Automated Testing of Image Captioning Systems

Boxi Yu
CHUKSZ, China

Zhiqing Zhong
SCUT, China

Xinran Qin
SCUT, China

Jiayi Yao
CHUKSZ, China

Yuancheng Wang
CHUKSZ, China

Pinjia He
CHUKSZ, China






香港中文大學(深圳) The Chinese University of Hong Kong, Shenzhen
数据科学学院 School of Data Science

Application: remote sensing image caption (examples from Arcgis)

a bridge connects the two sides of the river

The gym is red with green grass

A church near a building and a river

Application: automatically tag the photos

Man in black shirt is playing guitar.

Construction worker in orange safety vest is working on road.

Application: help visually impaired people understand their surroundin

Three people walking on the sidewalk with one of them carrying and using an umbrella

There is very colorful bus coming up the street

Modern image captioning systems

- Microsoft Azure Cognitive Services
- IBM-MAX-Image-Caption-Generator
- VINVL (Revisiting Visual Representations in Vision-Language Models)
- OSCAR (Object-Semantics Aligned Pre-training)

A typical captioning error

A bird sitting on a bench

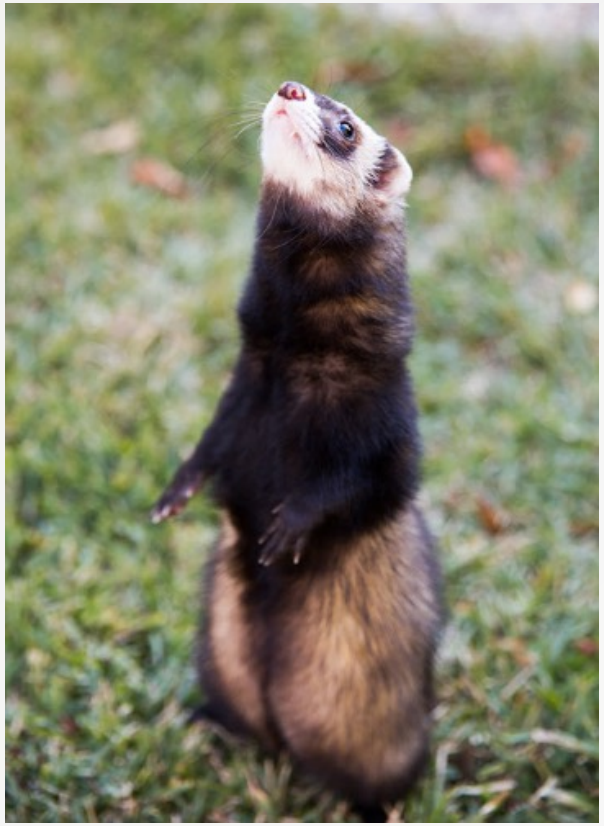
↓

Microsoft Azure Cognitive Services

↓

a person in a garment sitting on a bench

News that image captioning systems failed.



This, AI will have you know, is a squirrel.

Captionbot thinks Michelle Obama is a cell phone.

Effects caused by captioning errors

- For visually impaired people: threat to safety
- For people who use it to tag the images: bad experience
- For engineers who use it for remote sensing image caption: erroneous information

How man many captioning errors do we find by MetaIC?

Using **1000** seed images, we found **17,380** erroneous captions in *Microsoft Azure cognitive services & Several image captioning systems*

MetaIC

Idea: the object names should exhibit directional changes after object insertion.

