CoSIm

Sushma Akoju
Prof. Eduardo Blanco

CoSIm: Commonsense Reasoning for Counterfactual Scene Imagination

Target Questions to answer for Discussion

CoSIm: Commonsense Reasoning for Counterfactual Scene Imagination

1. What is a counterfactual?
2. What make a task / dataset multimodal?
3. “We collect 3.5K high-quality and challenging data instances,” How can you show that an instance is high-quality? And challenging?
4. Is it fair to characterize their dataset as a made-up challenge?
5. Do their experiments do anything in addition to traditional supervised learning?
6. You also read “Temporal Common Sense Acquisition with Minimal Supervision.” Can you think of reusing some ideas about minimal supervision to improve the models for this problem?

What is a counterfactual?

What is a counterfactual?

In Causal Inference:

counterfactual is that we do not observe from one of the potential outcomes.

We always only observe one outcome.

It is the difference between action taken vs action not taken

Example

Let X be a binary indicator if a subject eats breakfast in the morning

Let Y be a binary indicator that of being healthy

Let C1 be the value of Y if X = 1 i.e. breakfast taken

Let C0 be the value of Y if X = 0 i.e. breakfast **NOT** taken

C1 and C0 are potential outcomes of this experiment

Simulated World - All possible outcomes of this experiment

- 0* is the indicator that C1 did not occur when C0 occured
- 1* is the indicator that C0 did not occur when C1 occured

$$\begin{pmatrix} X & Y & C0 & C1 \\ 0 & 0 & 0 & 0* \\ 0 & 0 & 0 & 0* \\ 0 & 0 & 0 & 0* \\ 0 & 0 & 0 & 0* \\ 1 & 1 & 1* & 1 \\ 1 & 1 & 1* & 1 \\ 1 & 1 & 1* & 1 \\ 1 & 1 & 1* & 1 \end{pmatrix}$$

Remember:

Let C1 be the value of Y if X = 1 i.e. breakfast taken

Let C0 be the value of Y if X = 0 i.e. breakfast NOT taken

How is this relevant to CoSIm?

Capture: unseen changes that never occurred

What makes a task / dataset multimodal?

What makes a task / dataset multimodal?

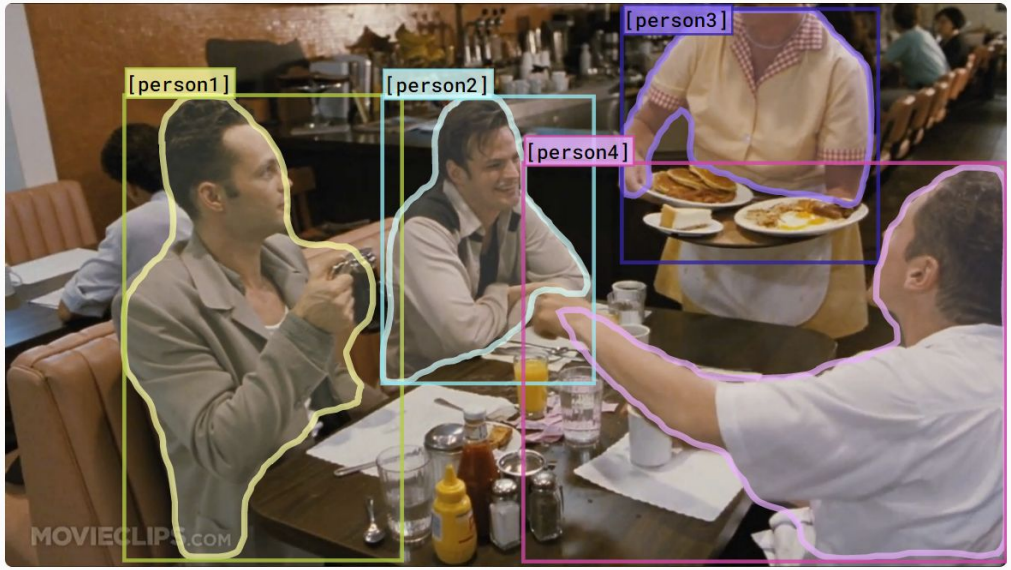
Having an Image and a Natural Language text associated with it is a Multimodal dataset.

