Convolutional Neural Network based for Automatic Text Summarization

Wajdi Homaid Alquliti[1], Norjihan Binti Abdul Ghani[2]
Faculty of Computer Science and Information Technology
University of Malaya
Malaysia

***Abstract*—In recent times, the apps for the processing of a natural language has been formed and generated through the use of intelligent and soft computing methods that allow computer systems to practically mimic practices related to the process of human texts like the detection of plagiarism, determination of the pattern as well as machine translation, Thereafter, Text summarization serves as the procedure of abridging writing within consolidated structures. 'Automatic text summarization' or the ATS is when a computer system is used to create a text summarization. In this study, the researchers have introduced a novel ATS system, i.e., CNN-ATS, which is a convolutional neural network that enables to Automatic text summarization using a text matrix representation. CNN-ATS is a deep learning system that was used to evaluate the improvements resulting from the increase in the depth to determine the better CNN configurations, assess the sentences, and determine the most informative one. Sentences deemed important are extracted for document summarization. The researchers have investigated this novel convolutional network depth for determining its accuracy during the informative sentences selection for each input text document. The experiment findings of the proposed method are based on the Convolutional Neural Network that uses 26 different configurations. It demonstrates that the resulting summaries have the potential to be better compared to other summaries. DUC 2002 served as the data warehouse. Some of the news articles were used as input in this experiment. Through this method, a new matrix representation was utilized for every sentence. The system summaries were examined by using the ROUGE tool kit at 95% confidence intervals, in which results were extracted by employing average recall, F-measure and precision from ROUGE-1, 2, and L.**

***Keywords*—Automatic text summarization; extracts summarization; information retrieval; deep learning; convolutional neural network**

I. INTRODUCTION

Recently, the formation and generation of apps for the processing of a natural language have taken place by using soft and intelligent computing methods that make it possible for computer systems to practically imitate practices associated with processing human texts, such as machine translation, detection of plagiarism, and identification of patterns. Intelligence methods such as genetic-based algorithm, evolution-based algorithm, swarm-based intelligence, fuzzy logic, and neural network are often involved. Thus, a main reason for enabling the mimicking is the utilisation of a precise computer system that performs quicker compared to the performance of individuals. So, automatically summarising texts is one type of natural language application that presumably makes use of such methods to optimise performance.

Literature has a fairly large amount of systems for automatic summarisation. Most of these systems manage the summarisation problem based on the desired kind of summaries. Numerous methods have been formulated to generate single-document summaries. Various methods were also presented, with machine learning being considered as the most visible. In numerous approaches, there is an assumption that the numerical representation of the text and the extracted features are demonstrations of how designing a method that is equipped with a powerful feature can produce a high-quality text summary. Scoring of the features for each sentence is performed in order to produce a summary for the input document. As a result, those chosen features affect the quality of the summary generated. Thus, there is a need to develop a mechanism to automatically obtain and calculate the feature.

The feature extraction stage is vital for data analysis in the NLP and machine learning processes. This step is helpful in identifying the interpretable representation of data for the machines that can enhance the performance of such learning algorithms. Applying unsuitable features could hinder the performance of even the best algorithms. On the other hand, simple techniques can have very good performance if appropriate features are implemented. The feature extraction stage is performed in a manual or unsupervised manner.

In the past few years, the deep learning technology has experienced massive developments. Empirical results have revealed that this is a better technique compared to other ML algorithms. This could be a result of the fact that this technique, like the brain model, copies the functioning of the brain and stacks several neural network layers on top of each other. According to [1], deep learning machines perform better than traditional machine learning tools since the feature extraction method is included. However, the deep learning methods get to know feature hierarchies by utilising features obtained from the higher hierarchical levels as a result of the organisation of low-level features. The learning features found at different levels of abstraction make it possible for the system to learn the complex functions. These complex functions are then responsible for using data to map the input and the resultant output without relying on human-developed features [1].

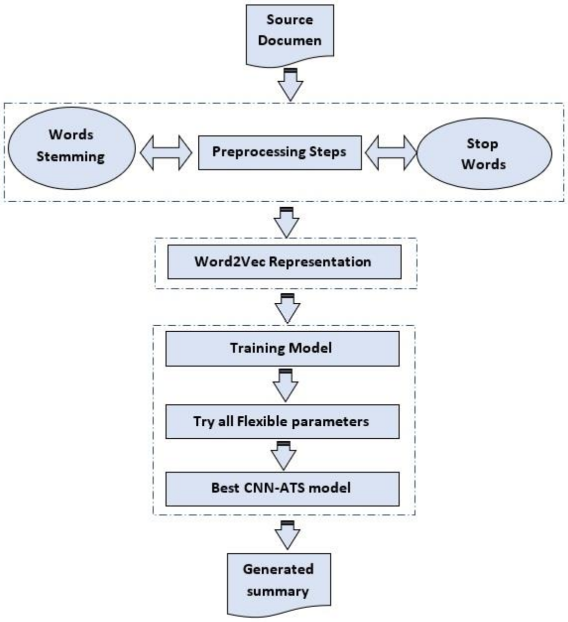

Fig. 1. Depicts the Exhaustive Framework of CNN-ATS Model.

Using deep learning for automatic text summarisation requires high investigation in the text representation. The texts are re-represented in such a way that they correspond with the basics of deep learning convolutional neural network, like the utilisation of the matrix format as a representation of the text structures in CNN architecture. In texts, the informative sentence is an adverse property that influences its capacity of being chosen for summarisation. The sentence's importance is determined based on their words, which have been identified as ontology and mining. Distributed representation makes it possible to build abstract features, which are helpful in choosing and measuring the informative sentence. Here, new sentences patterns that facilitate text mining interpretation were determined by the researchers. The new matrix representation of sentence demonstrated a similarity to the images that the deep learning technique represented. Fig. 1 depicts the exhaustive framework of the CNN-ATS model.

II. Definition and Related Work

A. Related Work

In the past few years, many new natural language processing applications were designed which applied intelligent and soft-computing methods. These applications can help the computational systems mimic all human text-based processing activities like pattern recognition, plagiarism detection and machine translation. Many intelligence techniques like swarm intelligence, genetic algorithms, evolutionary algorithms, neural networks and fuzzy logic have been used. These imitating techniques were used because the computer systems are often more precise than human performance. The automated text summarisation was a natural language application which applied these techniques for optimising its performance.

The text summarisation process includes summarising all texts into their condensed form [2]. The text summarisation conducted by a person is known as "manual text summarisation", however, a computer-assisted text summarisation process is known as "Automatic Text Summarisation" (ATS). In this study, the researchers have investigated the ATS techniques, and have discussed the various ATS processes, input sizes, styles and evaluation processes. The need for text summarisation has been described in Fig. 2.

