Deep Learning in Natural Language Processing

Poma Panezai[a]
Bushra Qayyum[b]
Abdul Qadeer[c]
Yahan Jan[d]

Abstract

Recent breakthroughs in machine learning have provided the artificial intelligence field with the necessary capabilities to address persistent challenges. This paper provides a brief overview of deep learning, classified into three main categories: discriminative, generative, and hybrid models. Natural language processing (NLP) is one of the domains that has considerably profited from these breakthroughs. Deep learning techniques have allowed the efficient handling of many complex tasks related to natural language processing. This paper emphasizes important NLP tasks and notable projects that have used deep learning to tackle them. Despite the wide range of NLP tasks, the results show that deep learning continuously surpasses conventional methods across numerous applications.

Keywords: Deep Learning, Natural Language Processing, Discriminative Models, Generative Models, Hybrid Models, Machine Learning

Article history:
Received on: November 12, 2024
Revised on: December 29, 2024
Accepted on: December 30, 2024
Published on: December 31, 2024

[a,b,d] Department of Computer Science, BUITEMS University, Quetta | poma.panezai@buitms.edu.pk
[c] BUET, Khuzdar

How to Cite
Panezai, P., Qayyum, B., Qadeer, A., & Jan, Y. (2024). Deep Learning in Natural Language Processing. *Journal of History and Social Sciences, 15(*2), 69-86. https://doi.org/10.5281/zenodo.15303476

INTRODUCTION

Machine learning technology is becoming increasingly involved in many aspects of the modern community, from identifying spam in an email provider to detecting credit card fraud to identifying heart failure[1]. Lately, most machine learning methodologies have used shallow learning architectures, which are considered extremely basic, often comprising a single hidden layer with nonlinear feature transformations. Figure 1 shows a shallow learning architecture consisting of an input, hidden, and output layer. This single layer is responsible for converting the unprocessed input data into a problem-specific feature space. Although shallow learning architectures have demonstrated efficacy in addressing numerous straightforward or well-structured problems, their constrained design is inadequate for tackling more intricate issues associated with real-world applications such as speech, pictures, and language. Therefore, there was a need for more advanced techniques to emerge and tackle these complicated problems, we call these techniques deep learning.

Deep learning is a category of machine learning methods characterized by multiple hidden layers compared to the single layer that makes up shallow learning architectures. Deep learning, also called representational learning, allows the machine to discover the representations in a raw input data, and use these representations for detection or classification. The multiple layers architecture could be seen as multiple levels of representation, where each level consists of simple nonlinear modules that transform the representation into a higher-level representation that is slightly more abstract.

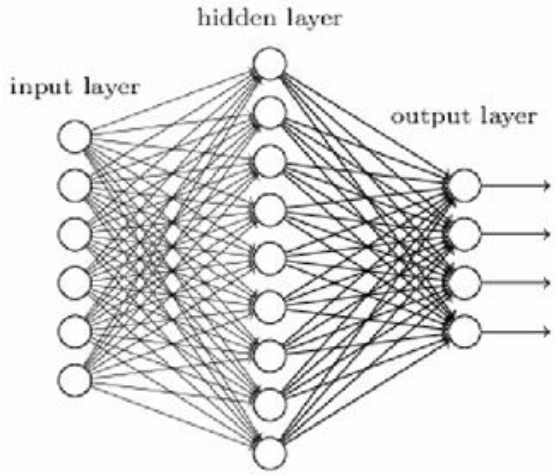

Fig. 1. Shallow learning with single hidden layer[2].

With deep learning, artificial intelligence community can tackle complicated problems that used to be a struggle beforehand. This has been shown through many applications using deep learning to address complex problems, including image and speech recognition. Deep learning also outperformed conventional machine-learning methods in forecasting the activity of possible

[1] R. Vijayakrishnan et al., "Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record," Journal of cardiac failure, 2014.

[2] M. Nielsen, "Neural Networks and Deep Learning," 2017.

therapeutic compounds[3]. Other examples for deep learning applications could be automatic machine translation, face recognition, self-driven car, and code completion. Deep learning is not a novel technique; however, as mentioned in[4], three advancements have revived it: significant improvements in chip processing power (e.g., GPU units), a marked decrease in computer hardware expenses, and recent advancements in machine learning research. The remaining sections of the paper has been organised as follows: An overview of the three categories of deep architectures is given in Section II. A brief overview of NLP is given in Section III. Section IV shows that deep learning outperformed traditional methods used in different NLP tasks, by presenting a number of recent research projects in NLP tasks. Section V, presents a comparison between using RNN and CNN in NLP tasks, specifically in text classification to provide an insight of the well suited deep learning techniques to be used.

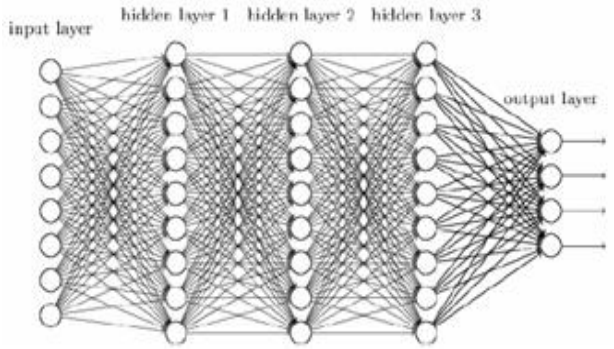

Fig. 2. Deep learning with multiple hidden layers[5].

THREE BROAD CLASSES OF DEEP ARCHITECTURES

Deep learning architectures are classified into three broad classes[6], generative, discriminative, and hybrid. Next, a brief introduction of each class is presented.

