Visual Storytelling

Ting-hao (Kenneth) Huang et al.

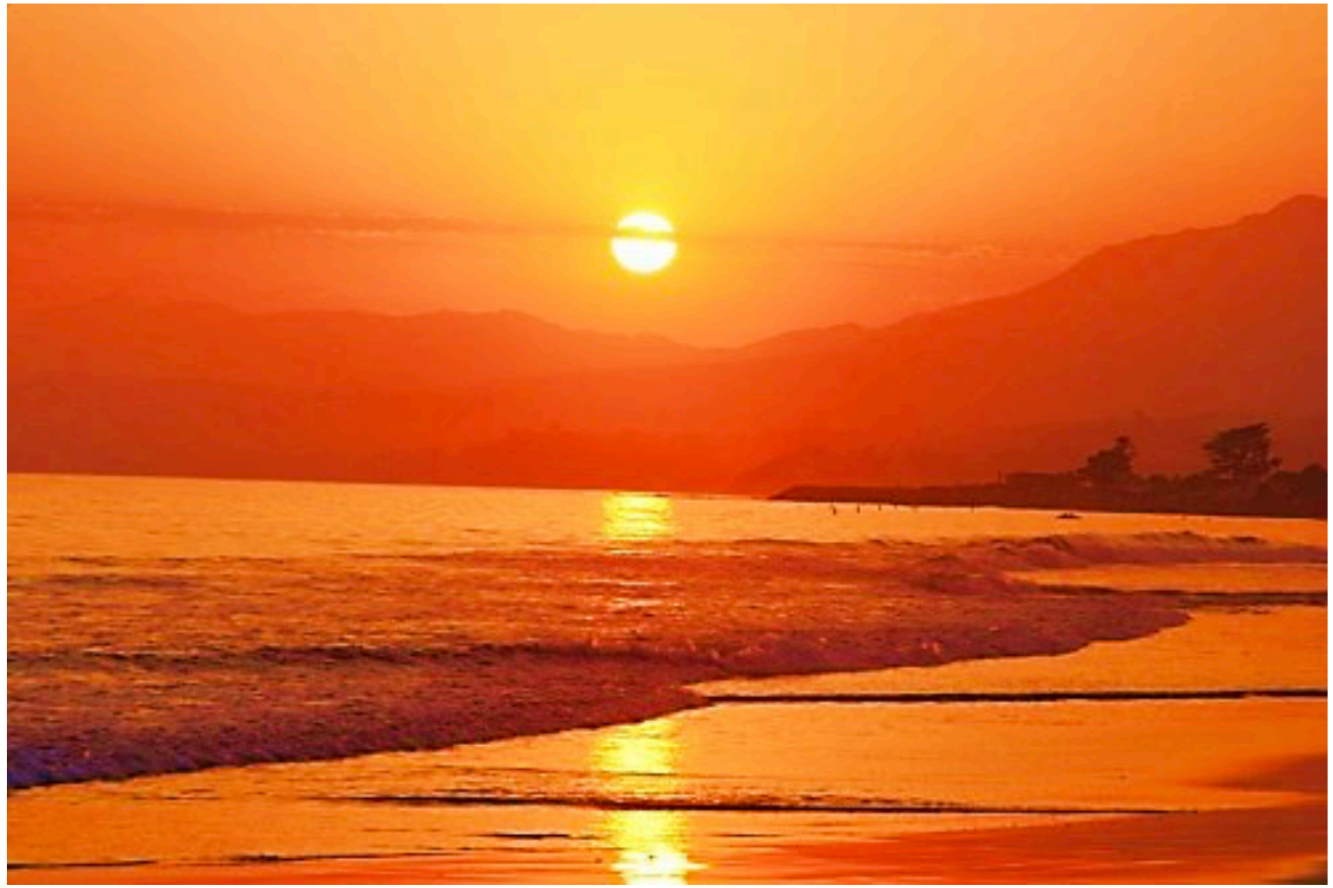
Presenter: Yiming Pang

There is a story behind every image

~~A group of people that are sitting next to each other.~~

Having a good time bonding and talking

There is another way to describe the scene



The sun is setting over the ocean and mountains.

