Grenze International Journal of Engineering and Technology, Jan Issue

Harmonizing Vision and Voice: A Review of Contemporary Research in Image Caption Generation and Text-to-Speech Synthesis

Yash Anand[1], Rudra Rawat[2], Aditya Sahu[3], Aayush Tandon[4] and Vaibhav E. Narawade[5]
[1-5]Ramarao Adik Institute of Technology (DY Patil Deemed to be University) Department of Computer Engineering
Email: vaibhav.narawade@dypatil.edu

***Abstract*— Automatically describing the content of a picture with meaningful and contextually appropriate textual descriptions is a difficult challenge at the interface of computer vision and natural language processing. This review paper provides a thorough summary of recent developments in picture caption generating methods, including both conventional strategies and cutting-edge deep learning-based solutions. We examine the changes in datasets, evaluation criteria, and model designs that have influenced this field's advancement. In order to improve the standard and variety of generated captions, we also integrate visual attention processes, transformer-based models, and reinforcement learning techniques. The focus of this study is on the algorithmic overlap between text to speech and visual.**

***Index Terms*— Deep learning, convolutional neural network, image captioning, text to speech, long short term memory, generative adversarial network.**

I. INTRODUCTION

Image captioning is a complex and challenging task that requires synthesis of visual and semantic information obtained from input images. To overcome this problem, [1] describes an innovative approach by introducing a unified multi-layer architecture. This new model combines bottom-up and top-down approaches in a visual aesthetic model that can instill habits that combine visual image-level features with semantic-level features to create brand images. In addition to image captions, image and video captions are also a big problem, as explained in [2]. These activities require not only understanding the visual content but also creating similar descriptions. The emergence of deep learning has varied across regions, resulting in models that deliver cutting-edge results. A standard subtitle model architecture typically includes an encoder that extracts visual features from the input image or video using a convolutional neural network (CNN). A decoder (often used as a convolutional neural network (RNN), such as a short-term neural network (LSTM)) then generates descriptive text. The impact of this technology is wide-ranging, from helping the visually impaired to improving video usability through subtitles to improving the accuracy of visual search engines. However, the current image collection method is described in [3]. This process addresses the problems of inadequate and misinformation. To alleviate these frustrations, the authors presented a successful R-LSTM model that proposes a new integrative concept by providing educational and generational visual material. Once trained, the model assigns different weights to words based on accuracy, part of speech, and context. During rendering, this information is used to reduce false positives and generate more

Grenze ID: 01.GIJET.10.1.116

descriptive information. The power of the R-LSTM model becomes evident when it demonstrates its ability to transform space, reaching performance levels of state-of-the-art benchmarks such as MS COCO and Flickr30k.
The review section [6] investigates the complexity of image captioning and highlights its importance for people with visual impairments. ECANN's model provides the ability to display better images by using deep learning as the light source. The ECANN model adds dimensionality to accessible images by generating multiple names for online images and using an iterative strategy to choose the most appropriate name.
Finally, the development of natural language processing (NLP) takes the middle of [12], where the evolution of deep learning is geared towards exploring depth. Covering the core areas and practical applications of NLP, this comprehensive study demonstrates the far-reaching results of deep neural networks equipped with thousands of parameters, GPU acceleration, and access to very large files. This research provides an example to guide researchers and practitioners throughout the evolution of NLP, provides important insights into the integration of information driven by computing and deep learning, and demonstrates the next great thing.

