# UNIVERSITY OF FORWARD THINKING WESTMINSTER⌗

## WestminsterResearch

http://www.westminster.ac.uk/westminsterresearch

**Are Deep Learning Approaches Suitable for Natural Language Processing?**

**Alshahrani, S. and Kapetanios, E.**

This is an author's accepted manuscript of a paper presented at NLDB 2016: 21st International Conference on Applications of Natural Language to Information Systems and published in Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings, Lecture Notes in Computer Science, vol. 9612, pp. 343-349, 2016.

The final publication is available at Springer via:

https://dx.doi.org/10.1007/978-3-319-41754-7_33

# Are Deep Learning Approaches Suitable for Natural Language Processing?

S. Alshahrani[*] and E. Kapetanios[*]

[*] Cognitive Computing Research Group, Computer Science Department, University of Westminster, London, UK

{w1484137@my.westminster.ac.uk, E.Kapetanios@westminster.ac.uk}

**Abstract.** In recent years, Deep Learning (DL) techniques have gained much attention from Artificial Intelligence (AI) and Natural Language Processing (NLP) research communities because these approaches can often learn features from data without the need for human design or engineering interventions. In addition, DL approaches have achieved some remarkable results. In this paper, we have surveyed major recent contributions that use DL techniques for NLP tasks. All these reviewed topics have been limited to show contributions to text understanding, such as sentence modelling, sentiment classification, semantic role labelling, question answering, etc. We provide an overview of deep learning architectures based on Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Recursive Neural Networks (RNNs).

**Keywords:** Deep Learning, Natural Language Processing, Artificial Neural Networks, Convolutional Neural Networks, Long Short-Term Memory, Recursive Neural Networks.

## 1    Introduction

Machine Learning (ML) is a robust AI tool, which has shown its usefulness in our daily lives, for example, with technologies used in search engines, image understanding, predictive analytics, transforming speech to text and matching relevant text. All ML approaches can be roughly classified as supervised, unsupervised and semi-supervised.

Building a machine-learning system with features extraction requires specific domain expertise in order to design a classifier model for transforming the raw data into internal representation inputs or vectors. These methods are called representation learning (RL), in which the model automatically feeds in raw data to detect the needed representation.

In particular, the ability to precisely represent words, phrases, sentences (statement or question) or paragraphs, and the relational classifications between them, is essential

to language understanding. Deep learning approaches are similar to RL methods using multiple levels of representational processing [1].

Deep Learning (DL) involves multiple data processing layers, which allow the machine to learn from data, through various levels of abstraction, for a specific task without human interference or previously captured knowledge. Therefore, one could classify DL as unsupervised ML approach. Investigating the suitability of DL approaches for NLP tasks has gained much attention from the ML and NLP research communities, as they have achieved good results in solving bottleneck problems [1]. These techniques have had great success in different NLP tasks, from low level (character level) to high level (sentence level) analysis, for instance, sentence modelling [12], Semantic Role Labelling [4], Named Entity Recognition [19], Question Answering [15], text categorization [11], opinion expression [8], and Machine Translation [9].

The focus of this paper is on DL approaches that are used for NLP tasks. ANNs are discussed in Section 2 below, while CNNs are considered in Section 3. Section 4 discusses the suitability of these techniques for NLP and the implications for future research. We then conclude with a brief overview in Section 5.

## 2    Artificial Neural Networks Approaches

A standard neural network consists of many connected units called neurons, each generating a sequence of real-valued activations. Neurons are activated by previous neurons in the circuit, via weighted connections. Each link transforms chains of computational sequences between neurons in a non-linear way. ANNs are robust learning models that are about precisely assigning weights across many levels. They are broadly divided into two types of ANN architectures that can be feed-forward networks (FF-NNs), Recurrent Neural Networks (RNNs) and Recursive Neural Networks [10]. FF-NNs architecture consists of fully connected network layers. The RNNs model, on the other hand, consist of a fully linked circle of neurons connected for the purpose of back-propagation algorithm implementation.

