Adaptability of Neural Networks on Varying Granularity IR Tasks

Daniel Cohen, Qingyao Ai, W. Bruce Croft
Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst, MA
{dcohen, aiqy, croft}@cs.umass.edu

ABSTRACT

Recent work in Information Retrieval (IR) using Deep Learning models has yielded state of the art results on a variety of IR tasks. Deep neural networks (DNN) are capable of learning ideal representations of data during the training process, removing the need for independently extracting features. However, the structures of these DNNs are often tailored to perform on specific datasets. In addition, IR tasks deal with text at varying levels of granularity from single factoids to documents containing thousands of words. In this paper, we examine the role of the granularity on the performance of common state of the art DNN structures in IR.

CCS Concepts

•Information systems → Retrieval models and ranking; *Question answering;* Document structure; **•Computing methodologies → Neural networks;**

Keywords

deep learning; Question Answering; ad-hoc retrieval

1. INTRODUCTION

Learning effective representations of data is a critical component of any system that ranks documents. Conventional approaches rely on transforming text into vectors consisting of lexical, semantic, and syntactic features that capture the information contained in text. This conversion depends on domain knowledge and is an independent step from the optimization process of the ranking method. As this process is separate from the loss function, potential information can be lost that negatively affects performance. Deep learning has been shown to learn internal representations directly from the text in natural language processing and specific IR tasks that yield state of the art performance. However, the deep learning models used for these IR tasks are often tailored for the individual task with the network structure making some assumption about the data, and little work has been done in examining how well these networks can adapt to collections with varying levels of granularity.

IR focuses on retrieval of information at differing levels of granularity whether at the single word level in the factoid task such as TREC QA, passage level involving community question and answer, or the document level dealing with ad-hoc retrieval. Each of these levels present unique challenges when fitting a model, and we show that DNNs are not exempt from this problem. In the following sections we examine state of the art DNNs on varying levels of granularity to demonstrate the efficacy of different neural structures at each level of granularity.

2. RELATED WORK

At each level of granularity, significant improvements have been made by introducing various DNN structures. Convolutional neural networks (CNN) have been used at various layers in the neural net, at the input level over word embeddings demonstrated by Severyn and Moschitti [8], as an intermediary layer within a feedforward network introduced by Feng et al. [2], or as a penultimate stage on top of a recurrent neural network (RNN) to provide more composite representations over the question and answer text by Tan et al. [11]. Regardless of the position of the convolutional layer, the motivation behind implementing a convolutional layer was to extract the most salient features from the input to allow easier similarity comparisons.

As language is sequential in nature, RNNs have been shown to work extremely well for IR tasks. Wang and Nyberg [13, 14] show that using a bidirectional Long Short-Term Memory (BiLSTM) network over query-answer pairs to determine relevance is an effective approach to the fine grain level of the TREC QA task as well as for passage level retrieval. Providing additional insight to how LSTM networks process text, Palangi et al. [7] demonstrate the use of a weakly supervised LSTM network to detemine answer sentence similarity and examine how individual cells attenuate information when processing query-answer pairs.

Another important attempt that applies DNN for IR tasks is the Deep Structured Semantic Model (DSSM) and its variations. Introduced by Huang et al. [4], DSSM uses a word hashing technique to project varied length text into fixed length vectors as the model's input and constructs a feed-

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Neu-IR '16 SIGIR Workshop on Neural Information Retrieval July 17–21, 2016, Pisa, Italy

© 2016 Copyright held by the owner/author(s).

Method	MRR	P@1
LSTM	0.6314	0.7849
CNN	0.3729	0.6225

Table 1: Comparison of a CNN and LSTM network after hyper-parameter tuning over Yahoo's Webscope L4 collection.

forward neural network above it. The relevance between documents and queries is measured by the cosine similarity between their output vectors. Recently, Shen et al. [9] proposed a convolution network (CLSM) and Palangi et al. [6] proposed a RNN-LSTM model with the same word hashing technique of DSSM. They showed positive results when applying these models on ad-hoc retrieval tasks with web page titles collections. However, the effectiveness of word hashing and DSSM models vary considerably as text length changes. Their performance on standard TREC collections are still poor according to our experiments.

