UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning

Abstract

Image Captioning, the task of automatic generation of image captions, has attracted attentions from researchers in many fields of computer science, being computer vision, natural language processing and machine learning in recent years. This paper contributes to research on Image Captioning task in terms of extending dataset to a different language - Vietnamese. So far, there is no existed Image Captioning dataset for Vietnamese language, so this is the foremost fundamental step for developing Vietnamese Image Captioning. In this scope, we first build a dataset which contains manually written captions for images from Microsoft COCO dataset relating to sports played with balls, we called this dataset UIT-ViIC. UIT-ViIC consists of 19,250 Vietnamese captions for 3,850 images. Following that, we evaluate our dataset on deep neural network models and do comparisons with English dataset and two Vietnamese datasets built by different methods. UIT-ViIC is published on our lab website for research purposes.

Introduction

Generating descriptions for multimedia contents such as images and videos, so called Image Captioning, is helpful for e-commerce companies or news agencies. For instance, in e-commerce field, people will no longer need to put much effort into understanding and describing products' images on their websites because image contents can be recognized and descriptions are automatically generated. Inspired by Horus BIBREF0 , Image Captioning system can also be integrated into a wearable device, which is able to capture surrounding images and generate descriptions as sound in real time to guide people with visually impaired.

Image Captioning has attracted attentions from researchers in recent years BIBREF1, BIBREF2, BIBREF3, and there has been promising attempts dealing with language barrier in this task by extending existed dataset captions into different languages BIBREF3, BIBREF4.

In this study, generating image captions in Vietnamese language is put into consideration. One straightforward approach for this task is to translate English captions into Vietnamese by human or by using machine translation tool, Google translation. With the method of translating directly from English to Vietnamese, we found that the descriptions are sometimes confusing and unnatural to native people. Moreover, image understandings are cultural dependent, as in Western, people usually have different ways to grasp images and different vocabulary choices for describing contexts. For instance, in Fig. FIGREF2, one MS-COCO English caption introduce about "a baseball player in motion of pitching", which makes sense and capture accurately the main activity in the image. Though it sounds sensible in English, the sentence becomes less meaningful when we try to translate it into Vietnamese. One attempt of translating the sentence is performed by Google Translation, and the result is not as expected.

Therefore, we come up with the approach of constructing a Vietnamese Image Captioning dataset with descriptions written manually by human. Composed by Vietnamese people, the sentences would be more natural and friendlier to Vietnamese users. The main resources we used from MS-COCO for our dataset are images. Besides, we consider having our dataset focus on sportball category due to several reasons:

By concentrating on a specific domain we are more likely to improve performance of the Image Captioning models. We expect our dataset can be used to confirm or reject this hypothesis.

Sportball Image Captioning can be used in certain sport applications, such as supportting journalists describing great amount of images for their articles.

Our primary contributions of this paper are as follows:

Firstly, we introduce UIT-ViIC, the first Vietnamese dataset extending MS-COCO with manually written captions for Image Captioning. UIT-ViIC is published for research purposes.

Secondly, we introduce our annotation tool for dataset construction, which is also published to help annotators conveniently create captions.

Finally, we conduct experiments to evaluate state-of-the-art models (evaluated on English dataset) on UIT-ViIC dataset, then we analyze the performance results to have insights into our corpus.

The structure of the paper is organized as follows. Related documents and studies are presented in Section SECREF2. UIT-ViIC dataset creation is described in Section SECREF3. Section SECREF4 describes the methods we implement. The experimental results and analysis are presented in Section SECREF5. Conclusion and future work are deduced in Section SECREF6.

Related Works

We summarize in Table TABREF8 an incomplete list of published Image Captioning datasets, in English and in other languages. Several image caption datasets for English have been constructed, the representative examples are Flickr3k BIBREF5, BIBREF6; Flickr 30k BIBREF7 – an extending of Flickr3k and Microsoft COCO (Microsoft Common in Objects in Context) BIBREF8.

Besides, several image datasets with non-English captions 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 BIBREF9 and image captioning BIBREF10, BIBREF3, Chinese for image tagging, captioning and retrieval BIBREF4. Each of these datasets is built on top of an existing English dataset, with MS-COCO as the most popular choice.

