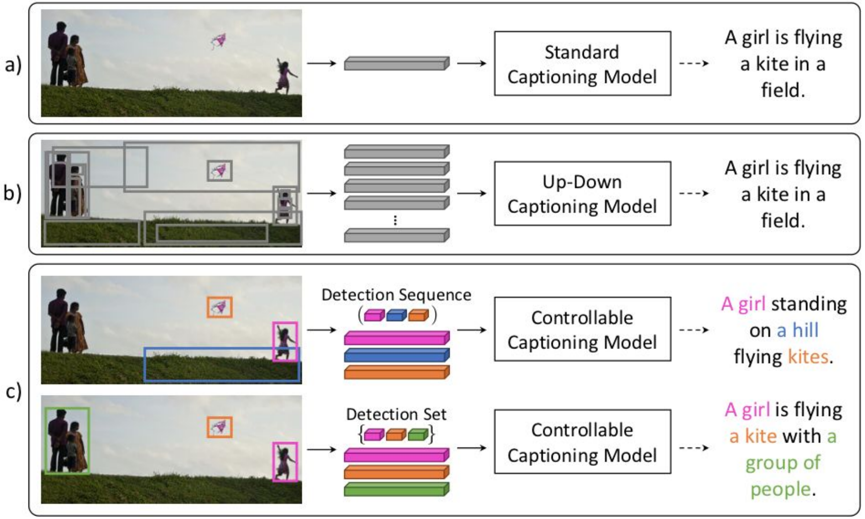
Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions

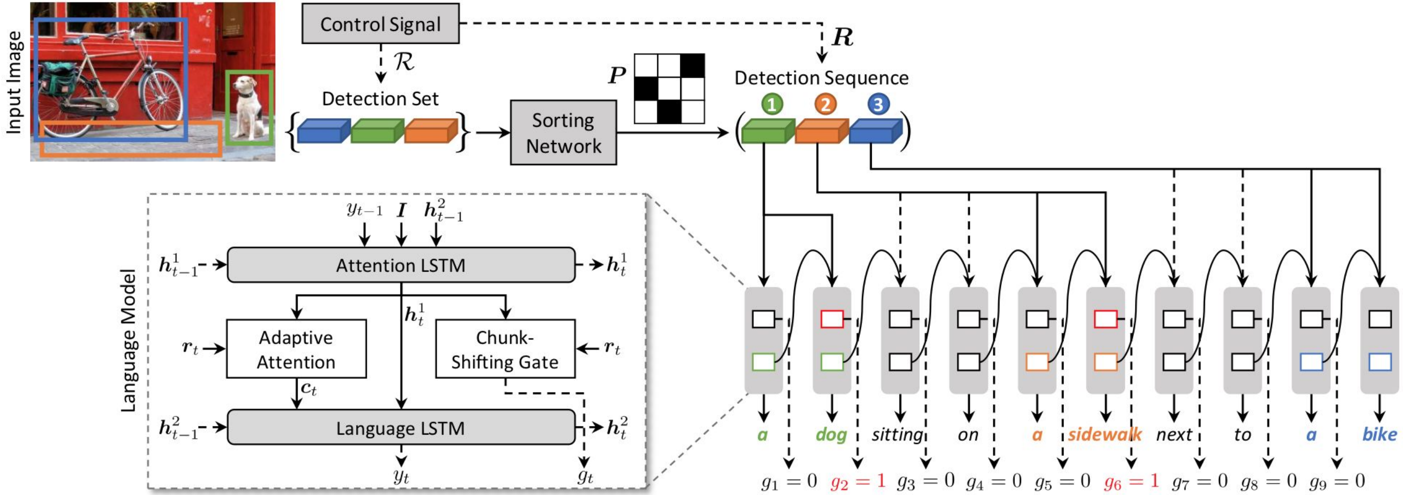
- Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara
University of Modena and Reggio Emilia
- Accepted to CVPR 2019
- https://arxiv.org/abs/1811.10652

Overview

- Given a control signal, in the form of a sequence or set of region sets, generate the corresponding caption together with its noun chunks grounded on region sets


a)
Standard Captioning Model
A girl is flying a kite in a field.
b)
Up-Down Captioning Model
A girl is flying a kite in a field.
c)
Detection Sequence
Controllable Captioning Model
A girl standing on a hill flying kites.
Detection Set
Controllable Captioning Model
A girl is flying a kite with a group of people.

