Structured Prediction and Image Reasoning Paper Presentation

Neural Baby Talk

By Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh (Georgia Tech and Facebook AI)

In Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018

14 March 2019

Presentors: Soojin Lee, Austin Rothwell, and Shane Sims
CPSC 532S
UBC Department of Computer Science

Motivation

https://www.organicauthority.com/.image/t_share/MTU5MzMwMTE3MjM5MTIxNTA0/dog-toy-cc.jpg

- Aid for the visually impaired
- Personal assistants
- Human robot interaction

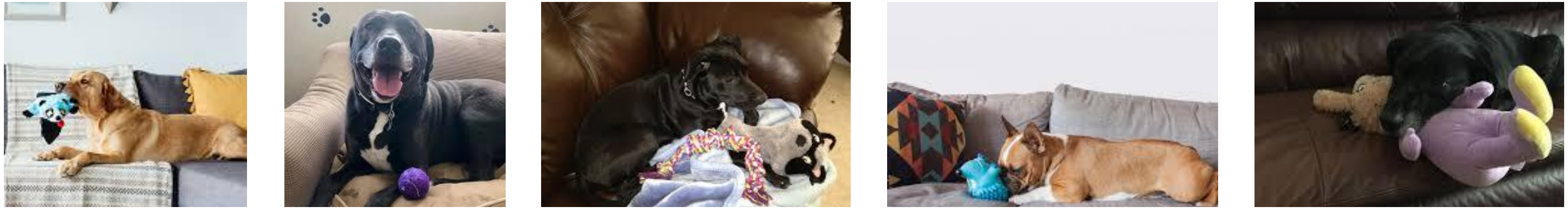
Current Methods

"A dog is sitting on a couch with a toy"

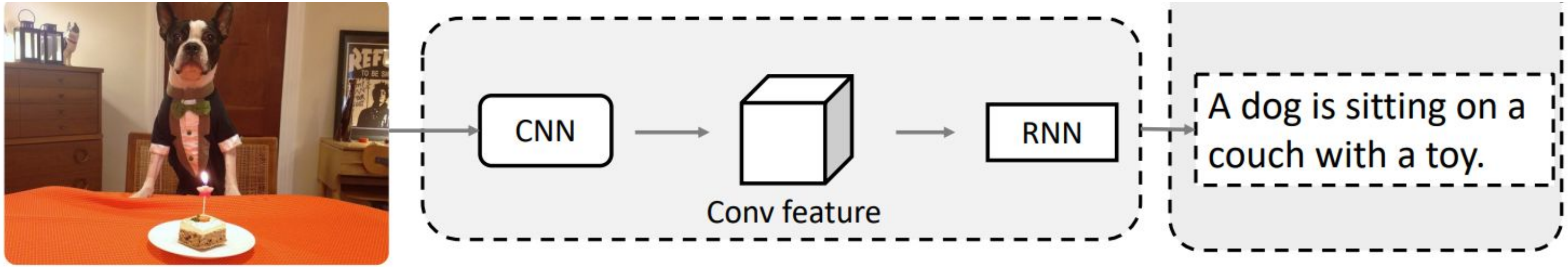
Source: Neural Baby Talk by Jiasen Lu et. al.

Current Issues

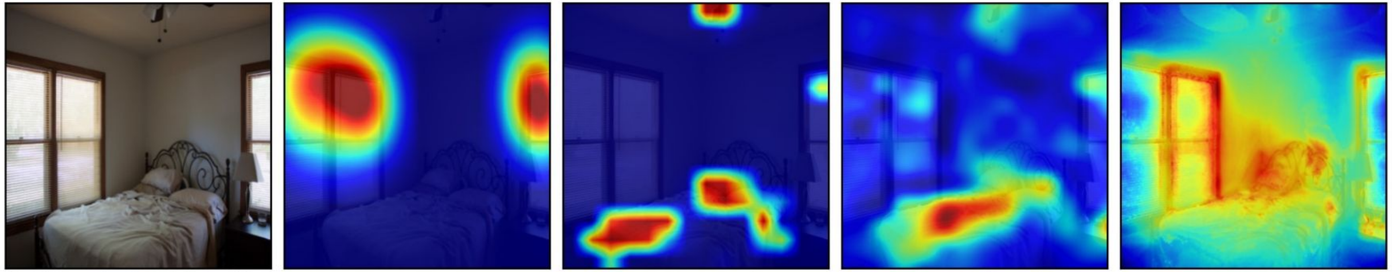
1.



2.