Our dataset UIT-ViIC is constructed using images from Microsoft COCO (MS-COCO). MS-COCO dataset includes more than 150,000 images, divided into three distributions: train, vailidate, test. For each image, five captions are provided independently by Amazon’s Mechanical Turk. MS-COCO is the most popular dataset for Image Captioning thanks to the MS-COCO challenge (2015) and it has a powerful evaluation server for candidates.

Regarding to the Vietnamese language processing, there are quite a number of research works on other tasks such as parsing, part-of-speech, named entity recognition, sentiment analysis, question answering. However, to the extent of our knowledge, there are no research publications on image captioning for Vietnamese. Therefore, we decide to build a new corpus of Vietnamese image captioning for Image Captioning research community and evaluate the state-of-the-art models on our corpus. In particular, we validate and compare the results by BLEU BIBREF11, ROUGE BIBREF12 and CIDEr BIBREF13 metrics between Neural Image Captioning (NIC) model BIBREF14, Image Captioning model from the Pytorch-tutorial BIBREF15 by Yunjey on our corpus as the pioneering results.

Dataset Creation

This section demonstrates how we constructed our new Vietnamese dataset. The dataset consists of

3,850 images relating to sports played with balls from 2017 edition of Microsoft COCO. Similar to most Image Captioning datasets, we provide five Vietnamese captions for each image, summing up to 19,250 captions in total.

Dataset Creation ::: Annotation Tool with Content Suggestions

To enhance annotation efficiency, we present a web-based application for caption annotation. Fig. FIGREF10 is the annotation screen of the application.

Our tool assists annotators conveniently load images into a display and store captions they created into a new dataset. With saving function, annotator can save and load written captions for reviewing purposes. Furthermore, users are able to look back their works or the others' by searching image by image ids.

The tool also supports content suggestions taking advantage of existing information from MS-COCO. First, there are categories hints for each image, displaying as friendly icon. Second, original English captions are displayed if annotator feels their needs. Those content suggestions are helpful for annotators who can't clearly understand images, especially when there are issues with images' quality.

Dataset Creation ::: Annotation Process

In this section, we describes procedures of building our sportball Vietnamese dataset, called UIT-ViIC.

Our human resources for dataset construction involve five writers, whose ages are from 22-25. Being native Vietnamese residents, they are fluent in Vietnamese. All five UIT-ViIC creators first research and are trained about sports knowledge as well as the specialized vocabulary before starting to work.

During annotation process, there are inconsistencies and disagreements between human's understandings and the way they see images. According to Micah Hodosh et al BIBREF5, most images' captions on Internet nowadays tend to introduce information that cannot be obtained from the image itself, such as people name, location name, time, etc. Therefore, to successfully compose meaningful descriptive captions we expect, their should be strict guidelines.

Inspired from MS-COCO annotation rules BIBREF16, we first sketched UIT-ViIC's guidelines for our captions:

Each caption must contain at least ten Vietnamese words.

Only describe visible activities and objects included in image.

Exclude name of places, streets (Chinatown, New York, etc.) and number (apartment numbers, specific time on TV, etc.)

Familiar English words such as laptop, TV, tennis, etc. are allowed.

Each caption must be a single sentence with continuous tense.

Personal opinion and emotion must be excluded while annotating.

Annotators can describe the activities and objects from different perspectives.

Visible “thing” objects are the only one to be described.

Ambiguous “stuff” objects which do not have obvious “border” are ignored.

In case of 10 to 15 objects which are in the same category or species, annotators do not need to include them in the caption.

In comparison with MS-COCO BIBREF16 data collection guidelines in terms of annotation, UIT-ViIC’s guidelines has similar rules (1, 2, 8, 9, 10) . We extend from MS-COCO’s guidelines with five new rules to our own and have modifications in the original ones.

In both datasets, we would like to control sentence length and focus on describing important subjects only in order to make sure that essential information is mainly included in captions. The MS-COCO threshold for sentence’s length is 8, and we raise the number to 10 for our dataset. One reason for this change is that an object in image is usually expressed in many Vietnamese words. For example, a “baseball player” in English can be translated into “vận động viên bóng chày” or “cầu thủ bóng chày”, which already accounted for a significant length of the Vietnamese sentence. In addition, captions must be single sentences with continuous tense as we expect our model’s output to capture what we are seeing in the image in a consise way.

On the other hand, proper name for places, streets, etc must not be mentioned in this dataset in order to avoid confusions and incorrect identification names with the same scenery for output. Besides, annotators’ personal opinion must be excluded for more meaningful captions. Vietnamese words for several English ones such as tennis, pizza, TV, etc are not existed, so annotators could use such familiar words in describing captions. For some images, the subjects are ambiguous and not descriptive which would be difficult for annotators to describe in words. That’s the reason why annotators can describe images from more than one perspective.

