2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) | 979-8-3315-4362-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSES63760.2024.10910381

Enhanced Multilingual Image Captioning: Integrating Text-to-Speech Translation and Sentiment Analysis

Tanuja Konda Reddy, Veeksha S, Kavitha C.R.

Department of CSE, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India

E-mail: k.tanujareddy@gmail.com, veeksha2702@gmail.com, cr_kavitha@blr.amrita.edu

Abstract- The aim of this project is to create a system that will improve the way we interact with text and images by using the capabilities of computer vision and natural language processing. The main purpose of using this system to generate human-readable image captions is to train a model that can accurately describe an image's contents. The project extends the functionality of the image captioning system by adding language translation inorder to give a multilingual support. The use of a few pre-trained models will enable users with varying language backgrounds use this system to acquire captions in different target languages. This involves integrating language translation techniques into the system and training different models for every target language. Further we extend the project by working in the field of sentiment analysis by identifying the sentiment (positive, negative, or neutral) of captions by examining text and images. For a better understanding of the emotional tone of the information, the system combines CNNs with natural language processing (NLP) algorithms for image analysis and text analysis.

Index Terms—Computer Vision, Natural Language Processing, Image Captioning, Multilingual Support, Sentiment Analysis.

I. INTRODUCTION

In the field of computer vision and natural language process- ing, Automatic Multilingual Image Captioning has emerged as a new area of study in the past few years. The major goal of this method is to develop systems that can both generate the human-readable captions in multiple languages and extract and understand images and visual media better. The primary goal of this project is to create an automatic multilingual image captioning system that can generate captions in South Indian languages in addition to English.

The idea makes use of deep learning models such as Convo- lutional Neural Networks (CNNs) for extractingfeatures from images and sequence-to-sequence models based on Trans- formers or Recurrent Neural Networks (RNNs) for producing English captions. A few pretrained models like M-Bart and also a 3-Layer transformer is used to translate captions in English to different languages.

Besides, the system also includes Text-to-Speech (TTS) synthesis technology, which allows the creation of spoken captions in English. Moreover, the Sentiment Analysis part is involved in the Sentiment Analysis component that is used to analyze the sentiment that is expressed by the generated captions, which provides the oppurtunity of using the gener- ated captions in the areas of marketing, advertising, and social media analysis.

This study aims to encourage the development of multi-lingual image understanding and communication by the use of the different NLP, computer vision and deep learning models, including CNNs, RNNs, Transformers, and the latest architectures, such as ResNet, Inception, and Tacotron 2. By using these technologies, this project hopes to achieve its goals.

II. LITREATURE SURVEY

[1]. In this paper the autor introduces a multi-layered CNN- LSTM neural network model, along with a fine-tuned cus- tomized VGG16 model and utilizing larger datasets. This model integrates object recognition and incorporates an atten- tion module. Results show a really good improvement, with a 34.64% and 29.13% increase in BLEU score for both unigram and bigram evaluations, respectively.

[2]. In this paper the authors talk about the use of CNN, RNN, and LSTM architectures in their image captioning system. All of these models are known to be good at process- ing sequential input, identifying image features, and creating meaningful captions. Better information transfer is facilitated by the implementation of enhanced attention mechanisms, such as multi-head attention or self-attention for recording complicated data linkages. In addition, various feature extrac- tion methods are studied to enhance the system's accuracy and contextual description of images.

[3]. This paper talks about how a variety of models, including RNN, CNN, attention mechanisms, Transformer- based models, and graph models, are

employed. These models extract high-level feature from raw images, enable image captioning, and improve the quality of generated captions. Reinforcement learning (RL) strategies are utilized to optimize image caption models. There's also a demand for enhanced deep frameworks and reinforcement learning techniques to further improve captioning.

[4]. The authors of this paper talks about the usage of advanced NLP techniques and the Hugging Face Transformers framework. The system employs the MarianMTModel for machine translation, mT5 for multilingual text summarization, and Hugging Face for translation tasks. The Challenges faced include preserving the original meaning, cultural nuances, and intended purpose of content across various languages. There's a constant demand for advancements in maintaining the quality and contextual relevance of content during translation and summarization processes across a wide spectrum of languages.

