Encoder-Decoder with Visual Attention Image Caption Generator

Yap Yun Onn[a], A.P. TS. Dr. Sri Devi A/P Ravana[b]

[a]*Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Wilayah Persekutuan Kuala Lumpur, Malaysia*
[b]*Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Wilayah Persekutuan Kuala Lumpur, Malaysia*

Abstract

This research paper presents a comparative analysis of two encoder-decoder image captioning models employing different attention mechanisms: Luong-style Attention and Bahdanau-style Attention. Evaluation metrics include BLEU scores (1 to 4), ROUGE-L, CIDEr, and METEOR. The Attention model exhibits varying performance with different search strategies, with Beam Search (n=1) achieving the highest BLEU-1 and BLEU-2 scores. The Additive Attention model consistently outperforms across all beam widths in Beam Search, achieving superior BLEU scores, ROUGE-L, and CIDEr scores. Greedy Search for the Additive Attention model also shows strong performance. The results emphasize the efficacy of Additive Attention, particularly with Beam Search (n=1), in enhancing image captioning model performance. **Keywords: image captioning, attention mechanisms, Encoder-Decoder, Additive Attention, Beam Search.**

1. INTRODUCTION

The mission to bridge the gap between human understanding and machine perception has inspired the development of sophisticated artificial intelligence systems that can not only recognize visual content but also describe it in natural language. One remarkable achievement in this pursuit is the realm of image caption generation, a field that has been significantly advanced through the innovative integration of deep learning techniques. Open domain based image captioning system is a very demanding task,as it requires a detailed understanding of the major and the minor entities in an image,as well as their attributes and relationships (Rennie et al. (2017)). Two seminal works, ”Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” (Xu et al. (2016)), and ”Show and Tell: A Neural Image Caption Generator” (Vinyals et al. (2015)), have illuminated the path towards more human-like machine-generated image descriptions.

This project will work on the development of an image caption generator that not only ”sees” images but also ”understands” them. We employ an encoder-decoder architecture enriched with attention mechanisms, enabling the model to attend to different parts of an image as it generates captions. This dynamic approach allows for more contextually relevant and informative descriptions, transcending the limitations of earlier traditional systems.

The project seeks to harness the strengths of both the ”Show, Attend and Tell” (Xu et al. (2016)) and ”Show and Tell” (Vinyals et al. (2015)) approaches to create a system that excels in generating detailed, coherent, and context-aware image captions. Through this research, we aim to contribute to the broader objective of fostering a deeper connection between artificial intelligence and human understanding, thereby making image caption generation a vital tool for various applications, from accessibility services for the visually impaired, medical imaging to content enrichment in the media and advertising industries.

2. LITERATURE REVIEW

Initially, image captioning relied on basic methods like template-based methods Farhadi et al. (2010) , where the authors presented a system that can compute a score linking an image to a sentence. This score can be used to attach a descriptive sentence to a given image or to obtain images that illustrate a given sentence. The score

Email addresses: s2105674@siswa.um.edu.my (Yap Yun Onn), sdevi@um.edu.my (A.P. TS. Dr. Sri Devi A/P Ravana)

is obtained by comparing an estimate of meaning obtained from the image to one obtained from the sentence, using a discriminative procedure that is learned using data. These methods, while straightforward, lacked the nuance and flexibility to handle the diversity and complexity of real-world scenes. Composition-based methods offer slightly more sophistication. In the year 2013, ranking-based method Hodosh et al. (2013) emerged as a slightly more complex way which uses nearest-neighbor search and various types of kernel canonical correlation analysis.The following year of 2014, composition-based methods Kuznetsova et al. (2014) presented an unusual approach for image caption generation. It focuses on creating a tree structure that represents the content of an image and then compressing this tree to form a coherent and concise sentence. The method involves analyzing existing image descriptions, extracting phrases, and combining them to form new descriptions.

A big shift began with the use of deep learning techniques, which introduced models that could understand the complex relationship between verbal patterns and visual features and learn to construct caption sequences. The encoder-decoder architecture, a centrepiece in sequence learning, has been crucial in neural machine translation (NMT) and image captioning advancements. Pioneering works by Cho et al. (2014a), Cho et al. (2014b) and Sutskever et al. (2014), along with advancements in image captioning Vinyals et al. (2015), have significantly contributed to the development and refinement of this architecture. The framework fundamentally consists of two parts: the encoder, which processes the input data into a fixed-dimensional context vector(state), and the decoder, which generates the output from this context Cho et al. (2014b). Cho et al. (2014a) and Sutskever et al. (2014) emphasize the framework's ability to handle variable-length input and output sequences, a significant improvement over previous fixed-length approaches.