Dataset Creation ::: Dataset Analysis

After finishing constructing UIT-ViIC dataset, we have a look in statistical analysis on our corpus in this section. UIT-ViIC covers 3,850 images described by 19,250 Vietnamese captions. Sticking strictly to our annotation guidelines, the majority of our captions are at the length of 10-15 tokens. We are using the term “tokens” here as a Vietnamese word can consist of one, two or even three tokens. Therefore, to apply Vietnamese properly to Image Captioning, we present a tokenization tool - PyVI BIBREF17, which is specialized for Vietnamese language tokenization, at words level. The sentence length using token-level tokenizer and word-level tokenizer are compared and illustrated in Fig. FIGREF23, we can see there are variances there. So that, we can suggest that the tokenizer performs well enough, and we can expect our Image Captioning models to perform better with Vietnamese sentences that are tokenized, as most models perform more efficiently with captions having fewer words.

Table TABREF24 summarizes top three most occuring words for each part-of-speech. Our dataset vocabulary size is 1,472 word classes, including 723 nouns, 567 verbs, and 182 adjectives. It is no surprise that as our dataset is about sports with balls, the noun “bóng” (meaning “ball") occurs most, followed by “sân” and "cầu thủ" (“pitch” and “athlete” respectively). We also found that the frequency of word “tennis” stands out among other adjectives, which specifies that the set covers the majority of tennis sport, followed by “bóng chày” (meaning “baseball”). Therefore, we expect our model to generate the best results for tennis images.

Image Captioning Models

Our main goal in this section is to see if Image Captioning models could learn well with Vietnamese language. To accomplish this task, we train and evaluate our dataset with two published Image Captioning models applying encoder-decoder architecture. The models we propose are Neural Image

Captioning (NIC) model BIBREF14, Image Captioning model from the Pytorch-tutorial BIBREF15 by Yunjey.

Overall, CNN is first used for extracting image features for encoder part. The image features which are presented in vectors will be used as layers for decoding. For decoder part, RNN - LSTM are used to embed the vectors to target sentences using words/tokens provided in vocabulary.

Image Captioning Models ::: Model from Pytorch tutorial

Model from pytorch-tutorial by Yunjey applies the baseline technique of CNN and LSTM for encoding and decoding images. Resnet-152 BIBREF18 architecture is proposed for encoder part, and we use the pretrained one on ILSVRC-2012-CLS BIBREF19 image classification dataset to tackle our current problem. LSTM is then used in this model to generate sentence word by word.

Image Captioning Models ::: NIC - Show and tell model

NIC - Show and Tell uses CNN model which is currently yielding the state-of-the-art results. The model achieved 0.628 when evaluating on BLEU-1 on COCO-2014 dataset. For CNN part, we utilize VGG-16 BIBREF20 architecture pre-trained on COCO-2014 image sets with all categories. In decoding part, LSTM is not only trained to predict sentence but also to compute probability for each word to be generated. As a result, output sentence will be chosen using search algorithms to find the one that have words yielding the maximum probabilities.

Experiments ::: Experiment Settings ::: Dataset preprocessing

As the images in our dataset are manually annotated by human, there are mistakes including grammar, spelling or extra spaces, punctuation. Sometimes, the Vietnamese’s accent signs are placed in the wrong place due to distinct keyboard input methods. Therefore, we eliminate those common errors before working on evaluating our models.

Experiments ::: Experiment Settings ::: Dataset preparation

We conduct our experiments and do comparisons through three datasets with the same size and images of sportball category: Two Vietnamese datasets generated by two methods (translated by Google Translation service and annotated by human) and the original MS-COCO English dataset. The three sets are distributed into three subsets: 2,695 images for the training set, 924 images for validation set and 231 images for test set.

Experiments ::: Evaluation Measures

To evaluate our dataset, we use metrics proposed by most authors in related works of extending Image Captioning dataset, which are BLEU BIBREF11, ROUGE BIBREF12 and CIDEr BIBREF13. BLEU and ROUGE are often used mainly for text summarization and machine translation, whereas CIDEr was designed especially for evaluating Image Captioning models.

