Vietnamese caption generation for images

Final Year Project Report
4th Year Student Names

Dinh Ba Khanh Trung
Nguyen Manh Tien

A thesis submitted in part fulfilment of the degree of BSc. in Computer Science with the supervision of
Le Dinh Huynh

Bachelor of Computer Science
Hoa Lac campus - FPT University
15 December 2021

ACKNOWLEDGEMENT

Firstly, we would like to express my deep and sincere gratitude to my research supervisor, Le Dinh Huynh, for his patience and time and instructing and advising us enthusiastically. Secondly, we would also like to thank my honorable teachers of the Department of Computer Science at the Hoa Lac Campus – FPT University who taught me during the course and giving us an excellent environment to study and grow over the years. Finally, we always remember our family's encouragement and support. We want to provide them with a special thanks because they motivate us to improve ourselves every day.

PROJECT SPECIFICATION

Dinh Ba Khanh Trung

- Research on methods of image captioning model and dataset rule
- Implement model on local machine
- Write code to apply rule on COCO-VN dataset and annotate data for Flickr900
- Make a desktop app to assist annotation process
- Write documentation (Related work, Methodology, Experimental, Conclusion)

Nguyen Manh Tien

- Research on methods of image captioning model and dataset rule
- Apply rule and annotate data for COCO-VN and Flickr900
- Write documentation (Introduction, Data, References)
- Make presentation slide

ABSTRACT

Automatic caption generation for images has attracted a great deal of attention from many machine learning researchers in recent years. However, a lot of work on this aspect is solved for English. This paper contributes to research on the Image Captioning task in terms of extending the existing dataset for Vietnamese descriptions for images and comparing different model approaches on Vietnamese caption generation. Since most of the available image captioning has been created for the English language and in other most spoken languages such as Chinese, there are very few dataset for Vietnamese. In this specific case, we create a dataset consisting of 4500 captions for 900 images that enlarge the current Vietnamese captions dataset and the translated version of preprocessed English captions from the train dataset of MS-COCO. We evaluated our extended dataset on a neural network-based image caption generation model, then compare it with the Vietnamese image captioning dataset UIT-ViIC, and we made an enhanced Vietnamese caption model based on the most famous image captioning model to improve the accuracy metrics.

Table of Contents

List Of Figures

List Of Tables

Chapter 1

INTRODUCTION

1.1 Introduction

Automated generation of multimedia content, for instance, videos and images are called image caption generation which is a rising research field that combines two fields of machine learning: computer vision and natural language processing. A description of an image must capture not only the objects visible in a picture but also the relationship between this object along with their attributes and the activities they are involved in. By improving image description quality, this will transcribe the surrounding scenes and output the caption into a text to speech model to support people with visual impairments. In the commerce field, the image captioning model can be used to automatically generate the description to understand and describe product images on their websites. Image captioning models can also be integrated to classify videos and images based on different scenarios therefore optimize the search quality for image based search engines. Although, a lot of work has been done in this field recently, and there has been promising efforts to overcome linguistic barriers by extending dataset captions into different languages base on their specific task such as YJ Captions [1] for Japanese, Multi30k [2] for German and for most spoken language like Chinese but there are a few public dataset or research for Vietnamese. It motivates us to work on our problem of creating Vietnamese captions to overcome the language barriers. Machine learning is a data driven method. Owing to the short amount of time, we choose to construct a captioning dataset based on Flickr30k [16]. The captions to Vietnamese were translated using a variety of methods, which are discussed in the subsections below.

1.2 Objective and Contribution

Firstly, we created Flickr900 to extend existing Vietnameses captioning dataset UIT-ViIC [4] which contain sport-ball images to balance this dataset.

Secondly, we built a full Vietnamese version of training dataset from the MS-COCO [38] dataset for Vietnamese caption.

Thirdly, we make a simple annotation tool for dataset construction to assist annotator to create caption efficiently.

Finally, we improve the model performance by combining the previous works with newly proposed techniques.

1.3 Related Work

Dataset	Release	Data source	Languages	Images	Sentences	Application
IAPR TC-12 [5]	2006	Internet	English/German	20,000	100,000	Image retrieval
Pascal sentences [6]	2015	Pascal sentences	Japanese/English	1,000	5,000	Cross-lingual document retrieval
YJ Captions [7]	2016	MS-COCO	Japanese/English	26,500	131,470	Image Captioning
MIC test data [8]	2016	MS-COCO	French/German/ English	1,000	5,000	Image retrieval
Bilingual caption [9]	2016	MS-COCO	German/English	1,000	1,000	Machine translation - Image Captioning
Multi30k [2]	2016	Flickr30k	German/English	21,014	186,084	Machine translation - Image Captioning
Flickr 8k-CN [10]	2016	Flickr 8k	Chinese/English	8,000	45,000	Image Captioning
AIC-ICC [11]	2017	Internet	Chinese	240,000	1,200,000	Image Captioning
Flickr30k-CN [12]	2017	Flickr30k	Chinese/English	1,000	5,000	Image Captioning
STAIR Captions [13]	2017	MS-COCO	Japanese/English	164,062	820,310	Image Captioning
COCO-CN [14]	2018	MS-COCO	Chinese/English	20,342	27,128	Image tagging - Image captioning - Image retrieval
WikiCaps [15]	2018	Wikimedia Commons	German/French/ Russian/English	3,816,940	3,825,132	Multimodal machine translation - Image retrieval - Image captioning
UIT-ViIC [4]	2020	MS-COCO	Vietnamese/English	3,850	19,250	Image Captioning
COCO-VN (this paper)	2021	MS-COCO	Vietnamese/English	118.344	591.720	Image Captioning
Flickr900 (this paper)	2021	Flickr30k	Vietnamese/English	900	4500	Image Captioning

Table 1.1: Non-English public image datasets with manually annotated

Table 1 provides a partial list of published Image Captioning datasets manually annotated and in different languages. Several image caption databases have been created in English, and the most famous examples are Flickr8k, Microsoft COCO(Microsoft Common in Objects in Context), and the enhanced version of Flickr8k - Flickr30k [16]. Along with these primary datasets, many other non-English caption datasets have been developed. Depending on their applications, the target languages of these datasets vary, including German and French for image retrieval, Japanese for cross-lingual document retrieval and image captioning, Chinese for image tagging, captioning, and retrieval. Each of these datasets is based on an existing English dataset, the most prominent of which is MS-COCO. There are two dataset AIC-ICC [11] and WikiCaps [15] that use data from the internet instead of the popular dataset from MS-COCO and Flickr. According to my knowledge, UIT-ViIC is the first image captioning dataset in Vietnamese, adopting Microsoft COCO as its data source. Each image has been reannotated with five Vietnamese sentences written by native speakers via the annotator team in Vietnam National University, Ho Chi Minh city. They write the sentences with the pre-set rules, so the quality is controlled, but their dataset seems unbalanced since there is too much image about a single subject such as image about tennis and baseball.