FF-NNs applied to NLP tasks consider syntax features as part of semantic analysis [26]. ANN learning models have been proposed that can be applied to different natural language tasks, such as semantic role labelling and Named Entity Recognition [23]. The advantage of these approaches is to avoid the need for prior knowledge and task specific engineering interventions. FF-NNs models have achieved an efficient performance in tagging systems with low computational requirements [4]. A Neural-Image-QA approach has been proposed, which combines NLP and image representation for a textual question answering system regarding images [16]. However, this system has shown incorrect answers regarding the size of the objects in an image, and also with spatial reasoning questions, such as "which is the largest object?". In another study [28], a novel approach was proposed to match queries with candidate answers (sentences) depending on their semantic meaning, by combing distributed representations and deep learning without the need for features engineering or linguistic expertise. State-of-the-art neural network-based distributional models have achieved a high performance in various natural NLP tasks; for instance, document classification [13] and

Entity Disambiguation and Recognition [23]. Another FF-NNS approach is the ANNABEL model [7], which aspires to simulate the human brain and to become able to process verbal and non-verbal natural language without prior knowledge. This work is based on a large-scale neural architecture which acts as the central executive and interacts with humans through a conversational interface. The lack of control over information flow through memory inspired the ANNABEL team to build their mental action to control the Short-Tern Memory system.

**Recurrent Neural Network Approaches**

The RNNs model fully links neurons in a circle of connections for the purpose of back-propagation algorithm implementation. Recurrent Neural Networks have been recommended for processing sequences [10], while Recursive Neural Networks are collections of recurrent networks that can address trees [6]. Another application uses Recurrent Neural Networks for question answering systems about paragraphs [15], and a Neural Responding Machine (NRM) has been proposed for a Short-Text Conversation generator, which is based on neural networks [21]. In addition, Recurrent Neural Networks models offered state-of-the-art performance for sentiment classification [13], target-dependent sentiment classification [25] and question answering [15]. Adaptive Recurrent Neural Network (AdaRNN) is introduced for sentiment classification in Twitter promotions based on the context and syntactic relationships between words [2]. Furthermore, Recurrent Neural Networks are used for the prediction of opinion expression [8].

## 3 Convolutional Neural Networks Architecture

CNNs evolved in the field of vision research, where the first use was for performing image classification learning to detect the edges of an image from its raw pixels, in the first layer, and then using the edges to identify simple shapes located in the second layer. These shapes are then used to identify higher-level features such as facial shapes in the higher layers of the network [27].

Applying a non-linear function over a sequence of words, by sliding a window over the sentences, is the key advantage of using CNNs architecture for NLP tasks [57]. This function, which is also called a 'filter', mutates the input (k-word window) into a d-dimensional vector that consists of the significant characteristic of the words in the window. Then, a pooling operation is applied to integrate the vectors, resulting from the different channels, into a single n-dimensional vector. This is done by considering the maximum value or the average value for each level across the different windows to capture the important features, or at least the positions of these features. For example, Figure 1 gives an illustration of the CNNs' structure where each filter executes convolution on the input, in this case a sentence matrix, and then produces feature maps, hence it showing two possible outputs. This example is used in the sentence classification model.

Most NLP classification tasks use CNN models and pioneering work has been done using these methods for semantic-role labelling [3], sentiment and question-type classification [14], text understanding from the basic character level [57] and text categorization [11].A further use of CNNs is in the visual question-answering model, known as mQA model, which is able to answer a question about image contents [5]. This model comprises of four parts: a Long Short-Term Memory (LSTM) to extract the query representation, a CNN, another LSTM for processing the context of the answers, and a process part for generating the answer, which can collect the information from the first three parts in order to provide answers. This can be in the form of a word, a phrase or a sentence. A new convolutional latent semantic approach for vector representation learning [22] uses CNNs to deal with ambiguity problems in semantic clustering for short text. However, this model can work appropriately for long text as well [25]. CNNs are proposed for sentiment analysis of short texts that learn features of the text from low levels (characters) to high levels (sentences) to classify sentences in positive or negative prediction analysis. However, this approach can be used for different sentence sizes [18]. The Dynamic CNN (DCNN) is embraced for semantic sentence modelling which includes extracting features from sentences. DCNN is trained by selectively weighting functions between linked network layers [12].

## 4    Discussion and Future Research

DL approaches are used for a variety of NLP tasks, as shown in Table 1. This gives an overview of different NLP tasks using NN, CNN and RNN (Recurrent Neural Network) approaches for NLP tasks related to semantic and context analysis. The RNNs model can achieve a competitive performance in NLP tasks with sequence input; but, if the problem deals with sparsity in its input, this can be solved using CNNs. The CNN implementation is most commonly used in NLP classification tasks, such as sentiment classification, question answering, and sentence selection. CNNs have achieved state-of-the-art performance in solving data sparsity problems in large structures of NLP tasks.

