic matching (deep-SM) method that uses the convolutional neural network and fully connected network to map images and texts into their label vectors, achieving state-of-the-art accuracy. The cross-media multiple deep network (CMDN) (Peng *et al.*, 2016a) is a hierarchical structure with multiple deep networks, and can simultaneously preserve intra-media and inter-media information to further improve the retrieval accuracy.

Although there are significant research efforts on uniform representation learning for cross-media analysis tasks, a large gap still exists between these methods and user expectations. This is caused by the fact that existing schemes still have not achieved a satisfactory performance; i.e., their accuracies are far

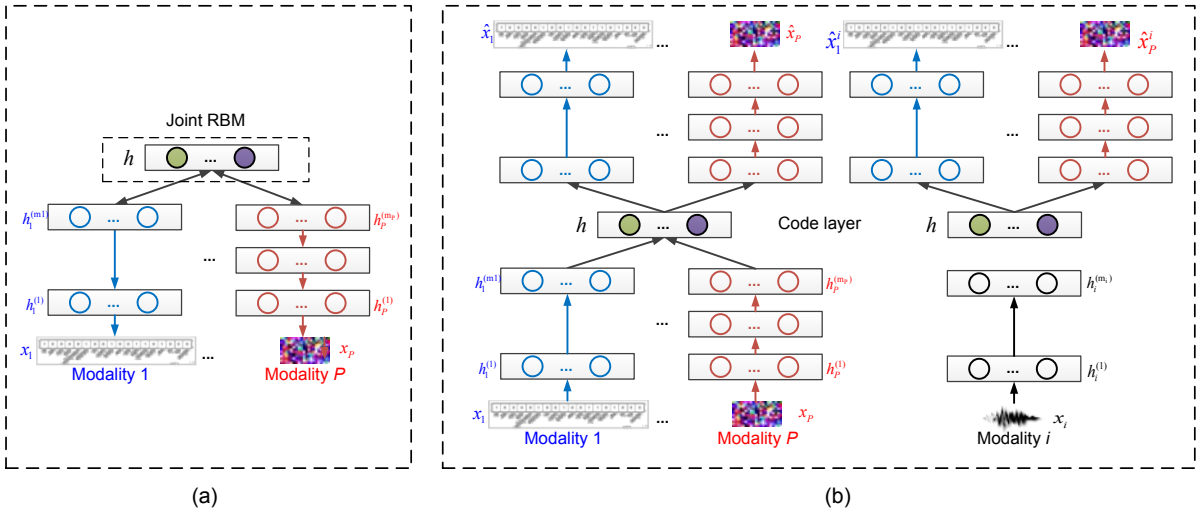

Fig. 2 The framework for compact multimodal representation learning, where (a) represents the pretraining stage and (b) represents the fine-tuning stage

from acceptable. Therefore, we still need to investigate better uniform representation methods for cross-media research.

2.2 Cross-media correlation understanding and deep mining

Cross-media correlations describe specific types of statistical dependencies among homogeneous and heterogeneous data objects. For example, if two images are taken from the same location, they may be intrinsically correlated from content, attribute, and topic perspectives, and thus they may share certain levels of intrinsic semantic consistency. The content in the paragraphs and social comments on a video webpage is semantically related to the content of the video itself. The aim of cross-media correlation learning is to construct metrics on heterogeneous data representation to measure how they are semantically relevant.

Existing cross-media correlation mining methods focus mainly on finding the common subspace where different modalities of data have semantic correlations. Researchers from the multimedia community have conducted extensive studies along this line. For example, in Feng *et al.* (2014), the correspondence autoencoder deep network was proposed to be trained on the raw features of different modalities, and then the combined multimodal deep feature was extracted for cross-media relevance measurement (Fig. 3). Zhang *et al.* (2014b) measured the correlations between visual and acoustic modalities by examining the visual-acoustic statistical relevance. However, cross-media correlation mining goes far beyond subspace learning. In many scenarios, the representation of cross-media data objects cannot be directly obtained. For example, there is no given feature representation for a structured cross-media object such as a set of hyperlinked multimedia documents or points-of-interest (POI). In such cases, the correlations can be inferred directly from the cross-media data objects by constructing appropriate information averaging mechanisms in a matrix completion framework to predict or complete the missing values in the object correlation description (Zhang *et al.*, 2015).

Following another line of research, researchers from the database community have investigated the correlations and fusion among unstructured, semi-structured, and structured data. However, most of these studies are based on low-level features and formats. Few studies are focusing on multimodal content and high-level correlations, e.g., generating a description for the entities by fusing semi-structured Wiki data and unstructured web data. Moreover, cross-media data is not only from different modalities and structures, but also from different sources. The study of associating and fusing cross-media data from different sources remains in its infancy, e.g., objective data and subjective user-generated content (UGC), user data from different online social networks (OSNs), and cross-space data from cyber and physical spaces.

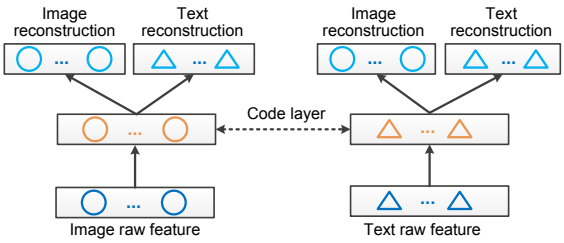

Fig. 3 Correspondence full-modal autoencoder (Feng *et al.*, 2014)

In cross-media deep mining, the knowledge base is manually and professionally edited by experts in traditional expert systems. Currently, many studies are focusing on extracting and learning knowledge from data automatically, e.g., Google Knowledge Vault (Dong *et al.*, 2014). However, similar to data, knowledge is essentially cross-media. Recently we have seen a rapid development of different types of intelligent perceptions, e.g., vision-based environmental perception in Visual SLAM (Fuentes-Pacheco *et al.*, 2015) and multimodal based human-computer interaction in gesture and action recognition (Rautaray and Agrawal, 2015). Moreover, ubiquitous perception has received increasing attention these days (Adib *et al.*, 2015). Development in the above areas provides opportunities to research the problem of cross-media knowledge mining. While critical challenges exist in constructing the cross-media knowledge base, it is of great theoretical and technical significance to combine perceptions from different modalities to supplement and improve the current text-based knowledge base.