Flickr900 dataset is constructed using 900 images and 4500 hand-written sentences from the Flickr30k dataset. Flickr30k is a dataset that includes more than 30,000 images and has five captions for each image and consists of 158,915 crowd-sourced captions in total. After MS-COCO, Flickr30k is the most well-known dataset for Image Captioning. We also created a large-scale Vietnamese caption dataset, COCO-VN that consists of Vietnamese captions for 118.344 images and 591.720 captions in the train dataset of MS-COCO which is the most famous dataset for Image Captioning.

Many research organisations have worked on picture captioning since 2010 and have claimed considerable improvements. The initial work was by Ali Farhadi et al.[17]. Their method evaluated the similarity between an image and a sentence by mapping each to the meaning space and comparing the result. Their model will learn the mapping from images to its meaning from images and assigned the meaning presentation. Kulkarni et al.[18] construct a conditional random field (CRF) to predict the label for an image, this architecture uses three unary potentials trained objects, attribute classification scores for an object and prepositional relationship score and high-order potentials from text

corpora, using an n-gram model for decoding and templates for constraints. Visually similar photos and captions are initially obtained from a huge database, followed by the generation of captions for the queried image. This approach generates valid and generalizable captions but not semantically correct captions.

To overcome this limitation, image captioning using deep neural networks (DNNs) methods was first introduced by Kiros [19]. The method using the log-bilinear language model (LBL) that use a convolutional neural network (CNN) to extract the feature from the image then feed to the neural language model which uses multimodal space to map image extracted feature with text feature then predict the word base on image feature and previously generated word. Later that year, Kiros et al.[20] also released a new method that extends their earlier work where the popular encoder-decoder model architecture was first used in image captioning. This model uses a convolutional neural network (CNN) to encode image features and Long Short Term Memory (LSTM) for textual data, and a neural language model to decode visual elements conditioned on text feature vectors. Mao et al. [21] structured a multimodal Recurrent Neural Network (m-RNN) for image caption generation. This model makes use of a deep neural network for images and deep RNN for captions. The sentence's description corresponding to the image will be put through two embedding layers, recurrent layer, multimodal layer and a softmax layer, and the image is put in a CNN. Then sentences and images are given as input to a multimodal layer, the model computes the probability distribution of the following word given previous words and the image. Karpathy et al.[22] presented a multimodal embedding approach combining visual and linguistic modalities for generating image descriptions.This model used R-CNN as an image encoder and then performed local similarity learning between image regions and sentence words by combining the similarity scores of all region-word pairings. This approach is finer level and embeds fragments of images and fragments of sentences into a common space. Detected objects produced by R-CNN are mapped to fragment embedding space and the sentences are embedded to dependency tree relations, and the inner product between them is called as a similarity score then be computed as a fixed function of their pairwise fragment scores.

Vinyals et al.[23] created a model called Neural Image Caption Generator also known as NIC. This method uses a convolutional neural network as a encoder to extract

features from images followed by a language generating RNN for caption generation. The output of the encoder's last hidden layer is utilised as the input to the LSTM. The model is trained by using maximum likelihood estimation (MLE) techniques and generates picture descriptions via joint embedding. Xu et al[24]. introduced an attention mechanism base image caption generation for the first time. This method uses the output of convolutional layers of the convolutional neural network rather than last hidden fully connected layers like NIC so it can concentrate on salient objects from images. Attention based methods can focus on different parts of the image and generate the corresponding words to that region. Then a Long Short Term Memory is used as a decoder to generate sentences.

Chapter 2

DATA

2.1 The COCO-VN

In this section, we will go over how we constructed the COCO-VN dataset. This dataset contained 118.344 images from the train dataset of MS-COCO 2017 dataset-one of the largest dataset for image captioning. Unlike other authors in [4][13][14], we try to implement a different approach by preprocessing the english dataset before using the Google Translate API to translate the captions. Influenced by STAIR Captions [13], UIT-ViIC[4] and other manual annotated image captioning dataset, we create the following rules to apply to the preprocessing stage of the english dataset:

1. Remove passive voice from the sentences using keywords: “that reads", “that says",”telling”, ..
2. Specific brand of items or company (Fedex, Mercedes-benz, . . .) is removed from captions
3. Name of people, places, street, national, breed(dalmatian, beagle, ...) and other ambiguous things that are not visible in the image are eliminated.
4. Rewrite sentences that are not in simple form (question, exclamation sentence, . . .)
5. Remove detailed information such as number on clock, number on licence plate, text on street sign or board, . . .
6. Fixing miss spelling word to make sure that Google translate can understand

7. Change all words in upper case to lower case to optimize for Google translate because Google translate will not translate the upper case word.

After preprocessing the english version and putting it into Google translate, we walk through all the images and fix sentences that have weird meanings or are not grammatically correct in Vietnamese.

Figure 2.1: Example from COCO-VN dataset

English: A cat stares at a chocolate topped donut, with the caption reading, "donut want."

Unpreprocessed: Một con mèo nhìn chằm chằm vào chiếc bánh donut phủ sô cô la với chú thích đọc là "muốn có bánh rán".

Preprocessed: Một con mèo nhìn chằm chằm vào chiếc bánh rán phủ sô cô la.

2.2 The Flickr900

In this section, we will go over how we constructed the Flickr900 dataset. This dataset contained 900 images of sports played with balls from the 30.000 images version of the Flickr dataset (Flickr30k). These images were chosen by extracting the image's object detection label in the Flickr30k annotation file. We will search for the keywords related to sports played with balls such as "soccer", "football", "volleyball". Following the rules of the published dataset created on Microsoft COCO and Flickr, we added some rules to be more suitable for the Vietnamese language. Each of the images in the dataset will consist of five handwritten captions, so the total number of captions was 4500.

