Automatic Abstraction Rating of Research Papers using Hierarchical Convolutional Neural Network

BY

Irina Tabassum

ID: 152-15-553

Humaira Rahaman

ID: 152-15-573

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Mahfujur Rahman

Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Tanvir Ahmed

Lecturer

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

APPROVAL

This Project titled "Automatic Abstraction Rating of Research Papers using Hierarchical Convolutional Neural Network", submitted by Irina Tabassum, ID No: 152-15-553 and Humaira Rahaman, ID No: 152-15-573 to the dept CSE has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Dr. Syed Akhter Hossain   **Chairman**

Professor and Head

Dept. CSE,DIU

Dr. S M Aminul Haque   **Internal Examiner**

Assistant Professor and Associate Head

Dept. of CSE, DIU

DECLARATION

We hereby declare that, this project has been done by us under the supervision of Md. Mahfujur Rahman, Lecturer, Department of CSE Daffodil International University.

Supervised by:

_____

Md. Mahfujur Rahman

Lecturer

Dept. of CSE

DIU

Co-Supervised by:

_____

Tanvir Ahmed

Lecturer

Dept of CSE

DIU

Submitted by:

_____

 Irina Tabassum

ID: 152-15-553

Dept of CSE

DIU

_____

 Humaira Rahaman

ID: 152-15-573

Dept. of CSE

DIU

ACKNOWLEDGEMENT

At first we thankful to Allah without the help of Allah we will never complete our final project.

We really grateful and wish our profound indebtedness to Md. Mahfujur Rahman, Lecturer, Department of CSE, Daffodil International University, Dhaka. He gives us a chance to work with him in machine learning project.

I take this opportunity to express gratitude to all of the DepAt first, we are thankful and grateful to almighty ALLAH for his kindness to enable us to fulfill this research effectively.
Department faculty members for their help and support. I also thank my parents for unceasing encouragement, support, and attention. I am also grateful to my partner who supported me through this venture.

We are also thankful to our be loving university DIU, Permanent Campus for providing an excellent   natural   environment.

ABSTRACT

In every there are many papers submitted to publishers. In that time, reviewer review that papers by reading and then give an output like accepted or rejected. Reviewers read every paper one by one and give a review about that paper is a time consuming. This research only worked for reviewing the abstraction of a paper. If we review an abstraction by using a model it will be very effective for saving time. In this research work, we discuss classification algorithms from machine learning algorithms for finding better result. We make a data set for our model from different review system. Many kinds of models or algorithms are used for this work. The most important usable algorithms are modularized Hierarchical Convolutional neural network and another is an attention-based convolutional neural network. We combine this two algorithms and named as HCNN for reviewing abstract of papers.

Contents

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In every day there are many papers submitted to publishers. In that time paper reviewer review that papers by reading and then give an output like accepted or rejected. So in this paper, we work for an automatic review that paper-based is it accepted or rejected. It matters the accuracy rate from a paper. we use some algorithms from data mining and machine learning for this . For tasting that algorithms we make a data set and divide that data set in two-part one is for tasting and another is trying. If the testing is passed then it's accepted. So in short times review, many papers. In machine learning, we use that are working to find out the accepts and rejects a paper with proper algorithms.

1.2 Motivation

Automatic paper reviewer reviews the papers abstract in an automatic way that uses some of the algorithms in data mining and machine learning that is read the data and get the proper results. If we us it we save our time so in those times we read many papers in proper data set.so our aim is reviewing the paper automatically using machine learning. And hope that this analysis will be very helpful for the paper review.In every day there are many papers submitted to publishers. In that time paper reviewer review that papers by reading and then give an output like accepted or rejected. So in this paper, we work for an automatic review that paper-based is it accepted or rejected. It matters the accuracy rate from a paper. we use some algorithms from data mining and machine learning for this . For tasting that algorithms we make a data set and divide that data set in two-part one is for tasting and another is trying. If the testing is passed then it's accepted. So in short times review, many papers. In machine learning, we use that are working to find out the accepts and rejects a paper with proper algorithms.