Sky illuminated with a brilliance of gold and orange hues.

Visual Storytelling: A solid next move in AI

Outline

- Motivation and Related Work
- Visual Storytelling 101
- Dataset: SIND
- Baseline Experiments
- Conclusion

Outline

- Motivation and Related Work
- Visual Storytelling 101
- Dataset: SIND
- Baseline Experiments
- Conclusion

From Vision to Language

Work in vision to language has exploded….

From Vision to Language

- Image Captioning
  - Given an image, describe it in natural language

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

two young girls are playing with lego toy.

boy is doing backflip on wakeboard.

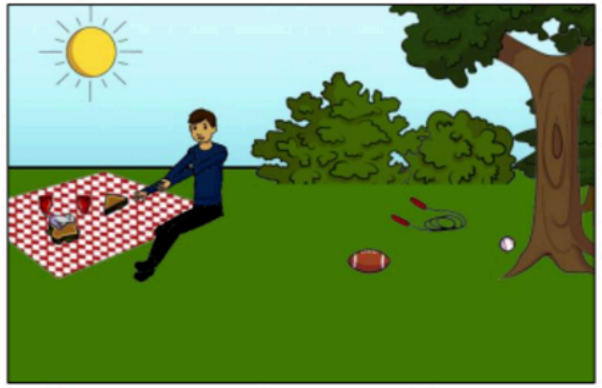
Deep Visual-Semantic Alignment for Generating Image Descriptions A. Karpathy, L. Fei-Fei

From Vision to Language

- Question Answering
  - Takes as input an image and a free-form, open-ended, natural language question about the image and produces a natural language answer as the output.

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



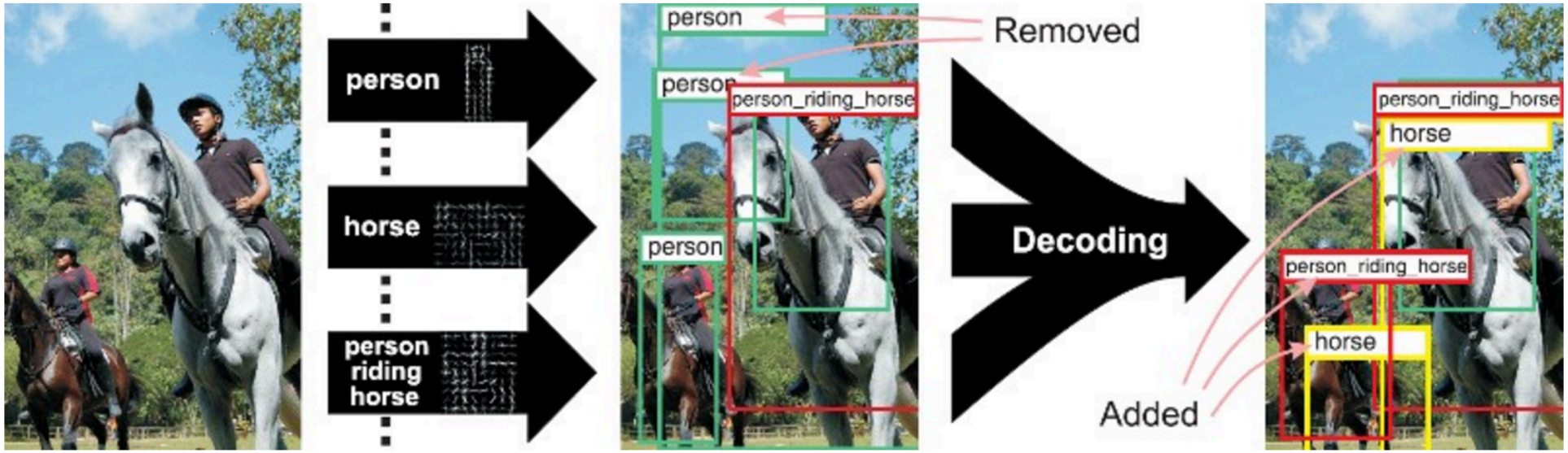
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA: Visual Question Answering A. Agrawal et al.

