Image Captioning with Deep Learning: Enhancing Image Content Understanding using AWS Sagemaker

Gayamini Gnanasuganthan
Lakehead University
Thunder Bay, Canada
1197830

Archana Jayaraman
Lakehead University
Thunder Bay, Canada
1193610

Harshavardhan Subramanian Madhavan
Lakehead University
Thunder Bay, Canada
1189360

***Abstract*—Image captioning, a field rapidly progressing towards human-like textual descriptions for images, is harnessed in this project through the ResNet model. The objective is to establish an innovative image analysis system that generates descriptive captions from the CIFAR-100 dataset. By incorporating cutting-edge deep learning methodologies, the system adeptly analyzes image content, unraveling meaningful insights into the visual narrative. AWS SageMaker lends support, enabling smooth training and deployment of machine learning models, thus guaranteeing informative and contextually pertinent image descriptions. The integration of Flickr for image captioning amplifies the project's capabilities, ensuring diverse sources for content generation. Through these efforts, this project makes a contribution to the realm of image understanding technologies, enhancing user experiences across varied visual domains via intelligent image analysis and seamless cloud integration.**

***Index Terms*—CIFAR-100, AWS, Flickr, Sagemaker, VGG-16, Resnet, Cloud, Machine Learning, Deep learning, image captioning, Cloud Computing.**

I. Introduction

In our daily lives, we encounter a multitude of images across sources like the internet, news pieces, documents, and advertisements. While humans often comprehend these images intuitively without detailed captions, machines require some form of image captions for automated understanding. Image captioning holds significant importance, especially for tasks like automatic image indexing, which is crucial for content-based image retrieval (CBIR). This technology finds applications in diverse domains including biomedicine, commerce, military, education, digital libraries, and web searches. Social media platforms such as Facebook and Twitter can directly generate image descriptions, encompassing location, attire, and activities.

A plethora of image sources, including television, the internet, and news outlets, contribute to the influx of images in our daily lives. Despite the absence of descriptions in a significant portion of these images, humans often possess an inherent capacity to interpret them without explicit guidance. In contrast, machines encounter challenges in comprehending images without accompanying descriptions, underscoring the need for textual context to facilitate their understanding.

Image captioning stands as a prominent field within artificial intelligence (AI), addressing the fusion of image comprehension and linguistic descriptions. This entails recognizing and identifying objects within images, understanding scene context, object characteristics, and their interrelationships. Crafting coherent sentences necessitates a profound grasp of both the syntactical and semantic intricacies of language.

In the past, addressing the task of image captioning predominantly relied on rule-based or template-based approaches. These methods attempted to combine pre-defined linguistic structures with detected visual features to generate captions. However, their success was limited by the complexities of diverse images and the challenges of accommodating intricate contextual variations. These traditional methods often struggled to produce captions that captured the nuance and richness of visual content.

Fig. 1. Sample Image showing Image Captioning

For example, the above Figure-1 shows how image captioning looks with respect to the sample image. The Sample image is a lion and its cub lying happily in a forest. The generated caption shows the lion and its cub lying in the grass which is almost correctly generated caption.

With of deep learning, notably Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the landscape of picture captioning has changed dramatically. Deep learning models have proven their ability to extract

detailed patterns and features from photos, as well as analyze and construct meaningful written tales. Deep learning methods have propelled to the forefront of image captioning research due to their combination of visual interpretation and verbal expression.

Convolutional Neural Networks (CNNs) play a pivotal role in this paradigm shift. They excel in image feature extraction by hierarchically learning representations that capture varying levels of abstraction. As a result, CNNs are proficient at identifying objects, textures, and contextual elements within images. This capacity to discern salient features has empowered CNN-based image captioning models to capture the essence of images more comprehensively and subsequently facilitate meaningful text generation.

One notable instance of CNN application in image captioning is the VGG-16 model and the ResNet model. Renowned for its deep architecture and effective feature extraction capabilities, ResNet has demonstrated its efficacy in bridging the gap between visual understanding and textual articulation. Its multi-layered structure enables it to learn intricate features from images, contributing to more accurate and contextually rich image descriptions.