2.2.1 Annotation tool

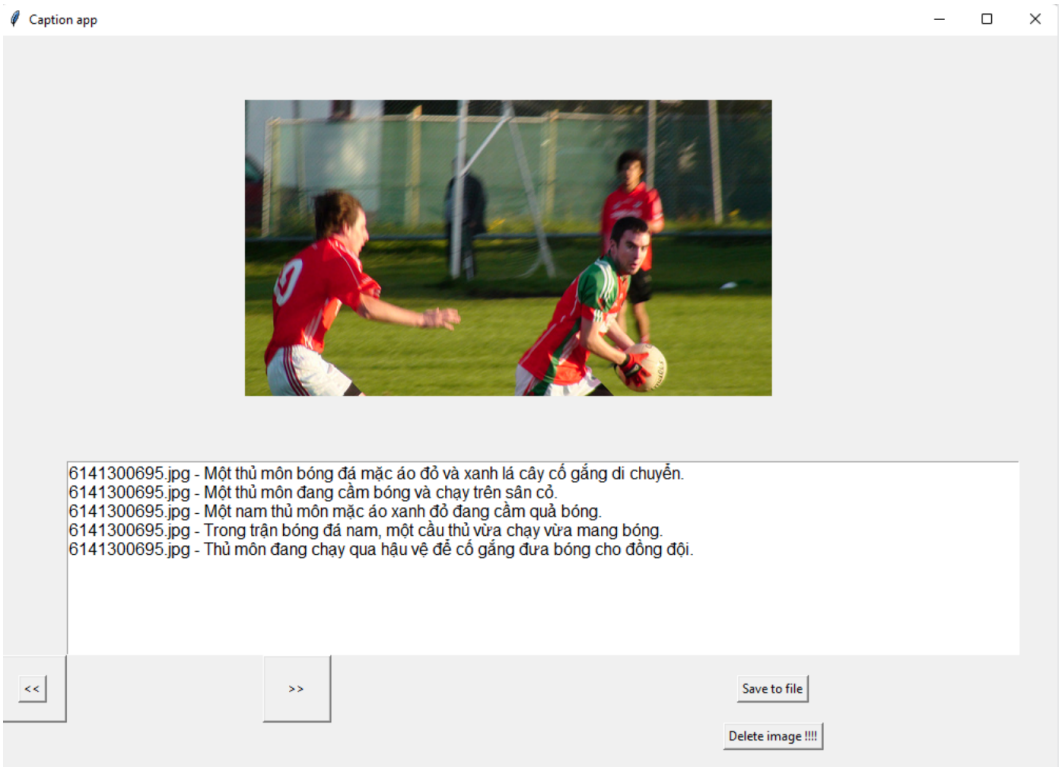

Figure 2.2: Example of user interface of caption annotation application

To assist the annotator in writing image captions efficiently, we first develop a simple desktop app for caption annotations. Figure 2.3 shows the example of the annotation screen in the application. Our applications assist the caption editor by loading images

and image captions into the user interface. With the saving function, the annotator can save and load written sentences to the dataset. The delete button will be used to delete the image that does not relate to our topic. Image captions given in the textbox are preprocessed by correcting spelling mistakes, removing some specific patterns, and translated to Vietnamese by Google translate. Original English sentences are also displayed to help the annotator if needed. Those content suggestions are helpful for the image that does not have a precise meaning or captions are obscure.

2.2.2 Annotation procedure

In this section, we describe how the data for our Flickr900 dataset is gathered and generated.

Inspired by MS-COCO annotation rules, Flickr, and another captioning dataset, we constructed our dataset by using the following guidelines:

1. Each image caption must contain at least eight words.
2. Describe all the essential parts of the scene, visible activities, and objects
3. Ignore all specific details like names of places, streets, manufacturers (New York, Mercedes-Benz, etc.), and number (times on the clock, exact time on TV, etc.)
4. Each caption must be a single statement that does not depict events that may or may not have occurred in the past or future.
5. When annotating, personal opinion and emotion must be avoided.
6. Remove all unclear items and describe visible objects.
7. While annotating, personal opinion and emotion must be eliminated

2.2.3 Dataset analysis

After finishing constructing the Flickr900 dataset, we have a look at statistical analysis of our dataset. Flick900 was made up of 900 images described by 4500 Vietnamese captions.

	UIT-ViIC	UIT-ViIC + Flick900
tennis	1658	1666
baseball	1389	1510
football	558	876
volleyball	119	223
American football (rugby)	22	28

Table 2.1: Statistics on sport categories in both dataset

Table 2.1 summarizes the most occuring sport categories in in UIT-ViIC and the combination of UIT-ViIC and Flickr900. The Flickr900 add more image about baseball, football, and volleyball to get better caption generation at more sport categories compare to the original UIT-ViIC. Since most of the image is about tennis, baseball, football and volleyball so we expect the model to generate best result for these sports.

Figure 2.3: Example from Flickr900 dataset

English caption: Two volleyball players standing next to a net that is part of an indoor court , celebrating a win or a point scored , with several people looking on .

Translated google translate: Hai cầu thủ bóng chuyền đứng cạnh lưới là một phần của sân trong nhà, ăn mừng một chiến thắng hoặc một điểm ghi được, với một số người đang nhìn.

Flickr900 manual annotated: Một nhóm cầu thủ bóng chuyền đang thi đấu trên sân trước đông đảo khán giả.

Chapter 3

METHODOLOGY

3.1 Network Architecture Overview

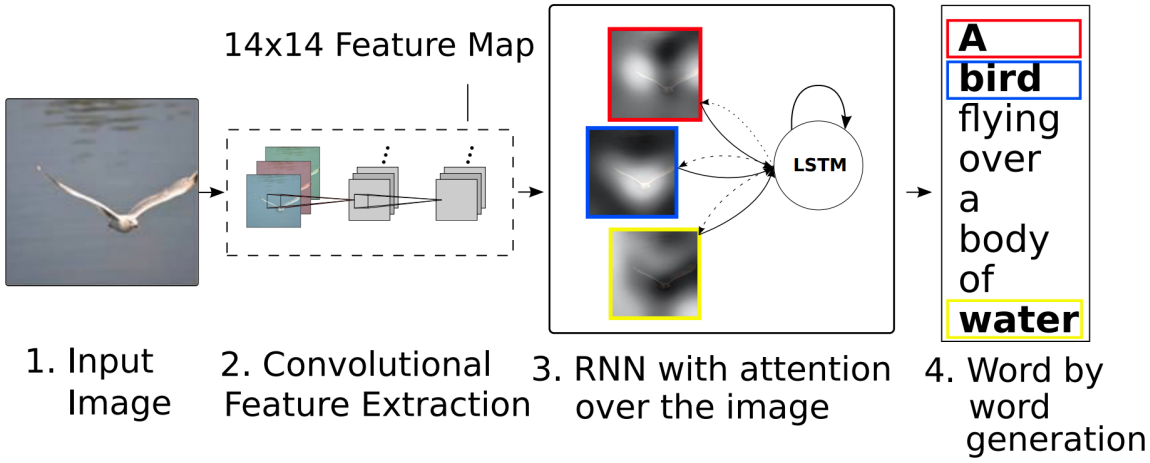

Figure 3.1: General encoder - decoder image captioning architecture with attention [24]

This image captioning model contain an encoder for image and a decoder to generate caption

1. The image is put into the encoder to extract important features from it. Typically a ImageNet pretrained convolutional neural network is used for this part of model.

