Generation in Image Captioning

Mark Johnson
Voicebox Technologies Australia / Macquarie University

Peter Anderson, Basura Fernando and Stephen Gould
Australian National University

Overview

- The image captioning problem
- Dependency evaluation of captions (SPICE)
- Controlling generation with specialised decoding
- Summary

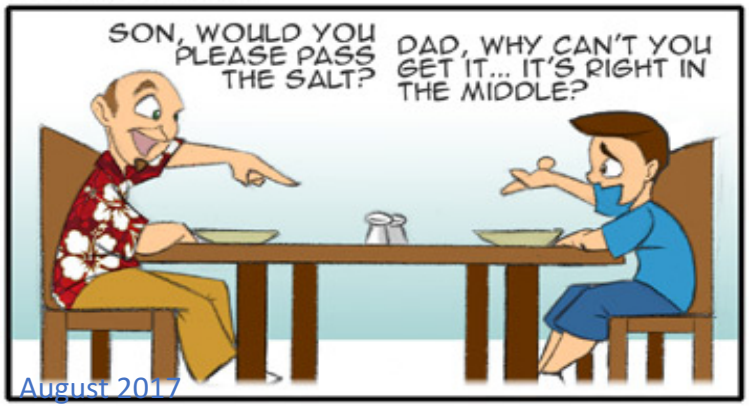
Image captioning

"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."



Evaluating automatic captions using SPICE

Evaluating captions automatically

- Benchmark datasets require fast to compute, accurate and inexpensive evaluation metrics
- Good metrics can be 'climbed' in the development-validation loop

- **The Evaluation Task:**
- Given a candidate caption c_i and a set of m reference captions $R_i = \{r_{i1}, \ldots, r_{im}\}$, compute a score S_i that represents similarity between c_i and R_i.

The state of the art

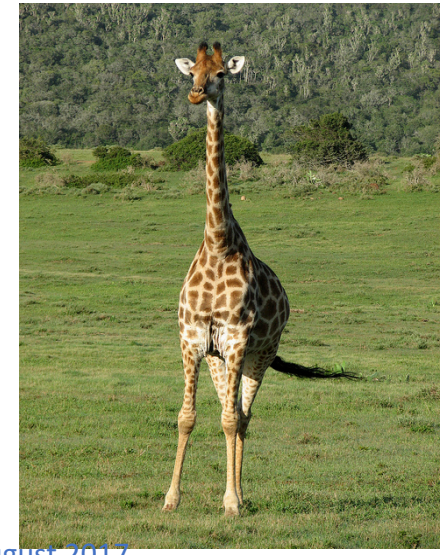
- **BLEU:** Precision with brevity penalty, geometric mean over n-grams
- **ROUGE-L:** *F*-score based on Longest Common Substring
- **METEOR:** Align fragments, take harmonic mean of precision & recall
- **CIDEr:** Cosine similarity with TF-IDF weighting

The current state of the art

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSR Captivator[9]	0.937	0.339	0.68	0.907	0.819	0.71	0.601
Google[4]	0.946	0.346	0.682	0.895	0.802	0.694	0.587
m-RNN (Baidu/ UCLA)[16]	0.896	0.32	0.668	0.89	0.801	0.69	0.578
m-RNN[15]	0.935	0.325	0.666	0.89	0.798	0.687	0.575
MSR[8]	0.925	0.331	0.662	0.88	0.789	0.678	0.567
PicSOM[13]	0.856	0.318	0.654	0.875	0.775	0.663	0.554
Nearest Neighbor[11]	5	9	9	9	8	7	7
Berkeley LRCN[2]	0.891	0.322	0.656	0.871	0.772	0.653	0.534
Montreal/Toronto[10]	0.878	0.323	0.651	0.872	0.768	0.644	0.523
MLBL[7]	0.752	0.294	0.635	0.848	0.747	0.633	0.517
Tsinghua Bigeye[14]	0.682	0.273	0.616	0.866	0.756	0.628	0.493
ACVT[1]	0.716	0.288	0.617	0.831	0.713	0.589	0.478
Human[5]	6	3	11	6	12	12	13

False positives in N-gram based evaluation

A young girl *standing on top of* a tennis court.



A giraffe *standing on top of* a green field.

N-gram overlap isn’t necessary

- A shiny metal pot filled with some diced veggies.
- A pan on the stove with chopped vegetables

Score this caption out of 10

"*A young girl standing on top of a basketball court*"

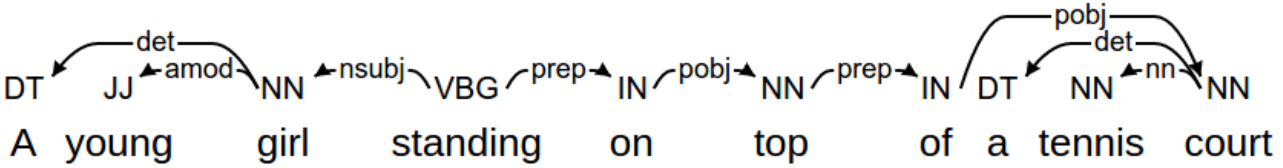
How would you score this caption?

