CLIC Lab

FOIL it! Find One mismatch between Image and Language caption

ACL, Vancouver, 31st July, 2017

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich,
Aurelie Herbelot, Moin Nabi, Enver Sangineto, Raffaella Bernardi
{firstname.lastname}@unitn.it
https://foilunitn.github.io

Overview

Research Question

- Do Language and Vision models genuinely integrate both modalities, plus their interaction?

Overview

Research Question

- Do Language and Vision models genuinely integrate both modalities, plus their interaction?
  - Image Captioning

- People riding bicycles down the road approaching a pigeon.
- A group of people on bicycles coming down a street

Image Captioning

Overview

Research Question

- Do Language and Vision models genuinely integrate both modalities, plus their interaction?
    - Visual Question Answering

- Question: How many people are riding a bicycle?
- Answer: three

Visual Question Answering

Overview

Research Question

- Do Language and Vision models genuinely integrate both modalities, plus their interaction?

Our contribution

- FOIL dataset and tasks as a (challenging) benchmark for SoA models

Take-home

- Current models fail in deeply integrating the two modalities

Related Work

- Binary Forced-Choice Tasks (Hodosh and Hockenmaier, 2016)
  - given two captions, original & distractor, an image captioning model has to pick one
  - model fails to pick the original caption
  - limitations
    - hard to pinpoint the reason for the model failure: due to multiple word change simultaneously
    - easier problem: due to selection between two captions

Micah Hodosh and Julia Hockenmaier "*Focused Evaluation for Image Description with Binary Forced-Choice Tasks.*" VL (ACL)' 16

Related Work

- CLEVR Dataset (Johnson et al., 2016)
  - artificial dataset to evaluate visual reasoning
  - analysed shortcoming of VQA models

  - limitations
    - task specific model achieves super human performance (Santoro et al., 2017)
    - some questions are hard to answer by human’s

Johnson et al. “*CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning.*” CVPR, 2017
Santoro et al. “*A simple neural network module for relational reasoning.*” Arxiv, 2017

Motivation

- Need of automatically generate resource with less effort
- Need tasks such that automatic and human evaluation have the same metric
- Need of diagnostics way to evaluate limitations of SoA models

FOIL Dataset

- For a given image and original captions, generate foil captions by replacing one NOUN in the original caption

A person on bike going through green light with red **bus** nearby in a sunny day.

Original Caption

Target Word : bus
Foil Word : truck
Target - Foil pair = bus - truck

A person on bike going through green light with red **truck** nearby in a sunny day.

Generated Foil Caption

FOIL Dataset

- For a given image and original captions, generate foil captions by replacing one NOUN in the original caption

- Original caption based on the MS-COCO (Lin et al., 2014) dataset for image and caption

- Target-Foil pair creation based on MS-COCO object super-category
  - replace objects within same super-category with each other
    - e.g. cat-dog, car-truck etc

Tsung-Yi Lin et al. "*Microsoft coco: Common objects in context.*" ECCV, 2014

FOIL Dataset : Criteria

- Foil not present
  - perform replacement only if the ‘foil’ word is not present
- Salient Target
  - replace a ‘target’ word only if it is visually salient
- Mining hardest foil caption
  - by using ‘neuraltalk’ (Karpathy and Fei-Fei, 2015) loss

Andrej Karpathy and Fei-Fei Li “*Deep Visual-Semantic Alignments for Generating Image Descriptions.*” CVPR, 2015