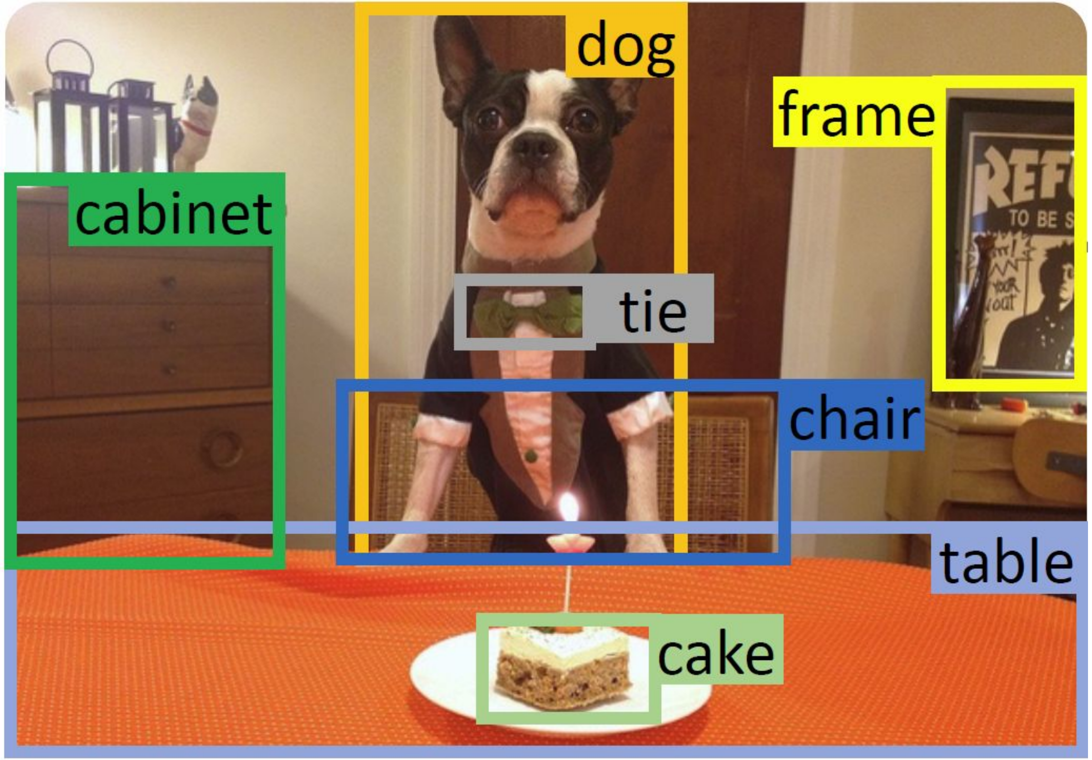
3.



What is covering the windows? blinds | Human Attention | SAN-2 (Yang et al.) | HieCoAtt-Q (Lu et al.) | Judd et al.

Das, Agrawal, Zitnick, Parikh, Batra. 2016

Proposed Method

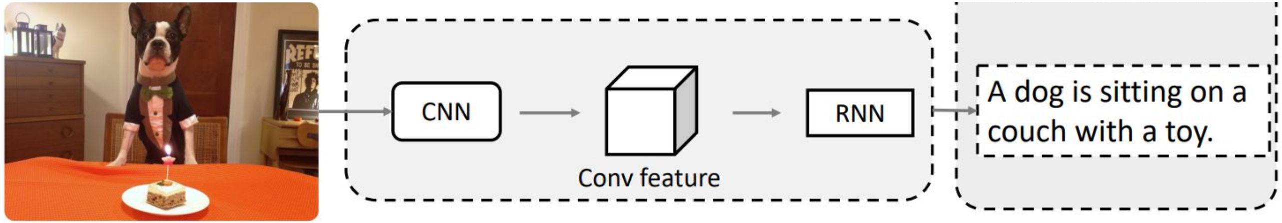


Source: Neural Baby Talk by Jiasen Lu et. al.

- Use object recognition to bolster image captioning
- 11% increase in average precision on COCO dataset in the last year
- Encourages visual grounding (i.e., associates named concepts to pixels in the image)

Proposed Method

Encoder-
Decoder



Natural but
inaccurate

Source: Neural Baby Talk by Jiasen Lu et. al.