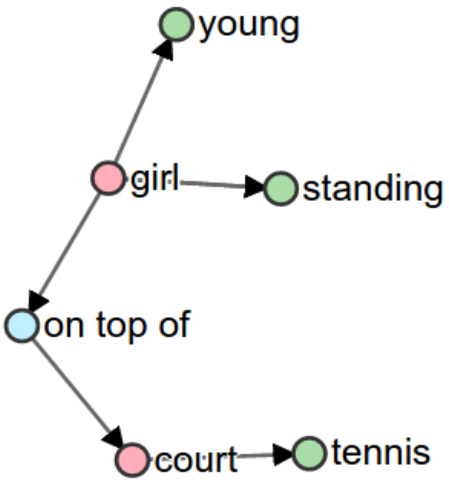
- There is girl
- Girl is young
- Girl is standing
- There is court
- Court is for basketball
- Girl is on court

These are the *propositional content* of the utterance

High level intuition behind SPICE

- Use a parser to identify sets of propositions in caption and gold labels
- Count the overlap between proposition sets





Related work

- **Syntactic dependency parsing**
  - Klein & Manning: *Accurate Unlexicalized Parsing*, ACL 2003
- **Scene graphs for image retrieval**
  - Johnson et. al: *Image Retrieval Using Scene Graphs*, CVPR 2015
- **Rule-based mapping from dependency parse to scene graph**
  - Schuster et. al: *Generating semantically precise scene graphs from textual descriptions for improved image retrieval*, EMNLP 2015

SPICE metric calculation

- Synonymous nodes merged in *G(S)*
- Wordnet synsets used for tuple matching

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|}$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|}$$

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$$

Example of scene graph

- Scene graph (right) parsed from a set of reference captions (left)

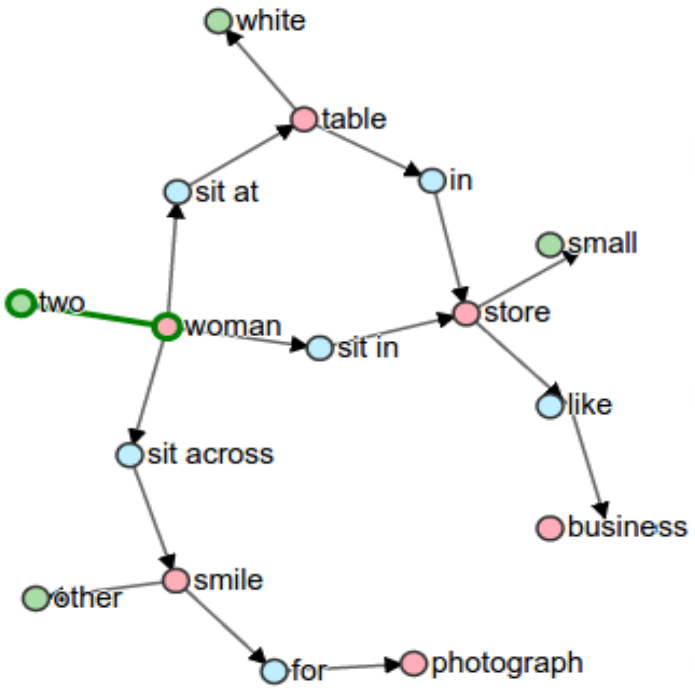
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

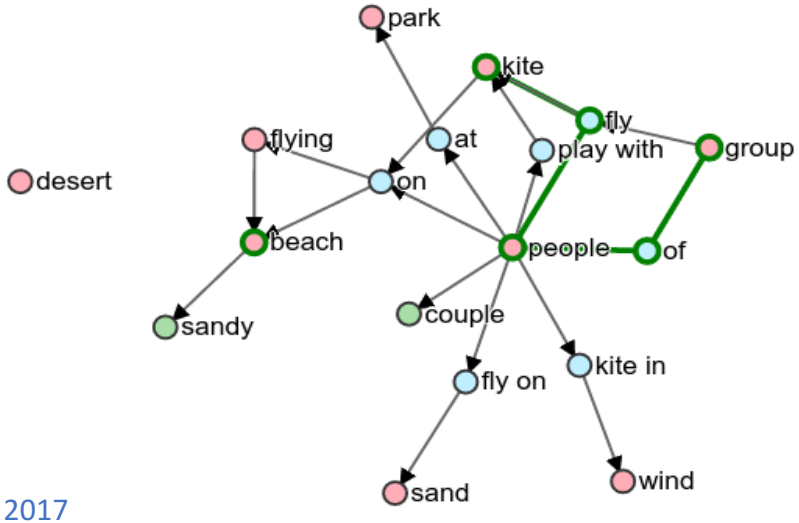


Good caption example (1)

Reference captions

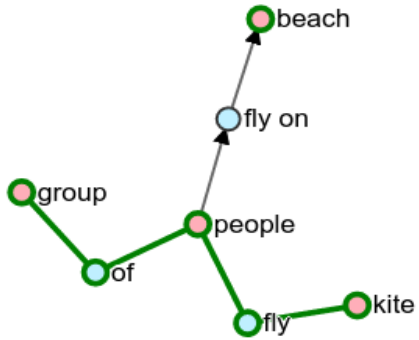
"People playing with kites outside in the desert."
"A group of people at a park flying a kite. "
"A group of people flying a kite on a sandy beach"
"People on the beach flying kites in the wind."
"A couple people out flying a kite on some sand."