In addition, DL learning techniques provide an enhanced performance for NLP tasks due to their ability to its distinguishing features. Firstly, they use powerful performance, provided by advanced CPUs, to improve training processes which are implemented in current deep learning techniques, for NLP tasks. Secondly, they provide functional evidence in terms of representation, as convolutional filters can automatically learn the term representations without requiring prior knowledge. Furthermore, RNNs deal with sequence inputs, and we therefore expect models which use these methods to have a large impact on natural language understanding over the next couple of years.

Conversely, conventional deep learning approaches that may be very appropriate for processing of raw pixels in images cannot work properly for text processing, due to the need for more features and data in the hidden layer parameters [24].

**Table 1.** Gives an overview of different NLP tasks using DL approaches for for NLP tasks related to semantic and context analysis.

| Text Understanding Task | ANNs | CNNs | RNNs |
|---|---|---|---|
| Sentence Classification | | ✓ | |
| Named Entity Recognition (NER) | ✓ | | |
| Text Categorization | | ✓ | |
| Semantic Role Labelling | | ✓ | |
| Semantic Clustering | | ✓ | |
| Short-Text Conversation | ✓ | | ✓ |
| Question answering | ✓ | ✓ | ✓ |
| Sentiment analysis | ✓ | | ✓ |
| Paraphrase Detection | ✓ | | |
| Document Classification | ✓ | | |
| Topic Categorization | | ✓ | |
| Opinion Expression Prediction | | | ✓ |

Popular word-embedding models are GloVe (Global Vectors for word representation) [17], Word2vec [20], and embedding algorithms, such as dynamically sized context windows [4]. Word2vec is an open-source Google text processing tool published in 2013. This approach is used for word representation. Word2vec relies upon two algorithms, skip-grams and continuous bag of words (CBOW). The skip-gram model process the current input word using a linear classifier to predict surrounding words in a specific scope, whereas CBOW is a trained model for understanding ambiguous words, based on their context [20]. GloVe is a vector representation word-learning method.

GloVe works by computing the large word co-occurrence matrix in memory, and is dependent on matrix factorization algorithms, making it a good model for optimization. In contrast, Word2Vec goes through sentences directly, treating each co-occurrence separately. The advantage of these approaches (word-embedding) can be combined with other deep learning models to enhance performance for NLP tasks. In addition, a gradient-based method can be used in NN training, which helps to reduce error over a training set.

In addition, further research may explore the suitability of DL for Conversational User Interfaces and/or for dynamic Question Answering, particularly for graph based pattern search and recognition, over Linked Open Data.
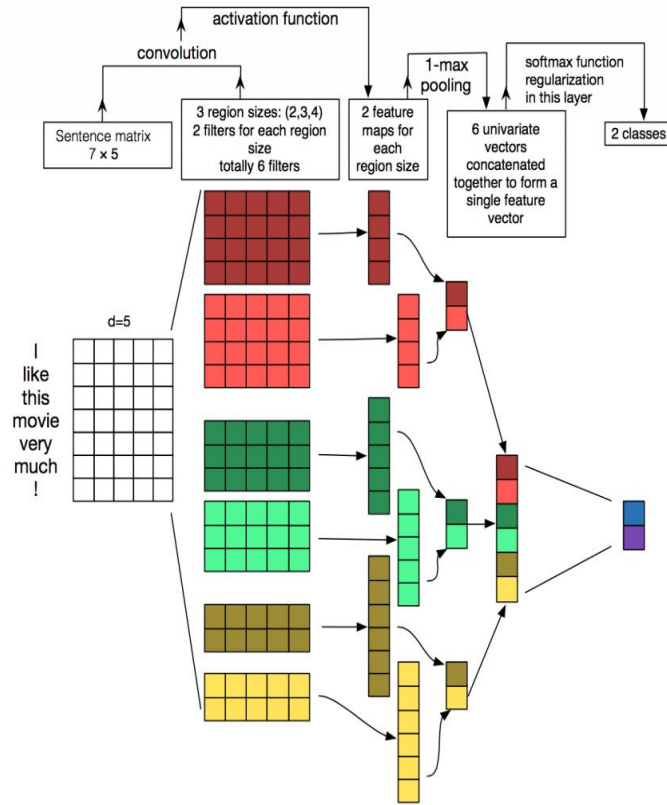
**Fig. 1.** Model of three filter division sizes (2, 3 and 4) of CNNs architecture for sentence classification. (Source: Zhang, Y. & Wallace, B. (2015)).