3. GRANULARITY TASKS

We examine the efficacy of deep learning on three distinct levels of granularity. First, at the fine grain level, retrieval focuses on a specific word or deals with a short sentence of text containing the relevant information. Second, at the medium granularity level, the information need of the query can no longer be addressed by a single sentence, and often requires multiple sentences to be relevant. Third, we address the coarse grain level, which we view as full document retrieval commonly found on ad-hoc retrieval tasks.

3.1 Fine Granularity

The focus of this section is on the TREC QA task. In this task, the length of individual documents are often no more than a single sentence, and queries consist of short questions such as *"When did James Dean die?"* or *"What is crips' gang color"*. The relevant information in each document is one or two words that directly address the information need of the query.

From the deep learning perspective, CNNs adapt well to the fine grained task as they are able to identify key aspects of an input matrix. This ability has resulted in these networks receiving widespread use in the computer vision task. The same principle can be applied to the sentence level by allowing convolutional layers to extract the most salient information over embeddings of a sentence. This approach has been used for semantic sentence level matching by Hu et al. [3]. Severyn and Moschitti [8] also take advantage of the matching ability of CNNs by implementing a convolutional layer to extract the most salient features between answer and query sentences to compute similarity scores for ranking.

An interesting note is the performance of RNNs at the same granularity level. As shown in Yin et al. [15] and Santos et al. [1], conventional CNNs often outperform equivalent LSTM networks at this level of granularity as filter lengths are able to capture the language dependencies and match keywords when the candidate answer sentences are short.

3.2 Medium Granularity

The medium granularity level, consisting of passages, contrasts sharply with the granularity of the previous section. Instead of identifying specific words contained within a sentence, the passage task deals with information related to the query that can span multiple sentences. However, relevance is not determined solely by topical similarity between document and query. Text in relevant passages can have little term overlap with the query, and conventional IR methods such as BM25 have reflected this in their performance.

Due to the span of relevant information across the length of candidate answers, LSTM networks are uniquely suited to this task as they are able to model syntactic and semantic dependencies across positions in a sequence and focus less on matching than the CNN does. We demonstrate this on Yahoo's Webscope L4 CQA collection [10] of "manner" type questions, where a LSTM model built in a similar fashion to [13] significantly outperforms an equivalent CNN network as shown in Table 1. The purpose of this test was to demonstrate the ability of the two network structures to retain information across long sequences, therefore the candidate answer pool for each query consisted of 10 randomly sampled answers from the collection. These candidate answers are significantly longer than those found in WikiQA [2] or the TREC QA task with a mean length of 75. The filter lengths of CNNs are unable to capture long term dependencies that span multiple sentences, which results in its poor performance relative to the LSTM network.

Palangi et al. [7] investigate the internal representation of text within a LSTM network. Internal cell states accumulate semantic information across sequences, and their corresponding inputs learn to respond to semantically related words specific to each cell. In addition, the LSTM network is capable of keyword recognition to directly match query and document similarities. This contrasts with a standard RNN without LSTM cells, where the length of the passage task results in the internal representation 'forgetting' previous information due to the vanishing gradient problem.

3.3 Coarse Granularity

Tasks with coarse granularity including Ad-hoc retrieval are usually concerned with collections of text with great variation in length. Although the queries tend to be shorter, the documents range from tens of words to thousands of terms. Accompanying the challenge that length variation poses, the concept of relevance varies from document to document as the relevant portion of a document might range from a few sentences to its entirety. These two unique properties of coarse granularity collections result in different challenges for DNNs which are not apparent at other granularity levels.

We applied conventional networks discussed in Section 3.2 which performed well on passage length answers, but were unable to perform better than random over the Robust04 collection and thus require a different approach.

Varied input length. In ad-hoc retrieval tasks, the length differences in documents are so large that they significantly affect the training of deep models. Most neural models include a step that converts varied length input into fixed length vectors (i.e. input layer for DNN, pooling in CNN and memory vectors in RNN). Without accounting for the original length of text, this process could introduce strong bias for short or long documents. For example, to train a deep structure semantic model (DSSM) for ad-hoc retrieval, Huang et al. [4] proposed a word hashing technique that aggregates n-grams of terms to produce a fixed length repre-

Method	MAP	nDCG@20	P@20
QL	0.253	0.415	0.369
WE	0.135	0.257	0.227
PV	0.177	0.288	0.264
WE-LM	0.255*	0.417*	0.370*
PV-LM	0.259*	0.418	0.371