Cho et al. (2014b) and Sutskever et al. (2014) utilize RNNs in their models, leveraging their capacity for temporal data processing. RNNs, fundamental in the initial encoder-decoder models, faced challenges with long-range dependencies, leading to the adoption of Long Short-Term Memory (LSTM), Greff et al. (2016) or Gated Recurrent Units (GRUs), Cho et al. (2014b). LSTMs and GRUs effectively retain information over longer sequences, making them ideal for complex sentence structures in translation tasks and detailed image descriptions. Sutskever et al. (2014) leveraged LSTMs in their sequence-to-sequence learning model, significantly improving the handling of long sequences in machine translation. Cho et al. (2014a) also underscored the effectiveness of GRUs and gating in capturing longer dependencies, crucial for translating complex sentences which showcased that GRUs are neither worse or better than gated LSTMs.

The applications of encoder-decoder architectures in neural machine translation (NMT), Cho et al. (2014a) highlight NMT's efficiency compared to statistical machine translation, particularly in joint training of encoder-decoder components. It was also highlighted by Sutskever et al. (2014), who demonstrated its efficiency in English-to-French translation tasks. The encoder in NMT converts a source language sentence into a dense vector representation, which the decoder then translates into the target language, maintaining the sentence's semantic integrity. Cho et al. (2014b)'s work on learning phrase representations further illustrates how the encoder-decoder model can effectively capture the semantic and syntactic structures of sentences. On the image captioning side, Vinyals et al. (2015) extends the encoder-decoder model to image captioning, where a CNN encodes an image into a feature-rich vector and an RNN decodes this vector into a descriptive caption. This adaptation illustrates the framework's versatility, bridging the gap between visual data processing and linguistic output.

The later stages involved integrating visual attention mechanisms in encoder-decoder architectures. This development, as discussed in the key papers by Bahdanau et al. (2014), Luong et al. (2015), Vaswani et al. (2017), and Xu et al. (2016), has revolutionized the way machines interpret and generate language based on visual and sequential data. The attention mechanism addresses the limitations of previous encoder-decoder models by allowing the model to focus on specific parts of the input sequence for each step of the output sequence. This approach has led to substantial improvements in the quality of machine-generated translations and image captions. Bahdanau-style attention Bahdanau et al. (2014), also known as additive attention, computes a context vector by focusing on different parts of the input sequence, which enhances the model's ability to manage long input sequences. Luong-style attention Luong et al. (2015), on the other hand, offers a more streamlined approach, often computationally efficient, with different alignment functions like dot, general, and concat. Xu et al. (2016) apply visual attention mechanisms in image captioning, enabling the model to focus on primary features within images dynamically. This method significantly improves the relevance and accuracy of generated captions, they were able to achieve a BLUE-1 score of 71.8 implementing Hard-Attention. While Vaswani et al. (2017) introduces the Transformer model, a new architecture that relies entirely on attention mechanisms, refraining from the need for recurrence and convolutions. This model was able to achieve the state of the art BLEU Score of 41.0 on the WMT 2014 English-to-French translation task using self-attention and position-wise feed-forward networks, providing improved parallelization and training time less than ¼ of the previous state-of-the-art model.

Furthermore, the advancements in neural networks, particularly in the context of image captioning, have been significantly influenced by the development of Convolutional Neural Networks (CNNs) and their efficient scaling, as highlighted in the works of EfficientNet, Tan and Le (2019) and EfficientNetV2, Tan and Le (2021). These advancements have direct implications for image captioning models, enhancing both their performance and efficiency. CNNs are vital in image captioning for extracting rich, hierarchical visual features from images. The depth, width, and resolution of CNNs determine their ability to capture fine-grained details, crucial for generating accurate captions. Tan and Le (2019)'s EfficientNet introduced a systematic approach to scaling CNNs, balancing network depth, width, and resolution, leading to models that are both efficient and powerful. The EfficientNetB7 was able to hit state-of-the-art 84.4% in top 1 accuracy while being 8.4x smaller and 6.1x faster than GPipe Huang et al. (2019) which was the best performing ConvNet at the time. This balance ensures that the feature extraction in image captioning models is not only more comprehensive but also computationally efficient. In 2021, Tan and Le (2021) further advanced EfficientNet architectures with EfficientNetV2, focusing on reducing training time and model size without sacrificing accuracy, EfficientNetV2 achieved 87.3% top-1 accuracy on ImageNet ILSVR2012, bested the recent ViT-L/16 by 2.0% accuracy while training for 5 to 11 times faster. Techniques like progressive image resizing and Fused-MBConv in EfficientNetV2 contribute to faster and more effective training of image captioning models. EfficientNets, with their optimized architectures, have shown remarkable performance on benchmarks like ImageNet, which directly translates to improvements in the quality of image captioning. EfficientNets also excel in transfer learning, effectively adapting pre-trained models to the specific task of image captioning. This adaptability enhances the model's performance, especially when dealing with diverse and complex datasets.