Generative Deep Architectures

Figure 2 shows the high-level architecture of a deep learning network, that consists of multiple hidden layers. A traditional deep network could use algorithms such as Back propagation as a learning algorithm. However, this algorithm suffers from some shortcomings. It did not work well in practice with neural networks that have more than three hidden layers. And the loss function in such learning is non-convex. These shortcomings were the reason behind moving toward the use of a shallow structure which has convex loss functions providing more efficient optimization, but

[3] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature vol. 521, 2014, pp. 436-444.
[4] L. Deng "Three classes of deep learning architectures and their applications: a tutorial survey," APSIPA transactions on signal and information processing, 2012.
[5] M. Nielsen, "Neural Networks and Deep Learning," 2017.
[6] L. Deng "Three classes of deep learning architectures and their applications: a tutorial survey," APSIPA transactions on signal and information processing, 2012.

again as stated earlier, shallow structures provide poor results when it comes to more complicated problems[7].

A deep belief network (DBN), a type of deep generative model developed in[8] is made of a stack of shallow structured networks called Restricted Boltzmann machines (RBM), where a greedy approach applies layer-by-layer training. Among the several benefits of this unsupervised learning method were the effective usage of unlabelled data and the optimization of DBN weights at a time complexity linear to the size and network depth. Deep autoencoders and Deep Boltzmann machines (DBM) are other generative deep architecture examples. When the output of recurrent neural networks (RNNs) is considered to be the future projected input data, they can also be categorized as deep generative architectures.

Discriminative Deep Architectures

Another deep architecture is discriminative deep architecture. Generally, the main goal of deep architecture models is to predict an output y given an input value x. The generative models aim to achieve that using a joint probability distribution p(x, y), and then through Bayes rules calculates the probability of y given x, p(y|x), and finally picks the most likely output y. A discriminative model will directly calculate the conditional probability p(y|x). Stated otherwise, the discriminative deep architecture aims to provide discriminative power for pattern classification, usually by characterizing the conditional distribution of classes given available data[9,10]. A second distinction between the generative model and the discriminative model would be the type of learning used by each one. The generative model uses unsupervised learning as mentioned previously while the discriminative model usually uses supervised learning.

Assume that we have the task of determining the language that someone is speaking, a generative approach solution would require us to learn each language and determine the spoken language according to the heard speech. On the other hand, a discriminative approach would be to determine the language based on linguistic differences without learning any language. An example of a discriminative model would be the Convolutional neural network (CNN) which goes back to a well-known research paper by LeCun and Bottou[11]. Like CNN, which is a feed-forward neural network made of a stack of separate layers, these networks are physiologically inspired from the way mammals visually interpret the environment around them using a layered architecture of neurones in the brain[12]. The convolutional layer preforms the core operations of the network, where a matrix called the 'kernel', slides over the input matrix of an image for example and

[7] J. Xu, H. Li and S. Zhou, "An Overview of Deep Generative Models," IETE Technical Review vol. 32, no. 2, 2014, pp. 131-139.

[8] G. Hinton, S. Osindero and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, 2006, pp. 1527-1554.

[9] A. Y. Ng and M. I. Jordan On Discriminative vs. Generative classifiers: A comparison of logistic regression and nave Bayes," Advances in neural information processing systems, vol. 2, 2002, pp. 841-848.

[10] L. Deng and N. Jaitly, Deep discriminative and generative models for pattern recognition," USENIX-Advanced Computing Systems Association, 2015.

[11] Y. LeCun et al., Gradient-based learning applied to document recognition," Proceedings of the IEEE vol. 86, no. 11, 1998, pp. 2278-2324.

[12] A. Saxena,"Convolutional Neural Networks: An Illustrated Explanation," Crossroads the ACM magazine for students, http://xrds.acm.org/blog/2016/06/convolutional-neural-networks-cnns-illustrated-explanation/, 2016, accessed on May 2017.

produces the so-called 'activation map', which is passed to the next layer. The pooling layer preforms an operation called 'down- sampling', which reduces the size of the 'activation maps'. It takes a sliding window moving it across the 'activation maps', transforming the values into representative values.

Since CNNs uses supervised learning, they need large amounts of labeled data to train, which restricts the applications in the fields that lack such data. Nevertheless, CNNs have proven very efficient in areas such as image classification [13, 14] and speech recognition[15,16]. Other examples for discriminative deep architectures include Deep stacking network (DSN) and RNN where the output is a labeled sequence associated with the input sequence[17].

File Formats for Graphics

The third class of deep architectures is hybrid deep architecture. This class combines the generative and the discriminative architectures to gain the benefits of both approaches. This could be achieved in one of two ways, either the goal would be discriminative but is assisted with the outcomes of generative architectures, or the other way around, where the goal is generative and discriminative architectures are used to learn the parameters[18]. A type of hybrid deep architecture is the Deep Neural Network (DNN), or hybrid Deep Belief Network (DBN). The goal of this network is discriminative, supported by the results of generative architectures. As outlined in the preceding section, the DBN serves as a generative deep architecture utilised for the initialisation of DNN weights through a process known as pre-training, as opposed to the conventional method of random initialisation. The pre-training of the DBN has been demonstrated to enhance the discrimination capabilities of the deep CNN compared to random initialisation[19].

Another example of hybrid deep architecture, wherein generative and discriminative frameworks are used to learn parameters, is a system that facilitates voice translation, converting spoken phrase in one language into text in another. This capability includes two tasks: the first is speech recognition, and the second is machine translation. Although both tasks are generative in nature, their parameters are learnt for discriminating[20,21].