Many studies used two approaches for investigating the ATS processes, i.e., the abstraction-based and extraction-based summarised techniques. The extraction-based technique generated a summary by choosing (copy-pasting) the significant sentences. All these sentences were later assessed using a scoring mechanism known as "features", wherein every sentence was assigned a different score. The high-scored sentences were then selected as the candidate summarised sentences. On the other hand, the abstraction-based technique summarised the texts by editing the significant text units (phrases or sentences) by appending, removing, segmenting or paraphrasing a few parts of the text units. This technique is more complex than the extraction-based approach [3].

The target text summary could be used and written in the following styles [4]: "indicative summary" or the "informative summary". The indicative summary offers brief information regarding what is present in the primary document which focuses on a specific topic. The summary that is generated is compressed between 5-10% of the primary text. On the other hand, the informative summary assesses a majority of the topics present in the primary text. This type of generated summary comprises of 20-30% of the original document. Furthermore, the ATS researchers have investigated two forms of document input sizes, i.e., a single document and a multi-document summarisation process.

Fig. 2. The Need to use a Text Summarisation Technique.

Based on every document's processing level, all summarisation techniques have been categorised into 3 approaches, surface, entity and discourse [2], [5]. A surface level technique applies a shallow feature set for extracting the relevant sentences from the document and including them in a text summary. The entity level process extracts the entities and derives their relationship in the text document, and models their extraction. For identifying the vital entity-entity relationship, many techniques could be used like the vector space model and the graph-based representation. A Discourse-Level process is based on the modelling of the global text structure and its correlation like the rhetorical text structure (e.g., narrative and argumentation structure), a document format (i.e., hypertext mark-ups or document outlines) and the various topics threads (as and when they get exposed in a text).

Some of the earlier proposed techniques that could be used for summarising the texts included the surface level (i.e., feature-scoring) approaches [6]–[8]. In one study, [6] proposed a novel term-frequency process for highlighting the term-importance in the context, while [7] proposed the sentence position technique for helping the summarizer identify the significance of a sentence in the text document. After 10 years, [8] investigated 2 processes and proposed the feature of some pragmatic words (including cue words like "key", "significant", "idea", etc.

As the feature scoring process helped in deriving significant results, the researchers have proposed many additional features for enhancing the text summarisation quality. Earlier literature showed that the text features process played a vital role in generating many qualified summaries [9], [10]. Hence, many studies enclosed the feature weighting technique for adjusting the feature scores in all summarisation-based issues [11]–[13]. This feature selection technique generated a higher solution quality. Also, the text summary quality was sensitive to the features which determined how the sentences were scored and weighted. Hence, there is a higher need for developing a mechanism which can differentiate the low and high significance features. As a result, several feature selection techniques were developed and proposed, however, there is a need to develop better mechanisms for obtaining good-quality results.

One other issue which must be addressed is related to the investigation of a majority of the text document subtopics. This helped in generating a summary, which can cover many themes in the document. For solving this issue, a cluster-based (or diversity) process can be used for diversifying the sentence selection technique, wherein the selected sentences can cover many topics in the text document. Several processes can be used for implementing the diversity-based approach for text summarisation [14]–[19]. This diversity during summarisation helps in controlling the sentence redundancy, which improves the summary quality.

Therefore, it is important to select a good similarity measure for adjusting the data clustering [20]. In their study, [21] estimated the sentence centrality score by sentence clustering. However, computing this score prevents any technique from determining the relationship between all sentences.

B. Deep learning

Deep learning is believed to drastically enhance the advanced artificial intelligent tasks such as object detection, speech recognition, and machine translation [22]. This technique's deep architectural nature can be used to solve complex artificial intelligence-related problems[23]. Thus, researchers have utilised this method in modern domains for numerous tasks such as object detection and face recognition. Application of this method to numerous language models has also been done. For example, [24] Using spiking deep belief network for Real-time classification and sensor fusion. [22] the recurrent neural networks has been used to denoise the speech signals and [25] stacked autoencoders have been used to determine the cluster pattern during gene expression. Also, they using deep learning methods for toxicity prediction [26]. Another study [27] utilised the neural model to produce images with varying styles. Furthermore, [28] the deep learning technology was used to simultaneously analyse sentiments from several modalities.

The deep learning technology went through massive developments in the past few years. Based on empirical results, it was determined that this technique was better compared to other ML algorithms. This could be a result of the fact that this technique, like the brain model, copies the functioning of the brain and stacks multiple neural network layers on top of each other. The author in [1] stated that the deep learning machines perform better than the conventional ML tools since they also utilise the feature extraction method. However, until now, there is no theoretical background for the deep learning technology. Feature hierarchies are learned by deep learning techniques using features obtained from the higher hierarchical levels, which have been formed through the organisation of the low-level features. The learning features found at the different abstraction levels make it possible for the system to gain an awareness of the complex functions that utilise the data to map the input and the resultant output without relying on the human-developed features [1]. For image recognition systems, the handcrafted features are extracted by the conventional setup extract and fed to the SVM. However, the deep learning technology performs better since it also conducts an optimisation of all the extracted features.

The main difference between deep learning technologies and ML is the difference in their performance when the volume of data increases. When the dataset is smaller, the deep learning method has an inefficient performance since it needs large data volume for proper comprehension [28].

C. Convolutional Neural Network

The convolutional neural network (CNN) is a kind of deep feed-forward network that can be generalised and trained easily compared to other networks that possess connectivity between adjacent layers [29], [30]. CNN has had successful usage when other neural networks were not as popular. Presently, it is being utilised in the computer vision community.

CNNs are formulated for data processing in the form of multiple arrays, such as a grey-scale image composed of 3×2D arrays with varying pixel intensities. Different data modalities are demonstrated as multiple arrays, such as 1D for signals and sequences, including language; 2D for image or audio spectrograms; and 3D for the video or volumetric images. The 4 main ideas that allow CNNs to utilise the features of the natural signals are shared weights, pooling, local connections, and use of multiple layers [29]–[31].

Numerous stages are included in a classic CNN architecture (Fig. 3). The initial stages consist of 2 kinds of layers: i.e. pooling and convolutional layers. Within the convolutional layer, one can organise the layers in the feature maps, where every unit is connected to the feature maps' local patches, which originate from the previous layers, through weights called as the filter bank. The result of the local weighted sum goes through the non-linearity, such as the ReLU [32]. All the units found in the feature map are observed to be sharing one filter bank. The different feature maps found in the layer utilise varying filter banks. This architecture was constructed for 2 purposes. Initially, for array data like images, it was considered that the local groups of values are highly correlated. They also form unique and easily noticeable local motifs. Secondly, the local statistics of other signals or images are considered invariant to the location. Thus, if the motif is observed within a certain section of the image, one may also find it elsewhere. This network therefore depends on the fact that the units found at various locations share the same weights and can therefore be detected through the utilisation of similar patterns from the other segments in the array. Mathematically, discrete convolution is considered as the primary filtering operation that is implemented in the feature maps; therefore, it is named so.