Reference scene graph



Candidate caption & scene graph

"a group of people flying kites on a beach"



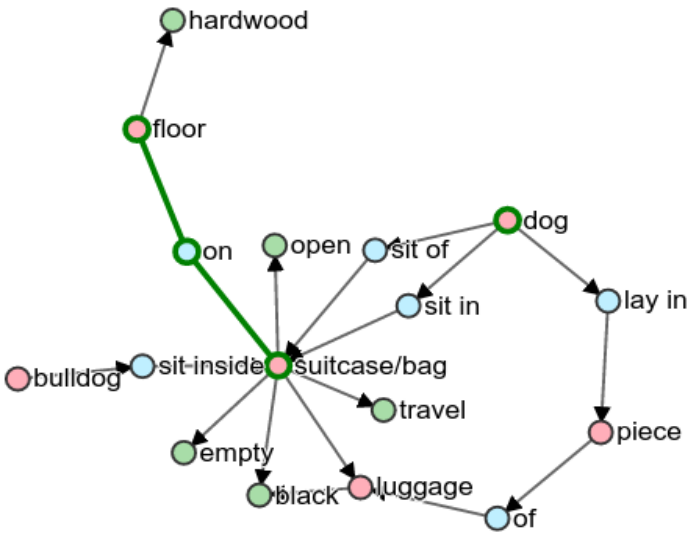
SPICE F-Score: 0.429, Pr: 0.857, Re: 0.286

Good caption example (2)

Reference captions

"a dog is sitting inside of a black suitcase"
"The bulldog is sitting inside the travel bag."
"A dog laying in a piece of black luggage."
"A dog sits in an open suitcase that is on a hardwood floor."
"A dog sitting inside an empty luggage bag on the floor"

Reference scene graph



Candidate caption & scene graph

"a dog sitting in a suitcase on the floor"

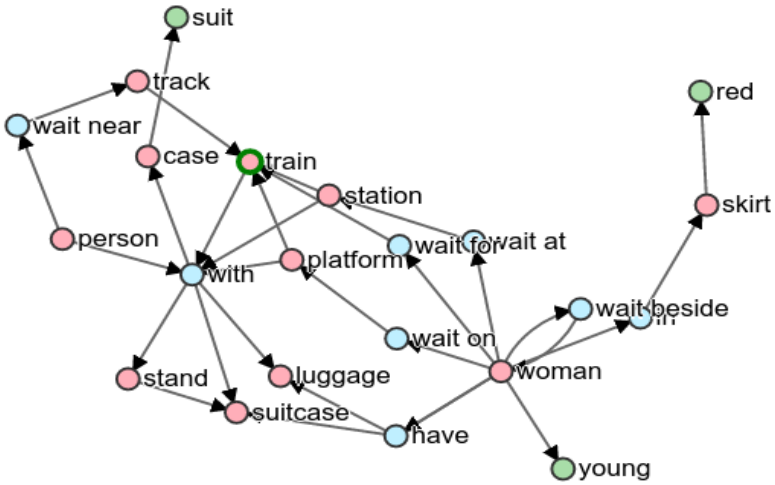
SPICE F-Score: 0.348, Pr: 1, Re: 0.211

Poor caption example (1)

Reference captions

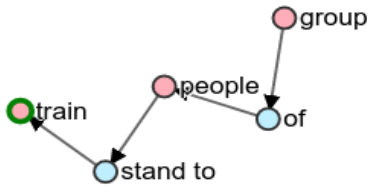
"A woman is waiting for a train. "
"A woman waiting at a train station with a suit case."
"A person with a suitcase stands waits near the train tracks. "
"A young woman in a red skirt is waiting on a train platform with her suitcase. "
"A woman waiting for a train with her luggage beside her."

Reference scene graph



Candidate caption & scene graph

"a group of people standing next to a train"



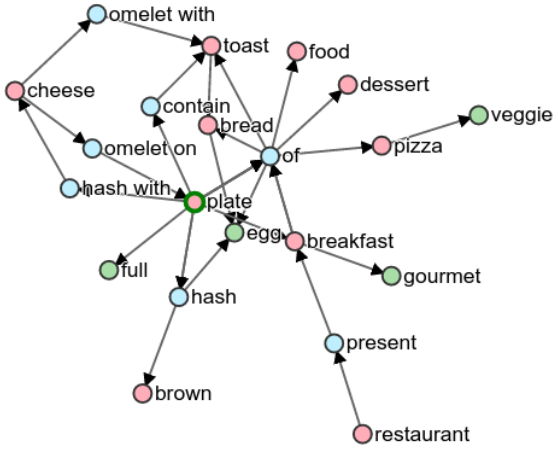
SPICE F-Score: 0.057, Pr: 0.2, Re: 0.033

Poor caption example (2)

Reference captions

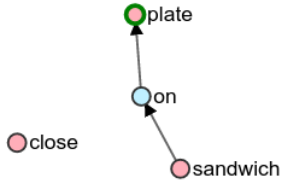
"The restaurant presents a gourmet breakfast of eggs and toast."
"A full plate of dessert, bread, and a veggie pizza. "
"A breakfast plate containing eggs, bread and french toast."
"A plate of food that includes toast, hash browns and eggs with cheese."
"A cheese omelet with toast on a plate."

Reference scene graph



Candidate caption & scene graph

"a close up of a sandwich on a plate"



