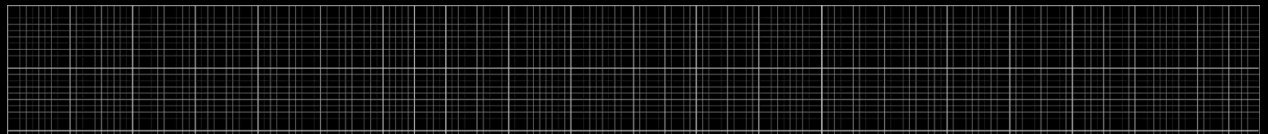
Neural Baby Talk (with Faster- RCNN)

VIRGINIA TECH™

Presentation by Himanshu Jahagirdar



A Preview

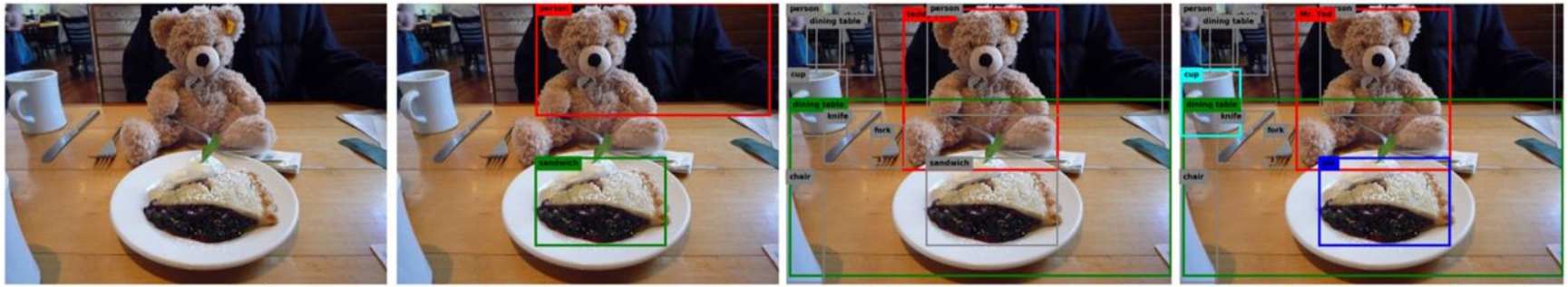
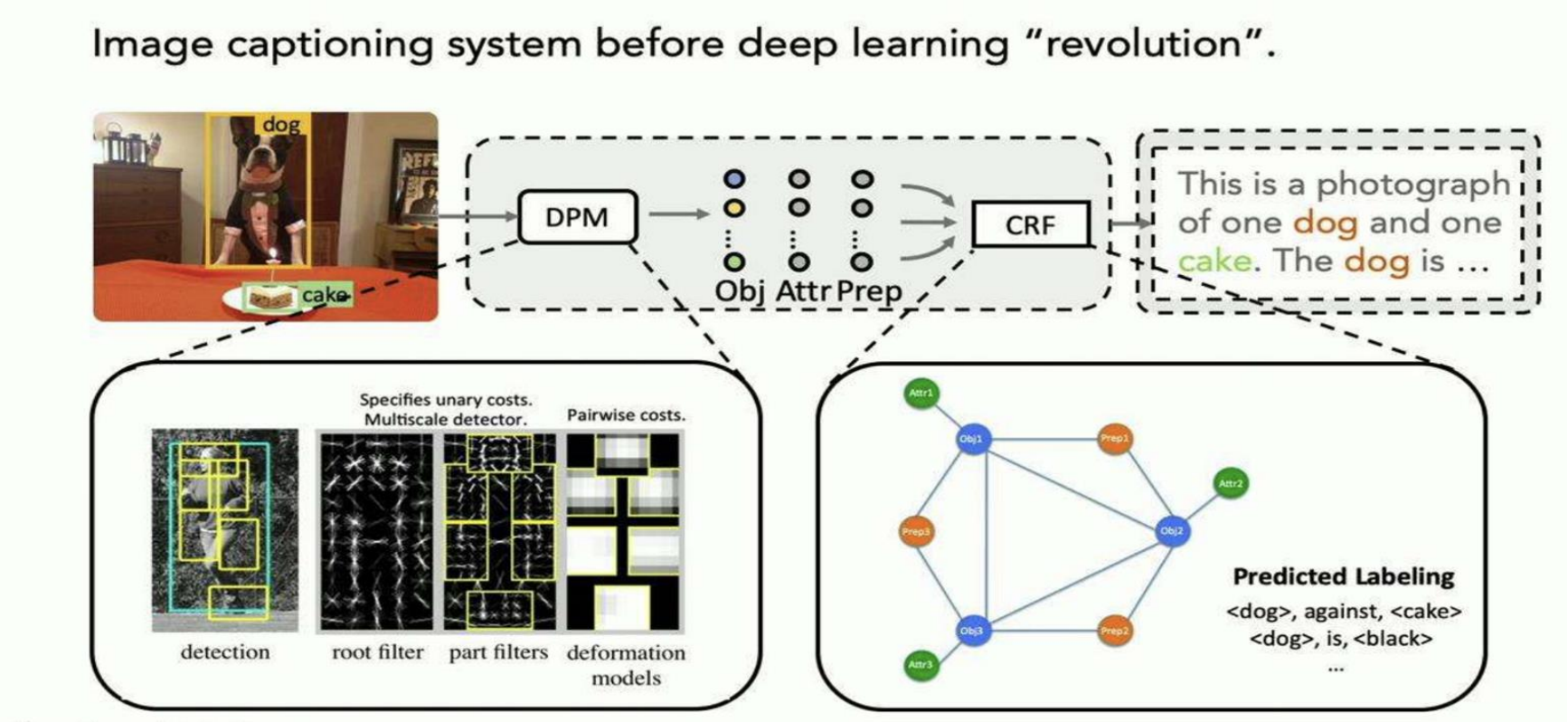


