Big Data and Deep Learning Final Project

Dufera Amanuel Tafese 3123999178, Saif Hamdi Mohammed Ahmed 3123999192

Mahmudul Hasan 3123999147, Zhou Zijian 3123154005, Md Jamir Hossain Tuhin 3123999070

May 2024

Pre-Trained CNN Architecture for Transformer-Based Image Caption Generation Model

1   Abstract

Automatic image captioning, a multifaceted task bridging computer vision and natural language processing, aims to generate descriptive textual content from visual input. While Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have achieved significant advancements, they present limitations. The inherent sequential nature of RNNs leads to sluggish training and inference times. LSTMs further struggle with retaining information from earlier sequence elements when dealing with very long sequences. This project presents a comprehensive guide to constructing and comprehending transformer models for image captioning. Transformers employ self-attention mechanisms, capturing both short- and long-range dependencies within the data. This facilitates efficient parallelization during both training and inference phases. We leverage the well-established Transformer architecture, recognized for its effectiveness in managing sequential data, and present a meticulous methodology. Utilizing the Flickr30k dataset, we conduct data pre-processing, construct a model architecture that integrates an Ef-

ficientNetB0 CNN for feature extraction, and train the model with attention mechanisms incorporated. Our approach exemplifies the utilization of parallelization for efficient training and inference.

2　Introduction

Many of the captioning algorithms now in use to capture the essence of an image in words are based on the encoder-decoder architecture, where a decoder infrastructure can use an attention method to predict words by employing a function obtained from an encoder network. The translation strategy using a visual encoder and a language decoder has been the major focus of studies on picture subtitling [13].

Image captioning serves many purposes, ranging from enabling autonomous driving and enhancing face recognition systems to assisting individuals with visual impairments and improving photo search quality. This challenging task involves extracting image features through computer vision (CV) techniques and generating natural language descriptions using natural language processing (NLP) technology. The impact of photo caption generation extends across diverse domains, including image search, software development for accessibility, video surveillance, security, and human-computer interaction [6].

The challenge of generating photo captions, a prominent task in sequence modeling, has seen the adoption of various methodologies. One popular approach combines Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) in a CNN-RNN framework. However, training RNN-based models, especially those like Long Short-Term Memory (LSTM) networks, can be challenging due to their sequential nature, leading to suboptimal performance [10]

To address this, recent advancements have introduced the Transformer architecture, overcoming the limitations of sequential processing through its parallelizable design and context-aware attention mechanism. Unlike traditional RNN-based models, Transformers can operate efficiently during both training and inference without relying on specific sequences, thus achieving state-of-the-art results in photo captioning tasks [15] [17]

The focus of caption generation studies has primarily been on English-language datasets, leading to the widespread adoption of attention-based mechanisms. While many utilize the VGG-16 architecture as the encoder in captioning models, alternative approaches incorporate pre-trained networks like AlexNet or Residual Networks (ResNet) to extract visual features. Additionally, some researchers leverage Bidirectional Long Short-Term Memory (BiLSTM) networks for language modeling in caption generation tasks[3] [17]

In addition to English, various datasets have been curated for caption generation in other languages such as Chinese, Japanese (e.g., Yoshikawa dataset), Arabic, and Bahasa Indonesia. Custom datasets have also been developed, combining sources like MS COCO and Flickr30k for languages like Indonesian. Moreover, specific language versions of popular datasets like Flickr30k, such as FEEH-ID Flickr8k, have been created to cater to diverse linguistic needs [1] [19][7]

[21] introduced a novel approach by leveraging Transformers in two distinct streams within their architecture: one for the visual component and another for the linguistic aspect. They combined a CNN model for encoding image features with a Transformer model for decoding captions. This architecture involved separate Transformer models for both the encoder and decoder, incorporating stacked self-attention mechanisms. [20] also demonstrated the effectiveness of

using CNN as an encoder to extract image features, generating a context vector that encapsulates essential visual information. This vector is then input into the Transformer for caption generation based on the extracted features.

