Image Captioning using VGG-16 Deep Learning Model

Vishal Jayaswal
Department of CSE
Ajay Kumar Garg Engineering College
Ghaziabad, India
vishaljayaswal026@gmail.com

Sharma Ji
Department of CSE
Ajay Kumar Garg Engineering College
Ghaziabad, India
saurabh12d@gmail.com

Satyankar
Department of CSE
Ajay Kumar Garg Engineering College
Ghaziabad, India
satyankargoelnbd@gmail.com

Vanshika Singh
Department of CSE
Ajay Kumar Garg Engineering College
Ghaziabad, India
vanshisingh2502@gmail.com

Yuvika Singh
Department of CSE
Ajay Kumar Garg Engineering College
Ghaziabad, India
yuvikasingh17@gmail.com

Vaibhav Tiwari
Department of CSE
Ajay Kumar Garg Engineering College
Ghaziabad, India
vaibhavtiwari0128@gmail.com

***Abstract*—Image captioning is a method that creates captions from images. It makes use of computer vision, natural language processing, and deep learning. The field of image captioning has evolved significantly in recent times, thanks to the application of both conventional and sophisticated deep learning approaches. This study uses two well-known benchmark datasets, Flickr8k and Flickr30k, to examine the use of the VGG-16 convolutional neural network (CNN) for producing descriptive captions. A recurrent neural network (RNN) is integrated to generate coherent and contextually relevant captions after the VGG-16 model is utilized for feature extraction. Human-provided references and generated captions are compared for quality using standard assessment metrics such as BLEU. The findings have applications in the areas of content indexing, assistive technology for the visually impaired, and enhancing user interfaces on image-centric systems. The comparison of the Flickr8k and Flickr30k datasets sheds light on the challenges posed by the different datasets and provides guidance for future image captioning research. A list of references, a synopsis of the key findings, and suggestions for future research topics are included in the paper's conclusion.**

***Index Terms*—Deep learning, Object Hallucination, Natural Language Processing, Computer Vision.**

I. Introduction

The increasing availability of optical perception data has led to its significant utility in various fields such as quality assurance, medical diagnostics, surveillance, autonomous vehicles, facial recognition, forensic investigations, biometrics, and 3D reconstruction. Image captioning, a developing subject at the intersection of natural language processing and computer vision, aims to create text-based descriptions that align linguistically [12], maintain syntactical accuracy, and adhere to semantic relevance. This technology can be used for improving content retrieval systems and helping visually impaired people. This research study explores the complexities of image captioning, providing a mathematical definition of the problem and addressing inherent difficulties such as maintaining variety in image content perception and producing cohesive captions [8]. It examines well-known datasets, preprocessing stages, and training procedures, including loss functions and optimization methods. The study also covers assessment metrics, focusing on the significance of both quantitative and qualitative evaluations. The study contrasts conventional methods with recent developments in deep learning to provide a comprehensive understanding of the field.

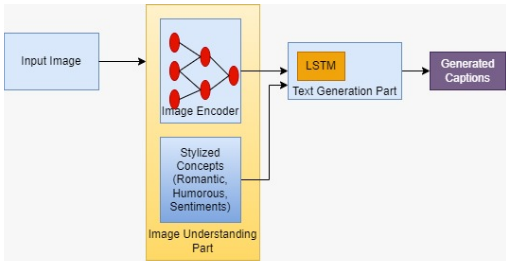


Fig. 1. Block Diagram representing Image Captioning process

In fig 1, we are explaining the process behind the model. Process includes steps i.e. image features are obtained using convolutional neural network, then attributes are extracted from visual features and multiple captions are created by language model and at last captions that are generated.

The techniques used for obtaining image features can be broadly divided into two categories:

(1) Traditional machine-learning based techniques (2) Deep machine-learning based techniques.

This study explores different methods for creating image captions, including template-based, retrieval-based, and creative approaches. Retrieval-based methods draw captions from pre-existing ones, while template-based methods use predetermined templates. Deep machine learning techniques improve semantic accuracy in each image. Learning methodologies like supervised and reinforcement learning can be used. Encoder-decoder or compositional architectures can be used for entire scenes or specific areas. Techniques use attention mechanisms, semantic ideas, and image-description styles.