SPICE F-Score: 0.059, Pr: 0.25, Re: 0.033

Properties of SPICE

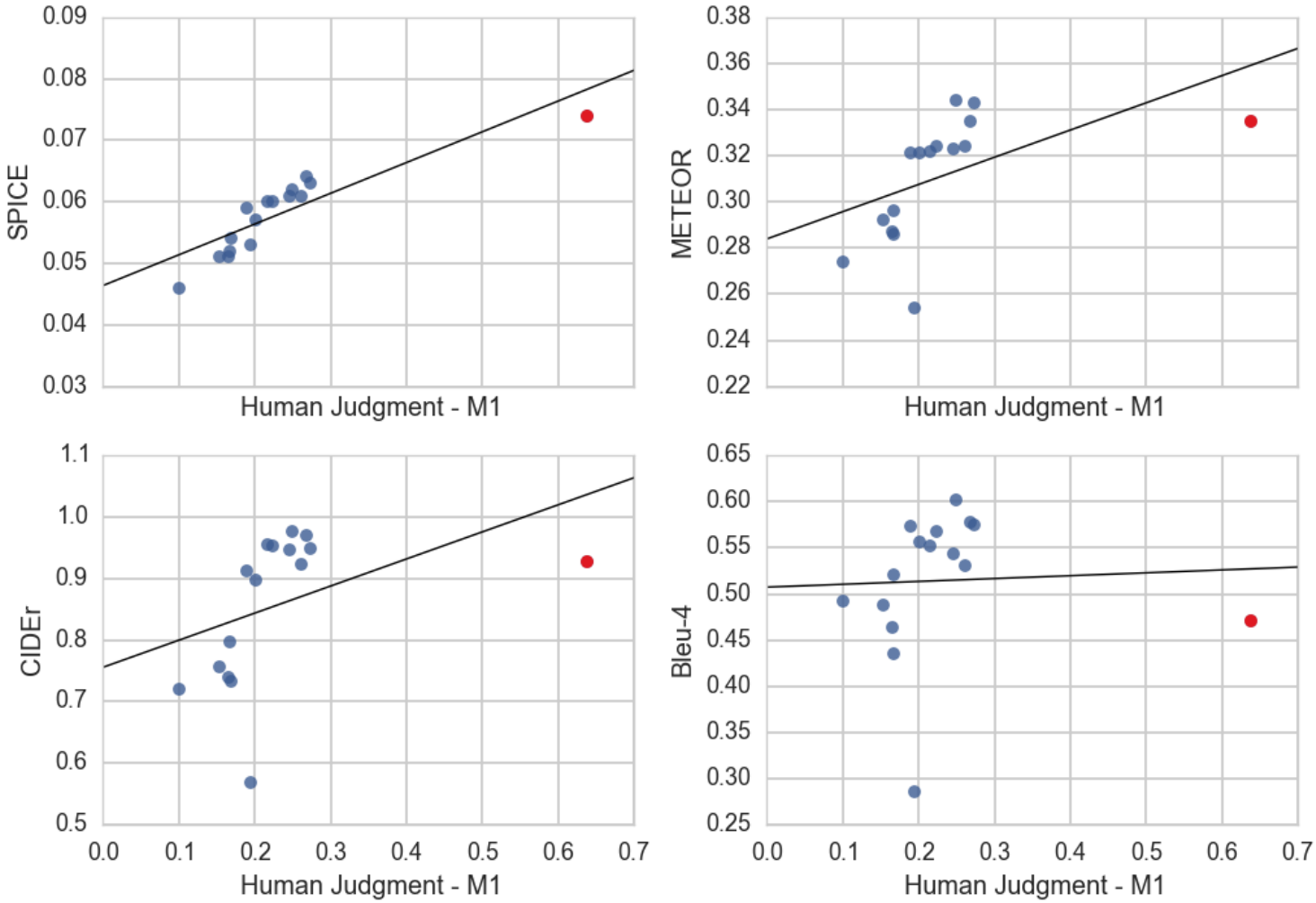
- SPICE measures how well caption models recover objects, attributes and relations
- Fluency neglected (as with n-gram metrics)
- If fluency is a concern, include a fluency metric such as surprisal*
- To model human judgement as closely as possible, build a task-specific metric ensemble

- *Hale, J: A probabilistic Earley Parser as a Psycholinguistic Model 2001; Levy, R: Expectation-based syntactic comprehension 2008

Evaluation on MS COCO data (1)

- Based on system level correlation between automatic scores and human judgments (using 255k human judgments)
- Pearson correlation with human judgments (M1) is 0.88 for SPICE, vs. 0.43 for CIDEr and 0.53 for METEOR.
- SPICE ranks human captions ahead of competition entries, and picks the same top-5 competition entries as humans.

Evaluation on MS COCO data (2)



Summary of SPICE

- SPICE measures how effectively image captions recover objects, attributes and relations
- Captures human judgment on model-generated captions better than CIDEr, BLEU, METEOR and ROUGE
- Tuples can be categorized to provide detailed error analysis
- Offers scope for further improvement as better parsers are developed