MetaIC

Insert a cat

A zebra standing in the field

A zebra standing in the field with a cat

Different overlapping ratios

$ratio_0 = 0\%$

$ratio_1 = 15\%$

$ratio_2 = 30\%$

$ratio_4 = 45\%$

MetaIC overview

Object Source Images

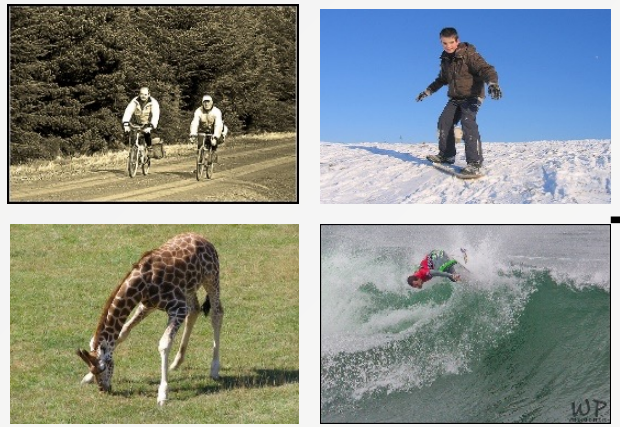
Object Extraction

Object Pool

Randomly Select

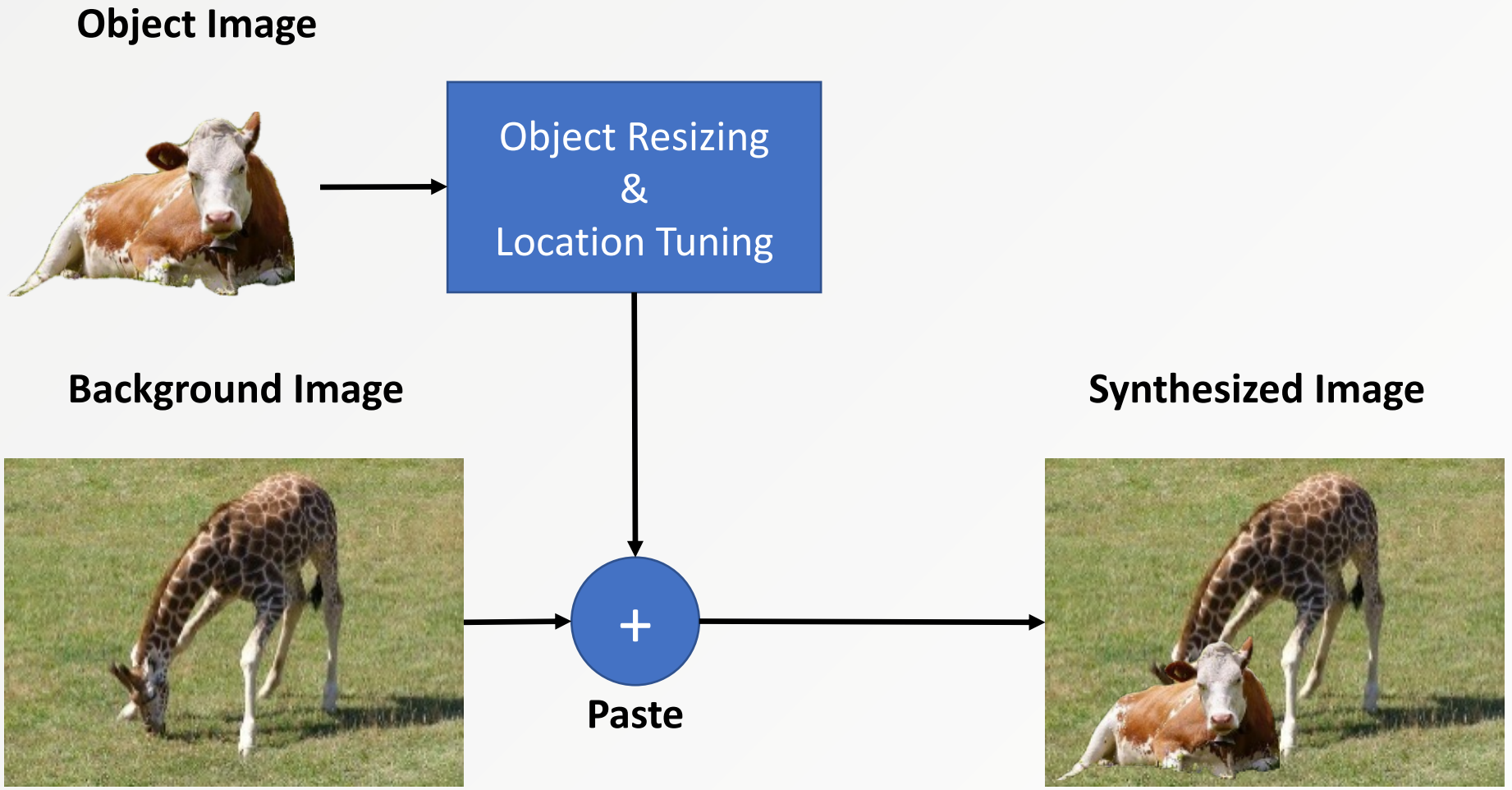
Object Image

Background Images



Randomly Select

MetaIC overview



MetaIC overview

Background Images

Synthesized Images

Caption Collection

Caption pair

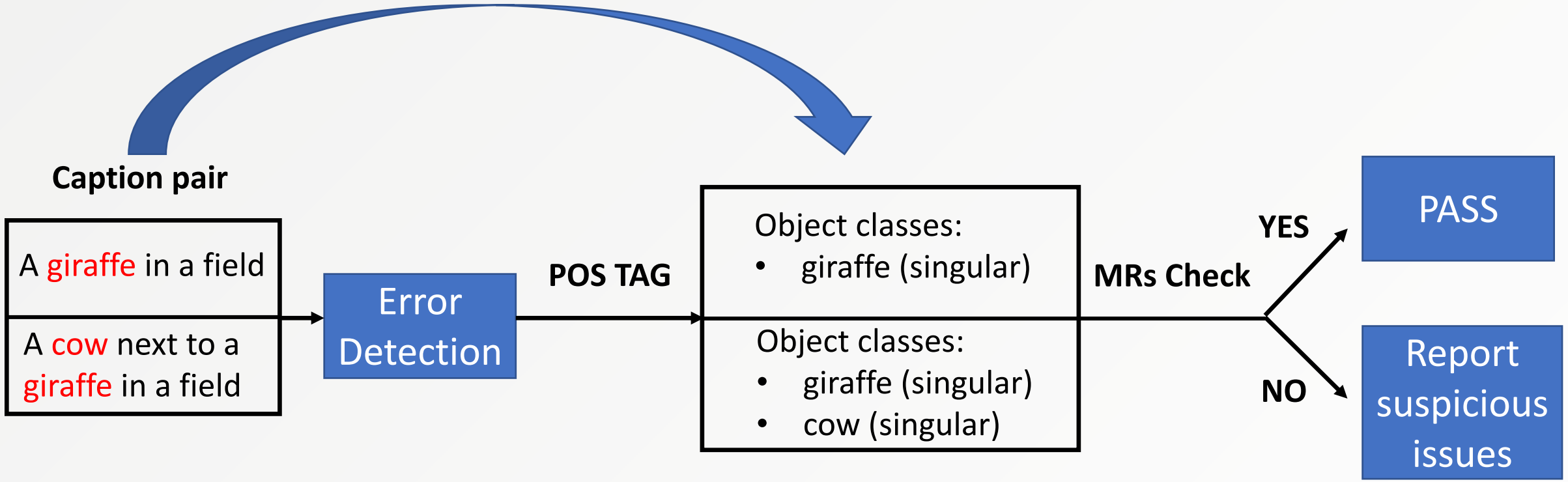