Architecture



- Input: a set / sequence of **region sets**. Output: sequence of noun chunks
- Two-stage pipeline. Objects sorting followed by joint prediction of token sequence and chunk shifting.
- Fine-tuned with REINFORCE. Reward defined as a combination of Meteor and Alignment between token sequence and detection sequence.

Objectives

MLE

$$\begin{aligned} L(\theta) = -\sum_{t=1}^{T} \Bigg(&\log \overbrace{p(y_t^* | \boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*)}^{\text{Word-level probability}} + \\ &+ g_t^* \log p(g_t = 1 | \boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*) + \\ &+ (1 - g_t^*)(1 - \log \underbrace{p(g_t = 1 | \boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*)}_{\text{Chunk-level probability}} \Bigg) \end{aligned}$$

REINFORCE

$$\nabla_\theta L(\theta) = -(r(\boldsymbol{w}^s) - b)(\nabla_\theta \log p(\boldsymbol{w}^s) + \nabla_\theta \log p(\boldsymbol{g}^s))$$

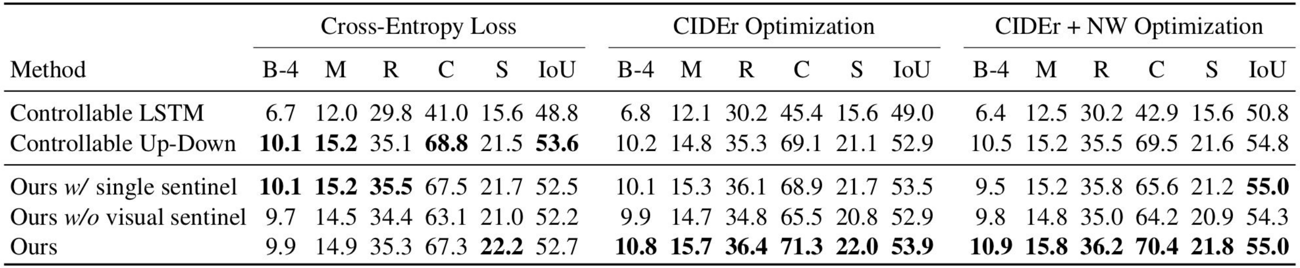
reward of the sentence obtained using regular inference procedure

Reward

- Rewarding caption quality - CIDEr
- Rewarding the alignment w.r.t. control input - Needleman-Wunsch score

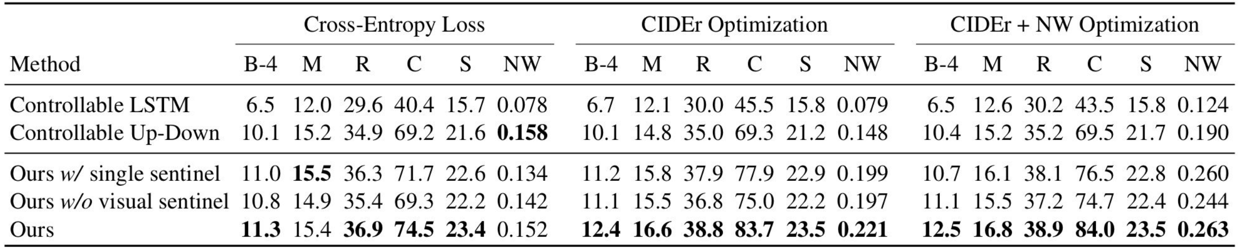
$$\mathrm{NW}(\boldsymbol{y}, \boldsymbol{y}^*) = \frac{al(\boldsymbol{y}, \boldsymbol{y}^*)}{\max(\#\boldsymbol{y}, \#\boldsymbol{y}^*)}$$

Performance on Flickr30k



	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU
Controllable LSTM	6.7	12.0	29.8	41.0	15.6	48.8	6.8	12.1	30.2	45.4	15.6	49.0	6.4	12.5	30.2	42.9	15.6	50.8
Controllable Up-Down	**10.1**	**15.2**	35.1	**68.8**	21.5	**53.6**	10.2	14.8	35.3	69.1	21.1	52.9	10.5	15.2	35.5	69.5	21.6	54.8
Ours *w/* single sentinel	**10.1**	**15.2**	**35.5**	67.5	21.7	52.5	10.1	15.3	36.1	68.9	21.7	53.5	9.5	15.2	35.8	65.6	21.2	**55.0**
Ours *w/o* visual sentinel	9.7	14.5	34.4	63.1	21.0	52.2	9.9	14.7	34.8	65.5	20.8	52.9	9.8	14.8	35.0	64.2	20.9	54.3
Ours	9.9	14.9	35.3	67.3	**22.2**	52.7	**10.8**	**15.7**	**36.4**	**71.3**	**22.0**	**53.9**	**10.9**	**15.8**	**36.2**	**70.4**	**21.8**	**55.0**

Table 9: Controllability via a set of regions, on the test portion of Flickr30K Entities.