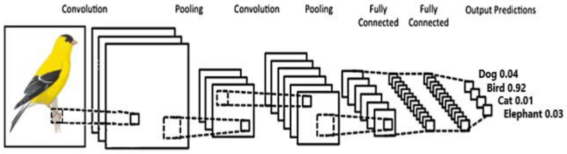

Fig. 3. Architecture of the CNN for Image Classification.

While the convolutional layer uses the earlier layer as the basis for detecting the local combination of the features, the pooling layer combines the semantically similar features into one feature. As a result of these features' relative position, there can be variations in the motif formation. Furthermore, reliably detecting the motif is possible by coarse-graining its position in each feature. The general pooling unit has the ability to calculate a maximal amount of the local patch of units into one feature map [33].

The CNN technique detects the edges based on the raw pixels in Layer 1. Afterwards, it utilises the edges to detect the simple shapes in Layer 2. Then, these shapes are used to detect the simpler shapes in Layer 2. These shapes are also used to determine high-level features, such as the face shape in the higher layers. The last layer is the classifier, which utilises these high-level features [34].

Over the years, a new automatic text summarisation technique possessing a higher degree of accuracy and the ability to automatically summarise any text is being perfected, when compared with the existing one. Developing a new automatic text summarisation technique has gained wide popularity because of its role in text mining for data mining programs. Much evolution is needed for the use of deep learning in automatic text summarisation in representation as text. Re-represented of the text is done that corresponds to the fundamentals pertaining to deep learning convolutional neural network like the utilisation of the matrix format for representing the text structures in the CNN architecture. In text, the informative sentence is regarded as an adverse property, which can impact its potential to be selected for summarisation. The significance of a sentence is determined based on the words, which are identified as ontology and mining. The ability of constructing abstract features is associated with distributed representation, which helps in the measurement and selection of the informative sentence. Here, the researchers have found out new sentence patterns that allow text mining interpretation. A similar representation was demonstrated by the new matrix representation of sentence when compared with the images seen with the deep learning technique.

III. Materials and Methods

Primarily, researchers provided a description of the construction of different experimental benchmarks that were utilized to test the system. Afterwards, they provided a description of the input representation and data encoding system, as well as the design of the deep convolutional network.

A. *Data Gathering and Preparation*

1) Data gathering: Another important part in this research phase is data gathering. Data gathering involves selection of data sets to be used for the purpose of research evaluation. There are important main data sets which we will require for these study for the evaluation our proposed summarization models. To test our summarization methods, we use the Document Understanding Conference (DUC) [15], [35] data collection. The DUC data collection which was created by the National Institute of Standards and Technology (NIST) of U.S. isa standard data set used by most researchers in the area of text summarization. Among the different data collections available in DUC, we have selected the DUC2002 data since it comes with summary extracts\abstracts for multi document articles. The segregation in DUC 2002 document collection is given Table I. For this study, in particular, we will use the document sets reporting on natural disaster events; which includes the document sets: D061j, D062j, D073b, D077b, D079a,D083a, D085d, D089d, D091c, D092c, D097e, D103g, D109h and D115i comprising.

TABLE I. Statistic of DUC 2002 Data Set

Category	**Document Category**
1	Single Natural disaster
2	Single event in any domain
3	Multiple distinct events of single type
4	Bibliographical information about a single individual

2) Data preparation: The step on pre-processing is essential in the range of computational phonetics, since the nature of the acquired outline relies on upon how proficient is the representation of content. In this proposition, few examinations will contain the preprocessing stage. For the most part, this stage will incorporate just two stages: dispensing with stop words and applying stemming as defined in [2].

B. Input Representation

Audio and image processing systems manage rich, high-dimensional datasets that are inputted as vectors of each raw pixel-intensity for image data, such as the power spectral density coefficients that correspond to the audio data. For tasks similar to speech or object recognition, we are aware that all the information needed for the successful performance of the task is encoded within the data (because these tasks can be performed by humans from the raw data). Nevertheless, natural language processing systems customarily consider words to be discrete atomic symbols. Thus, one can represent 'dog' as Id143 and 'cat' as Id537. These are arbitrary encodings, and they do not offer the system any useful information about the possible existing relationships between the individual symbols. This signifies that the model has the ability to leverage a very small portion of what it has gleaned about 'cats' when it is handling data about 'dogs' (i.e. they are both four-legged, animals, pets, etc.). Furthermore, representing words as discrete or unique results in data sparsity. This also typically means that more data may be needed in order to train statistical models successfully. Some of these obstacles can be overcome using vector representations.

Vector space models (VSMs) stand for (embed) words within a continuous vector space. In this space, mapping of semantically similar words is done to nearby points ('are embedded near each other'). VSMs are believed to have a long and rich NLP history. However, all methods rely on the Distributional Hypothesis in one way or another, which states that words share a semantic meaning when they appear in the same contexts.

This distinction is expounded on in greater detail by [36]. In a nutshell, it states that count-based methods calculate the statistics of how frequent a word will co-occur with its neighbor words within a large text. Then, these count-statistics are mapped down to a small, dense vector for every word. Predictive models directly attempt to predict a word from its neighbors based on learned dense and small embedding vectors (considered as the model’s parameters).

In particular, Word2vec is a computationally-efficient predictive model that can be used to learn word embeddings from raw text. There are two flavours, the Skip-Gram model and the Continuous Bag-of-Words model (CBOW). Algorithmically, a declarative example for an input document into the process of Word2vec is demonstrated in Fig. 4.

C. Network Architecture

After all the data are collected, the researchers examined the different model architectures. They considered the default architecture as the convolutional architecture having fully linked layers. Such architecture is suitable for the high-and multi-dimensional data, such as genomic data or 2D images. For evaluating the enhancement resulting from the increase in the depth of CNN-ATS, the researchers used the Krizhevsky principles to design the CNN-ATS layer configurations [29] that can view the source code [37].

In this study, all the CNN-ATS configurations examined are presented in Tables II to V, with one in each column. All further references made towards the configurations will be created depending on their names (A–Z). The configurations adhered to the generic design that was previously described in [29]. They also varied in depth ranged from 1 weight convolutional layer within the A network to 9 weight convolutional layers within the Z network. Both tables gave descriptions of the configurations. Here, the source text document traverses the stack of numerous convolutional (conv.) layers. The researchers utilised 2 different feature map sizes for use in the conv. layer of (3×3) and (5×5) (this is considered a good size for the up/down, left/right, centre) and different amounts of pooling and conv. layers.