Visual Commonsense reasoning:

Given a challenging question about an image, a machine must answer correctly and then provide a rationale justifying its answer.

Reference: From Recognition to Cognition: Visual Commonsense Reasoning - Rowan Zellers, Yonatan Bisk, Ali Farhadi, Yejin Choi

Example Visual Commonsense Reasoning



hide all | show all | [person1] | [person2] | [person3] | [person4]

more objects »

Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

Rationale: I think so because...

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

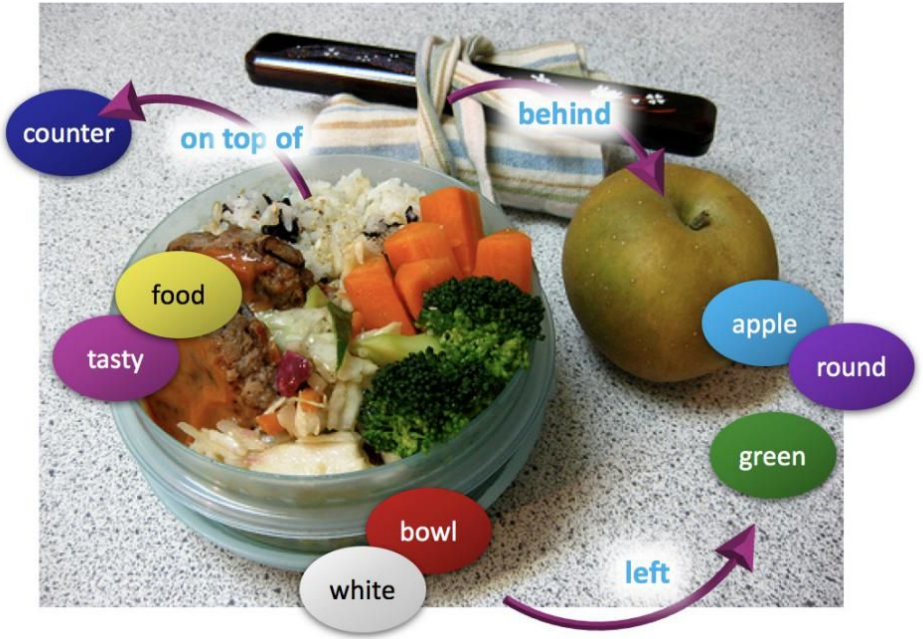
Visual Question Answer (VQA) Dataset

Free form,

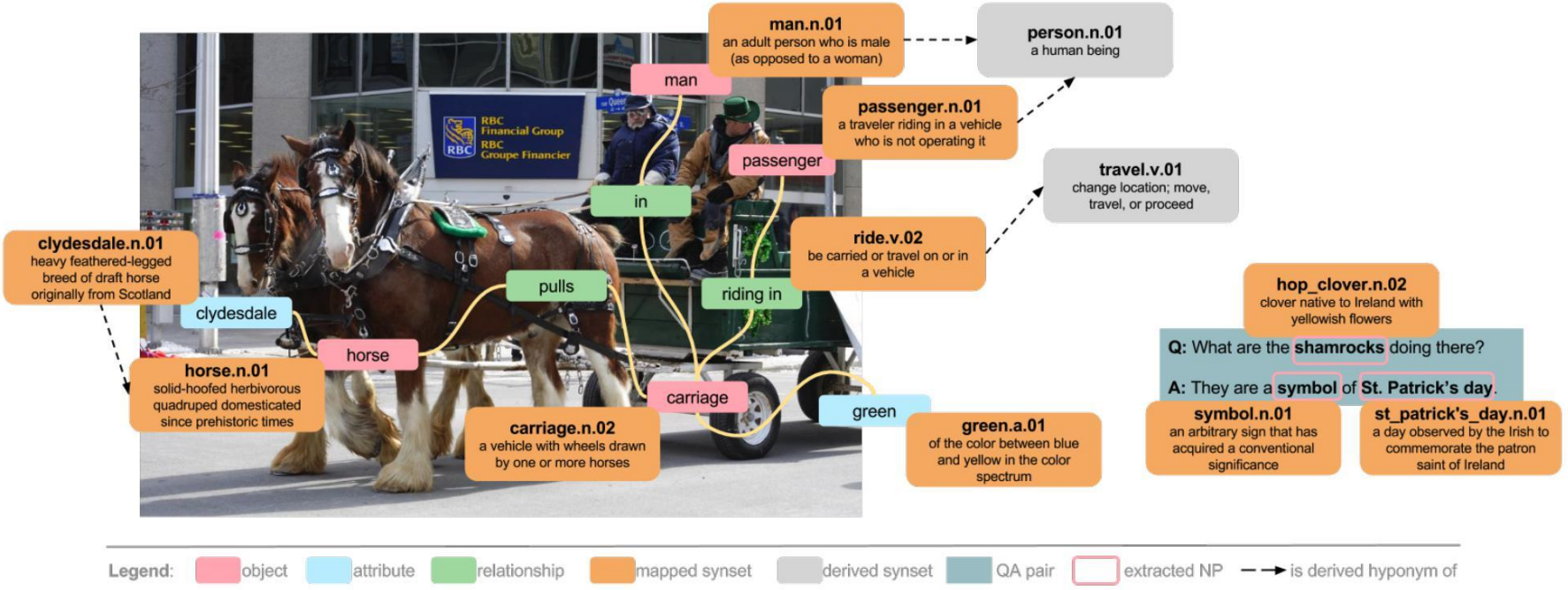
open-ended Visual Question Answering

Goal: to provide an accurate

natural language answer



Visual Genome Dataset : What, Where, When, Who, Why and How



How can you show that an instance is high-quality? And challenging?

To answer this question

Let us find out a bit more about CoSIm

CoSIm Dataset

Question / Initial Response

Is the railway line safe?

Yes because there are safety lights and crosswalk signs

Change

add a train to the tracks.

Answer Choices

New Response: no. although there are safety lights and crossing gates, they don't appear to be working and there is a train coming.

Distractor #1: yes. there are safety lights and crossing gates, they appear to be working and there is a train coming and it will stop.

Distractor #2: yes. although there are safety lights and crossing gates, they don't appear to be working and there is no train coming.

Distractor #3: no. although there are safety lights and crossing gates, there is a power outage and there is a train coming.

Figure 1: Example from our COSIM dataset. An image is associated with an initial commonsense question-response pair, a described counterfactual change to the image, and a new response to the question (randomly shuffled with three human-written distractors).

How can you show that an instance is high-quality? And challenging?

Claim: “We collect 3.5K high-quality and challenging data instances,”