Related Work

Slot Filling
- Farhadi et al., 2010
- **Kulkarni et al., 2011**

Obje
- A
- V
- A

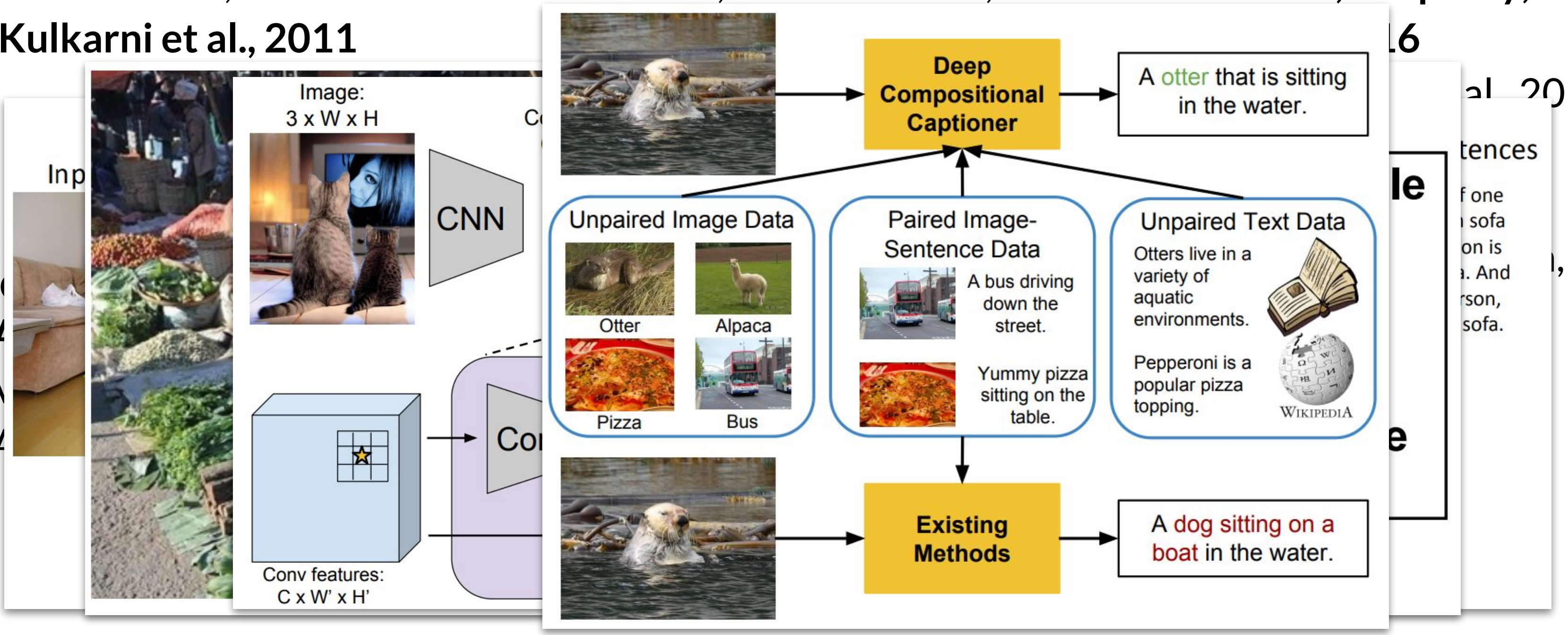
Encoder-Decoder
- Kiros, Salakhutdinov, &