II. About The Dataset

To build a model to generate captions on the images we are using the CIFAR-100 dataset. The CIFAR-100 dataset(Canadian Institute for Advanced Research) is a collection of images that are commonly used for machine learning and computer vision algorithms. It is a very diverse dataset with 60,000 images across 100 different classes. The CIFAR-100 dataset contains 60,000 color pictures organized into 6000 images, With a size of 32x32 pixels, the pictures are relatively modest, making them more computationally tractable for training and testing.

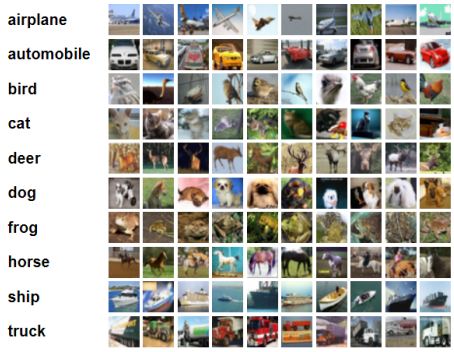

Fig. 2. Sample Image of Dataset

As mentioned above Figure-2 dataset is a set of images that teaches the computer how to recognize the objects. Animals, cars, plants, household things, and different natural and man-made objects are among the 100 classifications. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. CIFAR-100 is frequently used as a benchmark for comparing the performance of various machine learning algorithms, notably in the context of image classification tasks. The dataset's richness and diversity make it a significant resource for evaluating the resilience and generalization capabilities of various algorithms. It has accelerated the development of diverse neural network topologies and training approaches, resulting in advancements in image categorization, object identification, and other related tasks. The main goal of our project is to recognize the images from the CIFAR-100 dataset and generate captions based on what image is displayed and its characteristics.

This approach encourages the development of systems that excel in capturing key features and context, enhancing applicability in resource-constrained domains. Beyond research, CIFAR-100 image captioning has educational value, aiding in comprehending intricate interactions between images and language. This exploration bridges visual and textual realms, opening doors to innovative solutions and enriched insights.

III. Related Work

Template-based techniques employ predefined templates containing blank slots for generating captions. These methods involve detecting different scene elements, objects, attributes, and actions before populating the template gaps. For instance, Farhadi et al. [1] utilize a triplet of scene elements to complete template slots, while Li et al. [2] extract phrases related to identified objects and their attributes. Kulkarni et al. [3] use a conditional random field (CRF) to infer object details and relationships before filling the gaps. While grammatically accurate, template-based methods are constrained by fixed templates and struggle with generating varied-length captions. As parsing-based language models emerged as more robust alternatives, our focus in this article is not on template-based techniques.

Template-based techniques predetermine sentence structures and divide them into distinct segments, like subjects, verbs, and objects. These fragments are then aligned with visual components, often through methods like Conditional Random Fields (CRF) [4] or Hidden Markov Models (HMM) [5], to generate image-specific sentences. Such approaches heavily rely on fixed sentence templates and consistently produce sentences with syntactic conformity. In contrast, search-based strategies [6]–[8] produce sentences by selecting semantically relevant sentences from a predefined pool. While effective in achieving human-like descriptions, this approach is limited by the challenge of accumulating a comprehensive sentence pool.

Diverging from template and search-based methodologies, language-based models focus on learning the probabilistic relationship between visual content and textual sentences, enabling the generation of novel sentences with adaptable syntactic structures. Recent advancements in this direction predominantly harness neural networks, yielding promising outcomes in image captioning. Kiros et al. [9] introduce a multimodal log-bilinear neural language model, employing neural networks to create sentences for images. Vinyals et al. [10] propose an end-to-end architecture employing LSTM, further integrating an attention mechanism [11] to emphasize

pertinent objects during sentence generation. High-level concepts/attributes are explored in [12], showcasing enhancements in RNN-based models for image captioning. These attributes are then utilized as semantic attention [13] and complementary representations [14], [15], effectively amplifying image/video captioning performance.