The structure of this paper is outlined as follows: In Section 3, reviews the related work in the field. Section 4 provides an overview of the methodology of this study and section 5 details of our model design and implementation, while Section 6 offers a detailed performance analysis, highlighting the impact of different configurations. Finally, in Section 7, we draw conclusions.

3   Related works

Image description is an important problem in the field of Artificial Intelligence, in simple terms, the input is an image to the model, and the output of the model is a text sentence that can describe the image scene. There are many methods related to image description, the more famous ones are:

Traditional methods: early approaches to image caption generation typically used a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) or Recurrent Neural Networks (RNNs). These methods encode images into feature vectors and then use RNN or LSTM to generate subtitles[5].

Attention Mechanism Based Methods: later methods introduced an attention mechanism that allows the model to focus on different regions in the image when generating subtitles, thus improving performance and generation quality [8].

Transformer-based approaches: Recently, the Transformer architecture has achieved great success in natural language processing tasks, and some research has begun to apply it to image caption generation

tasks. These approaches use Transformer to capture complex relationships between images and text to generate more accurate and coherent captions[11]. When talking about work related to image caption generation, some classic papers and approaches are a must see, they give inspiring thoughts and give us a deeper understanding of the field, here are some of the classic papers as well as some of the latest technological advances.

Show and Tell: A Neural Image Caption Generator [16]: This is one of the classic work in the field of image caption generation. It introduces an end-to-end deep learning model that combines CNN and LSTM for encoding images into feature vectors and generating corresponding captions. The method was one of the pioneers in the task of image caption generation and laid the foundation for subsequent research.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [18]: This work introduces an attention mechanism based on the classical Show and Tell model, allowing the model to dynamically focus on different regions in the image when generating subtitles. In this way, the model can more accurately describe the image content and improve the quality of the generated subtitles.

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering [2] This work proposes a combined approach of bottom and top-level attention mechanisms for image caption generation and visual quizzing tasks. The bottom-level attention mechanism extracts local information from the image, while the top-level attention mechanism integrates this information in a global context. This combination allows the model to better understand images and generate more coherent captions.

BERT (Bidirectional Encoder Representations from Transformers) [4] BERT is a pre-trained Transformer-based language model that

has achieved great success in the field of natural language processing. While originally designed to process textual data, some research has also begun to explore the application of BERT or similar Transformer models to the task of image caption generation to capture the complex relationships between images and text.

Image Transformer [12] This work proposes a pure Transformer architecture for processing image data. It partitions an image into a series of blocks and processes these blocks through the Transformer model, which enables the learning of a representation of the image. Although this approach is initially used for image classification and segmentation tasks, it also provides a new idea for applying Transformer to image related tasks. These papers represent some of the classic and most recent work in the field of image caption generation, demonstrating advances in deep learning and natural language processing in the field. The language in which the captions are ultimately generated depends on the configured dataset, e.g., the processing of a language involves tasks such as word segmentation, speech recognition, machine translation, etc. Previous work may include building databases for that language, developing NLP tools and models based on that language, etc.

4   Methodology

The methodology adopted for this project incorporates experimenting of data on the deep learning model architecture. The Flicker30k datasets for image captions are available publicly, hence two procedures are followed for developing the image captioning system. The preparation of the dataset comes first, then comes the design and implementation of the model architecture. Dataset Preparation: Two tasks are involved in the production of the dataset: preprocessing and

<start> a bunch of people in camo pants run . <end>
<start> A group of people in army pants running . <end>
<start> A group of people wearing army clothes run together down the street . <end>
<start> A group of men and woman wearing army pants and a white shirt are running . <end>
<start> A group of men in camouflage pants and white t-shirts are running in a drill yard . <end>

<start> Three men are sitting at a wooden table . <end>
<start> Three men in a rustic cabin working with laptops . <end>
<start> Three man sitting around a table working with their laptops . <end>
<start> Three men are sitting around a picnic table , one of which is using a laptop computer . <end>
<start> Three men sitting at a table , one man pointing at the camera , with four red candles on the right . <end>

