Partially-Supervised Image Captioning

NeurIPS 2018

Peter Anderson Macquarie University (now Georgia Tech)
Stephen Gould Australian National University
Mark Johnson Macquarie University

Code release: www.panderson.me/constrained-beam-search

1. The Novel Object Captioning Task

- Describe images containing novel objects (not present in the available image-caption training data) by learning from image labels or object annotations
- **Motivation**: Scaling image captioning to many more visual concepts, without collecting expensive caption training data

Training data[1]:

Image-caption data (for 72 COCO classes)

Caption: An old fashioned yellow car waits at a stoplight.

(Classes considered In-Domain at test time)

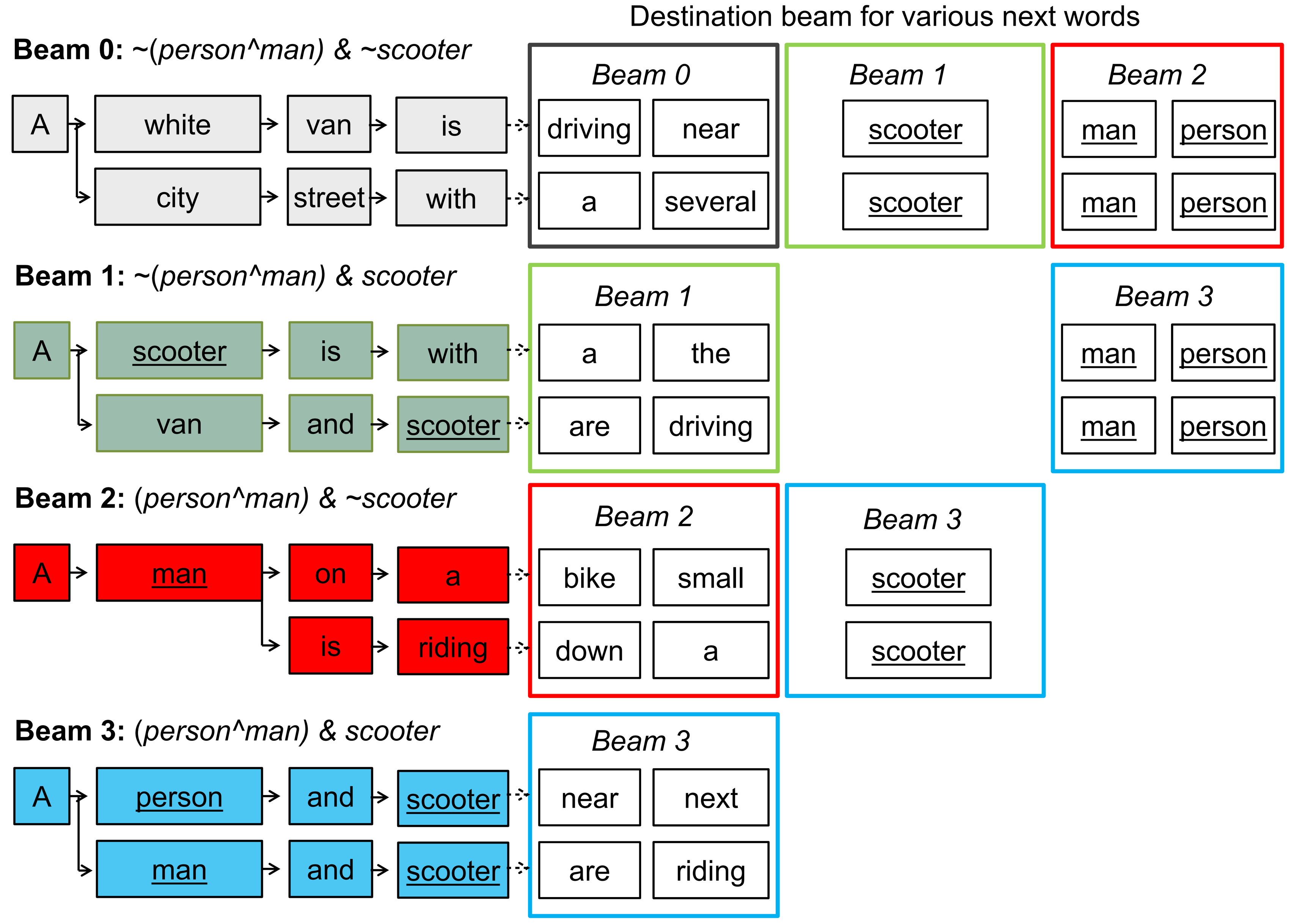
Image label data (for 8 COCO classes)