FOIL Dataset : Sample

- Sample Generated Example

1. An orange cat hiding on the wheel of a red **car**.
2. A **cat** sitting on a wheel of a vehicle.

Original Caption

1. An orange cat hiding on the wheel of a red **boat**.
2. A **dog** sitting on a wheel of a vehicle.

Generated Foil Captions

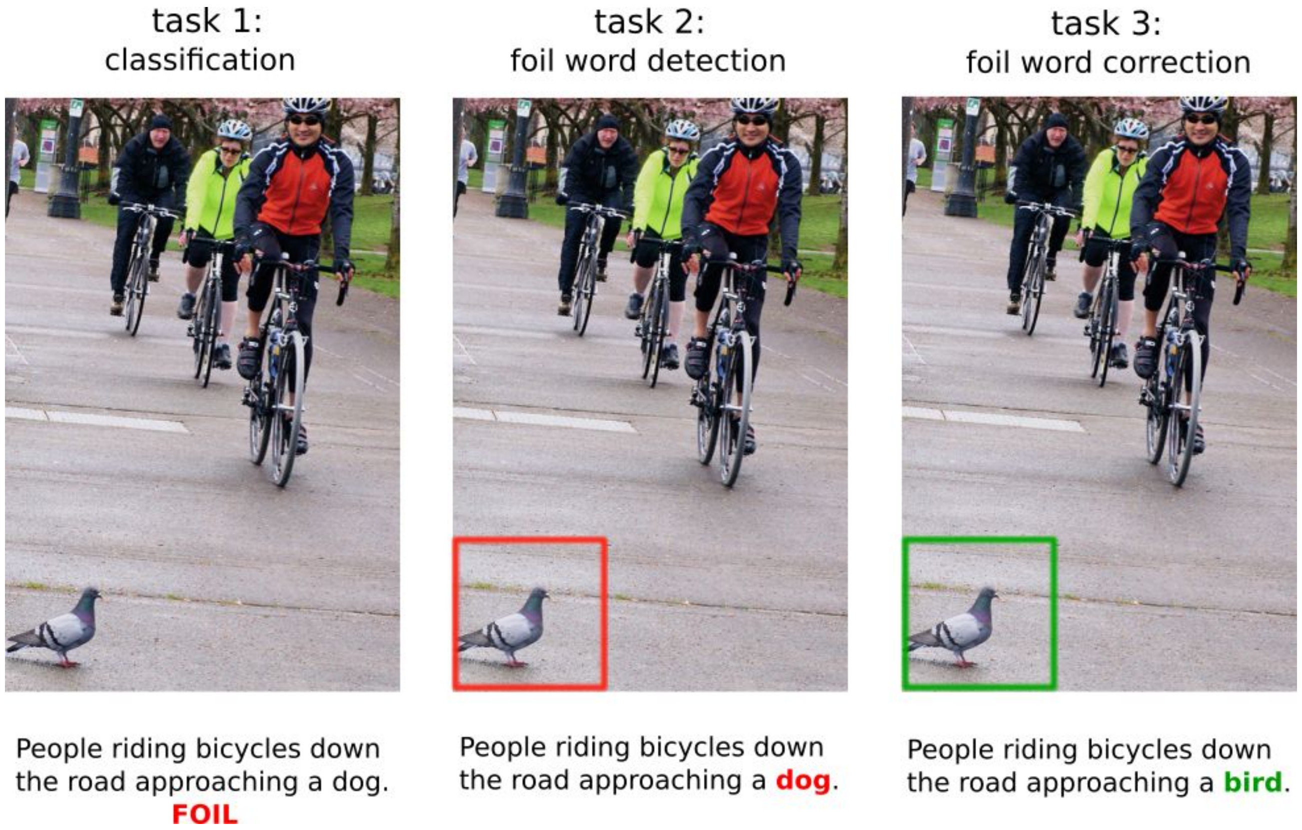
FOIL Dataset : Composition

- Composition of FOIL-COCO dataset

	# datapoints	# images	# captions	# target-foil pairs
Train	197,788	65,697	395,576	256
Test	99,480	32,150	198,960	216

FOIL Dataset : Proposed Tasks

- Task 1 : Binary classification : Original or Foil
- Task 2 : Foil word detection
- Task 3 : Foil word correction



Proposed Tasks : Task 1

- Binary classification: Original or Foil
  - Given an image and a caption decide original or foil caption

People riding bicycles down the road approaching a bird.

