Image Captioning using EfficientNetV2 based on Encoder-Decoder Framework

Dr. Rahul Gaikwad  
*Computer Engineering*   
 Amrutvahini College Of EngineeringSangamner, India  
rahul.gaikwad2k13@gmail.com

Ms. Akanksha Gaikwad  
*Computer Engineering*  
Amrutvahini College Of EngineeringSangamner, India  
akankshabgaikwad4@gmail.com Mr. Mayur Gadakh  
*Computer Engineering*Amrutvahini College Of EngineeringSangamner, India  
mayurgadakh12@gmail.com

Ms. Shivanjali Dhage  
*Computer Engineering*  
Amrutvahini College Of EngineeringSangamner, India  
shivanjalidhage123@gmail.com Mr. Gaurav Chaudhari  
*Computer Engineering*  
Amrutvahini College Of EngineeringSangamner, India  
gauravchaudhariga223@gmail.com

***Abstract-* In this work, a deep neural network-based framework consisting of a "Gated** **‎ Recurrent** **‎ Unit (GRU)" decoder** **‎ and an “EfficientNetV2B0-based** **‎ Convolutional** **‎ Neural** **‎ Network (CNN)” encoder** **‎ is used to** **‎ offer a unique method of** **‎ automatic picture captioning. The framework is designed** **‎ to perceive information** **‎ points within images and their** **‎ contextual relationships, facilitating the generation of meaningful and contextually relevant captions. The CNN** **‎ encoder built on the EfficientNetV2B0** **‎ architecture is very good at identifying objects in pictures and extracting features while preserving spatial** **‎ information. Next, a language describing** **‎ the** **‎ visual information** **‎ collected in** **‎ the photographs is created using these qualities. To improve the** **‎ captioning** **‎ process, the GRU** **‎ decoder is essential in word prediction and sentence** **‎ construction using the retrieved characteristics. The suggested** **‎ neural network system combines the GRU model with the** **‎ effectiveness and precision of the EfficientNetV2B0 model as** **‎ an image feature extractor to provide fixed-length output vectors for ultimate predictions.** **‎ Popular open-source datasets like** **‎ Flickr-8k and Flickr-30k are used in the study to train and evaluate** **‎ the model. Using Python-Keras and TensorFlow backend, the** **‎ framework is implemented, demonstrating the effectiveness of** **‎ the GRU-based model and** **‎ EfficientNetV2B0 in automatic** **‎ picture captioning tasks.** **‎ The suggested method for producing** **‎ correct and contextually appropriate picture captions is shown** **‎ to be successful and accurate when performance evaluation is** **‎ carried out using** **‎the BLEU‎ (BiLingual‎ Evaluation‎ Understudy)‎ measure.**

*Keywords- Image* *‎ Captioning, Convolutional* *‎Neural Networks (CNN), EfficientNetV2, Gated* *‎ Recurrent* *‎ Unit (GRU), Recurrent* *‎ Neural* *‎ Network (RNN).*

1. INTRODUCTION

Captioning an image involves providing a brief description of the image. When composing a caption for an image, it is necessary to first recognize the main components or elements, together with their attributes and connections, and then offer a fitting explanation. The employment of computer vision techniques in conjunction with a language‎ pattern built from Natural-Language-Processing‎ (NLP) is necessary to‎ verify picture words convey images of objects and their interactions.

The study is about, the ability to identify objects, actions, and relationships is necessary for image understanding. These methods mostly rely on‎ the‎ encoding and decoding framework, which is separated into 2 essential stages. Firstly,

CNN model is trained to make it as image encoder. 2nd , the input decoder which creates captions (picture description) is the hidden layer RNN model ImageNet‎ is‎ used‎ to‎ add‎ model‎ weights‎ after‎ CNN‎ is‎ used‎ to‎ extract‎ features‎ from‎ the‎ picture‎ using‎ the‎ EfficientNetV2 model.‎ Nonetheless,‎ the‎ bulk‎ of‎ contemporary‎ encoder-decoder‎ frameworks‎ encode‎ the‎ input‎ picture‎ using‎ a‎ CNN,‎ transform‎ it‎ into‎ a‎ dense‎ feature‎ vector,‎and‎ then‎ employ‎ an‎ architecture‎ called‎ a‎ "Gated-Recurrent-Unit‎ (GRU)"‎ to‎ translate‎ the‎ vector‎ into‎ an‎ illustrative‎ language.