Labels for this image: person, bus, scooter, van, white, yellow

(Classes considered Out-of-Domain at test time)

3. Constrained Beam Search[2] (CBS)

- CBS decoding example with beam size 2, at time step 4. There is one search beam for each state in the FSA. Beam 3 corresponds to the FSA accepting state s_3.

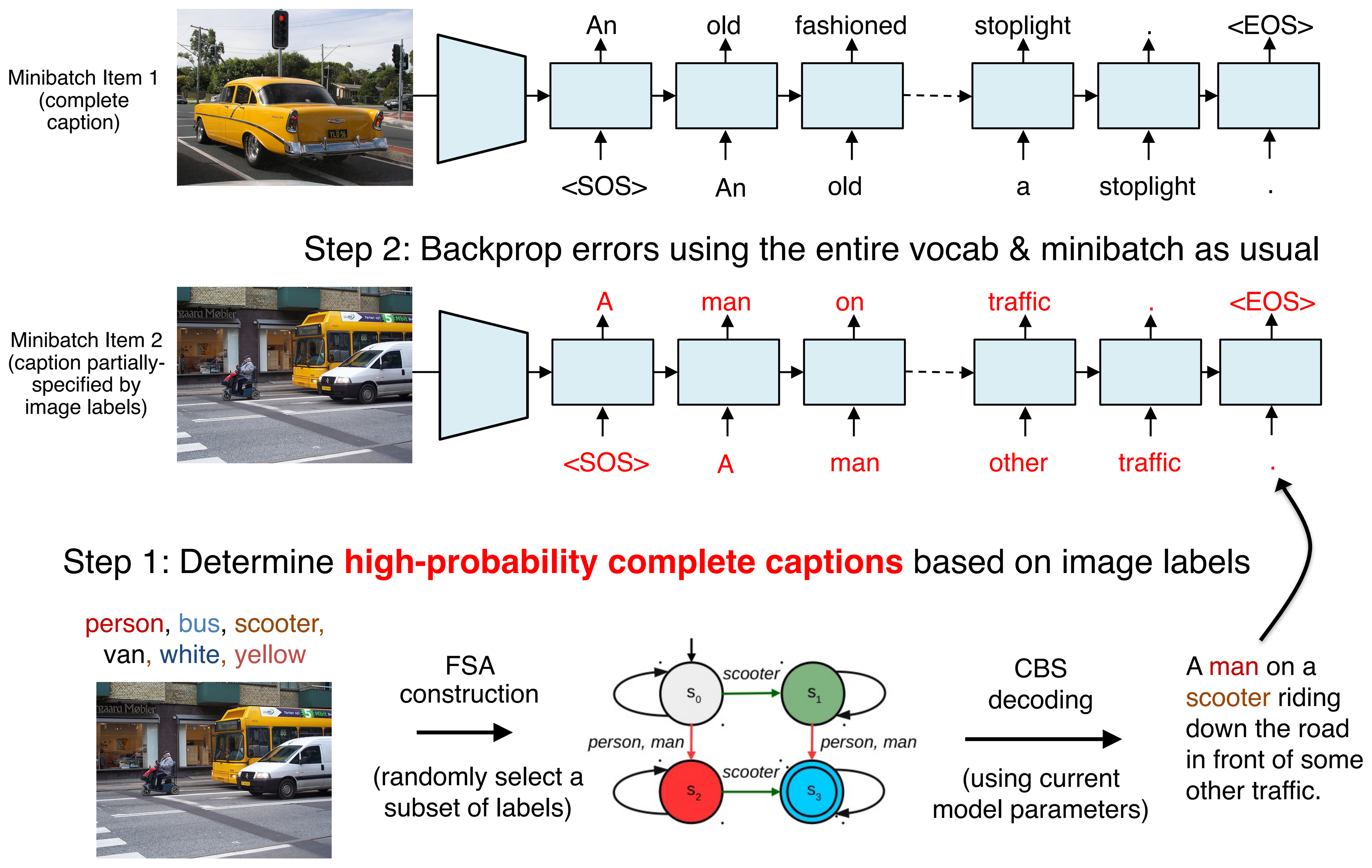


2. Approach: Partially-Specified Sequence Supervision (PS3)

- A general algorithm for training RNNs on partially-specified sequences
- Given a dataset of partially-specified training sequences X and current model parameters θ, iterate these two steps:

Step 1: Estimate the complete data Y by approximating $\boldsymbol{y}^i \leftarrow \mathrm{argmax}_{\boldsymbol{y}}\, p_\theta(\boldsymbol{y}|A^i)\ \forall\, \boldsymbol{x}^i \in X$ using constrained beam search[2] (CBS), where $A^i = (\Sigma, S^i, s_0^i, \delta^i, F^i)$ is an FSA that recognizes sequences that are consistent with the observed partially-specified sequence $\boldsymbol{x}^i$.
Step 2: Learn (or update) the model parameters by setting $\theta \leftarrow \mathrm{argmax}_\theta \sum_{\boldsymbol{y}\in Y} \log p_\theta(\boldsymbol{y})$