Guided Open Vocabulary Image Captioning with Constrained Beam Search

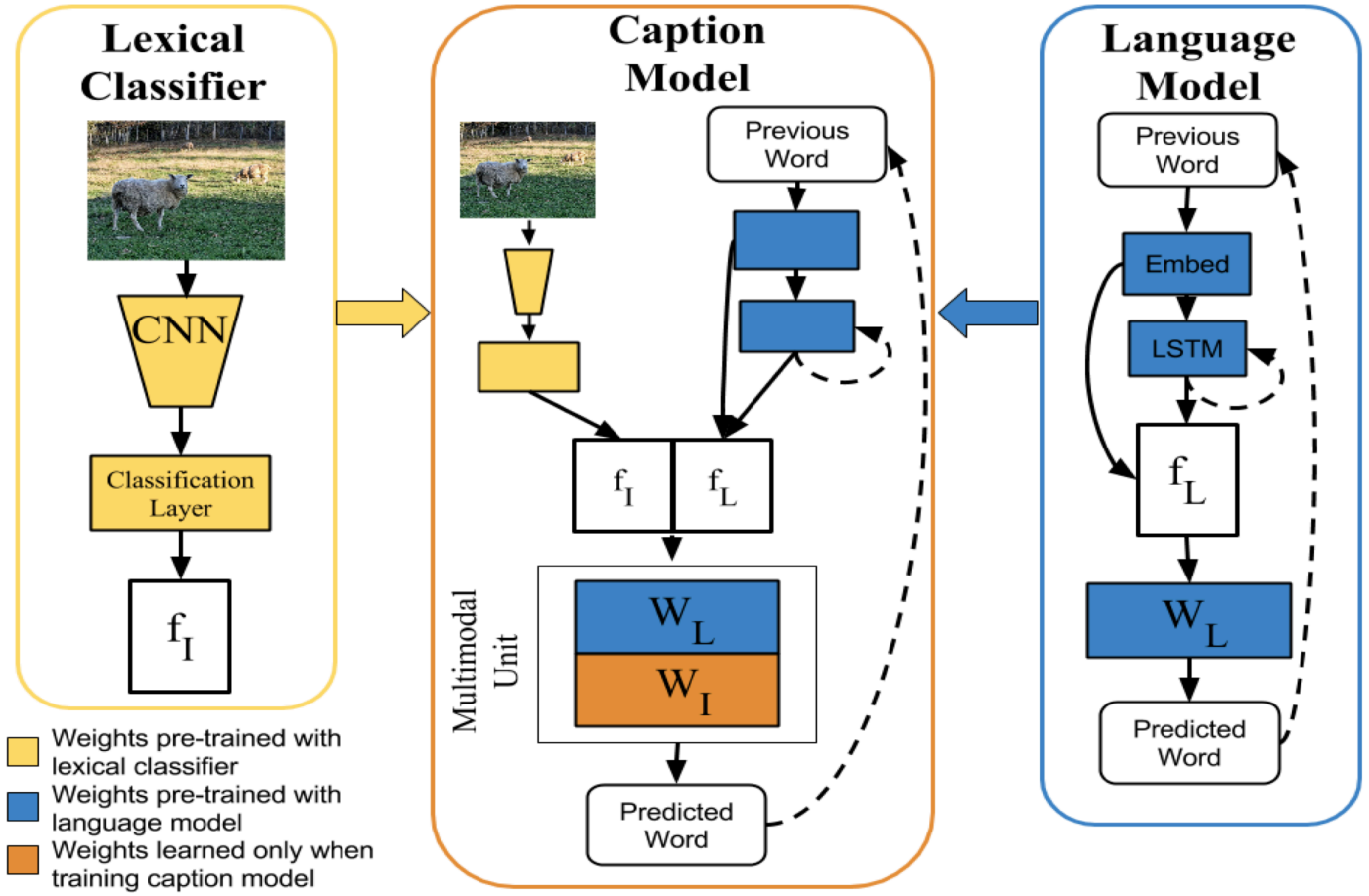
Motivation (1)

A close up of a pizza on the ground.

Motivation (2)

A bird standing on top of a grass covered field.

Prior work in out-of-domain captioning



- Source: '*Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data*', Hendricks et. al. CVPR 2016 oral

Guided open-vocabulary captioning

Out-of-Domain image containing unseen object ('suitcase')

CNN-RNN Captioning Model

Beam Search

A cat sitting on top of **a refrigerator**.

Open-vocab tags

cat, **suitcase**, **inside**

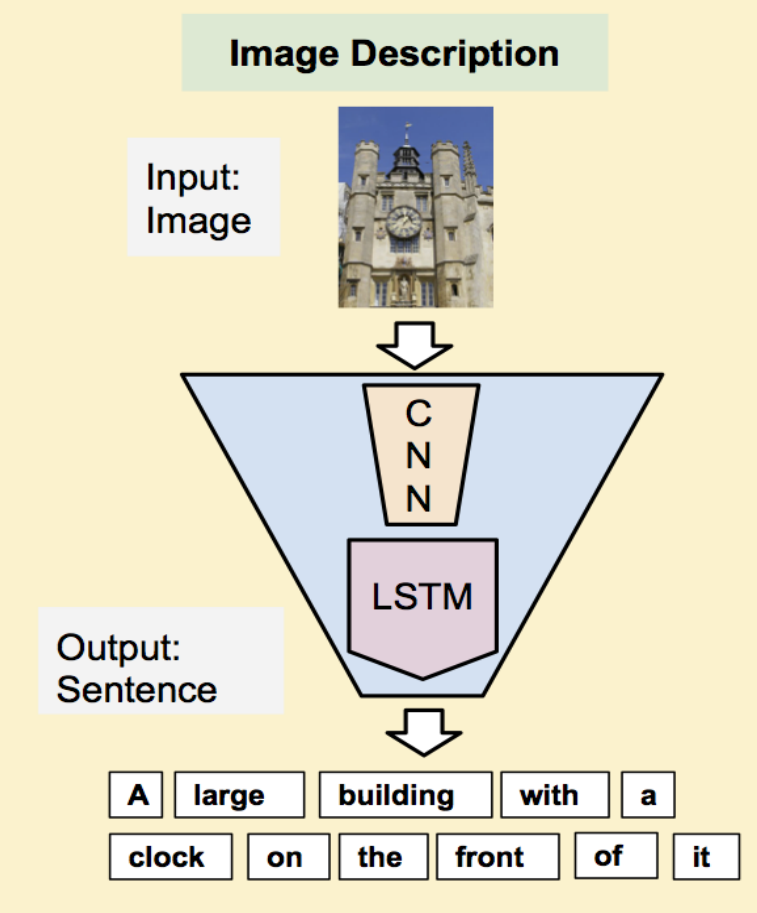
Constrained Beam Search

A **cat** sitting **inside** of a **suitcase**.

Overview of constrained decoding

- Caption generator uses pretrained word embeddings
  - Generates words not seen in training captions
- Image labeller trained on larger label vocabulary
- At test time:
  - Image labeller identifies key words that caption must contain
  - Construct a finite state automaton that accepts captions containing key words
  - Decoder has a beam for *each automaton state*
    - Minimises *label bias*
  - Output is *highest scoring string* in a final beam

Base model: LRCN



- 2-layer LSTM network, based on LRCN[1]
- LSTM inputs at level 1 and 2 given by:

$$x_t^1 = W_e \Pi_t$$

$$x_t^2 = (h_t^1, \mathrm{CNN}_\theta(I))$$

- where W_e is a word embedding, π_t is an indicator column vector, h^1_t is the output of the first layer, and I is the input image

[1] *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*, Donahue et. al. CVPR 2015. Figure reproduced from Donahue et. al.