<start> A man sews in a garment factory . <end>
<start> Man sewing with a machine surrounded by garments . <end>
<start> A man sits at a sewing machine surrounded by items . <end>
<start> A man is sitting beside a bunch of material while sewing . <end>
<start> A brown-haired man is sitting in front of a sewing machine looking down . <end>
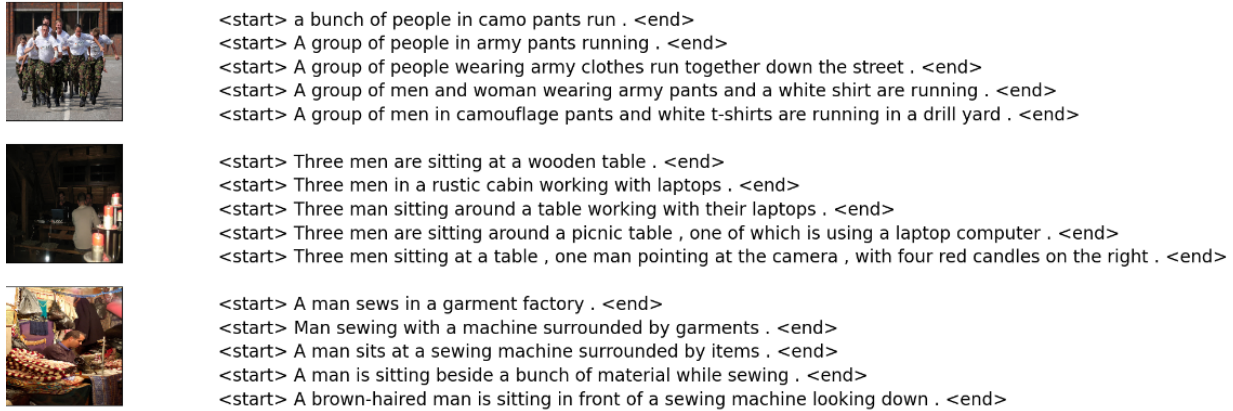
Figure 1: An examples from the dataset: Images and the five captions of each image

dataset collecting. The resulting dataset is split into three sections: one for testing, one for validation, and one for training. Dataset Collection: The Flickr30k public dataset provided the datasets gathered for this study. It comprises of more than 300,00 photos with 5 captions each. The public datasets are available for download.

Dataset Preprocessing: Pre-processing involves caption normalization, length filtering, and dataset partitioning. Image processing operations, such as decoding, and resizing, are implemented. The dataset is prepared using TensorFlow functions, including shuffling, mapping, batching, and prefetching for optimal parallelization. Dataset Creation: Following cleaning, the caption data are divided into three sets: a training set, a validation set, and a testing set. These sets contain, respectively, 20915, 5124, and 105 images. Using the TensorFlow "Dataset" library, the caption data are then mapped to the corresponding images and compressed to produce datasets for training and validation.