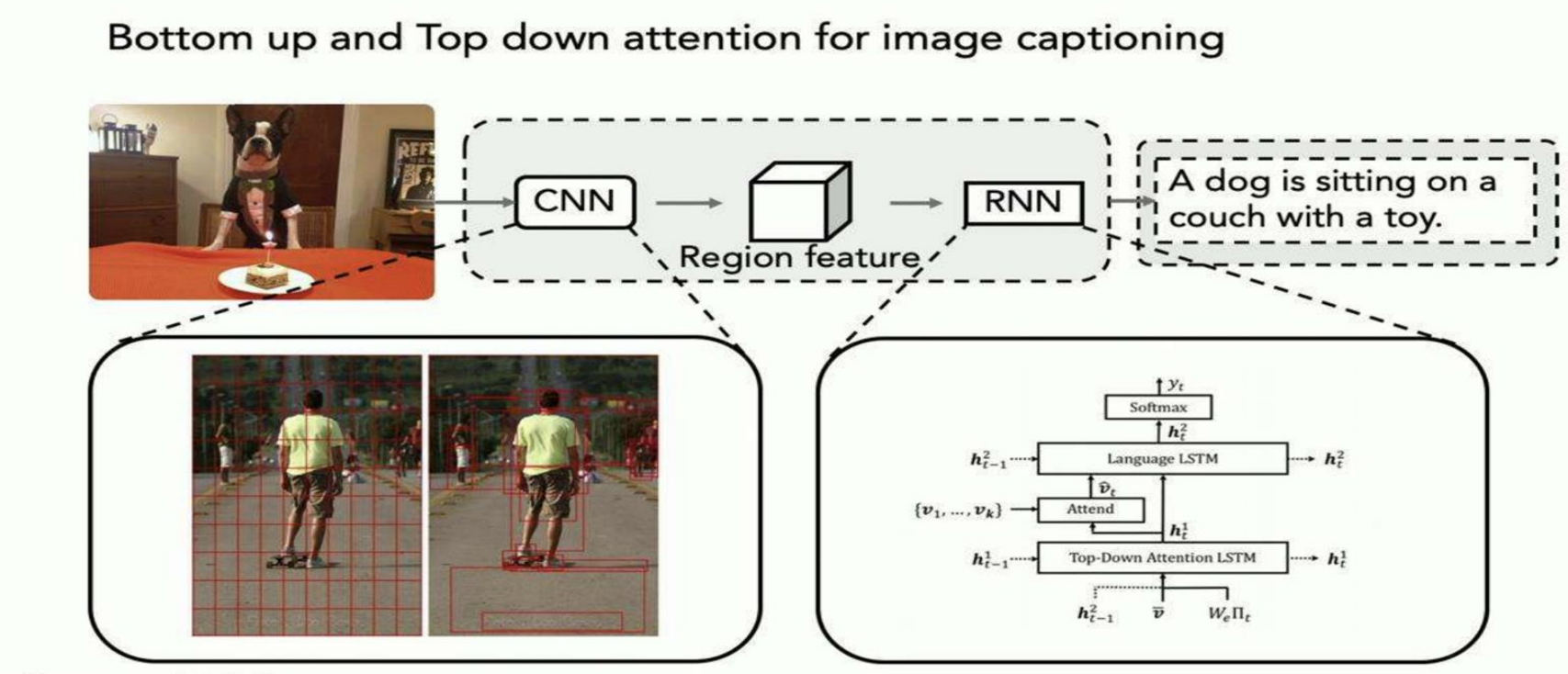
Figure 2. From left to right is the generated caption using the same captioning model but with different detectors: 1) No detector; 2) A weak detector that only detects "person" and "sandwich"; 3) A detector trained on COCO [26] categories (including "teddy bear"). 4) A detector that can detect novel concepts (e.g. "Mr. Ted" and "pie" that never occurred in the captioning training data). Different colors show a correspondence between the visual word and grounding regions.

Related Work - Baby Talk



[Kulkarni et. al 2011]

Related Work - Image Captioning



[Anderson et.al. 2018]

Related Work - Learning from object detection

SOTA neural image captioning system.

Two elephants and a baby elephant walking together.

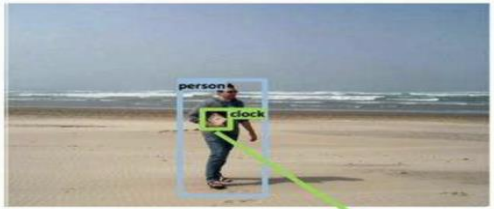
A cat is standing on a sign that says "UNK".

A man standing on a beach holding a surfboard.

Learn from object detection and OCR.

Two elephants and a baby elephant walking together.

A cat is standing on a sign that says "~~UNK~~"-"Abundzu".



A man standing on a beach holding a ~~surfboard~~ - clock.

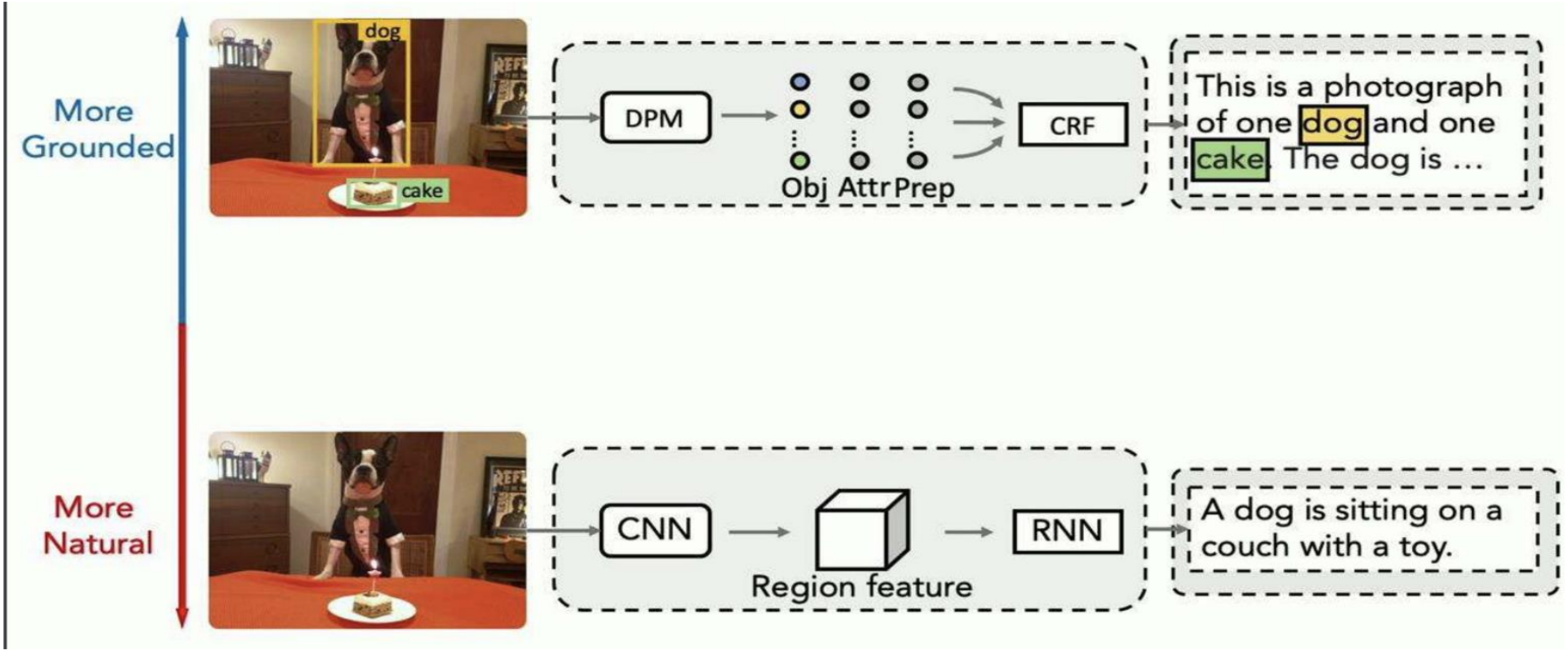
[Jiasen Lu Microsoft]

