Reviewing various techniques for Text and Image based Summarization

Manisha Rashinkar
Mit College of Engineering, Aurangabad

***Abstract*—Automated information retrieval and text summarization concept is a difficult process in natural language processing because of the infrequent structure and high complexity of the documents. The text summarization process creates a summary by paraphrasing a long text. Image captioning is to automatically describe an image with a sentence, which is a topic connecting computer vision and natural language processing. Research on image captioning has great impact to help visually impaired people understand their surroundings, and it has potential benefits for the sentence-level photo organization.Modern methods were mainly based on a combination of Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN). However, generating accurate and descriptive captions remains a challenging task. Accurate captions refer to sentences consistent with the visual content, and descriptive captions refer to those with diverse descriptions rather than plain common sentences. Generally, the vision model is required to encode the context comprehensively and the language model is required to express the visual representation into a readable sentence consistently. Additionally, the training strategy also affects the performance. Additionally, the results show that summaries generated using these semi supervised approaches lead indeed to higher ROUGE scores than n-gram language models reported in previous work. We propose a multi-modal based upon Skip gram word2vec mechanism is proposed to attend original sentences, images, and captions when decoding. The text summarization process creates a summary by paraphrasing a long text. Earlier models on information retrieval and summarization are based on a massive labeled dataset by the use of handcrafted features, leveraging on knowledge for a particular domain, and concentrated on the narrow sub-domain to improve efficiency.**

***Index Terms*—Information retrieval, text summarization, deep learning, word2vec, dense captioning, Stanford, NLP**

I. INTRODUCTION

Summarizing multi-modal documents to get multi-modal summaries is becoming an urgent need with rapid growth of multi-modal documents on the Internet. Text-Image summarization is to summarize a document with text and images to generate a summary with text and images. The summarization approach is different from pure text summarization. It is also different from image summarization which summarizes an image set to get a subset of images. An image worths thousands of words (Rossiter, et al., 2012). Image plays an important role in information transmission. Incorporating images into text to generate text-image summaries can help people better understand, memorize, and express information. Most of recent research focuses on pure text summarization, or image summarization. Little has been done on text-image summarization.

To generate such a text-image summary, the following problems should be considered: How to generate the text part? How to measure the importance of images, and extract important images to form the image summary? How to align sentences with images? In this paper, we will work with a neural text-image summarization model based on the word2vec and attentional hierarchical Encoder-Decoder model to solve the above problems. In the decoding stage, we combine text encoding and image encoding as the initial state, and use the attentional hierarchical decoder which attends original sentences, images and captions to generate the text summary. Each generated sentence is aligned with a sentence, an image, or a caption in the original document. Based on the alignment scores, images are selected and aligned with the generated sentences. In the inference stage, we adopt the multi-modal beam search algorithm which scores beams based on bigram overlaps of the generated sentences and the attended captions.

II. OUR MAIN CONTRIBUTION

The main contributions are as follows (We):

1. Propose the text-image summarization task, and extend the existing work by collecting images and captions of each news from the Web for the task.
2. Propose using word2vec model of Gensim library trained on Twitter Dataset and SeaNMF model, which discovers the accurate topics from short-text documents.
3. Propose using tokenization and Fast Text with three multi-modal attentional mechanisms which attend the text and the images simultaneously when decoding.
4. Perform experimental evaluation using ROUGE and other key index parameters to show that attending images when decoding can improve text summarization, and that our model can generate informative image summaries.

III. OUR MAIN OBJECTIVE

[1] To propose an effective short-text topic model named MMTextModel by combining the word2vec model of Gensim library trained on Twitter Dataset and SeaNMF model, which discovers the accurate topics from short-text documents.

[2] To incorporate both internal and external corpora to discover semantic relationships between words, resulting in more related terms under a single topic.

[3] To process qualitative SA (Semantic Analysis) that demonstrates the system's overall effectiveness by finding the meaningful topics and assigning appropriate labels to each topic using the document-term matrix.