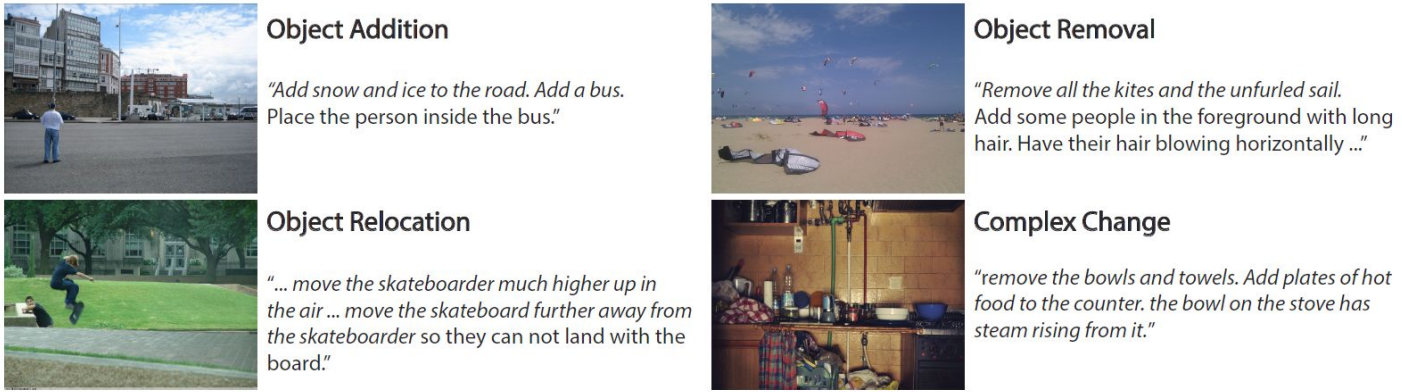


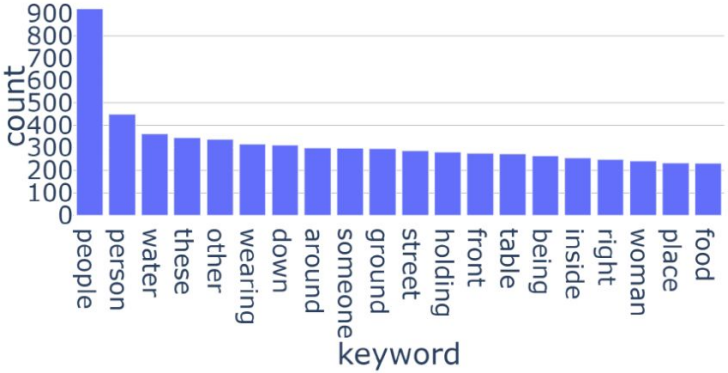
Figure 3: Scene change examples from our CoSIm dataset. The relevant portions of the change are in italics. Complex changes contain three or more changes within them (this example contains Object Removal, Object Addition, Object State Change).

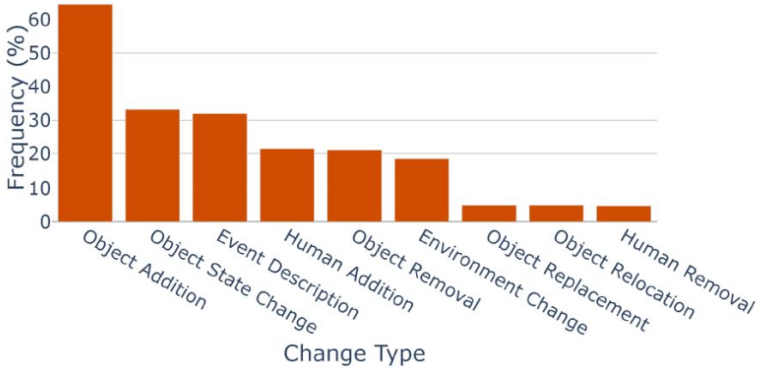
CoSIm

1. image,
2. an initial question-response pair
3. an imagined visual scene change
4. a new response with three distractors

imagined visual scene change : textual description of what to modify in the scene to alter the conditions.