Despite the achievements in cross-media correlation understanding, there is still a long way to go in

this research direction. Basically, existing studies construct correlation learning on cross-media data with representation learning, metric learning, and matrix factorization, which are usually performed in a batch learning fashion and can capture only the first-order correlations among data objects. How to develop more effective learning mechanisms to capture the high-order correlations and adapt to the evolution that naturally exists among heterogeneous entities and heterogeneous relations, is the key research issue for future studies in cross-media correlation understanding.

2.3 Cross-media knowledge graph construction and learning methodologies

The aim of cross-media knowledge graph construction is to represent framed rules, values, experiences, contexts, instincts, and insights with entities and relations from general to specific domains (Davenport and Prusak, 1998). In cross-media research, the entities and relations are defined and extracted from not only the textual data corpus, but also numerous loosely correlated data modalities including texts, images, videos, and other related information sources. Cross-media knowledge graphs provide essential computable knowledge representation structures for semantic correlation analysis and cognition-level reasoning in cross-media context, facilitating theoretical and technical development in cross-media intelligence and a diversified range of applications.

In recent decades, research efforts on knowledge graphs have been devoted to two aspects. First, knowledge graphs are used to represent general or domain-specific knowledge. Two primary elements in knowledge graphs are entities (a.k.a. ontologies) and relations. The set of entities for knowledge graph construction is defined by either domain expertise or existing entity sets, e.g., WordNet (Fellbaum and Miller, 1998), Wikipedia, and FreeBase. The relations, represented as edges with real values between the entities, are employed to reflect structural or statistical entity dependency in certain domain contexts. Most existing knowledge graphs are constructed on a textual data corpus using natural language processing (Carlson *et al*., 2010) and co-occurrence statistics (Cilibrasi and Vitanyi, 2007). In visual modalities, significant efforts have been devoted to constructing knowledge bases to describe relations between visual objects, scenes, and attributes (Deng *et al*., 2009; Chen X *et al*., 2013; Prabhu and Babu, 2015). For example, NEIL (Chen X *et al*., 2013) presents a never-ending learning system for visual ontology construction from image search engines, which iterates between concept relationship extraction, image instance recognition, and concept classifier/detector learning. Fang *et al*. (2016) proposed a multimodal ontology construction solution by considering both textual and visual information in extracting entity relationships. Zhu *et al.* (2015) proposed a scalable multimodal knowledge base construction system, and defined three types of relations: image-label, intra-correlations, and inter-correlations. Sadeghi *et al.* (2015) developed the visual knowledge extraction system (VisKE), which can extract some general relationships like ‘eat’ and ‘ride’ from the context of image and text. Hua *et al*. (2014) went beyond ontology co-occurrences (Cilibrasi and Vitanyi, 2007) in most of the existing visual knowledge bases, and proposed to measure the ontology similarity by combining visual, textual, and semantics cues. By designing human-expert-powered, semi-automatic, and fully automatic procedures, diverse types of knowledge graphs have been constructed and released for real applications, containing more than 60 billion ontologies and trillions of facts/relations, and covering a wide range of domains from geography to life science (Fig. 4). Unfortunately, none of them are specifically designed to represent knowledge in cross-media data.

The second area of focus on knowledge graphs is how to deploy knowledge graphs to enhance the performance and user experience in information retrieval and web applications, especially in the era of big data. As a pioneering work, Garfield (2004) developed the HistCite software to generate knowledge graphs in academic literature, which led to the birth of the academic search engine CiteSeer. The Knowledge Graph released by Google in 2012 (Singhal, 2012) provided a next-generation information retrieval service with ontology-based intelligent search based on free-style user queries. Similar techniques, e.g., Safari, were developed based on achievements in entity-centric search (Lin *et al*., 2012). However, existing entity-based search engines cannot perform fully automatic content parsing on heterogeneous modalities, and thus they cannot