From Vision to Language

- Visual Phrases
  - Chunks of meaning bigger than objects and smaller than scenes



Recognition using visual phrases M. Sadeghi and A. Farhadi

And the list keeps going on…

Why visual storytelling?

- Other works focus on direct, literal description of image content.
  - Useful, meaningful
  - But still, far from the capabilities needed by intelligent agents for naturalistic interactions
- However, with visual storytelling
  - More evaluative and figurative language
  - Brings to bear information about social relations and emotions

Outline

- Motivation and Related Work
- Visual Storytelling 101
- Dataset: SIND
- Baseline Experiments
- Conclusion

What is visual storytelling?

- Go beyond basic description **(literal description)** of visual scenes
- Towards **human-like** understanding of grounded event structure and **subjective expression (narrative)**.

Literal Description **Narrative**

Sitting next to each other **VS.** Having a good time

Sun is setting Sky illuminated with a brilliance…

Good story requires more information

Single Image

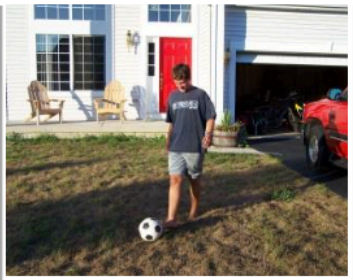
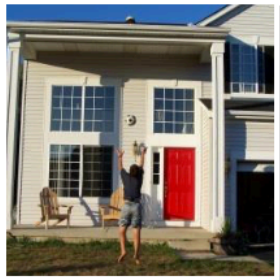
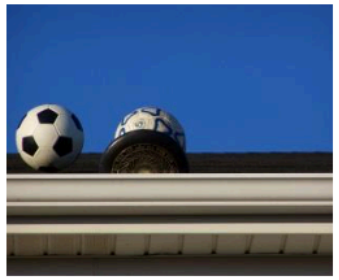
Sequence of Images

Three Tiers of Language for the Same Image

- Descriptions of Images-In-Isolation(DII):
  - Plain description as in image captioning
- Descriptions of Images-In-Sequence(DIS):
  - Same language style but images are displayed in a sequence
- Stories for Images-In-Sequence(SIS)
  - An ACTUAL story

Three Tiers of Language for the Same Image

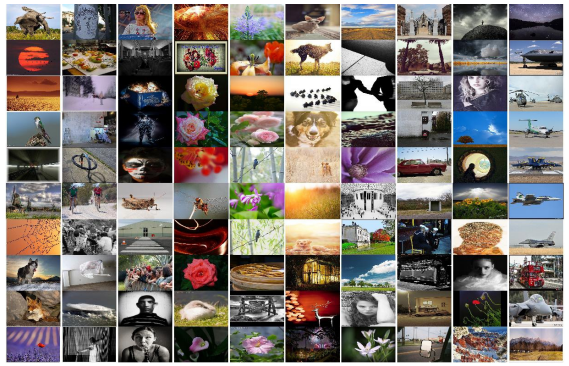
Descriptive Text ≠ Consecutive Captions ≠ Stories

					
DII	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
DIS	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Outline

- Motivation and Related Work
- Visual Storytelling 101
- Dataset: SIND
- Baseline Experiments
- Conclusion