In short, RNN performs effectively with any type of sequential knowledge, including building a group of words, while CNN is better at storing spatial data including recognize things in photos. Ultimately, performance comparisons against the most advanced approaches are shown, including a BLEU score for image captioning systems on datasets with multiple related datasets, such as Flickr30k and Flickr8k.The standard datasets from Flickr30k and Flickr8k, together with a few local datasets, have been used to study the suggested image caption generator.

1. LITERATURE SURVEY

Creating captions for characterization, regression, and prediction issues using natural language, the attention module, and CNN [1]. In order to integrate high-level semantic information into image captioning, this paper suggests a dynamic semantic attention technique. It enhances caption accuracy and richness by separating visual and non-visual word production, as demonstrated by encouraging trial findings [2]. “Natural language processing” and “machine learning” have the ability to automatically identify an image's content [3]. In order to improve the encoder-decoder model by incorporating geometric attention and capturing spatial relationships between identified objects, this study develops the Object Relation Transformer for picture captioning [4].

Provide a trainable, deterministically or stochastically trained attention-based model for visual description. The model gains the ability to create descriptive terms and concentrate on noticeable items [5]. Images are captioned using GRU [6]. Using EfficientNetB0 along with GRU, local tourist image captioning was built [7]. Three main issues plague the more recent approaches to traffic scene: vehicle detection, TSR recognizes traffic signs and detects pedestrians.VGG16 along with LSTM are utilized for sequential caption generation in order to comprehend traffic scenes with different vehicle types (autonomous automobiles)[8].

Image description methods are top-down and bottom-up [9] by producing a suitable natural-language explanation of the visual subject, deep neural network algorithms can effectively address the challenges associated with image captioning [10]. An encoder-decoder model utilizing Wavelet-based CNN for visual feature extraction and attention processes is presented. By combining contextual data, channel, and spatial attention, this approach creates captions [11]. The favored method makes use of deep-learning techniques, like Recurrent-Neural-Networks (RNNs) for caption creation and Convolutional-Neural-Networks (CNNs) for extracting features from images [12].

1. Methodology

The proposed encoder-decoder framework‎ is dependent on following steps:

1. *Data Acquisition and Exploration:*

Conduct the exploratory data analysis (EDA) which gains insights into the Flickr8k and Flickr30k datasets used in the project.

Validate data completeness and accuracy through data comprehension techniques, ensuring reliable inputs for the image captioning model.

1. *Data Preprocessing:*

Implement data preprocessing steps, including image resizing, normalization, and tokenization of captions, to prepare the data for model training.

1. *Model Architecture Design:*

Create a deep learning architecture for picture captioning that is built on the Encoder-Decoder framework and uses EfficientNetV2B0 and GRU layers. To maximize performance, try out several model iterations, adjust hyperparameters, and try out alternative architectures.

1. *Training and Evaluation:*

Train the deep learning model on the pre-processed data, utilizing technique like batch training to enhance convergence.

Evaluate the model's performance using BLEU metrics to assess caption quality and overall effectiveness.

Conduct‎ a‎ comparison‎ analysis‎ with‎ existing‎ techniques‎ to‎ demonstrate‎ the‎ advantages‎ of‎ our‎ proposed‎ model‎ in‎ generating‎ accurate‎ and‎ relevant‎ image‎ captions.

1. *Experimentation and results analysis:*

Conduct experiments to examine the impact of varying parameters, such as batch size, learning rate, and model complexity, on caption generation quality.

Demonstrate‎ the‎ model's‎ capacity‎ to‎ provide‎ a‎ variety‎ of‎ different‎ and‎ semantically‎ relevant‎ captions‎ for‎ a‎ range‎ of‎ images‎ by‎ presenting‎ specific‎ findings‎ and‎ performance‎ data.