Unique Words & Types of Scene changes in Text





Dataset format

question Q = {q1,q2 ...qn}

initial response Ri = {r1,r2...ri_nr_i

change C = {c1,...c_nc}

new response R_n = {r_n1, ...r_nc}

Length of question: NQ

Length of initial response: NRi

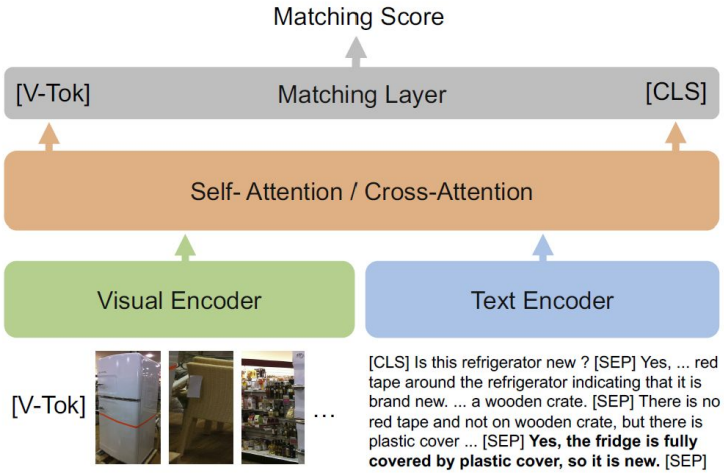
Length of scene change: NC and

Length of new response: NRn

About the features

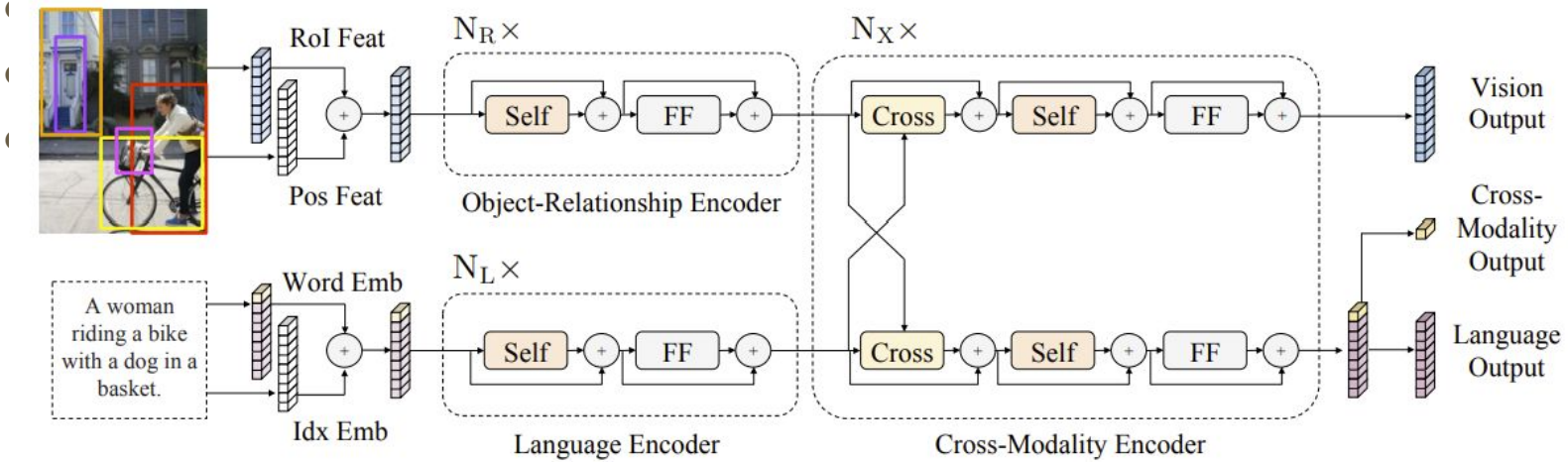
object-level visual features : Faster R-CNN

