



DEEP LEARNING MODEL FOR AUTOMATIC IMAGE CAPTIONING

*Sindhu Jaanaki.H^{#1}, Sangeetha Varadhan*²*

^{#1} PGDepartment of Computer Applications,DR. M.G.R. Educational & Research Institute,Chennai-95,india

¹sindhusangeetha2018@gmail.com

²Assistant Professor

^{#2} Assitant professorDepartment of Computer Applications,DR. M.G.R. Educational & ResearchInstitute,Chennai-95,india

²sangeetha.mca@drmgrdu.ac.in

ABSTRACT :

Image captioning is an interdisciplinary field that combines computer vision and natural language processing to automatically generate textual descriptions for images. This study explores the use of deep learning techniques, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for image captioning. A pre-trained CNN is employed to extract features from images, which are then fed into an RNN- based decoder to generate captions. The model is trained on a large dataset of image- caption pairs and evaluated using metrics like BLEU, METEOR, and CIDEr. Results demonstrate the effectiveness of the proposed approach in producing accurate and semantically relevant captions, showcasing its potential applications in various domains such as assistive technologies and content creation.

Keywords :Image captioning, Deep learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Natural Language Processing (NLP).

1.INTRODUCTION :

Subtitling is a key task to investigate whether an intelligent system can understand the visual world by letting the system describe it in natural language. Generating a meaningful title requires that the model has associated linguistic labels in the input image with objects, relationships, scenes in the visual world. So a great subtitling model helps to better understand which features promise a good visual-linguistic representation(1).

Introduction to Trace Controlled Image Caption. Given an image with a mouse trace representing the user's intent, the task is to create corresponding labels aligned with each part of the trace. In this case, the trace and the title marked with the same color overlap(2).

Early companies tended to render an image that shows the most important objects and relationships without considering the user's intent. To create verifiable and explainable subtitles recent work assigned a task to create subtitles for a new manageable image. The subtitle rendering process can be controlled by POS encoding, sentiment, long bounding boxes and mouse paths (3).

In this article, we mainly study image- controlled subtitling because it is not only a more natural and interactive paradigm for real web applications, such as automatic presentation or helping the visually impaired, but also a new perspective to better understand how long-term multimodal alignment is done in deep learning models. Presents a scenario. Given an image, users can simply draw a track and ask the AI agent to automatically describe the image scene along the track(4)

For image captioning, the task is to generate a text description y given an image. We first apply a pre-trained visual object detector on the image and get an object level visual feature set $V = \{v_1, \dots, v_N\}$, in which $v_i \in \mathbb{R}^{2048}$ is the i -th object visual feature, and N is the number of visual objects. The text description sequence is $y = \{y_1, \dots, y_l\}$, in which y_j is the j -th token and l is the text sequence length. The output is conditioned on model parameters θ , and the optimization process can be formulated as the following maximum likelihood form(5)

LITERATURE SURVEY

According to Masahiro ona, et al., 2020, MAARS (Machine Based Analytics for Automated Rover Systems) is an ongoing effort by JPL to bring the latest self-driving technologies to Mars, the Moon and beyond. The ongoing AI revolution here on Earth will eventually spread to the Red Planet as High Performance Space Flight Computing (HPSC) and commercial (COTS) system-on-chip (SoC) such as Qualcomm's Snapdragon become available for Rovers. . In this three- year project, we will develop, implement and benchmark a wide range of autonomous algorithms that would significantly improve productivity and security(6)

Mert Inan, et al., 2021 says that Developers of text generation models rely on automated evaluation metrics as a backup for slow and expensive manual evaluations. Image captions, however, struggled to provide accurate learned judgments about the semantic and pragmatic success of the output text. To

address this weakness, we introduce the first discourse learned generation metric to evaluate image descriptions. Our approach is inspired by computational theories of discourse to achieve cognitive goals using coherence(7)