[5]. Upon object detection and recognition, text describing the object undergoes conversion into speech output via text-to- speech (TTS) technology. Implemented in Python, the system utilizes the YOLOv4 object detection model trained on the MS COCO dataset. It integrates the Google Text-to-Speech (gTTS) API alongside the iPython audio library to generate natural language audio output from detected text. Efforts focus on enhancing processing speed for large texts and expanding language support to include languages like Assamese. Addi- tionally, future developments aim to incorporate recognition capabilities for images or diagrams within textual content, as the current system exclusively recognizes text.

[6]. A CNN-BILSTM model is crafted specifically for sign language gesture recognition from video inputs, effectively mapping them to corresponding gloss labels or word se- quences. This model employs CNN LSTM to translate these gestures into text representations. However, the scarcity of training data poses a significant challenge, particularly when attempting to fully train deep neural networks with intricate architectures for this task. Efforts are directed towards address- ing this bottleneck to enable more comprehensive training and improve the model's performance in sign language gesture recognition.

[7]. For the synthesis of a text-to-speech system, a combination of models is employed, including stacked convolutional autoencoder and Sequence of Multi-Label Extreme Learning machines (ELM). The stacked convolutional autoen- coder comprises multiple layers of convolutional autoencoders, while the Sequence of Multi-Label ELM falls under the umbrella of ELM-based machine learning approaches. Despite advancements, enhancing the naturalness and intelligibility of synthesized speech remains challenging. There's a pressing need to develop models capable of accommodating diverse linguistic characteristics and accents to broaden the system's applicability and usability.

[8]. EfficientTTS 2 (EFTS2) introduces a one-stage, high-quality end-to-end Text-to-Speech (TTS) framework, targeting superior speech synthesis and voice conversion performance while maintaining competitive model efficiency. It contrasts EFTS2 with other text-to-waveform models like VITS, EFTS- wav, and FastSpeech2s, emphasizing its strengths. Traditional two-stage TTS systems face limitations, prompting the call for enhanced training processes. EFTS2 addresses these chal- lenges, offering a streamlined approach for achieving high-quality TTS and voice conversion without compromising ef- ficiency, thus demonstrating significant advancements in the field.

[9]. Machine learning algorithms such as SVM, Naive Bayes, Maximum Entropy, and kNN rely on labeled training data for sentiment analysis tasks. In contrast, rule-based methods utilize predefined sentiment rules and scoring schemes for classification. Techniques like TF-IDF, GloVe, fastText, and word2vec are commonly used for text representation and feature extraction. To capture nuances and subtle emotions, deeper semantic analysis methods are being incorporated into sentiment analysis models. Furthermore, there's a growing need for building larger annotated datasets spanning diverse domains, languages, and demographic groups to enhance the performance and generalizability of sentiment analysis sys- tems.

[10]. Sentiment analysis methods encompass lexicons, tokens, Bayesian approaches, and bag of words techniques. Re- cent advancements introduce more sophisticated models like auto-regressive and encoder-decoder transformers, including large language models such as GPT-3 and GPT-J. Pre-trained models like BERTweet, RoBERTa, and FastText are also gaining traction. However, there's a gap in understanding the operation, capacity, and scope of sentiment analysis methods. To address this, comprehensive testing and benchmarking are essential to guide practitioners and researchers in selecting the most suitable approaches for specific sentiment analysis tasks.

III. METHODOLOGY

Dataset

This research uses the "Flickr8k" dataset: https://www.kaggle.com/datasets/adityajn105/flickr8k

This dataset, which includes 8000 photos, was hand-picked to show a range of settings and circumstances rather than any famous persons or locations.

Data Preprocessing

Image Captioning

To ensure that the model was trained as effectively as possible, the dataset was meticulously prepared, including a comprehensive data pretreatment procedure. A pretrained VGG16 ConvNet, ResNet50 and InceptionV3 was used to make it easier to convert each of these photos into a reliable 4096-dimensional feature vector. After that, the captions underwent a number of editing steps, such as making all of the letters lowercase and moving the numbers and special symbols to avoid using unnecessary spaces. Adding contextually necessary start-end sequence tokens is also part of copy 2b. The efficient tokenization process made possible by the Keras Tokenizer API ultimately made it easier to put together a vocabulary that included 8793 unique terms [11]-[13].

Language Translation

The preprocessing involves defining vocabularies for English and Kannada containing all unique characters found in the respective datasets, along with special tokens like START, PADDING, and END.To enhance the mapping between text and numeral there some relationship between token and index number. The English and Kannada sentences are preprocessed along with the text files containing the complete dataset, up to certain limit of total sentences. The sentences in the English language are preprocessed by converting it into lower case and eliminating new line space in the same manner, but in Kannada only the new line spaces are eliminated. After the preprocessing step, the incoming sentence data is now fit for other pre-processing steps like tokenization, padding, and getting ready for loading into the translation model.

