Vietnamese caption generation for images

Students

Dinh Ba Khanh Trung

Nguyen Manh Tien

Supervisor

M.S Le Dinh Huynh

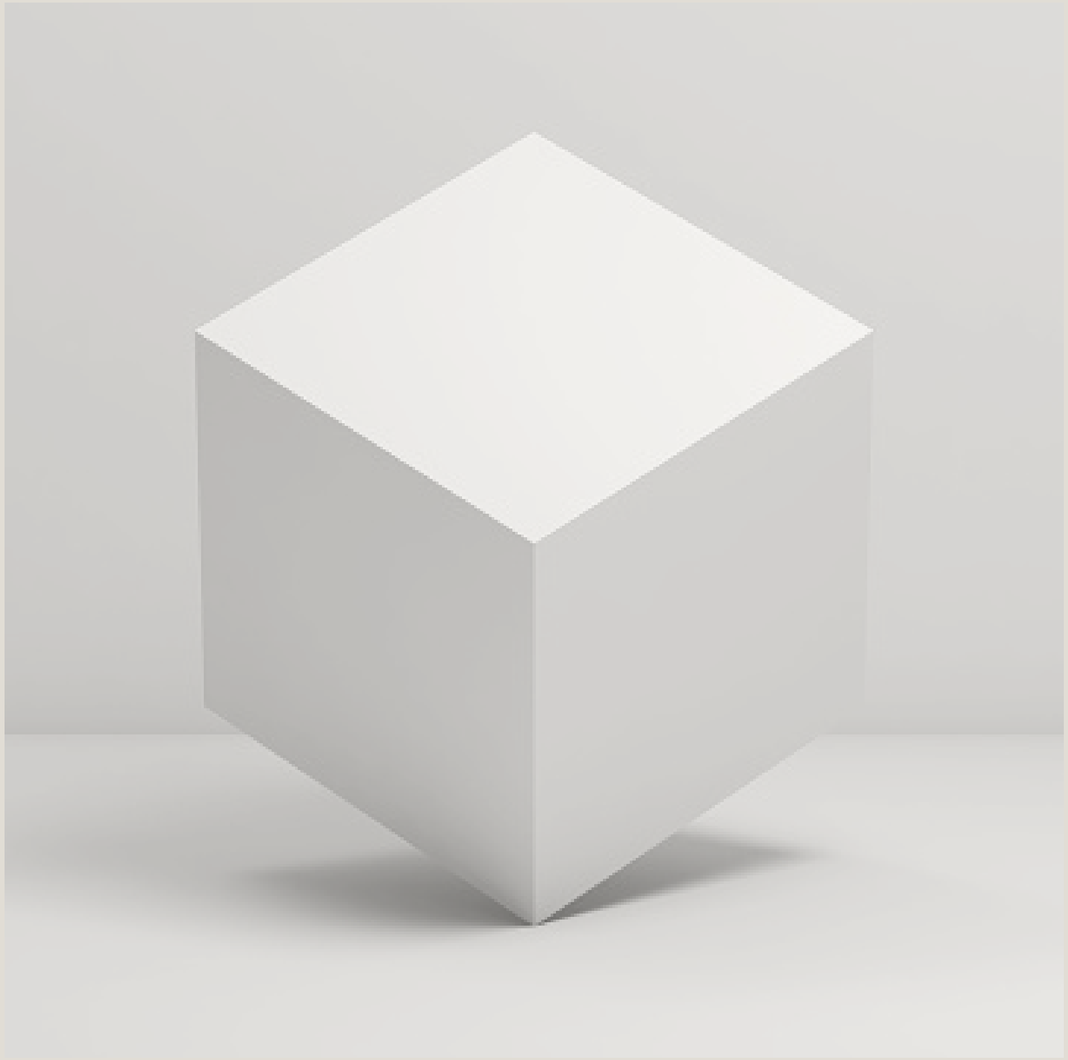


Table of Contents

I. Introduction

II. Data

III. Model

IV. Conclusion & Future Works

I.Introduction

I.Introduction

1. What is Image Caption Generation?
2. Applications of Images Caption Generation
3. Contributions

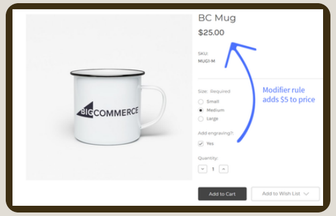
What is
Image Caption Generation?