3. DATA SCIENCE METHODOLOGY

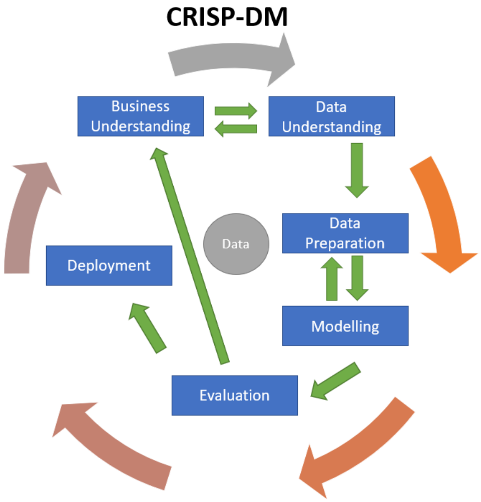


Figure 1: CRISP-DM (Huber et al. (2019))

3.1. Business Understanding

3.1.1. Problem Statements

The generation of human-like image captions is a challenging problem at the intersection of computer vision and natural language processing. With the goal of improving the quality and context-awareness of image captions, our data science project, "Encoder-Decoder with Visual Attention Image Caption Generator," aims to address the following key challenges:

1. Image Understanding: Existing image captioning systems often struggle to comprehensively understand the visual content, missing out on fine-grained details, subtle relationships between objects, and contextual nuances present in images.
2. Sequence Data: Traditional deep neural networks (DNNs) have face challenges in handling variable-length sequences.Sutskever et al. (2014) As they can only be applied to situations whose variables and targets that are encoder with a fixed dimension of vectors.
3. Adaptation to Diverse Contexts: The project aims to adapt the image captioning model to work effectively across various contexts and settings, challenging the

model's ability to understand and describe images in dynamic environments that differ significantly from the training data.

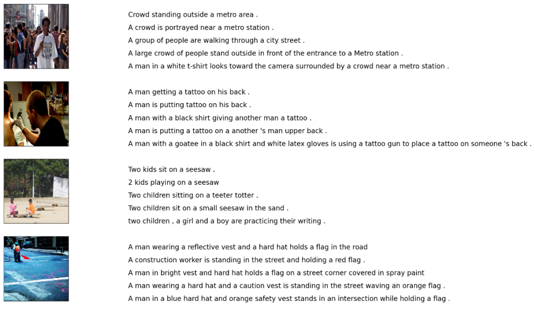
3.1.2. Objectives

The three objectives of this project are:

1. **To model Context-Aware Image Captioning:** Improve the contextual understanding of images by incorporating the encoder-decoder architecture (Rennie et al. (2017)). This architecture will capture relationships between the visual content and textual descriptions, leading to the generation of more coherent and context-aware image captions.
2. **To implement an Attention Mechanism into the Model:** Develop and integrate different visual attention mechanism (Bahdanau et al. (2014)) and (Luong et al. (2015)) into the image captioning model, allowing it to focus on relevant image regions during the caption generation process. This attention mechanism will be inspired by the principles outlined in the "Show, Attend and Tell" approach (Xu et al. (2016)) and aim to enhance the model's captioning accuracy and relevance
3. **User-Friendly Interface:** Develop a user-friendly interface to allow users to upload images and receive detailed and meaningful captions. The interface will serve as a practical application of the model's capabilities, benefiting users in various domains, including accessibility services and content enrichment.

3.2. Data Understanding

For this project, Flickr30K dataset (Plummer et al. (2016)) will be used as the main dataset to train and validate the model. The format of the dataset is image-caption pair values. Total of 31,783 images each contains 5 different captions, which amount to a grand total of 158,915 captions.



Crowd standing outside a metro area .
A crowd is portrayed near a metro station .
A group of people are walking through a city street .
A large crowd of people stand outside in front of the entrance to a Metro station .
A man in a white t-shirt looks toward the camera surrounded by a crowd near a metro station .

A man getting a tattoo on his back .
A man is putting tattoo on his back .
A man with a black shirt giving another man a tattoo .
A man is putting a tattoo on a another 's man upper back .
A man with a goatee in a black shirt and white latex gloves is using a tattoo gun to place a tattoo on someone 's back .

Two kids sit on a seesaw .
2 kids playing on a seesaw
Two children sitting on a teeter totter .
Two children sit on a small seesaw in the sand .
two children , a girl and a boy are practicing their writing .

A man wearing a reflective vest and a hard hat holds a flag in the road
A construction worker is standing in the street and holding a red flag .
A man in bright vest and hard hat holds a flag on a street corner covered in spray paint
A man wearing a hard hat and a caution vest is standing in the street waving an orange flag .
A man in a blue hard hat and orange safety vest stands in an intersection while holding a flag .

Figure 2: Visualization of Flickr 30K dataset

Before doing exploratory data analysis (EDA), the captions in the dataset have gone through a series of preprocessing tasks to improve and standardize the caption's quality. Several common methods were used, namely:

1. Converting all alphabets to lowercase.
2. Removed all punctuation.
3. Special character such as '@', '#', '&', '$' etc., non-alphabets such as numbers, multiples spaces were all removed.