Motivation -

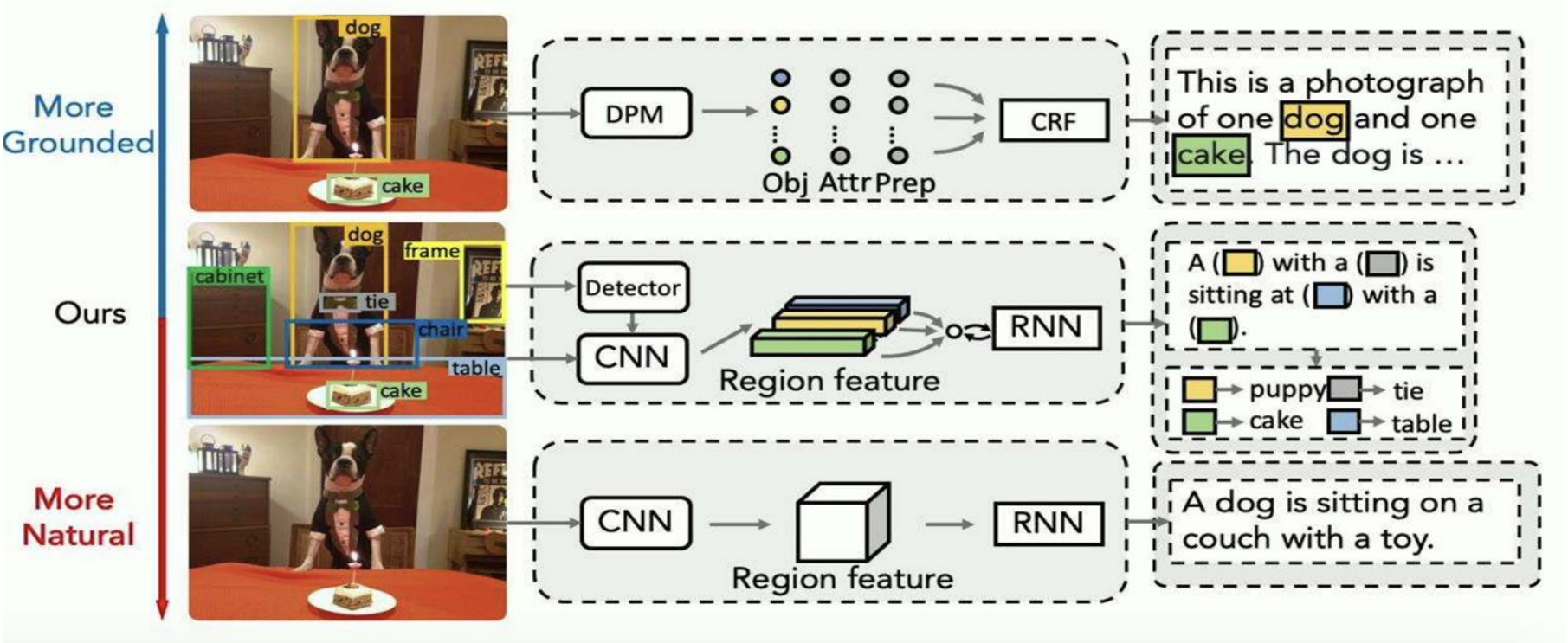
- Image Captioning is important
  - Aid for the visually impaired
  - Personal Assistants

- SOTA 2018 (Attention) based models
  - Improved image captioning performance
  - Novel object captioning is hard: Impossible to create training set that includes image-caption pairs for all concepts seen everyday
  - SOTA Models lack visual grounding

Motivation - Problems with classic slot-filling approaches



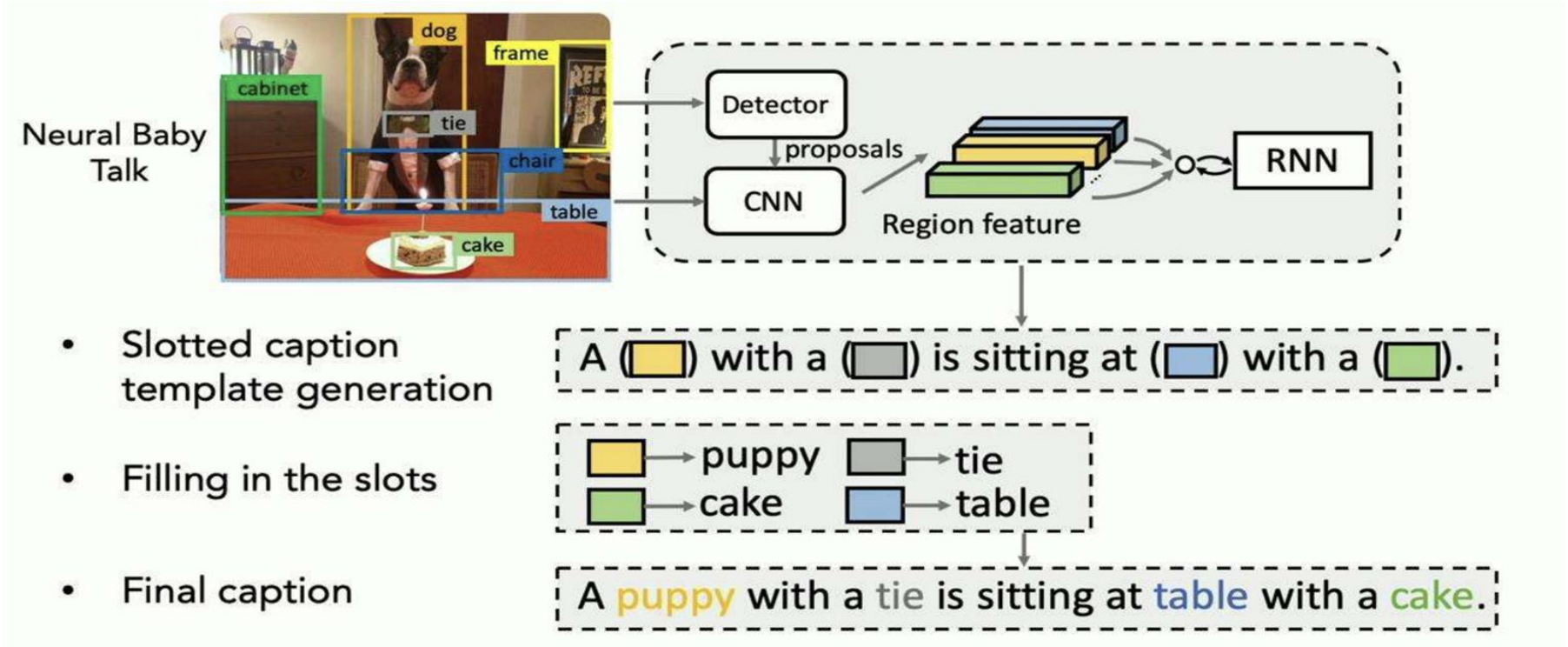
Motivation - Neural Baby Talk



Methodology

1. Create visually grounded captions by splitting the objective in 2 parts
    a. Maximize probability of generating the template
    b. Learn a model to find visual words to fill in the template slots
2. Define a latent vector r_t which can be
    a. Visual word y_{vis}
    b. Template word y_{txt}
3. Use a pre-trained Faster-RCNN to generate template captions
4. Fill in the slots using a standard detection framework

Methodology



- Slotted caption template generation
- Filling in the slots
- Final caption

Methodology - Generating Grounded Caption Templates

1. Model and Likelihood of correct caption - r_t is the latent vector

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{(\boldsymbol{I},\boldsymbol{y})} \log p(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\theta}) \qquad (1)$$

$$p(y_t|\boldsymbol{y}_{1:t-1},\boldsymbol{I}) = p(y_t|r_t,\boldsymbol{y}_{1:t-1},\boldsymbol{I})p(r_t|\boldsymbol{y}_{1:t-1},\boldsymbol{I}) \quad (3)$$

1. Generating Slotted Captions via a pointer network
   a. v_t is the region feature of r_t : computed using Faster-RCNN
   b. h_t= RNN(x_t, h_{t-1}) , P: distribution over grounding regions

$$u_i^t = \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_v \boldsymbol{v}_t + \boldsymbol{W}_z \boldsymbol{h}_t) \qquad (4)$$

$$\boldsymbol{P}_{\boldsymbol{r_I}}^t = \text{softmax}(\boldsymbol{u}^t) \qquad (5)$$

Faster RCNN

1. **RCNN** :

The R-CNN consists of 3 main modules:

1. The first module generates 2,000 region proposals using the **Selective Search** algorithm.
2. After being resized to a fixed pre-defined size, the second module extracts a feature vector of length 4,096 from each region proposal.
3. The third module uses a pre-trained SVM algorithm to classify the region proposal to either the background or one of the object classes.