- Image caption generation is the process of generating a textual description for given images

Một cầu thủ bóng đá đang chuẩn bị sút bóng

Applications of Images Caption Generation

- Support people with visual impairments
- Image captioning models transcribe the surrounding scenes and output the caption into a text to speech model



- Describe product images in the commerce field
- Image captioning models can be used to automatically generate the description to understand and describe product images on their websites

- Optimize the search quality for image based search engines
- 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

Contributions

01 Built a full Vietnamese version training dataset from the MS-COCO dataset for Vietnamese caption

02 Created Flickr900 to extend existing Vietnameses captioning dataset UIT-ViIC which contain sport-ball images to balance this dataset

03 Made a simple annotation tool for dataset construction to assist annotator to create caption efficiently

04 Improved the model performance by combining the previous works with newly proposed techniques

II.Data

II.Data

1. Related Works
2. Data Creation CoCo-Vn datasets
3. Data Creation Flickr900 datasets
4. Annotation Tool
5. Results

Related Works

- Each of these datasets is based on an existing English dataset, the most prominent of which is MS-COCO
- There are three datasets IAPR TC-12, AIC-ICC and WikiCaps that use data from the internet instead of the popular dataset from MS-COCO and Flickr
- UIT-ViIC is the first image captioning dataset in Vietnamese, adopting Microsoft COCO as its data source

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

Non-English public image datasets with manually annotated

Data Creation

CoCo-Vn datasets

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.

1. Preprocessing the english dataset
2. Remove passive voice (“that reads", “that says",”telling”)
3. Remove specific brand of items or company
4. Remove name of people, places, street, national,.
5. Rewrite sentences that are not in simple form
6. Remove detailed information
7. Fixing miss spelling word

Data Creation

Flickr900 datasets

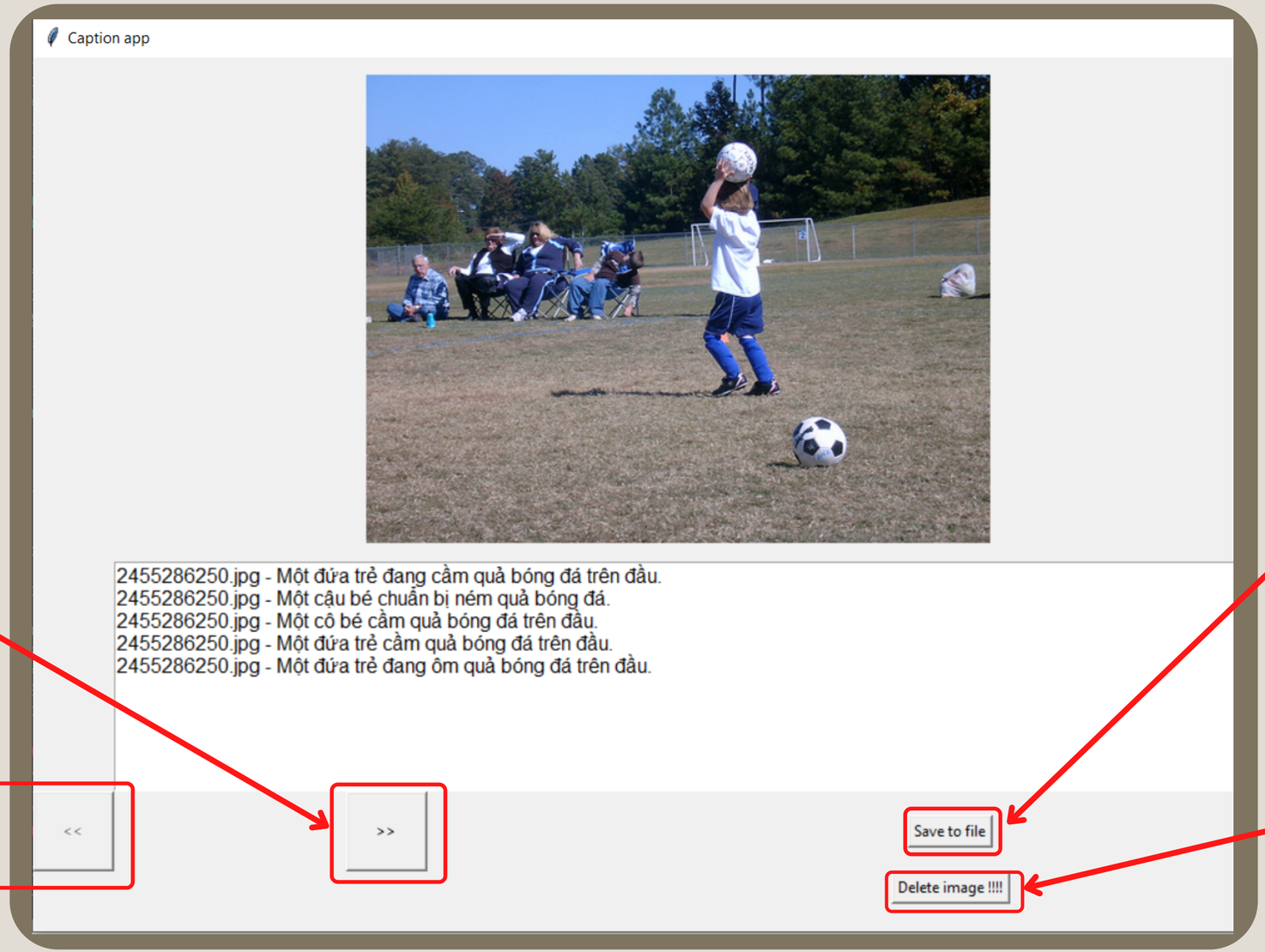
English caption: Two volleyball players standing next to a net that is part of an indoor court , celebrating a win or a point scored , with several people looking on .
Translated google translate: Hai cầu thủ bóng chuyền đứng cạnh lưới là một phần của sân trong nhà, ăn mừng một chiến thắng hoặc một điểm 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.