Stop words were not removed because it allows the model to learn in utilizing the stop words to establish sentence structure and coherence in everyday communication, literature, and content creation. Furthermore, stop words also facilitate proper grammar, ensuring generated captions remain fluent and easily understandable. This decision was inline of the objective of this project.

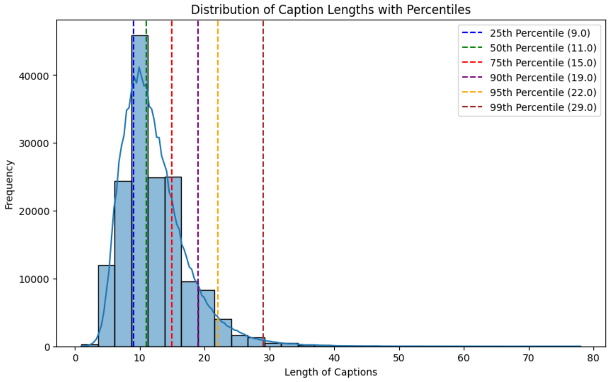


Figure 3: Caption Length Distribution

The histogram in Figure 3 shows that the distribution of caption length is concentrated around 10 to 15 words. The 25th percentile is at 9 words, and the 75th percentile is at 15 words, while the 50th percentile (the median) is at 11 words, which means half of the captions have 11 or fewer words. This implies that the dataset favors conciseness, which is often desirable in captioning to maintain user engagement and effectively convey information. The 95th and 99th percentiles show that only 5% of captions are longer than 22 words, and only 1% are longer than 29 words. For the project, the maximum caption length is set to 64 words, as this balances the usage of computation resources and avoid over truncating meaningful captions.

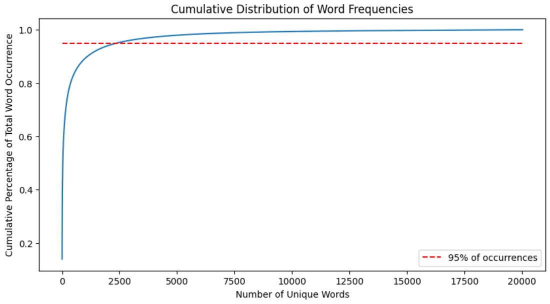


Figure 4: Cumulative Distribution of Word Frequencies

The line chart in Figure 4 shows that there a total of 20,028 unique words and 2,376 unique words make up 95% of the total word occurrences. This indicates that 2,376 is a good starting point for the size limit of the vocabulary. However, for this project we would set the vocabulary size to 12,500 to account for important and rare words that could represent primary object and actions that defines a picture, this can decrease the model's complexity without significantly sacrificing the ability to understand and generate captions. It also help reduces the computational cost and can also speed up the training process, as the model has fewer parameters to learn.

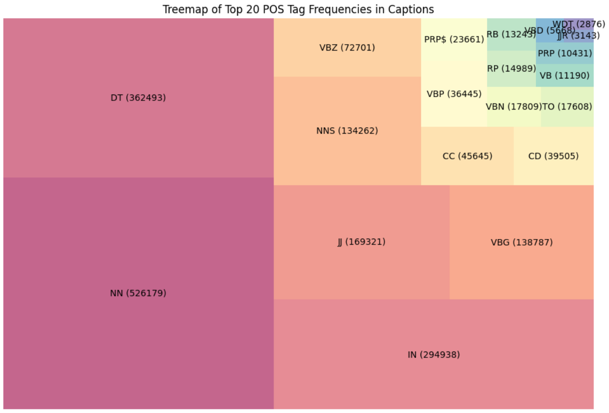


Figure 5: Top 20 Part-Of-Speech Tag Frequencies in Captions

The treemap in Figure 5 shows that dominant POS tags such as nouns (NN) and determiners (DT) implies that captions heavily focused on identifying and describing objects. The presence of adjectives (JJ) and various verb forms (VBG - present participles, VBN - past participles) indicates that the captions not only identify and describe objects but also their actions as well.

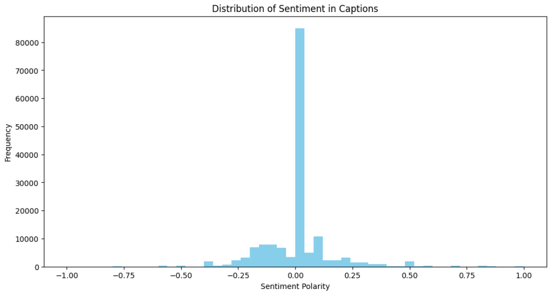


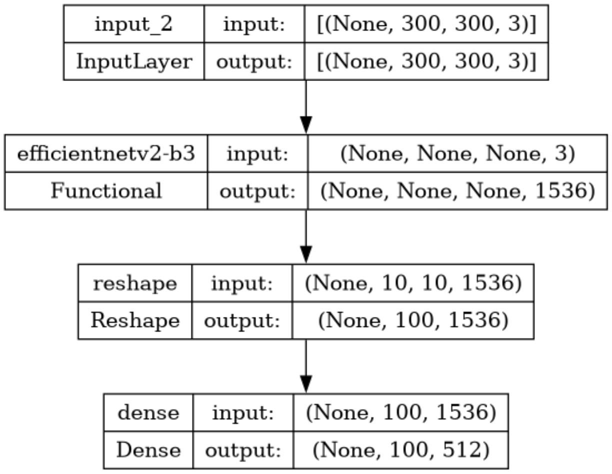
Figure 6: Distribution of Sentiment in Captions