II. Related Work

This study explores different methods for creating image captions, including template-based, retrieval-based, and creative approaches. Retrieval-based methods draw captions from pre-existing ones, while template-based methods use predetermined templates. Deep machine learning techniques can produce original captions from visual and multimodal [7] domains, improving semantic accuracy. Learning methodologies like supervised learning and reinforcement learning can be used. Encoder-decoder or compositional architectures can be used for entire scenes or specific areas of an image. Techniques use attention mechanisms, semantic ideas, and different image-description styles. [13]

A. Traditional Approaches

Prior to the invention of artificial neural networks (ANNs) [15], traditional segmentation and semantic segmentation techniques were primarily unsupervised. Traditional semantic segmentation techniques require less data and take less time to compute. The methods utilized by early researchers were template-based and retrieval-based to allow computers with visual learning abilities to generate coherent and understandable language for images.

Retrieval-based image captioning stores numerous image-caption pairs in a corpus using similarity comparison to find images similar to the input image. [6] Computers use annotations from previously obtained photos to characterize a given image, using the Tree-F1 rule and the nearest neighbor rule. Global similarity is calculated, and the caption of the matching image is sent.

Template-based image captioning is a model used in early projects, creating captions in a limited way, both syntactically and semantically. It involves using fixed templates with fixed blank spots to recognize the properties, features, actions, and surrounds of objects before filling them in. Some strategies include filling in blanks using a triplet, object identification techniques, and pretrained language models. [18]

Unlike retrieval models, template-based models produce descriptions more relevant to images but employ predefined templates and cannot produce captions of varying length, making the automated descriptions less organic compared to handwritten captions by humans.

B. Neural Network-Based Captioning

Comparing deep learning to other machine learning techniques and algorithms, image captioning has greatly improved. These techniques require a substantial quantity of training data, comprising images and caption labels, and are mostly used in supervised environments. These astounding results have been attained by applying a variety of models, such as autoencoders ,generative adversarial networks (GAN) [3], convolutional neural networks (CNN), recurrent neural networks [10], [2], artificial neural networks (ANN), and combinations of these.

1) Encoder-decoder framework: CNNs help anticipate words by assisting with object identification, localization, and interactions in conjunction with long-term dependencies from a recurrent network cell. Meaningful textual representations and descriptions are provided by a variety of CNN-based encoders; a novel recurrent fusion network (RFNet) [10] was presented. The sentiment of the image was labeled using a generative adversarial network (GAN) trained on a binary vector. [8]

- **Encoder:** Using the pre-trained VGG-16 convolutional neural network (CNN) as a feature extractor is the first step in the encoding process for picture captioning using VGG-16. Obtaining the VGG-16 model's pre-trained weights is the first step; this model is usually trained on extensive image classification datasets such as ImageNet [8].
- **Decoder:** The vanishing gradient problem is addressed with the Long Short-Term Memory (LSTM) [18] recurrent neural network architecture, which is utilized in sequence processing, mainly picture captioning. It captures and maintains long-term relationships in sequential data, using input images as input and predicting the probability distribution of subsequent words at each decoding step. The process iteratively continues until an end token is produced or the maximum sequence length is achieved.

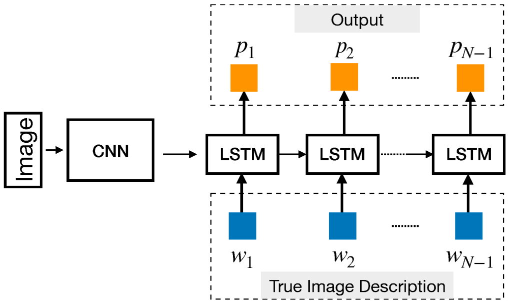


Fig. 2. LSTM as a decoder

In Fig 2 , we are defining LSTM model as decoder , in this an image is given as input to CNN , output of CNN is sent to LSTM which generates output sequences which are further used to generate captions.