- Flickr900 contained 900 images of sports played with balls from the 30.000 images version of the Flickr dataset (Flickr30k)
- Chosen by extracting the image's object detection label in the Flickr30k annotation file
- 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

Image Caption Generation Rules

1. Each image caption must contain at least eight word
2. Describe all the essential parts of the scene, visible activities, and objects
3. Ignore all specific details like names of places, streets, manufacturers
4. Each caption must be a single statement
5. While annotating, personal opinion and emotion must be eliminated
6. Remove all unclear items and describe visible objects.

Annotation Tool



- load next image and captions into interface
- load previous image and captions into interface
- save written sentences into the dataset
- The delete button will be used to delete the image that does not relate to our topic

Results

- COCO-VN consisting of all 118.344 images in the training dataset of MS-COCO with 591.720 captions
- Flick900 was made up of
  - 900 images
  - 4500 Vietnamese captions

	Flickr900	**UIT-ViIC + Flickr900**
tennis	8	1666
baseball	121	1510
football	318	876
volleyball	104	223
American football	6	28

III.Model

III.Model

1. Overview
2. Model Architecture
    a. Tokenizer
    b. Encoder
    c. Decoder
    d. Evaluation method

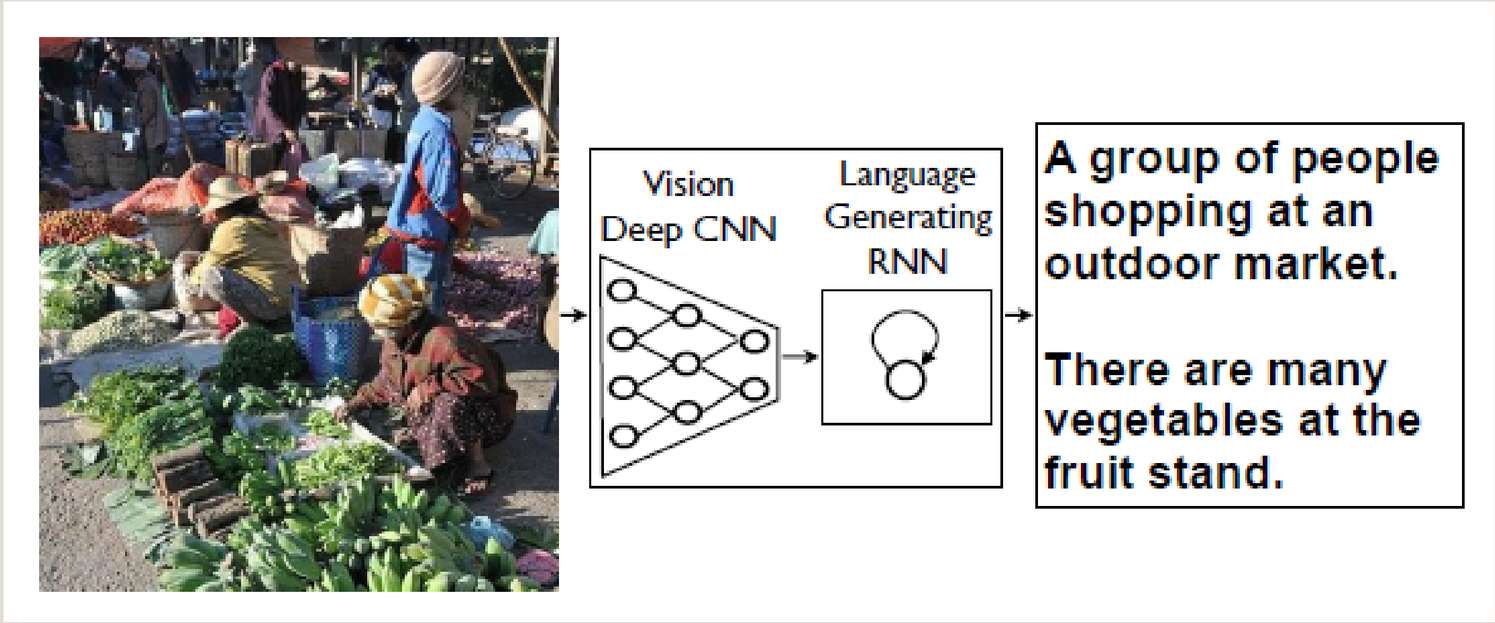
Related Work

Vinyals et al

Uses a convolutional neural network as a encoder to extract features from images followed by a language generating RNN for caption generation