Original Caption

People riding bicycles down the road approaching a dog.

Foil Caption

Proposed Tasks : Task 1

- Binary classification: Original or Foil
  - Given a image and a caption decide original or foil caption

People riding bicycles down the road approaching a bird.

Original Caption

People riding bicycles down the road approaching a dog.

Foil Caption

Human performance (AMT)

- **Majority (2/3) : 92.89**
- **Unanimity (3/3) : 76.32**

Proposed Tasks : Task 2

- Foil word detection
  - Given an image and a ‘foil’ caption identify the ‘foil’ word

People riding bicycles down the road approaching a dog..

Where is the mistake in caption?

People riding bicycles down the road approaching a dog..

Proposed Tasks : Task 2

- Foil word detection
  - Given an image and a ‘foil’ caption identify the ‘foil’ word

Human performance (AMT)
- **Majority (2/3) : 97.00**
- **Unanimity (3/3) : 73.60**

People riding bicycles down the road approaching a dog..

Where is the mistake in caption?

People riding bicycles down the road approaching a dog..

Proposed Tasks : Task 3

- Foil word correction
    - Given an image, a ‘foil’ caption and ‘foil’ word location, correct the ‘foil’ caption

People riding bicycles down the road approaching a dog..

Can you correct the mistake?

People riding bicycles down the road approaching a bird..

FOIL Dataset : is NOT Equal to

- Visual Question Answering

- In VQA, answers are highly dependent on the (linguistic) context of the question.

What man is riding?

$\neq$

A person on motorcycle going through green light with red bus nearby in a sunny day.

- In FOIL, we are asked a context independent fine-grained information about the image.

FOIL Dataset : is NOT Equal to

- Object Classification/Detection

- In computer vision tasks, generally question is, what objects are present in the image

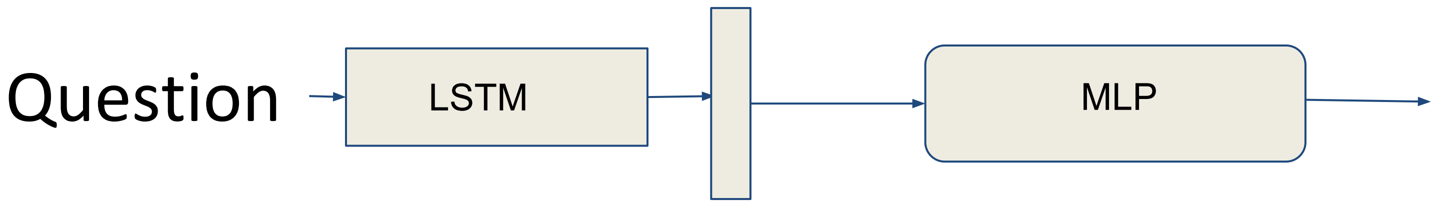
- In FOIL, question is "what object is NOT in the image (foil classification/detection) and understand what object is there based on the context(correction)?"