C. Deep Learning

Multi-layered neural networks are used in deep learning, a subfield of machine learning, to learn hierarchical data representations. It has significantly advanced picture captioning, with CNNs extracting features from images and identifying patterns and spatial hierarchies. Recurrent Neural Networks (RNNs) and Transformer models are used to produce logical and relevant captions. RNNs allow for sequential dependencies in language, while transformer models understand long-range dependencies and are popular for parallelization [14]. The deep learning framework simplifies end-to-end training, improving image captioning systems' performance and allowing models to generalize well to various unknown inputs. This combination of deep learning and picture captioning has led to more sophisticated and context-aware automatic description creation, with applications in image indexing, retrieval systems, and content accessibility. [5]

$$a_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{1}$$

$$e_{ij} = f(S_{i-1}, h_i) \tag{2}$$

III. DATASETS

For deep learning to fully understand patterns and optimize the number of parameters required for gradient convergence [12], a large volume of training data is required.To facilitate evaluation and testing, it is standard procedure to partition the dataset into training, validation, and testing sets. Because of this separation, the model can be trained on one subset of the data, validated on another to adjust hyperparameters and track performance, and then tested on an entirely different subset to evaluate its generalization abilities. In order to prepare a solid and organized dataset for the ensuing training of picture captioning models, it is imperative that the preprocessing and data loading processes are carried out carefully. As a result, the following datasets are readily available and were created especially for the task of semantic segmentation: [12]

A. *Flickr8k*

Flickr8k [14] is a crucial dataset in computer vision and natural language processing, containing 8,000 unique photographs labeled with five human-generated descriptions per image. It is essential for training and testing image captioning algorithms, teaching models the complex relationships between written descriptions and visual information. After training, models undergo a thorough evaluation process to produce semantically meaningful captions similar to human references. Flickr8k is compact, allowing easy training on budget laptops or desktop computers. It includes 82 JPEG photos, with 6000 used for development, 1000 for testing, and 1000 for training. The dataset contains 40460 captions, five for each image. [10]

B. *Flickr30k*

Flickr30k is a massive dataset of 31,000 photos from Flickr, each with five human-created captions, aimed at improving image captioning research. It contains 31,783 photos of people in their daily activities, each with a caption that describes it. The dataset fills the need for larger training and assessment sets, enabling the development of more reliable and scalable image captioning algorithms. [5]

C. *MS COCO*

To push the boundaries of computer vision, Microsoft Common Objects in Context was created using standard photos, annotation, and assessment.Bounding boxes or object annotated features are used in the dataset for the object recognition task. It has more than 800,000 pictures available for its training, test, and validation sets, more than 500,000 segmented object instances, and a total of 80 object categories. [5]

IV. Evaluation Metrics

We assess the automatically generated captions to make sure they accurately describe the provided image. Some popular metrics for evaluating image captioning in machine learning are listed below. [12]

A. *BLEU*

(BiLingual Evaluation Understudy) The BLEU (Bilingual Evaluation Understudy) [10] score is a statistical tool used to evaluate the quality of machine-generated text, including image captions. It assesses the similarity of a generated caption to human-annotated reference captions. The process involves tokenization [16] of generated and reference captions into individual words or n-grams, calculation of precision, geometric mean of precision scores, and brevity penalty to correct biases favoring shorter captions. The shortness penalty is calculated by dividing the difference between the generated caption's length and the average length of reference captions, and multiplying the overall precision by the brevity penalty. This helps in determining the accuracy of captions. The BLEU score can be expressed mathematically as given in Table 1:

$$\text{BLEU} = \text{BP} \times \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log(\text{precision}_n)\right) \tag{3}$$

where BP is the brevity penalty, N is the maximum n-gram length considered, and precision-n is the precision for n-grams. This comprehensive formula captures both precision and brevity aspects, providing a numerical representation of the quality of machine-generated captions in image captioning tasks.

B. *METEOR*

(Metric for Evaluation of Translation with Explicit Ordering) It is based on a weighted F-score calculation and a penalty function intended to verify the candidate sequence's order, and it tackles the shortcoming of BLEU. For the purpose of identifying sentence similarities, it matches synonyms. [1]

V. Implementation

Fig 3 shows the flow diagram of image captioning process , in this an image dataset is given as input to CNN(encoder) for feature extraction , after that caption sequences are generated using LSTM (decoder) and then resulting captions are generated.