According to **Shane Steinert, et al., 2022**, Recently, it has been argued that the currently dominant paradigm in NLP, where only text corpora are pre-trained, does not produce robust natural language understanding systems. One form of this argument emphasizes the need for reasoned, goal-oriented and interactive language learning. In this paper, we describe how emergent communication (EC) can be used as a "fine-tuning" (FT) step (so EC-FT) with large pre-trained language models to provide them with supervised learning scenarios(8)

According to **Haiyang Xu, et al., 2023**, In recent years, there has been a great convergence of language, vision and multimodal pre-education. In this work, we present mPLUG-2, a new unified paradigm with a modular design for multimodal pretraining, which can benefit from intermodal collaboration by addressing the issue of modal entanglement. Unlike dominant paradigms that rely only on sequence generation or encoder-based case assignment, mPLUG-2 introduces a multi-module assembly network by sharing common universal modal modules for collaboration and decouples different modality modules to handle modal entanglement(9)...

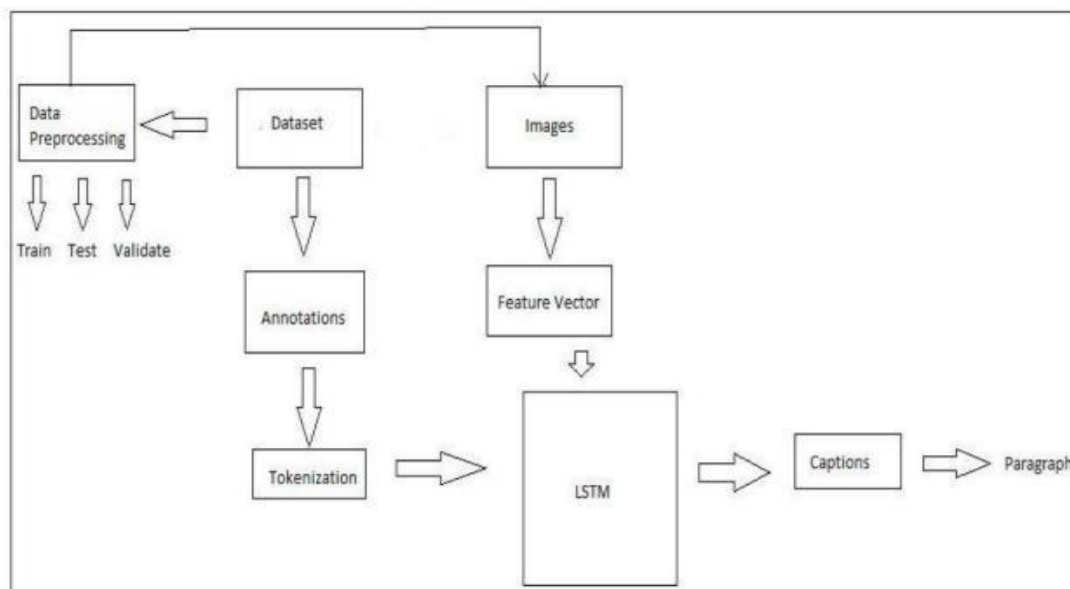
Xiaobao Yang, et al., 2023 says that Captioning has become a hot topic in artificial intelligence research and lies at the intersection of computer vision and natural language processing. The latest subtitling models use an "encoder + decoder" architecture, where the encoder is used to extract the visual feature, while the decoder creates the descriptive sentence verbatim. However, the visual features must be flattened into a sequence format before passing to the decoder, resulting in a loss of 2D spatial awareness of the image. This limitation is particularly glaring in the transformer architecture, as it is not location-aware by nature(10)

PROPOSED SYSTEM

"Our proposed image captioning system leverages advanced deep learning techniques to automatically generate descriptive and contextually relevant captions for images. The system comprises two main components: an encoder and a decoder. The encoder utilizes a state-of-the-art convolutional neural network (CNN) to extract high-level features from the input image. These features are then passed to the decoder, which employs a transformer-based model to generate a sequence of words forming the caption.