Grounded Language
- **Johnson, Karpathy, &**

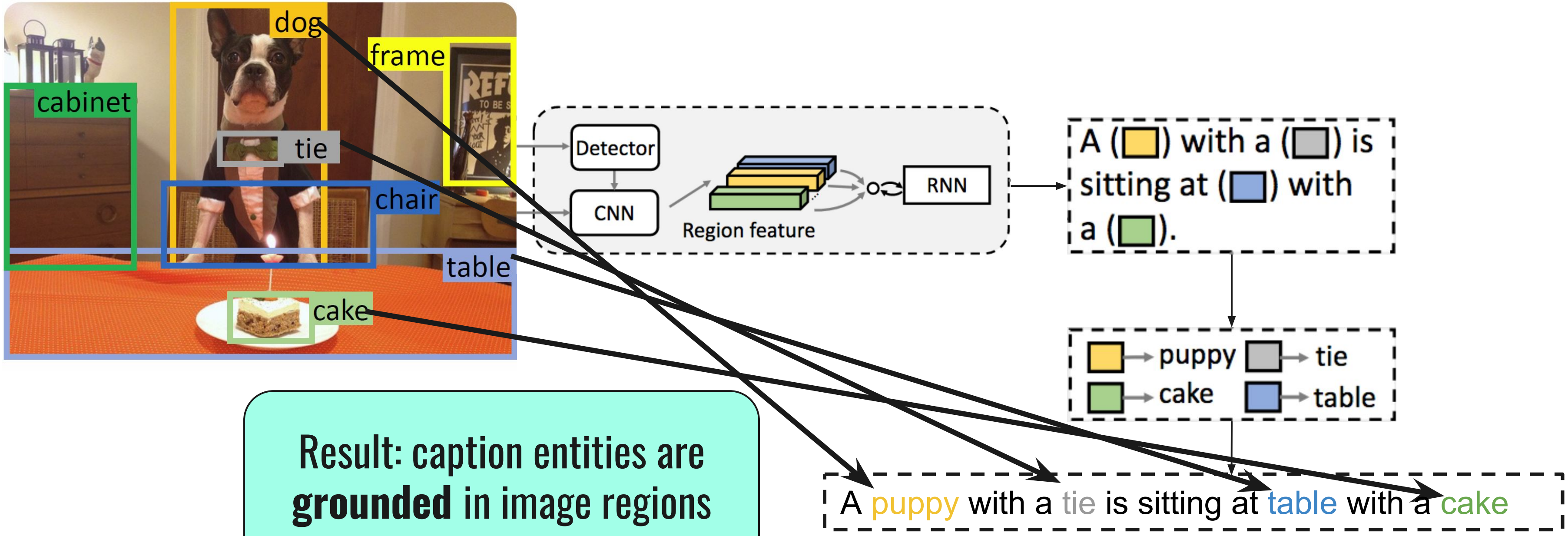
16

al., 2014

, 2017



High Level Methodology



Result: caption entities are **grounded** in image regions

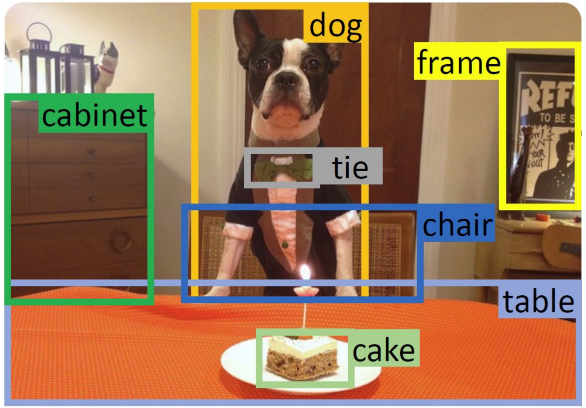
Image source: Neural Baby Tal

Slotted Caption Template Generation

1. Use CNN for object detection:
    a. Detected objects and their bounding boxes become ***candidate grounding regions***
2. Use CNN to generate feature map of input image (as in Assignment 3)
3. Use RNN to generate ***caption template***
    a. Initialize hidden state with CNN image features
    b. Training time: each input is ground truth caption word
    c. Inference time: each input is sampled from previous output
    d. Slots for *visual words* are generated using a pointer network [give blackbox description of Ptr-Net]
        i. Pointer networks map (point) tokens in the output sequence to tokens in the input sequence

Slotted Caption Template Generation

In:



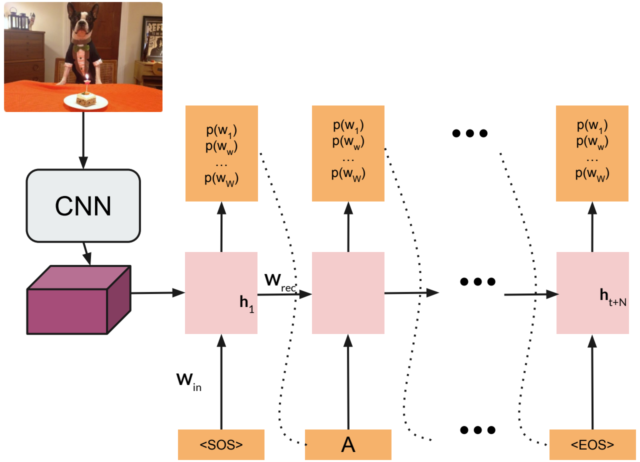
1) Obtain candidate grounding regions:
   cabinet, dog, tie, chair, table, cake, frame
2) Use a (Ptr-Net) RNN to generate caption template:
   a) Given candidate regions, Ptr-Net “points” from a token in the caption to an associated image region.

- **Note:** Ptr-Net can be used whenever we want to map output elements back to input elements *exactly*.

Out: A <region-2> with a <region-3> is sitting at <region-4> with a <region-5>

Pointer Networks (Vinyals O, Fortunato M, Jaitly N. NIPS 2015)

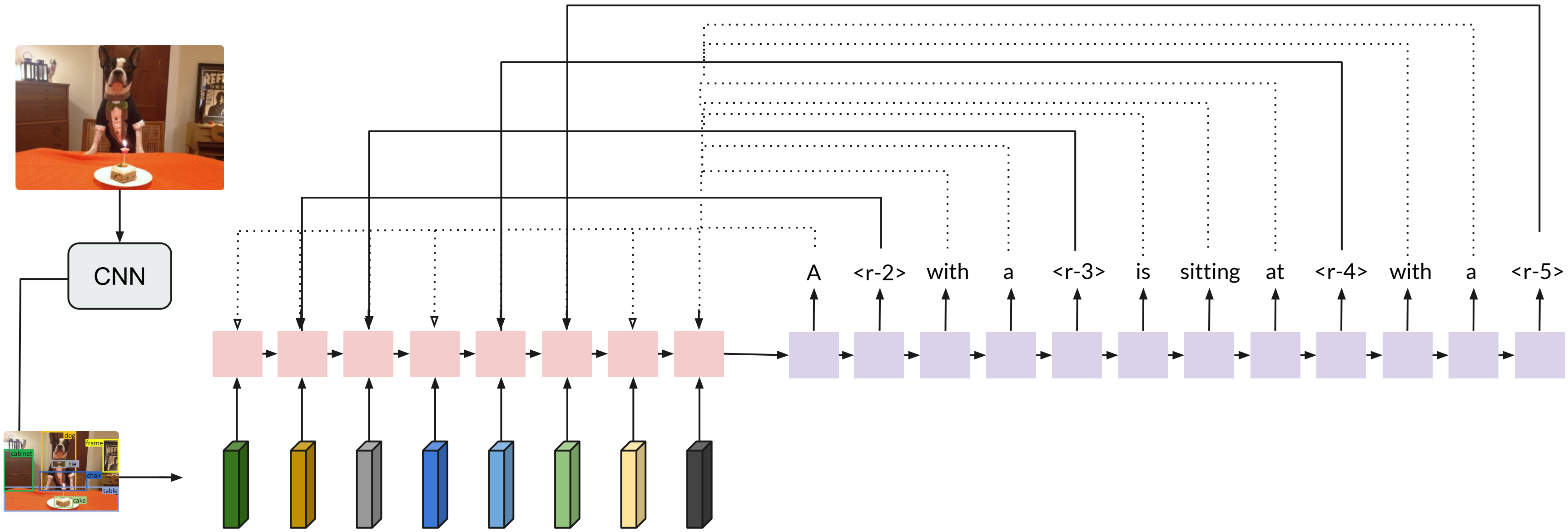
2) Use a (Ptr-Net) RNN to generate caption template