Models Tested

- VQA Models
- Image Captioning Model

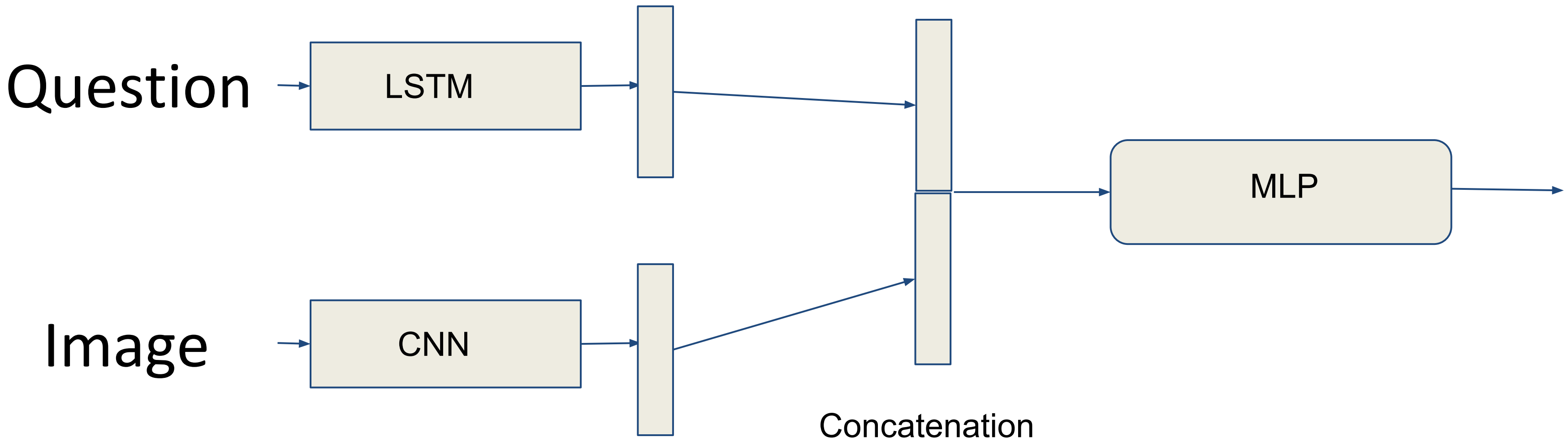
Models Tested

- Baseline Models
  - Language Only (Blind)
    - LSTM (Question) followed by MLP



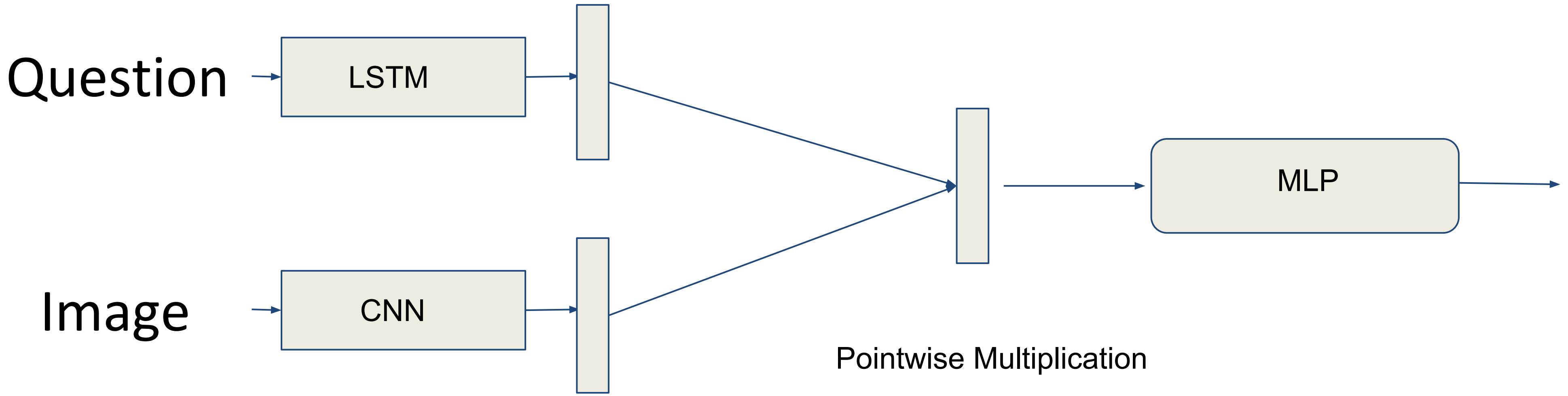
Models Tested

- Baseline Models
  - Language Only (Blind)
  - CNN + LSTM (Zhou et al., 2015)
    - CNN (Image), LSTM (Question) joined by concatenation followed by MLP