2. Those extracted features output from the encoder is taken as the input of the decoder and it will be used to generate the caption.

3. A language model Recurrent Neural Networks (RNNs) is used as the decoder that takes in image features with current word as input then outputs the next word in the caption.

4. The attention mechanism used in the decoder of image captioning models so the model can determine which part of the image is used to generate the next word.

5. The words are generated word by word in sequential to complete the caption.

3.2 Model Detail

Our model is inspired by the method of Xu et al. [24] We have some minor modifications to improve their performance on the Vietnamese captioning dataset.

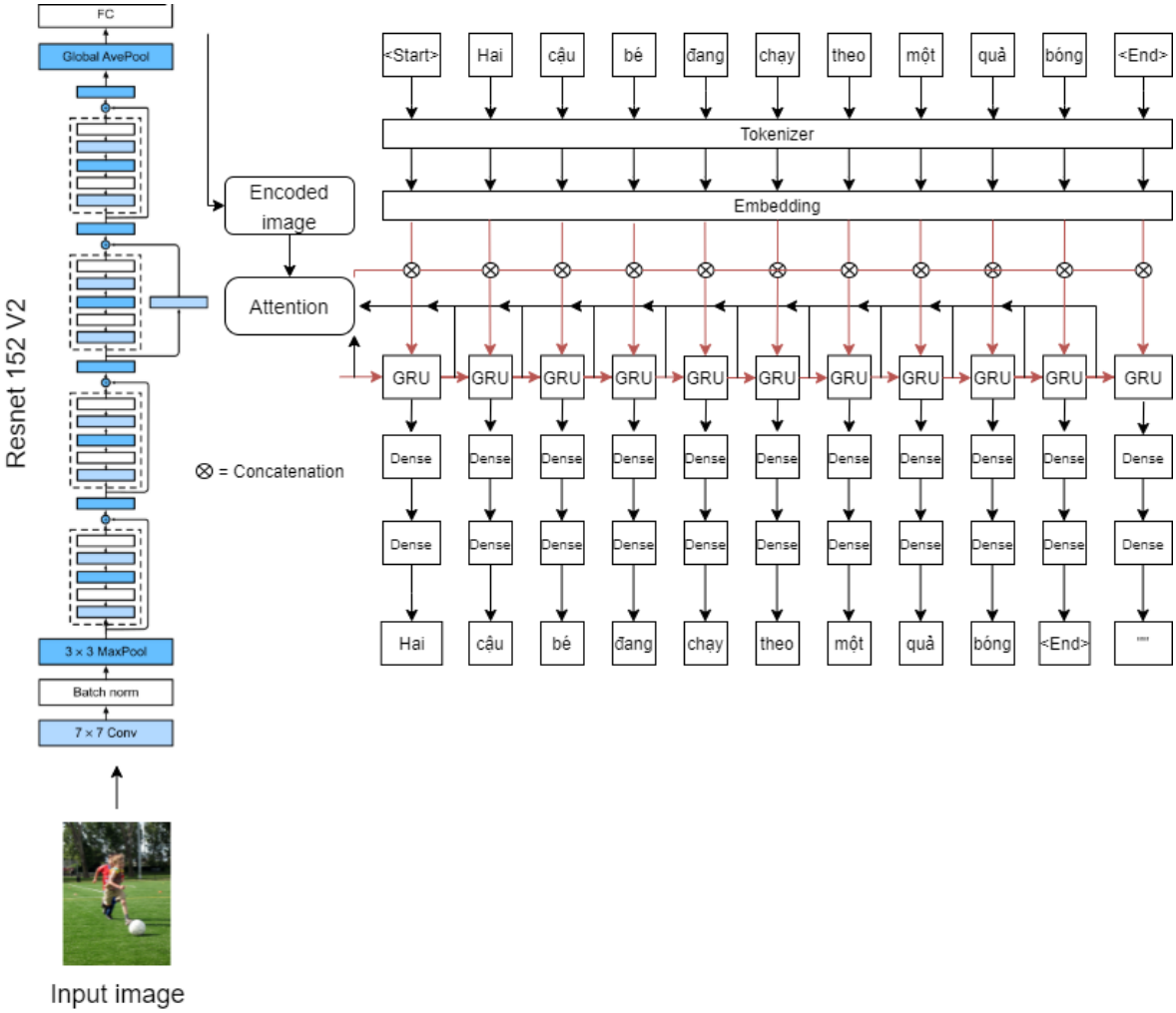

Figure 3.2: Image captioning model in Vietnamese

3.2.1 Tokenizer

Tokenization is a sort of segmentation used as the initial stage in any language processing task. Tokenizer breaks down a phrase, sentence, paragraph, or even an en-

tire text document into small fragments, such as words or terms. Each of these small fragments is called a token. Like many languages in Southeast Asia and East Asia, Vietnamese is an analytic and isolating language [25]. The smallest basic linguistic, meaningful unit in Vietnamese is morpheme similar to syllables or tokens in English. A Vietnamese word can consist of one, two, or even three tokens; therefore, to apply Vietnamese Image Captioning, we choose two different Vietnamese tokenizer tools: PyVI [35] and Coccoc-Tokenizer [36], both specialized for tokenization at the word level in the Vietnamese language. However, when we experiment with Coccoc-Tokenizer, it pretty slow when compare to PyVI to tokenize the dataset (it takes 1.2s per sentence compare to 0,1s with PyVI), and the way it breaks a sentence into words is the same as PyVI in most cases, so in this paper we only use the PyVI tool for model and benchmarking.