Xu et al

Uses the output of convolutional layers of the convolutional neural network to generate image caption base on an attention mechanism



Overview

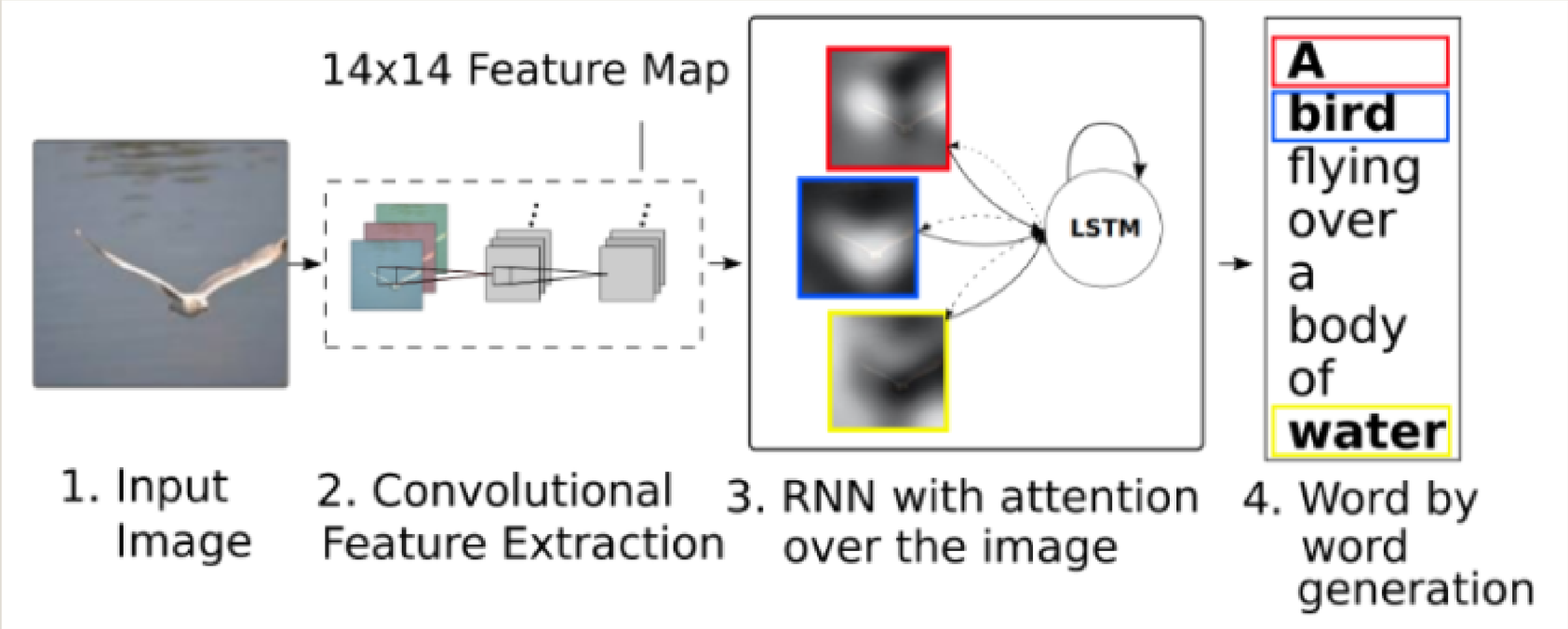
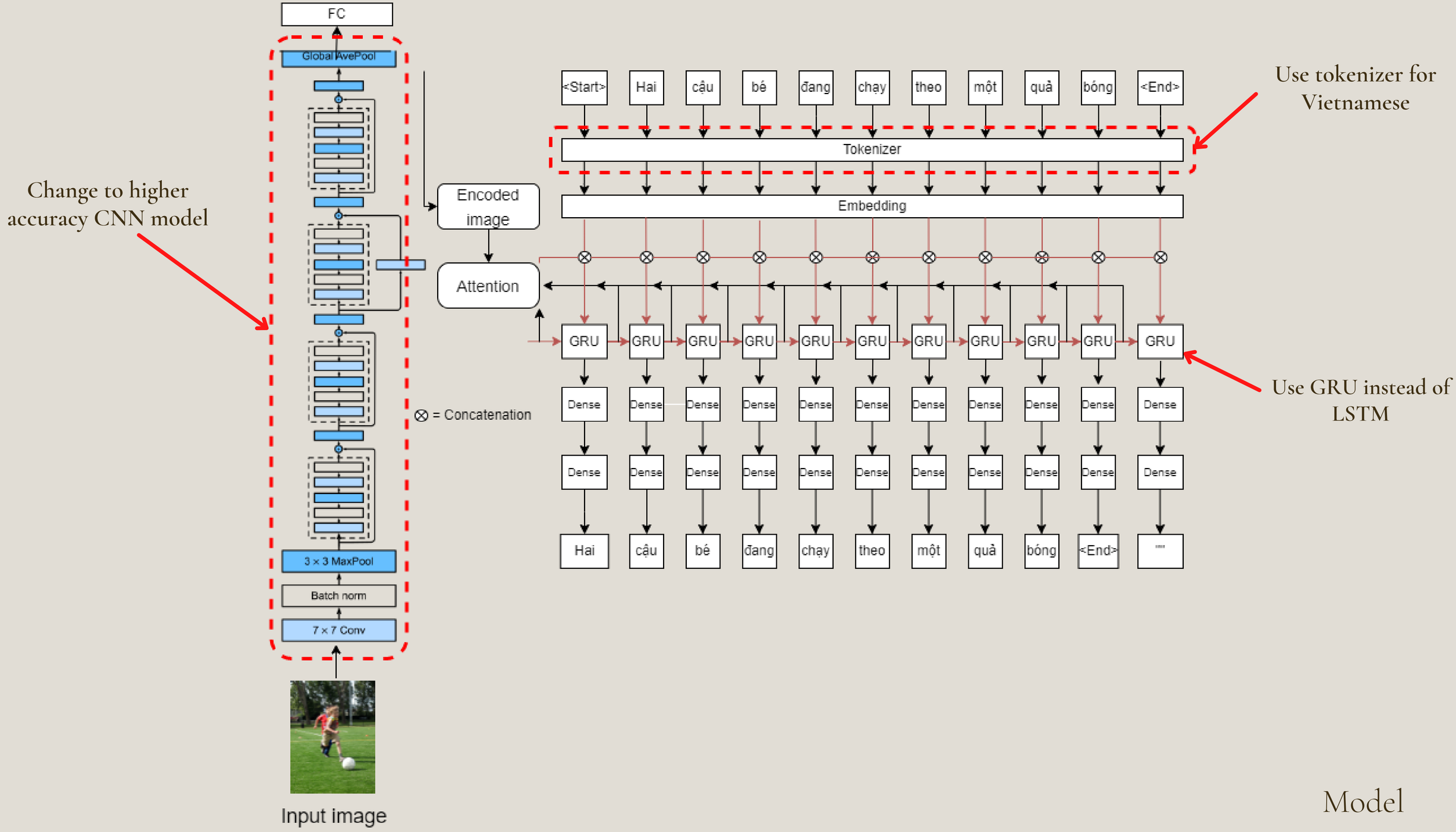
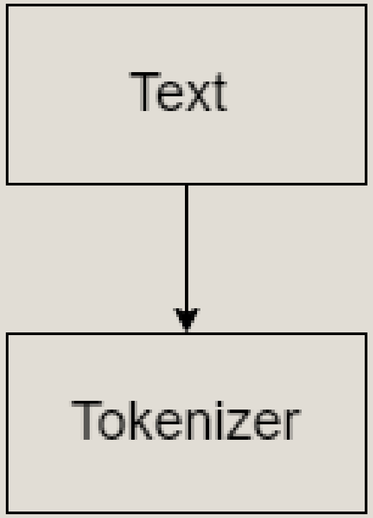