[4] To perform sufficient quantitative assessments on different short text real-world datasets to illustrate the enhanced performance of proposed MMTextModelover existing models.

IV. RELATED WORK

Approach techniques that have been used in realtime summarization are fuzzy-based and machine learning. An example of a method that uses a fuzzy-based approach is the fuzzy logic with classic Zadeh's calculus of linguistically quantified propositions [12] which addresses trend extraction and real-time problems where the results are superior in tnorm evaluation, but weak in semantic problems because the semantic results of other t-norms are unclear and unclear can be understood.

Fuzzy Formal Concept Analysis (Fuzzy FCA) [13] which addresses semantic and real time problems where the results excel at evaluaions in f-measures with optimal recall and comparable precision. An example of a method that uses a machine learning approach is Incremental Short Text Summarization (IncreSTS) by [14] which has better outlier handling, high efficiency, and scalability on target problems. Rank-biased precision-summarization (RBP-SUM) by [15] which has advantages in overcoming redundancy by evaluating using rouge, but this method can only produce extractive summaries.

Text summarization is a formidable challenge in the field of Natural Language Processing (NLP) [16] because it requires precise text analysis such as semantic analysis and lexical analysis to produce a good summary. A goodsummary, in addition, must contain important information and must be concise but also must consider aspects such as non-redundancy, relevance, coverage, coherence, and readability [17]; where to get all these aspects in a summary is a great challenge.

The review of papers on text summarization is important because summarizing extractive techniques has become a very broad research topic and is heading towards maturity [18]. Now research has shifted towards abstractive summarization [18] and real-time summarization. This is because abstractive summaries are more complex and complicated than extractive summaries.

So extractive summaries are easier to give expected and better results than abstractive summaries [19]. However, extractive summarization is also still in great demand as evident extractive research still exists in the last two years. This indicates the possibility that there are still opportunities or loopholes to improve.A clear literature study is demanded as a means for the advancement of research in the field of text summarization. Where literature studies are generally contained, analyzed, and compared in a review or survey paper.

Review paper made by [18] discusses popular components specifically about abstractive summarizing, such as research trends in the field of

abstractive summarization, general description of existing abstractive summarizing techniques, tools, and evaluations.

Other reviews were conducted by [20] gave a brief survey of the techniques of text summarization and specifically in Arabic. A survey conducted by [21] discussing text summarization focuses on the approach techniques and methods used in text summarization. [21] grouped approaches to statistics, machine learning, semantic-based, and swarm intelligence. Another survey was conducted by [22] which is about summarizing extractive texts that focus on unattended techniques, presents a list of strengths and weaknesses in a comparison table, alluding to a little about evaluations and future trends.

Some other review articles only cover smaller sections, for example only about approach techniques [21], the methods used, evaluations technique, or discuss the topic of extractive or abstractive text summarization [20].