1.3 Rational of the Study

In this current day, everything is coming to our hand in a click. In this situation, we wait weeks and weeks for confirmation of acceptance or researched. For using this model AAR we got our output very soon.on the other hand, there are many paper reviewer work for reviewing paper that's a boring thing. for using this model we are come out that boring problem.

1.4 Research Questions

1. How to read the paper in automatic?
2. How to find out the acceptance of the paper?
3. How to get the results properly in times?
4. How to read the title and abstract and understand the results automatically?
5. When the results are right or wrong?
6. What is the model working process when it is read?

1.5 Expected Output

We are trying to propose the paper to an automatic academic paper rating that automatically determines academic paper for the acceptance. We also use the novel modularized hierarchical CNN to make use of the structure of the paper. And the experimental results with a large margin show the outperforms baseline. Then we find out the abstract part for the working and the outcomes to identify the most source part for the result.so ultimately the expected output is to read the abstract and the title part and identify if it is right or wrong by the CNN and the others. To read the title and abstract identify if it is right or wrong and gives the proper results.

We are trying to propose the paper to an automatic academic paper rating that automatically determines academic paper for the acceptance. We also use the novel modularized hierarchical CNN to make use of the structure of the paper. And the experimental results with a large margin show the outperforms baseline. Then we find out the abstract part for the working and the outcomes

to identify the most source part for the result.so ultimately the expected output is to read the abstract and the title part and identify if it is right or wrong by the CNN and the others. To read the title and abstract identify if it is right or wrong and gives the proper results.

1.6 Report Layout Chapter

1. Background
2. Research methodology
3. Experimental results and discussion.
4. Conclusion and implication for future research.
5. Reference.

CHAPTER 2
BACKGROUND STUDY

2.1 Introduction

Work with this project we are study so many paper and journals and work with some algorithms related to our work. We read many kinds of papers from those papers we describe some work in this paper those are important for this model. we describe paper about automatic code review, automatic academic paper review, automatic newspaper review. And the other we read about some algorithms like convolutional neural networking, attention base neural networking, support vectors machine.

2.2 Related Works

when we start this work at first we start to read paper, journals, and books about machine learning. Our work is about deep learning and machine learning are related on the other hand we told that deep learning is a subset of machine learning. and all those models are related to each other. As like deep learning is inside of machine learning and like that machine learning is a pare fo artificial intelligence. artificial intelligence is a working procedure that behaves like the human brain.

In this section consists of various research articles which are applying Machine learning technique to read the abstract and title in automatic and gets the results.In ABCNN for Modeling Sentence

Pairs they work on how to improve the CNN model. They work stape by stape for improvising this model. In this paper, they work in three steps. And at last, they propose a batter model that is batter then convolutional neural networking. For this model work one step then another.To use many kinds of methods From those the important are Basic Bi-CNN, Attention Based BCNN, ABCNN1, ABCNN2, and ABCNN3.

In Building systemic they use automatic text classification techniques in this part represents classifying included or excluded and relevant and nonrelevant objectives. To maximize the number of relevant documents to ensure a systematic review. The next work is the reduce the over workload from the reviewers.They use bag-of-words, Unified Medical Language System, MetaMap, complement naive Bayes, Bi-Normal Separation.

In Intelligent code reviews It is automatic code reviewer so firstly it is broken the code part by part and then it is part all the code sequentially. Then review all the parts and gets the results. But its results are different from the others because its results system is too smart its describe all the part and get the proper results.It is use Pre-processing, Deep code review, Deep learning.

4

Systemic reviews for the automatic short answer grading in this stage mainly working about the teacher will find the task for the student. Then the student follows it and working about it and the teacher gets the results in after many times passed .so the time and the similar results percentage is very important for it. If we use it then get the result early and grading system is better then the first time.They use Conducting, Nature of datasets, Natural Language processing technique, Machine learning algorithms.Work with this project we are study so many paper and journals and work with some algorithms related to our work. We read many kinds of papers from those papers we describe some work in this paper those are important for this model. we describe paper about automatic code review, automatic academic paper review, automatic newspaper review. And the other we read about some algorithms like convolutional neural networking, attention base neural networking, support vectors machine.