Table 2: Comparison of different models over the Robust04 collection with title queries. * means significant difference over QL respectively at 0.005 significance level measured by Fisher randomization test.

sentation for each document. When applied to short text like web page titles, which have few n-grams, word hashing produces high quality representations without losing too much information. However, this technique becomes problematic as document length increases from tens of words to hundreds of words. According to our observations, the n-gram representations for documents with hundreds of words are dense and noisy. Unsurprisingly, our experiments with DSSM on standard ad-hoc retrieval collections were not effective. The MAP of DSSM and its convolution version (CLSM) are less than 0.1 on Robust04 title queries. Notice that the same metric for query likelihood (QL) model is 0.253 as shown in Table 2.

Varied relevance granularity. Another problem that makes ad-hoc retrieval difficult for existing deep models is the vague definition for relevance. A short document could be relevant to a query because its main topic is related to the query. Meanwhile, a long document could be relevant to a query if it has a subtopic that describes the query. This characteristic of ad-hoc retrieval presents challenges to both supervised and unsupervised neural models. For supervised models like DNN and RNN, the back propagation of relevance information affects the gradient computation on all input words. However, most of these words may not be related to the document's relevance with a specific query (especially for long documents). A considerable amount of labeled data is needed in order to learn the weights for a model that can understand the relevance of a document from different angles.

For unsupervised models like WE [12] and the paragraph vector model (PV) [5], the embedding representations of documents are constructed to capture their main topics. These representations lack discriminative ability at query time because we cannot distinguish the finer difference between semanticly related words and subtopics [16]. For example, Table 2 shows the performance of retrieval models including WE and PV. WE [12] aggregates embeddings of words to form document representations and ranks documents according to their cosine similarities with queries. PV estimates a language model with paragraph vector model [5] and ranks documents according to the likelihood of queries given document models. Using WE and PV solely did not perform well compared to QL with dirichlet smoothing. We only achieve positive results when we combined these models with language modeling approaches that explicitly capture the exact matching information of queries and documents. As these problems pose significant challenges from a deep learning perspective, one direction of future research is examining the role of the attention mechanism when dealing with documents at the coarse granularity level.

4. CONCLUSION

We have shown the efficacy and shortcomings of common neural architectures at varying levels of IR task granularity. When candidate answers are short, CNNs and LSTM networks perform at equivalent levels with differences attributed to attention methods and structure differences beyond the convolutional and LSTM layers. At the passage level, we demonstrate that LSTMs are able to store additional temporal information which an equivalent CNN is unable to accomplish. Lastly, we discuss the unique problem that ad-hoc retrieval poses for neural networks, and potential solutions to overcome these issues.

5. ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF IIS-1160894. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] C. N. dos Santos, M. Tan, B. Xiang, and B. Zhou. Attentive pooling networks. *CoRR*, abs/1602.03609, 2016.

[2] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou. Applying Deep Learning to Answer Selection: A Study and An Open Task. *Applying Deep Learning to*, aug 2015.

[3] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. *CoRR*, abs/1503.03244, 2015.

[4] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM, 2013.

[5] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[6] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward. Semantic modelling with long-short-term memory for information retrieval. *arXiv preprint arXiv:1412.6629*, 2014.

[7] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. K. Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(4):694–707, 2016.

[8] A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *SIGIR*, SIGIR '15, pages 373–382, New York, NY, USA, 2015. ACM.

[9] 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*

the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 101–110. ACM, 2014.

[10] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers on large online qa collections. In *ACL:HLT*, pages 719–727, 2008.

[11] M. Tan, B. Xiang, and B. Zhou. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108, 2015.

[12] I. Vulić and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372. ACM, 2015.

[13] D. Wang and E. Nyberg. A recurrent neural network based answer ranking model for web question answering. In *WebQA Workshop, SIGIR '15, Santiago, Chile*.

[14] D. Wang and E. Nyberg. A recurrent neural network based answer ranking model for web question answering. In *WebQA Workshop, SIGIR '15, Santiago, Chile*.

[15] W. Yin, H. Schütze, B. Xiang, and B. Zhou. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *CoRR*, abs/1512.05193, 2015.

[16] C. Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141, 2008.