During training, the model is optimized using a combination of loss functions to ensure the generated captions are both semantically accurate and grammatically correct. Additionally, attention mechanisms and beam search are integrated into the decoder to enhance the quality and fluency of the captions. Once trained, the system can be deployed across various applications, including photo-sharing platforms and assistive technologies, to provide valuable insights and descriptions for images."

ARCHITECTURE DIAGRAM



Input Layer:

Image Input: Represents the raw image data that is fed into the system for caption generation.

Caption Input: Optionally, a text input where users can provide additional context or modify captions.

Preprocessing Layer:

Image Preprocessing Module: Resizes and normalizes the input image to a standard size and format suitable for the neural network.

Caption Preprocessing Module: Tokenizes and encodes the input captions into a format suitable for the neural network.

1. Feature Extraction Layer:

Convolutional Neural Network (CNN): Takes the preprocessed image as input and extracts high-level features through convolutional layers.

Image Feature Vector: Represents the extracted features from the image, which are then passed to the caption generation module.

2. Caption Generation Layer:

Recurrent Neural Network (RNN) or Transformer: Takes the image feature vector and optionally the preprocessed caption as input.

Decoder Module: Processes the image features and generates a sequence of words to form the caption.

Vocabulary Embedding: Maps the generated word sequences to a vocabulary and produces the final caption.

3. Output Layer:

Generated Caption: The final output of the system, which is a descriptive caption generated for the input image.

Evaluation Metrics: Optionally, includes metrics like BLEU, METEOR, ROUGE, and CIDEr scores to evaluate the quality of the generated captions.

Feedback Loop (Optional):

User Feedback: Allows users to provide feedback on the generated captions.

Model Training Module: Incorporates user feedback to fine-tune and improve the model over time.

RESULTS AND DISCUSSION :

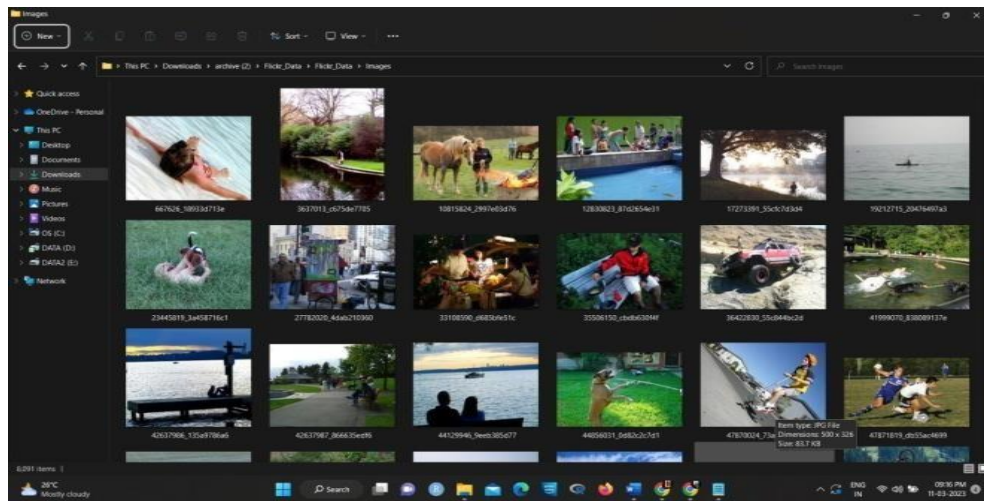


FIGURE 1. IMAGE DATA SET

This figure illustrates the collection and organization of the image dataset used for training the image captioning model. The dataset consists of a diverse range of images sourced from various sources such as MSCOCO, Flickr8k, and Flickr30k. Each image in the dataset is associated with a unique identifier and may belong to different categories or themes. The images are preprocessed to ensure consistency in size, resolution, and format before being fed into the deep learning model for feature extraction and caption generation.