[13] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural Networks," Advances in neural information processing systems, 2012, pp. 1097-1105.
[14] J. Dean et al., "Large Scale Distributed Deep Networks," Advances in neural information processing systems, 2012, pp. 1223-1231.
[15] L. Deng and X. Li "Machine learning paradigms in speech recognition: An overview," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, 2013, pp. 1060-1089.
[16] T. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Convolutional neural networks for LVCSR," IEEE International Conference on ICASSP, 2013, pp. 8614-8618.
[17] L. Deng "Three classes of deep learning architectures and their applications: a tutorial survey," APSIPA transactions on signal and information processing, 2012.
[18] Ibid.
[19] Ibid.
[20] H. Ney, "Speech translation: Coupling of recognition and translation," Proc. ICASSP, 1999.
[21] X. He and L. Deng "Speech recognition, machine translation, and speech translation — A unifying discriminative framework," IEEE Sig. Proc. Magazine, vol. 28, 2011.

BRIEF INTRODUCTION INTO NATURAL LANGUAGE PROCESSING

NLP is a multidisciplinary domain that merges artificial intelligence with linguistics, concentrating on the interaction between computational systems and human languages, encompassing both spoken and written forms. The primary goal of NLP research is to comprehend the processes by which people interpret language, in both textual and spoken formats. By acquiring this understanding, researchers aim to develop advanced tools and methodologies that enable computers to comprehend, process, and manipulate natural languages effectively[22]. This paper will list some of the well-known tasks in NLP presented in[23], and provide a review of each task and later on present some of the work on them:

Sentiment Analysis

This task aims to extract the perspective of the author or the speaker about specific objects or topics, determining the polarity or the emotional reaction towards them. It could be useful to compute customer satisfaction and identify trends of the public opinion in the social media. Textual entailment: This task represents a relation between two text fragments called 'text' and 'hypothesis', where readers of the 'text' would determine whether the 'text' entails the 'hypothesis', the 'text' contradicts the 'hypothesis', or the 'text' does not entail nor contradict the 'hypothesis'. Answer selection: Also called Question answering, this task aims to determine the answer to a given question. Some questions have a direct right answer like "what is the capital of Germany?", but other questions do not have a straight answer like "what is happiness?". This task targets the second type of questions, where it chooses the correct answer from a set of candidate sentences.

Part-of-Speech Tagging

This task assigns a part of speech to each word or term in a text, e.g. a noun, verb, adverb, etc. This assignment could depend on both the interpretation of the word itself and its context. Many words could take different parts of speech according to its context, e.g. "drive" could be a verb ("drive safely") or a noun ("the file is stored on the hard drive"), this makes assigning parts of speech more ambiguous. Author identification: This task aims to identify the author of a given text from a set of candidate authors. It could be applied to identify an anonymous author, detect plagiarism or find a ghost writer. A published writer usually has a unique writing style, and the primary goal of this task is to define an appropriate characterization of texts to capture that unique style of the writer, and use it in author identification[24,25]. Relation classification: This task aims to identify the relations between named entities in a given text, such as relations between people mentioned in the text, organizations, or locations. Other tasks also include question-relation matching, path query answering and coreference resolution.

[22] G. G. Chowdhury, S. Osindero and Y. Teh, "Natural language Processing," Annual review of information science and technology, vol. 37, no. 1, 2003, pp. 51-89.

[23] R. S. Dudhabaware and M. S. Madankar, "Review on Natural Language Processing Tasks for Text Documents," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2014.

[24] C. Qian, T. He and R. Zhang, "Deep Learning based Authorship Identification,".

[25] A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," IEEE International Conference on Machine Learning and Applications (ICMLA), 15th, 2016, pp. 898-903.

NLP TASKS

This section presents a set of recent research projects that employed different deep learning techniques to solve NLP tasks. The projects presented cover a range of NLP tasks for both text and speech. These tasks include: answer selection, sentiment analysis, author identification, part-of-speech tagging, spoken language understanding, and speech emotion recognition. In all the presented projects, deep learning techniques have shown to outperform the traditional methods used in each task.

Equations

Sentiment analysis is another previously discussed NLP task. As the amount of text data users generate on the Internet increases, sentiment analysis appears as an essential means for getting insights into public sentiments regarding specific topics or products, mainly when conducted on social media platforms like Facebook, Twitter, and Instagram. Advancements in deep learning algorithms offer more promising outcomes than conventional classification techniques.

Traditional methodologies employed for sentiment analysis, including Multinomial Naive Bayes (NB), Support Vector Machine (SVM), and Maximum Entropy (MaxEnt), utilize a bag-of-words model for text representation. Utilizing that approach yields a substantial text representation, as the bag-of-words is a vector whose length corresponds to the number of words in the lexicon. Simultaneously, employing a word vector with a length that is independent of the dictionary size will yield a more compact vector that represents only a single word rather than the entirety of the text. The bag-of-words approach has the disadvantage of ignoring the sequence of words, hence compromising the efficacy of sentiment analysis outcomes. Proposed the first study that applies deep learning techniques to classify sentiment of Thai Twitter data[26]. Mainly using two deep learning techniques, a Dynamic convolutional neural network (DCNN) followed by LSTM.