Tokenizer	**Compile time**	**Output tokenized sentence**
PyVI	0.1s	Một trận thi_đấu bóng_đá đang diễn ra trên sân
Coccoc-Tokenizer	1.2s	Một trận thi_đấu bóng_đá đang diễn ra trên sân
Nltk	0.1s	Một trận thi đấu bóng đá đang diễn ra trên sân

Table 3.1: Tokenized result from sentence: “Một trận thi đấu bóng đá đang diễn ra trên sân"

3.2.2 Encoder and attention mechanism

Our encoder takes an unprocessed image and puts it through a convolutional neural network in order to obtain a set of feature vectors from the image. Our model extracts features from lower convolutional layers instead of the fully connected layers. The advantage of using the output of the last convolutional layer is that the model's decoder can focus on characteristics that would otherwise be missed if the output of the fully connected layers was used. Then the production of the convolutional neural network will be put through an attention mechanism and GRU decoder. In this encoder module, we choose the latest CNN architectures Resnet-152v2 [26], InceptionV3 [27] and EfficientnetB7 [28] to extract more valuable features from the image.

We employ an approach by Bahdanau at el. [29] for the attention mechanism. So the model can learn to focus on essential regions of the image. This soft attention method corresponds to feeding in a soft weight context into the model and pays equal

attention to all regions of the picture. Image features extracted by convolutional neural networks and previous hidden state are passed through the attention mechanism. Then, at each step, we compute the alignment score of each encoder output with regard to the decoder input and hidden state.This score measures how much attention the decoder will spend on each of the encoder outputs while creating the next output and which area in the image should the decoder focus. The encoder outputs and decoder hidden state will be fed via their Linear layer, each with separate trainable weights then it will be added together and go through tanh activation function.

$$score_{alignment} = W_{combine}.tanh(W_{decoder}.H_{decoder} + W_{encoder}.H_{encoder})$$

Then this score is passed through a softmax function to get the weighted attention score between 0 and 1

$$AttentionWeights = softmax(alignmentscore)$$

After calculating the attention weights, the context vector is computed by performing an element-wise multiplication of the attention weights with the encoder outputs. Because of the softmax function, if the score of a certain input element is close to 1, its effect and influence on the decoder output is amplified, whereas if the score is close to 0, its influence is drowned out and eliminated.

$$ContextVector = attentionweights.extractedfeatures$$

After that the context vector is concatenated with the decoder's current input. The output of this equation will be passed to GRU to generate text words, this process is repeated until the whole caption is generated.

3.2.3 Decoder

The decoder part uses a language model Recurrent Neural Networks (RNNs) to generate word by word in sequential order to complete a caption.

Recurrent Neural Networks (RNN):

Recurrent neural networks (RNN) [39] are built to handle sequential data such as text and audio. It handles sequences by iterating through the sequence elements and keeping a state that contains information about what it has seen so far. A unit of

RNN will take input from the previous step with the current state and be incorporated with Tanh as an activation function, then the output is the new hidden state. This methodology faces short-term memory problems because of gradient vanishing or exploding problems. Recurrent neural networks have to backpropagate gradients over a long sequence, gradient value will decrease layer by layer and eventually vanish after some steps or become a very massive value in matrix multiplication. Therefore Gated Recurrent Unit and Long Short Term Memory with memory cell were created to cope with this weakness of RNNs.

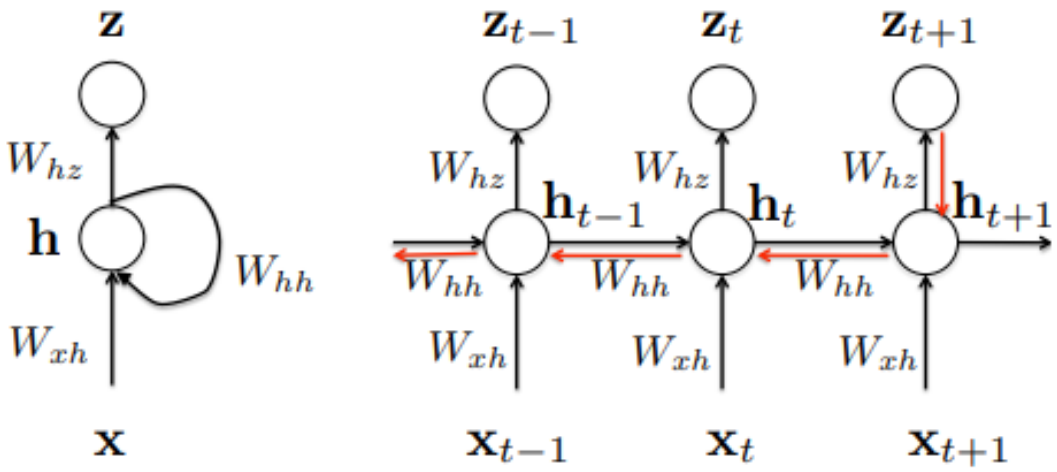

Figure 3.3: It is a RNN example: the left recursive description for RNNs, and the right is the corresponding extended RNN model in a time sequential manner. [36]

Long short-term memory (LSTM):

Forget state: This gate will decide how much data from the previous state should be preserved and how much should be forgotten by multiplying the incoming long-term memory by a forget vector formed by the current input and short-term memory.

Input state: This gate will take the output from the previous hidden state and the current input into a sigmoid function. The sigmoid outcome will determine which information from the tanh output should be kept.

Output gate: This gate will determine what the next hidden state should be. It will take the previous hidden state and the current input into a sigmoid function then pass the new cell state to the tanh function. After that, we combine the sigmoid output with the tanh output to get the hidden state. The new hidden state and new cell state are used in the next time step.

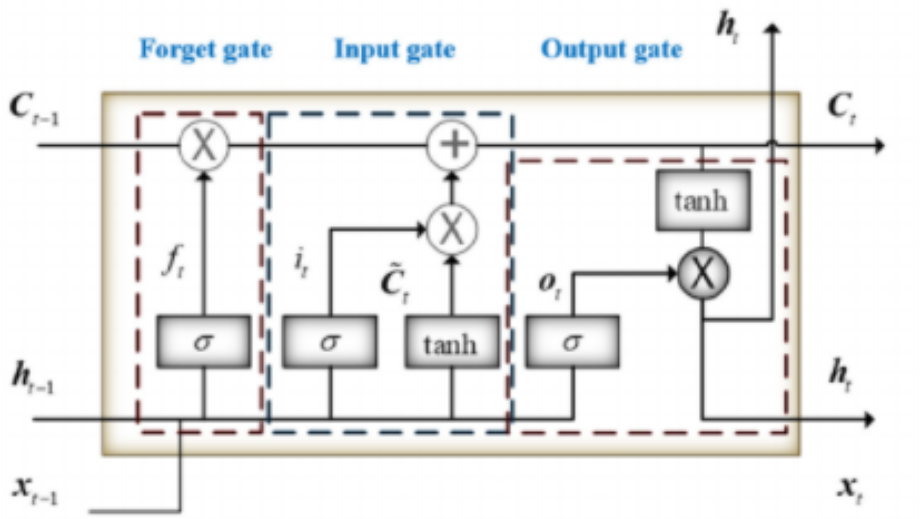

Figure 3.4: Basic structure of LSTM unit [37]

Gate recurrent unit (GRU):

Update gate: The update gate determine which information will be pass through the next state by calculating another representation of the input vector X and the previous hidden state

Reset state: This state is similar to the LSTM forget state, it decides what will be removed from the previous hidden time steps. This gate is calculated by multiplying the input vector and hidden state by their weights, followed by an element-wise combination of the reset gate and previously hidden state. In the end, it uses a non-linear activation function to get the result that lies between 0 and 1.