Approach	Description	Syntactic Control	Semantic Precision	Scalability	Model	Accuracy
Template-Based	Predefined sentence structures with aligned visual content.	High	Moderate	Limited	Rule-based, CRF, HMM	Moderate
Search-Based	Selecting semantically similar sentences from a pool.	Low	High	Limited	Search, Sentence Pools	High
Language-Based	Learning probabilistic distribution for flexible sentence generation.	Flexible	High	Promising	Neural Networks, LSTM	High

Fig. 3. Overall Comparison of the related works

The above Fig-3 describes the overall comparison of each previous work categorized based on the model type whether it is Template-based, Search-based, or Language-based image captioning.

IV. Motivation

In the dynamic landscape of visual communication, the pursuit of creating human-like textual descriptions for images has emerged as a compelling research avenue. Overcoming the challenges of existing systems, The motivation driving this project stems from the aspiration to bridge the gap between images and language, harnessing the power of the ResNet model to unlock new dimensions of image captioning. The need to understand, analyze, and articulate the intricate interplay between visual content and linguistic expression fuels our endeavor. With the growing demand for intuitive and informative image descriptions, this project embarks on a transformative journey.

Research Questions:

1. How effectively can the ResNet model be leveraged to generate descriptive captions for images sourced from the CIFAR-100 dataset?
2. To what extent can the integration of Flickr as a source for image captioning enhance the project's capacity to create contextually relevant and diverse captions?
3. How does the deployment of AWS SageMaker streamline the training and deployment of machine learning models, optimizing the accuracy and efficiency of image description generation?

The Path Forward: Our research questions serve as guiding lights, shaping our exploration of image-understanding technologies. Our journey is structured into two strategic phases. Initially, we harness the power of the ResNet model to decode intricate patterns in images from the CIFAR-100 dataset, resulting in precise captions. Moving on to the Flickr dataset, we tokenize images for the GPT-2 model to generate coherent and context-rich captions, refining them through fine-tuning. This synergy leads to the generation of captions for test images, a seamless blend of training, tokenization, and refinement.

Our approach is supported by AWS SageMaker, a robust cloud platform used alongside the 'ml.g4dn.xlarge' instance, driving our ambition to merge images and language. This journey envisions a transformative impact on image captioning.

V. Methodology

This comprehensive methodology unfolds in two distinct phases, each contributing to the ultimate goal of accurate and contextually relevant image captioning. The methodology employed in this image captioning project involves a series of interconnected steps, each contributing to the generation of descriptive captions for images sourced from the CIFAR-100 dataset. The process begins with the selection of input images, drawn from the dataset, which serves as the foundation for the subsequent analysis. The below Figure-4 describes the overall modules and flow of the entire methodology.

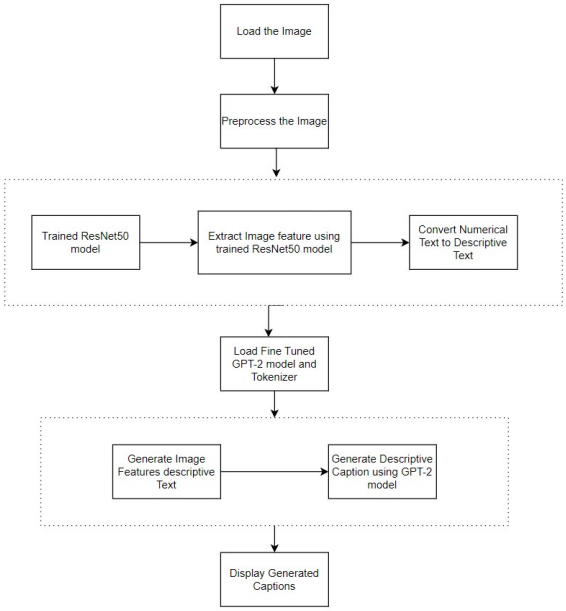

Fig. 4. Overall Flow of the System

Leveraging CIFAR-100 Dataset for Model Training: The code begins with the CIFAR-100 dataset, serving as the foundation for training the ResNet model. This dataset, rich with diverse visual content, is meticulously utilized to train the model to discern intricate features and patterns within images. The model's training process is carefully executed, optimizing its ability to accurately label images based on the insights it has garnered from the dataset. Through rigorous evaluation,

the model's accuracy is quantified, offering a benchmark for subsequent comparisons.