Table II provides a description of the combinations of two stacks of pooling and convolutional layers. Table III illustrates the combination of three stacks of pooling and convolutional layers. This combination is helpful for the models since they can benefit from each other and enhance the CNN-ATS configurations’ performance, producing the best selection of sentences for the automatic text summary approach.

The flatten layer comes after the conv. layer (with architectures possessing various depths) and helps turn the 2D matrix data into a vector. This makes it possible to conduct output processing using the fully connected layers, referred to as dense. In the regularisation layer, dropout is used and is configured for the random exclusion of 50% neurons for the decrease in overfitting. The last layer is formed by the soft-max layer [1], [29], [38]. All networks utilise the same fully linked layers configuration.

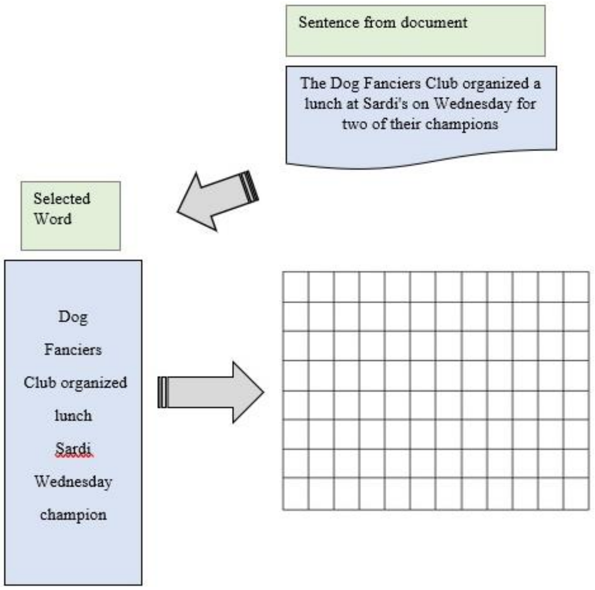
Sentence from document
The Dog Fanciers Club organized a lunch at Sardi's on Wednesday for two of their champions
Selected Word
Dog
Fanciers
Club organized
lunch
Sardi
Wednesday
champion

Fig. 4. A Declarative Example of an input Document into Word2vec Process.

For Tables IV and V, the “Best 2 layers” is considered as the best CNN-ATS configurations among the C to F in Table II. On the other hand, the “Best 3 layers” in Table III is considered the best CNN-ATS configurations from G to N.

Generally, one can conduct target prediction as follows.

The problem involves choosing the best Sentences that can be used for the automatic text summary method. If one can find the given sentence, i, in the document, t, one can encode this information in the binary form, y_{it}, wherein $y_{it} = 0$, for an unimportant sentence, and $y_{it} = 1$, for an important sentence, simultaneously. The training stage uses a standard back-propagation algorithm to determine the CNN and minimise the output layer activation and the cross-entropy of the targets.

D. Summarized Text Generation

The chosen sentence is obtained. Only sentences that possess 1 value of CNN-ATS output were chosen for consideration. Sentences deemed important are extracted for document summarisation. It has been demonstrated that a compression or an extraction rate near 20% of the core textual content is instructive of the contents as the complete text of the document [39]. In the last stage, one can organise the summarising sentences in the order of their conceptual occurrences as observed in the initial text.

E. The Benchmark Methods

The selected methods are standard benchmark methods that have been widely used [15], [35]. They are chosen for comparison purposes at each chapter and classified into five methods:

- Microsoft Word Summarizer.
- Copernic Summarizer (commercial products currently run in the market).
- Best System at DUC2002 Competition.
- Worst System at DUC2002 Competition.
- H2-H1: Human to human Summary.

TABLE II. CNN-ATS CONFIGURATION FOR A-F COLUMNS FOR THE 2 WEIGHT CONVOLUTIONAL LAYERS

CNN-ATS Configuration					
A	**B**	**C**	**D**	**E**	**F**
1 Weight Conv Layers	1 Weight Conv Layers	2 Weight Conv Layers	2 Weight Conv Layers	2 Weight Conv Layers	2 Weight Conv Layers
Input Text matrix					
Conv Layer (3X3)	Conv Layer (5X5)	Conv Layer (3X3) Conv Layer (3X3)	Conv Layer (5X5) Conv Layer (5X5)	Conv Layer (3X3) Conv Layer (5X5)	Conv Layer (5X5) Conv Layer (3X3)
Max Pooling (2X2)					
Flatten					
Dense					
Dropout (0.5)					
Dense					
Softmax					

TABLE III. CNN-ATS CONFIGURATION FOR G-N (COLUMNS) FOR THE 3 WEIGHT CONVOLUTIONAL LAYERS

CNN-ATS Configuration							
G	**H**	**I**	**J**	**K**	**L**	**M**	**N**
3 Weight Layers	3 Weight Layers	3 Weight Layers	3 Weight Layers	3 Weight Layers	3 Weight Layers	3 Weight Layers	3 Weight Layers
Input Text matrix							
Conv Layer (5X5) Conv Layer (5X5) Conv Layer (5X5)	Conv Layer (5X5) Conv Layer (5X5) Conv Layer (3X3)	Conv Layer (5X5) Conv Layer (3X3) Conv Layer (5X5)	Conv Layer (5X5) Conv Layer (3X3) Conv Layer (3X3)	Conv Layer (3X3) Conv Layer (5X5) Conv Layer (5X5)	Conv Layer (3X3) Conv Layer (5X5) Conv Layer (3X3)	Conv Layer (3X3) Conv Layer (3X3) Conv Layer (5X5)	Conv Layer (3X3) Conv Layer (3X3) Conv Layer (3X3)
Max Pooling (2X2)							
Flatten							
Dense							
Dropout (0.5)							
Dense							
Softmax							

TABLE IV. CNN-ATS CONFIGURATION FOR O-R COLUMNS FOR THE 4-6 WEIGHT CONVOLUTIONAL LAYERS

CNN-ATS Configuration			
O	**P**	**Q**	**R**
4 Weight Layers	5 Weight Layers	5 Weight Layers	6 Weight Layers
Input Text matrix			
Best 2 layers	Best 2 layers	Best 3 layers	Best 3 layers
Max Pooling (2X2)			
Best 2 layers	Best 3 layers	Best 2 layers	Best 3 layers
Max Pooling (2X2)			
Flatten			
Dense			
Dropout (0.5)			
Dense			
Softmax			