In Building systemic they use automatic text classification techniques in this part represents classifying included or excluded and relevant and nonrelevant objectives. To maximize the number of relevant documents to ensure a systematic review. The next work is the reduce the over workload from the reviewers.They use bag-of-words, Unified Medical Language System, MetaMap, complement naive Bayes, Bi-Normal Separation.

2.3 Research Summary

In this research From the normal convolutional neural networking or non-attention CNN gives a good result. But in this paper, the attention method is added for this is to give us a better result. And from comparing ABCNN1 and ABCNN2 from that two ABCNN2 is give us a better result than ABCNN1.and then ABCNN3 gives us a better result than ABCNN2. From all those works those models are not a big improvement. In the future, they work one this and give a better Model for this work.

This stage use the technique that is human-human workflow in the systematic review. Also calculated the evaluation measures that are kept the human-machine workflow and observe the recall the other reviewer and see that the 2-vote technique is supervised others technique in the global from.Update the systemic reviews by the investigate ways and update the result with more improve and select the training data set.

This part It so good for the result formate or result structure but its algorithms are not so good for the outcome result so they don't get the perfect result. Because the algorithms part is not working properly for the perfect result its drop more so the result is so much poor result.Update all the functions or algorithms part by part for the better results and sequentially the working procedure is changed and bring the update part of the algorithms .

In this platform, they use the synthetic review for the automatic grading system and they use a machine learning approach. And then the results show that the results come automatically for the systemic reviews. It is day by day updated its results for the working procedure.In this research, they update and create new techniques, datasets, and deep learning. And that has been creat a lot of different areas and underexplored .

5

2.4 Scope of the study

Most of the papers we saw that they are implemented all the algorithm in many ways but not clear the code side and they are used most of the time many structures and many other things that are not important for clear all the side . But in this stage, we use directly use many algorithms to find out and detect the automatic review in the abstract part. We mainly use all of this for automatic the review all the parts and find out the right. We have determined the accuracy of the review side to automatic review the part.

2.5 Challenges Chapter

We started working the real data set not use the synthetic dataset. In many papers, thy use the synthetic dataset but we find out or creat the real dataset for the identify the real things. The real dataset is huge and so difficult for the working part we faced many difficult problems when we are working for the identify many of the parts are missing many of the parts are not found, not matched so many other problems. Firstly we solved it and then we start the working part . In the working time, the whole dataset is not lode, in that case, we use the python and then it is loaded but all the function is not working then solved all the parts then it is working properly . We Read the parts and find out the output. With out those challenges we have another challenges are to work with when we work with HCNN this model give us a output but it gives us bad output for the resected class we can't train properly our model.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction:

Our work is all about machine learning and deep learning. Using machine learning how could we review a resource paper. In doing that work we propose a model name Hierarchical Convolutional Neural Networking. Using this HCNN model we can easily give the rating research paper. In this current world trying to do all kinds of time killing work using AI. For thinking about that fact we are trying to give a model that automatically gives us a desition to observe a paper is this paper accepted or resected. For this, we use machine learning technics. Machine learning is a very powerful way to work a machine-like human. Using that concept we want to make this mode to

rate abstract. From machine learning, we use deep learning algorithms in neural networking and some others.

3.2 Research Subject and Instrumentation:

Our research subject is an automatic abstract review using machine learning technics. For that, we use python with Pytorch framework and we also use anconda3 and jupyter notebook as a platform and also use Google Colab.

3.3 Data Collection Procedure:

Source of data: we collect data for online sources. This data set includes more than 17 thousands of academic papers. This data set is used for the text classification(Doen 2017). All the attributes of our data set in text formate. Data set is make with more than 19 thousand papers all of them more than 17 papers are use for train our model from training set 8889 papers are in accepted or yes class and 8329 papers are include rejected or no class.
The validation set is made with thousand of a research paper from those papers 513 are in yes class and 487 are in no class.
For testing, it uses also a thousand of papers in this section 506 are included in positive output or in yes class and other 496 are in no class.