The distribution shown in Figure 6 suggests that most captions won't express strong emotional responses. The tallest bar at sentiment polarity near 0 indicates that the majority of captions have a neutral sentiment, it could imply that model will have a tendency to use language with a neutral connotation. This is common in image captions, which tend to be descriptive rather than emotive. The spread of sentiment across captions, even if predominantly neutral, suggests varying degrees of emotional expression that the image captioning model might need to capture.

3.3. Data Preparations

- Preprocess the data to make it suitable for model training. This includes resizing images, image augmentations, and tokenizer fitted on the vocabulary.
- Split the data into training, validation, and testing sets for model development and evaluation.

3.4. Modelling


input_2 | input: | [(None, 300, 300, 3)]
InputLayer | output: | [(None, 300, 300, 3)]

efficientnetv2-b3 | input: | (None, None, None, 3)
Functional | output: | (None, None, None, 1536)

reshape | input: | (None, 10, 10, 1536)
Reshape | output: | (None, 100, 1536)

dense | input: | (None, 100, 1536)
Dense | output: | (None, 100, 512)

Figure 7: Encoder

- The encoder in Figure 7 will be responsible for converting an input image into a representation that captures the visual information present in the image. It will utilize a pre-trained convolution neural network (CNN) model, EfficientNetV2 (Tan and Le (2021)) and etc as the backbone structure. These pre-trained model are used to extract features from images, encode the features and output them in a fixed-size feature vector that represent the image content. The extracted features are then reshaped and passed to a dense layer for compression to prepare it for the attention layer. This feature vector serves as the initial input for the decoder.

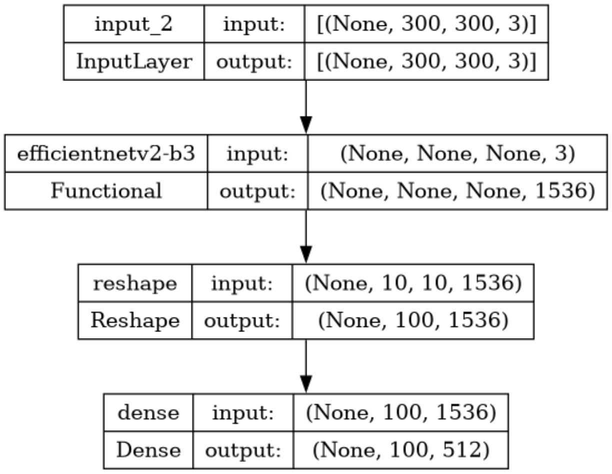


Figure 8: Decoder

- The decoder in Figure 8 is responsible for generating a caption, word by word, based on the feature representation provided by the encoder and its own internal state. Word inputs (tokens) are embedded into a higher dimensional vector space. It employs Gated Recurrent Unit (GRU),Cho et al. (2014b), a type of recurrent neural network (RNN), for sequential data generation. The GRU processes these embeddings, updates its internal state to track the sequence's context from previous words.The decoder is equipped with Luong-style attention, Luong et al. (2015) uses the GRU (decoder) outputs to attend over the encoder's output features, it works by first encoding them into a group of hidden states and then determining a series of attention weights by using a score function (dot function in this case) in Figure 9(c). These attention weights define the relevance of each hidden state in generating the output. The outputs are then formed by summing the hidden states and weighting them according to the generated attention weights. Then attention weight are then used to compute the context vector in Figure 9(b). The attention mechanism allows the model to focus on different regions of the image as it generates each word thus enhances the relevance and contextuality of the generated captions. A skip connection design was also implemented as the GRU output is passed to the attention layer, and also to the Add layer directly, this helps in generalization during training. Then, the Add & Normalization Layer combines the GRU output and the context vector (from attention), and normalize it to stabilize the learning process.