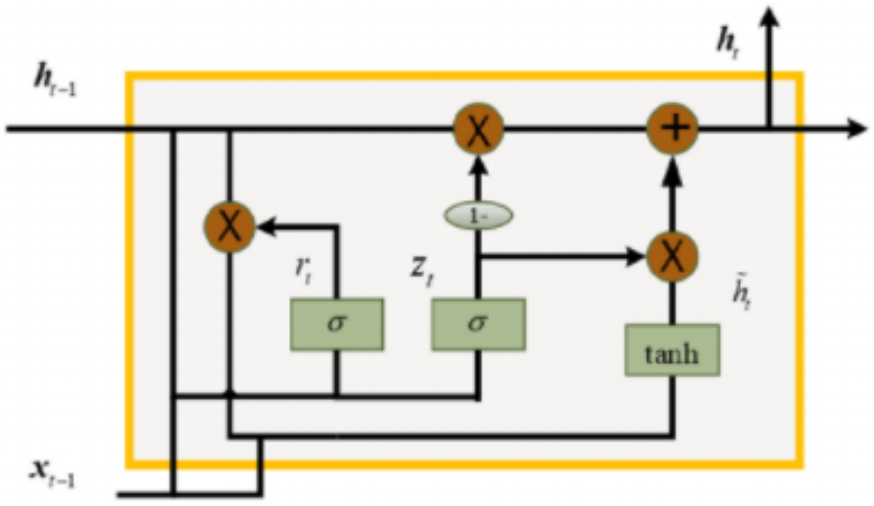

Figure 3.5: Basic structure of GRU unit [37]

Our decoder uses GRU to generate text for image captions. GRU is a similar algorithm to LSTM, except it contains fewer parameters. GRU lacks an output gate as well. These characteristics make GRU faster, less costly to compute, and more memory efficient. GRU generates one word at each time step. This created word is dependent on the prior hidden state of GRU, the previously generated word, and the context vector. Chung et al. [30] shown that the GRU outperforms the LSTMs on small datasets. As a result, GRU is the best option for our situation.

Chapter 4

EXPERIMENTAL

4.1 Experimental Setup

All the experiments are performed on my local machine and Tensorflow, using the hardware of GPU NVIDIA RTX 3060. We use the InceptionV3 , Resnet152-V2 , EfficientnetB7 pre-trained on Imagenet and set the input image size of the image feature extraction model to (299,299). We use two dataset UIT-ViIC and the combination of UIT-ViIC and Flickr900 to benchmark with the models. The batch size is set to 64 and use Adam optimizer with learning rate is 0.001 and SparseCategoricalCrossentropy loss. We train the model 60 epochs with the training time is about 1 hour .

4.2 Evaluation Metrics

To evaluate our data and model, we use metrics proposed by most papers in related works of image captioning model and dataset: BLEU [31], ROUGE [32], and CIDEr [33].

BLEU is a widely used metric for automation evaluation of machine translation that measures the n-grams precision of machine-generated sentences in comparison to human-generated sentences. BLEU seems like a suitable choice for our task at first glance however because it does not consider meaning and sentence structure which can make BLEU penalize many correctly generated sentences that cause a low correlation with human judgement of quality.

ROUGE was developed originally for evaluating text summarization and machine translation, whereas CIDEr was designed and used mainly for evaluating Image Cap-

tioning tasks by the organizers of the MS COCO Captioning challenge. It computes the consistency of n-gram occurrences in generated and reference texts, which is weighted by n-gram saliency and rarity.

4.3 Experimental Result

Table 5.1 are the model result experiment with different model's encoder and two dataset; UIT-ViIC and the combination of UIT-ViIC and Flickr900.

Encoder	**Attention**	**Dataset**	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Cider
InceptionV3	Yes	UIT-ViIC	0,824	0,702	0,596	0,522	0,657	0,622
InceptionV3	Yes	UIT-ViIC + Flickr900	0,776	0,663	0,566	0,498	0,668	0,641
Resnet152-V2	Yes	UIT-ViIC	0,782	0,632	0,508	0,414	0,651	0,584
Resnet152-V2	Yes	UIT-ViIC + Flickr900	0,781	0,659	0,553	0,484	0,677	0,638
Efficientnet B7	Yes	UIT-ViIC	0,829	0,719	0,619	0,550	0,679	0,770
Efficientnet B7	Yes	UIT-ViIC + Flickr900	**0,834**	**0,727**	**0,634**	**0,569**	**0,681**	**0,852**
Resnet152-V2	No	UIT-ViIC	0,777	0,597	0,483	0,396	0,626	0,369

Table 4.1: Experimental result with different encoder and dataset

As can be seen in table 5.1, with InceptionV3 as image feature extraction, UIT-ViIC dataset yields better results than the combination of UIT-ViIC and Flickr900 in all type of BLEU scores, however when compare in Rouge-L and specific score for the CIDEr that was designed especially for image captioning evaluation, our combined dataset start to beat UIT-ViIC. When changing the encoder to Resnet152-V2, the result drops a little when placed side by side with InceptionV3. With Resnet152-V2 as the encoder, the result show that the BLEU-1 score of UIT-ViIC is the same as our the combination of UIT-ViIC with Flickr900 but as the number of consecutive words considers(BLEU gram) increase, the BLEU scores of our combined dataset started to exceed UIT-ViIC and the Rouge-L and Cider scores for our combined dataset prove the same thing. After changing to one of the state of the art models trained on Imagenet - Efficientnet B7 for image feature extraction of the model's encoder, all of the benchmarks yield superior scores to the two previous models. From BLEU-1 to BLEU-4, our combined dataset gives a better score than the original UIT-ViIC, the Rouge-L and Cider also proved the same things. We also try to remove the attention mechanism from the model with Resnet152V2 and UIT-ViIC to observe if attention plays an essential role in our

architecture. The BLEU-1 of the model without attention decreases a little from the model with attention; however when increasing the BLEU score in range of 2 and 4, the gap between two models keeps growing. The Cider scores for models without attention decrease dramatically from 0,584 of the attention model to 0,369. And we also notice that the model without attention generates captions not as diverse as the model with attention.