1. All data includes 'data1.json', 'data2.json', 'data3.json' and 'data4.json' four files. The total number of samples is 37464.
2. After importing each json file, you will get a dictionary. Each item in the dictionary represents a sample. The key represents the 'paper-id', and the value represents the data of each paper.
3. The data of each paper is also a dictionary. The key of this dictionary is as follows:
    'abstract': the abstract of the paper;
    'title' : the title of the paper;
   'authors' : the authors of the paper;

3.4 Proposed Method:

For this model we use mainly CNN and LSTM. Train this model we use hierarchical convolutional neural networking for this model we combine attention based convolutional neural network, long short term memory, and alternative pooling layer. Using those three neural net algorithms we got a model that mode is our Hierarchical Convolutional Neural Networking model.

3.4.1 Hierarchical Convolutional Neural Networking:

We decide a model that is Hierarchical Convolutional Neural Networking when we want to describe this model first we want to know about the Hierarchical model than about the convolutional neural net.

The hierarchical model is a system to arrange a big model. Using this model we easily know what is the module comes fast and will be the next. In this model, modules are arranged in proper form that our model is needed. How Hierarchical method used in our model? In our model papers are divided into come module. For example, title is a module,

abstract is a module. which is module come fast and what module then we decide using the hierarchical model.

For this model, papers are divided into several modules (If a paper is P then p1, p2, ...pl are module).
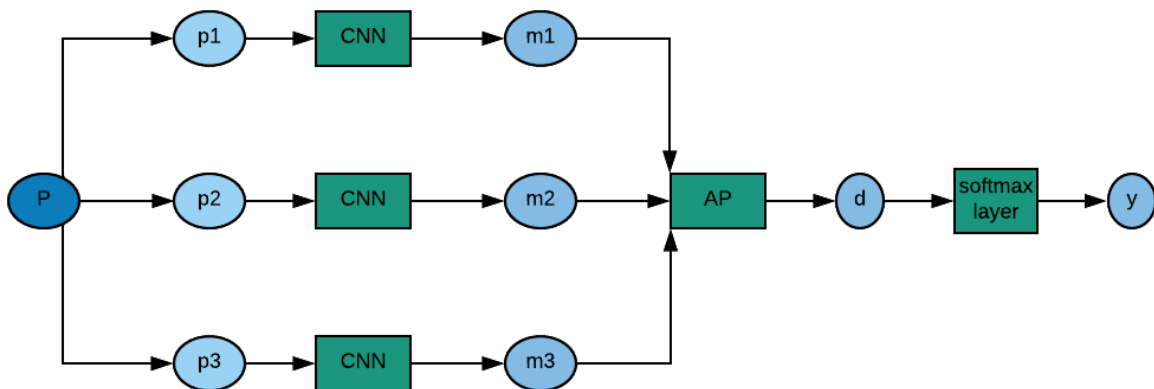

Figer : Hierarchical Convolutional Neural Network

Every module are apply into the attention-based convolutional neural network model. That we decide previously(p1,p2..). attention-based convolutional neural network is used to encode the sentence level. And the convolution neural networking represents those models an encoded model(m1,m2...).

From those m1, m2....ml use in attention pooling. Attentive pooling layer is different than normal pooling that we will discuss later in this report. Using of this pooling we got an aggregated representation d. We don't use d directly in our work final class.what is the valo we got form d this value is go through another method that call softmax layer. For that a softmax layer convert d into a usable form for final classification desition. Softmax is use to convert a number into 0 to 1 range that is important for our model. If we will get value in range 0 to 1 then we easily go with a class and that the goal of our classification using our proposed model.