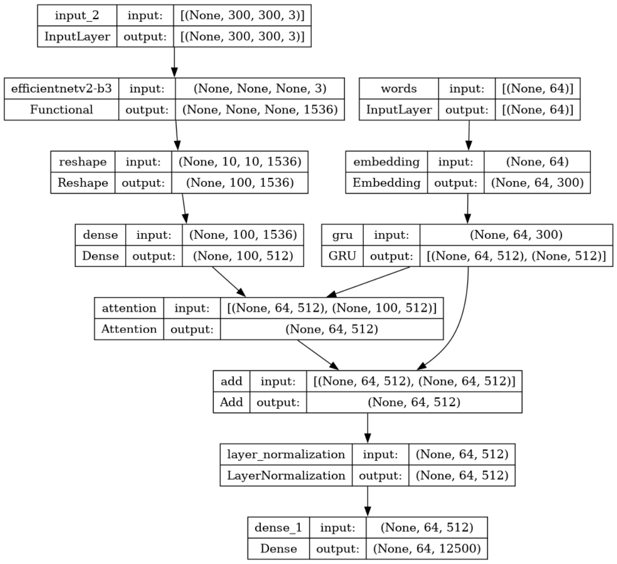


Figure 10: Encoder-Decoder Model

- The combination of the encoder and the decoder to learn to generate captions from images.The final output is a 3-D array of (none, 64, 12500) where 64 means the max caption length and 12500 is the vocabulary size. For each word in a 64 words caption, there are 12500 word probabilities.

3.5. Evaluations

- The model's performance was evaluated using 4 different metrics: BLEU (Papineni et al. (2002)), ROUGE (Lin (2004)), CIDEr (Vedantam et al. (2015)), and METEOR (Banerjee and Lavie (2005)).

$$\alpha_{ts} = \frac{\exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \qquad \text{[Attention weights]} \qquad (1)$$

(a) Attention Weight Equation

$$\boldsymbol{c}_t = \sum_{s} \alpha_{ts} \bar{\boldsymbol{h}}_s \qquad \text{[Context vector]} \qquad (2)$$

(b) Context Vector Equation

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^{\top} \boldsymbol{W} \bar{\boldsymbol{h}}_s & \text{[Luong's multiplicative style]} \\ \boldsymbol{v}_a^{\top} \tanh\left(\boldsymbol{W_1} \boldsymbol{h}_t + \boldsymbol{W_2} \bar{\boldsymbol{h}}_s\right) & \text{[Bahdanau's additive style]} \end{cases} \qquad (4)$$

(c) Score Function

Figure 9:
s is the encoder index.
t is the decoder index.
α_{ts} is the attention weights.
h_s is the sequence of encoder outputs being attended to (the attention ”key” and ”value” in transformer terminology).
h_t is the decoder state attending to the sequence (the attention ”query” in transformer terminology).
c_t is the resulting context vector.
a_t is the final output combining the ”context” and ”query”.

- The model’s parameter has been iterated through different configurations and the best one is kept.

3.6. Deployment

An open-source framework platform such as Streamlit will be used as a medium to deploy the image caption generator , thus allowing the public users to generate image captions by uploading images.

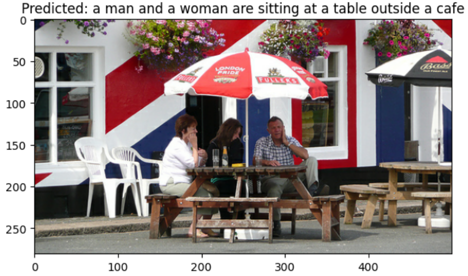
4. RESULTS

The evaluations result are reported on Flickr 8k dataset with a randomize selection of 2000 samples for each method. The same preprocessing and preparations were applied for this evaluations as well. By comparing two models with different attention mechanisms in Table 1 using different inference method, the performance of models was evaluated based on BLEU scores (1 to 4), ROUGE-L, CIDEr, and METEOR. The first model employed a standard Attention(Luong-style) mechanism, while the second utilized Additive Attention(Bahdanau-style). For the Attention model, we observed variations in performance across different search strategies. The Beam Search with n=1 outperformed others, achieving the highest BLEU-1 and BLEU-2 scores (0.6299 and 0.4572 respectively). However, the performance decreased with larger beam widths, indicating a trade-off between precision and diversity. Greedy Search demonstrated competitive results, and Stochastic Sampling showed lower scores across all metrics. In contrast, the Additive Attention model consistently demonstrated superior performance across all beam widths in Beam Search, outperforming the Attention model in BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores. The Additive Attention model with Beam Search (n=1) achieved the highest overall scores, indicating its effectiveness in generating captions that align closely with reference captions. Greedy Search for the Additive Attention model also exhibited strong performance, surpassing other strategies in BLEU-1, BLEU-2, and BLEU-3 scores. Notably, the Additive Attention model consistently outperformed the Attention model in ROUGE-L and CIDEr scores, suggesting its capability to capture longer sequences and generate more descriptive captions. METEOR scores were comparable between the two models, indicating similar overall translation quality. Overall, the findings highlight the efficacy of the Additive Attention mechanism, particularly in the context of Beam Search with a beam width of 1, for enhancing image captioning model performance. The choice of attention mechanism and search strategy significantly influences the model’s ability to generate diverse, contextually relevant, and grammatically sound captions.