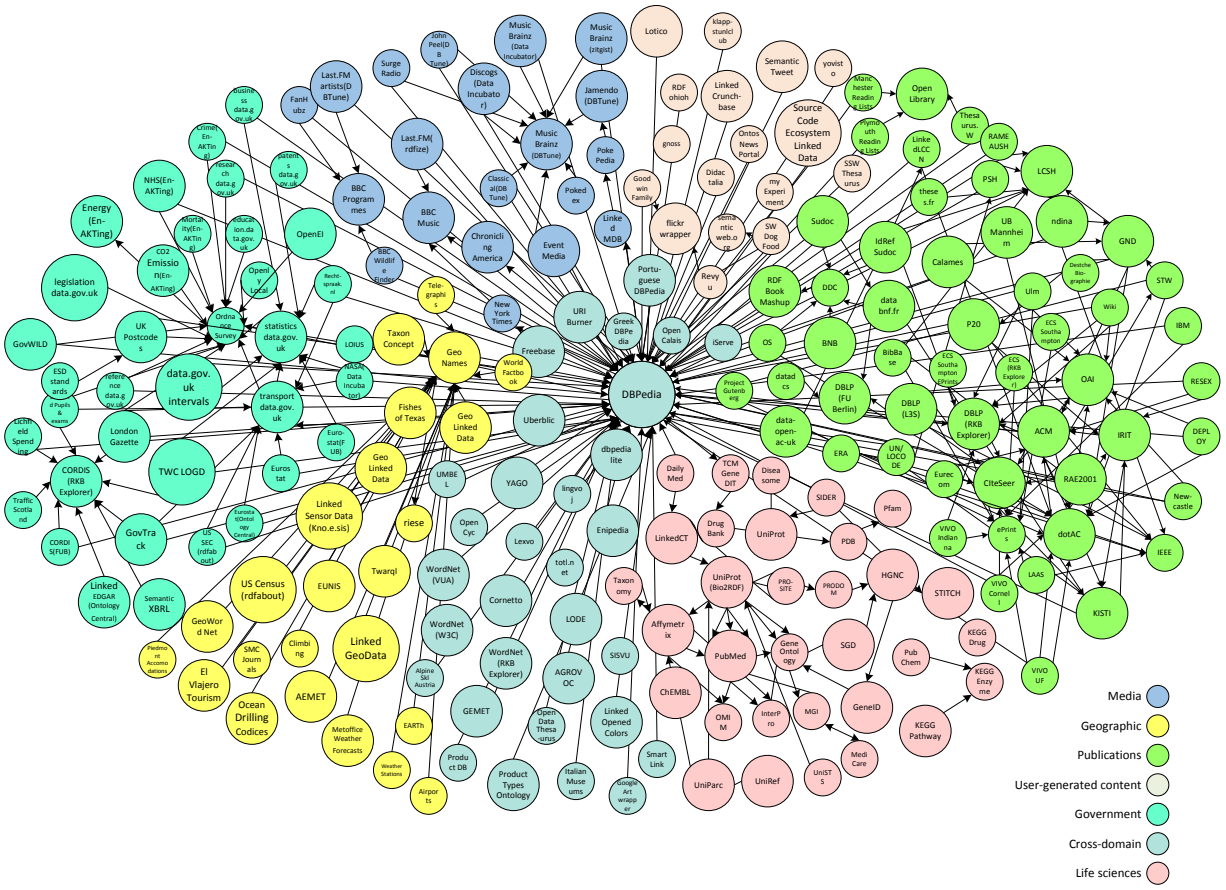

Fig. 4 An illustration of existing knowledge graphs

provide entity-based information retrieval for cross-media content.

To transform the web of data into a web of knowledge (Suchanek and Weikum, 2014), several issues should be considered in research on cross-media knowledge graphs. First of all, effective and efficient techniques for entity extraction and relation construction from heterogeneous cross-media information sources should be studied. Second, information search and retrieval based on cross-media knowledge graphs should be investigated to provide more effective knowledge harvesting and information seeking mechanisms for more diverse application contexts. Third, mining and reasoning in cross-media knowledge graphs should be developed to facilitate knowledge acquisition and high-level reasoning for real applications. Finally, knowledge-driven cross-media learning models will be required in the near future to achieve more generalization and learning capabilities, resulting in more advanced cross-media intelligence.

2.4 Cross-media knowledge evolution and reasoning

Early artificial intelligence systems mostly depend on texts to perform reasoning using predicates, propositions, and rules, under fully defined premises. Judea Pearl, as a National Academy of Engineering (NAE) member and Turing Award winner, was one of the pioneers in probabilistic and causal reasoning in artificial intelligence (Pearl, 2000). Radinsky *et al.* (2012) proposed a methodology for modeling and predicting future events by generalizing examples of causality pairs to infer a causality predictor. Some research institutions, such as Illinois State University and Michigan State University, attempted to estimate county health statistics (Culotta, 2014) and the unemployment rate (Antenucci *et al.*, 2013) by analyzing social media content. Google Flu Trends (Ginsberg *et al.*, 2009) analyzes search patterns on the web search engine to help predict the spread of influenza (Fig. 5).

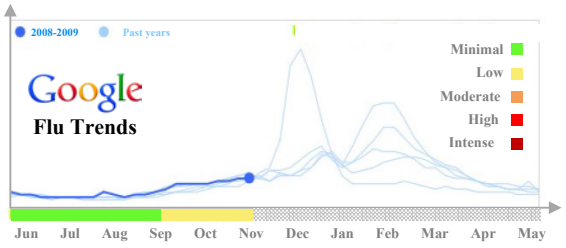

Fig. 5 Google Flu Trends

In addition, it has been shown that some learning mechanisms, such as reinforcement learning and transfer learning, can be helpful for constructing more complex intelligent reasoning systems (Lazaric, 2012). Furthermore, lifelong learning (Lazer *et al.*, 2014) is the key capability of advanced intelligence systems. For example, Google DeepMind has constructed a machine intelligence system based on a reinforcement learning algorithm (Gibney, 2015), which beat humans at classic video games. Recently, AlphaGo, developed by Google DeepMind, has been the first computer Go program that can beat a top professional human Go player. It even beat the world champion Lee Sedol in a five-game match. We have witnessed increasing numbers of intelligence systems winning human-machine competitions.

However, the knowledge and reasoning process in the real world usually involves collaboration among language, vision, and other types of media data. Most existing intelligent systems exploit only the information from a single media type, such as text, to perform reasoning processes. There have been some recent works involving reasoning on cross-media data. Visual question answering (VQA) can be regarded as a good example of cross-media reasoning (Antol *et al.*, 2015). VQA aims to provide natural language answers for questions given in the form of combination of the image and natural language. Johnson *et al.* (2015) attempted to improve the accuracy of image retrieval with the assistance of the scene graph, which also shows the idea of cross-media reasoning. A scene graph presents objects and their attributes and relationships, which can be used to guide image retrieval at the semantic level. However, it is still hard for these systems to make full use of the rich semantic information contained in complementary media types, and they cannot perform complex cross-media analysis and reasoning on multimedia big data. Therefore, the problem of performing cross-media reasoning based on multiple media types rather than on only text information, has become important in both research and application areas. Note that there is little research on cross-media knowledge evolution and reasoning, and many key problems need to be solved, which include, for instance, the acquisition, representation, mining, learning, and reasoning of cross-media knowledge, and the construction of large-scale cross-media knowledge bases. We still need to confront the significant challenges that are involved in constructing cross-media reasoning systems for real applications.

To address the problems noted above, several issues should be studied further. First, it is important to study data-driven and knowledge-guided cross-media knowledge learning methods. Second, cross-media reasoning frameworks based on semantic understanding should be constructed with technologies such as cross-media deep learning and multi-instance learning. Third, never-ending knowledge acquisition, mining, and evolution processes should be comprehensively investigated in future work.