1. *Data preprocessing and datasets:*

The accompanying subsections provide instructions on how to prepare associated photo and text data for the deep learning models after obtaining the relevant images and captions from datasets:

1. *Related Image/photos and Caption Datasets:* By using popular open-source datasets, including Flickr-8k, containing 8000 photographs, and Flickr-30k, containing 30000 images.
2. *Photographic data preparation:* The image quality is viewed using a pre-trained model. The goal of the study was to find out how the depth of Convolutional networks influences how accurate they are in large-scale picture recognition. To‎ minimise‎ the‎ amount‎ of‎ parameters‎ in‎ the‎ provided‎ network‎ models,‎ all‎ collected‎ pictures‎ are‎ reduced‎ to‎ 224‎ x‎ 224‎ pixels.‎ A‎ pre-trained‎ model‎ called‎ EfficientNetV2‎ is‎ then‎ used‎ to‎ extract‎ image‎ properties. Lastly, the suggested caption generating method is applied to a particular image of the dataset by means of these attributes.
3. *Text Data Preparation:* Each image in the both datasets have five captions, which results in a long vocabulary. Together with changing words to lowercase letters, the procedure also eliminates newlines and punctuation from the descriptions (captions).

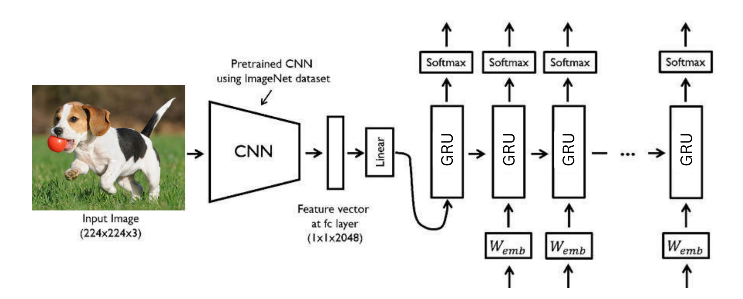


Fig.1. Merging GRU layers with CNN models for image captioning.[13]

1. *The Proposed methodology:*

The‎‎‎ CNN‎‎‎ encoder‎‎‎ and‎‎‎ the‎‎ GRU‎‎ decoder‎‎ are‎‎ the‎‎ two‎‎ main‎‎ deep‎‎ learning‎‎ models‎‎ included‎‎ in‎‎ the‎‎ suggested‎‎ framework.‎‎ Both‎‎ textual‎‎ data‎‎ and‎‎ sequences‎‎ are‎‎ covered‎‎ under‎‎ “Natural‎‎ Language‎‎ Processing‎‎ (NLP)”.

* CNN Encoder: Finally,‎ because‎ this‎ model‎ projects‎ the‎ picture's‎ attributes‎ and‎ advances‎ them‎ to‎ the‎ next‎ level,‎ it‎ discovers‎ that‎ extracting‎ features‎ from‎ an‎ image‎ is‎ simpler. The‎‎ ‎‎ CNN‎‎ ‎‎ architecture,‎‎ ‎‎ such‎‎ ‎‎ as‎‎ ‎‎ EfficientNetV2B0,‎‎ ‎‎ efficiently‎‎ ‎‎ captures‎‎ ‎‎ spatial‎‎ ‎‎ details‎‎ ‎‎ and‎‎ ‎‎ identifies‎‎ ‎‎ relevant‎‎ ‎‎ image‎‎ ‎‎ features,‎‎ ‎‎ facilitating‎‎ ‎‎ a‎‎ ‎‎ rich‎‎ ‎‎ representation‎‎ ‎‎ of‎‎ ‎‎ the‎‎ ‎‎ visual‎‎ ‎‎ content.
* GRU Decoder: A‎‎ Gated‎‎ Recurrent‎‎ Unit‎‎ (GRU)‎‎ decoder‎‎ must‎‎ first‎‎ scan‎‎ the‎‎ encoded‎‎ image‎‎ before‎‎ producing‎‎ a‎‎ sequential‎‎ caption. The‎‎ next‎‎ step‎‎ is‎‎ to‎‎ generate‎‎ a‎‎ text‎‎ description,‎‎ sometimes‎‎ called‎‎ the‎‎ "Language‎‎ Model"‎‎ via‎‎ word‎‎ sequence‎‎ formation.