Faster RCNN

2. **Fast RCNN** : Using fixed ROI using ROI pooling layer

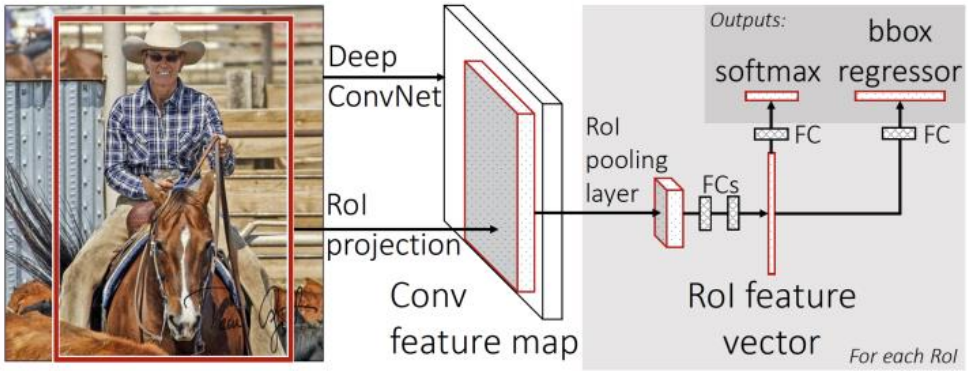


Image copyright Girshick, 2015

- Start with a small set of candidate ROIs (a few hundred per image)
- Each ROI feeds a neural net whose output is a 4-vector specifying the (x,y,w,h) of the nearest object.

Faster RCNN

3. **Faster RCNN** :

a. Instead of using pyramid of images or kernels, introduces anchor boxes
b. Introduces a region proposal network that generates proposals with various scales and aspect ratios.
c. Convolution computations are shared with RPN and Fast R-CNN.

Faster RCNN

3. **Faster RCNN** :

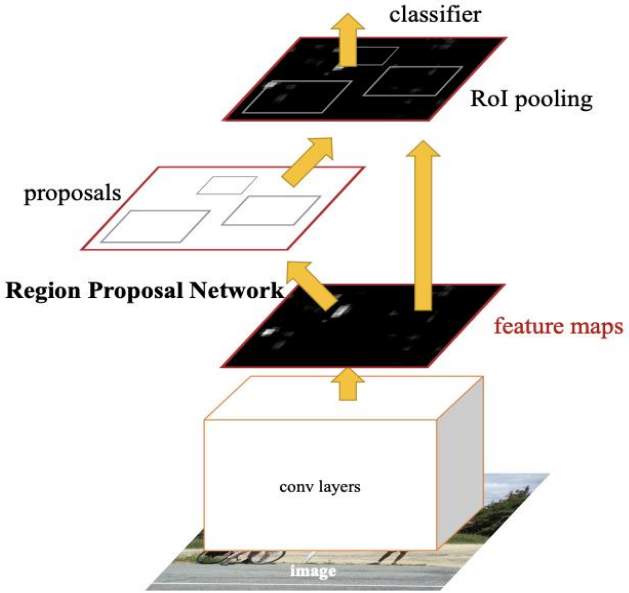


Figure 2: Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

"Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Ren, He, Girshick & Sun, 2016

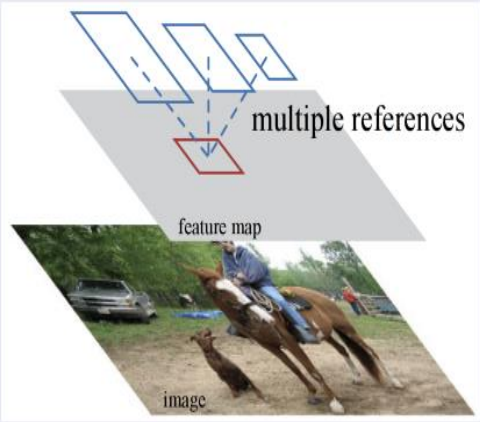


Image copyright Ren, He, Girchick & Sun, 2016

- Each candidate bounding box computes 9 different regression outputs, each of which is a 4-vector (x,y,w,h)
- The 9 different regression outputs from each bbox are w.r.t. 9 different "anchor" rectangles, each offset from the input ROI. Thus:

$$\text{anchor} = \text{ROI} + \text{known shift}$$

$$\text{object} = \text{anchor} + \text{regression}$$

Faster RCNN

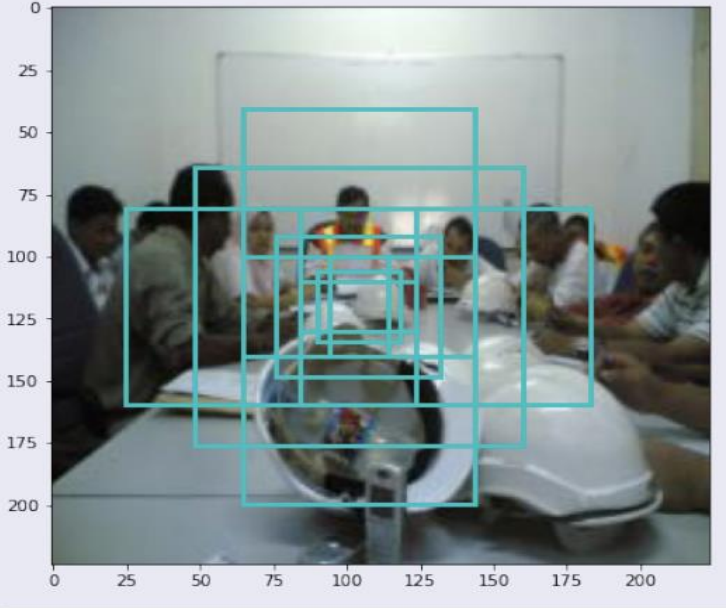
3. **Faster RCNN** :

3 sizes, 3 aspect ratios

The Faster RCNN paper described 9 anchors per ROI:

- 3 different anchor sizes: 128×128, 256×256, and 512×512.
- 3 different aspect ratios: $1:2$, $1:1$, and $2:1$