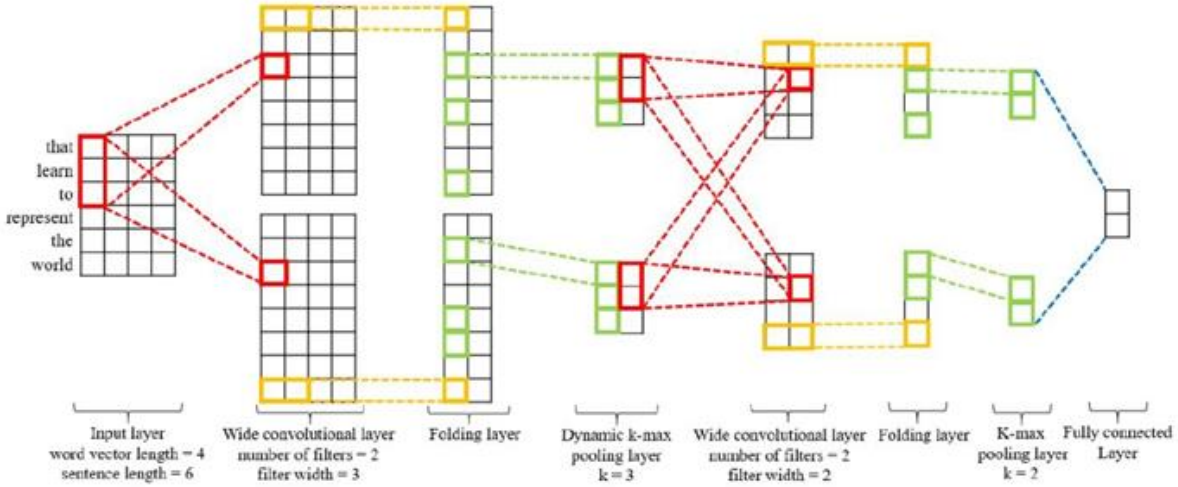

Fig. 3. Dynamic Convolutional Neural Network[27]

[26] P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," International Joint Conference on Computer Science and Software Engineering (JCSSE), 13th, 2016.
[27] Ibid.

DCNN is a CNN with a dynamic k-max pooling layer. Figure 3 shows the structure of DCNN. The first layer is a large convolutional layer in which filter matrices perform one-dimensional convolution on a row of sentence matrix. The second layer is a folding layer that combines two adjacency rows into a single row, followed by a dynamic k-max pooling layer that selects the highest maximum k-values from the sentence matrix's column. The network is made up of several convolutional, folding, and pooling layers. The network's top layer is fully interconnected and uses Softmax classification. The original word vectors were trained with word2vec rather than bag-of-words. Since DCNN is considered part of the discriminative deep architecture, and an unsupervised pre-training operation was performed to train the initial word vector, the approach utilized in [28] is an example of hybrid deep architectures. The experiment comparing their research results to NB, SVM, and MaxEnt, which are conventional approaches employing the bag-of-words model, revealed that LSTM and DCNN have higher accuracies than the other classifiers.

Another study using deep learning for sentiment analysis of short texts, described in[29], used two deep learning techniques: CNN and LSTM. An observation driven from[30] indicates that when using CNN architecture for text classification, many layers are needed to capture long-term dependencies in an input text, which could be avoided by taking advantage of RNNs ability to capture long-term dependencies with one single layer. But RNN suffers from vanishing gradient problem, and also since it is a biassed model, newer terms are more significant than older terms. even though important components could appear in any part of the text. These shortcomings of RNN were the reason behind using LSTM instead. Unlike [31]where LSTM followed DCNN, the researchers in[32] proposed ConvLstn neural net-work where LSTM substitutes the pooling layer in CNN, which reduces the loss of detailed local information and effectively reduces the number of the convolutional layers needed in order to capture long-term dependencies.

Similar to[33], the word vectors were initialized with pretrained vectors obtained from an unsupervised learning model. Experiments were carried out on two datasets, IMDB sentiment analysis dataset and Stanford sentiment treebank dataset. ConvLstm outperformed traditional methods such as NB and SVM, and achieved comparable results to other deep learning methods but with far less number of parameters. Both [34]and [35] used CNN and LSTM in their approaches, which indicates that hybrid deep architecture consists of these networks is well suited for sentiment analysis.

---

[28] Ibid.

[29] A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," International Conference on Control, Automation and Robotics (ICCAR), 3rd, 2017, pp. 705-710.

[30] X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," Advances in neural information processing systems, 2015, pp. 649-657.

[31] P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," International Joint Conference on Computer Science and Software Engineering (JCSSE), 13th, 2016.

[32] A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," International Conference on Control, Automation and Robotics (ICCAR), 3rd, 2017, pp. 705-710.

[33] P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," International Joint Conference on Computer Science and Software Engineering (JCSSE), 13th, 2016.

[34] P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," International Joint Conference on Computer Science and Software Engineering (JCSSE), 13th, 2016.

[35] A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," International Conference on Control, Automation and Robotics (ICCAR), 3rd, 2017, pp. 705-710.

Answer Selection

Answer selection is one of the tasks that has been attracting more attention recently, mainly because of the spread of question answering communities such as Quora, Yahoo! Answers, and Stackoverflow. And therefore, there is a need to develop an automatic mechanism to select answers. The challenge in this task lies in that selecting an answer does not solely depend on the semantic matching between answers and the question, but rather taking into consideration contextual factors as well. There may exist complex relations among the answers, for example, one answer could be a further elaboration on a previous answer, or it could simply be an expression of gratitude toward another answer. Supervised learning techniques are mainly used to tackle this difficulty, more specifically, DNNs are used for their strong learning abilities over other networks. Presented an attentive deep neural network architecture as an approach to this task. Next, an overview of the model used and the results derived from the experiments applied is presented[36].