Table 1: Evaluation Results of all metrics

		BLEU						
Model	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	METEOR
Attention	Beam Search, n=1	**0.6299**	**0.4572**	0.2882	0.2007	**0.3132**	0.6604	**0.2825**
	Beam Search, n=2	0.6256	0.4554	**0.2887**	**0.2024**	0.3131	**0.6635**	0.2817
	Beam Search, n=3	0.6043	0.4405	0.2777	0.1936	0.3087	0.6526	0.2710
	Greedy Search	0.6295	0.4560	0.2865	0.1987	0.3105	0.6203	0.2796
	Stochastic Sampling	0.5443	0.3802	0.2255	0.1534	0.2425	0.3403	0.2164
Additive Attention	Beam Search, n=1	**0.6377**	**0.4635**	**0.2935**	**0.2037**	**0.3161**	**0.6526**	0.2825
	Beam Search, n=2	0.6216	0.4526	0.2848	0.1970	0.3133	0.6457	0.2766
	Beam Search, n=3	0.6047	0.4413	0.2787	0.1953	0.3103	0.6394	0.2682
	Greedy Search	0.6350	0.4618	0.2917	0.2025	0.3146	0.6481	**0.2835**
	Stochastic Sampling	0.5454	0.3824	0.2285	0.1566	0.2431	0.3386	0.2163

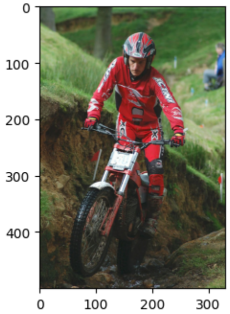
Ground Truth: three people are sitting at an outside picnic bench with an umbrella
Predicted: a man and a woman are sitting at a table outside a cafe



BLEU_1: 0.308
BLEU_2: 0.226
BLEU_3: 0.167
BLEU_4: 0.000
ROUGE-L: 0.308
METEOR: 0.310

Figure 11: Sample from Attention Model

Ground Truth: a man in red riding gear riding a dirt bike down a path
Predicted: a man in a red shirt and black helmet is riding a dirt bike



BLEU_1: 0.643
BLEU_2: 0.497
BLEU_3: 0.395
BLEU_4: 0.274
ROUGE-L: 0.597
METEOR: 0.585

Figure 12: Sample from Additive Attention Model

5. DISCUSSION

The evaluation findings highlight limitations in generating sensible captions for certain domains due to the small and less diverse training dataset (Flickr 30k). Future improvements could involve training on larger and more diverse datasets to enhance adaptation across various settings. Additionally, addressing the basic nature and the lack of human emotional engagement in generated captions suggests exploring advanced NLP techniques, pretrained word embeddings, and language models like Bert, GPT, or Gemini for more linguistically rich and engaging captions. Our current models also struggles when the image is slightly more chaotic and complex, as seen in Figure 11 where it fails to capture the existence of 3 person. Therefore, experimenting with different forms of attention, such as spatial or channel-wise attention, could potentially optimize the model's focus on relevant image regions.

6. CONCLUSION

The future of encoder-decoder models with visual attention in image captioning looks promising. With ongoing research addressing the challenges of generalization, abstract concept interpretation, and computational efficiency, these models are bound to become even more robust and versatile. The exploration of new neural architectures and attention mechanisms, coupled with efforts to reduce bias and improve evaluation methods, will further enhance their effectiveness. In conclusion, the results have shown that image captioning model with Additive Attention mechanism generally performs better than its counterpart that uses the Attention mechanism. As indicated in

Figure 12, it was able to capture slightly complex actions (riding), objects (man, dirt bike) and colours (red, black). The Github repository for this project can be accessed at this **link**.

7. APPENDIX/USER MANUAL

The web application has been deployed on Streamlit Community Cloud which can be accessed **here**.

Home Page: The main page that allows users to choose models, set the desired beam width and upload an image to generate a caption. **Note:** Only one image can be uploaded for each usage.

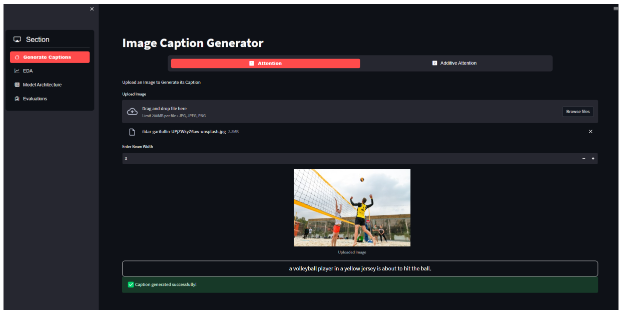


Image Caption Generator

Upload an Image to Generate its Caption

Upload Image

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

ildar-garifullin-UPjZWkyZ6aw-unsplash.jpg 2.3MB

Enter Beam Width

3

a volleyball player in a yellow jersey is about to hit the ball.

Caption generated successfully!

