International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Abstractive Text Summarization of Multimedia News Content using RNN

Vishal Pawar, Manisha Mali.

Abstract: ***Programmed content outline is a principle NLP procedure that plans to combine a source content into a shorter change. The fast expansion in all sort of information appear over the web requires abstractive summarization from non-concurrent accumulations of substance, picture, sound and video. Here propose an abstractive summarization procedure that joins the strategies of NLP, discourse handling, PC vision and Recurrent Neural Network (RNN) to examine the rich data contained from all type of data and to get better insight of the news content. The key plan is to associate the linguistics openings between multimodal substances. Sound and video frames are major properties in video clips. For sound data, we structure a way to manage explicitly manipulate its interpretation and to find the astounding nature of the translation with sound. For video data, we get acquainted with the both depictions of content and pictures with Computer Vision Technique. Previous researches done on Text Summarization mainly focuses on Extractive method of summarization. In this work, we put forward an Abstractive method of summarization with sequence to sequence architecture. Finally, all the multimodal points are considered to make a literary once-over by increasing the striking nature, non-reiteration, clarity and consideration through the arranged streamlining of sub isolated limits.***

Keywords: ***Summarization, Multimedia, RNN, NLP, Sequence-to-Sequence***

I. INTRODUCTION

Automatic textual summary plays a fundamental employment in ours step by step life and had been considered since long time. With the incident to the data era and the advancement of multiple datatype development, various data (counting text, picture, audio and video) have extended altogether. Multimedia data have unimaginably turned the way where people live and made it difficult for us to get noteworthy data capably. Most summarization structures base on just NLP, the opportunity to commonly improve the idea of the diagram with the guide of programmed discourse acknowledgment (ASR) and PC vision (CV) handling systems is commonly dismissed.

Abstractive textual summary aims to create a small summary of the complete article that covers all the important information. Summarization can be divided as *extractive* and *abstractive* methods.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Vishal Pawar*, studying M.Tech. in Computer Engineering at Vishwakarma Institute of Information Technology, Pune, Maharashtra, India.

Manisha Mali, working as an Assistant Professor in Department of Computer Engineering at Vishwakarma Institute of Information Technology, Pune, Maharashtra, India.

Extractive strategies build a synopsis by separating the notable words expressions, or sentences from the source content itself. Then again Abstractive techniques produce an outline that is like a human-composed theoretical by briefly summarizing the original content.

Such that, the previous guarantees the syntactic and semantic rightness of the produced synopses, while the last makes progressively differing and novel substance. In the paper, we center around abstractive content synopsis.

The quick headway in profound learning, urges many succession to grouping models, proposed to extend an information arrangement into another yield succession. These methodologies has been fruitful in numerous applications or undertakings like discourse acknowledgment, video subtitling and machine interpretation. In contrast to these undertakings, in content synopsis, the yield succession (outlined) is a lot shorter than the information source grouping (record). To understand the setting rundown, we proposes an attentional grouping to arrangement model dependent on RNNs. It anticipates the first report into low-dimensional embedding. In any case, RNNs will in general have low-productivity issues as they depend on the past advances when preparing. In this manner, however not normally utilized in succession model, we propose an arrangement to grouping model dependent on RNN to make the portrayals for the source writings.

II. RELATED WORK

Text summarization is to expel the noteworthy data from source archives. With the development of multimedia data on the web, a couple of pros revolve around multimodal summarization starting late. Existing examinations have exhibited that, stood out from text summarization; multimodal summarization can improve the idea of delivered abstract by using data in visual modality. Regardless, the yield of existing multimodal summarization systems is commonly addressed in a single modality.

Lion's share of work in the previous decade has spun around an extractive outline approach [4], [5], [6] where a synopsis comprises of catchphrases or sentences from the source content (article). Dissimilar to extractive techniques replicating contents from the original article legitimately, abstractive outline approach utilizes the intelligible way for human to abridge the key data of the first content. In this manner, abstractive methodologies can create significantly more extravagant and different synopses. Abstractive methodology of rundown undertaking has been institutionalized by the DUC2003 and DUC2004 rivalries [13]. Subsequently, there develop a progression of eminent techniques without neural systems on this undertaking, e.g., the best entertainer TOPIARY framework [14].