Figure 4.1: Demo image 1

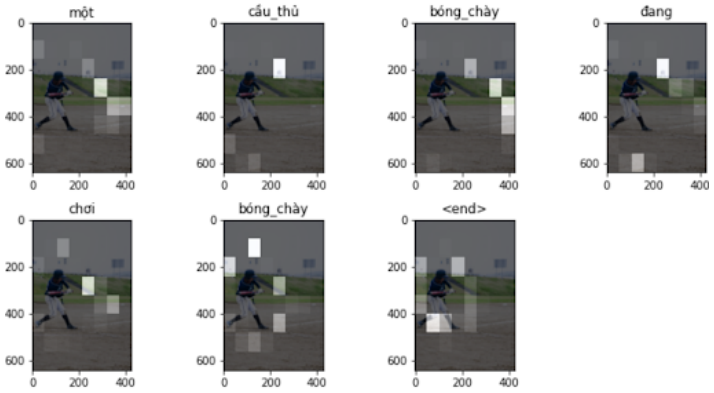

Figure 4.2: Demo attention 1

InceptionV3(UIT-ViIC): một cầu thủ đánh bóng đang vung gậy bóng chày trên sân

InceptionV3(UIT-ViIC + Flickr900): cầu thủ bóng chày đang vung gậy bóng chày đánh trả bóng

EfficientnetB7(UIT-ViIC): cầu thủ bóng chày đang cầm gậy thi đấu ở trong tay

EfficientnetB7(UIT-ViIC + Flickr900): cầu thủ bóng chày đang cầm gậy bóng chày chuẩn bị để đánh bóng

Figure 4.3: Demo image 2

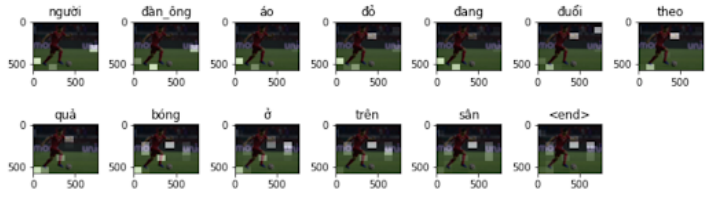

Figure 4.4: Demo attention 2

InceptionV3(UIT-ViIC): một cầu thủ bóng đá đang chuẩn bị sút bóng

InceptionV3(UIT-ViIC + Flickr900): một cầu thủ bóng đá đang chuẩn bị sút bóng

EfficientnetB7(UIT-ViIC): một trận thi đấu bóng đá ở trên sân

EfficientnetB7(UIT-ViIC + Flickr900): người đàn ông áo đỏ đang đuổi theo quả bóng ở trên sân

Figure 4.5: Demo image 3

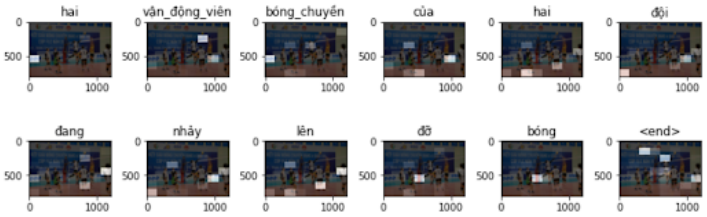

Figure 4.6: Demo attention 3

InceptionV3(UIT-ViIC): cầu thủ bóng rổ đang nhảy lên đánh bóng

InceptionV3(UIT-ViIC + Flickr900): một vận động viên bóng chuyền đang thi đấu trên sân

EfficientnetB7(UIT-ViIC): các cầu thủ tennis đang thi đấu ở trên sân

EfficientnetB7(UIT-ViIC + Flickr900): hai vận động viên bóng chuyền của hai đội đang nhảy lên đỡ bóng

Figure 4.7: Demo image 4

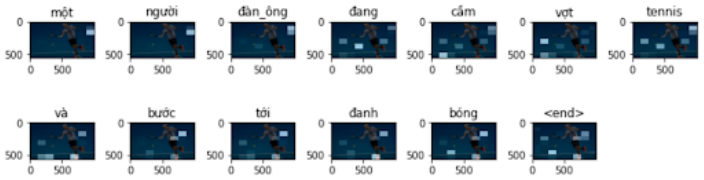

Figure 4.8: Demo attention 4

InceptionV3(UIT-ViIC): một nam vận động viên tennis đang cầm vợt thi đấu trên sân

InceptionV3(UIT-ViIC + Flickr900): vận động viên tennis nam đang thi đấu trên sân

EfficientnetB7(UIT-ViIC): một nữ vận động viên tennis đang bước dài trên sân

EfficientnetB7(UIT-ViIC + Flickr900): một người đàn ông đang cầm vợt tennis và bước tới đánh bóng

We test four images with four sport games in our dataset tennis, football, volleyball and baseball with different models. The model with EfficientnetB7 trained with UIT-ViIC combined with Flickr900 can describe images better than InceptionV3 in details such as colour of people's skirt, one or many people in the scene and people's gender. Moreover, EfficientnetB7 can tell the specific action of people more precisely (một người đàn ông đang cầm vợt tennis và bước tới đánh bóng) than model with InceptionV3. UIT-ViIC combined with Flickr900 can demonstrate more sport categories (hai vận động viên bóng chuyền của hai đội đang nhảy lên đỡ bóng); whereas UIT-ViIC is mistaken (cầu thủ bóng rổ đang nhảy lên đánh bóng). The confusion of UIT-ViIC is caused due to the size of the dataset and most images in this dataset are about tennis and baseball. However, there are case that the model and our dataset can not tell the expected background, gender or the age of the people in the picture that we need to improve in the future.

Chapter 5

Conclusion And Future Works

5.1 Conclusion

In this paper, we have prepared two dataset for Vietnamese language generation: COCO-VN consisting of all 118.344 images in the training dataset of MS-COCO with 591.720 captions translated with Google Translate with our modification rule for smoother sentences; and the Flickr900 including 900 images of sportball in Flickr-30K come along with 4500 manually-written Vietnamese captions to enrich UIT-ViIC dataset. Then we experimented the combined with our modified model to evaluate its efficiency when learning Vietnamese caption tasks.