[sequence-to-sequence (Assignment 3)]

Pointer Network (Vinyals O, Fortunato M, Jaitly N. NIPS 2015)

2) Use a (Ptr-Net) RNN generate caption template

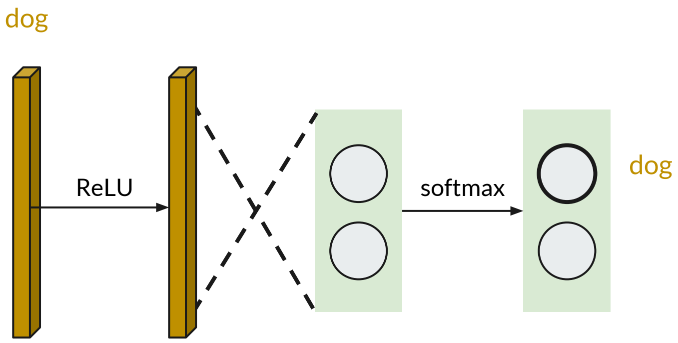


[Ptr-Net (Neural Baby Talk example)]

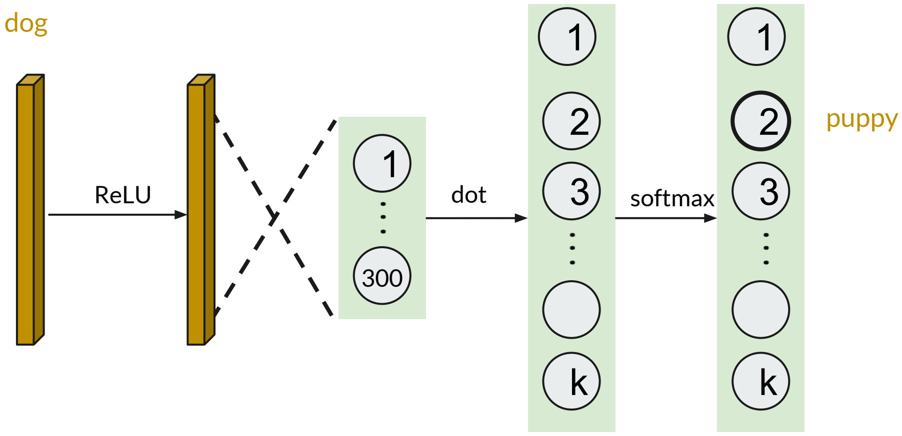
Filling in the Slots

In: - A <region-2> with a <region-3> is sitting at <region-4> with a <region-5>
- cabinet, dog, tie, chair, table, cake, frame (coarse names from object detector)

1) Classify Plurality



2) Determine Fine Grained Category



Out: A puppy with a tie is sitting at table with a cake

Objective

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log \Big(p(y_t^* | \tilde{r}, \boldsymbol{y}_{1:t-1}^*) p(\tilde{r} | \boldsymbol{y}_{1:t-1}^*) \mathbb{1}_{(y_t^* = y^{\text{txt}})} + p\left(b_t^*, s_t^* | \boldsymbol{r}_t, \boldsymbol{y}_{1:t-1}^*\right) \Big(\frac{1}{m} \sum_{i=1}^{m} p\left(r_t^i | \boldsymbol{y}_{1:t-1}^*\right) \Big) \mathbb{1}_{(y_t^* = y^{\text{vis}})} \Big)$$

Text word probability

Probability of predicting the correct text word, given the sentinel features, and the previous ground truth words and probability of generating the sentinel features given ground truth caption

Caption refinement prob.

Probability of predicting correct plurality and fine-grained name given image region features and previous ground truth words in caption

Averaged target region probability

Probability of produced the anchored image grounding region given previous characters of ground truth caption

Training: minimize this cross-entropy loss

Evaluation

Datasets

Flickr30k: 31,783 images, 5 captions per image, 275,555 annotated bounding boxes
COCO: 164,062 images, 5 captions per image

Object category to words

For COCO dataset. (e.g., mapping <person> to ["child", "baker", ...])

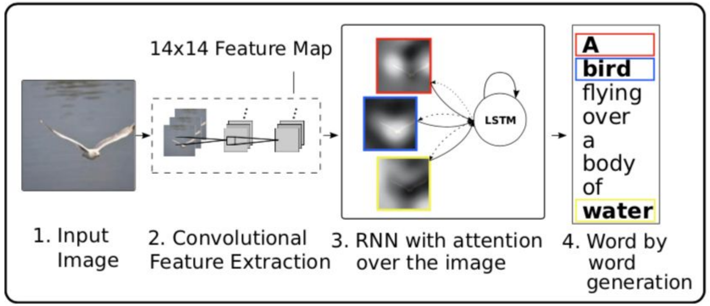
Caption pre-processing

Caption truncation (if > 16 words)
Building vocabulary (9,587 words for COCO, 6,864 words for Flickr30k)