5 Model Design and Implementation

The suggested method for image captioning incorporates a CNN model for image feature extraction and a Transformer network for
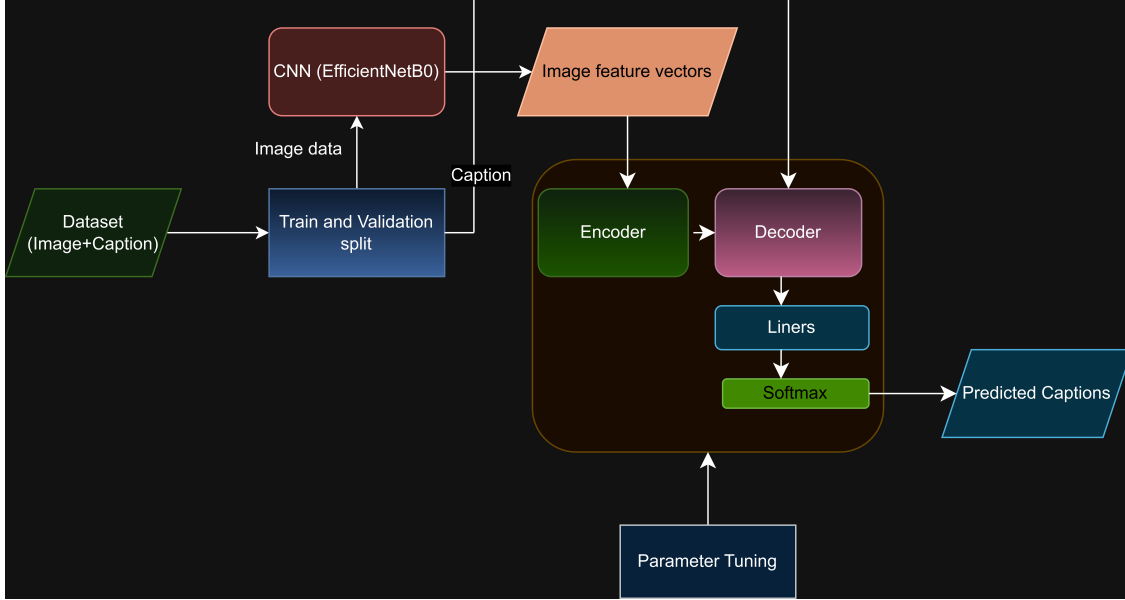
Figure 2: The model architecture

language modelling. Figure 2 displays the system architecture for this work. The CNN model can be developed from scratch but due to its advancement in recent years more accurate and efficient pre-trained models are available for use. Hence, 'EfficientNetB0' used to extract picture characteristics. The input image is encoded by CNN into a vector representation, which the decoder uses to produce captions. Since this job does not include categorization, the CNN's final softmax layer is eliminated

EfficientNet is a CNN model as well as a scaling technique that uses a set of preset scaling coefficients to uniformly scale depth, width, and resolution dimensions, in contrast to standard practice, which scales these variables arbitrarily [14] There are 8 variants of EfficientNet (B0 - B7). EfficeintNetB0 is used for this project because of its simplicity and relatively good performance. The transformer model is used to generate captions instead of the conventional Recurrent Neural Network-based architecture because RNN doesn't support parallelization and transformer networks have outperformed RNN in

Parameters	Value
Image Size	(299,299)
Max. Vocab Size	13000
Sequence Length	24
Embedding Size	512
Batch Size	128
Optimizer	Adam
Loss Function	Categorical-crossentropy

Table 1: Model Parameters

NLP tasks in recent years.Our proposed system follows the transformer architecture proposed by [15], The main components of the transformer networks consist of the encoder, decoder, positional encoding, embeddings, softmax, and multiheaded attention. Attention is utilized in the transformer model to find the relevant collection of values based on a few keys and queries. Attention weights, which are derived using the encoder hidden state (Key) and decoder hidden state (Value), have recently been used to give priority to distinct encoder hidden states (values) in processing the decoder states (query) [9].

Single attention-weighted values have been found to be insufficient to capture the many features of the input. The transformer model thus employs multiheaded attention for tackling this challenge. Similarly, positional encoding is used by the transformer networks to keep information about the order of sequence by adding the relative or absolute position of the tokens in sequence [15]. The positional encodings are added to the bottom of the encoder and decoder stacks. In order to implement our model, the datasets generated are passed to both of the CNNTransformer architecture. The input images of size (299x299) are passed to the CNN encoder to generate image vectors. The image vectors are then passed to the transformer encoder. The transformer decoder part is fed with the respective

captions to train the model. The encoder part comprises a single multi-headed attention head and a normalization layer whereas 2 multi-headed attention heads and 3 normalization layers are used in the decoder. These models are implemented using the TensorFlow Keras library. Table 1 shows the model parameters used in this work. The model parameters are chosen based on explicit experimentation on Flicker30K dataset. These are the optimal parameter values as per our research which can be further improved with the introduction of larger datasets and via parameter tuning.