2.5 Cross-media description and generation

Cross-media description and generation aims to realize cross-translation among text, image, video, and audio information, and link the multimodal understanding with natural language descriptions, where visual content description is the most challenging task. Therefore, we will stress this challenge in the following discussion. Visual content description is a new research direction integrating natural language processing and computer vision. It requires not only the recognition of visual objects and their semantic interactions, but also the ability to capture visual-language interactions and learn how to translate the visual understanding into sensible sentence descriptions. Fig. 6 shows some examples of visual content descriptions.

Existing studies on visual content description can be divided into three groups. The first group, based on language generation, first understands images in terms of objects, attributes, scene types, and their correlations, and then connects these semantic understanding outputs to generate a sentence description using natural language generation techniques, e.g., templates (Yang *et al.*, 2011), n-grams (Kulkarni *et al.*, 2011), and grammar rules

(Kuznetsova *et al.*, 2014). These methods are direct and intuitive, but the sentences generated are limited by their syntactic dependency and thus are inflexible.

Fig. 6 Examples of visual content descriptions, where (a)–(c) represent image descriptions and (d) represents a video description

The second group covers retrieval-based methods, retrieving content that is similar to a query and transferring the descriptions of the similar set to the query. According to the differences in the retrieval feature space, studies in this group include two types, i.e., retrieval in a uni-modal space (Ordonez *et al.*, 2011) and in a multimodal space (Hodosh *et al.*, 2013). The former aims to search for similar images or videos in the visual feature space, and the latter projects images or videos and sentence features into a common multimodal space, and searches for similar content in the projected space. Sentences obtained with these methods are more natural and grammatically correct, but they usually suffer with regard to generating variable-length and novel sentences.

The third group is based on deep neural networks, employing the CNN-RNN codec framework, where the convolutional neural network (CNN) is used to extract features from images, and the recursive neural network (RNN) (Socher *et al.*, 2011) or its variant, the long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997), is used to encode and decode language models. These methods typically use neural networks for both image-text embedding and sentence generation (Karpathy and Li, 2015; Vinyals *et al.*, 2015), and visual attention (Xu *et al.*, 2015) or semantic guidance (Jia *et al.*, 2015) is also integrated in the model learning to further improve the performance. Compared with the other methods, the deep models benefit from a stronger feature expression ability from CNN and capture dynamic spatio-temporal information with RNN, and thus they receive more attention. However, it is still a preliminary exploration and there exist many problems regarding further research: (1) As the parameter size of deep neural network is huge, it demands large amounts of annotated data for training and is easy to overfit, which makes sentence generation depend heavily on the training set; (2) The global features from CNN have difficulty in representing local objects accurately, which results in incorrect or missing descriptions of local objects, especially their correlation in images.

In conclusion, the current research is centered mainly on natural language descriptions of single-media content, and improvements are needed in the areas of training set collection and application, model building, and efficient learning and optimization modeling with human cognition. Furthermore, the cross-media descriptions of text, image, video, and audio are rarely involved, such as image generation from text and video generation from audio. Considering that human cognition is an integrated understanding procedure of different types of sensory information, it becomes a very challenging but valuable task to implement a comprehensive and accurate description of multimodal information with natural language processing. The connections with complex cognition, human emotion, and logical reasoning are also attractive areas for in-depth exploration.

2.6 Cross-media intelligent engines

The intelligent engine is a kind of intelligent analysis and reasoning system having specific purposes and common knowledge. With the rapid developments in artificial intelligence, some international companies and research institutions have implemented text-based artificial intelligent systems with specific capabilities. Technology companies such as Google, Baidu, and Microsoft have proposed the concept of intelligent search and the framework for search techniques (Uyar and Aliyu, 2015). Based on the highly effective indexing of big data, intelligent search attempts to realize intelligent and humanized information services, allowing users to retrieve whatever they want with input in natural language forms. It can provide more convenient and accurate search results than traditional search engines. In the field of medical treatment, researchers have also proposed the technological concept of the

intelligent medical search engine (Luo and Tang, 2008).

In the late 1990s, Deep Blue and Deeper Blue, developed by IBM, were the first computer chess-playing systems that won a chess match against a reigning world champion (Hsu, 2002). Siri, an intelligent personal assistant developed by Apple Inc. in 2010, is powered by natural language understanding and driven by entity or ontology based technologies. Later in 2011, IBM's DeepQA project developed a question answering computer system, named Watson as shown in Fig. 7 (Ferrucci *et al*., 2013), which was specifically designed to play on the quiz show Jeopardy, and it won the first-place prize. It has been further improved for the Q&A service in medical diagnosis. The chatbots Xiaobing and Tay on Twitter developed by Microsoft, can improve their own intelligence level through communication with human users. DeepMind has proposed an artificial intelligence system based on Q-learning and the convolutional neural network (Mnih *et al*., 2015), which can adapt to different application requirements.

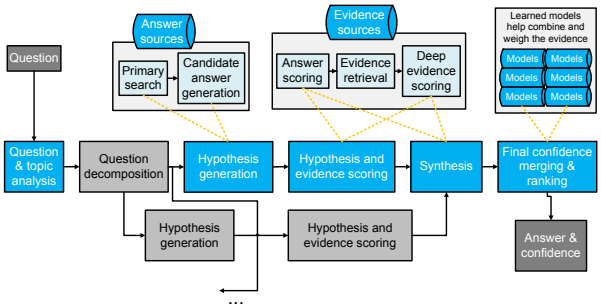

Fig. 7 The high-level architecture of IBM's DeepQA used in Watson

However, cross-media big data is naturally multimodal and cross-domain, employing sophisticated compositions, different representations, and complex correlations. Existing intelligent systems and frameworks depend heavily on the structured input and knowledge of specific domains. They cannot adapt to the characteristics of cross-media data, and cannot cope with the increasingly complex needs of general tasks (such as information retrieval) and specific tasks (such as content monitoring) in cross-media scenarios, which makes it very hard for them to realize cross-media intelligent analysis and reasoning. To address these problems, it is essential to develop an efficient cross-media intelligent engine with abilities in autonomic learning and evolution. The efficient intelligent engine would act as a bridge between technologies and applications, which could integrate cross-media uniform representation, correlation learning, knowledge evolution, reasoning, and so on. Such an engine would provide cross-media analysis and reasoning services, and be a computing platform for cross-media intelligent applications.