A. *Preprocessing and Data Loading*

The image captioning process involves preprocessing and data loading to convert unprocessed image and text data into a machine learning model. Preprocessing [17] involves cropping, normalization, and data augmentation techniques. Text preprocessing prepares written captions, while data loading organizes processed photos and captions into a training format. Batching speeds up the training process, while data shuffles

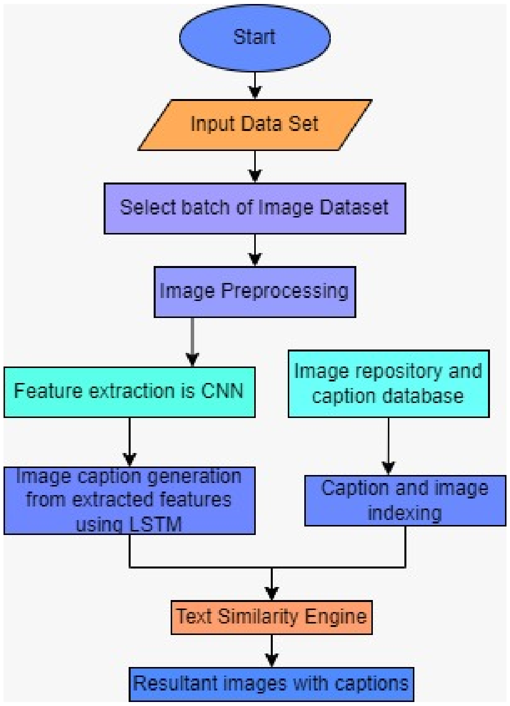

Start
Input Data Set
Select batch of Image Dataset
Image Preprocessing
Feature extraction is CNN
Image repository and caption database
Image caption generation from extracted features using LSTM
Caption and image indexing
Text Similarity Engine
Resultant images with captions

Fig. 3. Flow diagram of image captioning

prevent erroneous correlations. A data pipeline manages data loading and preprocessing on-the-fly during training.

- Load the Flickr8k dataset, which is made up of pictures with captions.
- Preprocessing: Adjust the photos' dimensions to a standard size and do any required normalization. Preprocess and tokenize the captions.

B. *VGG-16 Feature Extraction*

- Use a pre-trained VGG-16 model that was initially trained on ImageNet to extract high-level features from the images.
- The output of the final convolutional layer should be retained while the VGG-16 fully connected layers are removed. This output results in the visual representation.
- For the purpose of captioning photos, the VGG-16 model collects high-level information from the pictures. Rich representation of the input image is provided by the last convolutional layer of VGG-16, which preserves spatial information. Semantic aspects that include objects, forms, and patterns found in the image are recorded by VGG-16 [11]. These characteristics are essential for comprehending the visual context and producing insightful captions.

C. *Model Architecture*

- Construct an image captioning model architecture by fusing a language model for caption generation with the image attributes from VGG-16. [4]
- To turn the words in the captions into dense vectors, use an embedding layer.
- As the decoder, use an LSTM (Long Short-Term Memory) network to produce the captions' consecutive words.

D. *Training with Flickr8k*

The Flickr8k dataset, consisting of 8,000 photos with captions, is used to train a model for image captioning. Data preprocessing involves cleaning and tokenizing [13] captions, shrinking images, and normalizing pixel values. A Convolutional Neural Network extracts features, while tokenization and padding prepare captions for input. An encoder-decoder architecture with LSTM or GRU cells is used, with sequence-based loss functions and optimization approaches.

E. *Refine the model using Flickr30k*

- Adjust the model using the Flickr30k dataset. To do this, load the weights from the Flickr8k-trained model and carry on with the new dataset's training.
- By fine-tuning, the model may be made to work with a larger dataset and recognize a wider range of patterns in the image-caption pairs.