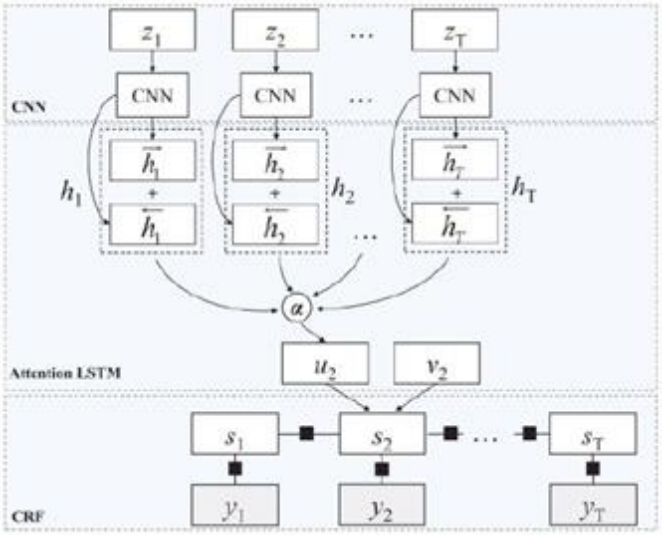

Fig. 4. Overview of the proposed attentive neural network[37]

Model Overview

The architecture of the model is composed of three different networks: a Convolutional Neural Network (CNN), an attention-based Recurrent Neural Network (RNN) utilizing Long Short-Term Memory (LSTM), and Convolutional Random Fields (CRF). These deep learning techniques are categorized within the class of discriminative deep architectures. Figure 4 illustrates the overview of the model. The initial component is the CNN, responsible for extracting features from each input sentence, whether it be a question or an answer, and subsequently compressing these features into a fixed-length vector. A CNN is structured with several components, including a word embedding layer, multiple convolutional layers utilizing various filters with window sizes of 3, 4, and 5, and a single max-pooling layer. The second component is the attention LSTM, which encodes the compressed vectors and learns dependencies from the sequential steps. The architecture includes a bidirectional LSTM layer (ht) succeeded by a soft attention layer. This

[36] Y. Xiang, Q. Chen, X. Wang and Y. Qin, "Answer Selection in Community Question Answering via Attentive Neural Networks," IEEE Signal Processing Letters, vol. 24, no. 4, 2017, pp. 505-509.
[37] Ibid.

configuration minimizes information loss associated with the LSTM and effectively captures correlations throughout the entire sequence. External elements serve as interfaces for the integration of additional features that extend beyond the established dependencies. The third component is the Conditional Random Field (CRF), used to produce final predictions by considering both the encoded representation (st) and the transitions between labels.

Experiments and Results

Using SemEval-2015 community question answering (cQA) dataset. The evaluation matrices used are macro-averaged Precision, Recall, F1, and Accuracy (Acc.). The model was compared to other methods that tackled answer selection but using other mechanisms. There methods are JAIST, ICRC, and RCNN. JAIST and ICRC, which employed statistical machine learning, are the Top-1 and Top-2 systems in this task, and RCNN uses a typical neural-network. The results are shown in Table I, where the optimal result is marked bold in each column. Three variants based on the architecture were tested, these variants differ in their input for- ms as follows, the concatenation of the question and answers, the answers, and the sequence of question answering (QA) pairs, namely A-ARC I, A-ARC II, and A-ARC III, respectively. The results show that the method that employs deep learning has much stronger learning ability in comparison to JAIST and ICRC, and the proposed global context modelling mechanism is superior to the existing deep context modelling method (RCNN).

Table 1
Comparison on Marco Average Results on Different Models

Model	Precision	Recall	Macro F1	Acc.
RCNN	56.41	56.61	56.14	72.32
ICRC	57.83	56.82	56.41	68.67
JAIST	57.31	57.20	57.19	72.57
A-ARC I	59.83	58.41	58.29	76.42
A-ARC II	60.12	58.22	58.21	76.32
A-ARC III	59.41	58.25	58.35	74.45

Part-of-Speech Tagging

The Part-of-Speech Tagging (POS) process tries to assign every word of the speech for a given text, a tag that categorizes its class according to the given language. The tagging process results in information that is beneficial to NLP in general. It can help understand the text, get the intention of the speaker. It also gives information about the neighbor words. POS has many applications in the domain of NLP such as parsing, translation, text-to-speech and information retrieval. The POS process, in general, relies on using several factors such as the meaning of the word, the position of the word with respect to its adjacent words, known as the context of the word, in addition to other factors that vary according to the natural language itself. The research literature demonstrated the use of several learning techniques to achieve POS. One of the main techniques used was SVM.

[38]and [39]demonstrate the use of SVM for Chinese and Bengali languages. Another approach used extensively for POS was Hidden markov model (HMM). HMM was used for general speech tagging such as in[40,41]. Besides, several special domain applications used HMM for POS such as POS for Twitter text[42].

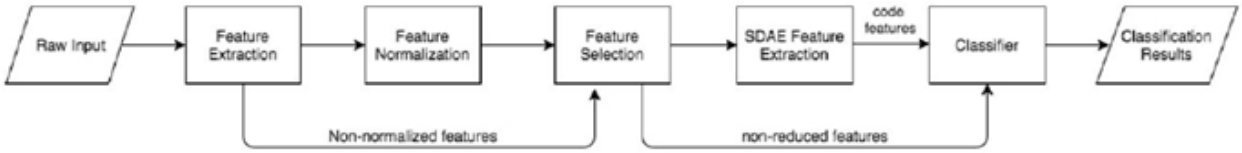

Fig. 5. The framework of the authorship identification[43]

Recently, several research projects applied deep learning for POS [44,45]. Demonstrates the use of DBNs for POS in the Bengali language. Their approach included two steps: Feature Vector Construction, and Classifier design[46]. The feature vector focused on three features: word length, suffix and prefix of the word, and POS of surrounding words. For the classifier, they used DBNs with a structure of three layers. DBNs resulted in 93.33% accuracy which is considered an improvement over other approaches for the same language.

