JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Multi Modal Text and Image Summarization using Deep Learning and Natural Language Processing Techniques – A Review

[1]Rucha R. Shaiva, [2]Prof. Dr. Bhagwan D. Phulpagar, [3]Prof. Yogita P. Narwadkar
[1]ME Student, [2]Associate Professor, [3]Assistant Professor
[1]Department of Computer Engineering
[1]PES's Modern College of Engineering, Pune, India

Abstract: The notion of automated information retrieval and text summarization is challenging in natural language processing due to the great complexity and irregular structure of the texts. A lengthy text is paraphrased during the text summarizing process to provide a summary. The automatic generation of a phrase to describe an image is known as image captioning, and it is a field that combines natural language processing with computer vision. The study of picture captioning has a significant influence on how visually impaired individuals comprehend their environment and may prove advantageous for sentence-level photo organizing. Convolution neural networks (CNN) and recurrent neural networks (RNN) were the primary building blocks of contemporary techniques. Making precise and evocative subtitles is still a difficult endeavor, though. Sentences that match the visual material are referred to as accurate captions; sentences that provide a variety of descriptions, as opposed to simple, everyday language, are referred to as descriptive captions. Generally speaking, the language model must consistently translate the graphical representation into a legible phrase, and the vision model must encode the context completely.

***IndexTerms* - Information retrieval, text summarization, deep learning, word2vec, dense captioning, Stanford, NLP**

I. Introduction

With the proliferation of multi-modal papers on the Internet, it is becoming increasingly necessary to summarize multi-modal materials in order to obtain multi-modal summaries. Text-image summarizing is the process of creating a summary using both text and images from a source. Pure text summarization is not the same as the summary technique. Additionally, it differs from picture summarization, which condenses an image collection into a smaller subset. A picture speaks a thousand words. In the transfer of information, image is crucial. Text-image summaries created by integrating visuals with text can improve people's comprehension, memorization, and expression of knowledge. The majority of current research concentrates on picture or pure text summarization.

To address this, in this study we will utilize a neural text-image summarization model built on the word2vec as well as hierarchical encoder-decoder models. During the decoding stage, we start with a combined text and picture encoding and utilize the attention hierarchy decoder to generate the text summary by taking into account the original phrases, images, and captions. Every sentence that is created is compared to a sentence, picture, or title found in the source text. Images are chosen and aligned to the output text based on alignment scores. In the inference phase, we employ the multimodal beam search strategy, which evaluates the beams based on the bigram overlap between the visited captions and the created words.

II. Related WORK

[1] Natural language processing is a discipline in which multimodal text summarization is a challenging and complex topic. Its objective is to integrate characteristics from multiple modalities in order to combine a specific set of input information into a brief yet useful summary. The author of this work conducted a detailed examination of several multimodal text summarizing techniques and procedures, looking at the consequences for academia and business. The author's model has proven to perform at the cutting edge on the MMSS dataset. The author also proposes an evaluation approach that uses semantics to determine the quality of the summaries that are generated.

[2] Multimodal sentiment analysis (MSA) research is important because multimodal data may efficiently represent user emotions and feelings. Nevertheless, semantic features and the connection between visual and textual material are frequently overlooked by present methods. This paper offers a novel Deep Multi-View Attention Network (DMVAN) for accurate multimodal emotion and feeling classification. The DMVAN model has three stages: cross-modal fusion learning, attentive interaction learning, and feature learning. It improves interaction, extracts textual and visual information, and combines data from intermediary features. Three real-world datasets are used to evaluate the model, and on Binary_Getty, Twitter, and the EMO-G dataset, it achieves accuracy rates of 99.801%, 96.867%, and 96.174%.

[3] The study suggests a multi-modal RNN-based neural-based extractive multi-modal summarization technique. In addition to encoding articles and photos, it gathers online images to expand the DailyMail corpus and uses a logistic classifier to compute summary probability. Tests reveal that the approach works better than the most advanced neural summarizing systems, demonstrating the power of deep learning approaches for summary.

[4] Multi-modal summarization is a complex problem in natural language processing and information retrieval, incorporating visual and aural aspects. Its flexible nature makes it challenging to understand existing works with uni-modal techniques alone. Humans create multi-modal summarization content using prior understanding and external knowledge, making automatic multi-modal summarization an interesting task.

[5] The author proposes a strategy for summarizing an event in multiple tweets, allowing users to quickly understand key moments. They develop a graph-based retrieval technique that detects popular discussion points from Twitter search engine tweets. Topical clustering is performed to ensure coverage of thematic variety. The technique can summarize the progress of different incidents with up to 81.6% precision and 80% recall.