II. Related Work

A famous paper [1] describes the architecture of Stack-VS, a multi-step process for generating detailed graphs. We use bottom-up and top-down approaches to create effective visuals and content. Complex decoder model improves subtitle quality. Another paper [2] discusses the potential of deep learning for image and video captioning. It demonstrates applications such as solving cognitive problems related to image and video captioning and helping the visually impaired. The paper [3] shows how to use short term memory (R-LSTM) to improve sentences using reference data. This reduces rejection issues and improves subtitle quality. A vision-based motion-assisted navigation system for the visually impaired was proposed in [4]. Immediately detect dynamic problems and adjust plans to improve safety. [5] proposed an inference model that generates complex, natural sentences from example images, emphasizing accuracy and linguistic quality. [6] presented a technique using deep learning to help visually impaired people recognize food while shopping online. [7] presented a Hierarchical Attention Fusion (HAF) model for RL-based images that combines multi-image mapping and compensation weights. [8] provides a comprehensive review of image matching techniques, traces their development, and presents their core components. The paper [3] conducted an in-depth study on the use of weight training and strength training and showed that strength training can improve performance standards when used as a benchmark for the effectiveness of R-LSTM models in generating accurate images. . . In Noun Promotion [6], the authors proposed an Evaluation Oriented Neural Network (ECANN) for automatic labeling of image names. Extensive testing shows that ECANN consistently outperforms existing deep learning models, achieving high accuracy and precision, especially when feature extraction is used. In [7], the authors introduced the Hierarchical Attention Fusion (HAF) model, which shows its best performance in entropy competition and learning Table 1. Depiction of different models of image caption support, referring to the ability to make documents proposed to generate the signature image. Additionally, this article introduces the concept of multi-reward granularity, which improves the efficiency of generated scripts. Taken together, these data demonstrate further advances in display technology, highlighting the importance of integrating multiple models of attention and reward into training models to create accurate graphics and relevant content. [11] presented a listening experiment showing the success of the GAN-based model compared to the acoustic model of spoken language. [17] conducted a preliminary study on TTS speech, demonstrating the effectiveness of VTN and RNN in both objective and evaluation methods. Latent representation visualization supports TTS pre-training execution. [12] briefly mentions NLP applications, focusing on sentiment analysis using LSTM and CNN. In [13], many teachers recognize that TTS uses distilled structures, thus developing naturalness, stability, and expressiveness, and many teachers' approaches show that this is important. These studies illustrate the diverse applications of natural language processing and deep learning.

III. Methodology

[1] The authors solve the problem of automatic image tagging in the field of multimodal translation by proposing a multi-layer architecture called “Stack-VS”, which is used to generate detailed and beautiful information. In order to accomplish this, top-down and bottom-up browsing models are combined in order to improve the input image's visual and semantic levels. A new stack decoder model with some decoder layers is also introduced. Each block has two LSTM layers that work together to adjust the placement weights of both layers. This is the best way to improve image quality. The effectiveness of this scheme has been confirmed through extensive testing on the well-designed MSCOCO dataset. [2] discussed the potential of deep learning to generate labels for images and videos.

The challenges that arise in captioning images and videos are considered a difficult skill in image research. It explains problems and applications, such as helping the visually impaired and improving search engine indexing. The paper [3] introduces a new method called R-LSTM (Link Long-Term Memory) in the encoder-decoder framework to solve this problem. This method attempts to improve the proposal using reference material. In addition, it reduces misrecognition problems by maximizing the consistency of the subtitle information generated by the subtitle model and the surrounding images. The suggested system's superiority was verified by extensive testing and comparisons with the results of the MS COCO and Flickr30k tests. Analysis of the recordings showed that the model gained the ability to combine the information used to understand the underlying content of the image and thereby create and influence meaningful messages. [4] demonstrated a cutting-edge integrated vision-based mobile navigation system made to help blind individuals find their way autonomously indoors.

Model	Architecture	Dataset & Evaluation	Comment
(2020, L. Chen et al.)	Multi-stage architecture (called Stack-VS) for image caption generation.	MSCOCO:- BLEU-4 / CIDEr / SPICE scores are 0.372, 1.226 and 0.216	Accuracy scores are in standardized metrics like Consensus-based Image Description Evaluation (CIDEr). Accuracy based compared to the labeled dataset is needed.
. (2020, Amirian et al.)	CNN and RNN,. Long-term Recurrent Convolutional Networks (LRCNs).	Charades, MSVD, ActivityNET captions, VideoStrory, MPII, M-VAD	System suffered from inaccuracies due to fundamental constraints; resulting in limited use in practical situations.
. (2019, Ding et al.)	Encoder-decoder framework, named Reference based Long Short Term Memory (R-LSTM)	MS COCO and Flickr30k:- 10.37% enhancement in terms of CIDEr on MS COCO.	It is only designed for single-image captioning.
(2018, B. Li et al.)	Time-stamped map Kalman filter (TSM-KF) algorithm	Google Maps, HERE Maps and AutoNavi Maps:- The subjects learned to use ISANA fairly easily.	System requires a high-quality RGB-D camera, which can be expensive, further testing is needed.
(2017, O. Vinyals et al.)	LSTM-Based Sentence Generator.	Pascal VOC,Flickr 8k, Flickr30k,MSCOCO SBU:- CIDER=0.854, METEOR=0.252, ROGUE=0.484,BLEU-4=0.217	The produced descriptions are one of many possible image interpretations, requiring better image resolution.
(2022, Tiwary et al.)	ARO,SDM (Spatial Derivative and Multi-scale),WPLBP (Weighted Patch Local Binary Pattern) ,ECANN (Extended Convolutional Atom Neural Network).	Flickr30k Freiburg Groceries, Grocery Store Datasets:- 99.46% on Grocery Store Dataset and 99.32% accuracy on Freiburg Groceries dataset.	This model focuses mainly on the grocery store items and does not perform well on other subjects.
(2020, C. Wu et al.)	A Hierarchical Attention Fusion (HAF), multi-level feature maps of Resnet, (REN)	MSCOCO 2014 caption dataset:- CIDeR=116.4, BLEU-1=80.5, BLEU-2=62.9, BLEU-3=47.7, BLEU-4=35.5, METEOR=27.3	The hierarchical attention mechanism and the multi-grained reward function are both computationally expensive.
(2021, Ma, J et al.)	Handcrafted methods that are trainable ones.	Lip6 Indoor, SUN3D, YFCC100M, Oxford Buildings:- Highest Accuracy 0.8986 on Character 2	The paper is a survey paper and does not provide detailed experimental results.

(2018, Karim et al.)	Fully convolutional neural networks (FCNs) long short-term memory recurrent neural network (LSTM RNN)	short-sequence UCR datasets:-0.9541	The model is not able to handle long time series. The model is limited to a time series of up to 1000 data points.
(2019, E. K. Wang et al.)	Faster recurrent convolutional neural network (Faster R-CNN),long short-term memory (LSTM)	Flickr 8K and Flickr 30K Microsoft COCO Caption Data Set:-	The model is not able to handle images with multiple objects.
(2020, Y. Zhao et al.)	End-to-end (E2E)speech-to-translation speech-to-image retrieval.	FlickR-real speech, Mboshi:- The PER of 70.4% for the same-language system (and 71.7% for the cross-language system).	Need to improve the quality of the discovered speech, image and translation encodings, finding the optimal acoustic feature set for the end-to-end systems
. (2021, D. W. Otter et al.)	sequence-to-sequence (seq2seq) voice conversion (VC) models appealing.	LibriSpeech dataset CMU ARCTIC dataset:- The CER and WER for the ground-truth validation set were 0.9% and 3.8%, respectively.	In the absence of appropriate data, seq2seq VC models may suffer from unstable training and mispronunciation issues.
. (2019, R. Liu et al.)	Automated speech recognizer (ASR)'s., Tacotron2	Speech corpus that contains 10 hours high-quality speech from a Chinese female speaker who mimics a little girl speaking	Tacotron2 has limitations, including challenges with rare or complex words, difficulty handling background noise..
(2022, Tjandra et al.)	RNN-based speech synthesis systems that combine the WaveNet vocoder with conditional Generative Adversarial Networks (GAN).	Custom high-quality multi-speaker synthesizing male voices	Need to improve multi-speaker text-to-speech synthesis systems by incorporating Wasserstein GAN and waveform loss-based acoustic model training.
(2021, D. Seo et al.)	Effective keyword spotting (KWS)	Google Speech Command Datasets V1 and V2, LibriSpeech 960 hours	The study only evaluated the method on a single task, keyword spotting. It is possible that the method would not be as effective for other tasks.

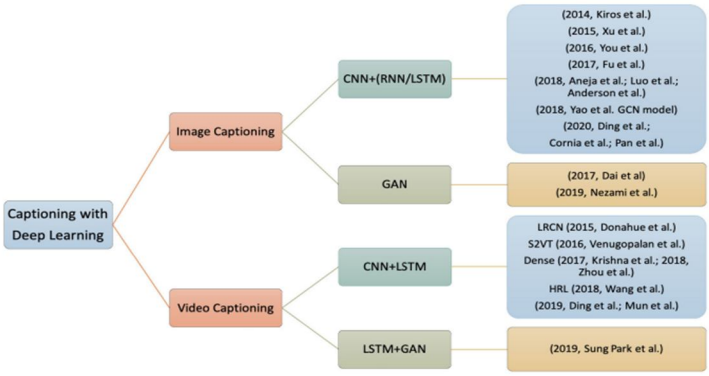

Fig. 1. LSTM generative adversarial network [(2020, Amirian et al.)].

The authors create a matching algorithm using the potential of Visual Positioning Service (VPS) on Google Tango devices. This algorithm facilitates semantic localization by bridging the gap between field-of-view description (ADF) data and semantic mapping. A Hierarchical Attention Fusion (HAF) model is introduced in [7] as a foundation for RL-based picture captioning. This model combines a hierarchical attention mechanism with Resnet's multi-level feature maps. The authors present the Revaluation Network (REN) for CIDEr grade re-evaluation. This involves assigning different weights to words based on their importance in the generated title. This weight compensation is treated as word-level compensation. We also use a Scoring Network (SN) to evaluate the generated sentences based on the truth value for the set of subtitles. This allows you to get rewards for different facts and keep track of your reward levels. In order to enhance time series categorization, we also investigate the incorporation of monitoring approaches. This monitoring aids in increasing efficiency and gives information about how the LSTM department makes decisions. Later, a different researcher presented a novel technique for acquiring images and suggested a layered thickness model. We employ Faster R-CNN (Faster Recurrent Convolutional Neural Network) to extract picture characteristics depending on coding level in order to accomplish this aim. The decision-making process involves generating descriptors using short-term transformations (LSTMs). The novelty of this approach is that it uses density mechanisms at the coding level. [11] The main purpose of this study is, firstly, to obtain a representation of meaning from speech, without using the text as a means of representation, and secondly, to evaluate the effectiveness of the representations obtained by speaking or interpreting the text and preserve the spoken word. [12] This article discusses the effectiveness of sequence-to-sequence (seq2seq) speech conversion (VC) models along with the ability to transform speech. However, these models can face problems such as unbalanced training and mispronunciation when dealing with limited resources, making them impossible. The notion suggests transferring data from other data-rich speech processes, such as text-to-speech (TTS) and automated speech recognition (ASR), to address these issues. The justification offered for this method is that VC models, initialized with pre-trained ASR or TTS model parameters, may provide efficient hidden representations for producing high-fidelity and highly understandable translated speech. [13] This research paper addresses these challenges.

IV. Conclusion

The collection of research papers covers a diverse range of topics in the fields of artificial intelligence, computer vision, natural language processing, and assistive technologies. These papers present innovative solutions to complex challenges and push the boundaries of their respective domains. Automatic image caption generation is taken to new heights with the "Stack-VS" architecture, which combines bottom-up and top-down attention models to generate detailed image descriptions. Deep learning's potential in generating multimedia captions is explored, highlighting the intellectual rigor required and its applications in aiding visually impaired individuals and improving search engine indexing. In the domain of assistive technologies, innovative solutions are presented, including a vision-based mobile navigation system for the visually impaired, dynamic obstacle detection, and path planning. A comprehensive survey of image matching techniques is provided, tracing their evolution and applications. Furthermore, research focuses on enhancing fully convolutional networks with LSTM modules for time series classification, emphasizing the importance of attention mechanisms. Innovative approaches to image captioning involve the use of dense attention mechanisms to selectively generate descriptive text. Exploration of

speech technology for unwritten languages and solutions to voice conversion challenges are also featured in the collection. The effectiveness of knowledge distillation in neural text-to-speech is highlighted, providing improved naturalness and stability in synthesized speech. Lastly, a deep neural network approach for constructing keyword spotting systems for voice-enabled devices with limited training data is presented. These papers collectively advance academic and practical knowledge, offering creative solutions to complex challenges in their respective fields.

REFERENCES

[1] Stack-Vs: Stacked Visual-Semantic Attention for Image Caption Generation, Ling Cheng, Wei Wei, Xianling Mao, Yong Liu, (Member, IEEE), And Chunyan Miao, (Senior Member, IEEE) (2020)

[2] Automatic Image and Video Caption Generation with Deep Learning: A Concise Review and Algorithmic Overlap, Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, and Hamid R. Arabnia (2020)

[3] Neural Image Caption Generation with Weighted Training and Reference, Guiguang Ding · Minghai Chen· Sicheng Zhao· Hui Chen· Jungong Han Qiang Liu. (2019)

[4] Vision-Based Mobile Indoor Assistive Navigation Aid For Blind People Bing Li, Member, IEEE, J. Pablo Munoz, Member, Ieee, Xuejian Rong, Qingtian Chen, Jizhong Xiao, Senior Member, Ieee, Yingli Tian, Fellow, IEEE, Aries Arditi, Mohammed Yousuf. (2018)

[5] Show And Tell: Lessons Learned from the 2015 Mscoco Image Captioning Challenge Oriol Vinyals, Alexander Toshev, Samy Bengio, And Dumitru Erhan. (2017)

[6] An Accurate Generation of Image Captions for Blind People Using Extended Convolutional Atom Neural Network Tejal Tiwary & Rajendra Prasad Mahapatra. (2022)

[7] Hierarchical Attention-Based Fusion for Image Caption with Multi-Grained Rewards Chunlei Wu, Shaozu Yuan, Haiwen Cao, Yiwei Wei, And Leiquan Wang. (2020)

[8] Image Matching from Handcrafted to Deep Features: A Survey Jiayi Ma· Xingyu Jiang· Aoxiang · Junjun Jiang· Junchi Yan. (2021)

[9] Multilayer Dense Attention Model for Image Caption Eric Ke Wang, Xun Zhang, Fan Wang, Tsu-Yang Wu, And Chien-Ming Chen. (2019)

[10] Lstm Fully Convolutional Networks for Time Series Classification Fazle Karim, Somsubhra Majumdar, Houshang Darabi, (Senior Member, IEEE), and Shun Chen. (2018) Wasserstein Gan And Waveform Loss-Based Acoustic Model Training for Multi-Speaker Text-To-Speech Synthesis Systems Using

[11] A Wavenet Vocoder Yi Zhao, (Student Member, Ieee), Shinji Takaki2, (Member, IEEE),Hieu-Thi Luong2 , (Student Member, IEEE), Junichi Yamagishi2,3, (Senior Member, IEEE),Daisuke Saito1, (Member, IEEE), And Nobuaki Minematsu1, (Member, IEEE). (2020)

[12] A Survey of The Usages of Deep Learning ForNatural Language Processing Daniel W. Otter, Julian R. Medina, And Jugal K. Kalita. (2020)

[13] Decoding Knowledge Transfer For Neural Text-To-Speech Training Rui Liu, Member, Ieee, Berrak Sisman , Member, IEEE, Guanglai Gao, And Haizhou Li , Fellow, IEEE. (2022)

[14] Machine Speech Chain Andros Tjandra, Sakriani Sakti, Member, IEEE, And Satoshi Nakamura, Fellow, IEEE. (2021)

[15] Pre-Alignment Guided Attention For Improving Training Efficiency And Model Stability In End-To-End Speech Synthesis Xiaolian Zhu, Yuchao Zhang, Shan Yang, Liumeng Xue, And Lei Xie, (Senior Member, IEEE). (2019)

[16] Pre-Training Techniques For Sequence-To-SequenceVoice Conversion Wen-Chin Huang, Student Member, Ieee, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, Senior Member, Ieee, And Tomoki Toda, Senior Member, IEEE. (2021)