1. *Model Creation:*

The‎ three‎ main‎ parts‎ of‎ the‎ suggested‎ deep‎ learning-based‎ merging‎ model‎ are‎ an‎ image‎ feature‎ extraction,‎ ‎ sequence‎ processor,‎ and‎ a‎ decoder,‎ as‎ seen‎ in‎ Figure 1.

* *Image feature extractor:*The‎ process‎ of‎ extracting‎ features‎ from‎ photographs‎ is‎ known‎ as‎ image‎ feature‎ extraction.‎ With‎ the‎ exception‎ of‎ the‎ output‎ layer,‎ a‎ pre-trained‎ 237-layer‎ EfficientNetB2V0‎ model‎ is‎ used‎ to‎ extract‎ the‎ content‎ from‎ the‎ photos.‎ After‎ pre-processing Ultimately,‎ the‎ model‎ discovers‎ that‎ it‎ is‎ easier‎ to‎ extract‎ features‎ from‎ a‎ picture‎ because‎ it‎ projects‎ the‎ image's‎ attributes‎ and‎ advances‎ them‎ to‎ the‎ next‎ level.
* *Sequences processor:* Text‎ data‎ is‎ handled‎ by‎ a‎ word‎ embedding‎ layer,‎ which‎ is‎ followed‎ by‎ a‎ GRU-based‎ recurrent‎ layer.‎ Sequence‎ modelling‎ challenges‎ are‎ well-known‎ for‎ the‎ competitive‎ performance‎ and‎ computational‎ efficiency‎ of‎ GRU‎ units.‎ When‎ combined,‎ these‎ elements‎ allow‎ for‎ the‎ accurate‎ creation‎ of‎ captions‎ by‎ capturing‎ the‎ textual‎ data's‎ sequential‎ relationships.
* *Decoder:* The decoder takes charge of producing the ultimate textual descriptions, while a “Dense layer” integrates and processes the fixed-length vectors from the sequence processor and captions feature extractor.

Our‎ objective‎ was‎ to‎ establish‎ an‎ expansive‎ lexicon‎ while‎ ensuring‎ brevity,‎ resulting‎ in‎ a‎ more‎ streamlined‎ model‎ conducive‎ to‎ quicker‎ training.‎ Each‎ description‎ undergoes‎ segmentation‎ into‎ individual‎ words.‎ The‎ model‎ proceeds‎ to‎ analyze‎ each‎ word‎ in‎ conjunction‎ with‎ its‎ corresponding‎ photo,‎ predicting‎ subsequent‎ words‎ accordingly.‎ The next word in the series is then predicted by the model using the first two words in the text description together with the corresponding image. This deep learning architecture is intended to function as a "merge-model," where a thick layer creates a condensed representation of the picture.‎ In‎ addition,‎ the‎ sequence‎ processor‎ part‎ accepts‎ input‎ sequences‎ with‎ predetermined‎ lengths.‎ These‎ sequences‎ are‎ then‎ sent‎ via‎ an‎ embedded‎ layer,‎ that‎ employs‎ a‎ masking‎ method‎ to‎ ignore‎ padded‎ data.‎ Dropout‎ regularization‎ is‎ used‎ to‎ reinforce‎ both‎ input‎ models‎ in‎ order‎ to‎ prevent‎ overfitting‎ to‎ the‎ data‎ set‎ for‎ training‎ and‎ take‎ advantage‎ of‎ the‎ model's‎ quick‎ learning‎ speed.