2.7 Cross-media intelligent applications

The advent of the artificial intelligence era and the availability of huge amounts of cross-media data have been revolutionizing the landscape in all industry sectors. Among these, cross-media web content monitoring, web information trend analysis, and healthcare data fusion and reasoning are three key applications, which if well addressed would present important models and demonstration significance to all other areas. We will briefly review the preliminary background, previous studies, as well as the existing challenges to be confronted.

iMonitor: The Internet is recognized as one of the most influential factors for the stability of human society. Many countries have built intelligent systems to monitor the content propagating or streaming over the Internet, such as the PRISM system in the US, the Tempora system in the UK, and the SORM system in Russia. At the same time, China is developing a set of web content monitoring systems, such as the Golden Shield Project for the Ministry of Public Security of China. However, existing monitoring systems work mainly in the form of passive sampling-post hoc analysis, which limits the usefulness of existing systems, and raises three challenges in the intelligent systems community, namely (1) time lag, (2) insufficient coverage, and (3) high cost, especially considering the diversity of cross-media data.

iTrend: Trend analysis of cross-media web information is the key to improve the stability of human society, by alleviating unnecessary social panic and understanding the evolution of public opinion. There are numerous existing studies on social media analysis, sentiment analysis, and news verification. For example, the Xinhua News Agency explores verification techniques on UGC data, and there is also the PHEME project in the EU, the Tian-Ji system developed by the Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS), and the TRS analysis system. However, existing systems for

cross-media trend analysis suffer from the following three main limitations: (1) They are unable to efficiently collect cross-media data; (2) They have the disadvantage of under-utilization of cross-media data; (3) The sequential characteristics of public opinion are usually ignored in the analysis. To address these challenges, trend analysis systems must be carefully designed with three additional components, namely fusion, reasoning, and decision making. An advanced framework is shown in Fig. 8.

iCare: Data-driven healthcare analytics (MIT Technology Review, 2014), based on the fusion of massive cross-media data, is reforming the experience diagnostics and evidence-based medicine (Brownson *et al.*, 1999) toward the next stage, namely personalized and precision medicine (Aamodt and Plaza, 1994). Healthcare analytics is a key technique for a wide range of real-world applications (Fig. 9).

Many IT giants have joined the healthcare analytics community; e.g., IBM released Watson Healthcare (http://spectrum.ieee.org/computing/software/ibms-watson-goes-to-med-school), Google announced DeepMind (https://deepmind.com/health), and Baidu just released Baidu Medical Brain. In spite of their usefulness in certain areas, the applicability of existing models and algorithms (Kumar *et al.*, 2012; Chen Y *et al.*, 2013; Yuan *et al.*, 2014) is limited due to (1) inability to perform cross-media fusion and analysis (Chen *et al.*, 2007), (2) lack of supervision from domain experts (Chen Y *et al.*, 2013), and (3) poor adaptability toward different medical paradigms.