TABLE V. CNN-ATS CONFIGURATION FOR S-Z COLUMNS FOR THE 6-9 WEIGHT CONVOLUTIONAL LAYERS

CNN-ATS Configuration							
S	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
6 Weight Layers	7 Weight Layers	7 Weight Layers	8 Weight Layers	7 Weight Layers	8 Weight Layers	8 Weight Layers	9 Weight Layers
Input Text matrix							
Best 2 layers	Best 2 layers	Best 2 layers	Best 2 layers	Best 3 layers	Best 3 layers	Best 3 layers	Best 3 layers
Max Pooling (2X2)							
Best 2 layers	Best 2 layers	Best 3 layers	Best 3 layers	Best 2 layers	Best 2 layers	Best 3 layers	Best 3 layers
Max Pooling (2X2)							
Best 2 layers	Best 3 layers	Best 2 layers	Best 3 layers	Best 2 layers	Best 3 layers	Best 2 layers	Best 3 layers
Max Pooling (2X2)							
Flatten							
Dense							
Dropout (0.5)							
Dense							
Softmax							

IV. RESULTS AND DISCUSSION

To test the application of the suggested CNN-ATS based approach for single-document extractive summarisation, 100 articles/documents were taken from the DUC dataset (DUC, 2002) [15], [35]. The standard corpus is used widely in text summarisation studies, which has documents and human model summaries. Firstly, pre-processing is conducted on the collection of documents. In this step, sentence splitting, stop words elimination, tokenisation, and word stemming are all involved. Once the documents complete the pre-processing process, word2vec input representation is applied in order for every sentence to be presented as matrix. Then, the CNN-ATS model is applied for testing and training. Sentences with value 1 were chosen for the output prediction in the suggested model. Lastly, the chosen sentences were chosen as a summary of the main text with the compression rate (20%) as the basis. We used three pyramid assessment measures –precision, mean coverage score (recall), and F-measure to assess our proposed approach. This metric evaluates the system summary's quality by making a comparison with human model summaries and other systems for benchmark summarisation.

To make a comparison of the proposed approach's performance, various comparison configurations models were set up. First, the results for the configurations that are presented in Tables II and III are compared by making a comparison of the F-measure value taken by ROUGE-1. The Best 2 and Best 3 configurations were also selected. Then, the results between the various configurations presented in Tables IV and V were compared by comparing the F-measure value obtained by ROUGE-1. They indicate that the configurations can generate a better summary.

Afterwards, the best configurations results were compared with five benchmark summarisers using three evaluation measures as the basis - Precision, Recall, and F-measure. The five benchmark summarisers were best automatic summarisation system in DUC2002, Copernic summariser, worst automatic systems in DUC2002, Microsoft Word 2007 summariser, and the average of human model summaries (Models) H1:H2.

The proposed code was applied in the Theano [40], which refers to a public deep learning software that uses the Keras as the basis [41]. All layers within the deep network went through simultaneous initialisation with the ADADELTA[42]. Training of the complete network was done using the Dell Precision T1700 CPU system equipped with a 6 GB memory. Two weeks were needed to test and train the deep network.

F. Evaluation Measures

When creating and updating a system for summarisation, it is important to have a method or a tool to monitor the system performance and the modifications therein. For the summarisation process, there are two primary types of assessments: intrinsic as well as extrinsic. Intrinsic evaluations are utilised for assessing the summaries' quality. They may be helpful in answering queries regarding a summary like its grammar, coherence and whether it indicates incorrect knowledge deductions with reference to the original text, or repeated information within its summary itself. Alternatively, extrinsic assessment can be helpful in answering if the summary fulfils its designated purpose. For example, it helps determine if a summary replaces the original text well and conveys the most significant information.

The judgments of the intrinsic assessments are based on the summary output. Human intrinsic assessments measure a summary's cohesion, clarity and informativeness [43]. Automatic intrinsic assessments compare the summaries produced by the systems with those produced by humans. BLEU, ROUGE, Precision/Recall and Pyramid are some of the primary intrinsic tools. The Pyramid needs manual annotation of system-generated summaries and human-generated summaries prior to their comparison. BLEU, Precision/Recall and ROUGE, conversely, are completely automated and only need reference summaries.

F-Measure, Precision and Recall are some of the simplest assessment techniques present that measure the summary relevance with reference to the significance of the sentences it contains. Precision (P) is the quantity of sentences occurring in both the human and the system generated summaries divided by the quantity of sentences present in the system generated summary. Recall (R) is the quantity of sentences present in both the human and system summaries divided by the quantity of the sentences in the human summary. F-Score is a combination integrating both P and R [5]. The F-Score can be calculated with the following formula:

$$F = \frac{(1 + \beta^2)PR}{\beta^k P + R}$$

Where β represents a weighting variable, which is adjustable to influence precision and recall.

The ROUGE (Recall Oriented Understudy for Gisting Evaluation) was introduced in 2004 [44] in order to solve the drawbacks of BLEU at ISI (Information Science Institute). It is approximately based on BLEU; however, it focuses instead on recall. Moreover, it quantifies overlapping of words in sequences and was discovered to correlate in a better way with human assessments compared to several other systems.

Many variants of ROUGE have been recommended [10]:

- ROUGE-N: counts contiguous n-gram. N ranges from 1 to 4.
- ROUGE-L: Longest Common Subsequence (LCS) based metric.

G. Results

In this research, CNN-ATS was proposed by the researchers. This is a novel-based approach that can be used for single-document extractive summarisation. CNN-ATS is considered a convolutional neural network, which has a new text matrix representation. It is also utilised for automatic text summarisation. It is also a deep learning system that integrates the information about the words and the concept about them. Hence, a comparison of the proposed CNN-ATS technique and 5 other benchmark summarisers was done.

This new deep learning system was evaluated in terms of its computing F-measure obtained through ROUGE-1 for all the different CNN-ATS configurations using the procedure stated in the Subsection 3.5. All results gathered from the A, B, C, D, E and F models have undergone comparison using the boxplot technique. Then, the models that demonstrated the best configurations were referred to as the Best 2 layer models. Then, the analysis results that were gathered from the G, H, I, J, K, L, M and N models were compared in terms of their boxplot data. The best 3-layer model was then chosen. This refers to Stage 1 of the result analysis. A more complete description is given below. Meanwhile, the researchers in Stage 2 evaluated the results' prediction accuracy for the CNN-ATS configuration models corresponding to O, P, Q, R, S, T, U, V, W, X, Y and Z. Their boxplot results were then compared to determine the best CNN-ATS architecture. During Stage 3, the researchers conducted a comparison of the results for best configurations that were gathered from the previous 2 stages with the 5 standard benchmark summarisers.

1) Stage 1: During Stage 1, the researchers conducted an evaluation and a comparison of the F-measure values for 100 documents summaries found in the DUC 2002 dataset. Fig. 5 contains the values for the comparison of F-measure during the Stage 1 experiments that included the CNN-ATS A to N configurations.