Vocabulary expansion

- Introduce pretrained GloVe[2] 300D embeddings at both the LSTM input and output layers (W_e):

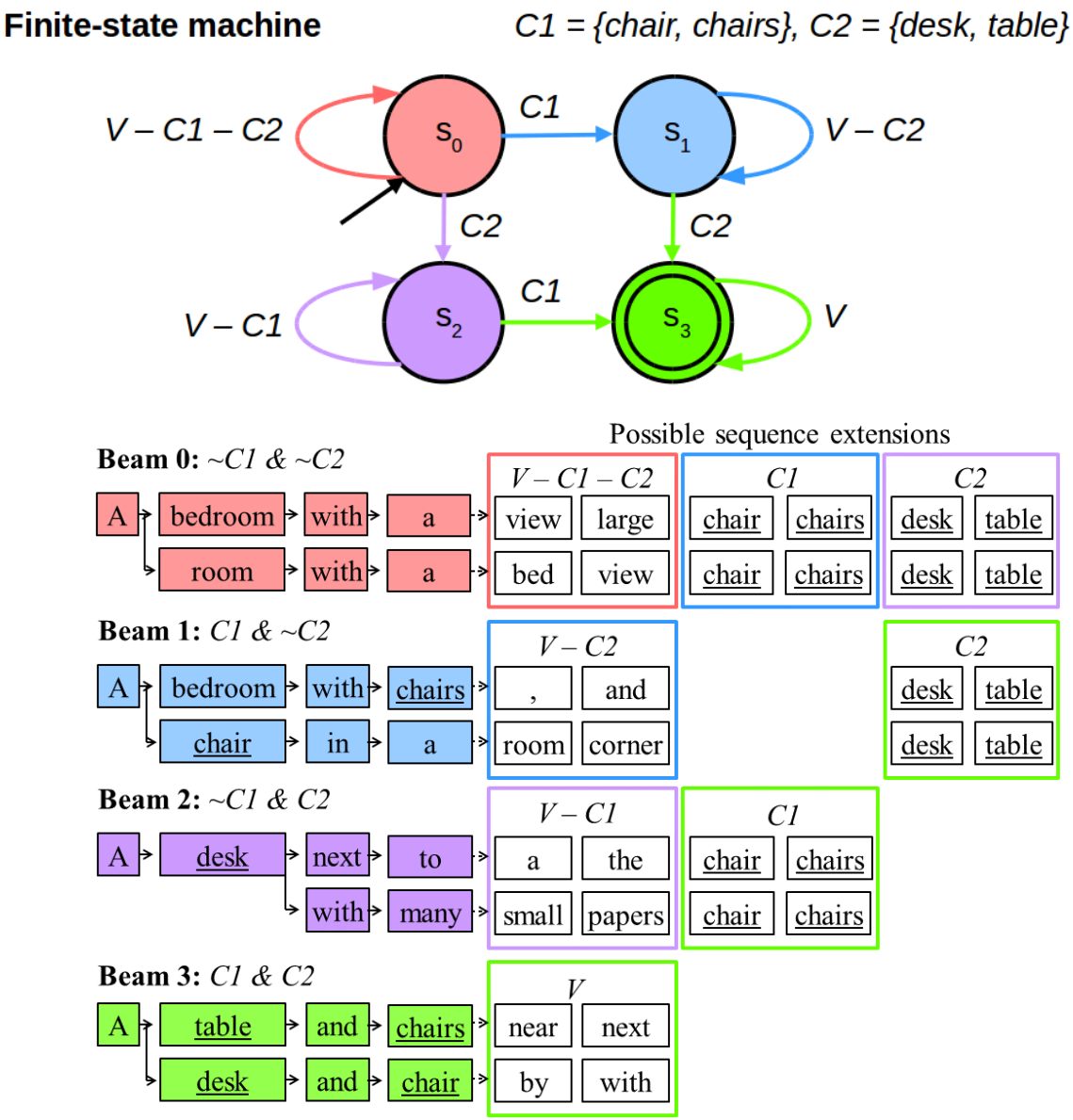
$$v_t = \tanh\left(W_v h_t^2 + b_v\right)$$

$$p(y_t \mid y_{t-1}, \ldots, y_1, I) = \text{softmax}\left(W_e^T v_t\right)$$

- W_e fixed during training with minimal performance impact (using conventional cross-entropy loss).
- Model learns to predict 300D vectors v_t with a high dot-product similarity with the GloVe embedding of the correct output word.
- New vocabulary introduced at test time by concatenating the GloVe vector as an additional column to W_e

[2] 'GloVe: Global Vectors for Word Representation', Pennington et. al. EMNLP 2014