F. *Retuning the Model on Flickr8k*

- Adjust the model once more using the Flickr8k dataset. This step aims to improve generalization by maximizing the model's output on the original dataset. We can see Table II shows images and a comparison between reference caption of the dataset and predicted captions by our model. [9]

G. *Testing and Evaluation*

- Analyze the trained model's performance with metrics like BLEU, METEOR, etc., or on a different validation set.
- The accuracy of translations produced by machines is gauged by the BLEU (Bilingual Evaluation Understudy) score [8]. Based on word sequences (n-grams) that overlap between the proposed and reference translations, it determines precision. Brevity Penalty [5] takes into consideration translations that are shorter. The geometrical median of n-gram precisions, shortened for clarity, is the BLEU score. Higher scores denote higher quality translation; scores range from 0 to 1.

H. *Optimization and adjustments*

- Model Complexity: Modify the LSTM decoder's and other model elements' complexity. A model that is too simple could have trouble capturing complicated relationships, whereas a model that is too complex could result in overfitting [3].
- Training epochs and Batch Size: Try different batch sizes and epoch counts. In order to prevent overfitting or underfitting, keep an eye on both training and validation performance across epochs.
- Assessment Measures: To obtain a more thorough grasp of caption quality, take into account utilizing additional

assessment metrics in addition to BLEU, such as METEOR, CIDEr, and ROUGE [5].

- Data Augmentation: Use methods such as picture flipping, rotation, and scaling to add more images to the training dataset. The resilience and generalization of a model can be strengthened through augmentation.
- Hyperparameter Search: Look for the best possible hyperparameters by methodically examining various architectural parameters, LSTM units [13], and embedding dimensions. Either a random or grid search may be used in this investigation.
- Error Analyzation: Analyze model errors in test or validation sets using error analysis. Recognize the different kinds of errors the model commits and modify the training plan accordingly.

VI. Results and Findings

A. Epoch vs Accuracy Graphs

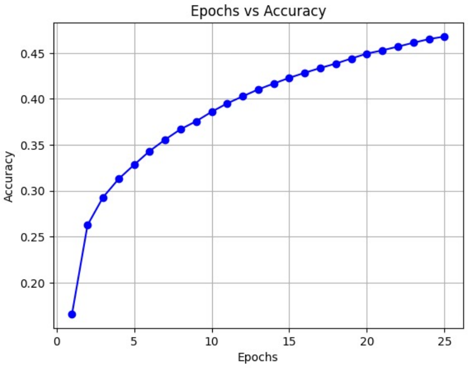


Fig. 4. For Flickr8k

Fig 4 shows a graph of Epochs vs Accuracy for Flickr8K dataset, graph shows an exponential increase in accuracy when model training is done for increased epochs values. Fig 5 shows a graph of Epochs vs Accuracy for Flickr8K on Flickr30k dataset , it shows a dynamic increase and decrease in the accuracy values on increasing epochs values for training of model.

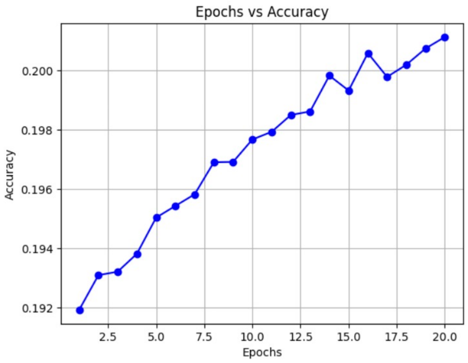
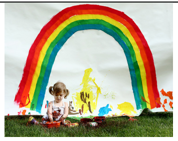
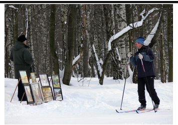
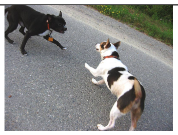
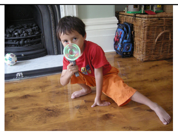


Fig. 5. For Flickr8k on Flickr30k

TABLE I
BLEU Scores for Datasets

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Flickr8k	0.53	0.30	0.11	0.17
Flickr30k	0.53	0.27	0.16	0.08
Flickr8k trained on Flickr30k	0.56	0.34	0.22	0.13