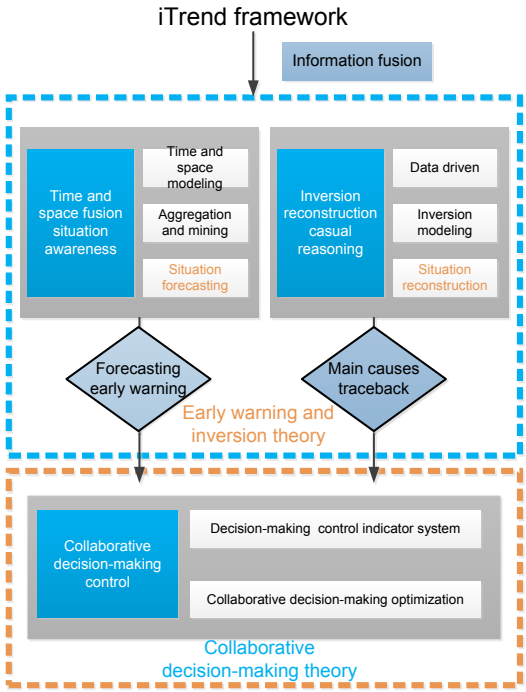

Fig. 8 iTrend framework

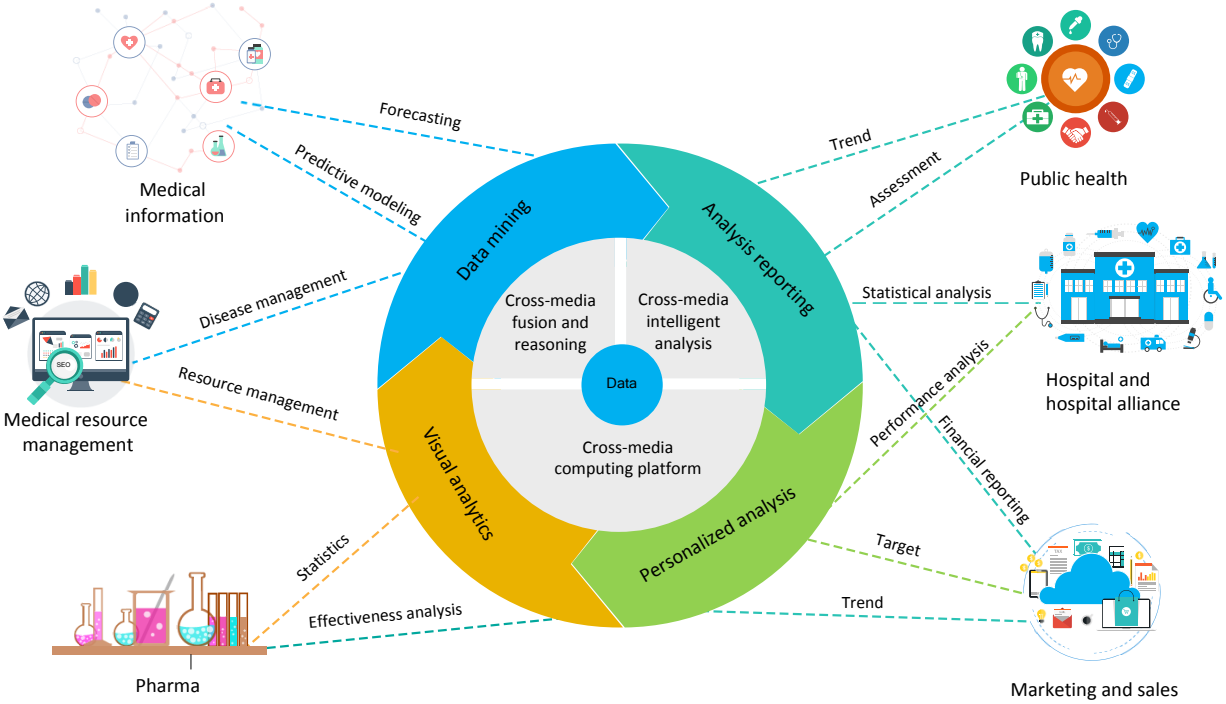

Fig. 9 Existing applications in healthcare analytics

3 Conclusions

In this paper, we have presented an overview of cross-media analysis and reasoning. The advances achieved by existing studies, as well as the major challenges and open issues, have been shown in the overview. From the seven parts of this paper, it can be seen that cross-media analysis and reasoning has been a key problem of research, and has wide prospects for application. The introduction and discussion in this paper are expected to attract more research interest to this area, and provide insights for researchers on the relevant topics, so as to inspire future research in cross-media analysis and reasoning.

Acknowledgements

The authors would like to thank Peng CUI, Shi-kui WEI, Ji-tao SANG, Shu-hui WANG, Jing LIU, and Bu-yue QIAN for their valuable discussions and assistance.

References

Aamodt, A., Plaza, E., 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Commun.*, **7**(1):39-59. http://dx.doi.org/10.3233/AIC-1994-7104

Adib, F., Hsu, C.Y., Mao, H., *et al*., 2015. Capturing the human figure through a wall. *ACM Trans. Graph.*, **34**(6):219. http://dx.doi.org/10.1145/2816795.2818072

Andrew, G., Arora, R., Bilmes, J., *et al*., 2013. Deep canonical correlation analysis. Int. Conf. on Machine Learning, p.1247-1255.

Antenucci, D., Li, E., Liu, S., *et al*., 2013. Ringtail: a generalized nowcasting system. *Proc. VLDB Endow.*, **6**(12): 1358-1361. http://dx.doi.org/10.14778/2536274.2536315

Antol, S., Agrawal, A., Lu, J., *et al*., 2015. VQA: visual question answering. IEEE Int. Conf. on Computer Vision, p.2425-2433. http://dx.doi.org/10.1109/ICCV.2015.279

Babenko, A., Slesarev, A., Chigorin, A., *et al*., 2014. Neural codes for image retrieval. European Conf. on Computer Vision, p.584-599. http://dx.doi.org/10.1007/978-3-319-10590-1_38

Brownson, R.C., Gurney, J.G., Land, G.H., 1999. Evidence-based decision making in public health. *J. Publ. Health Manag. Pract.*, **5**(5):86-97. http://dx.doi.org/10.1097/00124784-199909000-00012

Carlson, C., Betteridge, J., Kisiel, B., *et al*., 2010. Towards an architecture for never-ending language learning. AAAI Conf. on Artificial Intelligence, p.1306-1313.

Chen, D.P., Weber, S.C., Constantinou, P.S., *et al*., 2007. Clinical arrays of laboratory measures, or "clinarrays", built from an electronic health record enable disease subtyping by severity. AMIA Annual Symp. Proc., p.115-119.

Chen, X., Shrivastava, A., Gupta, A., 2013. NEIL: extracting visual knowledge from web data. IEEE Int. Conf. on Computer Vision, p.1409-1416. http://dx.doi.org/10.1109/ICCV.2013.178

Chen, Y., Carroll, R.J., Hinz, E.R.M., *et al*., 2013. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J. Am. Med. Inform. Assoc.*, **20**(e2):253-259. http://dx.doi.org/10.1136/amiajnl-2013-001945

Cilibrasi, R.L., Vitanyi, P.M.B., 2007. The Google similarity distance. *IEEE Trans. Knowl. Data Eng.*, **19**(3):370-383. http://dx.doi.org/10.1109/TKDE.2007.48

Culotta, A., 2014. Estimating county health statistics with twitter. ACM Conf. on Human Factors in Computing Systems, p.1335-1344. http://dx.doi.org/10.1145/2556288.2557139

Daras, P., Manolopoulou, S., Axenopoulos, A., 2012. Search and retrieval of rich media objects supporting multiple multimodal queries. *IEEE Trans. Multim.*, **14**(3):734-746. http://dx.doi.org/10.1109/TMM.2011.2181343

Davenport, T.H., Prusak, L., 1998. Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press, Boston, p.5.

Deng, J., Dong, W., Socher, R., *et al*., 2009. ImageNet: a large-scale hierarchical image database. IEEE Conf. on Computer Vision and Pattern Recognition, p.248-255. http://dx.doi.org/10.1109/CVPR.2009.5206848

Dong, X., Gabrilovich, E., Heitz, G., *et al*., 2014. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.601-610. http://dx.doi.org/10.1145/2623330.2623623

Fang, Q., Xu, C., Sang, J., *et al*., 2016. Folksonomy-based visual ontology construction and its applications. *IEEE Trans. Multim.*, **18**(4):702-713. http://dx.doi.org/10.1109/TMM.2016.2527602

Fellbaum, C., Miller, G., 1998. WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA.

Feng, F., Wang, X., Li, R., 2014. Cross-modal retrieval with correspondence autoencoder. ACM Int. Conf. on Multimedia, p.7-16. http://dx.doi.org/10.1145/2647868.2654902

Ferrucci, D., Levas, A., Bagchi, S., *et al.*, 2013. Watson: beyond jeopardy! *Artif. Intell.*, **199-200**:93-105. http://dx.doi.org/10.1016/j.artint.2012.06.009

Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M., 2015. Visual simultaneous localization and mapping: a survey. *Artif. Intell. Rev.*, **43**(1):55-81. http://dx.doi.org/10.1007/s10462-012-9365-8

Garfield, E., 2004. Historiographic mapping of knowledge domains literature. *J. Inform. Sci.*, **30**(2):119-145. http://dx.doi.org/10.1177/0165551504042802

Gibney, E., 2015. DeepMind algorithm beats people at classic video games. *Nature*, **518**(7540):465-466.