Evaluation

1. Standard Image Captioning

BLEU: precision
METEOR: averaged precision and recall
CIDEr: averaged cosine similarity
SPICE: defined over scene graphs



Flickr30k dataset

Method	BLEU1	BLEU4	METEOR	CIDEr	SPICE
Hard-Attention [46]	66.9	19.9	18.5	-	-
ATT-FCN [50]	64.7	23.0	18.9	-	-
Adaptive [27]	67.7	25.1	20.4	53.1	14.5
NBT	**69.0**	**27.1**	**21.7**	**57.5**	**15.6**
NBT$^{\text{oracle}}$	72.0	28.5	23.1	64.8	19.6

COCO dataset

Method	BLEU1	BLEU4	METEOR	CIDEr	SPICE
Adaptive [27]	74.2	32.5	26.6	**108.5**	19.5
Att2in [39]	-	31.3	26.0	101.3	-
Up-Down [3]	74.5	33.4	26.1	105.4	19.2
Att2in* [39]	-	33.3	26.3	111.4	-
Up-Down† [3]	79.8	36.3	27.7	120.1	21.4
NBT	**75.5**	**34.7**	**27.1**	107.2	**20.1**
NBT$^{\text{oracle}}$	75.9	34.9	27.4	108.9	20.4

Xu et al. (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

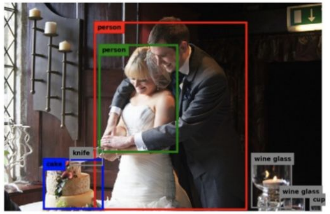
source: Neural Baby Talk by Jiasen Lu et. al.

Evaluation

Success | **Failure**

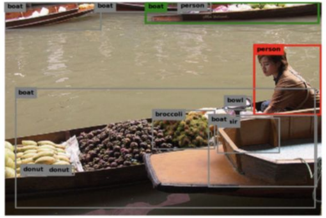
COCO

A dog is laying in the grass with a Frisbee.



A bride and groom cutting a cake together.

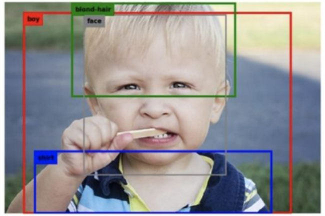
A little girl holding a cat in her hand.



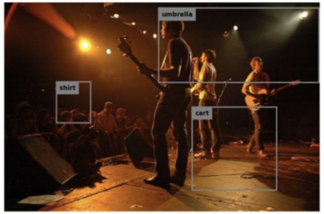
A woman sitting on a boat in the water.

Flickr30k

A cat is standing on a sign that says “UNK”.



A young boy with blond-hair and a blue shirt is eating a chocolate



A band is performing on a stage.

Two people are sitting on a boat in the water.

* Different colours show a correspondence between the visual words and grounding regions.

* Grey regions are the proposals not selected in the captions.

source: Neural Baby Talk by Jiasen Lu et. al.

Evaluation

2. Robust Image Captioning

To evaluate image captioning for novel scene compositions

Robust-COCO split

- Distribution of co-occurring objects in train data is different from test data
- Calculate the co-occurrence statistics for 80 object categories
- Sufficient examples from each category in train set
- Novel compositions (pairs) of categories in test set

Accuracy

- Whether or not a generated caption includes the new object combination
- 100% accuracy for at least one mention of the novel category pair

Evaluation

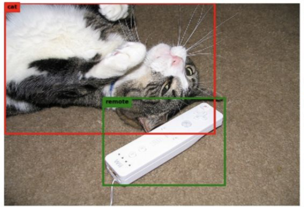
Robust-COCO split

: worse (2~3 points drop) performance for all models

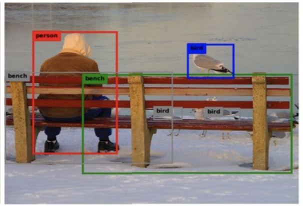
COCO dataset with Robust split

Method	BLEU4	METEOR	CIDEr	SPICE	Accuracy
Att2in [39]	31.5	24.6	90.6	17.7	39.0
Up-Down [3]	31.6	25.0	92.0	18.1	39.7
NBT	**31.7**	**25.2**	**94.1**	**18.3**	**42.4**
NBT$^{\text{oracle}}$	31.9	25.5	95.5	18.7	45.7

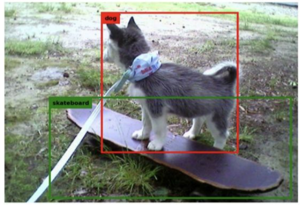
Success



A cat laying on the floor next to a remote control.

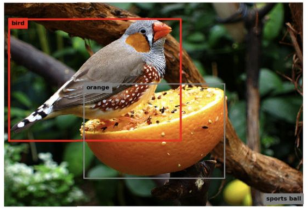


A man sitting on a bench next to a bird.



A dog is standing on a skateboard in the grass.

Failure



A bird sitting on a branch in a tree.

Image source: Neural Baby Talk by Jiasen Lu et. al.