Example: A minibatch of images with both complete and partially-specified captions



4. Results and Examples

COCO novel object captioning validation set scores.

	Training Captions	PS3 Labels	CBS Labels	Out-of-Domain Scores SPICE	Out-of-Domain Scores CIDEr	Out-of-Domain Scores F1	In-Domain Scores SPICE	In-Domain Scores CIDEr
Baseline	◐			14.4	69.5	0.0	**19.9**	**108.6**
CBS	◐		▲	15.9	74.8	26.9	19.7	102.4
PS3	◐	●		**18.3**	**94.3**	**63.4**	18.9	101.2
PS3 + CBS	◐	●	▲	18.2	92.5	62.4	19.1	99.5
CBS (GT)	◐		★	18.0	82.5	30.4	22.3	109.7
PS3 + CBS (GT)	◐	●	★	20.1	95.5	65.0	21.7	106.6
Baseline (GT)	●			20.1	111.5	69.0	20.0	109.5

● = full training set, ◐ = impoverished training set, ▲= constrained beam search (CBS) decoding with predicted labels, ★= CBS decoding with ground-truth labels

Imposing FSA constraints during training using PS3 always improves the Out-of-Domain scores

Imposing FSA constraints at test time using CBS adds no further improvement

Performance on the COCO novel object captioning test set. PS3 applied to the Bottom-Up and Top-Down captioning model[4] outperforms all prior work.