Author Identification

Deep learning techniques were applied to another NLP task which is author identification. Author identification is the task of identifying the author of a given piece of text from a set of candidates, by identifying the writing style of the author. Several features have been used to capture the writing style of a given author, such as lexical feature, syntactic features, and content-specific features. However, character n-grams approach is considered the state-of-the-art feature for author identification. In [20], a variable length character n-gram features were fed to a Stacked denoising autoencoder (SDAE) to present an approach to this task.

[38] A. Ekbal and S. Bandyopadhyay, "Part of speech tagging in bengali using support vector machine," Information Technology, 2008. ICIT08. International Conference on. IEEE, 2008, pp. 106-111.

[39] X. Wang, J. Zhang and Y. Yan, "Support Vector Machine for Chinese Part-Of-Speech Tagging in SpeechSynthesis Systems," International Con- ference on Biomedical Engineering and Computer Science, 2010.

[40] A. Ekbal, S. Mondal, and S. Bandyopadhyay, "Pos tagging using hmm and rule-based chunking," The Proceedings of SPSAL, 2007, pp. 25-28.

[41] A. Paul, B. Purkayastha and S. Sarkar, "Hidden Markov Model based Part of Speech Tagging for Nepali language," International Symposium on Advanced Computing and Communication (ISACC), 2015, pp. 149-156.

[42] S. Sun, H. Liu and H. Lin, "Twitter part-of-speech tagging using pre-classification Hidden Markov model," IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012, pp. 1118-1123.

[43] A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," IEEE International Conference on Machine Learning and Applications (ICMLA), 15th, 2016, pp. 898-903.

[44] X. Zheng, H. Chen, and T. Xu, Deep learning for chinese word segmentation and pos tagging," EMNLP, 2013, pp. 647-657.

[45] Y. Tsuboi, "Neural networks leverage corpus-wide information for part-of-speech tagging," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 938-950.

[46] A. Ekbal, S. Mondal, and S. Bandyopadhyay, "Pos tagging using hmm and rule-based chunking," The Proceedings of SPSAL, 2007, pp. 25-28.

SDAE is a deep learning technique that is considered a part of the generative deep architectures. It is formed by stacking Denoising autoencoders (DAEs), where the output of the encoding layer of the current DAE is used as the input for the next one. The training of SDAE consists of two procedures: unsupervised pre-training then supervised fine- tuning. An overview of the proposed system in [20] is shown in Fig.5, feature extraction component converts the raw data into a vector of features, character n-gram and frequent words are the most widely used features. Afterwards, feature values are either normalized using min max normalization or passed directly to feature selection. Feature selection is a technique that is used to remove redundant features and keep the most relevant features. Then, a higher level feature extraction is preformed using SDAE. The extracted features from SDAE are passed as input to the classifier, which takes another input, that is the selected features without further extraction. The classifier finally produces the classification results.

Experiments used a subset of the Reuters Corpus Volume 1 (RCV1) as a dataset, and the results showed that the system using SDAE for feature extraction has outperformed the state- of-the art of author identification techniques.

Speech Emotion Recognition

Speech is considered one of the main communication methods for humans, as it also includes the emotional state of the speaker. Speech emotion recognition (SER) research has been gaining more attention recently for the variety of its possible applications.

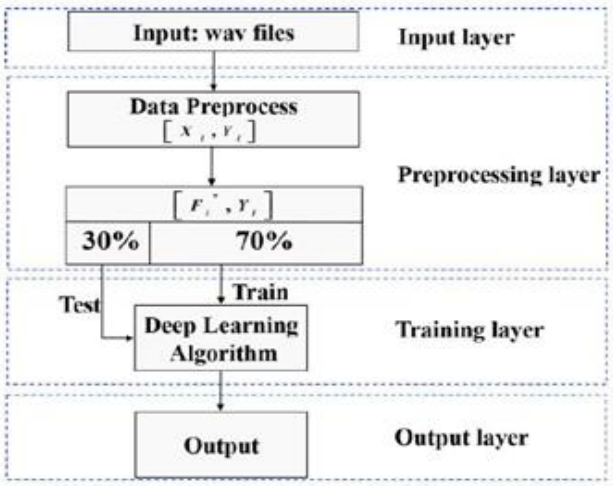

Fig. 6. Overview of the proposed attentive neural network[47].

For instance, it could help in preventing fatigue driving accidents by identifying the emotional state of the driver from real time conversations with the driver, and offer a warning whenever a state of fatigue is recognized. Another application could be emotion management for teenagers, warning the teenager of possible emotional outburst by recognizing the emotional state in the

[47] X. Zhou, J. Guo and R. Bie, "Deep Learning Based Affective Model for Speech Emotion Recognition," Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communi- cations, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences, 2016, pp. 841-846.

collected speech signals. Researchers in their earlier attempts in emotion recognition, applied hand-tuned features into traditional classifiers. Mostly used features in SER are prosodic features, spectral features and acoustic features. Although traditional classifiers were preforming well, they suffered from the complexity of implementing artificial multi-dimension features extraction. In [48]deep learning techniques were introduced for feature extraction to overtake issues of traditional classifiers. The deep learning techniques used are Stacked autoencoder (SAE) and DBN, both of which belongs to the generative deep architectures.

The structure used in [49]to recognize emotion state from speech signals is shown in Fig.6. It contains four layers. The input layer where the input is a speech signal wav file. The preprocessing layer where the sampling and framing of the speech signals is performed. The training layer where SAE and DBN are used for feature extraction. The output layer uses sigmoid function to predicate the emotional state. The experiments used the German Berlin Emotional Speech Database (Emo-DB) as the dataset. Seven kinds of emotion are to be identified, which are anger, boredom, disgust, anxiety, happiness, sadness and neutral state, respectively. The baseline method used raw spectral features and SVM which was introduced by [50]and had accuracy of 22.4%.