[6] Suggested a framework for automatically creating visual summaries from the microblog feeds of various media types. The three steps that comprise the proposed framework are as follows: Firstly, a noise reduction technique is developed to remove potentially noisy photos. Cross-Media-LDA (CMLDA), a cross media stochastic model, is suggested to simultaneously detect microblog sub events from various media sources. Ultimately, a method for creating multimedia microblog summaries is developed to collaboratively find pertinent visual and textual samples. These are then combined to create a comprehensive visualized summary that draws on cross-media understanding of all the sub events that were found.

Extensive tests on two real world micro blogging datasets were carried out to illustrate the advantages of the suggested framework over the most sophisticated approaches. In order to develop summaries that perform better than any of the basic algorithms, [7] suggested ensemble approaches to aggregate the output of many base summarizing algorithms. It also looked into the possibility of integrating existing summarization techniques to produce summaries of greater quality. [8] The author suggested creating excellent summaries for microblog summaries by applying the concept of multiobjective optimization. The suggested framework studies several types of genetic operators, such as the self-organizing map (a kind of neural network) orientated generator that was recently found. To measure the similarity of tweets, word mover distance is utilized, which may capture semantic similarity.

Multiple statistical quality indicators, including length, the tf-idf value of tweets, non repeative, and rating various aspects of summary, are improved concurrently using the search capacity of a multiobjective variation technique. For assessment, four reference sets of data on catastrophic events are used. The results are compared with many cutting-edge approaches using ROUGE measures. Using statistical measures, the ROUGE are applied to four benchmark data sets for catastrophic events and outperform state-of-the-art techniques by 62.37% and 5.65%, respectively.

[9] Looked at machine learning-based rumor detection methods. It was shown that selection and feature selection had a higher impact on the accuracy of rumor identification than machine-learning technique selection. Furthermore, rumor detection strategies were investigated by combining five entirely novel user behavior-based components—such as reposting and comments from followers—with well-established, highly successful user behavior-based characteristics to evaluate whether a microblog post is a rumor. The findings of the experiment, which used real data from Sina Weibo, show that the proposed technique and features are successful and efficient. The study results lead the authors to the conclusion that rumor identification based on mass behaviors outperforms identification based on microblog features.

[10] To identify twitter spam, a novel spam detection approach, Glove, combines convolutional neural networks (CNNs) with a based on features model. As a meta-classifier, the model employs based on content, user-based, and n-gram characteristics, as well as a multilayer neural network. On balanced and imbalanced information sets, the strategy surpasses current approaches, indicating the usefulness of this ensemble approach. By employing a multilayer neural network as a meta-classifier, the technique integrates deep learning with traditional feature-based models. The authors tested their methods using two sets of data, one balanced and one unbalanced. The experimental results suggest that the proposed solution outperforms previous approaches.

[11] Two modules make up the suggested framework: a batch model update module and a real-time spam detection module. Four light detectors are used by the spam detection module: a multi classifier-based detector, a near-duplicate detector, a blacklisted domain detector, and a dependable ham detector. Based on earlier tweets, the detecting module's data is updated in batch mode. The system maintains high accuracy in identifying spam in a Twitter stream by adaptively learning new trends in spam behavior. Based on the tweets tagged over the preceding time, the detection module's data is modified in batch mode. Extensive experiments on large data sets demonstrate that the framework adaptively learns the trends of novel spam activities while maintaining exceptional accuracy in detecting spam in a twitter stream.

[12] The authors proposed non-dominated sorting and sharing, which were criticized for their computational expense, non-elitism approach, and need to specify a sharing parameter. They present NSGA-II, a non-dominated sort-based Multi-Objective Evolutionary Algorithm (MOEA), which solves all three problems. NSGA-II is faster than two other elitist MOEAs, the strength-Pareto natural algorithm and the Pareto-archived development strategy, and can find a broader variety of solutions while achieving greater convergence close to the true Pareto-optimal front. The authors modify the idea of dominance to address limited multiple challenges. The NSGA-II performs better on restricted simulation results, particularly a five-objective, seven-constraint not linear issue, compared to another constrained multifaceted optimizer.

[13] The authors created the dense captioning issue, which calls for a system that uses computer vision to identify important areas in images and provide plain English descriptions of them. For this purpose, they suggested using a Fully Convolutional Localization Network (FCLN) structure, which can be trained end-to-end in a single optimization cycle, analyzes pictures rapidly, and doesn't require any external region recommendations. Convolutional networks, dense localization layers, and recurrent neural networks as language models are all part of the design.