Image captioning model contains an encoder for image and a decoder to generate caption


Change to higher accuracy CNN model
FC
Global AvePool
3 × 3 MaxPool
Batch norm
7 × 7 Conv
Input image
Encoded image
Attention
⊗ = Concatenation
<Start>
Hai
cậu
bé
đang
chạy
theo
một
quả
bóng
<End>
Tokenizer
Embedding
GRU
Dense
Use tokenizer for Vietnamese
Use GRU instead of LSTM
Model

Model Architecture



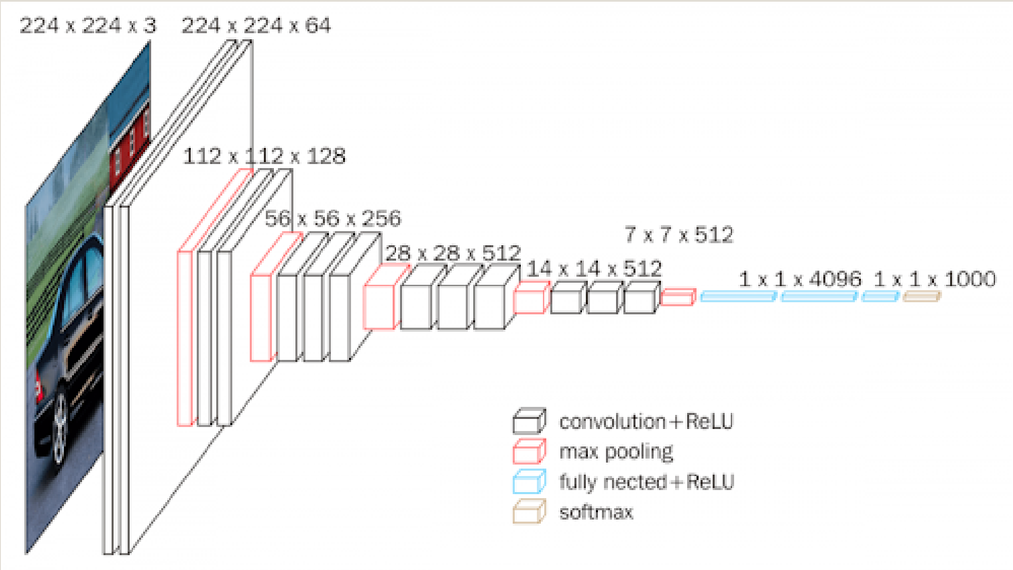
Tokenizer

- Tokenizer breaks down a phrase, sentence, paragraph, or even an entire text document into small fragments, such as words or terms
- Each of these small fragments is called a token

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

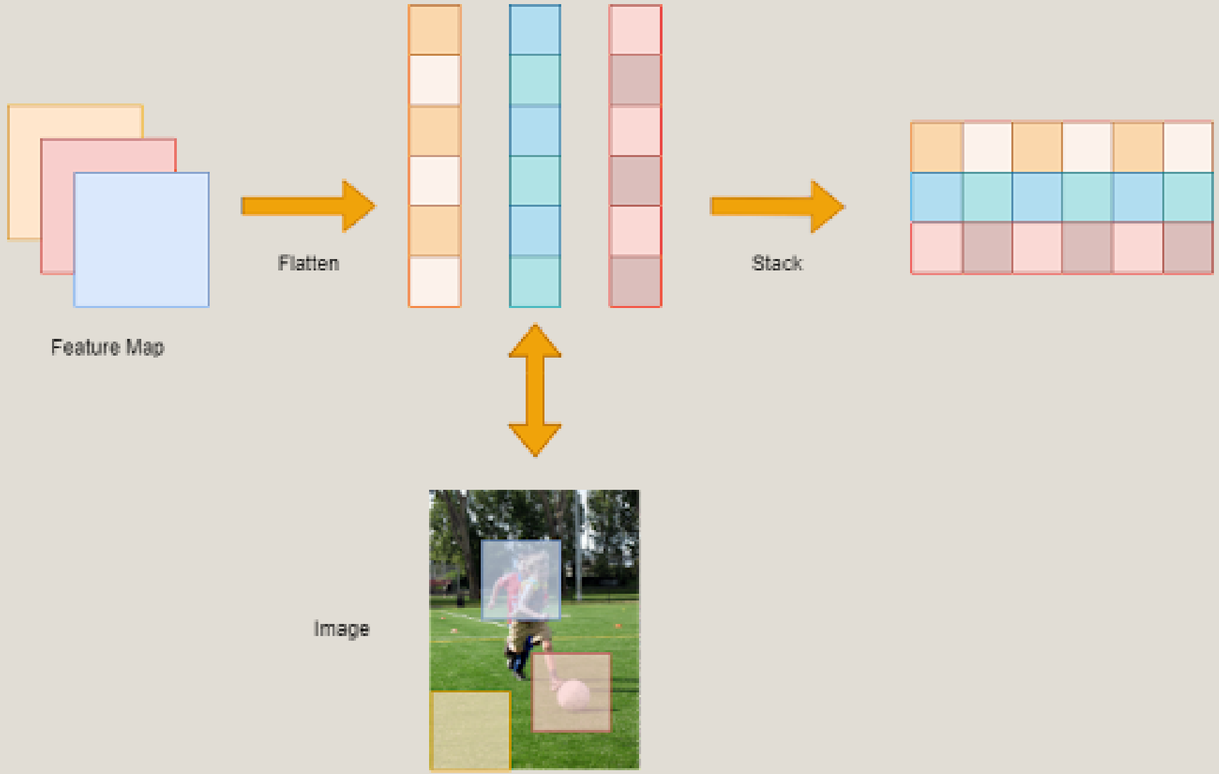
- PyVI and Coccoc-Tokenizer , both specialized for tokenization at the word level in the Vietnamese language