V. SUMMARIZATION OF LITERATURE

Author(s)	Title	Methodology Proposed	Evaluation / Remarks
M. A. H. Khan, D. Bollegala, G. Liu and K. Sezaki	Multi-tweet Summarization of Real-Time Events	A method for multi-tweet summarization of an event. It allows the search users to quickly get an overview about the important moments of the event. Also proposed a graph-based retrieval algorithm that identifies tweets with popular discussion points among the set of tweets returned by Twitter search engine in response to a query comprising the event related keywords. To ensure maximum coverage of topical diversity, we perform topical clustering of the tweets before applying the retrieval algorithm.	Evaluation performed by summarizing the important moments of a real-world event revealed that the proposed method could summarize the proceeding of different segments of the event with up to 81.6% precision and up to 80% recall.
Bian, Y. Yang, H. Zhang and T. -S. Chua	Multimedia Summarization for Social Events in Microblog Stream	A framework to automatically generate visualized summaries from the microblog stream of multiple media types. Proposed framework comprises three stages, as follows 1) A noise removal approach is first devised to eliminate potentially noisy images. 2) A novel cross-media probabilistic model, termed Cross-Media-LDA (CMLDA), is proposed to jointly discover subevents from microblogs of multiple media types. 3) Finally, based on the cross-media knowledge of all the discovered subevents, a multimedia microblog summary generation process is designed to jointly identify both representative textual and visual samples, which are further aggregated to form a holistic visualized summary.	Conducted extensive experiments on two real-world microblog datasets to demonstrate the superiority of the proposed framework as compared to the state-of-the-art approaches.
Dutta, V. Chandra, K. Mehra, A. K. Das, T. Chakraborty and S. Ghosh	Ensemble Algorithms for Microblog Summarization	Proposed ensemble schemes that can combine the outputs of multiple base summarization algorithms, to produce summaries better than what is generated by any of the base algorithms.	Investigated whether off-the-shelf summarization algorithms can be combined to produce better quality summaries.
Saini, S. Saha and P.	Multiobjective-Based Approach	Employed the concept of multiobjective optimization in microblog summarization to produce good quality summaries. Different types of genetic operators including recently developed self-organizing map (a type of neural	Different statistical quality measures namely, length, tf-idf score of the tweets, antiredundancy, measuring different aspects of summary, are

Bhattacharya	for Microblog Summarization	network) based operator, are explored in the proposed framework. To measure the similarity between tweets, word mover distance is utilized which is capable of capturing the semantic similarity between tweets.	optimized simultaneously using the search capability of a multiobjective differential evolution technique.For evaluation, four benchmark data sets related to disaster events are used, and the results obtained are compared with various state-of-the-art techniques using ROUGE measures. It has been found that our algorithm improves by 62.37% and 5.65% in terms of ROUGE-2 and ROUGE-L scores, respectively, over the state-of-the-art techniques. Results are also validated using statistical significance t-test.
G. Liang, W. He, C. Xu, L. Chen and J. Zeng	Rumor Identification in Microblogging Systems Based on User's Behavior	Investigated the machine-learning-based rumor identification approaches.Observed that feature design and selection has a stronger impact on the rumor identification accuracy than the selection of machine-learning algorithms. Also investigated rumor identification schemes by applying five new features based on users' behaviors, and combine the new features with the existing well-proved effective user behavior-based features, such as followers' comments and reposting, to predict whether a microblog post is a rumor.	Experiment results on real-world data from Sina Weibo demonstrate the efficacy and efficiency of proposed method and features. From the experiments, authors conclude that the rumor detection based on mass behaviors is more effective than the detection based on microblogs' inherent features.
Madisetty and M. S. Desarkar	A Neural Network-Based Ensemble Approach for Spam Detection in Twitter	Proposed an ensemble approach for spam detection at tweet level. Developed various deep learning models based on convolutional neural networks (CNNs). Five CNNs and one feature-based model are used in the ensemble. Each CNN uses different word embeddings (Glove, Word2vec) to train the model. The feature-based model uses content-based, user-based, and n-gram features. Theapproach combines both deep learning and traditional feature-based models using a multilayer neural network which acts as a meta-classifier.	Authors have evaluated method on two data sets, one data set is balanced, and another one is imbalanced. The experimental results show that our proposed method outperforms the existing methods.
Sedhai and A. Sun	Semi-Supervised Spam Detection in Twitter Stream	The proposed framework consists of two main modules: spam detection module operating in real-time mode and model update module operating in batch mode. The spam detection module consists of four lightweight detectors: 1) blacklisted domain detector to label tweets containing blackli2) near-duplicate detector to label tweets that are near-duplicates of confidently prelabeled tweets; 3) reliable ham detector to label tweets that are posted by trusted users and that do not contain spammy words; and 4) multi classifier-based detector labels the remaining tweets. The information required by the detection module is updated in batch mode based on the tweets that are labeled in the previous time window.	Experiments on a large-scale data set show that the framework adaptively learns patterns of new spam activities and maintain good accuracy for spam detection in a tweet stream.
Deb, A. Pratap, S. Agarwal and T.	A fast and elitist multiobjective genetic algorithm: NSGA-II	Multi-objective evolutionary algorithms (MOEAs) that use non-dominated sorting and sharing have been criticized mainly for: (1) their O(MN/sup 3/) computational complexity (where M is the number of objectives and N is the population size); (2) their non-elitism approach; and (3) the need to specify a sharing parameter. In this paper, we suggest a non-dominated sorting-based MOEA, called NSGA-II (Non-dominated Sorting Genetic Algorithm II), which alleviates all	Simulation results on difficult test problems show that NSGA-II is able, for most problems, to find a much better spread of solutions and better convergence near the true Pareto-optimal front compared to the Pareto-archived evolution strategy and the strength-Pareto evolutionary