6   Results and Discussions

In order to create captions, we employed Adam's optimization algorithm to train a Transformer-based model. This model was trained with a 30-epoch cross-entropy loss on our Flickr30k dataset. Prior to this training procedure, we utilized transfer learning by including components of a pre-trained model into our new model. Using a pre-trained encoder for photo captioning is usually better than starting from scratch with a fresh model and no prior knowledge. It is done within fifty epochs and halted if there has been no progress in BLEU-4 over the most recent 10 epochs (the stopping training criterion), while the total training procedure is done in 5-12 hours on each pre-trained CNN architecture. Python with Tensorflow's library is the performance analysis environment for each CNN model that contains three phases: (1) Training period. (2) Validation. (3) Testing. In other words, we apply the parallelization to it, as the Transformer's design permits the simultaneous process.

   A dataset of over 30,000 image-caption pairings is developed which are separated into training, validation and testing sets. The "test" results were assessed using our evaluation approach, where each sen-

Generated Caption: a group of people are standing outside of a building

BLEU-1: 0.25769

BLEU-2: 0.50763

BLEU-3: 0.66578

BLEU-4: 0.71248

Generated Caption: a surfer in a black wetsuit is riding a wave

BLEU-1: 0.33636

BLEU-2: 0.57997

BLEU-3: 0.72118

BLEU-4: 0.76156

Figure 3: Model output that shows the prediction caption with score of BLEU.

tence created by the model in the provided picture was compared to the target sentences used as sentence references to calculate the BLEU score. These scores measure only matching grams of a given order, such as single words (1-gram) or word pairs (2-gram or bigram), and so forth. BLEU score goes from 0 to 1 where a score between 0.6 to 0.7 is believed to be the greatest attainable outcome while at the same time, a score between 0.3 to 0.4 is considered an obviously decent translation and a score more than 0.4 is called high-quality translation(Google, 2022). In this work, BLEU score is generated on the complete test data at once using the NLTK bleu library.

Figure 3 shows the qualitative outcomes of the Transformer-based model's inference, as predicted directly. The picture does not exist in the prior training dataset, where specific images and words from our Transformer-based model create grammatically valid descriptions. On the other hand, the first two BLEU ratings are not as very good, but may nonetheless serve as a reference for picture captioning. It is also observed that the final two BLEU scores are
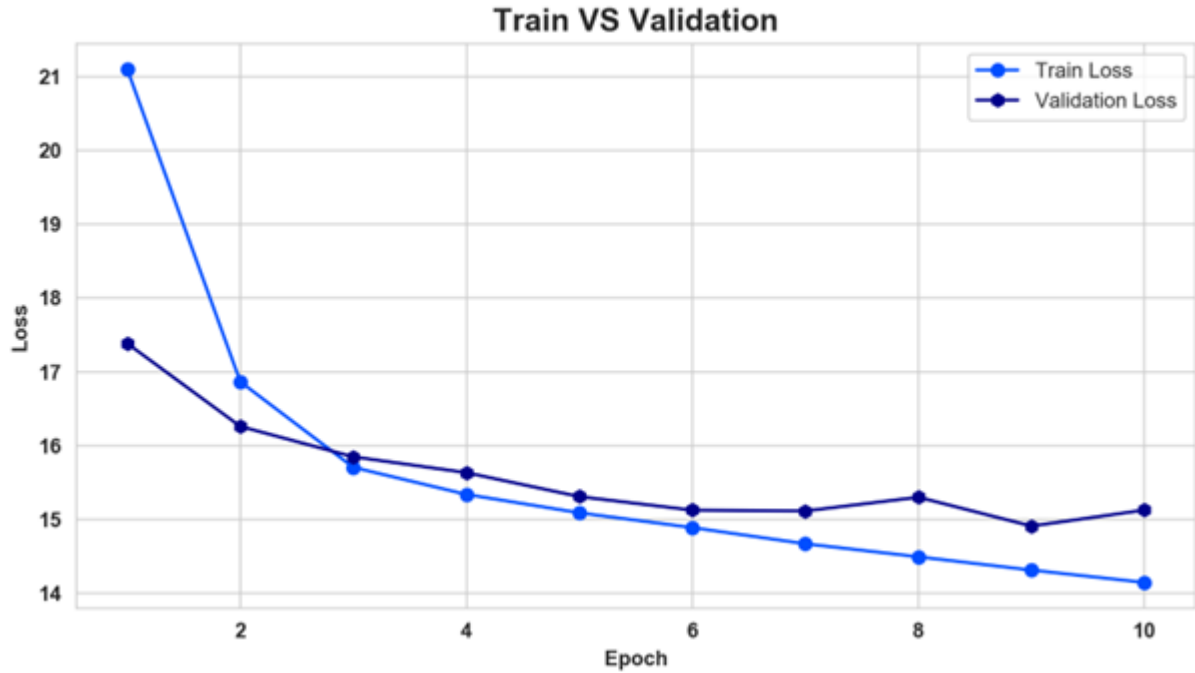
Figure 4: The difference between Training loss and validation loss