Model Architecture

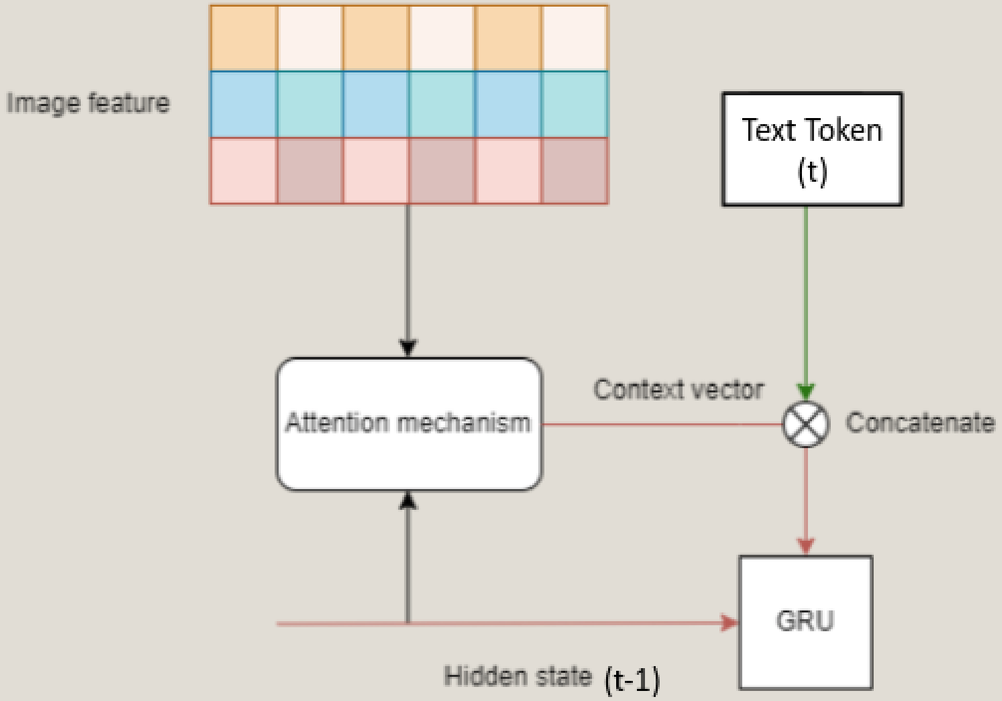


Encoder

Encoder take raw image and put into a CNN to extract valuable feature from the image



Model Architecture



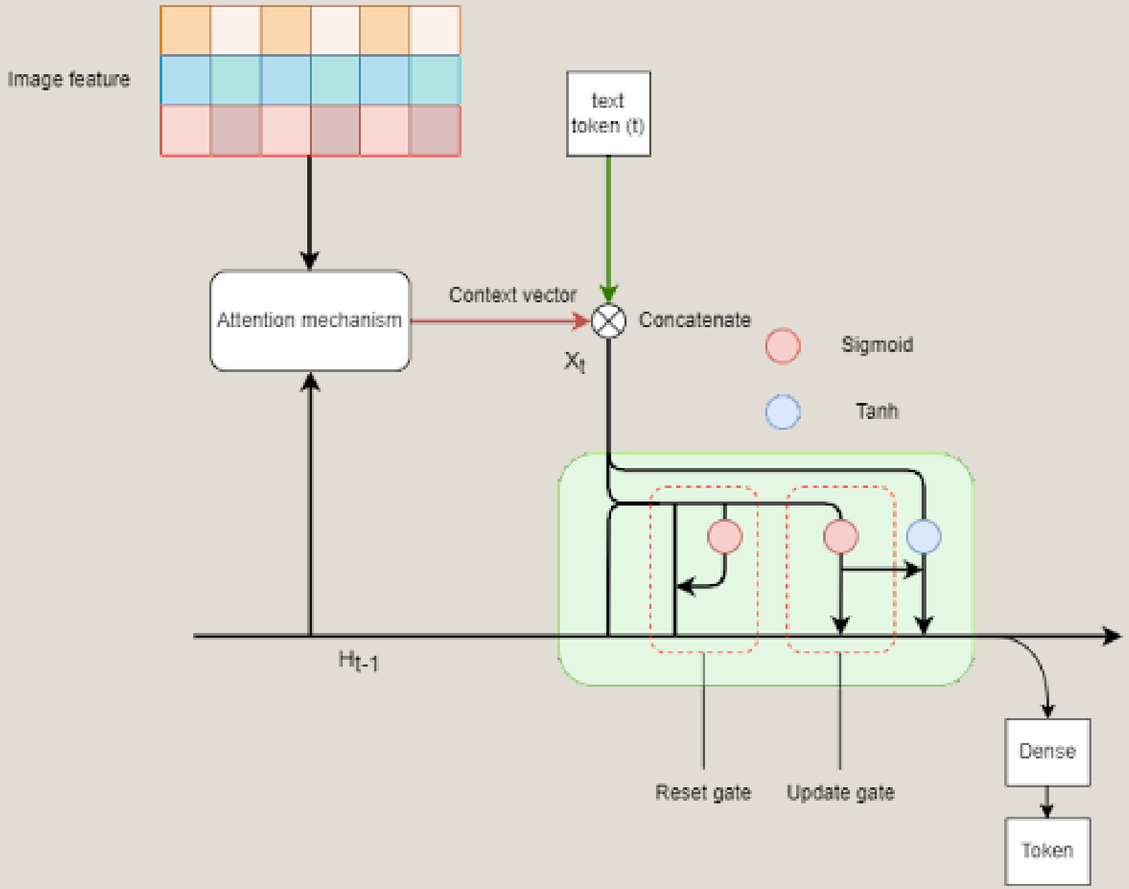
Attention mechanism

Approach by Bahdanau at el for the attention mechanism

Image features extracted by convolutional neural networks and previous hidden state are passed through the attention mechanism