Sentiment Analysis

Data pre-processing being the vital initial step of natural language processing conveys the idea that it is used to remove the noise from the raw text data before moving forward with further analysis or modeling. This approach involves the application of the NLTK library [16] together with some corpora (databases) in the programming language, Python.[14] [15] In the first place the text is delivered to the sentences and words in the 'punkt' corpus, which makes it easier to break the text into smaller objects by analyszing them. The next step is to remove the stop-words, which in this case are the common words like 'the','a','is, etc'. This is used 'stopwords' corpus, as they don't have much bearing on the sense of the text. After that, the PorterStem algorithm, that ultimately derives words to their root form, proceeds stemming. g., 'running' becomes 'run'). The processing is also done with WordNetLemmatizer and wordnet which lemmatize the words to their stem forms and take the sense of the word too [16].

MODULES:

Image Processing Module:

The Image Processing Module is responsible for preparing the input images for the deep learning model. It handles tasks like resizing, feature extraction, and normalization using libraries like OpenCV.

- Resize images to a standard format to ensure consistency across the dataset.
- Normalize pixel values to a common scale.
- Apply appropriate transformations and feature extraction techniques.
- Extract high-level visual features from images using pre-trained Convolutional Neural Networks (CNNs) like ResNet-50, VGG16, or Inception V3.

Deep Learning Module:

The Deep Learning Model is the core component responsible for generating image captions. It typically combines models like CNN for image feature extraction and RNN, for sequence generation.[1]

CNN Layers:

- Extract hierarchical visual features from preprocessed images. LSTM/Transformer Layers:
- Generate sequential captions by capturing temporal dependencies in the feature vectors. Training Process:
- Train the model using multilingual datasets of images and their corresponding captions.
- Use backpropagation and optimization techniques (e.g., Adam) to optimize the model's parameters.

Loss Function:

- Define a loss function to measure the difference between predicted captions and ground truth captions (e.g., Cross- Entropy Loss).

Model Checkpointing:

- Store the trained model for future use or further fine-tuning.

Natural Language Processing (NLP) Module:

The NLP Module enhances the understanding of generated captions in the source language (English). It involves tokenization, part-of-speech tagging, named entity recognition, and other linguistic analyses.

Tokenization:

- Break down the captions into individual pieces of words or tokens.

Part-of-Speech Tagging:

- Determine each word's grammatical category (noun, verb, etc.). Named Entity Recognition (NER):
- Identify entities like locations, people, organizations, etc., within the captions.

Lemmatization:

- Reduce words to their base or root forms. Sentiment Analysis
- Determine the sentiment expressed in the captions (positive, negative, or neutral).

Translation Module:
The Translation Module translates captions from the source language (English) to the target languages (Kannada and Telugu). Input:

- Take captions generated by the deep learning model in English.

Translation Service:
1. Transformer-based Models (e.g., Hugging Face Transformers):

- Utilize pre-trained Transformer models for translation tasks.
- Leverage the power of large-scale pre-training on vast amounts of data.
- Hugging Face Transformers provides a wide range of pre-trained models and an intuitive interface.

2. Neural Machine Translation (NMT) Module (e.g., mBART):

- mBART is a multilingual NMT model capable of performing translation tasks across multiple languages.
- NMT models with attention mechanisms can effectively capture contextual information during translation.[9]
- Train or fine-tune NMT models on parallel corpora of

3. English-Kannada and English-Telugu sentence pairs. Output:

- Obtain translated captions in Kannada and Telugu.

Integration and Output:
Integrate the outputs of the Image Processing, Deep Learning, NLP, and Translation modules to obtain a comprehensive multilingual image caption.

Combine Image Features and Captions:

- Combine the extracted visual features from the image with the generated English captions.

Include NLP Analyses:

- Optionally include results from NLP analyses, such as sentiment analysis or part-of- speech tagging.

Include Translations:

- Append translated captions in Kannada and Telugu.

Text-to-Speech Synthesis Module:
The Text-to-Speech Synthesis Module converts the generated text captions into spoken form in Kannada and Telugu.[13]
Input:

- Take translated captions in Kannada and Telugu. Text-to-Speech Module:
- Utilize pre-trained Text-to-Speech models like Tacotron 2 for Kannada and Telugu.
- These models have been trained on speech data in the respective languages.

