Learning Deep Representations for Natural Language Processing Applications

Ivano Lauriola[1,2]

[1] University of Padova - Department of Mathematics
Via Trieste, 63, 35121 Padova - Italy

[2] Fondazione Bruno Kessler
Via Sommarive, 18, 38123 Trento - Italy

ivano.lauriola@phd.unipd.it

Abstract. Recently, the literature shows that the representation of the data plays a crucial role in machine learning applications. Hence, several methods were born to learn the best representation for a given problem, as is the case of Deep Neural Networks and Multiple Kernel Learning. These methods reduce the human effort in designing good representations while increasing the expressiveness of the learning algorithms. In this project, the representation learning is analyzed from two different viewpoints. The former aims to develop novel technologies and models to learn the representation, mainly focusing on Embeddings, Multiple Kernel Learning, Deep Neural Networks, and their combination. The latter aims to provide a proof-of-concept of these methods on real-world Natural Language Processing tasks, such as the Named Entity Recognition and large-scale document classification in the biomedical domain.

Keywords: Representation Learning · Natural Language Processing · Named Entity Recognition · Deep Learning · Multiple Kernel Learning

1 Introduction

When dealing with Machine Learning methods, one of the most expensive steps is the definition of the representation which describes the shape of the data. An extensive literature [2, 4, 9, 11] shows that the choice of the representation is a key step for building good predictors. Different representations emphasize different aspects of the main problem and could entail different results.
In the context of textual analysis and document classification, a document can be represented as the Set-Of-Words that compose it, potentially by including the number of occurrences of each word, as in the well-known Bag-Of-Words representation. These representations are focusing on the content of the text, by analyzing the presence/absence of words in the document. Otherwise, the same document can be expressed as a set of *n-grams*, aiming to catch the dependencies between groups of words. A representation is good if the task can be "easily" solved. However, the selection of the most suitable representation for a given problem is a hard task.

In a typical learning pipeline, the user tries several representations, guided by some prior knowledge or via a validation procedure. However, this process is computationally expensive when the number of possible representations is large. Besides, the pool of representations taken into consideration is not exhaustive, and it defines some bias, bounding the expressiveness of the learning algorithm with a sub-optimal representation. To overcome the aforementioned issue, methods to directly learn the best representation for a given problem have been recently proposed [2]. Several representation learning paradigms exist in the literature. In this project, we are focusing mainly on Deep Neural Networks (NNs) and Multiple Kernel Learning (MKL). The former is a very popular approach due to its expressiveness and empirical effectiveness at learning the representation among a hierarchy of features with increasing complexity. The latter aims at learning the representation as a combination of several weak implicit representations, named kernels [6]. Each method has its own advantages and bottlenecks. Usually, Deep NNs achieve better results with respect to classical MKL algorithms, but they require a huge amount of training data, and they are less scalable. Moreover, the MKL is supported by several theoretical properties [9], and algorithms find an optimal solution instead of a local minimum.

In this work, the representation learning problem is analyzed from two different viewpoints. The former aims at understanding, developing and improving theoretically sound representation learning models, algorithms and tools. In this step the focus is on MKL, Deep NN, Neural Embeddings and their cooperation, aiming at combining the key aspects of these methods. The latter step is more practical and aims at understanding and evaluating the empirical effectiveness of such methods in complex Natural Language Processing applications. The two main applications that we are considering are large-scale online biomedical semantic indexing of PubMed documents based on the Medical Subject Headings (MeSH) [13], and the Biomedical Named Entity Recognition (BNER) task, whose purpose is to recognize and extract relevant entities and concepts from the biomedical literature. These entities can be the name of proteins, cellular components, diseases, species and so on.

This project is a joint work between the University of Padova and the Bruno Kessler Foundation. The main advisor is Fabio Aiolli, from the University of Padova, Dept of Mathematics. Co-advisors are Alberto Lavelli from Fondazione Bruno Kessler, and Giuseppe Sartori from the University of Padova, Dept of General Psychology. The doctoral course of the candidate is *Brain, Mind and Computer Science* of the University of Padova. This work is partially supported by grant CR30I1_162758 of the Swiss National Science Foundation.

2 State of the art

Representation Learning is one of the most challenging fields in machine learning research [2]. Two well-known approaches for this purpose consider the application of Deep NN [17], or Multiple Kernel Learning [6].
Due to their theoretical and empirical effectiveness, representation learning ap-

proaches have been widely applied to several domains, especially in large-scale applications where there is a lack of prior knowledge. Some examples of applications are sentence classification [8] and multimodal sentiment analysis [14]. Recently, the literature provides mechanisms to learn effective representations for Natural Language Processing applications [18]. This is the case of 1D Convolutional Neural Networks [16, 7], or dedicated Embeddings to map words, sentences, and documents into dense vectors. One of the most known algorithms for this purpose is Word2Vec [11].

In this project, one of the main interesting applications of Natural Language Processing (NLP) is the Named Entity Recognition (NER) [12] in the biomedical domain. Lately, standard NLP techniques have been combined with machine learning tools in order to solve this task, including the usage of Support Vector Machines and Neural Networks [1, 3]. State-of-the-art representations for the BNER task consist of hand-crafted features based on a strong prior knowledge [15, 1], and word-embeddings. Each representation has their own advantages. General-purpose word-embeddings can be easily pre-trained on large-scale corpora, and they do not require a lot of prior knowledge. Hand-crafted representations instead, could better represent the problem by means of a powerful prior knowledge, but they require a lot of human effort to extract relevant features.