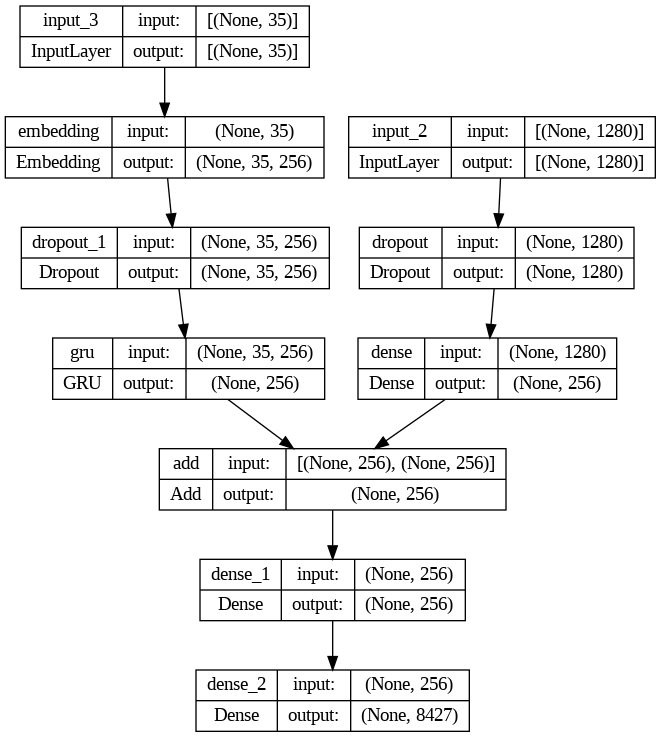


Fig.2. Model Summary

1. *Training Procedure:*

The‎ training‎ process‎ starts‎ by‎ taking‎ useful‎ information‎ from‎ images‎ using‎ a‎ model‎ that‎ has‎ already‎ been‎ trained,‎ and‎ then‎ proceeds‎ to‎ process‎ text‎ data.‎ This‎ process‎ includes‎ changing‎ captions‎ to‎ lowercase,‎ eliminating‎ non-alphabetic‎ characters,‎ and‎ inserting‎ start‎ and‎ end‎ tokens. Afterwards,‎ the‎ text‎ undergoes‎ tokenization‎ using‎ the‎ Tokenizer‎ class,‎ and‎ the‎ vocabulary‎ size‎ is‎ determined‎ by‎ the‎ number‎ of‎ distinct‎ words.‎ within‎ the‎ dataset‎ known‎ as‎ corpus.‎ The‎ data‎ is‎ split‎ into‎ test‎ and‎ training‎ sets,‎ with‎ 80%‎ of‎ the‎ data‎ designated‎ for‎ training,‎ after‎ the‎ maximum‎ caption‎ length‎ for‎ padding‎ is‎ established.‎ During‎ the‎ training‎ of‎ the‎ model,‎ a‎ data‎ generator‎ function‎ is‎ used‎ to‎ produce‎ batches‎ of‎ training‎ data.‎ This‎ function‎ pre-processes‎ text‎ and‎ generates‎ pairs‎ of‎ input‎ and‎ output‎ for‎ the‎ training.‎ The‎ model‎ architecture‎ is‎ built‎ using‎ TensorFlow‎ Keras‎ and‎ consists‎ of‎ an‎ encoder-decoder‎ structure.‎ It‎ includes‎ dropout‎ regularization‎ and‎ dense‎ layers.‎ The‎ image‎ features‎ are‎ handled‎ by‎ the‎ encoder,‎ while‎ the‎ decoder‎ utilizes‎ GRU‎ layers‎ to‎ incorporate‎ sequence‎ features.‎ This‎ particular‎ model‎ consists‎ of‎ distinct‎ input‎ layers‎ for‎ image‎ attributes‎ and‎ encoded‎ text.‎ These‎ layers‎ are‎ combined‎ by‎ utilizing‎ the‎ add‎ function‎ to‎ create‎ the‎ network‎ model.‎ A‎ predetermined‎ number‎ of‎ epochs‎ are‎ processed‎ to‎ train‎ the‎ model,‎ during‎ which‎ the‎ Adam‎ optimization‎ algorithm‎ and‎ categorized‎ cross-entropy‎ loss‎ are‎ used‎ to‎ construct‎ the‎ model.‎ Every‎ period‎ is‎ made‎ up‎ of‎ several‎ stages,‎ where‎ the‎ data‎ generator‎ supplies‎ batches‎ of‎ training‎ data‎ to‎ be‎ used‎ for‎ model‎ adjustment.‎ The‎ model‎ learns‎ to‎ create‎ descriptions‎ for‎ images‎ using‎ the‎ given‎ training‎ data‎ by‎ going‎ through‎ this‎ repeated‎ process.