A giraffe in a field

A cow next to a giraffe in a field

Error Detection

MetaIC overview

Turn natural language sentences into categories



Challenges: why object resizing?

With object resizing

Without object resizing

Challenges: why location tuning?

With location tuning

Without location tuning

Challenges: difficult to verify the correctness of the captions!

Deeptest (Brightness)

Add Brightness

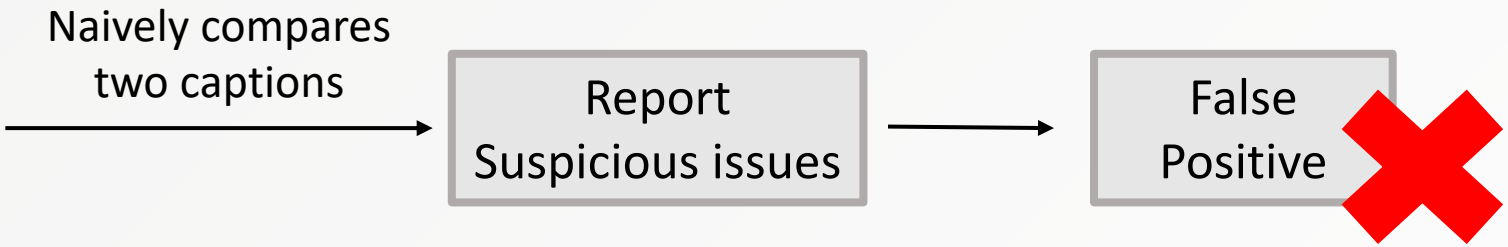
a man holding a skateboard next to a skate park.

a man holding a skateboard in front of a skate park.