Evaluation

3. Novel Object Captioning
Excludes all the image-sentence pairs that contain at least one of the eight objects in COCO (“bottle”, “bus”, “couch”, “microwave”, “pizza”, “racket”, “suitcase”, and “zebra”)

Test set is split into in-domain and out-of-domain subsets

F1 score
: checks if the excluded object is correctly mentioned in the generated caption

Evaluation

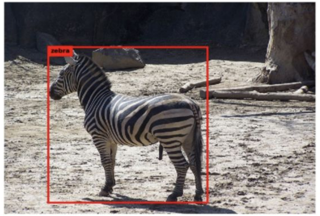
Method	Out-of-Domain Test Data												In-Domain Test Data		
	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg	SPICE	METEOR	CIDEr	SPICE	METEOR	CIDER
DCC [4]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8	13.4	21.0	59.1	15.9	23.0	77.2
NOC [43]	17.8	68.8	25.6	24.7	69.3	68.1	39.9	89.0	49.1	-	21.4	-	-	-	-
C-LSTM [49]	29.7	74.4	38.8	27.8	68.2	70.3	44.8	91.4	55.7	-	23.0	-	-	-	-
Base+T4 [2]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0	15.9	23.3	77.9	18.0	24.5	86.3
NBT*+G	7.1	73.7	34.4	61.9	59.9	20.2	42.3	88.5	48.5	15.7	22.8	77.0	17.5	24.3	87.4
NBT†+G	14.0	74.8	42.8	63.7	74.4	19.0	44.5	92.0	53.2	16.6	23.9	84.0	**18.4**	25.3	94.0
NBT†+T1	36.2	77.7	43.9	65.8	70.3	19.8	51.2	93.7	57.3	16.7	23.9	85.7	**18.4**	**25.5**	**95.2**
NBT†+T2	**38.3**	**80.0**	**54.0**	**70.3**	**81.1**	**74.8**	**67.8**	**96.6**	**70.3**	**17.4**	**24.1**	**86.0**	18.0	25.0	92.1

Table 4. Evaluation of captions generated using the proposed method. G means greedy decoding, and T1−2 means using constrained beam search [2] with 1−2 top detected concepts. $*$ is the result using VGG-16 [41] and $\dagger$ is the result using ResNet-101.

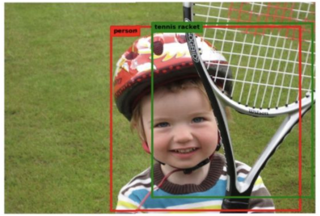
source: Neural Baby Talk by Jiasen Lu et. al.

Evaluation

Success



A zebra that is standing in the dirt.



A little girl wearing a helmet and holding a tennis racket.

A woman standing in front of a red bus.

Failure

A plate of food with a bottle and a cup of beer.

Image source: Neural Baby Talk by Jiasen Lu et. al.

Conclusion

A novel image captioning framework

: natural language + grounded in detected objects

Two-stage approach

: 1) generate hybrid template

: 2) fills the slots with categories recognized by object detector

NBT outperforms the state-of-art models on standard, robust, and novel object captioning

Limitations

<donut>	donut, doughnut, bagel
<cake>	cake, cheesecake, cupcake, shortcake, coffeecake, pancake
<bird>	bird, ostrich, owl, seagull, goose, duck, parakeet, falcon, robin, pelican, waterfowl, heron, hummingbird, mallard, finch, pigeon, sparrow, seabird, osprey, blackbird, fowl, shorebird, woodpecker, egret, chickadee, quail, bluebird, kingfisher, buzzard, willet, gull, swan, bluejay, flamingo, cormorant, parrot, loon, gosling, waterbird, pheasant, rooster, sandpiper, crow, raven, turkey, oriole, cowbird, warbler, magpie, peacock, cockatiel, lorikeet, puffin, vulture, condor, macaw, peafowl, cockatoo, songbird
<chair>	chair, seat, recliner, stool
<couch>	couch, sofa, recliner, futon, loveseat, settee, chesterfield
<potted plant>	potted plant, houseplant
<bed>	bed
<dining table>	dining table, table
<toilet>	toilet, urinal, commode, lavatory, potty
<tv>	tv, monitor, televison, television
<laptop>	laptop, computer, notebook, netbook, lenovo, macbook
<mouse>	mouse
<remote>	remote

the CNN to

2. Not clear how useful fined grained category name assignment is.
    a. In general, authors could have compared performance with and without this sub-model

Table source: Neural Baby Talk by Jiasen Lu et. al.

Image source: CPSC 532S lecture slides

Possible Extensions

1. Compare NBT performance with and without end-to-end training of the CNN
2. Perform object detection in model that maximizes accuracy-specificity tradeoff.
    a. Critical for real world tasks that authors use as their motivation (ex: helping visually impaired)
        i. Likely harm COCO evaluation metrics -> points to the need for a new metric for this real world task
    b. Accomplish this by:
        i. Pretrain CNN pre-trained on ImageNet
        ii. Classify according to semantic hierarchy (eliminating part of slot filling model)
            1. Already organized by WordNet hierarchy
            2. See my course project for doing this with modern CNN architectures

Thank you.