Three types of classification experiments were performed. First, for different sizes of training dataset, where SAE and DBN both outperformed the baseline method, obtaining ac- curacy of 25% and 38%, respectively. Second, for different speakers, where deep learning techniques also outperformed the baseline method with accuracy of 29% in SAE, and 39% in DBN. The same result was concluded in the third experiment which presented the classification performance under different numbers of emotion states. From the previous results another conclusion could be driven, which is that DBN performed better than SAE, and that is because DBN has a deeper structure from SAE.

Spoken Language Understanding

Spoken language understanding (SLU) represents one of the new areas that combines NLP and speech processing tasks. It allows language understanding via a computerized approach. SLU allows human to machine communication. Examples of applications that benefit from SLU are phone calls routing, and interactive voice response (IVR). Call routing systems are typically structured into two components: the speech-processing component, and the action classifier. Speech processing techniques focus on obtaining the text corresponding to the speech. SLU focuses more on obtaining the intention out of the text. Hence, SLU can help in the action classifier component.

Different research projects-built action classifiers using several machine learning techniques. Boosting, Maximum Entropy Modeling (MEM), and SVM are machine learning techniques used as action classifiers. In general, these techniques share a mutual concern; they are supervised

[48] X. Zhou, J. Guo and R. Bie, "Deep Learning Based Affective Model for Speech Emotion Recognition," Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communi- cations, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences, 2016, pp. 841-846.

[49] Ibid.

[50] Z. w. Huang, W. t. Xue, and Q. r. Mao, peech emotion recognition with unsupervised feature learning," Frontiers of Information Technology & Electronic Engineering, vol. 16, 2015, pp. 358-366.

learning techniques, which require a large amount of labeled data to produce action classifiers with high accuracy. In many cases, creating such labeled data require major efforts. The problem becomes bigger when dealing with application domains that have many words. Neural networks are also common in solving speech processing and NLP tasks. As stated earlier in the introduction section, traditional neural networks (shallow networks) perform poorly when dealing with complicated problems.

The recent developments added to DBNs allowed it to gain two significant advantages: learn layer by layer efficiently using a greedy algorithm, and learn using unlabeled data, which is plenty in the field of call routing for instance. In their research work [28], they discuss the application of DBNs on call routing problem and compare the results to Boosting, MEM, and SVM machine learning techniques. Their research work shows the use of DBNs in three different ways. First, the use of DBNs solely. Second, they combined DBN along with SVM. Finally, they combined DBNs, SVM, and RBM into one solution where they mix supervised learning with unsupervised learning. The results showed accuracy improvements at different data sizes sets in comparison with Boosting, MEM, and SVM. Besides, the result indicates that DBNs technique produces much better results when the size of labeled data is small.

COMPARISON BETWEEN CNN AND RNN FOR NLP TASKS

Looking at the deep learning techniques employed in the previously presented NLP tasks, we can notice that many tasks were achieved using CNN and RNN, such as sentiment analysis and answer selection. Some tasks used LSTM, while another type of gated RNN called gated recurrent unit (GRU) could also be used in some of NLP tasks. However, choosing the best suited DNN for a NLP task is not easy. In this section, a comparison between CNN and RNN is presented to shed some light on the main differences between the two networks and give some insight on when to use which network. CNN in its hierarchical architecture is supposed to be good at extracting position-invariant features, which makes it a good candidate to be used in tasks such as sentiment analysis. RNN, on the other hand, in its sequential architecture is supposed to be good at modeling units in sequence, making it a good candidate to be used in tasks such as language modeling[51]. But this assumption is not supported in the current NLP literature. For instance, RNNs perform well on document-level sentiment analysis in, and gated CNNs outperform LSTMs on language modeling tasks in[52].

In[53], they presented a systematic comparison between CNN and RNN on a range of NLP tasks including sentiment analysis, relation classification, path query answering, part-of- speech tagging and others. The experiments were carried out on different dataset for each NLP task. Results have shown that, for path query answering and part-of-speech tagging, both GRU and LSTM outperform CNN, which supports the assumption about RNNs. while other results do not support the assumption about CNNs, such that for sentiment analysis and relation classification, where

[51] R. Sarikaya, G. Hinton and A. Deoras, "Application of deep belief networks for natural language understanding," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 4, 2014, pp. 778-784.

[52] Y. N. Dauphin, A. Fan, D. Grangier and M. uli, "Language modeling with gated convolutional networks," arXiv preprint arXiv:1612.0808, 2016.

[53] R. Sarikaya, G. Hinton and A. Deoras, "Application of deep belief networks for natural language understanding," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 4, 2014, pp. 778-784.

GRU outperformed CNN. An explanation for that could be that GRU performs better when the sentiment is determined by the entire sentence rather than some key-phrases. For example, the sentence "it's a movie that you won't want to miss" includes "won't" and "miss" which could make it identified as a negative sentiment, while the whole sentence actually represents a positive sentiment. In summary, choosing the best suited DNN for text classification tasks depends on how it is important to semantically under- stand the whole sequence of the text.

CONCLUSION

Deep learning methods are being used more and more frequently to solve complex problems that are difficult to solve using conventional methods. Promising outcomes were observed when deep learning techniques were applied to NLP tasks. An overview of the three distinct classes of deep learning architectures was provided in this paper. The research projects that used various deep learning networks to complete NLP tasks were also presented in the paper. In handling NLP tasks, all these methods performed better than their state-of-the-art.