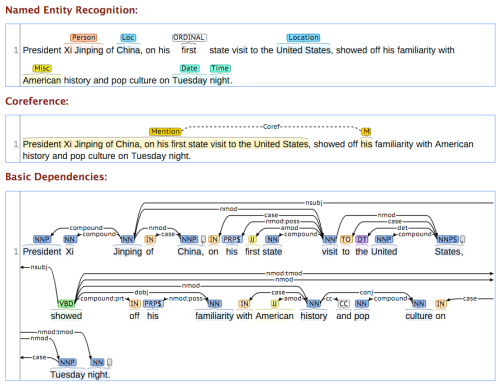
Extracting Photos



Flickr Data Release

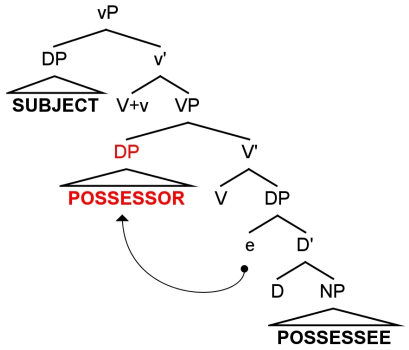
Descriptions

Feed into



Stanford CoreNLP

Extract



Possessive Dependence Patterns

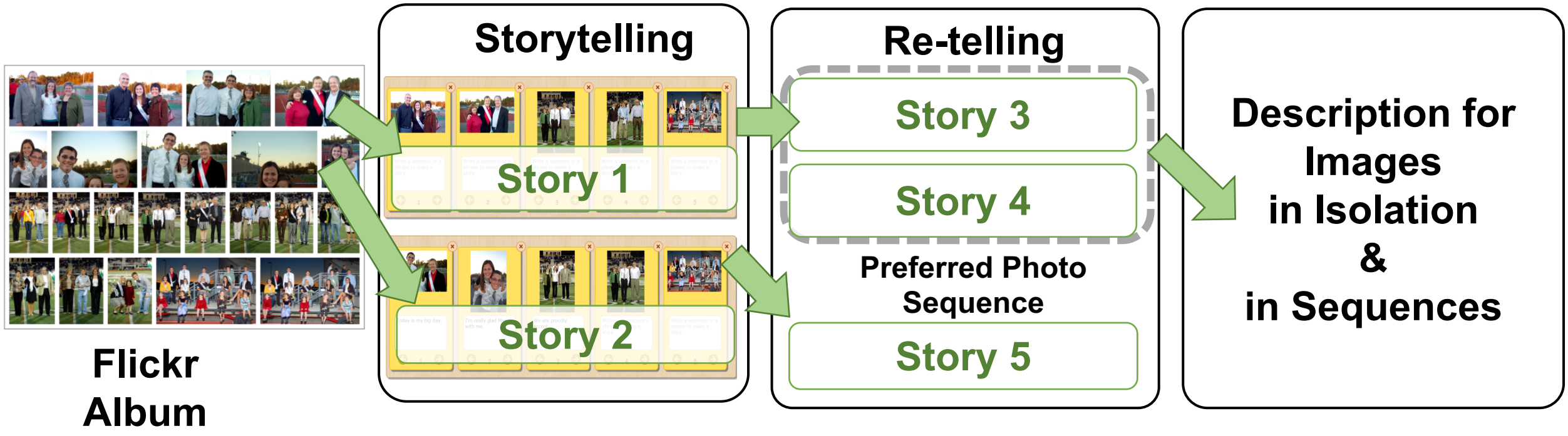
Flickr API

Only include albums within a 48-hour span

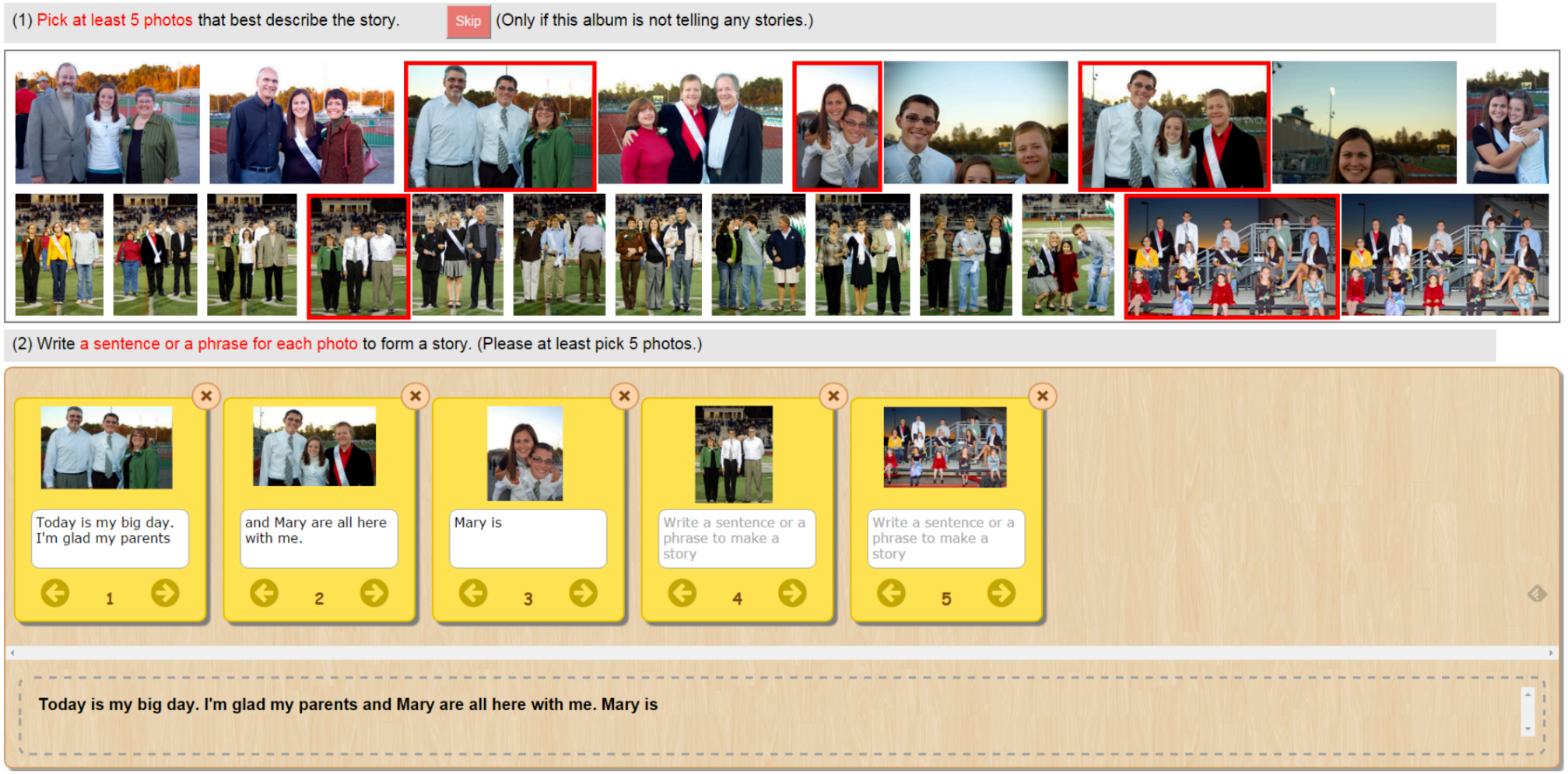
Filter by

Classify as `EVENT`

Dataset Crowdsourcing Workflow



Interface for Storytelling



Data Analysis

- 10,117 Flickr albums
- 210,819 unique photos
- 20.8 photos per album on average
- 7.9 hours time span on average

Top Words Associated with Each Tier

Desc.-in-Iso.			Desc.-in-Seq.			Story-in-Seq.		
man	sitting	black	chatting	amount	trunk	went	[female]	see
woman	white	large	gentleman	goers	facing	got	today	saw
standing	two	front	enjoys	sofa	bench	[male]	decided	came
holding	young	group	folks	egg	enjoying	took	really	started
wearing	image		shoreline	female		great	time	

Outline

- Motivation and Related Work
- Visual Storytelling 101
- Dataset: SIND
- Baseline Experiments
- Conclusion

What's the best metric to evaluate the story?