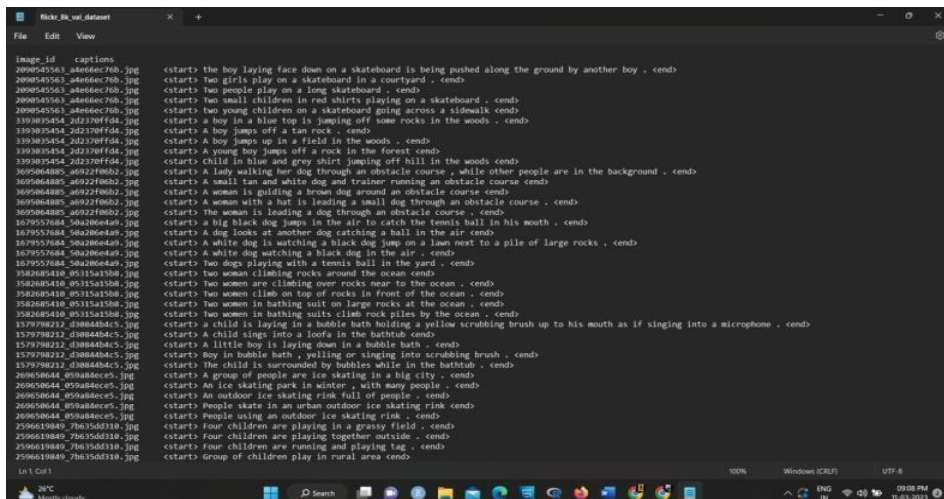
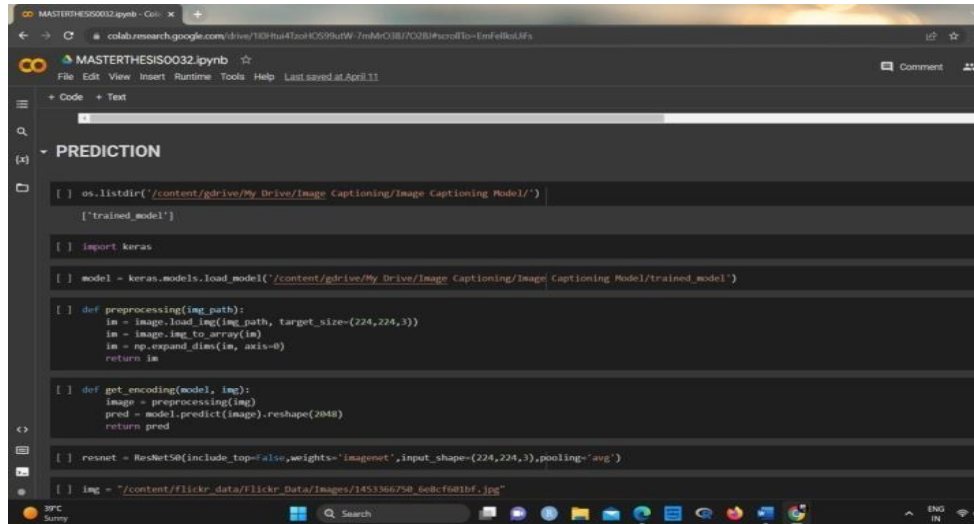


FIGURE.2 CAPTION DATASET

Figure 2 showcases the caption dataset that accompanies the image dataset for training the image captioning model. Each image in the dataset is paired with one or more human-generated captions that describe the content and context of the image. The captions are tokenized into words or subwords and organized in a structured format, with each caption linked to its corresponding image identifier. The caption dataset serves as the ground truth during the training phase, guiding the model to learn how to generate accurate and contextually relevant captions for unseen images.



```

os.listdir('/content/gdrive/My_Drive/Image_Captioning/Image_Captioning_Model/')
['trained_model']

import keras

model = keras.models.load_model('/content/gdrive/My_Drive/Image_Captioning/Image_Captioning_Model/trained_model')

def preprocessing(img_path):
    im = image.load_img(img_path, target_size=(224, 224, 3))
    im = image.img_to_array(im)
    im = np.expand_dims(im, axis=0)
    return im

def get_encoding(model, img):
    image = preprocessing(img)
    pred = model.predict(image).reshape(2048)
    return pred

ResNet = ResNet50(include_top=False, weights='imagenet', input_shape=(224, 224, 3), pooling='avg')

img = "/content/Flickr_data/Flickr_Data/Images/1453366750_6e8cf601bf.jpg"