Deep learning has been growing quick as of late, and it can deal with numerous NLP errands. So specialists have started to consider such system as a successful, totally information driven option for content synopsis. Reference [7] utilized convolutional way to encode the first content, and an attentional feed-forward neural system to produce outlines. In any case, it didn't utilize various leveled CNNs, hence not so successful. Reference [10] exhibited a gigantic dataset in Chinese langu.age for short substance outline and it applied a RNN based seq2seq model. Reference [11] utilized a practically identical seq2seq model with RNN encoder and decoder. Reference [12] utilized the generative ill-disposed systems to deliver the dynamic content outline, furnished with a period rot consideration component. Reference [13] relied upon BERT to abuse the pre-prepared language model in the seq2seq structure, and arranged a two-organize decoder method to consider the different sides' setting information of each word in a synopsis. Reference [14] proposed a two-arrange sentences determination model dependent on grouping and advancement methods. Reference [15] proposed a substance framework model subject to a LSTM- CNN structure that can create rundowns by researching semantic articulations. Reference [16] displayed a modified methodology for substance consultation which relied upon fluffy standards on an assortment of separated highlights features to find the most huge information from the source content.

A. Multi-Document Summarization

MDS tries to expel huge data from a great deal of reports related to an event to create an outline of much more diminutive size. MDS can be abstractive or extractive in approach. Extractive- based models use distinctive phonetic features, for instance, sentence position [3], [4] and tf*idf [5], to perceive the most surprising sentences in a ton of reports. Diagram based methods [6] are generally used extractive-set up together MDS models based. Finally, the top-situated sentence is picked to manufacture outlines.

B. Multi-modal Summarization

As of now, much work has been performed to consolidate meetings narratives, sports recordings, films, picture storylines and socio media. Reference [2] hope to make huge parts of a social event capturing subject to an examination of sound, textual and visual activities. Reference [4] propose a technique to consolidate a game by separating the textual data removed from multiple assets and recognizing the huge substance. They consolidate news pictures by text and visualize text by pictures.

By then, a news story and a picture are picked to address each topic. For electronic long range interpersonal communication summarization, [2] propose to layout the genuine events subject to multimedia content. A multimodal LDA to recognize topic by getting the connections among the text and image features' of little scale online diaries with introduced pictures. The yield of their procedure is a great deal of specialist pictures that portray the events.

III. PROBLEM DEFINATION

The present applications related to Text Summarization consolidate get-together of summarization, sport video summarization, film summarization, pictorial storyline summarization, course of occasion's summarization and social multimedia summarization. Past examinations on these topics overwhelmingly revolve around sketching out synchronous multimodal substance. Pictorial storylines involve a great deal of pictures with text depictions. None of these applications revolve around sketching out multimedia content with non-simultaneous data of event.

A. Implementation Model Overview:

There are various basic edges in making a not too bad textual framework for multimedia content. The prominent substance in records should be held, and the main convictions in recordings and pictures must be verified. Then, the once-over should be clear and non-dull and should seek after the fixed length prerequisite. All of these perspectives can be as one overhauled by the arranged expansion of sub particular capacities.

Max S T {F(S) :X sϵS ls <=L }

Above T is the course of action of sentences, S is the blueprint, ls is length words, L is spending p lan, i.e., length prerequisite for the summary, and sub measured limit F(S) is the summation score related to the recently referenced points of view. Text is the essential modality of archives, and on occasion, pictures are embedded in records. Recordings include at any rate two sorts of modalities: sound and visual. Next, we give all things considered handling strategies for different modalities.

Sound, i.e., discourse can be consequently converted into text by using an ASR system2. For visual, which is extremely a progression of pictures (diagrams), in light of the fact that most of the adjacent housings hold abundance data, we first concentrate the most significant edges, i.e., key frames. We become acquainted with the joint depictions for textual and visual modalities and would then have the option to recognize the sentence that is appropriate to the picture. Thusly, we can guarantee the consideration of made framework for the visual data.

By then this text output is given to our trained RNN model.

IV. ARCHITECTURE MODEL

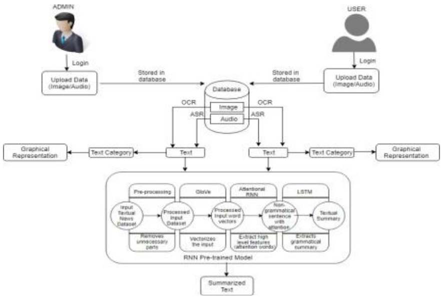


Fig1: Architecture Diagram.