greater than the first two values. As show figure 4, validation data is used during model development to guide training and make decisions about the model. On the other hand, the accuracy of the model increasing with each epoich training.

7 Conclusion

This study focuses on the utilization of a Transformer model for image captioning tasks on a small dataset, offering a comprehensive resource for understanding and building Transformer models. The Transformer architecture, known for its efficiency in handling sequential data, particularly in NLP tasks, is explored. Key components such as input embedding, positional encoding, multi-head attention, and masking are explained in detail.

The methodology involves data preprocessing using the Flickr30k dataset, including caption normalization and image processing func-

tions. The model architecture consists of a pre-trained Efficient-NetB0 CNN for image feature extraction, followed by an encoder-decoder setup incorporating multi-head attention and layer normalization. Training involves updating weights to minimize loss, with the model evaluated using sparse categorical cross-entropy loss during validation and the Greedy algorithm combined with BLEU score for caption generation during inference.

Overall, this research contributes to the understanding and implementation of Transformer-based models for image captioning, offering insights into the architecture, training process, and evaluation techniques. The approach presented in this paper lays the foundation for further advancements in image captioning tasks using Transformer models on small datasets.

References

[1] Huda A. Al-muzaini, Tasniem Nasser Al-Yahya, and Hafida Benhidour. Automatic arabic image captioning using rnn-lstm-based language model and cnn. *International Journal of Advanced Computer Science and Applications*, 9, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2017.

[3] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3), oct 2023.

[6] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6), feb 2019.

[7] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1549–1557, New York, NY, USA, 2017. Association for Computing Machinery.

[8] Maofu Liu, Lingjun Li, Huijun Hu, Weili Guan, and Jing Tian. Image caption generation with dual attention mechanism. *Information Processing Management*, 57(2):102178, 2020.

[9] Shital Kumar Mishra and Han Wang. Computational analysis predicts hundreds of coding lncrnas in zebrafish. *Biology*, 10(5), 2021.

[10] K. C. Nithya and V. Vinod Kumar. A review on automatic image captioning techniques. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 0432–0437, 2020.

[11] Oscar Ondeng, Heywood Ouma, and Peter Akuon. A review of transformer-based approaches for image captioning. *Applied Sciences*, 13(19), 2023.

[12] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018.

[13] R. Subash, R. Jebakumar, Yash Kamdar, and Nishit Bhatt. Retracted: Automatic image captioning using convolution neural networks and lstm. *Journal of Physics: Conference Series*, 1362(1):012096, nov 2019.

[14] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[15] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[16] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015.

[17] Cheng Wang, Haojin Yang, and Christoph Meinel. Image captioning with deep bidirectional lstms and multi-task learning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 14(2s), apr 2018.

[18] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[19] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset, 2017.

[20] Wei Zhang, Wenbo Nie, Xinle Li, and Yao Yu. Image caption generation with adaptive transformer. In *2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 521–526, 2019.

[21] Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. Captioning transformer with stacked attention modules. *Applied Sciences*, 8:739, 2018.