3.4.2 Convolutional Neural Networking:

For this model CNN use in every module (p1, p2 ...). It is used for encoded modules using of convolution layer and pooling layer. In the first layer, it encoded the words and every code is different for every word. then it found a relation between those words and build relationships with those words.
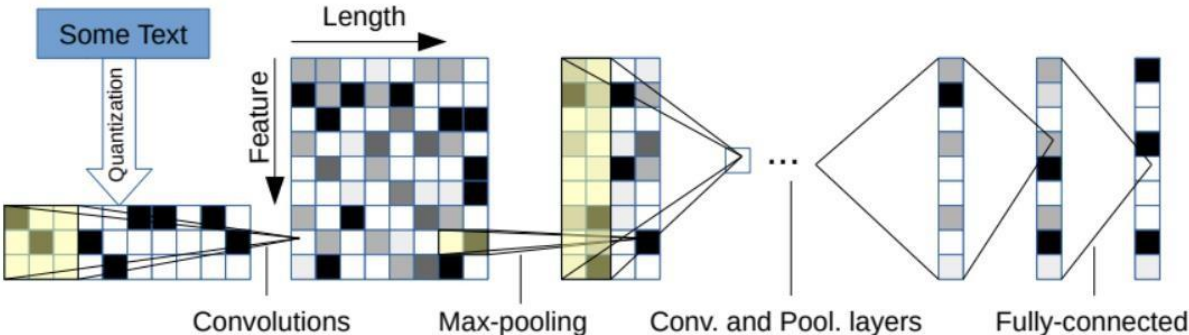

Figer: CNN for text classification

4.3.3 Convolution layer:

we are very familiar with 2D image classification in 2D image classification we use 2D kernel or filter. and then those kernels perform one by one. For image, classification performs more then one kernel.

We told about image classification but in text classification, it performs 1D kernel or filters because text data are always performed in 1D it will never perform like 2D.
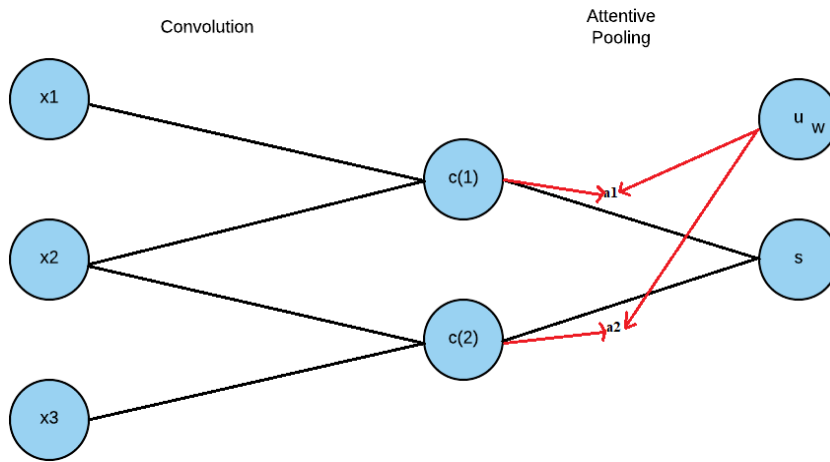
Figer: CNN for Hierarchical Convolutional Neural Network

In our model, CNN is performed efficiently than traditional CNN. It is different when the pooling layer is performed in our mode to use an attentive pooling layer. we know that after convolution pooling lear is performed in every CNN.

3.3.4 Pooling layer:

another important layer for CNN is the pooling layer. In this layer value is calculated using some calculation called max pooling.

For this model we use an attentive pooling layer for this we got c (c1, c2...). the attentive pooling is used to aggregate the representation of the sequence to know how the words are formatted high-level representation.
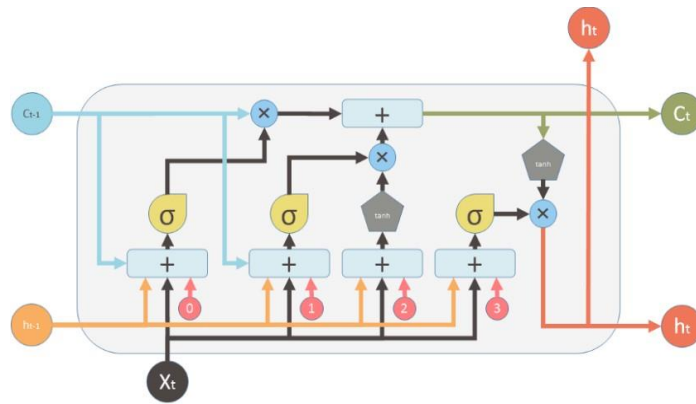
Figer: working process of LSTM

