Visual Madlibs:
Fill-in-the-blank Description and Question Answering

Licheng Yu, Eunbyung Park, Alexander C. Berg, Tamara L. Berg

http://tamaraberg.com/visualmadlibs

Dataset Overview

The place is a(n) **road**.
When I look at this picture, I feel **free**.
The most interesting aspect of this picture is **the motorcycles**.
One or two seconds before this picture was taken, **they stopped to chat and decided where to go**.
One or two seconds after this picture was taken, **the bikers ride down the road**.

The people are **sitting at** the dining table.

The people are **playing with** the teddy bears.

Person B is **wearing a grey T-shirt**.
Person B is **talking to two other people**.
Person B is **next to an elephant**.

Person C is **a blonde woman**.
Person C is **petting an elephant**.
Person C is **at a zoo**.

The cake is **iced in chocolate**.
The cake is **on the table**.
People could **eat** the cake.

The table is **covered with a white cloth**.
The table is **in a restaurant**.
People could **have a cake at** the table.

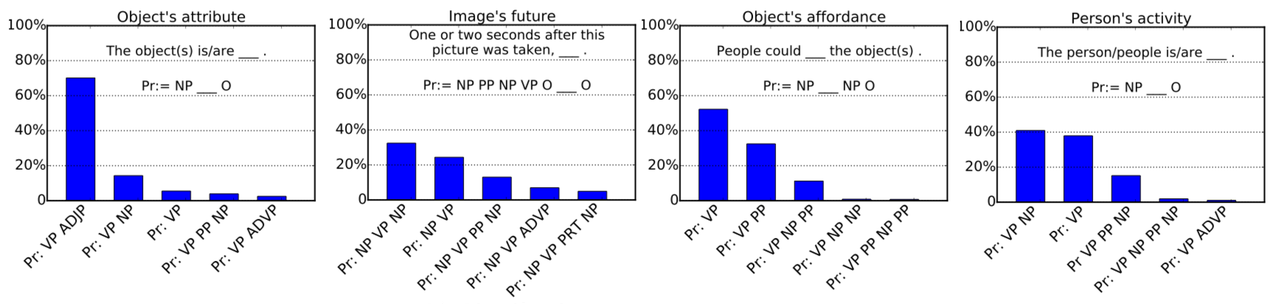
The **Visual Madlibs** dataset is collected using automatically produced fill-in-the-blank templates designed to gather targeted descriptions. Different from previous datasets that contain generic descriptions for the entire image, our dataset has:

- **360,001** focused descriptions for **10,738** images.
- **12 types** of fill-in-the-blank Madlib questions:
  - General scene
  - Emotional content
  - What happened before
  - What happened next
  - The most interesting/unusual content
  - Appearance, activity, and location of **each** person
  - Appearance, affordance, and position of **each** object
  - Interactions between people and objects

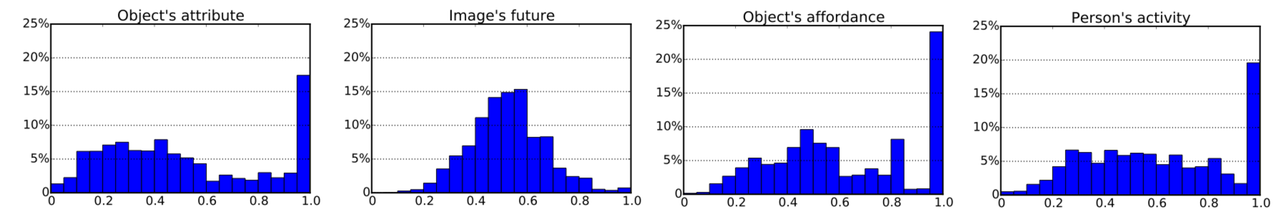
Instruction	Prompt
Describe the type of scene/place shown in this picture.	The place is a(n) ____ .
Describe the emotional content of this picture.	When I look at this picture, I feel ____ .
Describe the most interesting or unusual aspect of this picture.	The most interesting aspect of this picture is ____ .
Describe what happened immediately before this picture was taken.	One or two seconds before this picture was taken, ____ .
Describe what happened immediately after this picture was taken.	One or two seconds after this picture was taken, ____ .
Describe the appearance of the indicated object.	The object(s) is/are ____ .
Describe the function of the indicated object.	People could ____ the object(s).
Describe the position of the indicated object.	The object(s) is/are ____ .
Describe the appearance of the indicated person/people.	The person/people is/are ____ .
Describe the activity of the indicated person/people.	The person/people is/are ____ .
Describe the location of the indicated person/people.	The person/people is/are ____ .
Describe the relationship between the indicated person and object.	The person/people is/are ____ the object(s).