9 anchors per ROI



Methodology - Generating Grounded Caption Templates

3. **Visual Sentinels** - r̂ serves as dummy grounding for visual word. Therefore, Probability of a textual word is:

$$p(y_t^{txt}|\boldsymbol{y}_{1:t-1}) = p(y_t^{txt}|\tilde{r}, \boldsymbol{y}_{1:t-1})p(\tilde{r}|\boldsymbol{y}_{1:t-1}) \quad (6)$$

Visual sentinels s_t can be obtained using an LSTM:

$$\boldsymbol{g}_t = \sigma\left(\boldsymbol{W}_x\boldsymbol{x}_t + \boldsymbol{W}_h\boldsymbol{h}_{t-1}\right) \quad (7)$$

$$\boldsymbol{s}_t = \boldsymbol{g}_t \odot \tanh\left(\boldsymbol{c}_t\right) \quad (8)$$

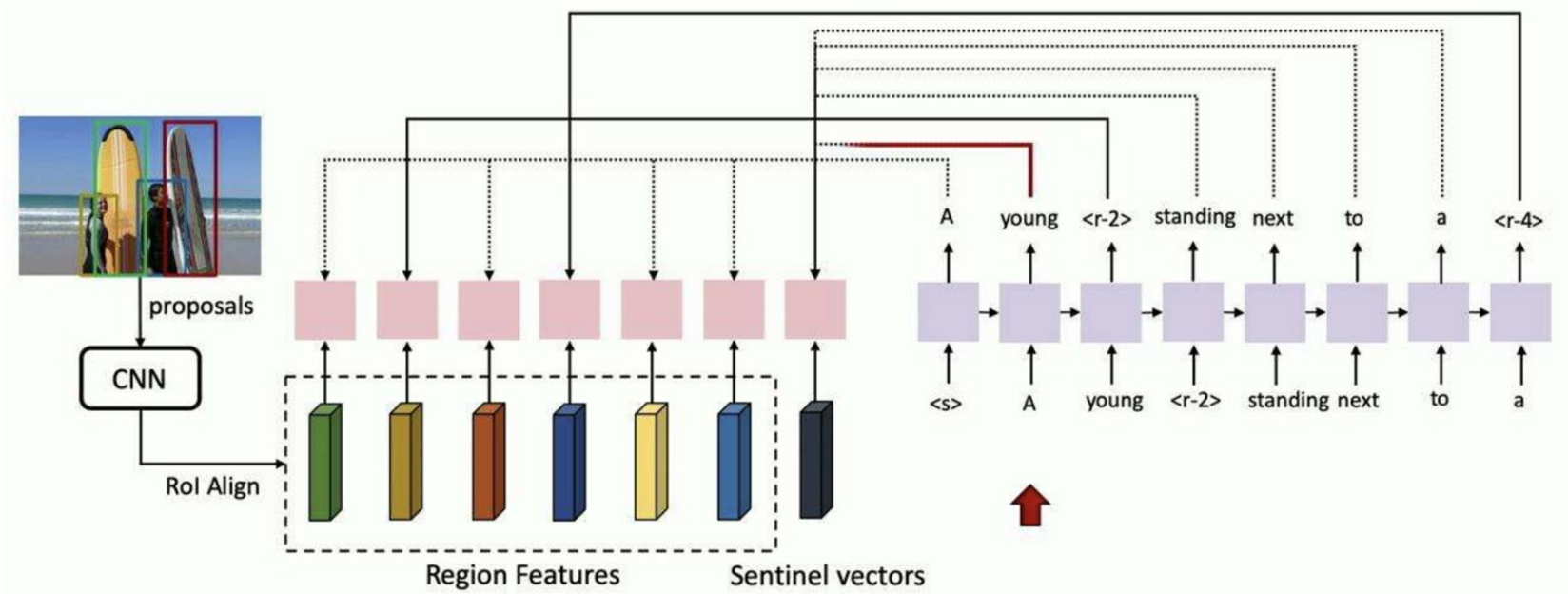
4. Therefore, pointer network equation becomes

$$\boldsymbol{P}_{\boldsymbol{r}}^t = \text{softmax}([\boldsymbol{u}^t; \boldsymbol{w}_h^T \tanh(\boldsymbol{W}_s\boldsymbol{s}_t + \boldsymbol{W}_z\boldsymbol{h}_t)]) \quad (9)$$

Methodology - Visualizing the pointer network

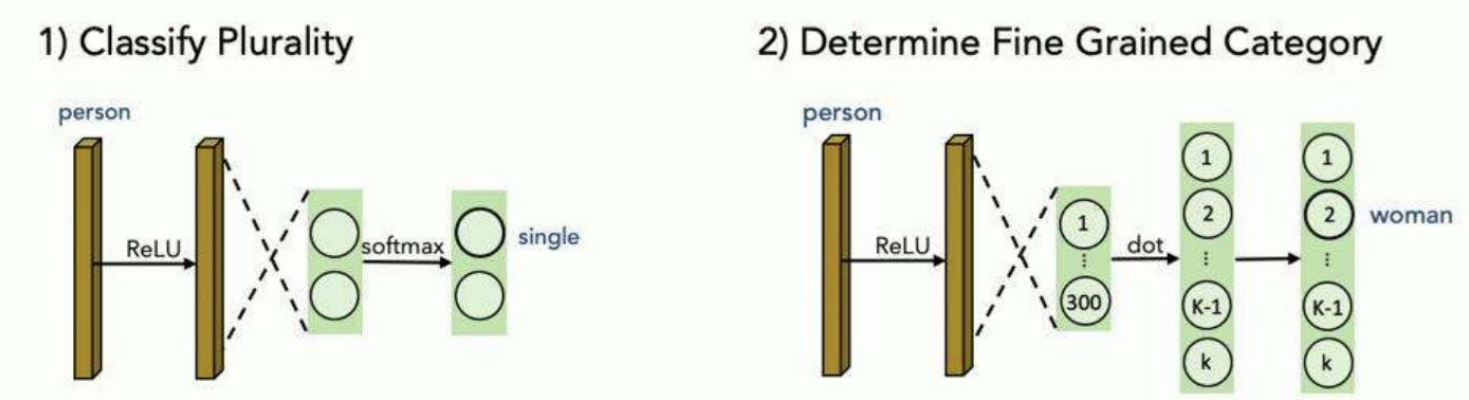
5. Finally, probability of textual words conditioned on the image is

$$\boldsymbol{P}_{txt}^{t} = \text{softmax}\left(\boldsymbol{W}_q \boldsymbol{h}_t\right) \qquad (10)$$



Methodology - Caption Refinement

6. **Object Detection Network to fill in slots** -



$$\boldsymbol{P}_b^t = \text{softmax}\left(\boldsymbol{W}_b f_b\left([\boldsymbol{v}_t; \boldsymbol{h}_t]\right)\right) \quad (11)$$

$$\boldsymbol{P}_g^t = \text{softmax}\left(\boldsymbol{U}^T \boldsymbol{W}_g f_g\left([\boldsymbol{v}_t; \boldsymbol{h}_t]\right)\right) \quad (12)$$

Methodology - Final Objective

Minimize this cross-entropy loss

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

Template word prob.

Refinement prob.

target region prob.

Methodology - Final Objective

Minimize this cross-entropy loss

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

Template word prob. | **Refinement prob.** | **target region prob.**

Co-reference when different kind of supervision exists.