Textual feature encoding: BERT

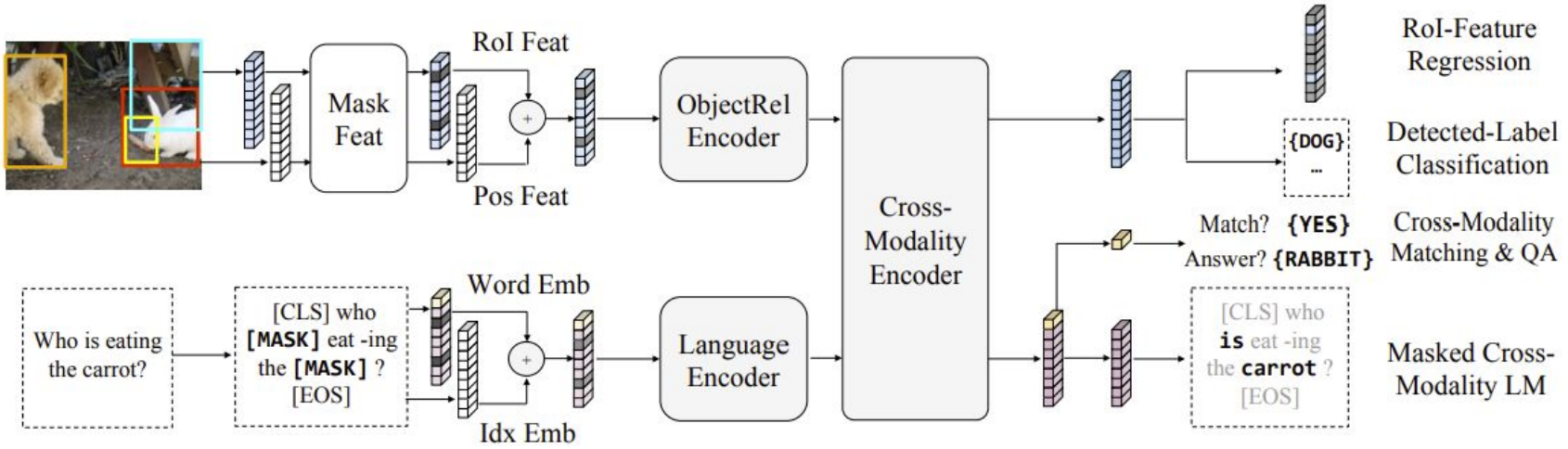


LXMERT

- vision and language feature matching scores via multi-head self-attention layers
- cross-modal attention layers: e cross-attention sub-layer is used to exchange



LXMERT : example



Object encoder: taking the features of detected objects as the embeddings of images

the word w_i and its index i (w_i's absolute position in the sentence) are projected to vectors by embedding sub-layers, and then added to the index-aware word embeddings

Contribution

- Visual Language Transformer and ablation studies.
- Mainly used in scene understanding tasks
- Over video understanding to auto-generate scene descriptions

ple, we can easily anticipate the implications of the sun being overcast by rain clouds (e.g., the street will get wet) and accordingly prepare for that. In this paper, we introduce a new task/dataset called Commonsense Reasoning for **C**ounterfactual **S**cene **Im**agination (COSIM) which is designed to evaluate the ability of AI systems to reason about scene change imagination. In this task/dataset, models are given an image and an initial question-response pair about the image. Next, a counterfactual imagined scene change (in textual form) is applied, and the model has to predict the new response to the initial question based on this scene change. We collect 3.5K high-quality and challenging data instances, with each instance consisting of an image, a commonsense question with a response, a description of a counterfactual change, a new response to the question, and three distractor responses. Our dataset contains various complex scene change types (such as object addition/removal/state change, event description, environment change, etc.) that require models to imagine many different scenarios and reason about the changed scenes. We present a baseline model based on a vision-language Transformer (i.e., LXMERT) and ablation studies. Through human evaluation, we demonstrate a large human-model performance gap, suggesting room for promising future work on this challenging counterfactual, scene imagination task.[1]

How can you show that an instance is high-quality? And challenging?

Seems like lack of sufficient prior works on quality metrics of the data (Scene understanding)

Comparison between other existing datasets

What benchmarking strategies exist to provide insights about quality of the scene understanding task?

How generalizable is the dataset over existing Counterfactual Visual Commonsense reasoning models?

Is it fair to characterize their dataset as a made-up challenge?

Attempt to Create Counterfactual Commonsense data

several diverse types of imagined scene changes

prior work or what foundation may be required to define scene changes?

Attempt to find intersection between: multi-task learning, visual commonsense reasoning with counterfactual (unseen changes that never occurred)

Could have been better?

- Comparison of Prior works towards Visual Commonsense reasoning
- prior works on Counterfactual reasoning for commonsense reasoning
- Could be formulated more simpler terms: such as we just want to provide

  What would the answer be, if system receives a question, but had not extracted Knowledge or seen the Image ?