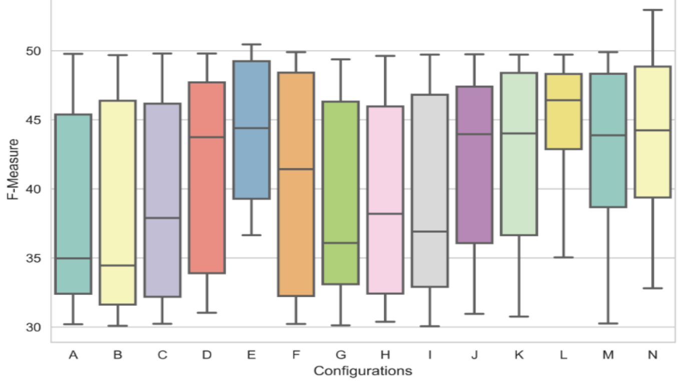

Fig. 5. Comparison of F-Measure Values for the CNN-ATS A to N Configurations Model using Boxplot.

As demonstrated in the figure, the 14 models exhibited a distinct difference in their mean values. Furthermore, the E configuration model gives a mean value of 43.85, while the L model has the best accuracy value of 46.52 (in Stage 1). A smaller variance value of 13.81 was observed in the E model. Therefore, it is considered better compared to the A, B, C, D and F models. The results are also indicative of the fact that E can be considered the best 2-layer model and may be utilised in the CNN-ATS O to Z model configurations to enhance the outcomes of the comprehensive convolutional neural network. The L model was found to have a variance value of 14.67, which indicates that L is the best among the 3-layered models and that it can be utilised further in the CNN-ATS O to Z model configurations to enhance the results of the comprehensive convolutional neural network.

2) Stage 2: During Stage 2, the researchers conducted an analysis of the values of the F-measure for the O, P, Q, R, S, T, U, V, X, Y and Z models of the configurations given in Tables IV and V. Table IV gives the combinations for the 2 stacks of the pooling and convolutional layers. Here, the results from Stages 1 and 2 were utilised by the researchers, where it was seen that the E model had the best 2-layer configuration. On the other hand, the L model was discovered to have the best 3 layered configuration. Table V offers a description of the combination of 3 stacks of the pooling and convolutional layers using the same best configuration models that were gathered for the 2 and 3 layers. Fig. 6 illustrates the comparison of the values of the F-measure using Boxplot for the 11 configurations.

Based on the Boxplot results, one can observe that the O, P, Q models had the best F-measure values equal to 35.07, 36.25 and 36.98, respectively. These values are indicators that the models having 2 stacks of pooling and convolutional layers performed better than the models that have 3 stacks of pooling and convolutional layers. Moreover, the O, P, and Q models exhibited the lowest variance in comparison to the other models. This fact emphasises that the O, P, and Q models can be considered as the combination models in Stage 2.

3) Stage 3: During the final Stage 3, the researchers conducted a comparison of all the best 6 configuration outcomes that were gathered from the earlier 2 stages in order to determine the best CNN-ATS architecture. Moreover, the researchers made a comparison of the best CNN-ATS architecture with the 5-standard and the 3-standard benchmark summarisers – Copernic summariser, Microsoft Word 2007 summariser, the best automatic summarisation system in DUC 2002, the average of human model summaries (Models) H1:H2, and the worst automatic systems in DUC 2002. The basis of the comparison was the three evaluation measures - precision, recall, and F-measure based on the three types: ROUGE-1, ROUGE-2 and ROUGE-L. Fig. 7 shows the outcomes of the comparison performed with the E, L, N, P and Q configuration models. It was observed that model L performed the best and had a corresponding F-measure with a value of 46.52. Moreover, the E N, P and Q models had a larger variance, which indicates that L can be considered the best combination model based on the various experiments.

As is apparent from all figures and stages in the earlier experiment, a CNN-ATS model L is considered as the best model configuration for generating an improved summary. For comparative evaluation, Table VI presents the average precision, mean coverage score (recall), and average F-measure gathered from the DUC 2002 dataset for the proposed approach using five benchmark summarisers: Copernic summariser, Microsoft Word 2007 summariser, the best automatic summarisation system in DUC, the average of human model summaries (Models), and the worst automatic systems in DUC 2002 using ROUGE-1, ROUGE-2 and ROUGE-L, respectively.

As presented in Fig. 8, 9 and 10, the researchers gave the outcome of the boxplot test after a comparison of the F-Measure, precision and Recall values, respectively was performed for the CNN-ATS, Copernics, MS Word, worst system, best system, and human model summaries H2:H1. Furthermore, one can observe that the CNN-ATS algorithm exhibited a good recall value and F-Measure precision. They also noted a larger variance between the other methods compared to the CNN-ATS algorithm. This emphasises the superiority in the F-Measure values that were observed in the algorithms.

TABLE VI. COMPARISON OF SINGLE EXTRACTIVE DOCUMENT SUMMARIZATION USING ROUGE-1, ROUGE-1 AND ROUGE-L RESULT AT THE %95 CONFIDENCE INTERVAL

ROUGE Model	Evaluation Measure	MS-Word	Copernic	Best-System	Worst-System	H2:H1	CNN-ATS
ROUGE-1	precision	0.47705	0.46144	0.50244	0.06705	0.51656	0.50325
	Recall	0.40325	0.41969	0.40259	0.68331	0.51642	0.50492
	F-Measure	0.42888	0.43611	0.43642	0.1209	0.51627	0.50379
ROUGE-2	precision	0.22138	0.19336	0.24516	0.38344	0.23417	0.28718
	Recall	0.17441	0.17084	0.1842	0.03417	0.23394	0.28896
	F-Measure	0.19041	0.17947	0.20417	0.06204	0.23395	0.28862
ROUGE-L	precision	0.44709	0.29031	0.46677	0.66374	0.484	0.48465
	Recall	0.36368	0.25986	0.37233	0.06536	0.48389	0.48397
	F-Measure	0.39263	0.27177	0.40416	0.11781	0.48374	0.48423

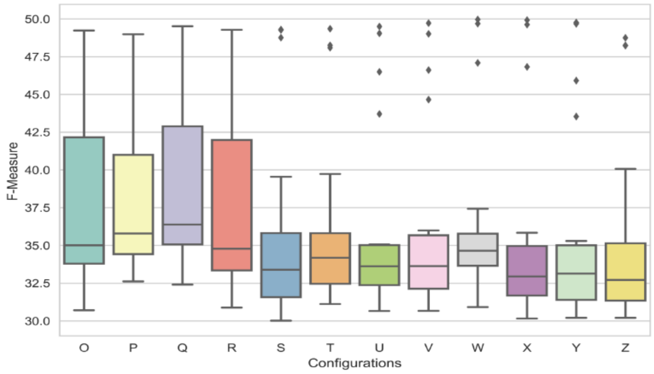

Fig. 6. Comparison of F-Measure Values for the CNN-ATS Models: O, P, Q, R, S, T, U, V, X, Y and Z Models Configurations Model using Boxplot.