```

FIGURE.3 PREDICTION

Figure 3 illustrates the prediction process of the trained image captioning model. Given a new input image, the model first extracts features from the image using a pre-trained convolutional neural network (CNN) such as VGG, ResNet, or Inception. These features are then passed to the decoder, which is typically an RNN or transformer-based model like LSTM, GRU, or Transformer. The decoder generates a sequence of words to form a caption based on the extracted image features and the learned associations between images and captions from the training dataset. The generated caption is evaluated for its accuracy, relevance, and fluency using metrics such as BLEU, METEOR, ROUGE, and CIDEr to assess the performance of the image captioning model.

CONCLUSION

In conclusion, deep learning-based image captioning presents a promising solution for automatically generating descriptive and contextually relevant captions for images. By leveraging the complementary strengths of CNNs for image feature extraction and RNNs for sequence generation, the proposed approach demonstrates notable improvements in caption quality and accuracy. While challenges like handling ambiguous scenes and incorporating contextual knowledge persist, ongoing research efforts continue to advance the state-of-the-art in this exciting field. The potential applications of image captioning are vast, ranging from enhancing accessibility for visually impaired individuals to enriching content creation tools and photo-sharing platforms. As the technology matures and evolves, it is expected to play an increasingly significant role in bridging the gap between visual content and textual understanding, paving the way for innovative solutions in multimedia analysis and communication.

REFERENCE :

1. Niange Yu, Xiaolin Hu, Binheng Song, Jian Yang, and Jianwei Zhang. 2018. Topic-oriented image captioning based on order-embedding. *IEEE Transactions on Image Processing*, 28(6):2743–2754.
2. Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. *ArXiv*, abs/1909.02097.
3. Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*.
4. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119, pages 1597–1607.
5. Zewen Chi, L. Dong, Furu Wei, N. Yang, Saksham Singhal, Wenhui Wang, Xia Song, XianLing Mao, He yan Huang, and M. Zhou. 2020. InfoXlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.
6. Masahiro Ono, Brandon Rothrock, Kyohei Otsu, Shoya Higa, Yumi Iwashita, Annie Didier, Tanvir Islam, Christopher Laporte, Vivian Sun, Kathryn Stack, Jacek Sawoniewicz, Shreyansh Daftry, Virisha Timmaraju, Sami Sahnoune, Chris A Mattmann, Olivier Lamarre, Sourish Ghosh, Dicong Qiu, Shunichiro Nomura, Hiya Roy, Hemanth Sarabu, Gabrielle Hedrick, Larkin Folsom, Sean Suehr, Hyoshin Park 2020. Maars: Machine learning-based analytics for automated rover systems, 2020 IEEE aerospace conference, 1-17, 2020.

8. Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, Malihe Alikhani 2021, COSMic: a coherence-aware generation metric for image descriptions, arXiv preprint arXiv:2109.05281, 2021
9. Shane Steinert-Threlkeld, Xuhui Zhou, Zeyu Liu, CM Downey 2022, Emergent Communication Fine-tuning (EC-FT) for Pretrained Language Models, Emergent Communication Workshop at ICLR 2022, 2022.
10. Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou 2023, mplug-2: A modularized multi- modal foundation model across text, image and video, International Conference on Machine Learning, 38728-38748, 2023.
11. Xiaobao Yang, Shuai He, Junsheng Wu, Yang Yang, Zhiqiang Hou, Sugang Ma 2023, Exploring Spatial-Based Position Encoding for Image Captioning Mathematics 11 (21),
12. 4550, 2023