Figure 13: Home Page

EDA Page: The page that consists of all the EDA visualisations done in this project.

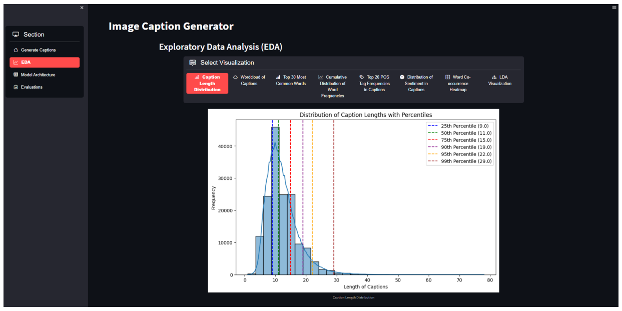


Image Caption Generator

Exploratory Data Analysis (EDA)

Figure 14: EDA Page

Model Architecture Page: The page that consists of the architectures of all models.

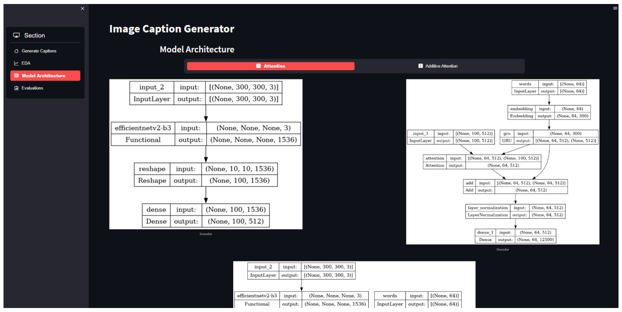

Section
Generate Captions
EDA
Model Architecture
Evaluations
Attention
Additive Attention
input_2 | input: | [(None, 300, 300, 3)]
InputLayer | output: | [(None, 300, 300, 3)]
efficientnetv2-b3 | input: | (None, None, None, 3)
Functional | output: | (None, None, None, 1536)
reshape | input: | (None, 10, 10, 1536)
Reshape | output: | (None, 100, 1536)
dense | input: | (None, 100, 1536)
Dense | output: | (None, 100, 512)
Encoder
words | input: | [(None, 64)]
InputLayer | output: | [(None, 64)]
embedding | input: | (None, 64)
Embedding | output: | (None, 64, 300)
input_3 | input: | [(None, 100, 512)]
InputLayer | output: | [(None, 100, 512)]
gru | input: | (None, 64, 300)
GRU | output: | [(None, 64, 512), (None, 512)]
attention | input: | [(None, 64, 512), (None, 100, 512)]
Attention | output: | (None, 64, 512)
add | input: | [(None, 64, 512), (None, 64, 512)]
Add | output: | (None, 64, 512)
layer_normalization | input: | (None, 64, 512)
LayerNormalization | output: | (None, 64, 512)
dense_1 | input: | (None, 64, 512)
Dense | output: | (None, 64, 12500)
Decoder
input_2 | input: | [(None, 300, 300, 3)]
InputLayer | output: | [(None, 300, 300, 3)]
efficientnetv2-b3 | input: | (None, None, None, 3)
Functional | output: | (None, None, None, 1536)
words | input: | [(None, 64)]
InputLayer | output: | [(None, 64)]

Image Caption Generator

Model Architecture

Figure 15: Model Architecture Page

Evaluations Page: The page that consists of all the evaluations charts on all models for each method.

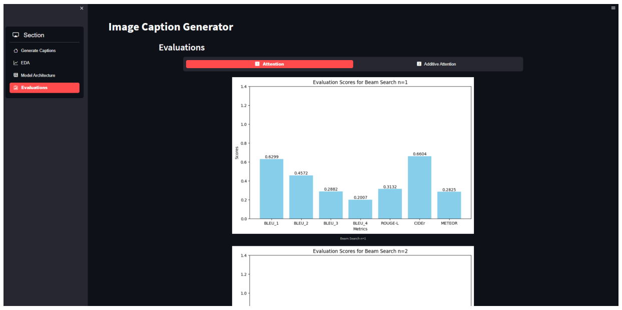


Image Caption Generator

Evaluations

Figure 16: Evaluations Page

8. ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. Sri Devi A/P Ravana for her support, guidance and invaluable suggestions throughout the course of this research project. Her expertise and mentorship have played a huge role in shaping the outcome of this project. Lastly, it is my hope that the outcome of this project will benefit both the academic community and practitioners, continuing advancements in greater techniques and enhancing applications in real-world scenarios.

References

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

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop* *on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2016). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., et al. (2019). Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32.

Huber, S., Wiemer, H., Schneider, D., and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications–a holistic extension to the crisp-dm model. *Procedia Cirp*, 79:403–408.

Kuznetsova, P., Ordonez, V., Berg, T. L., and Choi, Y. (2014). Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2016). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2016). Show, attend and tell: Neural image caption generation with visual attention.