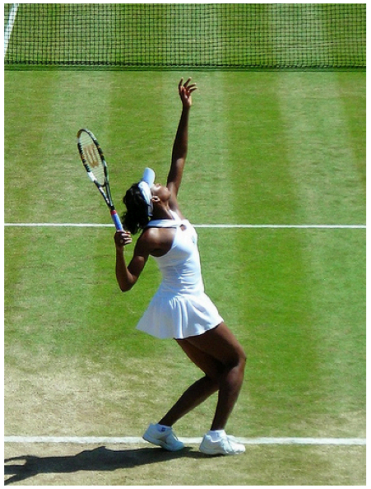
Finite-state multi-beam decoder



Experimental setup – MS COCO

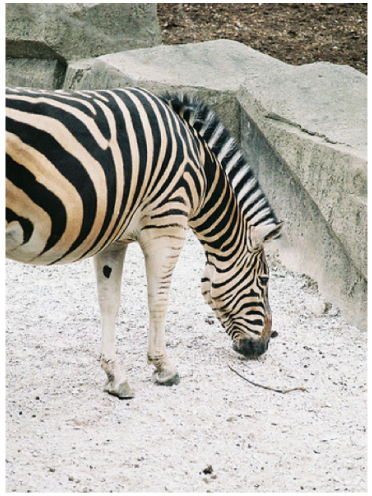
- Following Hendricks et. al., 8 objects removed from the caption training set: *bus, racket, couch, suitcase, bottle, microwave, pizza, zebra* (incl. plurals, synonyms)
- Image tag training set is unrestricted (formed by tokenizing captions)
- We re-use the image-tagger (Lexical Classifier) from trained by Hendricks et. al.

Examples: MS COCO (1)



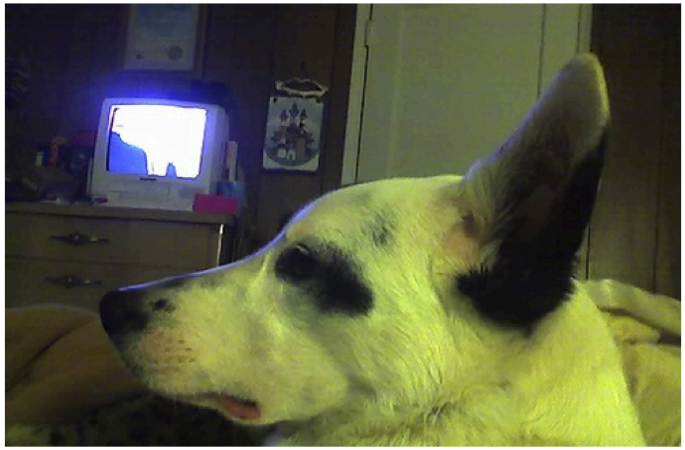
Base: A woman is playing tennis on a tennis court. **LC4 Tags:** tennis, player, ball, racket. **Base + LC4:** A tennis player swinging a racket at a ball.

Base: A man standing next to a yellow train. **LC4 Tags:** bus, yellow, next, street. **Base + LC4:** A man standing next to a yellow bus on the street.



Base: A close up of a cow on a dirt ground. **LC4 Tags:** zebra, zoo, enclosure, standing. **Base + LC4:** A zebra standing in front of a zoo enclosure.

Examples: MS COCO (2)



Base: A dog is sitting in front of a tv. **LC4 Tags:** dog, head, television, cat. **Base + LC4:** A dog with a cat on its head watching television.

Base: A group of people playing a game of tennis. **LC4 Tags:** pink, tennis, crowd, ball. **Base + LC4:** A crowd of people standing around a pink tennis ball.

Results on MS COCO

Model	CNN	Out-of-Domain SPICE	METEOR	CIDEr	F1	In-Domain SPICE	METEOR	CIDEr
DCC [10]	VGG-16	13.4	21.0	59.1	39.8	15.9	23.0	77.2
NOC [28]	VGG-16	-	20.7	-	50.5	-	-	-
Base	VGG-16	12.4	20.4	57.7	0	17.6	24.9	93.0
Base + LC1	VGG-16	13.6	21.7	68.9	27.2	17.9	**25.0**	**93.4**
Base + LC2	VGG-16	14.8	22.6	75.4	38.7	**18.2**	**25.0**	92.8
Base + LC3	VGG-16	15.5	23.0	77.5	48.4	**18.2**	24.8	90.4
Base + LC4	VGG-16	**15.9**	**23.3**	**77.9**	**54.0**	18.0	24.5	86.3
Base + GT3	VGG-16	18.7	27.1	119.6	54.5	22.0	29.4	135.5
Base All Data	VGG-16	17.8	25.2	93.8	59.4	17.4	24.5	91.7
Base	ResNet-50	12.6	20.5	56.8	0	18.2	24.9	93.2
Base + LC1	ResNet-50	14.2	21.7	68.1	27.3	18.5	25.2	**94.6**
Base + LC2	ResNet-50	15.3	22.7	74.7	38.5	**18.7**	**25.3**	94.1
Base + LC3	ResNet-50	16.0	23.3	**77.8**	48.2	**18.7**	25.2	92.3
Base + LC4	ResNet-50	**16.4**	**23.6**	77.6	**53.3**	18.4	24.9	88.0
Base + GT3	ResNet-50	19.2	27.3	117.9	54.5	22.3	29.4	133.7
Base All Data	ResNet-50	18.6	26.0	96.9	60.0	18.0	25.0	93.8

Captioning ImageNet

- Can we leverage existing image labels?
- Base model using ResNet-50 CNN, trained on MS COCO + Flickr 30k (150k captions)
- Constrained beam search using the ground-truth synset
- Intend to release captions for 1.2M images (ILSVRC 2012)

ImageNet examples

Base: A close up of a pizza on the ground. **Synset:** rock crab. **Base + Synset:** A large rock crab sitting on top of a rock.

Base: A close up shot of an orange. **Synset:** pool table, billiard table, snooker table. **Base + Synset:** A close up of an orange ball on a billiard table.

Base: A man and a woman standing next to each other. **Synset:** colobus, colobus monkey. **Base + Synset:** Two colobus standing next to each other near a fence.