Competing Interests

The authors did not declare any known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

A. Ekbal and S. Bandyopadhyay, "Part of speech tagging in bengali using support vector machine," Information Technology, 2008. ICIT08. International Conference on. IEEE, 2008, pp. 106-111.

A. Ekbal, S. Mondal, and S. Bandyopadhyay, "Pos tagging using hmm and rule-based chunking," The Proceedings of SPSAL, 2007, pp. 25-28.

A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," International Conference on Control, Automation and Robotics (ICCAR), 3rd, 2017, pp. 705-710.

A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural Networks," Advances in neural information processing systems, 2012, pp. 1097-1105.

A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," IEEE International Conference on Machine Learning and Applications (ICMLA), 15th, 2016, pp. 898-903.

A. Paul, B. Purkayastha and S. Sarkar, "Hidden Markov Model based Part of Speech Tagging for Nepali language," International Symposium on Advanced Computing and Communication (ISACC), 2015, pp. 149- 156.

A. Saxena,"Convolutional Neural Networks: An Illustrated Explanation," Crossroads the ACM magazine for students, http://xrds.acm.org/blog/2016/06/convolutional-neural-networks-cnns-illustrated-explanation/, 2016, accessed on May 2017.

A. Y. Ng and M. I. Jordan On Discriminative vs. Generative classifiers: A comparison of logistic regression and nave Bayes," Advances in neural information processing systems, vol. 2, 2002, pp. 841-848.

C. Qian, T. He and R. Zhang, "Deep Learning based Authorship Identification,".

E. Kumar "Natural Language Processing," I. K. International Pvt Ltd, 2011.

G. G. Chowdhury, S. Osindero and Y. Teh, "Natural language Processing," Annual review of information science and technology, vol. 37, no. 1, 2003, pp. 51-89.

G. Hinton, S. Osindero and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, 2006, pp. 1527-1554.

H. Ney, "Speech translation: Coupling of recognition and translation," Proc. ICASSP, 1999.

J. Dean et al., "Large Scale Distributed Deep Networks," Advances in neural information processing systems, 2012, pp. 1223-1231.

J. Xu, H. Li and S. Zhou, "An Overview of Deep Generative Models," IETE Technical Review vol. 32, no. 2, 2014, pp. 131-139.

L. Deng "Three classes of deep learning architectures and their applications: a tutorial survey," APSIPA transactions on signal and information processing, 2012.

L. Deng and N. Jaitly, Deep discriminative and generative models for pattern recognition," USENIX-Advanced Computing Systems Association, 2015.

L. Deng and X. Li "Machine learning paradigms in speech recognition: An overview," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, 2013, pp. 1060-1089.

M. F. Kabir, K. Abdullah-Al-Mamun and M. N. Huda, "Deep learning based parts of speech tagger for Bengali," International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 26-29.

M. Nielsen, "Neural Networks and Deep Learning," 2017.

P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," International Joint Conference on Computer Science and Software Engineering (JCSSE), 13th, 2016.

R. S. Dudhabaware and M. S. Madankar, "Review on Natural Language Processing Tasks for Text Documents," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2014.

R. Sarikaya, G. Hinton and A. Deoras, "Application of deep belief networks for natural language understanding," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 4, 2014, pp. 778-784.

R. Sarikaya, G. Hinton and A. Deoras, "Application of deep belief networks for natural language understanding," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 4, 2014, pp. 778-784.

R. Vijayakrishnan et al., "Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record," Journal of cardiac failure, 2014.

S. Sun, H. Liu and H. Lin, "Twitter part-of-speech tagging using pre-classification Hidden Markov model," IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012, pp. 1118-1123.

T. Duyu, B. Qin and T. Liu, "Document modeling with gated recurrent neural network for sentiment classication," In Proceedings of EMNLP, 2015, pp. 1422-1432.

T. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Convolutional neural networks for LVCSR," IEEE International Conference on ICASSP, 2013, pp. 8614-8618.

X. He and L. Deng "Speech recognition, machine translation, and speech translation — A unifying discriminative framework," IEEE Sig. Proc. Magazine, vol. 28, 2011.

X. Wang, J. Zhang and Y. Yan, "Support Vector Machine for Chinese Part-Of-Speech Tagging in SpeechSynthesis Systems," International Con- ference on Biomedical Engineering and Computer Science, 2010.

X. Zhang, J. Zhao and Y. LeCun, "Character-level convolutional networks for text classification," Advances in neural information processing systems, 2015, pp. 649-657.

X. Zheng, H. Chen, and T. Xu, Deep learning for chinese word segmentation and pos tagging," EMNLP, 2013, pp. 647-657.

X. Zhou, J. Guo and R. Bie, "Deep Learning Based Affective Model for Speech Emotion Recognition," Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communi- cations, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences, 2016, pp. 841-846.

Y. LeCun et al., Gradient-based learning applied to document recognition," Proceedings of the IEEE vol. 86, no. 11, 1998, pp. 2278-2324.

Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature vol. 521, 2014, pp. 436-444.

Y. N. Dauphin, A. Fan, D. Grangier and M. uli, "Language modeling with gated convolutional networks," arXiv preprint arXiv:1612.0808, 2016.

Y. Tsuboi, "Neural networks leverage corpus-wide information for part-of-speech tagging," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 938-950.

Y. Xiang, Q. Chen, X. Wang and Y. Qin, "Answer Selection in Community Question Answering via Attentive Neural Networks," IEEE Signal Processing Letters, vol. 24, no. 4, 2017, pp. 505-509.

Z. w. Huang, W. t. Xue, and Q. r. Mao, peech emotion recognition with unsupervised feature learning," Frontiers of Information Technology & Electronic Engineering, vol. 16, 2015, pp. 358-366.