1. *Computational resources used & hyperparameters:*

We made use of Colab Pro's powerful NVIDIA Tesla V100 GPU with 32GB VRAM, a subscription service that offers enhanced GPU capabilities and longer runtime. Our model training process was greatly expedited by this configuration, with the training and testing stages being finished in around three hours. To provide a stable development environment, we used TensorFlow, Keras, NumPy, pandas, tqdm, matplotlib, NLTK, and pickle in our software stack. We used Colab Pro's cloud data storage management feature and batch processing with a 64-batch size to achieve effective training iterations. The performance of our model was further enhanced by adjusting hyperparameters such as the optimizer (Adam), embedding dimensions (256), and dropout rate (0.4). Once we reached convergence after 11 epochs of training, we saved the trained model as 'best\_model.h5' for further analysis and assessment.

1. Experimental results
2. *Datasets:*

The Flickr8k, Flickr30k shared datasets were the two common datasets utilized in the research. These datasets were chosen because of their reasonable size-small enough to be generated on a desktop computer with a single CPU-realism, and open-source nature. The dataset contains eighty different categories of items depicted in the photos. Each image is associated with five captions stored in token.txt. Additionally, the proposed model underwent testing on a local dataset. Prior to processing, Every image is cropped to 224 by 224 pixels. Additionally, the suggested model was integrated into datasets used for testing & training.

1. *Outcomes:*

The proposed approach focuses on providing image captions that describe the pictures. A few anticipated results of deep learning-driven image description generator are shown in the figure, which depicts the user interfaces of the proposed system. It was observed that images featuring people or other human subjects achieved the highest accuracy, as most training images are individual photos. Local images sourced from local camera shots, university websites, and other platforms were also utilized to evaluate the suggested model, which yielded satisfactory results in captioning the images.



Fig.3. Actual and predicted captions for image from the dataset

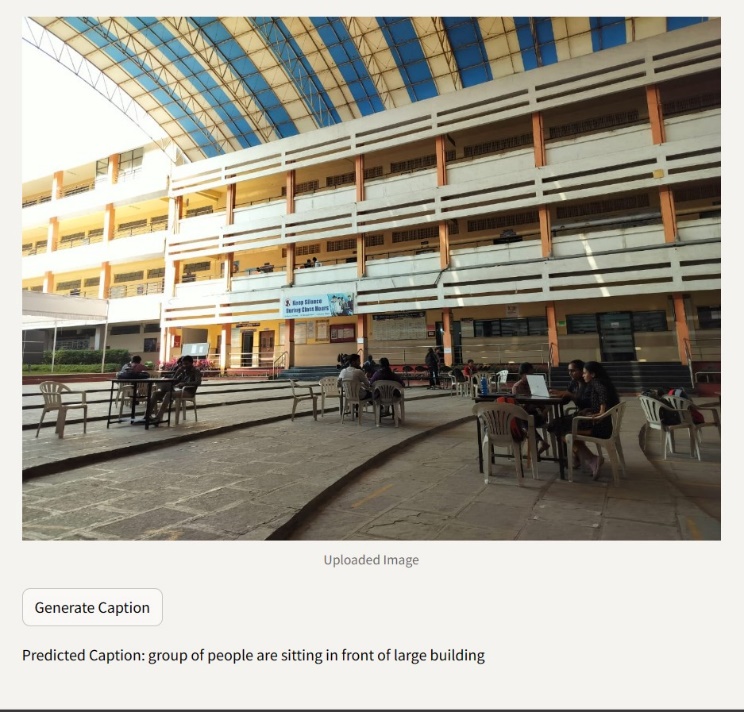


Fig.4. Predicted caption for custom image

1. *Performance Evaluation:*

|  |  |  |
| --- | --- | --- |
| BLEU Metric | Flickr8k | Flickr30k |
| BLEU-1 | 0.625841 | 0.598135 |