A young woman is sitting inside a boat.

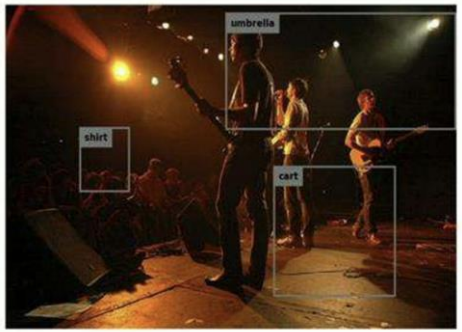
Problem: which boat is the caption referred to.

Solution: maximize the averaged target region prob.

Methodology - Visual word extraction

In case there are no bounding boxes corresponding to target label, then the model dynamically predicts the textual word instead- hence avoids problems even when the wrong label is predicted.

How to handle inaccurate object detection?



A man is playing guitar on the stage.

- Treat object labels as caption template.

  Template: *A man is playing guitar on the stage.*

- Dynamically identify the visual words.

  IoU(*detected bbox, GT bbox*) > 0.5 *and labels == words*

Dataset details

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., “person” mapping to [“child”, “baker”, …])

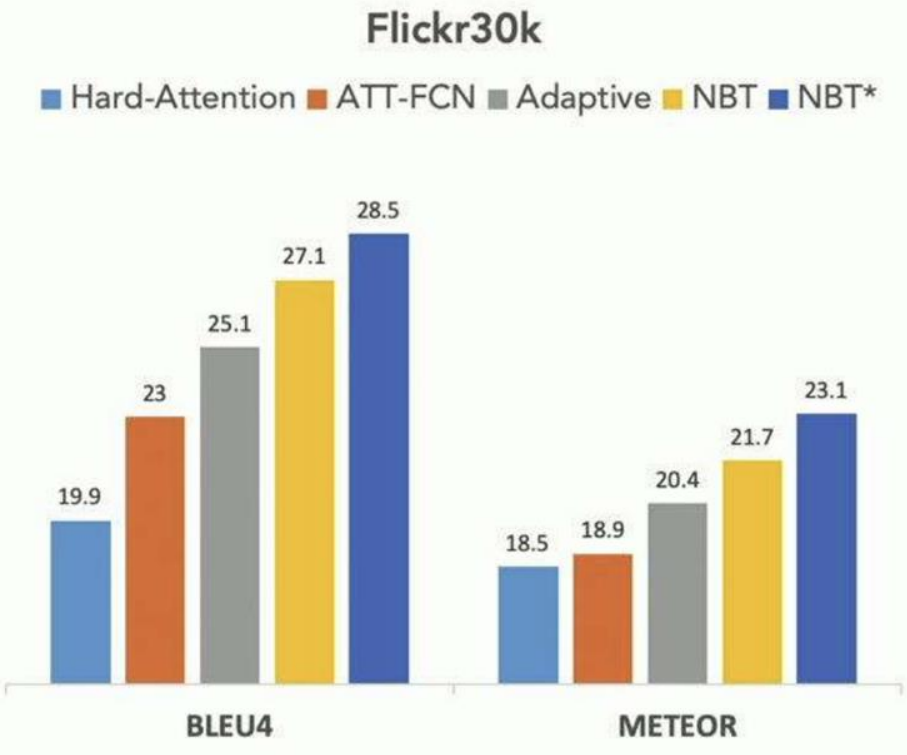
Caption pre-processing

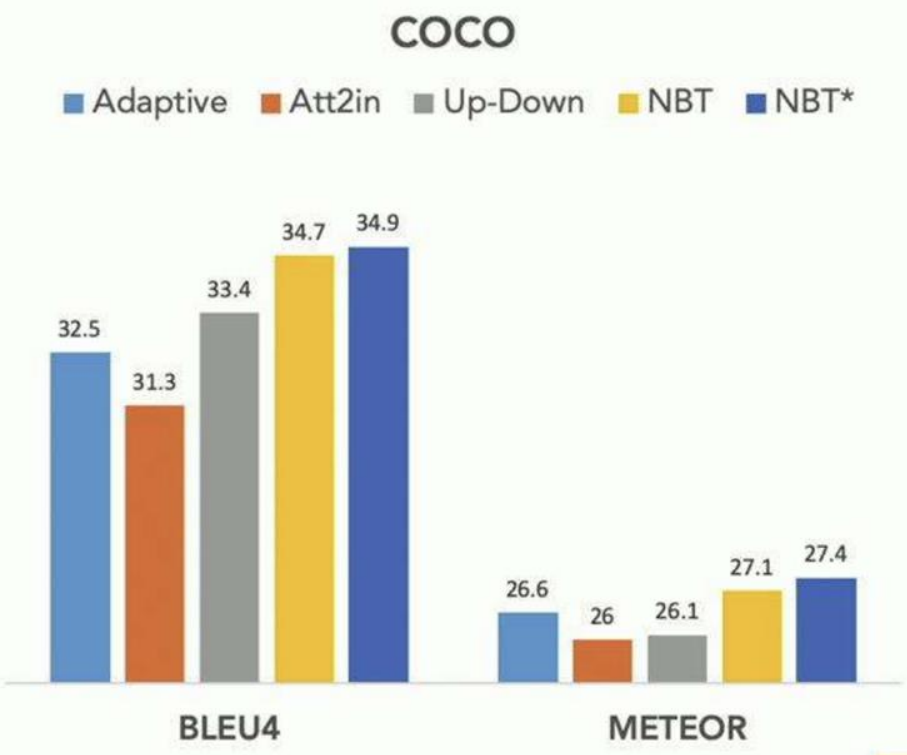
- Caption truncation (if > 16 words)
- Building vocabulary (9,587 words for COCO, 6,864 words for Flickr30k)

Results - Standard image captioning

1. Report test performances on image captioning task on COCO and Flickr 30K datasets
2. Use additional setting NBT Oracle - better object detector
3. Evaluation metrics used are
    a. BLEU : Captures the amount of n-gram overlap between the output sentence and the reference ground truth sentence.
    b. METEOR: Precision-based metric to measure quality of generated text. Allows synonyms and stemmed words to be matched with the reference word
    c. CIDEr:
    d. SPICE : F-score over Tuples.