Models Tested

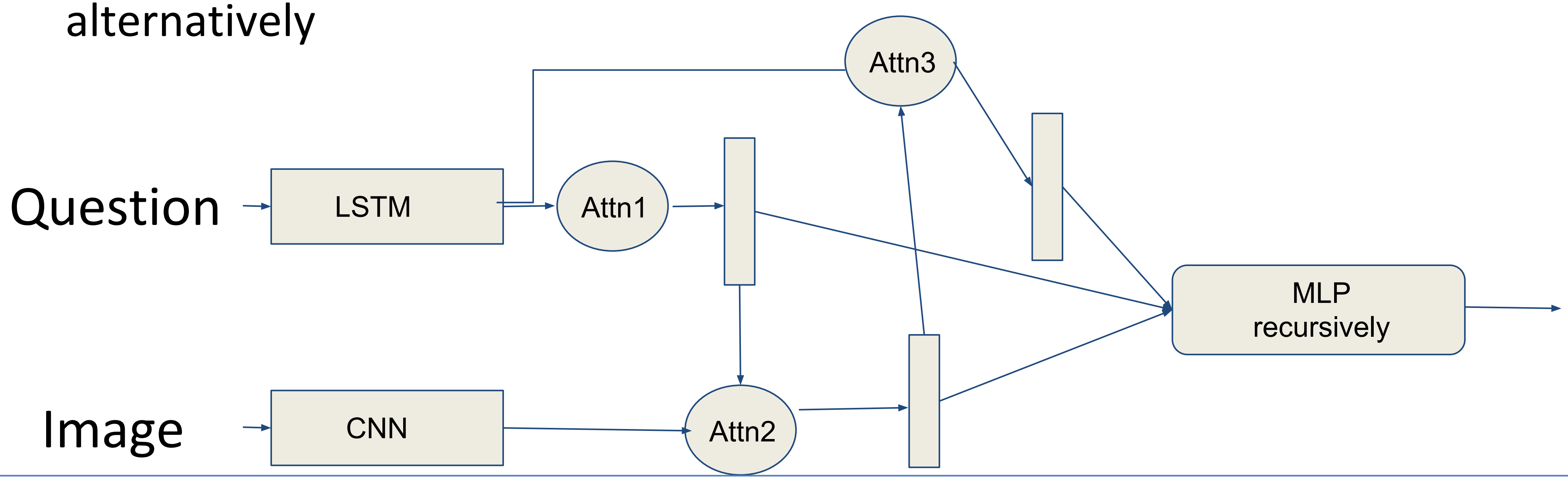
- VQA Models
  - LSTM + norm I (Antol et al., 2015)
    - CNN (Image), LSTM (Question) joined by pointwise multiplication followed by MLP



Antol et al. “*VQA: Visual Question Answering.*” ICCV, 2015

Models Tested

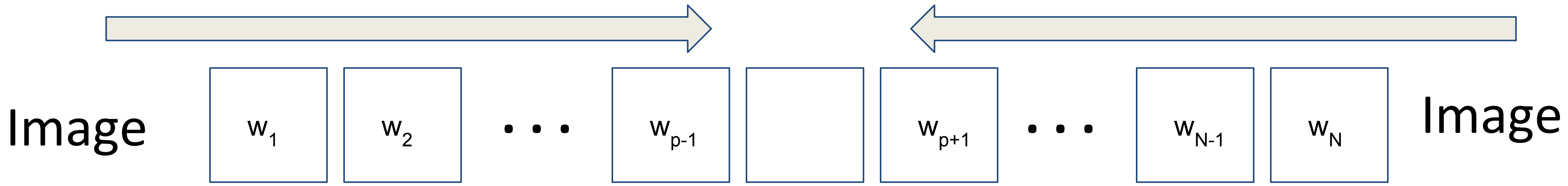
- VQA Models
  - LSTM + norm I (Antol et al., 2015)
  - Hierarchical Co-attention (HieCoAttn) (Lu et al., 2016)
    - CNN (Image), LSTM (Question), both Image & Question is co-attended in alternatively