[14] The author makes use of a breakthrough computer vision technique called dense captioning, which facilitates the interpretation of images with detailed verbal descriptions. The aim is to identify visual concepts (i.e., objects, object fragments, and connections among them) in images and label each with a short, descriptive term. We draw attention to two significant problems with dense captioning that need to be resolved in order to solve the problem. First, it is challenging to precisely localize visual concepts since each image has rich annotations related to visual concepts that considerably overlap target areas. Second, it is challenging to identify any visual concept based just on look due to the vast quantity of them. The authors propose a new model pipeline based on two novel concepts, collaborative inference and context fusion, to overcome these two problems. We carefully consider architectural modifications as we steadily create our model architecture. The compact and efficient final model we have developed delivers the highest accuracy for dense captioning on Visual Genome, achieving an approximate gain of 73% over the state-of-the-art method. Qualitative studies also demonstrate the semantic possibilities of our technology in crowded settings.

[15] The author uses Google as a search engine and the World Wide Web (WWW) as a database. This tactic also functions with other databases and search engines. This idea is then applied to the construction of a system that automatically calculates the Google similarity distance, or similarity, between words and phrases found on the World Wide Web based on Google page counts. The World Wide Web is the largest database on the planet, and the contextual data contributed by millions of unique users adds up to produce useful automatic semantics. The author uses the WordNet file as an objective baseline against which to evaluate the efficacy of our approach, and then conducts a massive randomized trial in binary classification using support vector machine learning to learn categories based on our Google distance, resulting in an eighty-seven percent mean agreement with the specially crafted WordNet.

Fuzzy-based and machine learning techniques are utilized in real-time summarization, particularly in the use of fuzzy logic in Zadeh's calculus for trend extraction. This method excels in norm evaluation but struggles with semantic issues due to unclear results from other t-norms. Understanding the unclear results helps in comprehending the method's effectiveness. Fuzzy logic in Zadeh's calculus is a prime example of a fuzzy-based approach used to address trend extraction and immediate issues, despite its strengths in norm evaluation. Fuzzy Formal Concept Assessment (Fuzzy FCA) and Incremental Short Text Summarization (IncreSTS) are machine learning approaches that improve recall and precision in semantic and real-time issues. IncreSTS offers increased outlier management, efficiency, and scalability on target difficulties, while RBP-SUM avoids repetition by assessing with rouge but only provides extractive summaries.

In the realm of Natural Language Processing (NLP), text summarizing is a challenging endeavor [20] because producing a strong summary requires in-depth text analysis, including semantic and lexical analysis. A good summary should be concise, contain just the most important information, and take into account readability, coherence, non-redundancy, coverage, and readability [21]. It might be difficult to include all of these elements in a summary. Because summarizing methods for extraction has grown into a huge study field that is almost finished, evaluating studies on the summarization of texts is essential [22]. Abstractive summarization [22] or actual time summarization are now the main areas of study. Since extractive techniques have developed into a mature study area, text summarization evaluation is essential. Because abstract summary and real time summarizing have more detailed and extensive descriptions, they are the major subjects of attention.

Consequently, compared to abstractive summaries, extractive summaries yield better and more consistent outcomes [23]. The presence of extractive research throughout the preceding two years, however, indicates that summarized extractive is still in great demand. This implies there may yet be opportunities or loopholes to shut. To conduct additional studies in the field of text summarization, a thorough evaluation of the literature is necessary. In contrast, most literature research are constrained, analyzed, and analyzed in an assessment or survey study.

The review paper focuses on abstractive summarization, with an emphasis on research advances, current methodologies, tools, and assessments. It also gives an overview of text summarization methods, notably in Arabic. [25] Conducts a survey of text summarization methodologies and approaches, covering statistical, machine learning, semantic-based, and swarm intelligence techniques. Another study [26] focuses on extractive text summarization employing unsupervised techniques, presenting pros and cons in comparative tables. Some review articles concentrate on certain topics, such as approach tactics, methodologies employed, assessment approaches, or the issue of extractive or abstractive text summary.

III. CONCLUSION

This paper aims to present the latest research and advancements in this field using the novel methods discussed also it also aims to show that it can provide a more structured, comprehensive, and unique review covering a wide range of topics and issues, including trends, data sets, preprocessing, characteristics, approach techniques, challenges, methods, and evaluation that can serve as a roadmap for future research. The relationship between the problems, obstacles, and challenges in each topic, approach, and methods is discussed in depth.

REFERENCES

[1] M. A. H. Khan, D. Bollegala, G. Liu and K. Sezaki, "Multi-tweet Summarization of Real-Time Events," 2013 International Conference on Social Computing, Alexandria, VA, USA, 2013, pp. 128-133, doi: 10.1109/SocialCom.2013.26.J.