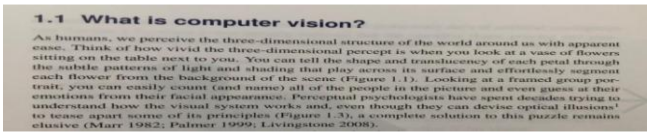
A. Text-Image Matching Model

The key casings in the accounts and the photos installed in the files as frequently as conceivable catch news incorporates that address the immense information that the diagram should cover. Before surveying the fuse for the photos, we need a model to vanquish any counteractive action among content and picture. We can manage this issue by cross-modular assessment. Crossmodal semantic planning can be better explored when multi-modular information is anticipated into the joint subspace.

Here the text-picture coordinating model is readied, for each text-picture pair (si, pj) in our endeavor, we can discover the coordinating score m(si, pj). We decide the edge as a typical coordinating score for the true text-picture pair.

B. Frame Level Text-Image Matching

The principle thought is every movement word from the sentence is separate with its recommendation debates, and the naming for every specific action word is known as a "plot". Each bundling tends to an occasion, and the contentions express the noteworthy information about this occasion. There is a lot of questions demonstrating the semantic action of each term in a bundling. An instance of edge semantic parsing is exhibited wail figure. The primary sentence "President Bush avowed government catastrophe help for the influenced spaces and made blueprints for an assessment voyage through the state" is changed into two adjusted sentences "President Bush embraced bureaucratic debacle help" and "President Bush made game arrangements for an assessment voyage through the state". The improved sentences have less middle of the road variety in importance, which central focuses the content picture planning.



1.1 What is computer vision?

As humans, we perceive the three-dimensional structure of the world around us with apparent ease. Think of how vivid the three-dimensional percept is when you look at a vase of flowers sitting on the table next to you. You can tell the shape and translucency of each petal through the subtle patterns of light and shading that play across its surface and effortlessly segment each flower from the background of the scene (Figure 1.1). Looking at a framed group portrait, you can easily count (and name) all of the people in the picture and even guess at their emotions from their facial appearance. Perceptual psychologists have spent decades trying to understand how the visual system works and, even though they can devise optical illusions[1] to tease apart some of its principles (Figure 1.3), a complete solution to this puzzle remains elusive (Marr 1982; Palmer 1999; Livingstone 2008).

Fig2: Example for simplified sentence based on frame-semantic parsing.

C. Multi-modal Topic Modeling

After the content picture planning model is prepared, we secure the joint delineation of content and pictures. Next, we see the subjects of content and pictures. The inspiration driving this strategy is that printed outlines of pictures as regularly as conceivable give vital information about semantic edges (themes), and picture highlights are reliably associated with semantic subjects. Wang et al. make a course of events outline for Tweet streams by perceiving subject progression. For our errand, the multi-modular theme model can uncover different bits of content and pictures; by then we can investigate an agent course of action of content covering the bits of the photos. Subject models, for example, LDA, can together learn latent points and theme assignments of reports. To uncover the semantic points of view, we make the multi-modular theme model subject to a neural point model (NTM). The multimodal topic model figures the unforeseen probability p(w|d) using the apportionment of the word (or picture)- topic p(w|t) and topic-report (or video) p(t|d).

$$p(w|d) = \sum^{T}_{i=1} p(w \mid ti)p(ti \mid d)$$

D.

E. Seq to Seq RNN Model

Our model we completed are relied upon a succession to arrangement encoder-decoder RNN model with a attention component. For abstractive summarization, it is ordinary to utilize either GRU or LSTM cells for the RNN encoder and decoder. We chose for use LSTM cells for their additional control by means of their memory unit, albeit many top models use GRU cells for their less expensive calculation time.

F. Preprocessing of the training dataset

We prepared our model on sentence-feature sets from the Unannotated English Gigaword Corpus, which is regularly utilized for preparing models for abstractive synopsis. During preprocessing, we hauled out 700,000 sets of features and first sentences, and we prepared our model to anticipate the feature given the primary sentence.