Ginsberg, J., Mohebbi, M., Patel, R.S., *et al*., 2009. Detecting influenza epidemics using search engine query data. *Nature*, **457**(7232):1012-1014.

Gong, Y., Ke, Q., Isard, M., *et al*., 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vis.*, **106**(2):210-233. http://dx.doi.org/10.1007/s11263-013-0658-4

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neur. Comput.*, **9**(8):1735-1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

Hodosh, M., Young, P., Hockenmaier, J., 2013. Framing image description as a ranking task: data, models and evaluation metrics. *J. Artif. Intell. Res.*, **47**(1):853-899.

Hotelling, H., 1936. Relations between two sets of variates. *Biometrika*, **28**(3-4):321-377. https://doi.org/10.1093/biomet/28.3-4.321

Hsu, F., 2002. Behind Deep Blue: Building the Computer that Defeated the World Chess Champion. Princeton University Press, Princeton, USA.

Hua, Y., Wang, S., Liu, S., *et al*., 2014. TINA: cross-modal correlation learning by adaptive hierarchical semantic aggregation. IEEE Int. Conf. on Data Mining, p.190-199. http://dx.doi.org/10.1109/ICDM.2014.65

Jia, X., Gavves, E., Fernando, B., *et al*., 2015. Guiding long-short term memory for image caption generation. arXiv:1509.04942.

Johnson, J., Krishna, R., Stark, M., *et al*., 2015. Image retrieval using scene graphs. IEEE Conf. on Computer Vision and Pattern Recognition, p.3668-3678. http://dx.doi.org/10.1109/CVPR.2015.7298990

Karpathy, A., Li, F.F., 2015. Deep visual-semantic alignments for generating image descriptions. IEEE Conf. on Computer Vision and Pattern Recognition, p.3128-3137. http://dx.doi.org/10.1109/CVPR.2015.7298932

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet: classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, p.1097-1105.

Kulkarni, G., Premraj, V., Dhar, S., *et al*., 2011. Baby talk: understanding and generating simple image descriptions. IEEE Conf. on Computer Vision and Pattern Recognition, p.1601-1608. http://dx.doi.org/10.1109/CVPR.2011.5995466

Kumar, S., Sanderford, M., Gray, V.E., *et al*., 2012. Evolutionary diagnosis method for variants in personal exomes. *Nat. Meth.*, **9**(9):855-856. http://dx.doi.org/10.1038/nmeth.2147

Kuznetsova, P., Ordonezz, V., Berg, T.L., *et al*., 2014. TREETALK: composition and compression of trees for image descriptions. *Trans. Assoc. Comput. Ling.*, **2**:351-362.

Lazaric, A., 2012. Transfer in reinforcement learning: a framework and a survey. *In*: Wiering, M., van Otterlo, M. (Eds.), Reinforcement Learning: State-of-the-Art. Springer Berlin Heidelberg, Berlin, p.143-173. http://dx.doi.org/10.1007/978-3-642-27645-3_5

Lazer, D., Kennedy, R., King, G., *et al*., 2014. The parable of Google flu: traps in big data analysis. *Science*, **343**(6176):1203-1205. http://dx.doi.org/10.1126/science.1248506

Lew, M.S., Sebe, N., Djeraba, C., *et al*., 2006. Content-based multimedia information retrieval: state of the art and challenges. *ACM Trans. Multim. Comput. Commun. Appl.*, **2**(1):1-19. http://dx.doi.org/10.1145/1126004.1126005

Lin, T., Pantel, P., Gamon, M., *et al*., 2012. Active objects: actions for entity-centric search. ACM Int. Conf. on World Wide Web, p.589-598. http://dx.doi.org/10.1145/2187836.2187916

Luo, G., Tang, C., 2008. On iterative intelligent medical search. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.3-10. http://dx.doi.org/10.1145/1390334.1390338

Mao, X., Lin, B., Cai, D., *et al*., 2013. Parallel field alignment for cross media retrieval. ACM Int. Conf. on Multimedia, p.897-906. http://dx.doi.org/10.1145/2502081.2502087

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, **264**(5588):746-748. http://dx.doi.org/10.1038/264746a0

MIT Technology Review, 2014. Data driven healthcare. https://www.technologyreview.com/business-report/data-driven-health-care/free [Dec. 06, 2016].

Mnih, V., Kavukcuoglu, K., Silver, D., 2015. Human-level control through deep reinforcement learning. *Nature*, **518**(7540):529-333. http://dx.doi.org/10.1038/nature14236

Ngiam, J., Khosla, A., Kim, M., *et al*., 2011. Multimodal deep learning. Int. Conf. on Machine Learning, p.689-696.

Ordonez, V., Kulkarni, G., Berg, T.L., 2011. Im2text: describing images using 1 million captioned photographs. Advances in Neural Information Processing Systems, p.1143-1151.

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

Pearl, J., 2000. Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge, UK.

Peng, Y., Huang, X., Qi, J., 2016a. Cross-media shared representation by hierarchical learning with multiple deep networks. Int. Joint Conf. on Artificial Intelligence, p.3846-3853.

Peng, Y., Zhai, X., Zhao, Y., *et al*., 2016b. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Trans. Circ. Syst. Video Technol.*, **26**(3):583-596. http://dx.doi.org/10.1109/TCSVT.2015.2400779

Prabhu, N., Babu, R.V., 2015. Attribute-Graph: a graph based approach to image ranking. IEEE Int. Conf. on Computer Vision, p.1071-1079. http://dx.doi.org/10.1109/ICCV.2015.128

Radinsky, K., Davidovich, S., Markovitch, S., 2012. Learning causality for news events prediction. Int. Conf. on World Wide Web, p.909-918. http://dx.doi.org/10.1145/2187836.2187958

Rasiwasia, N., Costa Pereira, J., Coviello, E., *et al*., 2010. A new approach to cross-modal multimedia retrieval. ACM Int. Conf. on Multimedia, p.251-260. http://dx.doi.org/10.1145/1873951.1873987

Rasiwasia, N., Mahajan, D., Mahadevan, V., *et al*., 2014. Cluster canonical correlation analysis. Int. Conf. on Artificial Intelligence and Statistics, p.823-831.

Rautaray, S.S., Agrawal, A., 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artif. Intell. Rev.*, **43**(1):1-54. http://dx.doi.org/10.1007/s10462-012-9356-9

Roller, S., Schulte im Walde, S., 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. Conf. on Empirical Methods in Natural Language Processing, p.1146-1157.

Sadeghi, F., Divvala, S.K., Farhadi, A., 2015. VisKE: visual knowledge extraction and question answering by visual verification of relation phrases. IEEE Conf. on Computer Vision and Pattern Recognition, p.1456-1464. http://dx.doi.org/10.1109/CVPR.2015.7298752

Singhal, A., 2012. Introducing the knowledge graph: things, not strings. Official Blog of Google.

Socher, R., Lin, C., Ng, A.Y., *et al*., 2011. Parsing natural scenes and natural language with recursive neural networks. Int. Conf. on Machine Learning, p.129-136.

Socher, R., Karpathy, A., Le, Q., *et al*., 2014. Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Ling.*, **2**:207-218.

Srivastava, N., Salakhutdinov, R., 2012. Multimodal learning with deep Boltzmann machines. Advances in Neural Information Processing Systems, p.2222-2230.

Suchanek, F., Weikum, G., 2014. Knowledge bases in the age of big data analytics. *Proc. VLDB Endow.*, **7**(13):1713-1714. http://dx.doi.org/10.14778/2733004.2733069

Uyar, A., Aliyu, F.M., 2015. Evaluating search features of Google Knowledge Graph and Bing Satori: entity types, list searches and query interfaces. *Onl. Inform. Rev.*, **39**(2):197-213. http://dx.doi.org/10.1108/OIR-10-2014-0257

Vinyals, O., Toshev, A., Bengio, S., *et al*., 2015. Show and tell: a neural image caption generator. IEEE Conf. on Computer Vision and Pattern Recognition, p.3156-3164. http://dx.doi.org/10.1109/CVPR.2015.7298935

Wang, D., Cui, P., Ou, M., *et al*., 2015. Learning compact hash codes for multimodal representations using orthogonal deep structure. *IEEE Trans. Multim.*, **17**(9): 1404-1416. http://dx.doi.org/10.1109/TMM.2015.2455415

Wang, W., Ooi, B.C., Yang, X., *et al*., 2014. Effective multimodal retrieval based on stacked auto-encoders. *Proc. VLDB Endow.*, **7**(8):649-660. http://dx.doi.org/10.14778/2732296.2732301

Wang, Y., Wu, F., Song, J., *et al*., 2014. Multi-modal mutual topic reinforce modeling for cross-media retrieval. ACM Int. Conf. on Multimedia, p.307-316. http://dx.doi.org/10.1145/2647868.2654901

Wei, Y., Zhao, Y., Lu, C., *et al*., 2017. Cross-modal retrieval with CNN visual features: a new baseline. *IEEE Trans. Cybern.*, **47**(2):449-460. http://dx.doi.org/10.1109/TCYB.2016.2519449

Wu, W., Xu, J., Li, H., 2010. Learning similarity function between objects in heterogeneous spaces. Technique Report MSR-TR-2010-86, Microsoft.

Xu, K., Ba, J., Kiros, R., *et al*., 2015. Show, attend and tell: neural image caption generation with visual attention. Int. Conf. on Machine Learning, p.2048-2057.

Yang, Y., Zhuang, Y., Wu, F., *et al*., 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Trans. Multim.*, **10**(3):437-446. http://dx.doi.org/10.1109/TMM.2008.917359

Yang, Y., Teo, C.L., Daume, H., *et al*., 2011. Corpus-guided sentence generation of natural images. Conf. on Empirical Methods in Natural Language Processing, p.444-454.

Yang, Y., Nie, F., Xu, D., *et al*., 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. Patt. Anal. Mach. Intell.*, **34**(4):723-742. http://dx.doi.org/10.1109/TPAMI.2011.170

Yuan, L., Pan, C., Ji, S., *et al*., 2014. Automated annotation of developmental stages of Drosophila embryos in images containing spatial patterns of expression. *Bioinformatics*, **30**(2):266-273. http://dx.doi.org/10.1093/bioinformatics/btt648

Zhai, X., Peng, Y., Xiao, J., 2014. Learning cross-media joint representation with sparse and semi-supervised regularization. *IEEE Trans. Circ. Syst. Video Technol.*, **24**(6):965-978. http://dx.doi.org/10.1109/TCSVT.2013.2276704

Zhang, H., Yang, Y., Luan, H., *et al*., 2014a. Start from scratch: towards automatically identifying, modeling, and naming visual attributes. ACM Int. Conf. on Multimedia, p.187-196. http://dx.doi.org/10.1145/2647868.2654915

Zhang, H., Yuan, J., Gao, X., *et al*., 2014b. Boosting cross-media retrieval via visual-auditory feature analysis and relevance feedback. ACM Int. Conf. on Multimedia, p.953-956. http://dx.doi.org/10.1145/2647868.2654975

Zhang, H., Shang, X., Luan, H., *et al*., 2016. Learning from collective intelligence: feature learning using social images and tags. *ACM Trans. Multim. Comput. Commun. Appl.*, **13**(1):1. http://dx.doi.org/10.1145/2978656

Zhang, J., Wang, S., Huang, Q., 2015. Location-based parallel tag completion for geo-tagged social image retrieval. ACM Int. Conf. on Multimedia Retrieval, p.355-362.

Zhu, Y., Zhang, C., Ré, C., *et al*., 2015. Building a large-scale multimodal knowledge base system for answering visual queries. arXiv:1507.05670.