How naïve comparison gives false positive?

Original image: a man holding a skateboard next to a skate park.

Synthesized image: a man holding a skateboard next to a skate park.



The reason why naïve comparison failed?

An image may have multiple correct caption sentences!

- a man and his dog playing frisbee in the snow ✓
- a dog leaping in the air to grab a frisbee from its owner on a snowy day. ✓
- a man tossing a frisbee to a brown dog. ✓
- a man playing frisbee with a dog in the snow. ✓
- a man holds up a frisbee and a dog jumps for it in a snowy clearing. ✓

MetaIC with Pos Tagging to solve the test oracle problem

Naive comparison:
compares the words in caption sentences **one by one**

MetaIC with POS Tagging:
compares the **key objects** in the caption sentences

Error type: Wrong singular-plural form

a couple of **dogs** running on the beach.

a couple of **dogs** running on the beach with two **cows**.

Error type: Misclassifying the object

a group of **elephants** walking in a field

a group of **elephants** with a **person** in a garment

Error type: Omission of the object

a herd of **zebra** standing on top of a lush green field.

a herd of **zebra** standing on top of a lush green field.

Evaluation: precision

IC Systems	Deeptest (Perturbation Method)	MetaIC (Overlapping Ratio)
	Blur	0%
$VinVL_L$	21.8% (131/602)	86.7% (535/617)
Microsoft Azure API	39.8% (181/455)	96.6% (858/888)

Compare with Baseline

IC Systems	Deeptest (Perturbation Method)				MetaIC (Overlapping Ratio)			
	Blur	Brightness	Contrast	Shear	0%	15%	30%	45%
Attention	35.0% (185/528)	37.9% (120/317)	37.2% (196/527)	41.1% (245/596)	98.0% (948/967)	97.7% (937/959)	98.4% (948/963)	98.2% (948/965)
$Oscar_B$	20.2% (127/630)	14.7% (38/258)	18.4% (96/521)	21.4% (119/555)	91.3% (652/714)	91.4% (637/697)	91.2% (667/731)	92.2% (694/753)
$Oscar_L$	19.8% (121/610)	12.9% (36/279)	17.4% (91/522)	18.5% (100/542)	92.3% (624/676)	91.7% (620/676)	91.2% (625/685)	91.6% (647/706)
$VinVL_B$	34.2% (207/606)	26.2% (113/431)	29.1% (167/574)	28.6% (185/646)	88.0% (563/640)	87.3% (552/632)	88.4% (571/646)	88.5% (598/676)
$VinVL_L$	21.8% (131/602)	16.5% (60/363)	16.9% (98/579)	19.3% (113/586)	86.7% (535/617)	86.0% (535/622)	84.9% (535/630)	85.1% (560/658)
Microsoft Azure API	39.8% (181/455)	41.2% (56/136)	41.2% (163/396)	38.6% (197/511)	96.6% (858/888)	96.1% (852/887)	96.5% (859/890)	97.4% (860/883)

Case Study on IC Errors via Visualization

$\mathbf{bg}$	$ratio_0$	$ratio_1$	$ratio_2$	$ratio_3$
				
A black **bear** standing in a grassy field.	A black **bear** and a baby black **bear** in a field.	A black **bear** and a baby black **bear**.	Two black **sheep** grazing in a field of grass.	A couple of black **horses** grazing in a field.