G. Encoder – Decoder with Attention Mechanism

Our model is based on the Neural Machine Translation model used in Bahdanau (2014). The encoder contains bidirectional LSTM-RNN. Whereas the decoder consist of a uni-directional LSTM-RNN with the same hidden state size as encoder and an attention mechanism over the source hidden state. Furthermore, this method likewise accelerates assembly by centering the demonstrating exertion as it were on the words that are fundamental to a given model. This strategy is especially appropriate to summarize since a huge extent of the words in the synopsis originate from the source record in any case.

For encoder-decoder neural systems, the utilization of attention mechanism takes into consideration the creation of a setting vector at each timestep, given the decoder's current concealed state and a subset of the encoder's shrouded states. For worldwide consideration, the setting vector is adapted on the majority of the encoder's shrouded states, though nearby consideration utilizes a severe subset of the encoder's concealed states.

V. DATA COLLECTION AND ANNOTATION

Here we build up a dataset as seeks after. We have taken TOI dataset [17], [18] available publically on kaggle dataset website for news articles in image format. Further, audio clips of All India Radio News are extracted from there official youtube channel [19]. We use python FFMpeg multimedia framework to extract these audio clips in wav format. The criteria for gathering records are (1) hold the huge substance of the information reports (2) avoid monotonous data; (3) have a respectable lucidity; (4) satisfy quite far.

VI. EXPERIMENTAL STUDIES

Two or three models are considered in our assessments, joining conveying once-overs with various modalities and utilizing various ways to deal with oversee effect pictures.

Text: The model produces diagrams using just the text in archives.

Audio: The model produces blueprints using just the discourse translations from recordings.

The going with models produce layouts using the two archives and recordings anyway adventure pictures in different ways. The striking quality scores for text are gained with heading techniques.

A. Implementation Details

We first convert the multimedia news data (Audio and Images) into text form. For this a GUI is developed with a user login and an admin login. We also categorized the news articles into different categories such as Politics, Education, Movies, Electronics, Fashion, Others, etc. Further the extracted text data is fed as an input to our pre-trained RNN model. The trained model gives an abstractive summarization as an output.

For the training of RNN model, we have divided the data into 80-20 manner training and testing respectively.

B. Statistics of the Training Dataset

Total Data Entries	7189
Training Size	5068
Testing Size	1267
Total Epoch	100

C. Experimental Results

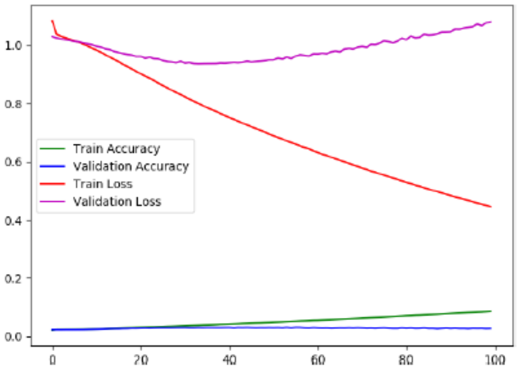


Fig3: Seq to seq RNN accuracy and loss at epoch 50.

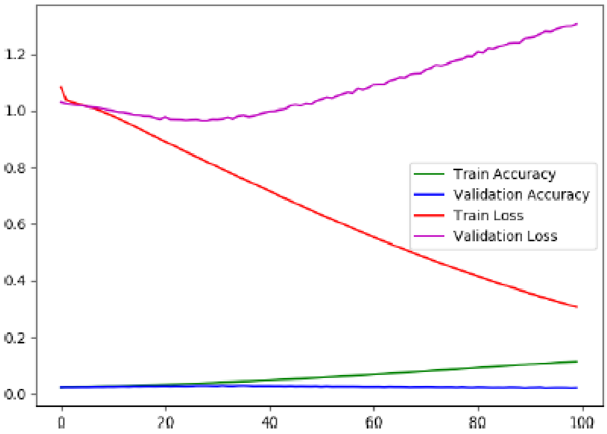


Fig4: Seq-to-seq RNN accuracy and loss at epoch 100.

VII. CONCLUSION

This paper tends to an offbeat Abstractive Summarization task, to be specific, how to utilize related text and sound data to create a textual outline. We apply OCR to convert image to text and ASR for audio to text. The sequence to sequence RNN framework is applied for the task of abstractive summarization with very promising results. Our model is built upon an encoder- decoder model with attentional mechanism. As a part of future work, we will concentrate on developing more advance model that deal with the huge range of multimedia data to extract more comprehensive summarization.