- The best and most reliable evaluation is human judgment
  - Crowdsourcing on MTurk

Strongly disagree Disagree Neutral Agree Strongly agree

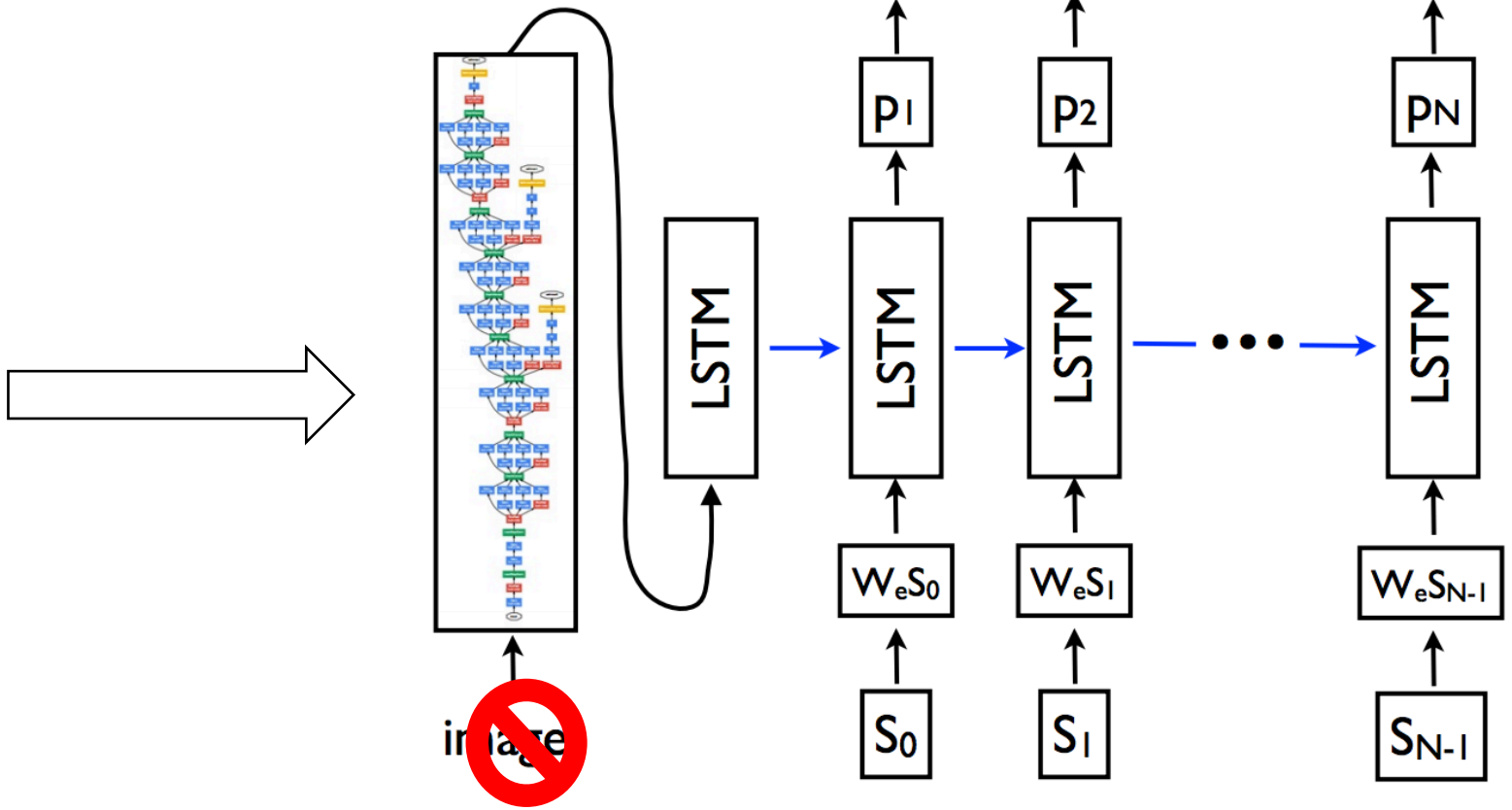
- For quick benchmark progress: automatic evaluation metric
  - METEOR
    - The Meteor automatic evaluation metric scores machine translation hypotheses by aligning them to one or more reference translations. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases.
  - Smoothed-BLEU
    - Bilingual evaluation user study
  - Skip-Thoughts

Which one is the best?