3.3.5 Long Short Term Memory:

Our work we use LSTM but it is not the regular LSTM model it is the C-LSTM that proposed for text classification. this LSTM mode use in convolution network for higher-level text classification representation. In this work before getting m this output from pooling is going through this C-LSTM model. In this model LSTM is use for previous and data the relation between text.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

we show a model in previse chapter based on that model now we see some results. In this chapter, we see a confusion matric and compersion between CNN and HCNN.

This paper is all about an automatic abstract reviewDepend on those goes with some experimental result We got those result and those will show and discuss below. In this result section, we present the accuracy of this model and the confusion matrices for the accuracy check method. Discuss how much this model is appropriate and consistent.

4.2 Experimentation result:

4.2.1 Table for Confusion matrix:

Total =1000	True	False
Positive	383	121
Negative	202	294

Description:

Accuracy of HCNN= 67.7%

For testing our proposed model we use 1000 paper. Our total number of data is n= 1000. True positive =383, True Negative = 294, False Positive = 121, False Negative = 202 In this result, we see that for the accept paper we gote more perfect results than the rejected paper. we are work on that problem.

Compare Accuracy:

Accuracy of CNN = 61.3%
Accuracy of HCNN = 67.7%

For those two model accuracy of CNN is 61.3% and when we use HCNN than accuracy is 67.7%. The accuracy of HCNN is improving 6.4% than traditional CNN. Because of this kind of change is when we apply a hierarchical model that time CNN is applied in every hierarchical model differently and then we got a combined output from that work.

CHAPTER 5

CONCLUSION

5.1 Summary:

In this paper, we using the deep learning process through machine learning for automatic abstract review. For that work, we train this model using convolutional neural networking, long short term memory and some other technics of machine learning. we

use more than seventy thousands of data set train this model. In this paper, we discuss all the algorithms that are used for this model.

5.2 Limitations:

Our paper is review only the abstract from the whole paper. Our paper model is trained only to review the abstract. And what the result we gel that is not so good.

5.3 Future study:

In future, we want to develop this model for review whole paper automatically. For that, we improvise the model using a more deep learning algorithm or optimize this model. we improve the parents of the accuracy of the papers review.

CHAPTER 6
REFERENCES

L. Dong, F. Wei, S. Liu, M. Zhou, K. Xu. 2014. A Statistical Parsing Framework for Sentiment Classification. CoRR, abs/1401.6330

Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 101–110. ACM, 2014.

Rui Zhang, Honglak Lee, and Dragomir Radev. 2016. Dependency sensitive convolutional neural networks for modeling sentences and documents. In Proceedings of NAACL-HLT, pages 1512–1521.

Y. Shen, X. He, J. Gao, L. Deng, G. Mesnil. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search. In Proceedings of WWW 2014.

 J. Silva, L. Coheur, A. Mendes, A. Wichert. 2011. From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, 35(2):137–154.

O. Liu, O.L., Rios, J.A., Heilman, M., Gerard, L., Linn, M.C.: Validation of automated scoring of science assessments. J. Res. Sci. Teach. 53(2), 215–233 (2016).

L. Magooda, A., Zahran, M.A., Rashwan, M., Raafat, H., Fayek, M.B.: Vector based techniques for short answer grading. In: International Florida Artificial Intelligence Research Society Conference Ahmed, pp. 238–243 (2016)

R. Roy, S., Narahari, Y., Deshmukh, O.D.: A perspective on computer assisted assessment techniques for short free-text answers. In: Ras, E., Joosten-ten Brinke, D.

J. Santos, J.C.A.D., et al.: Avalia c̃ao autom´atica de quest̃oes discursivas usando lsa. Universidade Federal do Par´a (2016)