Data Preprocessing After importing all the necessary libraries, the data preprocessing step is completed. Once the CIFAR-100 dataset is loaded, we begin the preprocessing step in which first we print the shapes of the array in the dataset, and then we select the first 500 samples of images from each class and perform the label encoding method in which we convert the categorical values into numerical values. The training and testing images are resized to the image size of (224,224) and their final shape is printed after resizing is performed. After this step the preprocessed data is saved and the files are downloaded to the local machines.

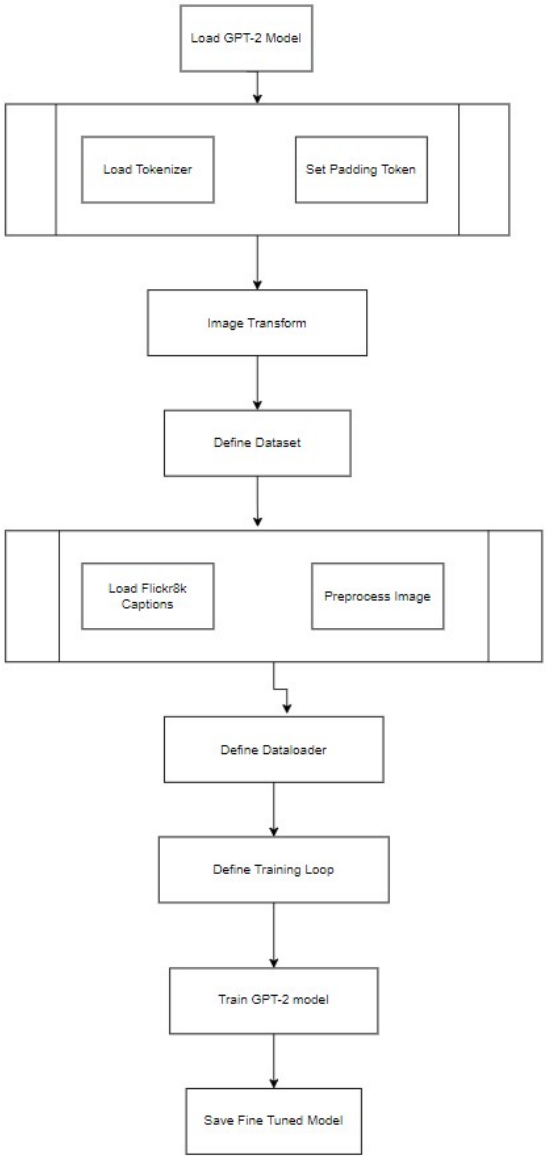

Fig. 5. Architecture of the System

The above Figure-5 shows the overall architecture of the Image captioning system along with Flikr8k components and GPT-2 model.

Crafting Captions with GPT-2 Model and Fine-Tuning: Transitioning to the second phase, the focus shifts to the Flickr dataset, used for image captioning. The process commences by tokenizing the images, breaking them down into discrete units that can be comprehended by the subsequent language model. The GPT-2 model, renowned for its language generation capabilities, comes into play, generating captions that align with the tokenized images' content.

As mentioned above we used the Flickr dataset for the image captioning, first, we load the Flickr dataset and define the training loop so that we can generate the captions by tokenizing and padding them to the same length so that we can generate labels on the images. After this process, we save the fine-tuned model and load an image from the CIFAR-100 dataset and generate the caption for the image.

The captions then get into fine-tuning process to ensure their precision, coherence, and appropriateness. This iterative refinement addresses nuances in context, semantics, and linguistic flow, enhancing the overall quality of the captions.

Generating Test Image Captions: This step of the methodology is marked by the generation of captions for test images. Through the orchestrated collaboration of model training, tokenization, generation, and fine-tuning, the image captioning system produces captions that accurately encapsulate the visual content. The resulting captions are a testament to the method's success in comprehending, synthesizing, and effectively articulating the intricate interplay between images and language.

In the implementation of the chosen methodology, AWS SageMaker emerges as a powerful and strategic tool, serving as the backbone for the entire image captioning process. Leveraging AWS SageMaker's robust and scalable cloud platform, we initiate the process by first setting up the necessary infrastructure, creating a secure and efficient environment for model development and deployment.