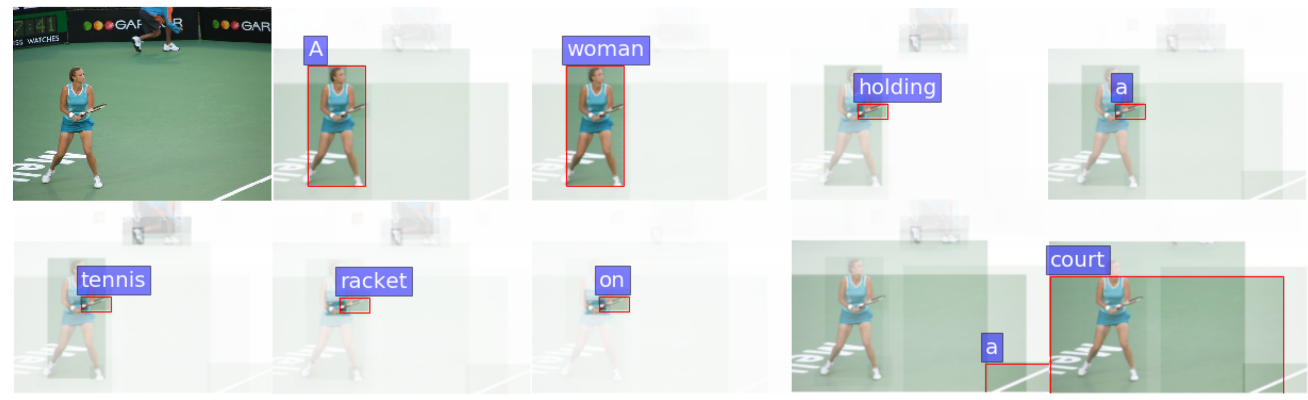
Model	CNN	Out-of-Domain Scores SPICE	Out-of-Domain Scores METEOR	Out-of-Domain Scores CIDEr	Out-of-Domain Scores F1	In-Domain Scores SPICE	In-Domain Scores METEOR	In-Domain Scores CIDEr
DCC	VGG-16	13.4	21.0	59.1	39.8	15.9	23.0	77.2
NOC	VGG-16	-	21.3	-	48.8	-	-	-
C-LSTM	VGG-16	-	23.0	-	55.7	-	-	-
LRCN + CBS	VGG-16	15.9	23.3	77.9	54.0	18.0	24.5	86.3
LRCN + CBS	Res-50	16.4	23.6	77.6	53.3	18.4	24.9	88.0
NBT	VGG-16	15.7	22.8	77.0	48.5	17.5	24.3	87.4
NBT + CBS	Res-101	17.4	24.1	86.0	**70.3**	18.0	25.0	92.1
PS3 (ours)	Res-101	**17.9**	**25.4**	**94.5**	63.0	**19.0**	**25.9**	**101.1**

Baseline: A food truck parked on the side of a road.
Ours: A white bus driving down a city street.

Baseline: A collage of four pictures of food.
Ours: A set of pictures showing a slice of pizza.

Baseline: A zebra is laying down in the grass.
Ours: A tiger that is sitting in the grass.

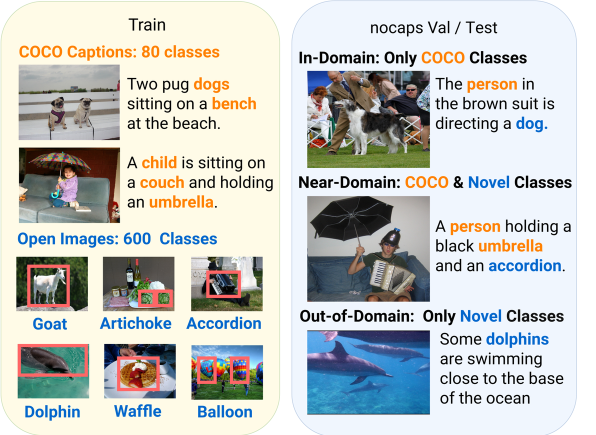
Attention visualization: Novel objects (such as racket) are correctly grounded.



ONGOING WORK

nocaps: novel object captioning at scale

- A large-scale dataset for novel object captioning based on Open Images[3]
- See nocaps.org



References
[1]Hendricks *et al.* Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. CVPR 2016
[2]Anderson *et al.* Guided Open Vocabulary Image Captioning with Constrained Beam Search. EMNLP 2017
[3]Kuznetsova *et al.* The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982, 2018
[4]Anderson *et al.* Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018