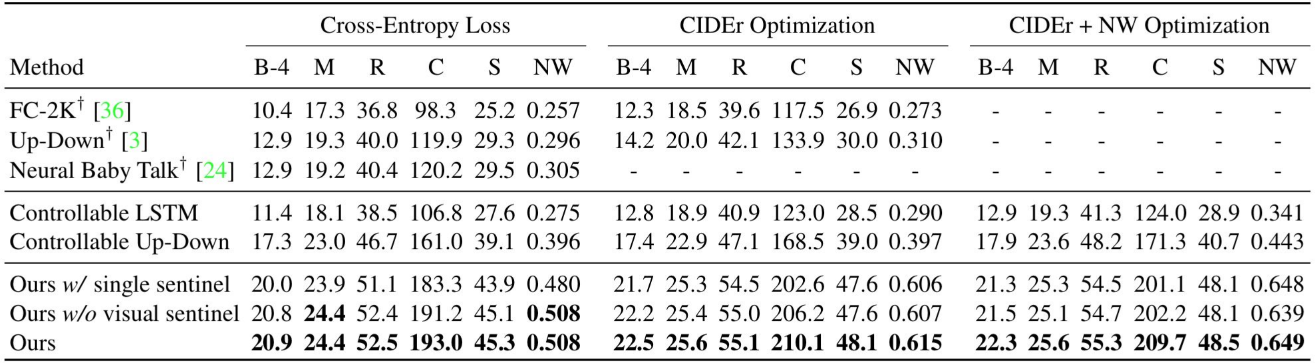


	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW
Controllable LSTM	6.5	12.0	29.6	40.4	15.7	0.078	6.7	12.1	30.0	45.5	15.8	0.079	6.5	12.6	30.2	43.5	15.8	0.124
Controllable Up-Down	10.1	15.2	34.9	69.2	21.6	**0.158**	10.1	14.8	35.0	69.3	21.2	0.148	10.4	15.2	35.2	69.5	21.7	0.190
Ours *w/* single sentinel	11.0	**15.5**	36.3	71.7	22.6	0.134	11.2	15.8	37.9	77.9	22.9	0.199	10.7	16.1	38.1	76.5	22.8	0.260
Ours *w/o* visual sentinel	10.8	14.9	35.4	69.3	22.2	0.142	11.1	15.5	36.8	75.0	22.2	0.197	11.1	15.5	37.2	74.7	22.4	0.244
Ours	**11.3**	15.4	**36.9**	**74.5**	**23.4**	0.152	**12.4**	**16.6**	**38.8**	**83.7**	**23.5**	**0.221**	**12.5**	**16.8**	**38.9**	**84.0**	**23.5**	**0.263**

Table 7: Controllability via a sequence of regions, on the test portion of Flickr30K Entities.

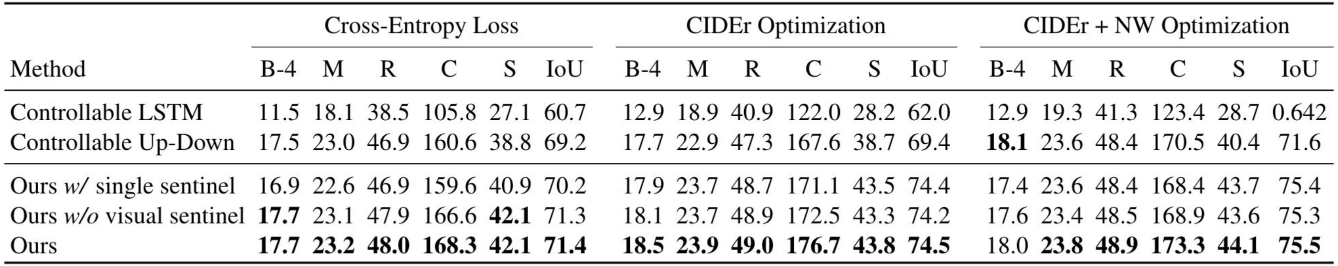
- Metrics are collected using references that are describing the same set of objects as the control input (only 1 reference in most cases)