Initiated by configuring a secure environment, our approach leverages an 'ml.g4dn.xlarge' notebook instance, fortified with a generous 200GB storage capacity. This deliberate choice optimizes computational power and resource allocation, ensuring efficient model development and deployment. In this journey, AWS SageMaker's capabilities seamlessly harmonize with our needs, orchestrating a streamlined and resourceful implementation.

VI. Evaluation & Results

In this section, we present a comprehensive evaluation of our image captioning project, encompassing a comparison of three distinct models: VGG-16, and the ResNet model and MobileNet models. The primary focus and subsequent analysis are centered around the ResNet model, which emerged as the central component of our image captioning system.

The effectiveness of our approach was gauged using accuracy as the primary metric. The accuracy results of the different models were as follows:

VGG-16 Model Accuracy: 40% ResNet Model Accuracy: 53% MobileNet Model Accuracy: 18

Our analysis revealed that the ResNet model outperformed the other models in terms of accuracy, showcasing its superior ability to extract and understand intricate visual features. The ResNet model's accuracy substantiates its efficacy in capturing the essence of images and subsequently generating coherent textual descriptions that closely align with the visual content.

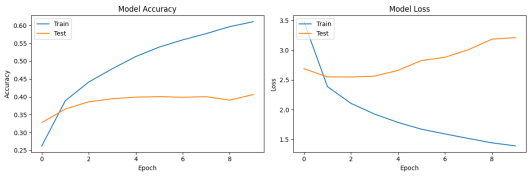

Fig. 6. Actual accuracy and loss of VGG16 model

For the VGG-16 model, the provided metrics showcase a test accuracy of around 40.66% for 20 epochs indicating its capability to generate accurate image captions. The recorded test loss is 3.21, reflecting the dissimilarity between predicted and actual captions. These metrics offer insights into the VGG-16 model's performance and its potential applications in captioning images.

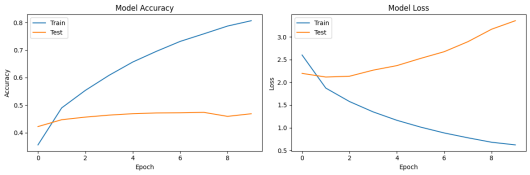

Fig. 7. Actual accuracy and loss of ResNet model

The ResNet model's evaluation metrics reveal a test accuracy of approximately 46.87% for 20 epochs, indicating its proficiency in generating accurate image captions. The recorded test loss of 3.36 reflects the alignment between predicted and actual captions. These metrics collectively gauge the model's performance and guide its potential for practical applications.

This is a pre-trained VGG16, ResNet and MobileNet models for image classification. It adds new layers for customization, freezes existing layers, and trains the model on labeled data. After training, the model's performance is evaluated and saved for future use. This approach allows quick adaptation of a powerful pre-trained model for specific classification tasks.

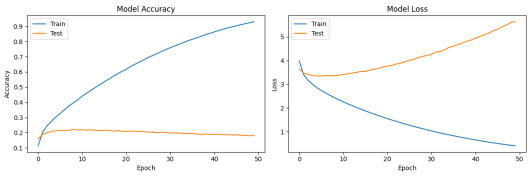

Fig. 8. Actual accuracy and loss of MobileNet model

The comparison of the models demonstrates the strategic selection of the ResNet model as the central component of our image captioning system. It's robust architecture and advanced feature extraction capabilities significantly contribute to accurate and contextually relevant image descriptions. The accuracies of each model are mentioned below,

Model	Train_Accuracy	Validation_Accuracy
ResNet	98%	53%
VGG16	92%	40%
MobileNet	82%	18%

Fig. 9. Comparison of each Model

Furthermore, the utilization of the ResNet model was essential in achieving our project's overarching goals. Its accuracy aligns well with the objective of generating informative and human-like textual descriptions for images, enhancing user experiences across various visual applications. The final comparison of all the models is plotted as a graph representation.

