Image Captioning in Vietnamese Language

Supervisor:
Dr. Phan Việt Anh

Trainees:
Nguyễn Minh Dũng
Nguyễn Duy Nhất
Lê Công Pha
Võ Minh Tâm

Mar 1st, 2021

VINBIGDATA

CONTENTS

I. OVERVIEW

Image Captioning?

Src: MS COCO Dataset

1. A woman with a racket goes to hit a tennis ball.

2. The tennis player is ready to hit the ball.

I. OVERVIEW

Image Captioning?

Vietnamese?

1. A woman with a racket goes to hit a tennis ball.

2. The tennis player is ready to hit the ball.

Src: MS COCO Dataset

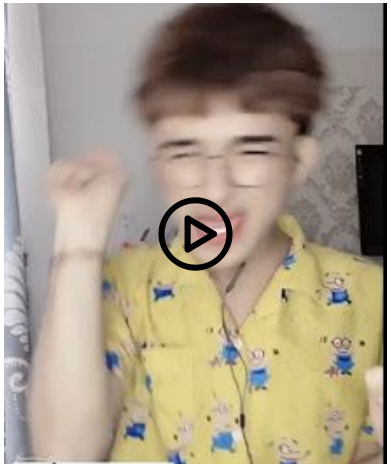
I. INPUT - OUTPUT

Input

Vận động viên tennis nữ đang vung vợt đỡ bóng.

Output

I. APPLICATIONS



Src: Internet

II. DATASET

UIT-ViIC (UIT-Vietnamese Image Captioning)		
Authors	Quan Hoang Lam, Quang Duy Le, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen (The UIT NLP Group, University of Information Technology - VNU HCM) [1]	
Ori-src	MS COCO - sportball	
Train set	**2695** imgs	**3850** imgs 5 captions/img 10-15 tokens/caption mostly
Val set	**924** imgs	
Test set	**231** imgs	

[1] https://arxiv.org/pdf/2002.00175.pdf

II. Dataset

1. Các cầu thủ bóng đá dang di chuyển ở trên sân để tranh bóng.
2. Một nam thủ môn đang truwowc trên sân để bắt bóng.
3. Các cầu thủ bóng đá đang thi đấu ở trên sân. /
4. Một cầu thủ áo xanh đang khuỵu gối ở cạnh quả bóng.
5. Những cậu bé đang chơi bóng đá ở trên sân.

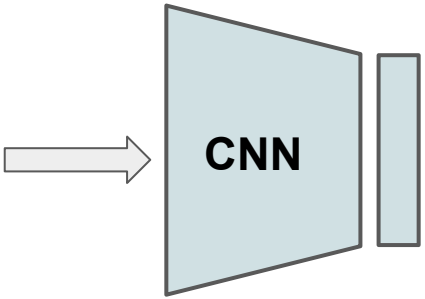
III. APPROACH

- **Preprocessing mistakes:**
  - Grammar
  - Spelling
  - Extra spaces
  - Punctuation
  - Vietnamese’s accent signs
- **Image Data Augmentation**
  - Random Cropping
  - Random Flipping