	METEOR	BLEU	Skip-Thoughts
r	0.22 (2.8e-28)	0.08 (1.0e-06)	0.18 (5.0e-27)
ρ	0.20 (3.0e-31)	0.08 (8.9e-06)	0.16 (6.4e-22)
τ	0.14 (1.0e-33)	0.06 (8.7e-08)	0.11 (7.7e-24)

Train

Sequence of Images



Show and tell: a neural image caption generator O. Vinyals et al.

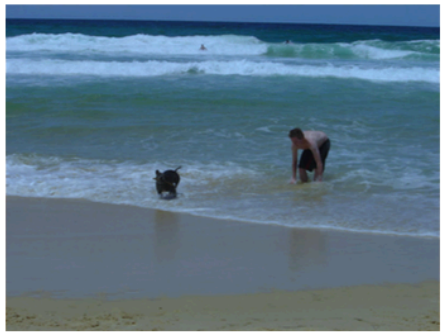
Generate the story

- Simple beam search (size=10)
- However, it does not work very well…

This is a picture of a family.

This is a picture of a cake.

This is a picture of a dog.



This is a picture of a beach.

This is a picture of a beach

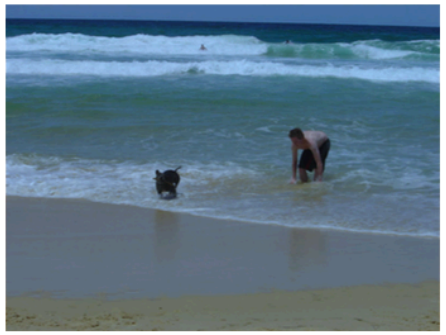
Generate the better story

- Greedy beam search (size=1)
- Resulting in a 4.6 gain in METEOR score

The family gathered together for a meal

The food was delicious.

The dog was excited to be there.



The dog was enjoying the water.

The dog was happy to be in the water.

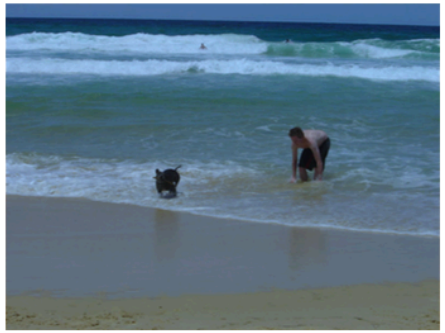
Generate the better story (cont.)

- A very simple heuristic: the same content word cannot be produced more than once within a given story.
- Resulting in a 2.3 gain in METEOR score

The family gathered together for a meal

The food was delicious.

The dog was excited to be there.



The kids were playing in the water

The boat was a little too much to drink.

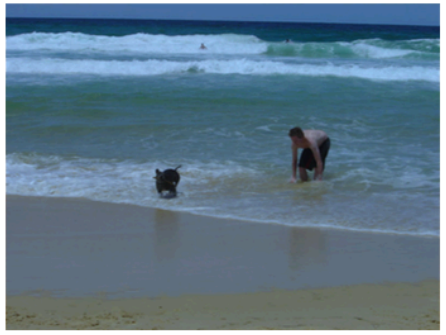
Generate the better story (cont.)

- Additional baseline: visually grounded words
- $$\frac{P(w|T_{caption})}{P(w|T_{story})} > 1.0$$
- Resulting in a 1.3 gain in METEOR score

The family got together for a cookout

They had a lot of delicious food.

The dog was happy to be there.



They had a great time on the beach.

They even had a swim in the water.

Final Results

- METEOR scores for different methods

Beam=10	Greedy	-Dups	+Grounded
23.13	27.76	30.11	31.42

Outline

- Motivation and Related Work
- Visual Storytelling 101
- Dataset: SIND
- Baseline Experiments
- Conclusion

Conclusion

- The first dataset for sequential vision-to-language.
- Images-in-isolation to stories-in-sequence.
- Evolving AI towards more human-like understanding

Q&A