Dataset analysis

1) Structure of Madlibs



2) Similarity among Answers



3) Template parsing

a) person's attribute

The person is an old man in red coat.
- refer name: man + person
- general attribute: old
- affiliate attribute: red
- affiliate name: coat

b) object's attribute

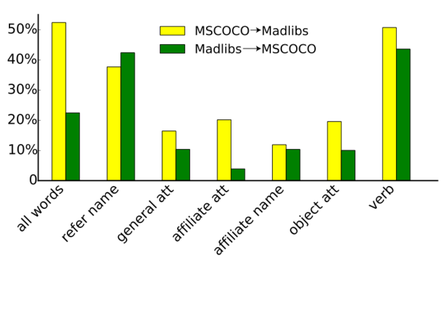
The dog is large.
- object attribute: large

c) person's activity/pair's relationship

The person is running.
The person is waiting on the boat.
- verb: running, waiting

	MSCOCO	Mad Libs
Refer	95.1%	46.5 % (100%)
General att	15.5%	37.3%
Affiliate att	1.8%	16.3%
Affiliate obj	7.4%	29.6%
Verb	79.0%	95.4%
Object att	18.7%	86.8%

4) vs. general description



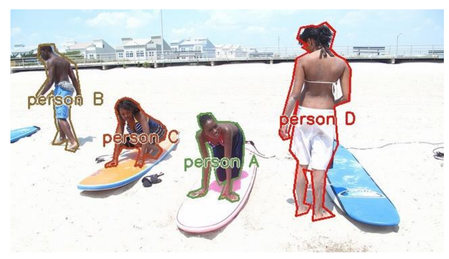
Two Evaluation Tasks

Question-Answering

Choose one answer to describe what happened before this picture was taken.

One or two seconds before this picture was taken, ___.
- ☐ they were eating
- ☑ they poured their wine
- ☐ a bottle was broken
- ☐ a group of people are talking

Focused Description



Describe the activity of the Person D.
Person D is ____.

Describe the appearance of Person B.
Person B is ____.

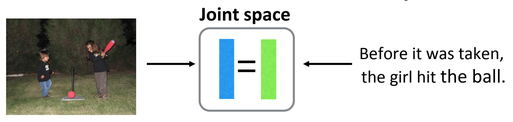
Describe the location of Person A.
Person A is ____.

Experiments

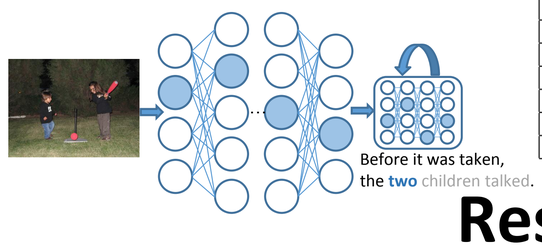
Hard Task

	#Q	n-gram	CCA	nCCA	nCCA (place)	nCCA (bbox)	nCCA (all)	CNN+LSTM (madlibs)	CNN+LSTM(r) (madlibs)	Human
1. scene	6277	22.8%	63.8%	70.1%	70.7%	–	68.2%	63.6%	64.2%	75.6%
2. emotion	5138	25.1%	33.9%	37.2%	38.3%	–	33.2%	34.6%	37.6%	38.4%
3. past	4903	22.4%	47.9%	52.8%	49.5%	–	54.0%	42.2%	39.5%	73.9%
4. future	4658	24.4%	47.5%	54.3%	50.5%	–	53.3%	41.1%	39.5%	75.1%
5. interesting	5095	27.6%	51.4%	53.7%	50.5%	–	55.1%	44.0%	37.1%	76.7%
6. obj attr	7194	29.5%	42.2%	43.6%	41.5%	49.8%	39.3%	41.6%	42.3%	70.5%
7. obj aff	7326	32.2%	54.5%	63.5%	60.9%	63.0%	48.5%	–	69.4%	52.7%
8. obj pos	7290	29.2%	49.0%	55.7%	53.3%	50.7%	53.4%	46.7%	50.2%	70.8%
9. per attr	6651	23.3%	33.9%	38.6%	35.5%	46.1%	31.6%	35.5%	42.4%	70.5%
10. per act	6501	24.0%	59.7%	65.4%	62.6%	65.1%	66.6%	57.3%	53.7%	85.1%
11. per loc	6580	22.3%	56.8%	63.3%	65.5%	57.8%	62.6%	50.4%	56.8%	72.9%
12. pair rel	7595	30.1%	49.4%	54.3%	52.2%	56.5%	52.0%	–	54.6%	74.7%