Base: A herd or horses standing on a lush green field. **Synset:** rapeseed. **Base + Synset:** A group of horses grazing in a field of rapeseed.

Base: A black bird is standing in the grass. **Synset:** oystercatcher, oyster catcher. **Base + Synset:** A black oystercatcher with a red beak standing in the grass.

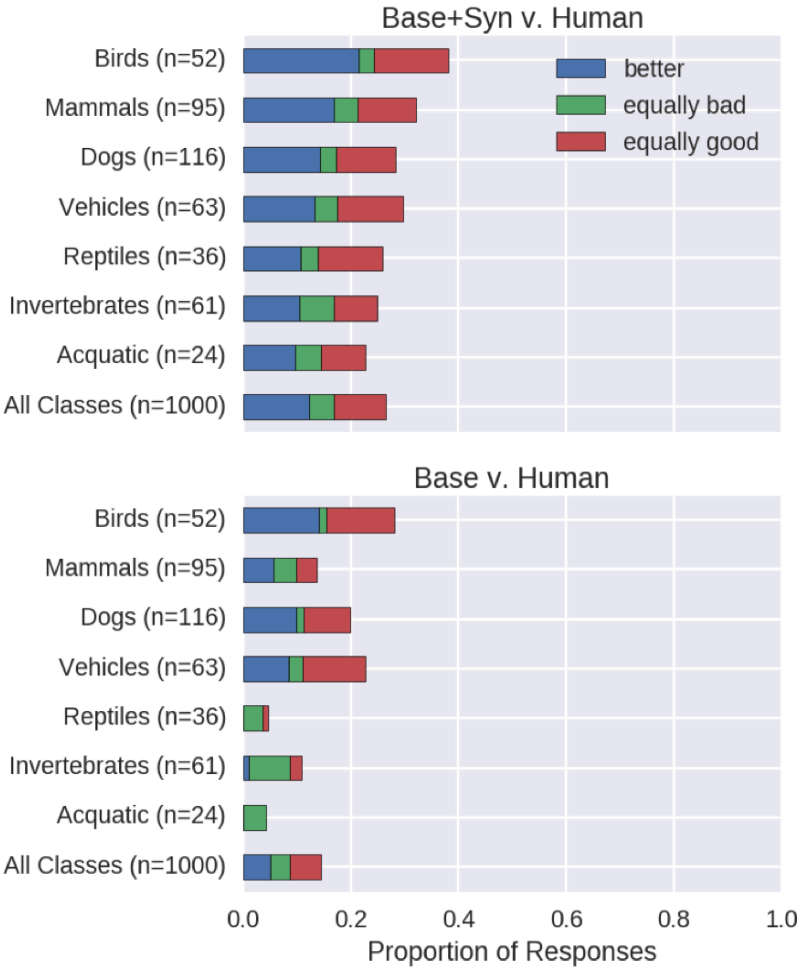
Base: A bird standing on top of a grass covered field. **Synset:** cricket. **Base + Synset:** A bird standing on top of a cricket field.

Human evaluation on ImageNet (1)

- AMT evaluations, protocol identical to MS COCO Captioning Challenge 2015
- Workers compare two captions, 3 evaluations x 5k samples images
- For context, the best 2015 in-domain model achieved 11% ‘better’, 17% ‘equally good’

	Better	Equally Good	Equally Poor	Worse
Base+Syn v. Human	0.13	0.09	0.05	0.73
Base+Syn v. Base	0.39	0.06	0.42	0.13
Base v. Human	0.05	0.06	0.04	0.86

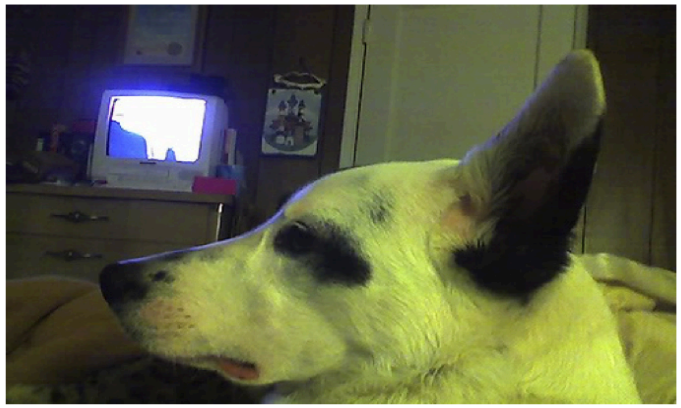
Human evaluation on ImageNet (2)



- Clustering class labels illustrates improvements across all categories
- 38% equal or better than human on birds
- Promising for combining fine-grained object detectors with general captioning models

Future work on constrained decoding

- Couple with Expectation-Maximization (EM) algorithm to learn from weakly-labelled images
- Ground tags in the image to tackle these failures:



Base: A dog is sitting in front of a tv. **LC4 Tags:** dog, head, television, cat. **Base + LC4:** A dog with a cat on its head watching television.

Base: A group of people playing a game of tennis. **LC4 Tags:** pink, tennis, crowd, ball. **Base + LC4:** A crowd of people standing around a pink tennis ball.

Conclusions

- Vision + language / zero-shot learning
- Base model using ResNet-50 CNN, trained on MS COCO + Flickr 30k (150k captions)
- Constrained beam search using the ground-truth synset
- Intend to release captions for 1.2M images (ILSVRC 2012)

Conclusions and future work

Conclusions and future work

- SPICE evaluates captions by comparing their *propositional content* to the propositional content of reference captions
- Our guided decoding algorithm uses a high-precision image labeler to constrain the decoder
  - Finite state constraints on decoder
  - Multiple beams minimise label bias
- Is there a better decoding algorithm than left to right decoding?