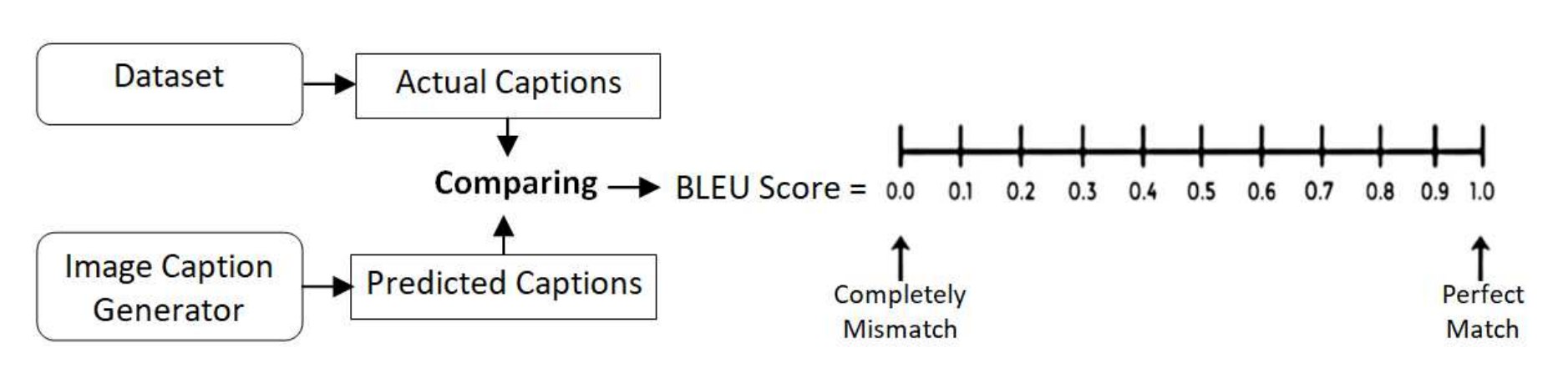


Fig.5. The process of generating BLEU score [13]

The‎ test‎ set‎ is‎ used‎ to‎ predict‎ picture‎ captions,‎ and‎ these‎ predictions‎ are‎ then‎ evaluated‎ using‎ an‎ established‎ metric‎ to‎ evaluate‎ the‎ proposed‎ model.‎ BLEU‎ ('Bilingual-Evaluation-Understudy'),‎ a‎ bilingual‎ assessment‎ metric,‎ was‎ used‎ to‎ gauge‎ how‎ effective‎ the‎ proposed‎ photo‎ captioning‎ method‎ was.

To calculate BLEU score, a predicted sentence is compared to a reference sentence. The corresponding BLEU score perfection and examples of relevant captions are displayed. The Python NLTK module was used to produce the BLEU score for the evaluation of the candidate text. The quality of machine-generated versions is evaluated using the BLEU (Bilingual-Evaluation-Understudy) score. Phrase BLEU Score for each sentence and Corpus BLEU Rating for groups of sentences are the two levels at which it works. Comparing‎ comparable‎ grammes‎ in‎ a‎ preset‎ order—for‎ example,‎ one‎ gramme‎ for‎ single‎ words‎ and‎ two‎ grammes‎ for‎ word‎ pairs—determines‎ the‎ N-gram‎ scores. Every N-gram match receives a weight, usually 0 for non-matches and one for matches. Balanced geometric averages of N-gram scores over a range of orders—from 1 to n—are computed to determine the BLEU score. The‎ Collective‎ N-gram‎ values‎ (BLEU-N),‎ which‎ are‎ produced‎ by‎ calculating‎ the‎ weighted‎ geometrical‎ average‎ of‎ the‎ individual‎ N-gram‎ scores,‎ are‎ a‎ crucial‎ factor‎ in‎ determining‎ the‎ overall‎ BLEU‎ score.

1. CONCLUSION

Constructing a system for encoding and decoding for deep neural networks for a ‎ practical photo captioning system was the aim of this effort. After processing and passing through the RNN layers for all ‎ relevant keywords, text strings, and captions, The CNN algorithm layer aids in extracting features from the images. Adding anticipated words to the feedforwarding model is the last step. This produces a final word description that is based ‎ on the input image's features (derived by CNN) and the data source's words (words from a dataset by RNN). Suggested deep-learning model (BLEU metrices) for picture caption ‎ creation was assessed in the experiments with two standard datasets (Flickr-8k and Flickr-30k) and images from various local sources. The image captioning test produced satisfactory results and showed that photographs with people or other ‎ humans in them are the most accurate. The suggested system, however, has significant computational expenses and needs powerful GPU hardware in order to operate. This ‎ prevented us from training the entire dataset, resulting in a vocabulary reduction that is insufficient for precise item detection.