III. APPROACH

- **Model architecture**

Input



Encoder

Resnet-152
Inception-V3
EfficientNet

Decoder

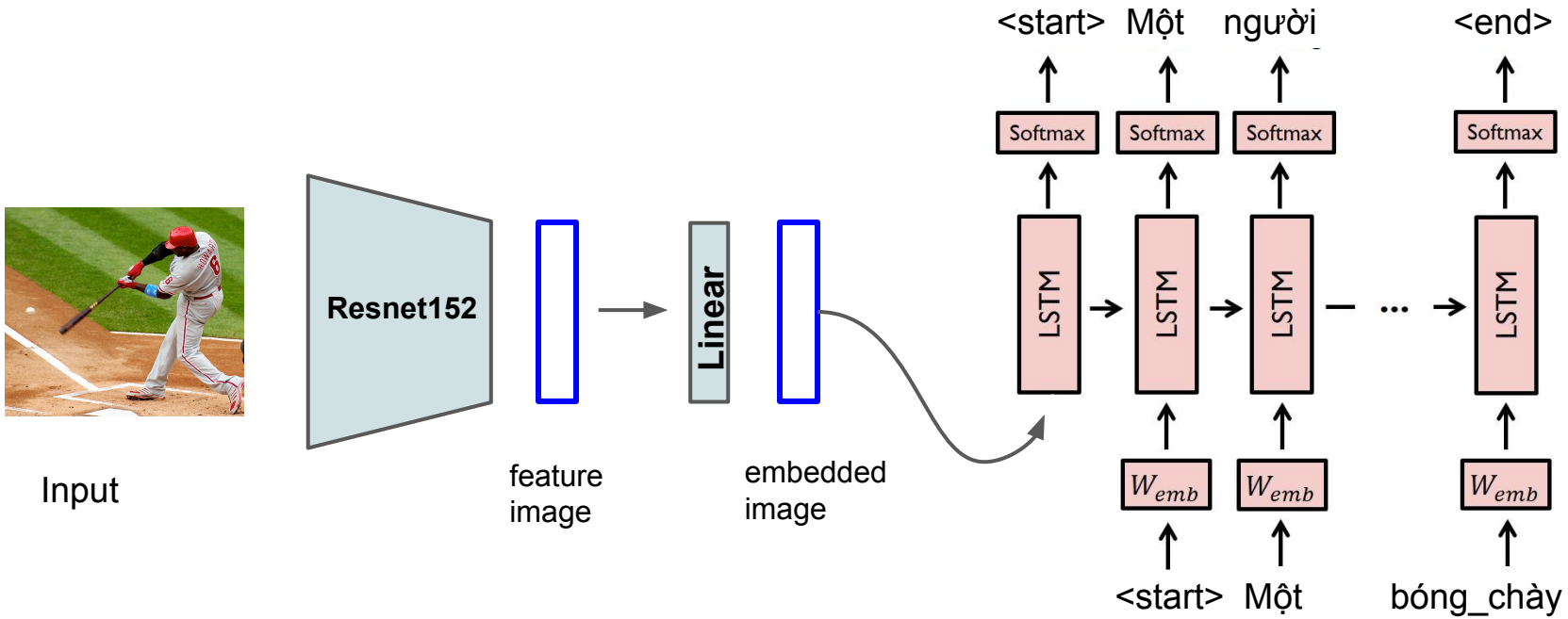
GRU
LSTM

Một người đàn ông đang xoay người để đánh quả bóng chày .

Caption

Resnet152 + LSTM

Caption segmentation: <start> Một người đàn_ông đang xoay người để đánh_quả bóng_chày <end>



Resnet152 + Attention+ LSTM

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$

$a_{t,:,:} = \text{softmax}(e_{t,:,:})$

$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

Resnet152

Alignment scores

$e_{1,1,1}$ $e_{1,1,2}$ $e_{1,1,3}$
$e_{1,2,1}$ $e_{1,2,2}$ $e_{1,2,3}$
$e_{1,3,1}$ $e_{1,3,2}$ $e_{1,3,3}$

Attention scores

$a_{1,1,1}$ $a_{1,1,2}$ $a_{1,1,3}$
$a_{1,2,1}$ $a_{1,2,2}$ $a_{1,2,3}$
$a_{1,3,1}$ $a_{1,3,2}$ $a_{1,3,3}$

$h_{1,1}$ $h_{1,2}$ $h_{1,3}$
$h_{2,1}$ $h_{2,2}$ $h_{2,3}$
$h_{3,1}$ $h_{3,2}$ $h_{3,3}$

S_0

S_1

Một

y_1

c_1 y_0

[START]

Resnet152 + Attention+ LSTM

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$
$$a_{t,:,:} = \text{softmax}(e_{t,:,:})$$
$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$

Resnet152

$h_{1,1}$ $h_{1,2}$ $h_{1,3}$
$h_{2,1}$ $h_{2,2}$ $h_{2,3}$
$h_{3,1}$ $h_{3,2}$ $h_{3,3}$

S_0 → S_1

c_1 y_0

[START]

Một

y_1

Resnet152 + Attention+ LSTM

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$

$a_{t,:,:} = \text{softmax}(e_{t,:,:})$

$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

Resnet152

$e_{1,1,1}$ $e_{1,1,2}$ $e_{1,1,3}$
$e_{1,2,1}$ $e_{1,2,2}$ $e_{1,2,3}$
$e_{1,3,1}$ $e_{1,3,2}$ $e_{1,3,3}$

$a_{1,1,1}$ $a_{1,1,2}$ $a_{1,1,3}$
$a_{1,2,1}$ $a_{1,2,2}$ $a_{1,2,3}$
$a_{1,3,1}$ $a_{1,3,2}$ $a_{1,3,3}$

$h_{1,1}$ $h_{1,2}$ $h_{1,3}$
$h_{2,1}$ $h_{2,2}$ $h_{2,3}$
$h_{3,1}$ $h_{3,2}$ $h_{3,3}$

S_0 S_1 S_2

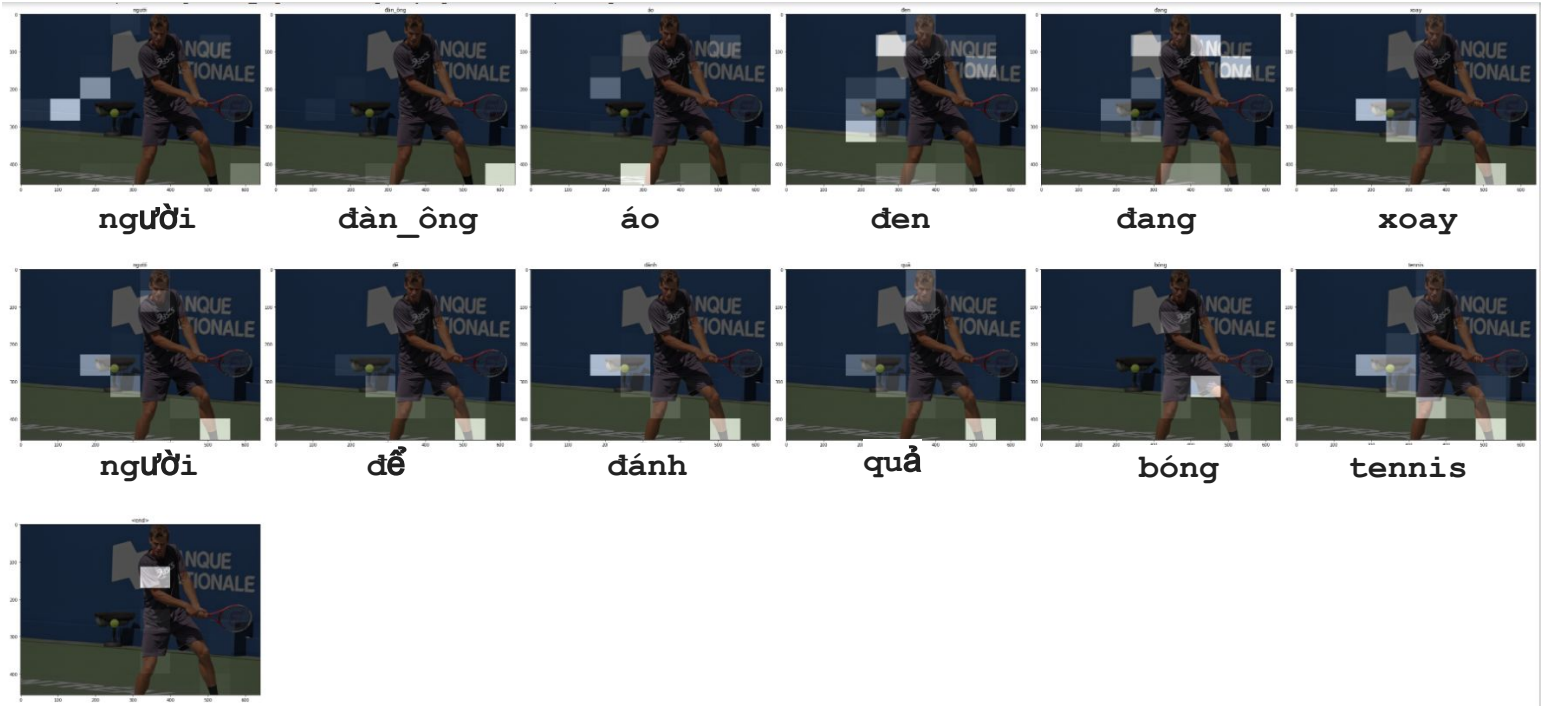
y_1 y_2

c_1 y_0 c_2 y_1

Một người

[START] Một

người đàn_ông áo đen đang xoay người để đánh quả bóng tennis <end>



IV. RESULT

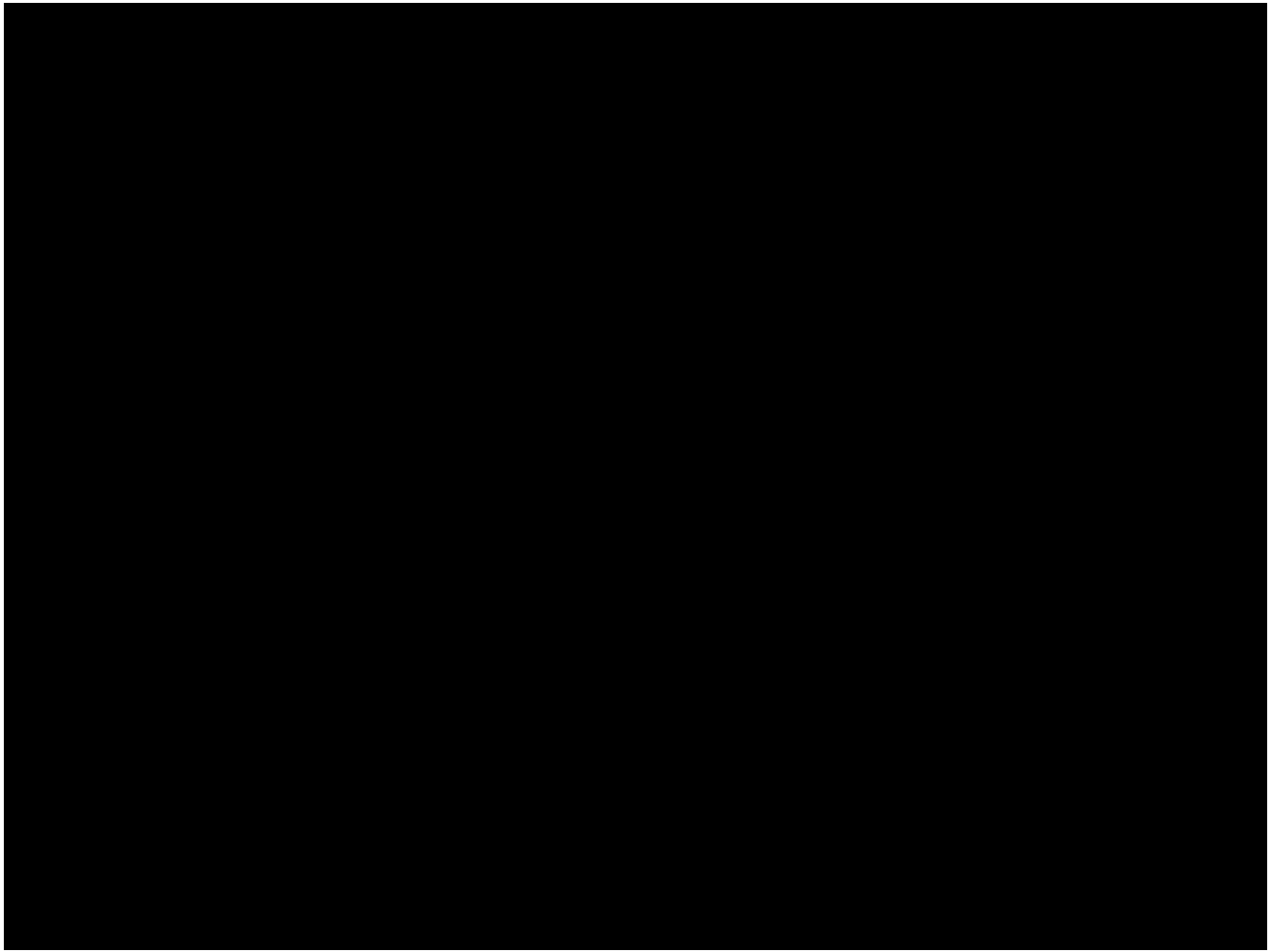
TN	Encoder	Decoder	Att?	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
1	InceptionV3	GRU	x	0.563	0.438	0.344	0.270	0.483	0.645
2	EfficientNetB4	GRU	x	0.571	0.455	0.366	0.296	0.504	0.662
3	Resnet152	GRU	x	0.630	0.517	0.426	0.353	0.541	0.762
4	Resnet152	LSTM	x	0.551	0.433	0.340	0.266	0.486	0.586
5	Resnet152	LSTM	-	**0.669**	**0.544**	**0.452**	**0.380**	**0.575**	**1.031**

Hình: Một số kết quả thực nghiệm

Dataset	Tokenizer	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
English-sportball	nltk	**0.689**	0.501	0.355	0.252	0.585	0.667
GT-sportball	PyVI	0.643	0.481	0.368	0.281	0.565	0.567
UIT-ViIC	PyVI	0.682	**0.561**	**0.411**	**0.327**	0.599	**0.818**

Hình: Kết quả của nhóm tác giả (UIT)

DEMO




THANK YOU