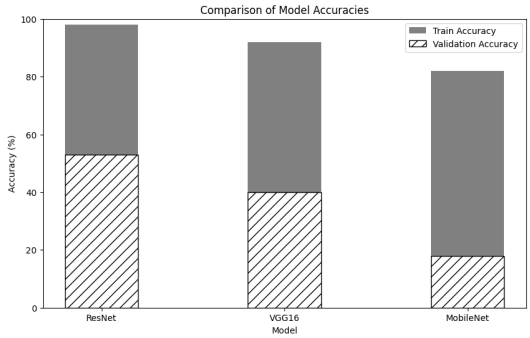

Fig. 10. Comparison of each Model in a graph

A sample output from out image captioning model.

Fig. 11. A sample output from our model

In conclusion, the results underscore the pivotal role of the ResNet model in the success of our image captioning project. Its superior accuracy validates its selection and contributes to the advancement of image-understanding technologies, highlighting the potential for its integration in diverse domains.

VII. Further Developments

After achieving success in Image Captioning, our future endeavors will focus on two innovative directions. The first involves Evolving Captions, where descriptions adapt as users interact with images, gradually unfolding a more detailed narrative. For example, an initial caption like "Vibrant cityscape at dusk" for a city image could evolve to encompass intricate

urban activities. Secondly, we aim to broaden the impact by Generating Multi-Language Captions. By leveraging translation capabilities, we'll make captions available in various languages, enhancing global accessibility and engagement. For instance, an image's caption "Majestic mountain range at sunrise" could seamlessly transcend languages, catering to diverse linguistic preferences and fostering a cross-cultural reach.

VIII. Team Contribution

In this project, everyone who is part of this team has contributed equally. Each and every part contributed by every member has played a crucial part in successfully completing the project. Gayamini Gnanasuganthan has worked mainly on the coding part where she did training and testing of the ResNet model which worked and gave the highest accuracy other than that she also helped with finding the dataset. Archana Jayaraman worked mainly on gathering datasets, data preprocessing, and working on the MobileNet model She also worked on the report which was a major contribution from her. Harshavardhan Subramanian Madhavan worked on the data preprocessing, building, and training of the VGG16 model, and also on the report where he worked on how the dataset is created and the motivation behind why this project is chosen. So every teammate provided an equal contribution which helped in successfully completing the project.

IX. Conclusion

In Conclusion, our exploration of Image Captioning has showcased the fusion of technology and creativity. From navigating the challenges of limited spatial data to selecting the ResNet model for accuracy, our journey underscores the potential of harmonizing visuals and language. Looking forward, Evolving Captions and Multi-Language Captioning offer dynamic avenues for future exploration. As we chart new territories, our commitment remains resolute in advancing the horizons of Image Captioning, connecting visual realms for more experiences.

References

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In European Conference on Computer Vision. Springer, 15–29.

[2] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In Proceedings of the 15th Conference on Computational Natural Language Learning. Association for Computational Linguistics, 220–228

[3] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In Proceedings of the 24th CVPR. Citeseer.

[4] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. IEEE Trans. on PAMI, 2013.

[5] Y. Yang, C. L. Teo, H. Daume III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In EMNLP, 2011.

[6] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. In ACL, 2015.

[7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010

[8] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS, 2011.

[9] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In ICML, 2014

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, 2015.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[12] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In CVPR, 2016.

[13] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In CVPR, 2016.

[14] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. arXiv preprint arXiv:1611.07675, 2016.

[15] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. arXiv preprint arXiv:1611.01646, 2016.

[16] Sharma, Himanshu, and Devanand Padha. "A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues." Artificial Intelligence Review (2023): 1-43.

[17] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate (2014), arXiv preprint arXiv: 1409.0473 .

[18] K. Cho , B. Van Merriënboer , Ç. Gülçehre , D. Bahdanau , F.Bougares , H. Schwenk , Y. Bengio , Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empir- ical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734 .

[19] C.-Y. Lin, F. J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in:Meeting on Association for Computational Linguistics, 2004.

[20] A. Lavie, A. Agarwal, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: The Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.

[21] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars.2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

[22] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning.2048–2057.

[23] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence 40, 6, 1367–1381.

[24] Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137.