Performance on COCO



	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW
FC-2K† [36]	10.4	17.3	36.8	98.3	25.2	0.257	12.3	18.5	39.6	117.5	26.9	0.273	-	-	-	-	-	-
Up-Down† [3]	12.9	19.3	40.0	119.9	29.3	0.296	14.2	20.0	42.1	133.9	30.0	0.310	-	-	-	-	-	-
Neural Baby Talk† [24]	12.9	19.2	40.4	120.2	29.5	0.305	-	-	-	-	-	-	-	-	-	-	-	-
Controllable LSTM	11.4	18.1	38.5	106.8	27.6	0.275	12.8	18.9	40.9	123.0	28.5	0.290	12.9	19.3	41.3	124.0	28.9	0.341
Controllable Up-Down	17.3	23.0	46.7	161.0	39.1	0.396	17.4	22.9	47.1	168.5	39.0	0.397	17.9	23.6	48.2	171.3	40.7	0.443
Ours *w/* single sentinel	20.0	23.9	51.1	183.3	43.9	0.480	21.7	25.3	54.5	202.6	47.6	0.606	21.3	25.3	54.5	201.1	48.1	0.648
Ours *w/o* visual sentinel	20.8	**24.4**	52.4	191.2	45.1	**0.508**	22.2	25.4	55.0	206.2	47.6	0.607	21.5	25.1	54.7	202.2	48.1	0.639
Ours	**20.9**	**24.4**	**52.5**	**193.0**	**45.3**	**0.508**	**22.5**	**25.6**	**55.1**	**210.1**	**48.1**	**0.615**	**22.3**	**25.6**	**55.3**	**209.7**	**48.5**	**0.649**

Table 2: Controllability via a sequence of regions, on test portion of COCO Entities. NW refers to the visual chunk alignment measure defined in Sec. 3.2. The † marker indicates non-controllable methods.



	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU
Controllable LSTM	11.5	18.1	38.5	105.8	27.1	60.7	12.9	18.9	40.9	122.0	28.2	62.0	12.9	19.3	41.3	123.4	28.7	0.642
Controllable Up-Down	17.5	23.0	46.9	160.6	38.8	69.2	17.7	22.9	47.3	167.6	38.7	69.4	**18.1**	23.6	48.4	170.5	40.4	71.6
Ours *w/* single sentinel	16.9	22.6	46.9	159.6	40.9	70.2	17.9	23.7	48.7	171.1	43.5	74.4	17.4	23.6	48.4	168.4	43.7	75.4
Ours *w/o* visual sentinel	**17.7**	23.1	47.9	166.6	**42.1**	71.3	18.1	23.7	48.9	172.5	43.3	74.2	17.6	23.4	48.5	168.9	43.6	75.3
Ours	**17.7**	**23.2**	**48.0**	**168.3**	**42.1**	**71.4**	**18.5**	**23.9**	**49.0**	**176.7**	**43.8**	**74.5**	18.0	**23.8**	**48.9**	**173.3**	**44.1**	**75.5**

Table 8: Controllability via a set of regions, on the test portion of COCO Entities.

Sorting Network

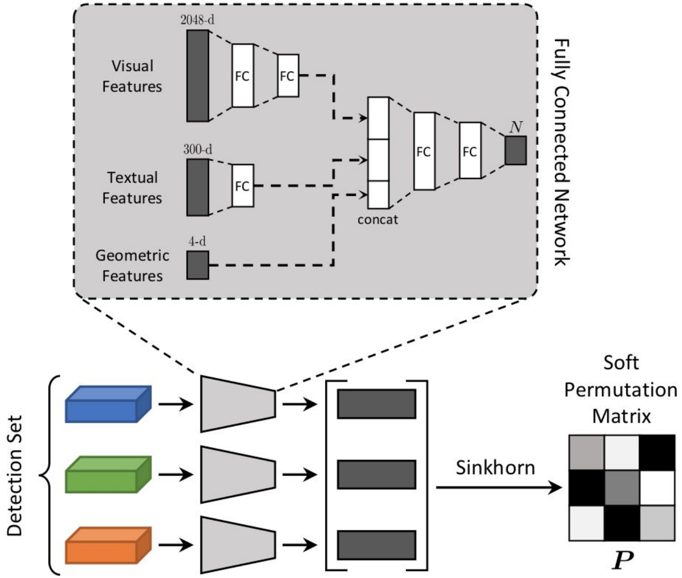


Figure 7: Schema of the sorting network.

- Learns a soft permutation matrix of the input sequence
- Soft permutation matrix is converted to permutation matrix using Hungarian algorithm on inference
- Given a set of region sets R = {r1, r2, ..., rN }, each region set produces a vector of length N