We also tweak the most famous encoder-decoder with attention to image captioning model with a more accurate image feature extraction model for encoder, different recurrent neural network and finetune a few hyperparameters to work well in the Vietnamese language.

5.2 Future Work

For future improvement, we currently work on a encoder-decoder with attention model that uses a Transformer encoder for self-attention on visual features and a Transformer decoder for masked self-attention on caption tokens.

Second, we try to improve the encoder by finetune the pre-trained CNN with our sport dataset and replace greedy search with a beam search decoder for better caption generation performance and quality.

Third, we will try to reduce the size of the model such as smaller image feature extraction model or reduce decoder layer to help the model to be runnable on embedded devices with nvidia jetson nano or raspberry PI.

Fourth, we will try to train our model with the COCO-VN on a more powerful computer to evaluate this large dataset that helps the model to understand visuals in many real world applications.

References

[1] Miyazaki, T. and Shimizu, N. (2016). Cross-Lingual Image Caption Generation. pages.1780–1790.

[2] Elliott, D., Frank, S., Sima'an, K. and Specia, L. (2016). Multi30K: Multilingual English-German Image Descriptions. arXiv:1605.00459 [cs]. [online] Available at: https://arxiv.org/abs/1605.00459.

[3] 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, [online] 47, pp.853–899. Available at: https://www.jair.org/index.php/jair/article/view/10833

[4] Lam, Q.H., Le, Q.D., Van Nguyen, K. and Nguyen, N.L.-T. (2020). UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning. arXiv:2002.00175 [cs]. [online] Available at: https://arxiv.org/abs/2002.00175.

[5] Michael Grubinger, Paul Clough, Henning Müller, Thomas Deselaers (2006). The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems

[6] Funaki, R. and Nakayama, H. (2015). Image-Mediated Learning for Zero-Shot Cross-Lingual Document Retrieval. [online] ACLWeb. Available at: https://aclanthology.org/D15-1070/

[7] Miyazaki, T. and Shimizu, N. (2016). Cross-Lingual Image Caption Generation. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics

[8] Rajendran, J., Khapra, M.M., Chandar, S. and Ravindran, B. (2016). Bridge Correlational Neural Networks for Multilingual Multimodal Representation Learning. arXiv:1510.03519 [cs]. [online] Available at: https://arxiv.org/abs/1510.03519.

[9] Hitschler, J., Schamoni, S. and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), [online] pp.2399–2409. Available at: https://arxiv.org/abs/1601.03916.

[10] Li, X., Lan, W., Dong, J. and Liu, H. (2016). Adding Chinese Captions to Images. Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval.

[11] Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., Wang, Y. and Wang, Y. (2019). AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. 2019 IEEE International Conference on Multimedia and Expo (ICME), [online] pp.1480–1485.

[12] Lan, W., Li, X. and Dong, J. (2017). Fluency-Guided Cross-Lingual Image Captioning. Proceedings of the 25th ACM international conference on Multimedia, [online] pp.1549–1557. Available at: https://arxiv.org/abs/1708.04390.

[13] Yoshikawa, Y., Shigeto, Y. and Takeuchi, A. (2017). STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics

[14] Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G. and Xu, J. (2019). COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval. arXiv:1805.08661

[15] Schamoni, S., Hitschler, J. and Riezler, S. (2018). A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions.

[16] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, pp.67–78

[17] 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. [online] experts.illinois.edu. Available at: https://experts.illinois.edu/en/publications/every-picture-tells-a-story-generating-sentences-from-images

[18] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C. and Berg, T.L. (2011). Baby talk: Understanding and generating simple image descriptions. CVPR 2011.

[19] Ryan Kiros, Ruslan Salakhutdinov, Richard Zemel (2014). Multimodal neural language models

[20] Kiros, R., Salakhutdinov, R. and Zemel, R.S. (2014). Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. arXiv:1411.2539 [cs]. [online] Available at: https://arxiv.org/abs/1411.2539

[21] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuille, A. (2015). Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). arXiv:1412.6632 [cs]. [online] Available at: https://arxiv.org/abs/1412.6632

[22] Karpathy, A., Joulin, A. and Fei-Fei, L. (2014). Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. arXiv:1406.5679 [cs]. [online] Available at: https://arxiv.org/abs/1406.5679

[23] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2014). Show and Tell: A Neural Image Caption Generator. [online] arXiv.org. Available at: https://arxiv.org/abs/1411.4555.

[24] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. [online] arXiv.org. Available at: https://arxiv.org/abs/1502.03044.

[25] Cong, S.N.D., Ngo, Q.H. and Jiamthapthaksin, R. (2016). State-of-the-Art Vietnamese Word Segmentation. 2016 2nd International Conference on Sci-

ence in Information Technology (ICSITech), [online] pp.119–124. Available at: https://arxiv.org/abs/1906.07662

[26] He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep Residual Learning for Image Recognition. [online] arXiv.org. Available at: https://arxiv.org/abs/1512.03385.

[27] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. [online] arXiv.org. Available at: https://arxiv.org/abs/1512.00567.

[28] Tan, M. and Le, Q.V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. [online] arXiv.org. Available at: https://arxiv.org/abs/1905.11946.

[29] Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.0473.

[30] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. [online] arXiv.org. Available at: https://arxiv.org/abs/1412.3555.

[31] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2001). BLEU. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. [online] Available at: https://dl.acm.org/citation.cfm?id=1073135.

[32] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. [online] aclanthology.org. Available at: https://aclanthology.org/W04-1013/.

[33] Vedantam, R., Zitnick, C.L. and Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726 [cs]. [online] Available at: https://arxiv.org/abs/1411.5726.

[34] Tran, V.T. (2021). trungtv/pyvi. [online] GitHub. Available at: https://github.com/trungtv/pyvi.

[35] GitHub. (2021). C++ tokenizer for Vietnamese. [online] Available at: https://github.com/coccoc/coccoc-tokenizer

[36] Chen, G. (2018). A Gentle Tutorial of Recurrent Neural Network with Error Backpropagation. arXiv:1610.02583 [cs]. [online] Available at: https://arxiv.org/abs/1610.02583.

[37] Jiang, C., Chen, Y., Chen, S., Bo, Y., Li, W., Tian, W. and Guo, J. (2019). A Mixed Deep Recurrent Neural Network for MEMS Gyroscope Noise Suppressing. Electronics, 8(2), p.181.

[38] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Lawrence, Z.C. and Dollár, P. (2014). Microsoft COCO: Common Objects in Context. [online] arXiv.org. Available at: https://arxiv.org/abs/1405.0312.

[39] LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature, 521(7553), pp.436–444.