## 5 Conclusion

In this paper, we have presented an overview of the deep learning based approach to natural language processing tasks. Different studies are included in this review, covering various NLP tasks that implemented ANNs, CNNs, RNNs and LSTM. These approaches can be combined with other deep learning models to develop improved performance for NLP tasks. The advantage of using these approaches is to avoid the need for prior knowledge and human engineering interventions. The key conclusion from this overview is that deep learning approaches are ideal for solving data sparsity problems in large structures, while CNNs have the advantage of fast performance and also providing functional evidence for representational learning and feature extraction.

We have not discussed statistical surveys or performance comparisons relating to deep learning techniques used for NLP tasks. This should be explored comprehensively as part of future research.

## References:

1. Ba, L. & Caurana, R.: Do Deep Nets Really Need to be Deep ? arXiv preprint arXiv:1312.6184, 521(7553), pp.1-6 (2013).
2. Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K.: Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In: Acl-2014, 49–54 (2014).
3. Collobert, R. & Weston, J.: Fast semantic extraction using a novel neural network architecture. In: Proceedings of the 45th Annual Meeting of the ACL (pp. 560– 567) (2007).
4. Collobert, R. Weston, J. Bottou, M. Karlen, K. Kavukcuglu, P. Kuksa.: Natural Language Processing (Almost) from Scratch. In: Journal of Machine Learning Research 12:2493–2537(2011).
5. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L. & Xu, W.: Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In: Arxiv, 1–10 (2015).
6. Goller, C. & Kuchler, A.: Learning Task-Dependent Distributed Representations by Backpropagation through Structure. In: Proc. of the ICNN-96, pp. 347–352. IEEE (1996).
7. Golosio, B., Cangelosi, A., Gamotina, O. & Masala, G. L.: A Cognitive Neural Architecture Able to Learn and Communicate through Natural Language. Plos One, 10(11) (2015).
8. Irsoy, O., & Cardie, C.: Opinion Mining with Deep Recurrent Neural Networks. Emnlp-2014, 720–728 (2014).
9. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. In Proc. ACL-IJCNLP (2015).
10. J. L. Elman.: Finding structure in time. In: Cognitive science, 14(2):179-211 (1990).
11. Johnson, R., Zhang, T.: Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding, 1-12 (2015).
12. Kalchbrenner, N., Grefenstette, E., Blunsom, P. : A Convolutional Neural Network for Modelling Sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, {ACL} 2014, June 22-27, 2014, Baltimore, MD, USA, Vol 1: Long Papers (pp. 655–665) (2014).
13. Karl Moritz Hermann, Phil Blunsom.: Multilingual Models for Compositional Distributional Semantics. In Proceedings of ACL (2014).
14. Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746-1751(2014).
15. M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, H. D. III.: A neural network for factoid question answering over paragraphs. In EMNLP (2014).
16. Malinowski, M., Rohrbach, M. & Fritz, M. : Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. (2015). In: IEEE International Conference on Computer Vision, 1-9 (2015).
17. Pennington, J., Socher, R., Manning. : GloVe: Global Vectors for Word Representation. In: EMNLP (2014).

18. Santos, C. N. dos, Gatti, M. : Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. Coling-2014, 69–78 (2014).
19. Santos, C.N. Dos, Guimarães, V.: Boosting Named Entity Recognition with Neural Character Embeddings. In Acl, (2014), pp.25–33 (2015).
20. Schmidhuber, J.: Deep Learning in neural networks: An overview. Neural Networks, 61, 85-117 (2015).
21. Shang, L., Lu, Z. Li, H.: Neural Responding Machine for Short-Text Conversation.In: Acl-2015, 1577-1586 (2015).
22. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: 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 - CIKM '14, 101-110 (2014).
23. Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., Wang, X.: Modelling Mention, Context and Entity with Neural Networks for Entity Disambiguation, (Ijcai), 1333–1339 (2015).
24. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013) (2013).
25. Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F. & Hao, H.: Semantic Clustering and Convolutional Neural Network for Short Text Categorization. In: Proceedings ACL 2015, 352-357 (2015).
26. Weston, J., America, N. E. C. L., Way, I.: A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. Icml2008, 160-167(2008).
27. Y. LeCun, Y. Bengio: Convolutional networks for images, speech, and time-series, In: MIT Press, pp. 255-258 (1995).
28. Yu, L., Hermann, K. M., Blunsom, P., Pulman, S.: Deep Learning for Answer Sentence Selection. NIPS Deep Learning Workshop, 9 (2014).
29. Zhang, X., LeCun, Y.: Text Understanding from Scratch. (2015).