- The applications of this dataset or such counterfactual commonsense reasoning is not clear
- Since they mentioned scene understanding : https://openaccess.thecvf.com/content_cvpr_2018/html/Ramanishka_Toward_Driving_Scene_CVPR_2018_paper.html

Possible use cases

An App at a Skii resort:

If I had a camera that tells me if I should skii today during my trip to Switzerland?

To check if it is good to skii today or not

How much snow is sufficient to skii?

If there is storm in the same scene, is it safe to skii?

Closed-world assumptions - not included

It is generally considered change in scenarios can cause change in inference using “closed-world assumption” : (nonmonotonicity into logic)

Example:

- know whether there is a flight from London to Porto Alegre on 6 August 2012
- Assume assume that all flights from London to Porto Alegre are listed on this database
- later, a new flight is entered into the database
- Changes earlier conclusion

Reference Book: Neural-Symbolic Cognitive Reasoning, Artur S. d’Avila Garcez, Luís C. Lamb, Dov M. Gabbay

Textual Scene change vs Imagined Scene Changes

Questions like: How many blue objects will be present in this scene?

The examples used in this dataset seem more familiar towards ablation studies but from Closed World Assumptions from Neuro Symbolic reasoning.

Do their experiments do anything in addition to traditional supervised learning?

Implementation

Tested on two models:

1. Full model with LXMERT & Image, Textual context, response
2. Multi task learning with VCR & CoSIm alternating mini-batch
3. Contrastive Learning (with data augmentation)

LXMERT: Learning Cross-Modality Encoder Representations from Transformers , Hao Tan, Mohit Bansal

Claims

Various complex scene change types:

- Object addition/removal/state change, event description,
- environment change

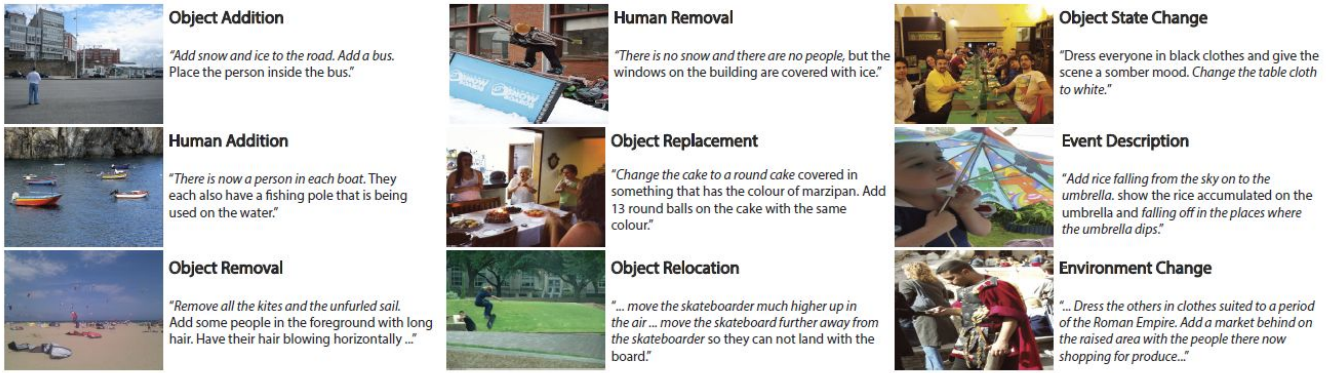


Figure 12: Scene change examples from our dataset. The relevant portions of the change are in italics.

Temporal Common Sense Acquisition with Minimal Supervision

Temporal Common Sense Acquisition with Minimal Supervision

1.
2. **Taco LM - Temporal Common sense Language Model**
3. Commonsense-level understanding
4. If doctor went for a vacation, doctor **will not** see patients
5. But if doctor went for a walk, doctor **will see** the patients soon

Information extraction

Extract entities from text (Information extraction) & classify the sequences for duration, frequency class types

Improvements to CoSIm models using Information extraction

To use specific objects detected

And variations/antonyms of the objects

Using language synonyms, antonyms for entities identified between image and text