Canonical Correlation Analysis



CNN+LSTM

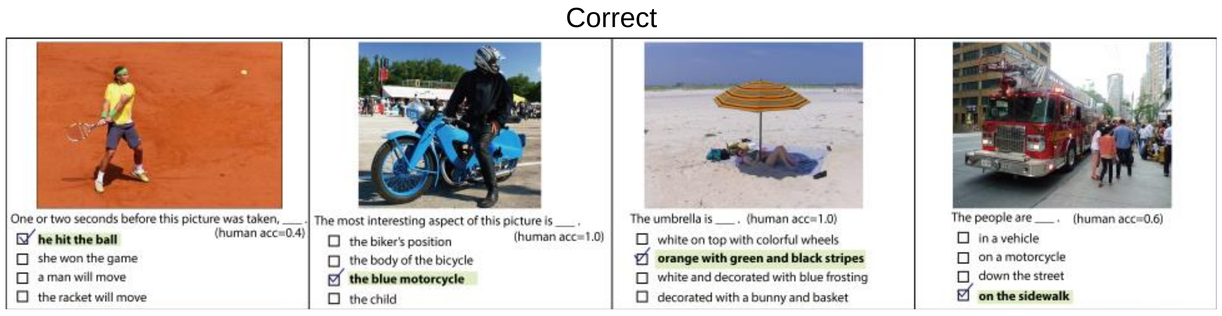


	BLEU-1			BLEU-2		
	nCCA	nCCA (box)	CNN+LSTM (madlibs)	nCCA	nCCA (bbox)	CNN+LSTM (madlibs)
1. scene	0.52	–	0.62	0.17	–	0.19
2. emotion	0.17	–	0.38	0	–	0
3. future	0.38	–	0.39	0.12	–	0.13
4. past	0.39	–	0.42	0.12	–	0.12
5. interesting	0.49	–	0.65	0.14	–	0.22
6. obj attr	0.28	0.36	0.48	0.02	0.02	0.01
7. obj aff	0.56	0.60	–	0.10	0.11	–
8. obj pos	0.53	0.55	0.71	0.24	0.25	0.49
9. per attr	0.26	0.29	0.57	0.06	0.07	0.25
10. per act	0.47	0.41	0.53	0.14	0.11	0.20
11. per loc	0.52	0.46	0.63	0.22	019	0.39
12. pair rel	0.46	0.48	–	0.07	0.08	–

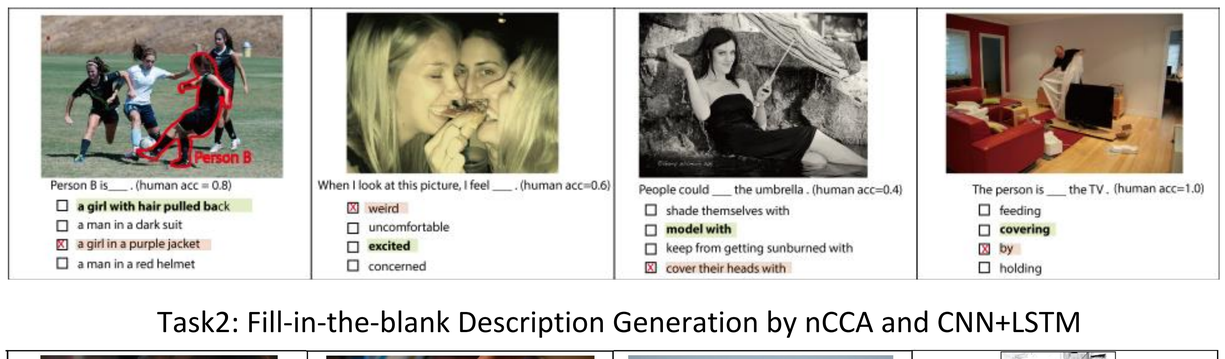
Results

Task1: Multiple-choice Question Answering by nCCA

Correct



Incorrect



Task2: Fill-in-the-blank Description Generation by nCCA and CNN+LSTM