[2] Bian, Y. Yang, H. Zhang and T. -S. Chua, "Multimedia Summarization for Social Events in Microblog Stream," in IEEE Transactions on Multimedia, vol. 17, no. 2, pp. 216-228, Feb. 2015, doi: 10.1109/TMM.2014.2384912.S.

[3] Dutta, V. Chandra, K. Mehra, A. K. Das, T. Chakraborty and S. Ghosh, "Ensemble Algorithms for Microblog Summarization," in IEEE Intelligent Systems, vol. 33, no. 3, pp. 4-14, May./Jun. 2018, doi: 10.1109/MIS.2018.033001411.

[4] N. Saini, S. Saha and P. Bhattacharyya, "Multiobjective-Based Approach for Microblog Summarization," in IEEE Transactions on Computational Social Systems, vol. 6, no. 6, pp. 1219-1231, Dec. 2019, doi: 10.1109/TCSS.2019.2945172.

[5] G. Liang, W. He, C. Xu, L. Chen and J. Zeng, "Rumor Identification in Microblogging Systems Based on Users' Behavior," in IEEE Transactions on Computational Social Systems, vol. 2, no. 3, pp. 99-108, Sept. 2015, doi: 10.1109/TCSS.2016.2517458.

[6] S. Madisetty and M. S. Desarkar, "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter," in IEEE Transactions on Computational Social Systems, vol. 5, no. 4, pp. 973-984, Dec. 2018, doi: 10.1109/TCSS.2018.2878852.

[7] S. Sedhai and A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," in IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 169-175, March 2018, doi: 10.1109/TCSS.2017.2773581.

[8] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, April 2002, doi: 10.1109/4235.996017.

[9] J. Johnson, A. Karpathy and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4565-4574, doi: 10.1109/CVPR.2016.494.

[10] L. Yang, K. Tang, J. Yang and L. -J. Li, "Dense Captioning with Joint Inference and Visual Context," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1978-1987, doi: 10.1109/CVPR.2017.214.

[11] R. L. Cilibrasi and P. M. B. Vitanyi, "The Google Similarity Distance," in IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pp. 370-383, March 2007, doi: 10.1109/TKDE.2007.48.

[12] Kacprzyk, J., Wilbik, A., Zadrozny, S., 2008. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets Syst. 159, 1485–1499. https://doi.org/10.1016/j.fss.2008.01.025.

[13] Maio, C De, Fenza, G., Loia, V., Parente, M., 2015. Time aware knowledge extraction for microblog summarization on twitter. Inf. Fusion.

[14] Liu, C., Tseng, C., Chen, M., 2015a. IncreSTS: Towards real-time incremental short text summarization on comment streams from social network services. IEEE Trans. Knowl. Data Eng. 60, 1–14. https://doi.org/10.1109/TKDE.2015.2405553.

[15] Rodríguez-Vidal, J., Jorge, C.-D.-A., Amigó, E., Plaza, L., Gonzalo, J., 2019. Automatic generation of Entity -oriented Summaries for Reputation Management. J. Ambient Intell. Humaniz. Comput. https://doi.org/10.1007/s12652-019-01255-9.

[16] Rane, N., Govilkar, S., 2019. Recent trends in deep learning based abstractive text summarization. Int. J. Recent Technol. Eng. 8, 3108–3115 https://doi.org/10. 35940/ijrte.C4996.098319.

[17] Verma, P., Om, H., 2019. MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. Expert Syst. Appl. 120, 43– 56. https://doi.org/10.1016/j.eswa.2018.11.022.

[18] Gupta, S., Gupta, S.K., 2019. Abstractive summarization: An overview of the state of the ARt. Expert Syst. Appl. 121, 49–65. https://doi.org/10.1016/jeswa.2018.12.011.

[19] Elrefaiy, A., Abas, A.R., Elhenawy, I., 2018. Review of recent techniques for extractive text summarization. J. Theor. Appl. Inf. Technol. 96, 7739–7759.

[20] Abualigah, L., Bashabsheh, M.Q., Alabool, H., Shehab, M., 2020. Text summarization: A brief review. Stud. Comput. Intell. 874, 1–15. https://doi.org/10.1007/978-3-030-34614-0_1.

[21] Nazari, N., Mahdavi, M., 2018. A survey on automatic text summarization. J. AI Data Min. 0, 121–135. https://doi.org/10.22044/jadm.2018.6139.1726.

[22] Elrefaiy, A., Abas, A.R., Elhenawy, I., 2018. Review of recent techniques for extractive text summarization. J. Theor. Appl. Inf. Technol. 96, 7739–7759