1. FUTURE SCOPE

* *Attention mechanisms:* Attention mechanisms will be leveraged in future image captioning developments to allow models to concentrate on important picture aspects for improved relevance and accuracy in captions.
* *Teacher forcing:* Implementation of teacher forcing techniques to improve training stability and convergence, ensuring more accurate and coherent caption generation by leveraging ground truth information during model training.
* *Real-time and optimization:* Innovations in real-time captioning will make it easier to generate captions instantly. This will be especially helpful for live events and video streams that need precise captions right away.
* *Fine-tuning:* To increase model performance, fine-tuning approaches will be used to incorporate transfer learning, improve hyperparameters, and guarantee flexibility across a variety of datasets.
* *Scalability:* Improvements in scalability will guarantee effective handling of big datasets and instantaneous picture analysis, rendering the model adaptable and useful in a range of industries, including media, healthcare, and industrial automation.

All things considered, these developments hold up the possibility of more precise, contextually appropriate, and adaptable picture captions, increasing the image captioning technology's usefulness in a variety of fields and applications.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Amrutvahini College of Engineering (@Maharashtra, India) for their invaluable support and assistance throughout the course of this study. Their contributions have played a significant role in the successful completion of this project.

##### REFERENCES

1. Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128-3137. 2015.
2. Liu, Xiaoxiao, and Qingyang Xu. "Adaptive attention-based high-level semantic introduction for image caption." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, no. 4 (2020): 1-22.
3. Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An overview of image caption generation methods." *Computational intelligence and neuroscience* 2020 (2020).
4. Herdade, Simao, Armin Kappeler, Kofi Boakye, and Joao Soares. "Image captioning: Transforming objects into words." *Advances in neural information processing systems* 32 (2019).
5. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, Conference Proceedings, pp. 3156–3164.
6. Khan, Rashid, M. Shujah Islam, Khadija Kanwal, Mansoor Iqbal, Md Imran Hossain, and Zhongfu Ye. "A deep neural framework for image caption generation using gru-based attention mechanism." *arXiv preprint arXiv:2203.01594* (2022).
7. Fudholi, Dhomas Hatta, Yurio Windiatmoko, Nurdi Afrianto, Prastyo Eko Susanto, Magfirah Suyuti, Ahmad Fathan Hidayatullah, and Ridho Rahmadi. "Image captioning with attention for smart local tourism using efficientnet." In *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 012038. IOP Publishing, 2021.
8. Li, Wei, Zhaowei Qu, Haiyu Song, Pengjie Wang, and Bo Xue. "The traffic scene understanding and prediction based on image captioning." *IEEE Access* 9 (2020): 1420-1427.
9. Verma, Akash, Arun Kumar Yadav, Mohit Kumar, and Divakar Yadav. "Automatic image caption generation using deep learning." *Multimedia Tools and Applications* 83, no. 2 (2024): 5309-5325.
10. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, Conference Proceedings, pp. 4651– 4659.
11. Sasibhooshan, Reshmi, Suresh Kumaraswamy, and Santhoshkumar Sasidharan. "Image caption generation using visual attention prediction and contextual spatial relation extraction." *Journal of Big Data* 10, no. 1 (2023): 18.
12. Dongare, Yashwant, Bhalchandra M. Hardas, Rashmita Srinivasan, Vidula Meshram, Mithun G. Aush, and Atul Kulkarni. "Deep Neural Networks for Automated Image Captioning to Improve Accessibility for Visually Impaired Users." *International Journal of Intelligent Systems and Applications in Engineering* 12, no. 2s (2024): 267-281.
13. Rahman, Md Mijanur, Ashik Uzzaman, and Sadia Islam Sami. "Implementing Deep Neural Network Based Encoder-Decoder Framework for Image Captioning." In *2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pp. 26-31. IEEE, 2021.