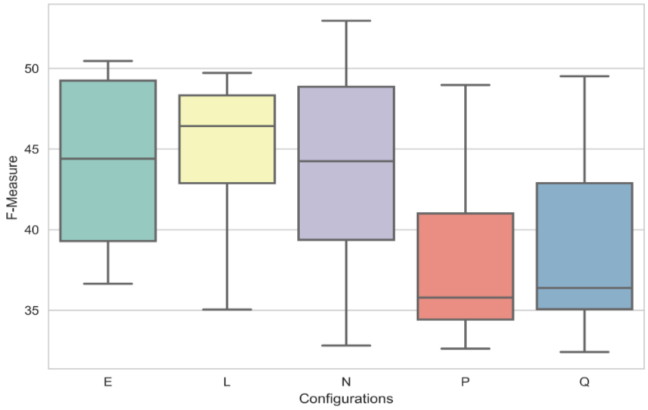

Fig. 7. Comparison of F-Measure Values for the CNN-ATS Models: E, L, N, P and Q Models Configurations Model using Boxplot.

The experiment findings of the proposed method are based on the Convolutional Neural Network that uses 26 different configurations. It demonstrates that the resulting summaries have the potential to be better compared to other summaries. DUC 2002 served as the data warehouse. Some of the news articles were used as input in this experiment. It utilised three pyramid evaluation metrics (average precision, mean coverage score (recall), and average F-measure) to comparatively evaluate the proposed approach as well as other summarisation systems. Through this method, a new matrix representation was utilised for every sentence. It was also used to evaluate the improvements resulting from the increase in the CNN-ATS depth to determine the better CNN configurations, assess the sentences, and determine the most informative one. The chosen sentences were then utilised to establish the summary. The sentences' scoring process was done based on the prediction output values of the CNN network for every sentence.

Based on the results given in Fig. 8, one can observe that F-Measure that uses configuration L produces better summarisation results compared to other configurations. Based on the experimental results of the proposed approach, it can be said that determining a good CNN configuration for text summarisation and utilising the word2vec techniques generates a good summary. Furthermore, CNN-ATS can be considered as the best automatic text summarisation among the six benchmarked methods. This can then be used to create the ideal summary and compare it to the human summary.

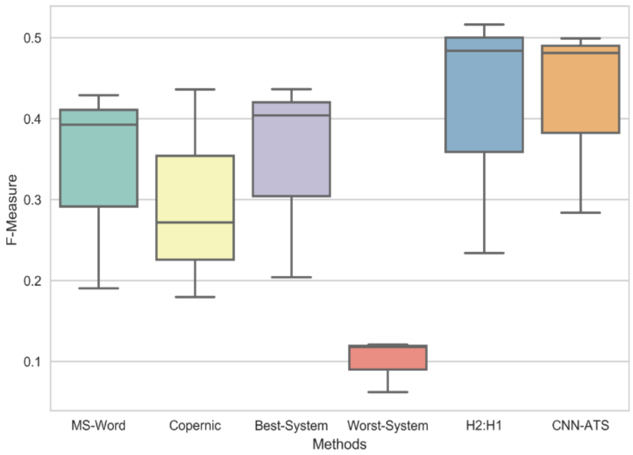

Fig. 8. Comparison of F-Measure Values for the CNN-ATS, MS Word, Copernics, Best System, Worst-System and Human Model Summaries H2:H1.

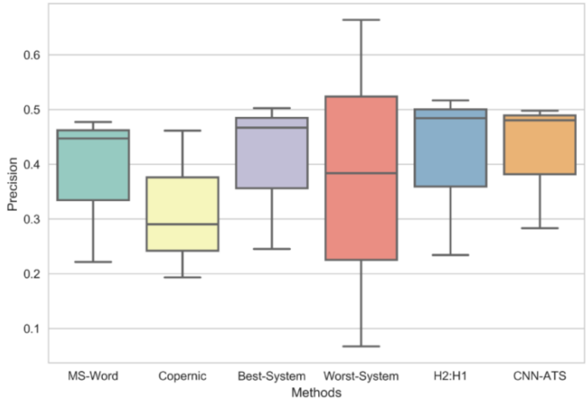

Fig. 9. Comparison of Precision Values for the CNN-ATS, MS Word, Copernics, Best System, Worst System and Human Model Summaries H2:H1.

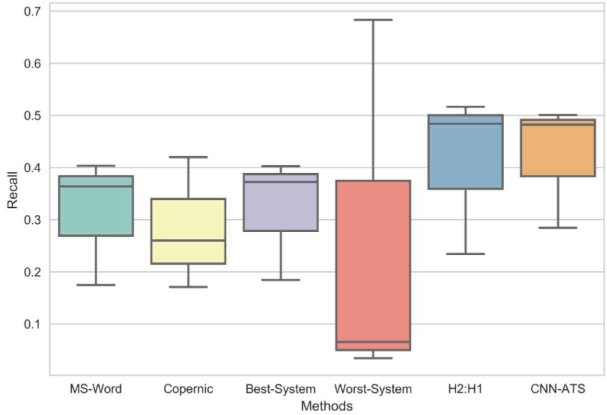

Fig. 10. Comparison of Recall Values for the CNN-ATS, MS Word, Copernics, Best System, Worst System and Human Model Summaries H2:H1.

V. CONCLUSION

To summarise, this chapter presented the utilisation of word2vec for building matrix representations that can be used to determine the summary of the single documents. The convolutional neural network will be the basis. Training and testing of this model was performed using a collection of 100 documents obtained from the DUC2002 dataset. A total of 26 different CNN configurations were used to identify the best architecture that will predict the informative sentences. The researchers examined the deep convolutional networks (possessing up to 9 weight layers) that will predict the informative sentences. They showed that for better selection accuracy, there is a corresponding lower representation depth. The influence that CNN depth has on the summarisation task is investigated. The study then used the selected sentences to group and create text summaries. The outcomes of the proposed summariser in this research were compared with various summarisers, such as Copernic summariser, Microsoft Word 2007 summariser, worst system, and best system. The ROUGE tool kit was then utilised for assessing the system summaries at 95% confidence intervals. The results were extracted using average recall, precision, and F-measure from ROUGE-1, 2, and L, respectively. The F-measure served as a selection criterion since it is a balance of the precision and the recall for the results of the system. The results indicate that the best average precision, recall and F-Measure are generated by our proposed methodology.

VI. Availability of Data and Materials

The dataset is available for download [45]. ROUGE Library [44]. NLP Library [46]. Boxplot library [47]. Convolutional Neural Networks [48].

Acknowledgment

This research was supported by the University of Malaya (UM). I'm thankful to my supervisor Prof. Norjihan Binti Abdul Ghani who provided expertise that greatly assisted the research, although they may not agree with all of the interpretations provided in this paper.

References

[1] Wang and B. Raj, "On the Origin of Deep Learning," Arxiv, pp. 1–72, 2017.