Experiments ::: Evaluation Measures ::: Comparison methods

We do comparisons with three sportball datasets, as follows:

Original English (English-sportball): The original MS-COCO English dataset with 3,850 sportball images. This dataset is first evaluated in order to have base results for following comparisons.

Google-translated Vietnamese (GT-sportball): The translated MS-COCO English dataset into Vietnamese using Google Translation API, categorized into sportball.

Manually-annotated Vietnamese (UIT-ViIC): The Vietnamese dataset built with manually written captions for images from MS-COCO, categorized into sportball.

Experiments ::: Experiment Results

The two following tables, Table TABREF36 and Table TABREF36, summarize experimental results of Pytorch-tutorial, NIC - Show and Tell models. The two models are trained with three mentioned datasets, which are English-sportball, GT-sportball, UIT-ViIC. After training, 924 images from validation subset for each dataset are used to validate the our models.

As can be seen in Table TABREF36, with model from Pytorch tutorial, MS-COCO English captions categorized with sportball yields better results than the two Vietnamese datasets. However, as number of consecutive words considered (BLEU gram) increase, UIT-ViIC's BLEU scores start to pass that of English sportball and their gaps keep growing. The ROUGE-L and CIDEr-D scores for UIT-ViIC model prove the same thing, and interestingly, we can observe that the CIDEr-D score for the UIT-ViIC model surpasses English-sportball counterpart.

The same conclusion can be said from Table TABREF36. Show and Tell model's results show that MS-COCO sportball English captions only gives better result at BLEU-1. From BLEU-3 to BLEU-4, both GT-sportball and UIT-ViIC yield superior scores to English-sportball. Besides, when limiting MS-COCO English dataset to sportball category only, the results are higher (0.689, 0.501, 0.355, 0.252) than when the model is trained on MS-COCO with all images, which scored only 0.629, 0.436, 0.290, 0.193 (results without tuning in 2018) from BLEU-1 to BLEU-4 respectively.

When we compare between two Vietnamese datasets, UIT-ViIC models perform better than sportball dataset translated automatically, GT-sportball. The gaps between the two results sets are more trivial in NIC model, and the numbers get smaller as the BLEU's n-gram increase.

In Fig. FIGREF37, two images inputted into the models generate two Vietnamese captions that are able to describe accurately the sport game, which is soccer. The two models can also differentiate if there is more than one person in the images. However, when comparing GT-sportball outputs with UIT-ViIC ones in both images, UIT-ViIC yield captions that sound more naturally, considering Vietnamese language. Furthermore, UIT-ViIC demonstrates the specific action of the sport more accurately than GT-sportball. For example, in the below image of Fig. FIGREF37, UIT-ViIC tells the exact action (the man is preparing to throw the ball), whereas GT-sportball is mistaken (the man swing the bat). The confusion of GT-sportball happens due to GT-sportball train set is translated from original MS-COCO dataset, which is annotated in more various perspective and wider vocabulary range with the dataset size is not big enough.

There are cases when the main objects are too small, both English and GT - sportball captions tell the unexpected sport, which is tennis instead of baseball, for instance. Nevertheless, the majority of UIT-ViIC captions can tell the correct type of sport and action, even though the gender and age identifications still need to be improved.

Conclusion and Further Improvements

In this paper, we constructed a Vietnamese dataset with images from MS-COCO, relating to the category within sportball, consisting of 3,850 images with 19,250 manually-written Vietnamese captions. Next, we conducted several experiments on two popular existed Image Captioning models to evaluate their efficiency when learning two Vietnamese datasets. The results are then compared with the original

MS-COCO English categorized with sportball category.

Overall, we can see that English set only out-performed Vietnamese ones in BLEU-1 metric, rather, the Vietnamese sets performing well basing on BLEU-2 to BLEU-4, especially CIDEr scores. On the other hand, when UIT-ViIC is compared with the dataset having captions translated by Google, the evaluation results and the output examples suggest that Google Translation service is able to perform acceptablly even though most translated captions are not perfectly natural and linguistically friendly. As a results, we proved that manually written captions for Vietnamese dataset is currently prefered.

For future improvements, extending the UIT-ViIC's cateogry into all types of sport to verify how the dataset's size and category affect the Image Captioning models' performance is considered as our highest priority. Moreover, the human resources for dataset construction will be expanded. Second, we will continue to finetune our experiments to find out proper parameters for models, especially with encoding and decoding architectures, for better learning performance with Vietnamese dataset, especially when the categories are limited.