REFERENCES

1. Nallapati, R.; Zhou, B.; dos Santos, C.N.; Gülçehre, Ç.; Xiang, B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, 11–12 August 2016; pp. 280–290.
2. G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," IEEE Transactions on Multimedia, vol. 15, no. 7, pp. 1553–1568, 2013.
3. R. R. Shah, A. D. Shaikh, Y. Yu, W. Geng, R. Zimmermann, and G.Wu, "Eventbuilder: Real-time multimedia event summarization by visualizing social media," in Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015, pp. 185–188.
4. J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, "Automatic text summarization using a machine learning approach," in *Proc. Adv. Artif. Intell. 16th Braz. Symp. Artif. Intell. (SBIA)*, Nov. 2002, pp. 205–215.
5. G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *CoRR*, vol. abs/1109.2128, 2011.[Online].Available: http://arxiv.org/abs/1109.2128.
6. K. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proc. Conf. 22nd Int.*
7. *Conf. Comput. Linguist. (COLING)*, Manchester, U.K., Aug. 2008, pp. 985–992.
8. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Lisbon, Portugal, Sep. 2015, pp. 379–389.
9. S. Khuller, A. Moss, and J. S. Naor, "The budgeted maximum coverage problem," Information Processing Letters, vol. 70, no. 1, pp. 39–45, 1999.
10. K. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proc. Conf. 22nd Int.*
11. *Conf. Comput.*
12. Hu, Q. Chen, and F. Zhu, "LCSTS: A large scale Chinese short text summarization dataset," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Lisbon, Portugal, Sep. 2015, pp. 1967–1972.
13. Nallapati, R.; Xiang, B.; Zhou, B. Sequence-to-Sequence RNNs for Text Summarization. arXiv 2016, arXiv:1602.06023.
14. Rekabdar, B.; Mousas, C.; Gupta, B. Generative Adversarial Network with Policy Gradient for Text Summarization. In Proceedings of the 13th IEEE International Conference on Semantic Computing, ICSC 2019, Newport Beach, CA, USA, 30 January–1 February 2019; pp. 204–207, doi:10.1109/ICOSC.2019.8665583.
15. Zhang, H.; Gong, Y.; Yan, Y.; Duan, N.; Xu, J.; Wang, J.; Gong, M.; Zhou, M. Pretraining-Based Natural Language Generation for Text Summarization. arXiv 2019, arXiv:1902.09243.
16. D. Zajic, B. Dorr, and R. Schwartz, "Bbn/umd at DUC-2004: Topiary," in *Proc. Doc. Understanding Conf. NLT/NAACL*, 2004, pp. 112–119.
17. Song, S.; Huang, H.; Ruan, T. Abstractive text summarization using LSTM-CNN based deep learning. Multimed. Tools Appl. 2019, 78, 857–875, doi:10.1007/s11042-018-5749-3.
18. Goularte, F.B.; Nassar, S.M.; Fileto, R.; Saggion, H. A text summarization method based on fuzzy rules and applicable to automated assessment. Expert Syst. Appl. 2019, 115,264–275, doi:10.1016/j.eswa.2018.07.047.
19. https://epaper.timesgroup.com/Olive/ODN/Tim esOfIndia/#
20. https://www.kaggle.com/snapcrack/all-the- news
21. https://www.youtube.com/watch?v=hr1vrOqmKHs&list=PLcDghvQhYD9J7mvAruAoDuT08Mu6b5ITS

AUTHORS PROFILE

Vishal Pawar, studying M.Tech. in Computer Engineering at Vishwakarma Institute of Information Technology, Pune, Maharashtra, India. His area of interests are Machine Learning, Data Analytics, Data mining, Data Science and Artificial Intelligence.
Email Id: vishal.17p007@viit.ac.in

Manisha Mali, working as an Assistant Professor in Department of Computer Engineering at Vishwakarma Institute of Information Technology, Pune, Maharashtra, India. Her area of interests are Block Chain, Data Analytics, Text Mining, Data mining, Sentiment/Opinion Mining, Information Extraction, and Machine Learning.
Email Id: Manisha.mali@viit.ac.in