Meyarivan		of the above three difficulties. Specifically, a fast non-dominated sorting approach with O(MN/sup 2/) computational complexity is presented.	algorithm - two other elitist MOEAs that pay special attention to creating a diverse Pareto-optimal front. Moreover, we modify the definition of dominance in order to solve constrained multi-objective problems efficiently. Simulation results of the constrained NSGA-II on a number of test problems, including a five-objective, seven-constraint nonlinear problem, are compared with another constrained multi-objective optimizer, and the much better performance of NSGA-II is observed.
Johnson, A. Karpathy and L. Fei-Fei	DenseCap: Fully Convolutional Localization Networks for Dense Captioning	Authors introduced the dense captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. To address the localization and description task jointly authors proposed a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single, efficient forward pass, requires no external regions proposals, and can be trained end-to-end with a single round of optimization. The architecture is composed of a Convolutional Network, a novel dense localization layer, and Recurrent Neural Network language model that generates the label sequences.	We evaluate our network on the Visual Genome dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. We observe both speed and accuracy improvements over baselines based on current state of the art approaches in both generation and retrieval settings.
Yang, K. Tang, J. Yang and L. -J. Li	Dense Captioning with Joint Inference and Visual Context	Dense captioning is a newly emerging computer vision topic for understanding images with dense language descriptions. The goal is to densely detect visual concepts (e.g., objects, object parts, and interactions between them) from images, labeling each with a short descriptive phrase. We identify two key challenges of dense captioning that need to be properly addressed when tackling the problem. First, dense visual concept annotations in each image are associated with highly overlapping target regions, making accurate localization of each visual concept challenging. Second, the large number of visual concepts makes it hard to recognize each of them by appearance alone.	We propose a new model pipeline based on two novel ideas, joint inference and context fusion, to alleviate these two challenges. We design our model architecture in a methodical manner and thoroughly evaluate the variations in architecture. Our final model, compact and efficient, achieves state-of-the-art accuracy on Visual Genome for dense captioning with a relative gain of 73% compared to the previous best algorithm. Qualitative experiments also reveal the semantic capabilities of our model in dense
L. Cilibrasi and P. M. B. Vitanyi	The Google Similarity Distance	Author uses the World Wide Web (WWW) as the database, and Google as the search engine. The method is also applicable to other search engines and databases. This theory is then applied to construct a method to automatically extract similarity, the Google similarity distance, of words and phrases from the WWW using Google page counts. The WWW is the largest database on earth, and the context information entered by millions of independent users averages out to provide automatic semantics of useful quality.	Author uses the WordNet database as an objective baseline against which to judge the performance of our method and conducts a massive randomized trial in binary classification using support vector machines to learn categories based on our Google distance, resulting in a mean agreement of 87 percent with the expert crafted WordNet

VI. CONCLUSION

Summarising texts is an intriguing study area in the NLP field that aids in the production of succinct information. The purpose of this paper is to present the most recent research and progress in this field using the proposed method; the goal is to demonstrate that it can offer a more organised, broad, and different review, covering everything from trends/topics, data sets, preprocessing, characteristics, approach techniques, difficulties, methods to evaluation that can be used as a guide for future work; the relationship between trends/topics, problems, and hurdles in each topic, technique, and method used is spelled out.

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