The context vector is concatenated with the decoder's current input then pass through GRU

Model Architecture



Decoder

Uses GRU to generate text for image captions

GRU contain:

Reset gate: Decides what will be removed from the previous hidden time steps

Update gate: The update gate determine which information will be pass through the next state

GRU calculate the hidden state for next step and the input for Dense layer to generate text tokens

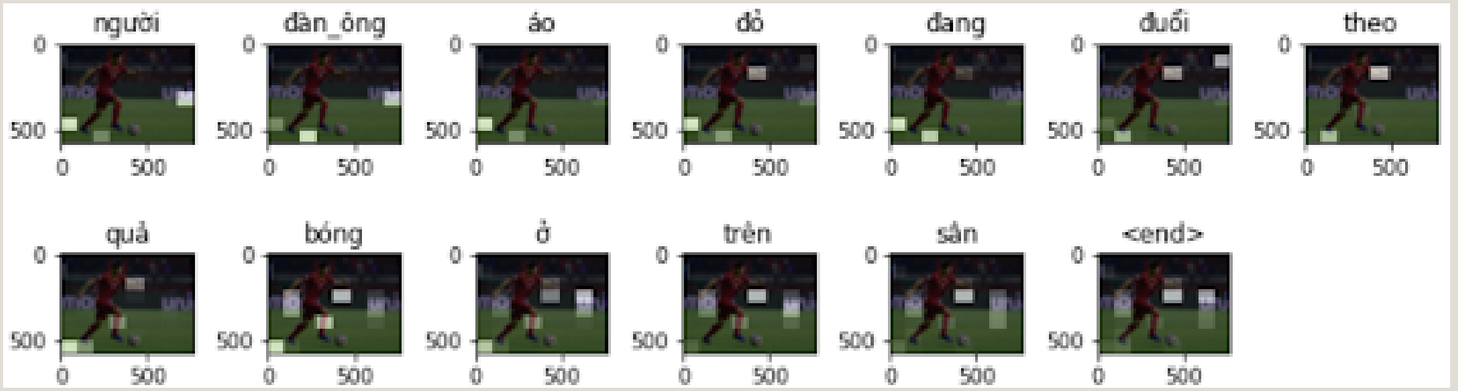
Model Evaluation

Evaluation method

- Metrics:
  - BLEU
  - ROUGE
  - CIDEr
- Experiment results

Evaluate on 924 image in validation set of UIT-ViIC

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**

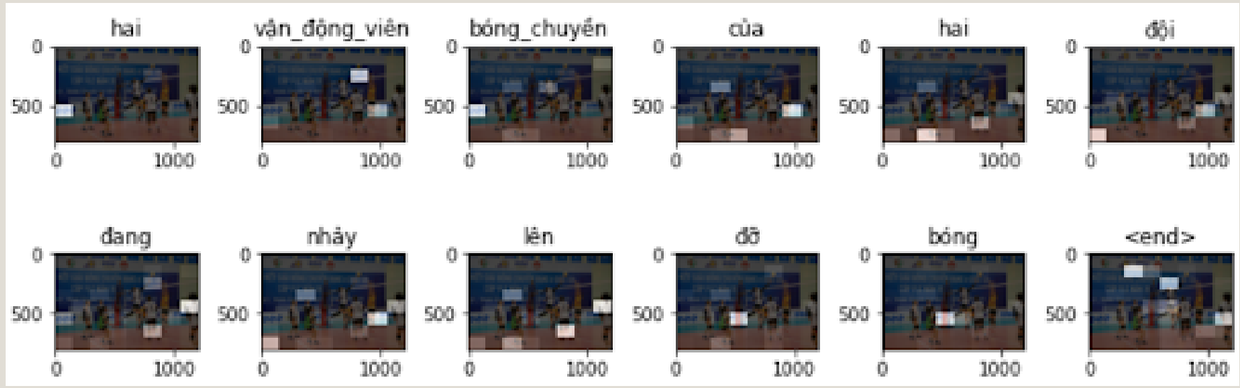


InceptionV3(UIT_ViIC):
một cầu thủ bóng đá đang chuẩn bị sút bóng

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



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

IV.Conclusion & Future Works

Conclusions

Data

- Publish 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
- Flickr900 including 900 images of sportball in Flickr-30K come along with 4500 manually-written Vietnamese captions to enrich UIT-ViIC dataset

Model

- Tweak the most famous encoder-decoder with attention to image captioning model with a more accurate image feature extraction model for encoder and use the GRU for the decoder

Future Works

- Working 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
- Replace greedy search with a beam search for better caption generation performance and quality
- 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
- 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


Thanks for watching