3 Direction, Methodology and Practical impact

Nowadays, the literature considers Deep Neural Networks as the state-of-the-art of representation learning approaches, without taking into account the limits of such methods, such as the lack of prior knowledge, the lack of training data, and the computational cost. In this work, the representation learning paradigm is considered from a more general point of view, without any bias on methodologies and without focusing exclusively on Deep Neural Networks. We expect to better understand the potential between shallow and Deep learning techniques, with a consequent improvement of classification accuracy on NLP applications and machine learning tasks in general.

As discussed before, this work is spread over two different phases. The former consists of analyzing, evaluating and improving novel technologies, models and algorithms to learn the representation from data directly. The main mechanisms taken into account for this purpose are NNs, MKL and Embedding strategies. Empirical effectiveness and a comparison between these and classical approaches is mandatory, aiming to analyze the limits and pros of representation and Deep learning. This step includes the study of novel algorithms, efficient optimization procedures, the analysis of theoretical bounds, an exhaustive empirical evaluation, and a deep analysis of scalability, robustness, and efficiency of the proposed algorithms. Anyhow, unlike the classical representation learning methodologies, this work also aims to combine these paradigms. For instance, MKL methods could combine hidden representations computed from NNs.

Usually, the effectiveness of the large part of these methods is analyzed by using sand-box environments or benchmark datasets. However, these datasets do not reflect the complexity of real-world applications, where there are a lot of unexpected problems, such as noise, missing data or lack of prior knowledge. In order to assess the effectiveness and robustness of our methods, in the latter phase of the research project the acquired knowledge, methods and techniques will be applied to complex Natural Language Processing tasks.

4 Preliminary results

Preliminary results on the application of representation learning techniques on the BNER task are showed and described in the paper *Learning Representations for Biomedical Named Entity Recognition*, accepted at the NL4AI workshop of the AI*IA (2018) conference.
In that work, a comparison of domain-specific and general purpose representations in the BNER task has been performed. Each of the considered representations emphasizes different viewpoints of the problem. However, each ontology (proteins, diseases...) has different complexity, and it requires a proper representation instead of a global one. Even if these representations achieve individually comparable results, they express orthogonal information, and the cooperation between these pieces of information could further improve the performance. A general framework based on the MKL paradigm has been considered to learn the representation for each ontology automatically. Results show that the combination through the MKL paradigm improves the accuracy of the correct recognition. Besides, our solution achieves better results than other state-of-the-art approaches, including Convolutional Neural Networks. Moreover, results clearly show that the complexity of the representation plays a key role in this application, and it must be considered in the learning procedure.

For this purpose, we proposed a novel MKL algorithm which takes into account the expressiveness/complexity of the obtained representation in its objective function in such a way that a trade-off between large margins and simple hypothesis spaces can be found. Broadly speaking, the algorithm, named MEMO [10] (Minimum Effort Maximum Output), tries to maximize the margin between classes and minimize the Spectral Ratio of the solution simultaneously. The Spectral Ratio is an empirical measure of the expressiveness of a kernel, which has been proposed in [4]. The algorithm has been compared with several baselines, including other state-of-the-art margin-based MKL methods.

However, margin-based algorithms do not consider the spread of the data in the feature space, which is a relevant aspect of a good representation [5]. For this purpose, several MKL algorithms exist in the literature which try to minimize the ratio between the radius of the Minimum Enclosing Ball (MEB) which contains data in the feature space, and the margin between classes. However, these algorithms perform some relaxations of the main problem to make it tractable. As far as we know, we propose the first MKL algorithm which optimizes the exact ratio, through an alternate optimization procedure. The algorithm, dubbed GRAM, has been proposed at the ICANN conference [9], and an extension for the *Machine Learning Journal* is currently under review.

References

1. Basaldella, M., Furrer, L., Tasso, C., Rinaldi, F.: Entity recognition in the biomedical domain using a hybrid approach. Journal of biomedical semantics **8**(1), 51 (2017)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
3. Crichton, G., Pyysalo, S., Chiu, B., Korhonen, A.: A neural network multi-task learning approach to biomedical named entity recognition. BMC bioinformatics **18**(1), 368 (2017)
4. Donini, M., Aiolli, F.: Learning deep kernels in the space of dot product polynomials. Machine Learning **106**(9-10), 1245–1269 (2017)
5. Gai, K., Chen, G., Zhang, C.s.: Learning kernels with radiuses of minimum enclosing balls. In: Advances in neural information processing systems. pp. 649–657 (2010)
6. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of machine learning research **12**(Jul), 2211–2268 (2011)
7. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: Advances in neural information processing systems. pp. 2042–2050 (2014)
8. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
9. Lauriola, I., Polato, M., Aiolli, F.: Radius-margin ratio optimization for dot-product boolean kernel learning. In: International Conference on Artificial Neural Networks. pp. 183–191. Springer (2017)
10. Lauriola, I., Polato, M., Aiolli, F.: The minimum effort maximum output principle applied to multiple kernel learning. Proceedings of the 26th European Symposium of Artificial Neural Networks, Computational Intelligence and Machine Learning (2018)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
13. Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G., Kakadiaris, I.: Results of the fifth edition of the BioASQ challenge. In: BioNLP 2017. pp. 48–57. Association for Computational Linguistics, Vancouver, Canada, (August 2017)
14. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2539–2544 (2015)
15. Saha, S.K., Sarkar, S., Mitra, P.: Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of biomedical informatics **42**(5), 905–911 (2009)
16. dos Santos, C., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 69–78 (2014)
17. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)
18. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. arXiv preprint arXiv:1708.02709 (2017)