Questions?

Back-up Slides

Compared Models

Hard-attention

: Attention-based image caption

(“soft” or “hard” attention)

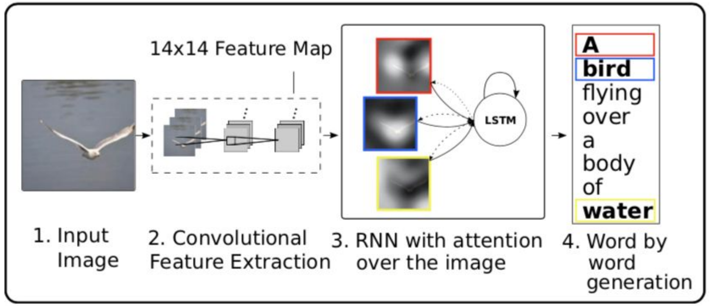
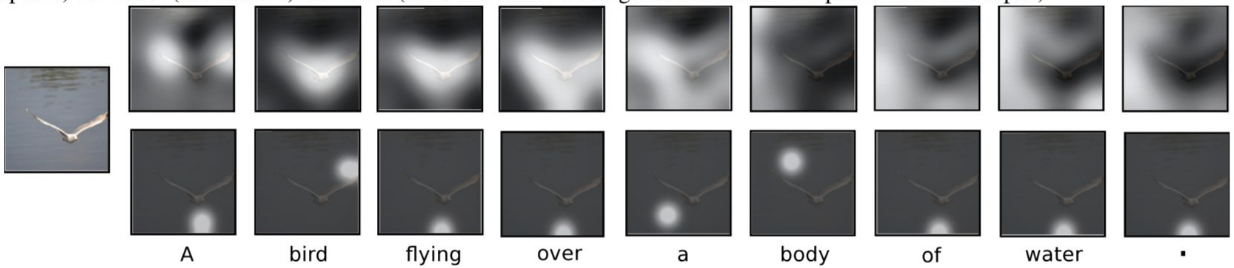


Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. “soft” (top row) vs “hard” (bottom row) attention. (Note that both models generated the same captions in this example.)

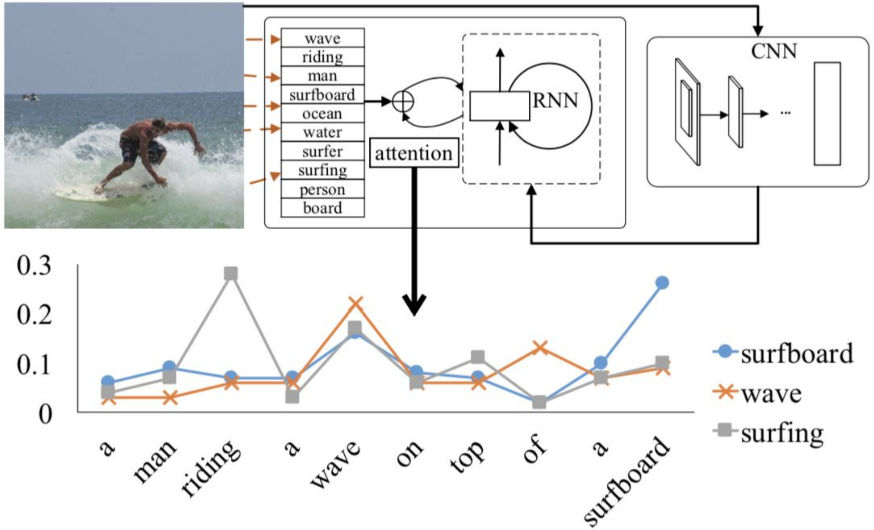


Xu et al. (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Compared Models

ATT-FCN

: Attention-based image caption
(semantically important regions)



You et al. (2016). Image Captioning with Semantic Attention

Compared Models

Adaptive
: Attention-based model with a visual sentinel
(“when” to look at + “which” region)

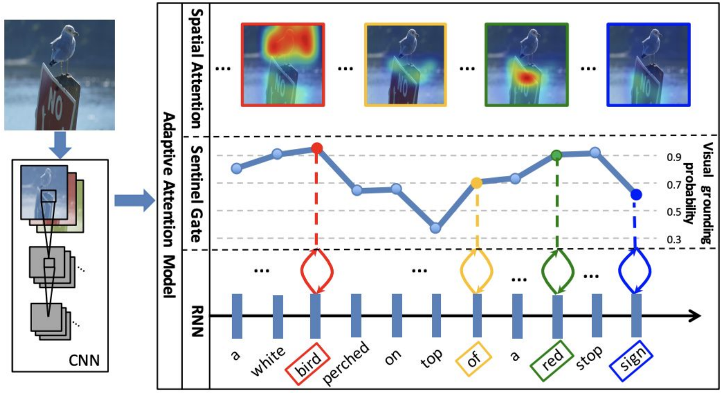


Figure 1: Our model learns an adaptive attention model that automatically determines when to look (**sentinel gate**) and where to look (**spatial attention**) for word generation, which are explained in section 2.2, 2.3 & 5.4.

Lu et al. (2017). Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning