Deep Learning for Information Retrieval

Hang Li
Noah's Ark Lab, Huawei Technologies
Hong Kong Science Park
Shatin, Hong Kong
hangli.hl@huawei.com

Zhengdong Lu
Noah's Ark Lab, Huawei Technologies
Hong Kong Science Park
Shatin, Hong Kong
zhengdong.lu@huawei.com

ABSTRACT

Recent years have observed a significant progress in information retrieval and natural language processing with deep learning technologies being successfully applied into almost all of their major tasks. The key to the success of deep learning is its capability of accurately learning distributed representations (vector representations or structured arrangement of them) of natural language expressions such as sentences, and effectively utilizing the representations in the tasks. This tutorial aims at summarizing and introducing the results of recent research on deep learning for information retrieval, in order to stimulate and foster more significant research and development work on the topic in the future.

The tutorial mainly consists of three parts. In the first part, we introduce the fundamental techniques of deep learning for natural language processing and information retrieval, such as word embedding, recurrent neural networks, and convolutional neural networks. In the second part, we explain how deep learning, particularly representation learning techniques, can be utilized in fundamental NLP and IR problems, including matching, translation, classification, and structured prediction. In the third part, we describe how deep learning can be used in specific application tasks in details. The tasks are search, question answering (from either documents, database, or knowledge base), and image retrieval.

Keywords

Deep Learning, Information Retrieval, Search, Question Answering, Image Retrieval

1. INTRODUCTION

"Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources" -Wikipedia. Figure 1 gives an overview of the major tasks in IR, including search, question answering (from either documents, database, or knowledge base), and image retrieval. The user expresses her information need (intent) in keywords or a question and submits it to the information retrieval system; the system accesses the information and/or knowledge base (content), finds relevant result, and returns to the user. The key questions are how to represent the intent & content and how to conduct matching between the intent & content.

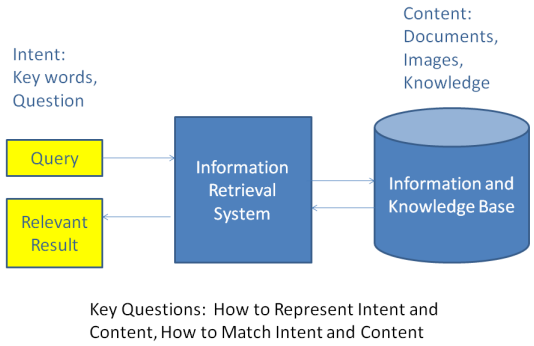

Figure 1: An overview of information retrieval.

The easiest task in IR would be document retrieval. In the traditional IR approach, the intent is represented as key words in a query, and the content consists of documents and each document is represented as a bag of words. Cosine similarity is calculated between the tf-idf vectors of the query and each document, and is viewed as the matching (relevance) score between them.

The traditional approach works well to some extent, but it also suffers from the term mismatch drawback. In fact, semantic matching between query and document is necessary in order to help the user to effectively find relevant information [21]. Deep learning techniques have been successfully employed to conduct semantic matching in web search, and significant improvement in relevance have been observed (e.g., [11, 31, 29]).

There are harder tasks in IR, such as question answering from documents, question answering from relational database, question answering from knowledge base, and image retrieval. The challenge is that the representations in the tasks are in different forms, for example, unstructured data, structured data, multimedia data, or mix of them, and it is very diffi-

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914800

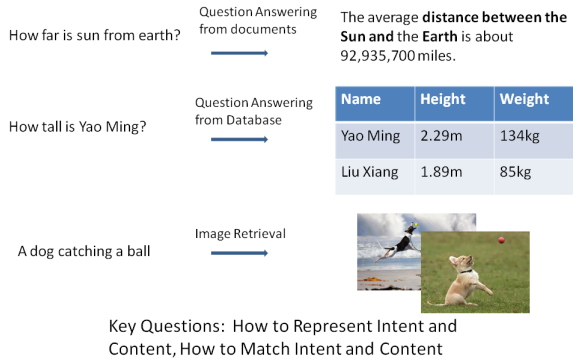

Figure 2: Hard problems in information retrieval and natural language processing.

cult or even impossible to perform direct matching between the intent and content in the tasks. Although some methods had been proposed for question answering, image retrieval, etc, they tended to be ad-hoc, and their performances were not satisfactory. Recently, significant progresses have been made in solving the hard problems in IR, with deep learning as the major machinery.

Deep learning is powerful, because it can help automatically learn representations of different data in different tasks [18]. The learned representations are all in the same form, namely real valued vectors, also referred to as distributed representations. In this way, the matching in IR can be conducted through the vector representations, and therefore the performance of some IR tasks can be significantly enhanced and the other tasks which were previously considered impossible can be successfully carried out.

For example, it is found possible to directly learn representations from images and their associated texts, exploit the representations in matching between questions and images, and achieve high accuracy in image retrieval (e.g., [15, 14, 17, 24, 36, 23]). Deep learning can also be employed to carry out question answering from database or knowledge base. Recent work shows that given only question answer pairs, a relational database or knowledge base, as well as the 'grounding' relations between the answers and the database or knowledge base, one can learn a deep neural network to automatically conduct question answering from the database or knowledge base. No human effort is needed in construction of a semantic parser for analyzing the questions and the performance can be even higher (e.g., [4, 28, 38, 41, 1, 43, 42]). It is also found that deep learning can improve traditional question answering from documents, which can be formalized as a problem of matching between two sentences (question and answer) each having a complicated syntactic and semantic structure. The advantage of the approach is that usually one does not need to use linguistic knowledge to build the system (e.g., [32, 22, 10, 37]). Another new and interesting finding is that with deep learning and a large amount of question answer pair data one can build a system which automatically *generates* an answer given a question in question answering, or in general in single turn dialogue (e.g.,[30, 35]). This was considered as a very hard problem.

Indeed, deep learning has opened many new opportunities to IR. There are a large number of new problems which one can try, and a large number of existing problems on which we can make improvement with deep learning. Certainly, there are also many challenges. The biggest question is how to combine neural computing (or deep learning) with traditional symbolic processing, both appears necessary for IR.

The objectives of the tutorial are as the following. First, to help the students and researchers who want to make further investigations on deep learning for IR to enhance the state of the art of the field; second, to help the practitioners who want to apply the existing deep learning for IR technologies into real world problems.

The tutorial is set at an intermediate level. It is assumed that the attendees have certain knowledge on machine learning and information retrieval. It is not a requirement, however, that the attendees know much about deep learning.

Several wonderful tutorials were given at related conferences, for example, Richard Socher, Chris Manning, Yoshua Bengio, Deep Learning for NLP, at ACL 2012 and NAACL 2013; Kevin Duh, Deep Learning for Natural Language Processing and Machine Translation, at CWMT 2014; Xiaodong He, Jianfeng Gao, and Li Deng, Deep Learning for Natural Language Processing: Theory and Practice, at CIKM 2014; Wen-tau Yih, Xiaodong He, and Jianfeng Gao, Deep Learning and Continuous Representations for NLP, at NAACL-HLT 2015. This tutorial is significantly different from the previous tutorials in the sense that it is organized from the viewpoint of IR.

2. OUTLINE OF TUTORIAL

The outline of the tutorial is as follows.

1. Basics of Deep Learning (50 minutes)
   - Word Embedding
   - Recurrent Neural Network
   - Convolutional Neural Network
   - Training of Models
2. Fundamental Problems in Deep Learning for IR and NLP (50 minutes)
   - Learning of Distributed Representation
   - Matching
   - Translation
   - Classification
   - Structured Prediction
3. Applications of Deep Learning to IR (50 minutes)
   - Search
   - Question Answering (from Documents)
     - Retrieval-based Question Answering
     - Generation-based Question Answering
   - Question Answering from Relational Database
   - Question Answering from Knowledge Base
   - Image Retrieval
4. Future Directions of Deep Learning for IR (10 minutes)

3. CONTENT OF TUTORIAL

The tutorial consists of four parts. We introduce the content of each part.

3.1 Basics of Deep Learning

We start with introducing the basic tools in deep learning for information retrieval and natural language processing, including word embedding [25, 27, 19, 20], recurrent neural network (RNN) [26, 9, 6], convolutional neural network (CNN) [7, 10, 13, 31], as well as training of deep neural network models.

3.2 Fundamental Problems in Deep Learning for IR and NLP

We formalize many tasks in IR and NLP into a number of fundamental problems including matching, translation, classification, and structured prediction, described below.

- Matching: matching two strings
  $s, t \rightarrow R^+$
- Translation: transforming one string to another
  $s \rightarrow t$
- Classification: assigning a label to a string
  $s \rightarrow c$
- Structured Prediction: mapping a string to its structure
  $s \rightarrow s'$

We first point out when we apply deep learning to the problems, we in fact learn representations of natural language in the problems. The learned representations can be used in realizing the tasks, with often enhanced performance. We explain methods that can be used for learning the representations in matching [22, 10, 37], translation [33, 6, 2, 8], classification [13, 16, 44], and structured prediction [7, 34, 5].

3.3 Applications of Deep Learning to IR and NLP

We introduce the recent work on applications of deep learning to IR tasks. First, we describe deep learning models which have been successfully applied to search to enhance relevance [11, 31, 29], as extensions of conventional linear models [3, 40, 39].

We also talk about deep learning for question answering including the retrieval-based setting in which answers are retrieved and returned from a large repository of question answer pairs [32, 22, 10, 12, 37], as well as the generation-based setting in which answers are automatically generated from a system which is trained with a large number of question answer pairs [30, 35].

We explain the work about question answering from database or knowledge base using deep learning in which only question answer pairs and the database or knowledge base are used in construction of the system [4, 28, 38, 41, 1, 43, 42]

We introduce the recent progress in image retrieval using deep learning in which only images and their associated texts (questions) are used as training data [15, 14, 17, 36, 24, 23].

3.4 Future Directions of Deep Learning for IR

We conclude the tutorial by summarizing the major challenges and opportunities in deep learning for information retrieval.

4. ACKNOWLEDGEMENT

The work is partially supported by China National 973 project 2014CB340301. The authors are grateful to Xin Jiang, Xi Zhang, Jun Xu, Shengxian Wan, and Liang Peng for providing related materials for this tutorial.

5. REFERENCES

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv:1601.01705*, 2016.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR'15*, 2015.

[3] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger. Supervised semantic indexing. In *Proceedings of CIKM'09*, 2009.

[4] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale simple question answering with memory networks. *arXiv:1506.02075*, 2015.

[5] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP'14*, 2014.

[6] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP'14*, 2014.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[8] J. Gu, Z. Lu, H. Li, and V. O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv:1603.06393*, 2016.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Proccedings of NIPS'14*, 2014.

[11] P.-S. Huang, X. He, and J. Gao. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of CIKM'13*, 2013.

[12] Z. Ji, Z. Lu, and H. Li. An information retrieval approach to short text conversation. *arXiv:1408.6988*, 2014.

[13] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of ACL'14*, 2014.

[14] A. Karpathy, A. Joulin, and F. Li. Deep fragment embeddings for didirectional image sentence mapping. In *Proceedings of NIPS'14*, 2014.

[15] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of CVPR'15*, 2015.

[16] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP'14*, 2014.

[17] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539*, 2014.

[18] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[19] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS'14*, 2014.

[20] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of Association for Computational Linguistics*, 3:211–225, 2015.

[21] H. Li and J. Xu. Semantic matching in search. *Foundations and Trends's in Information Retrieval*, 7(5):343–469, 2014.

[22] Z. Lu and H. Li. A deep architecture for matching short texts. In *Proceedings of NIPS'13*, 2013.

[23] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of ICCV'15*, 2015.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632*, 2014.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[26] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Proceedings of InterSpeech'10*, 2010.

[27] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS'13*, 2013.

[28] A. Neelakantan, Q. V. Le, and I. Sutskever. Neural programmer: Inducing latent programs with gradient descent. *arXiv:1511.04834*, 2015.

[29] A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of SIGIR'15*, 2015.

[30] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *Proceedings of ACL'15*, 2015.

[31] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of CIKM'14*, 2014.

[32] R. Socher, E. H. Huang, and A. Y. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NIPS'11*, 2011.

[33] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of NIPS'14*, 2014.

[34] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. Grammar as a foreign language. *arXiv:1412.7449*, 2014.

[35] O. Vinyals and Q. V. Le. A neural conversational model. *arXiv:1506.05869*, 2015.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: a neural image caption generator. *arXiv:1411.4555*, 2014.

[37] M. Wang, Z. Lu, H. Li, and Q. Liu. Syntax-based deep matching of short texts. In *Proceedings of IJCAI'15*, 2015.

[38] J. Weston, S. Chopra, and A. Bordes. Memory networks. In *Proceedings of ICLR'15*, 2015.

[39] W. Wu, H. Li, and J. Xu. Learning query and document similarities from click-through bipartite graph with metadata. In *Proceedings of WSDM'13*, 2013.

[40] W. Wu, Z. Lu, and H. Li. Learning bilinear model for matching queries and documents. *The Journal of Machine Learning Research*, 14(1):2519–2548, 2013.

[41] W. Yih, M. Chang, X. He, and J. Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of ACL'15*, 2015.

[42] J. Yin, X. Jiang, L. Shang, Z. Lu, H. Li, and X. Li. Neural generative question answering. In *Proceedings of IJCAI'16*, 2016.

[43] P. Yin, Z. Lu, H. Li, and B. Kao. Neural enquirer: Learning to query tables. In *Proceedings of IJCAI'16*, 2016.

[44] H. Zhao, Z. Lu, and P. Poupart. Self-adaptive hierarchical sentence model. In *Proceedings of IJCAI'15*, 2015.