Results - Standard image captioning





Results - Standard image captioning- Flickr30K

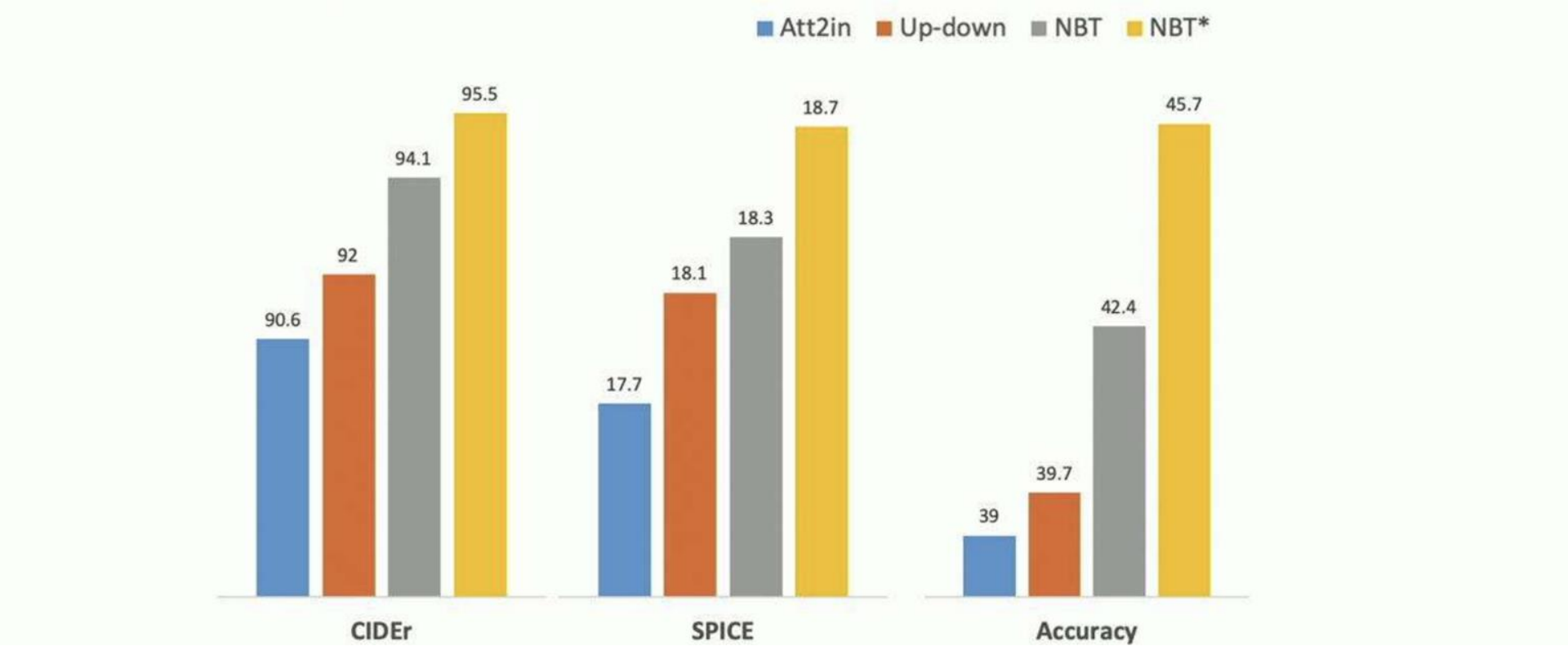
Method	BLEU1	BLEU4	METEOR	CIDEr	SPICE
Hard-Attention [46]	66.9	19.9	18.5	-	-
ATT-FCN [50]	64.7	23.0	18.9	-	-
Adaptive [27]	67.7	25.1	20.4	53.1	14.5
NBT	**69.0**	**27.1**	**21.7**	**57.5**	**15.6**
$\text{NBT}^{\text{oracle}}$	72.0	28.5	23.1	64.8	19.6

Table 1. Performance on the test portion of Karpathy *et al.* [20]'s splits on Flickr30k Entities dataset.

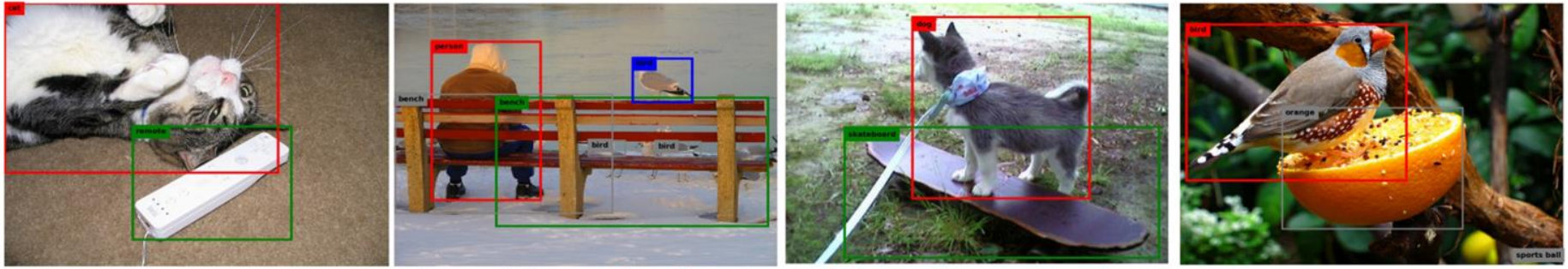
Results - Standard image captioning- COCO

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

Results - Robust Image Captioning



Results - Robust Image Captioning



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

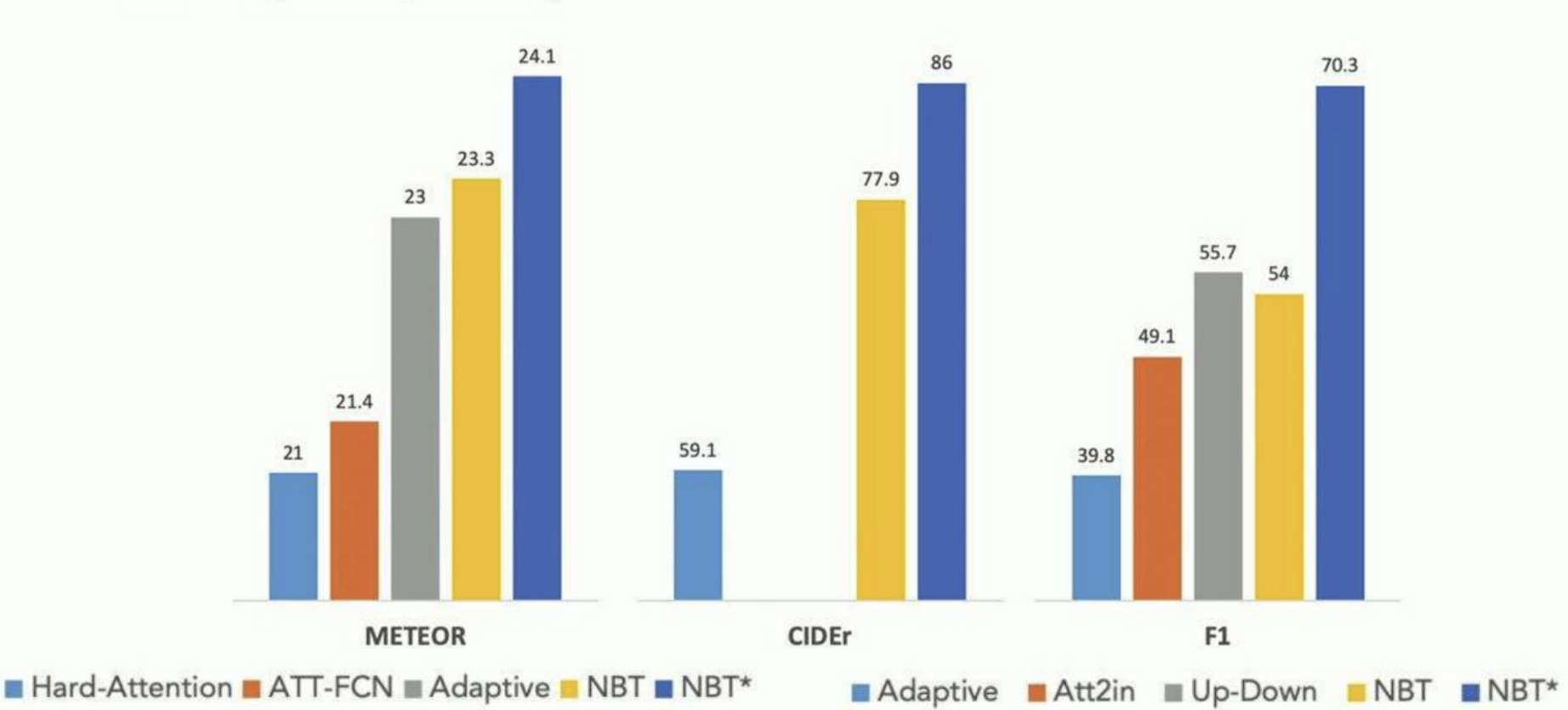
A man sitting on a bench next to a bird.