[2] H. Saggion and T. Poibeau, "Automatic Text Summarization: Past, Present and Future," Multi-source, Multiling. Inf. Extr. Summ. Springer, pp.3-13, pp. 3–13, 2016.

[3] G. Armano, A. Giuliani, and E. Vargiu, "Studying the impact of text summarization on contextual advertising," Proc. - Int. Work. Database Expert Syst. Appl. DEXA, pp. 172–176, 2011.

[4] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, "A comprehensive survey on text summarization systems," Proc. 2009 2nd Int. Conf. Comput. Sci. Its Appl. CSA 2009, 2009.

[5] I. Mani, M. T. Maybury, and M. Sanderson, "Advances in Automatic Text Summarization," Comput. Linguist., vol. 26, no. 2, pp. 280–281, 1999.

[6] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM J. Res. Dev., vol. 2, no. 2, pp. 159–165, 1958.

[7] P. B. Baxendale, "Man-made index for technical literature - an experiment," I.B.M. J. Res. Dev., vol. 2, no. 4, pp. 354–361, 1958.

[8] H. P. Edmundson, "New methods in automatic extracting," J. Assoc. Comput. Mach., vol. 16, no. 2, pp. 264–285, 1969.

[9] R. Ferreira et al., "Assessing sentence scoring techniques for extractive text summarization," Expert Syst. Appl., vol. 40, no. 14, pp. 5755–5764, 2013.

[10] M. M. Haque, S. Pervin, and and Zerina Begum, "Literature Review of Automatic Single Document Text Summarization Using NLP," Int. J. Innov. Appl. Stud., vol. 3, no. 3, pp. 857–865, 2013.

[11] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," Comput. Speech Lang., vol. 23, no. 1, pp. 126–144, 2009.

[12] M. S. Binwahlan, N. Salim, and L. Suanmali, "Swarm Based Features Selection for Text Summarization," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 9, no. 1, pp. 175–179, 2009.

[13] L. Suanmali, N. Salim, and M. S. Binwahlan, "Genetic algorithm based sentence extraction for text summarization," 2011.

[14] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '98, pp. 335–336, 1998.

[15] K. Filippova, M. Mieskes, and V. Nastase, "Cascaded Filtering for Topic-Driven Multi-Document Summarization," Proc. Doc. Underst. Conf., pp. 30–35, 2007.

[16] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," SIGIR'01, Sept. 9-12,2001,New Orleans, Louisiana, USA, 2001.

[17] W. Kraaij, M. Spitters, and M. H. der Van, "Combining a mixture language model and naive bayes for multi-document summarisation," SIGIR2001 Work. Text Summ., pp. 1–10, 2001.

[18] J. Steinberger and K. Ježek, "Text Summarization: An Old Challenge and New Approaches," Found. Comput. Intell., vol. 206, pp. 127–149, 2009.

[19] M. S. Binwahlan, N. Salim, and L. Suanmali, "Swarm Diversity Based Text Summarization," ICONIP 2009, pp. 216–225, 2009.

[20] R. L. Cilibrasi and P. M. B. Vitányi, "The Google similarity distance," IEEE Trans. Knowl. Data Eng., vol. 19, no. 3, pp. 370–383, 2007.

[21] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "Multiple documents summarization based on evolutionary optimization algorithm," Expert Syst. Appl., vol. 40, no. 5, 2013.

[22] Y. LeCun, B. Yoshua, and H. Geoffrey, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[23] Y. Bengio, Learning Deep Architectures for AI, vol. 2, no. 1. 2009.

[24] P. O'Connor, D. Neil, S. C. Liu, T. Delbruck, and M. Pfeiffer, "Real-time classification and sensor fusion with a spiking deep belief network," Front. Neurosci., vol. 7, no. 7 OCT, pp. 1–13, 2013.

[25] A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," 2015 IEEE Int. Conf. Bioinforma. Biomed., pp. 1328–1335, 2015.

[26] T. Unterthiner, A. Mayr, G. Klambauer, and S. Hochreiter, "Toxicity Prediction using Deep Learning," Front. Environ. Sci., vol. 3, no. February, 2015.

[27] L. a Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," arXiv Prepr., pp. 1–16, 2015.

[28] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis," vol. 1, 2016.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., pp. 1–9, 2012.

[30] T. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 8614–8618, 2013.

[31] M. Atzori, M. Cognolato, and H. Müller, "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," Front. Neurorobot., vol. 10, no. SEP, pp. 1–10, 2016.

[32] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," Proc. 27th Int. Conf. Mach. Learn., no. 3, pp. 807–814, 2010.

[33] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," Front. Plant Sci., vol. 7, no. September, pp. 1–10, 2016.

[34] Chen, Yu-Hsin and Krishna, Tushar and Emer, Joel and Sze, Vivienne, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in IEEE International Solid-State Circuits Conference, ISSCC 2016, Digest of Technical Papers, 2016, pp. 262–263.

[35] W. Kraaij, M. Spitters, and A. Hulth, "Headline extraction based on a combination of uni-and multidocument summarization techniques," in Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002). ACL, 2002.

[36] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap., pp. 238–247, 2014.

[37] V. GUPTA, "Image Classification using Convolutional Neural Networks in Keras," 2017. [Online]. Available: https://www.learnopencv.com/image-classification-using-convolutional-neural-networks-in-keras/.

[38] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," Mol. Syst. Biol., vol. 12, no. 7, pp. 1–16, 2016.

[39] A. H. Morris, G. M. Kasper, and D. A. Adams, "The effect and limitations of automatic text condensing on reading comprehension performance," Inf. Syst. Res., vol. 3, no. 1, pp. 17–35, 1992.

[40] F. Bastien et al., "Theano: new features and speed improvements," pp. 1–10, 2012.

[41] F. Chollet, "Keras Documentation," Keras.Io, 2015.

[42] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," 2012.

[43] U. Hahn and I. Mani, "Challenges of automatic summarization," Computer (Long. Beach. Calif)., vol. 33, no. 11, pp. 29–36, 2000.

[44] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," Proc. Work. text Summ. branches out (WAS 2004), no. 1, pp. 25–26, 2004.

[45] "Document Understanding Conferences DUC 2002 Dataset." [Online]. Available: https://duc.nist.gov/data.html.

[46] R. Rehurek and P. Sojka, "Gensim - Statistical Semantics in Python," EuroScipy, vol. 6611, no. May 2010, p. 25.–28. 8. 2011, 2011.

[47] "boxplot python library." [Online]. Available: https://seaborn.pydata.org/generated/seaborn.boxplot.html.

[48] "Convolutional_neural_network." [Online]. Available: https://github.com/llSourcell/Convolutional_neural_network/blob/master/convolutional_network_tutorial.ipynb.