bear

bear(1)

bear(2)

bear(1)

bear(2)

sheep

horses

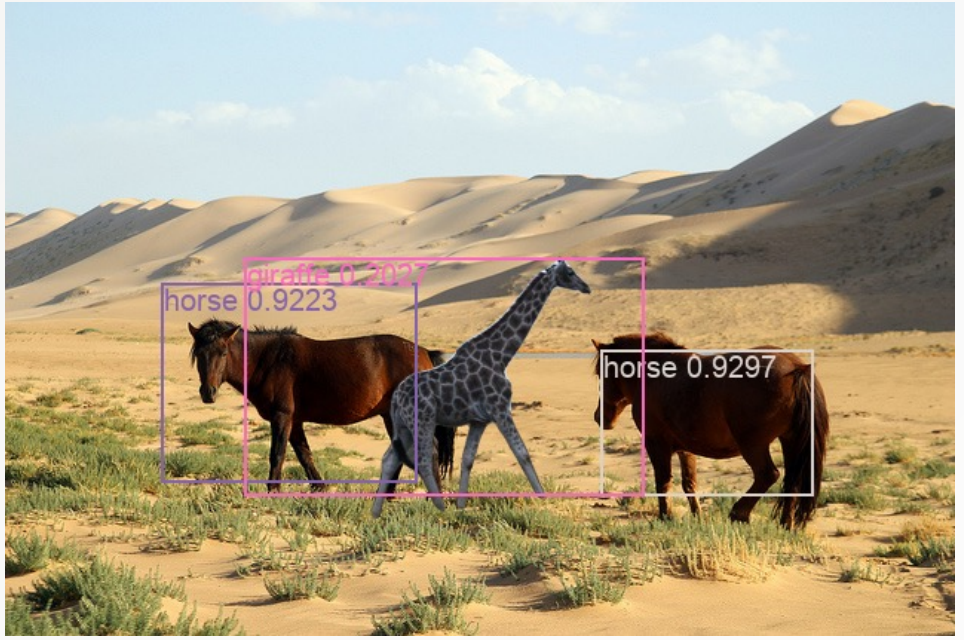
Visualization of the attention for each generated word (Show, Attend and Tell)

Case Study on IC Errors via Visualization

Background image

(a) a couple of **horses** standing in the grass in a field.

Synthesized image

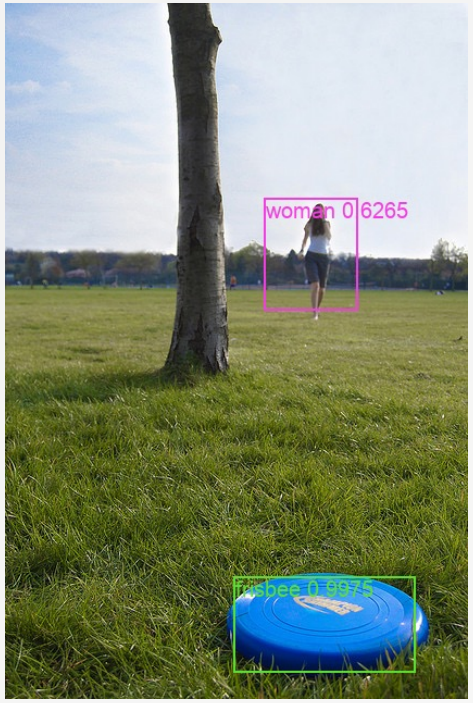


(b) a couple of **giraffes** and a **horse** walking in a field.

Errors caused by the detection component (OSCAR model)

Case Study on IC Errors via Visualization

Background image



a person standing in a field with a blue frisbee.

Synthesized image

(c) a woman standing in a field with a blue frisbee.

Errors caused by the language component (Oscar model)

Finding Labeling Errors in the Training Corpus

Idea: what doesn't change during the metamorphic procedure should be the important feature of the image, thus they should appear in ground truth Labels

A zebra standing in the field

A zebra standing in the field with a cat

The description of the zebra is the essential part in the image pairs

How many label errors we find?

Results: Finding 151 label errors in 6,662 captioned images in coco caption dataset

Examples of label errors in coco caption datase

A **at** licking its lips in a pantry.

a **dig** with a red freeze be walking in some grass.

A young girl who is throwing a piece of **pizza**.

a **man** that is sitting by a window staring out the window

There is no image here to provide a caption for.

a **camel** walking on a beach towards the water