TABLE II
Captions Mapping

Image Id	Image	Caption
1002674143_1b742ab4b8		• **Reference Caption:** A little girl covered in paint sits in front of a painted rainbow with her hands inside bowl • **Predicted Caption:** two children enjoy the grass with fingerpaints in the background
101669240_b2d3e7f17b		• **Reference Caption:** A man in hat is displaying pictures next to a skier in a blue hat • **Predicted Caption:**Two people are the snow white in the snow
1001773457_577c3a7d70		• **Reference Caption:** A black dog and a spotted dog are fighting • **Predicted Caption:**Two dogs are playing in the grass
1096395242_fc69f0ae5a		• **Reference Caption:** A boy with a toy gun pointed at the camera • **Predicted Caption:** Young girl in the camera

VII. Discussion

The results obtained from algorithms for deep learning and machine learning and backbones differ according on how well they can acquire mappings from image inputs to labels. CNN-based approaches are the most efficient in tasks involving images, although they can be computationally expensive. Traditional machine learning algorithms like Markov random field, random forest, ensemble modeling ,naive Bayes, and support vector machines (SVMs)are overly basic and mainly dependent on handcrafted feature engineering or domain feature expertise. Clustering algorithms like K means and fuzzy C-means are not particularly successful when there are several boundaries. CNN is useful for spatial data, object detection, and localization due to its invariant property. Supervised learning approaches remain a popular technique, but data augmentation and generative adversarial networks have played significant roles in image captioning. Deep learning models have yielded the best performance across nearly all criteria, with the encoder-decoder architecture being the most common implementation. Attention mechanism concepts have been applied to improve feature computation, while generative adversarial networks and Autoencoders have been doing a great job of producing succinct image annotation. Reinforced learning techniques have also generated sequences that succinctly describe images in a timely manner.

VIII. Conclusion

This review study has explored the field of picture caption creation and highlighted the effectiveness of a hybrid CNN-LSTM model. Modern developments in the fields of image captioning and feature extraction have made it possible to examine important methods in detail and reveal their features and efficacy simultaneously. In this survey, methods that have produced outstanding results were explained, highlighting their function in effectively extracting, recognizing, and localizing objects in the image data. Simultaneously, the complex feature extraction procedure and its smooth conversion into a language model for picture captioning have been examined. A key accomplishment in this investigation is the incorporation of a CNN-LSTM hybrid model, which illustrates the cooperative relationship between convolutional neural networks and long short-term memory networks. This combination makes visual content easier to understand and makes it possible to create subtitles that make sense within their context. The results demonstrate how far natural language comprehension and computer vision have come, and how the hybrid approach has aided in the improvement of captioning performance. The survey results provide valuable insights for future research attempts in the dynamic convergence of computer vision and language modeling. This will help shape innovation in picture captioning as the field develops.

References

[1] Tushar Aggarwal. *Image Descriptive Summarization by Deep Learning and Advanced LSTM Model Architecture*. PhD thesis, 2019.

[2] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

[3] Ahmed Elhagry and Karima Kadaoui. A thorough review on recent deep learning methodologies for image captioning. *arXiv preprint arXiv:2107.13114*, 2021.

[4] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.

[5] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39, 2023.

[6] Ansar Hani, Najiba Tagougui, and Monji Kherallah. Image caption generation using a deep architecture. In *2019 International Arab Conference on Information Technology (ACIT)*, pages 246–251. IEEE, 2019.

[7] Xiaodong He and Li Deng. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine*, 34(6):109–116, 2017.

[8] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

[9] Licheng Jiao and Jin Zhao. A survey on the new generation of deep learning in image processing. *Ieee Access*, 7:172231–172263, 2019.

[10] Xiaoxiao Liu, Qingyang Xu, and Ning Wang. A survey on deep neural network-based image captioning. *The Visual Computer*, 35(3):445–470, 2019.

[11] Yue Ming, Nannan Hu, Chunxiao Fan, Fan Feng, Jiangwan Zhou, and Hui Yu. Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1339–1365, 2022.

[12] Ariyo Oluwasammi, Muhammad Umar Aftab, Zhiguang Qin, Son Tung Ngo, Thang Van Doan, Son Ba Nguyen, Son Hoang Nguyen, and Giang Hoang Nguyen. Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning. *Complexity*, 2021:1–19, 2021.

[13] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.

[14] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489, 2018.

[15] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022.

[16] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[17] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[18] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38, 2022.