A dog is standing on a skateboard in the grass.

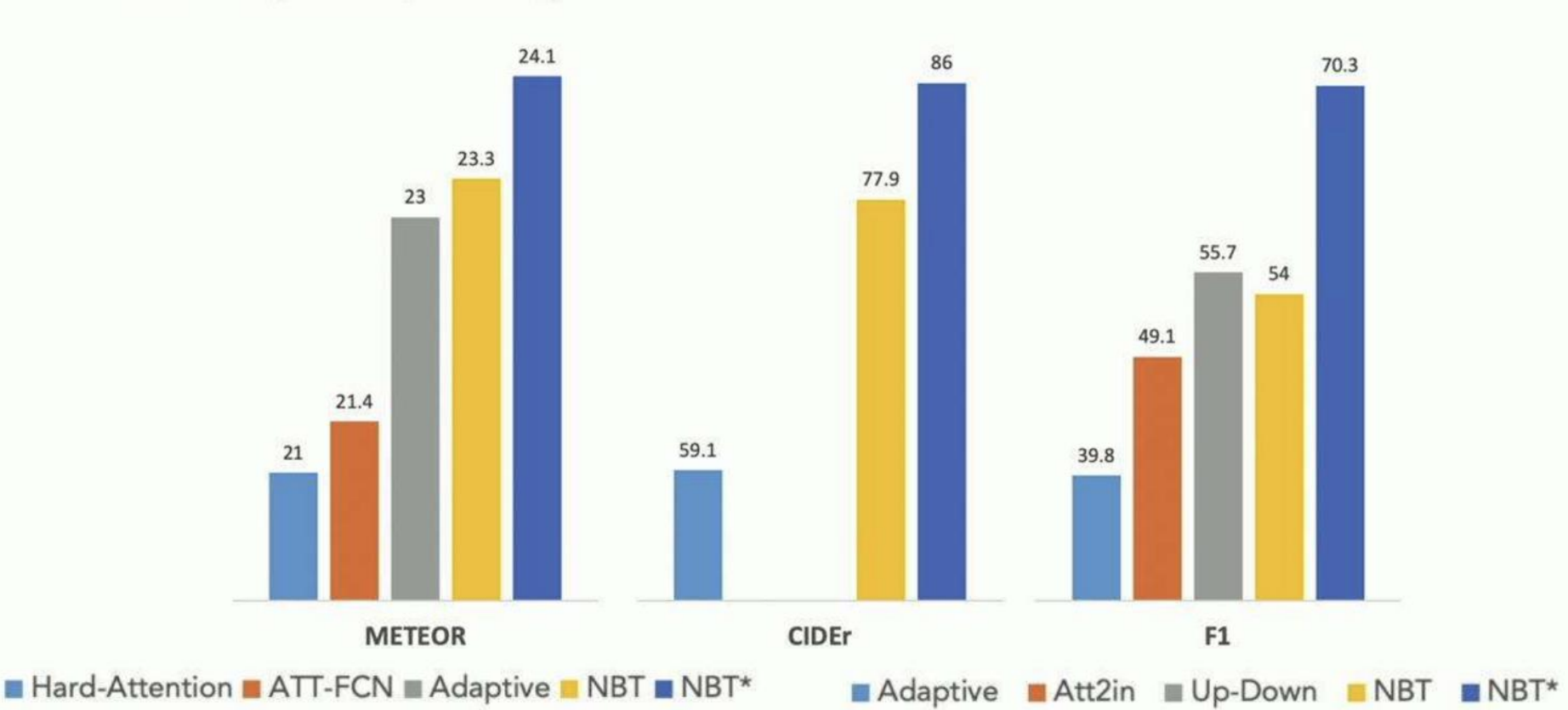
A bird sitting on a branch in a tree.

Figure 6. Generated captions and corresponding visual grounding regions for the robust image captioning task. “cat-remote”, “man-bird”, “dog-skateboard” and “orange-bird” are co-occurring categories excluded in the training split. First 3 columns show success and last column shows failure case (orange was not mentioned).

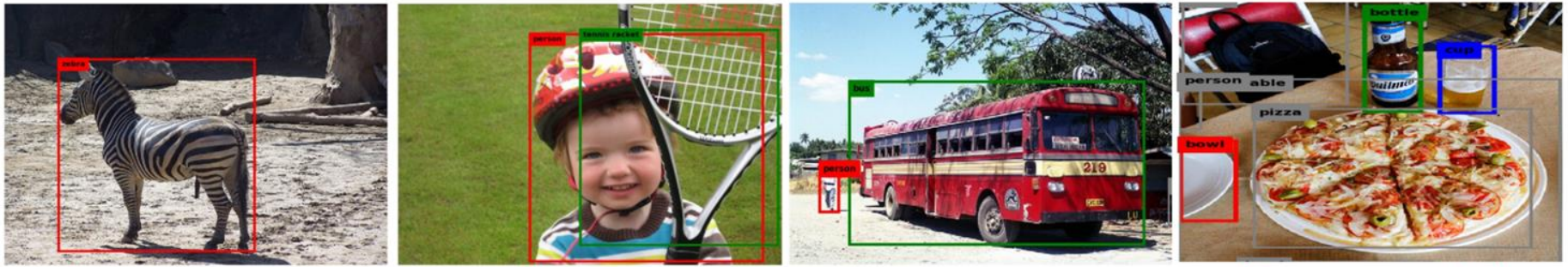
Results - Novel Object Captioning



Results - Novel Object Captioning



Results - Novel Object Captioning



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.

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

Figure 7. Generated captions and corresponding visual grounding regions for the novel object captioning task. "zebra", "tennis racket", "bus" and "pizza" are categories excluded in the training split. First 3 columns show success and last column shows a failure case.

Strengths

1. Beats the state of the art using a complex and well-thought out solution

2. Provides unique template based approach that takes advantage of natural language while grounding to visual features

2. Results are more promising in the novel object captioning section

Weaknesses

1. Only ground on noun words found by object detector (no actions)

2. Natural Language limited to image captioning dataset

References

1. https://courses.engr.illinois.edu/ece417/fa2020/slides/lec10.pdf
2. https://www.microsoft.com/en-us/research/uploads/prod/2019/11/Visually-Grounded-Language-Understanding-and-Generation-SLIDES.pdf
3. https://blog.paperspace.com/faster-r-cnn-explained-object-detection/
4. https://arxiv.org/pdf/1506.01497.pdf