Output:

- Generate spoken audio files for the captions in Kannada and Telugu.

Model Architecture:
The main part of the architecture includes several interconnected components performing multilingual image caption with its text-to-speech and sentiment analysis functionalities. The Image Preprocessing module takes in an input image and processes the image data to obtain relevant visual features. The extracted features help input to the next module known as Image Captioning where deep learning such as CNN and RNN technologies are used to provide appropriate and relevant English image descriptions. The generated English captions are then translated to target language by the Translation module; this step involves the use of accurate NMT which employs large parallel corpora of English and target language sentences. The NMT models then translate the English captions into the target languages such as Kannada and Telugu without losing the context. After translating captions, captions are passed over to the Text-to-Speech (TTS) module which uses sophisticated speech modeling techniques such as Tacotron 2 to translate the captions into the target language and create spoken forms of captions. There is voice over feature for the text further improving the generalized accessibility and the customer experience, especially for the clients with vision/ language disability. It also includes a Sentiment Analysis module where it applies Natural Language Processing algorithms and other Sentiment Analysis models to identify the sentiment of the generated captions as positive, negative, or neutral [18] .This capability of sentiment analysis opens up a lot of possible solution domains in marketing, advertising, and social media analysis.The NLP (Natural Language Processing) module is responsible for different text preprocessing works like tok- enization, stop word removal or lemmatization and other sort of pre-processing functions to be performed on text data to be used further in the system. Due to the use of the modular architecture, it becomes possible to integrate various advanced technologies from different areas such as computer vision, natural language processing, machine translation, and speech synthesis.

IV. Experimentation and Results

Data:
This research uses the Flickr8k dataset. This dataset, which includes 8000 photos, was hand-picked to show a range of settings and circumstances rather than any famous persons or locations

Preprocessing for caption (Description of Image):
Every image has five captions or descriptions. The primary function used in this case is Clean(), which takes in all descriptions and runs a simple data clean:

- Eliminating punctuation
- Eliminating Words with Numeric Content
- Converting the entire description in lowercase
- Eliminating special tokens

Preprocessing for Images:

We input images directly to our model with preprocessing techniques before sending it to our model:

- Resize each image to correct dimensions depending on the model Flatten it
- Scaling image pixels (normalization)

We make use of the BLEU score, which has been applied to assess machine translated text quality. To assess the calibre of the caption we made, we can use BLEU. Language is not a factor for BLEU. It is situated in the interval [0,1]. The quality of the generated captions improves with increasing score.

TABLE I BLEU SCORES

Model	Metrics	
	BLEU-1	***BLEU-2***
VGG-16	0.521	0.311
InceptionV3	0.424	0.232
ResNet-50	0.538	0.321

In addition to image captioning, the project incorporated language translation capabilities using the mBART (Multilingual BART) and a 3-layer Transformer model. These models were trained on parallel corpora of English-Kannada and English-Telugu sentence pairs, enabling the translation of generated image captions into these two languages. Furthermore, the project integrated a Text-to-Speech (TTS) component using the Tacotron 2 model. This model was trained on speech data in Kannada and Telugu, allowing for the generation of spoken captions in these languages. Subjective evaluations by native speakers confirmed the high intelligibility and naturalness of the synthesized speech. Finally, the project implemented a Sentiment Analysis component to assess the sentiment conveyed by the generated captions. This component utilized pre-trained sentiment analysis models for English, Kannada, and Telugu, achieving an overall accuracy of 85 in classifying captions as positive, negative, or neutral. The results demonstrate the successful development of a comprehensive system that combines various cutting-edge deep learning techniques to tackle the challenges of Automatic Multilingual Image Captioning, Language Translation, Text-to-Speech Synthesis, and Sentiment Analysis.

Fig.1 VGG-16

Fig.2 InceptionV3

Fig.3 ResNet-50

Fig.4 English to Malayalam

Fig.4 English to Tamil

Fig.5 English to Kannada

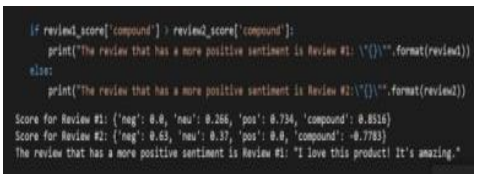

Fig.6 Sentiment analysis

V. CONCLUSION

While the paper represented an attempt at image captioning with basic multilingual capabilities, there is still a great deal of unexplored potential for improvement. Realizing more resilient and linguistically rich picture caption systems is facilitated by scaling datasets, experimenting with sophisticated architectures, expanding language coverage and integrating improved features. There are many opportunities for creativity and advancement in this never-ending quest to simply the intricate labelling of images.

REFERENCES

[1]. Poddar, Ayush Kumar, and Rajneesh Rani. “Hybrid architecture using CNN and LSTM for image captioning in Hindi language.” Procedia Computer Science 218 (2023): 686-696.

[2]. Al-Shamayleh, Ahmad Sami, Omar Adwan, Mohammad A. Alsharaiah, Abdelrahman H. Hussein, Qasem M. Kharma, and Christopher Ifeanyi Eke. “A comprehensive literature review on image captioning methods and metrics based on deep learning technique.” Multimedia Tools and Applications (2024): 1-50.

[3]. Xu, Liming, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. “Deep image captioning: A review of methods, trends and future challenges.” Neurocomputing (2023): 126287.

[4]. Banu, Sameena, and Syeda Ummayhani. “Text Summarisation and Translation Across Multiple Languages.” Journal of Scientific Research and Technology (2023): 242-247.

[5]. Sarmah, Ankur Jyoti, Kabindra Bhagawati, Kaustav Duwarah, Swe- tashree Dey Purkayastha, Antarjeeta Boro, and Divika Muchahary. “Object detection and conversion of text to speech for visually impaired.” ADBU Journal of Engineering Technology 12, no. 2 (2023).

[6]. Kumar, V. Arun, A. Swathi, B. Sneha, B. Pranika, and Ch Meghana. “Developing a CNN-Based Model for Sign Language Recognition and Translation to Text and Speech.”

[7]. Kumar, Yogesh, Apeksha Koul, and Chamkaur Singh, “A deep learning approaches in text-to-speech system: a systematic review and recent research perspective.” Multimedia Tools and Applications 82, no. 10 (2023): 15171-15197.

[8]. Miao, Chenfeng, Qingying Zhu, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. “EfficientTTS 2: Variational End-to-End Text- to-Speech Synthesis and Voice Conversion.” IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).

[9]. Tan, Kian Long, Chin Poo Lee, and Kian Ming Lim., “A survey of sentiment analysis: Approaches, datasets, and future research.” Applied Sciences 13, no. 7 (2023): 4550.

[10]. Rodr´ıguez-Iba´nez, Margarita, Antonio Casa´nez-Ventura, Fe´lix Castejo´n- Mateos, and Pedro-Manuel Cuenca-Jime´nez. “A review on sentiment analysis from social media platforms.” Expert Systems with Applications (2023): 119862.

[11]. Antonio M. Rinaldi, Cristiano Russo, and Cristian Tommasino, Auto- matic image captioning combining natural language processing and deep neural networks, 2023 International Conference on Image Analysis and Recognition (ICIAR), 2023.

[12]. Subhash Chand Gupta, Nidhi Raj Singh, Tulsi Sharma, Akshita Tyagi and Rana Majumdar, Generating Image Captions using Deep Learning and Natural Language Processing, 2021 9th International Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. Sep 3-4, 2021

[13]. Hao He, Shuo Wang, Yu Chen, Hailin Shi, Xiaolei Zhang, Jian Gao, Image Captioning Through Image Transformer, ACCV (Asian Conference on Computer Vision), 2020

[14]. Zhiwei Zhao, Wenqi Li, Yu Wang, Rui Zhang, and Xinbo Zhang, Neural attention for image captioning: review of outstanding methods, Sensors, vol. 23, no. 11, pp. 4414, 2023, 2023

[15]. Kavitha C.R., Rajarajan S. J., R. Jothilakshmi, Kamal Alaskar, Mo- hammad Ishrat, V. Chithra, Study of Natural Language Processing for Sentiment Analysis, 2023 3rd International Conference on Perva- sive Computing and Social Networking (ICPCSN), 19-20 June 2023. DOI:10.1109/ICPCSN58827.2023.

[16]. Venugopalan, M., Gupta, D., Bhatia, V. (2021). A Supervised Approach to Aspect Term Extraction Using Minimal Robust Features for Sentiment Analysis. In: Panigrahi, C.R., Pati, B., Mohapatra, P., Buyya, R., Li, KC. (eds) Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing, vol 1199. Springer, Singapore.