Lu et al. "*Hierarchical Question-Image Co-Attention for Visual Question Answering.*" NIPS, 2016

Models Tested

- Image Captioning Model
    - Bi-directional IC Model (IC-Wang) (Wang et al., 2016)
        - Given Image, and past and future context model predicts current word



Results

- Task 1 : Binary Classification

	Overall	**Correct**	**Foil**
Blind	55.62	86.20	25.04
CNN + LSTM	61.07	89.16	32.98
LSTM + norm I	63.26	**92.02**	34.51
HieCoAttn	**64.14**	91.89	**36.38**
IC-Wang	42.21	38.98	45.44
Human (Majority)	92.89	91.24	94.52
Human (Unanimity)	76.32	73.73	78.90

Results

- Task 2 : Foil word detection

	Only Nouns	All Words
Chance	23.25	15.87
LSTM + norm I	26.32	24.25
HieCoAttn	**38.79**	**33.69**
IC-Wang	27.59	23.32
Human (Majority)	_	97.00
Human (Unanimity)	_	73.60

Results

- Task 3 : Foil word correction

	All Target Words
Chance	1.38
LSTM + norm I	4.7
HieCoAttn	4.21
IC-Wang	**22.16**

Conclusion

- Created a challenging dataset and corresponding challenging tasks
  - used to evaluate limitations of language and vision models
  - can be extended to other part of speech (see Shekhar et al., 2017), scene etc
  - by knowing source of error, will help in designing better models

- Need fine-grained joint understanding of language and vision

Shekhar et al. "*Vision and Language Integration: Moving beyond Objects.*" IWCS, 2017

Thank You !!!
Q & A







Dataset https://foilunitn.github.io

Crowdflower

- **Read** and **understand** the caption and carefully **watch** the image
- **Determine** if the caption provides a correct description of what is depicted in the image
- If you judge the caption as "wrong", you will be asked to type **the word** that makes the caption incorrect

Crowdflower

Caption:
a man riding a bull through part of a parking lot

Does the caption provide a correct or wrong description of the image? (required)

- correct
- wrong

Crowdflower

Caption:
a man riding a bull through part of a parking lot

Does the caption provide a correct or wrong description of the image? (required)

- [x] correct
- [ ] wrong

Crowdflower

Caption:
a man riding a bull through part of a parking lot

Does the caption provide a correct or wrong description of the image? (required)

- [ ] correct
- [x] wrong

Type the wrong word (one word) (required)

FOIL Dataset : Criteria

- Foil not present
- Salient Target

FOIL Dataset : Criteria

- Foil not present
  - Perform replacement only if ‘Foil’ word is not present in the image
    - Check that ‘Foil’ word is not used by any other ms-coco annotator

For e.g.,

I. “A **boy** is running on the beach”
II. “A boy and a little girl are playing on the beach”

- Target - Foil = Boy - Girl

FOIL Dataset : Criteria

- Salient Target
  - Replace 'Target' words only if it is visually salient in the image
    - Based on annotator agreement i.e. more than one annotator used 'Target' word

For e.g.,

I. Two **zebras** standing in the grass near rocks.
II. Two **zebras** grazing together near rocks in their enclosure.
III. Two **Zebras** are standing near some rocks.
IV. two **zebras** in a field near one another
V. A grassy area shows artificially arranged rocks and two **zebras**, as well as part of the lower half of a **deer**.

- Target - Foil = Zebra - Dog (Used)
- Target - Foil = Deer - Dog (Not Used)

FOIL Dataset : Mining Hardest Foil Caption

- To eliminate visual-language bias
  For every original caption could produce one or more foil caption

- Neuraltalk loss is used to mine hardest foil caption
  Eliminates both visual and language bias