Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel University of Toronto

Computer Science
UNIVERSITY OF TORONTO

The problems to solve

- Retrieve and generate descriptions of images

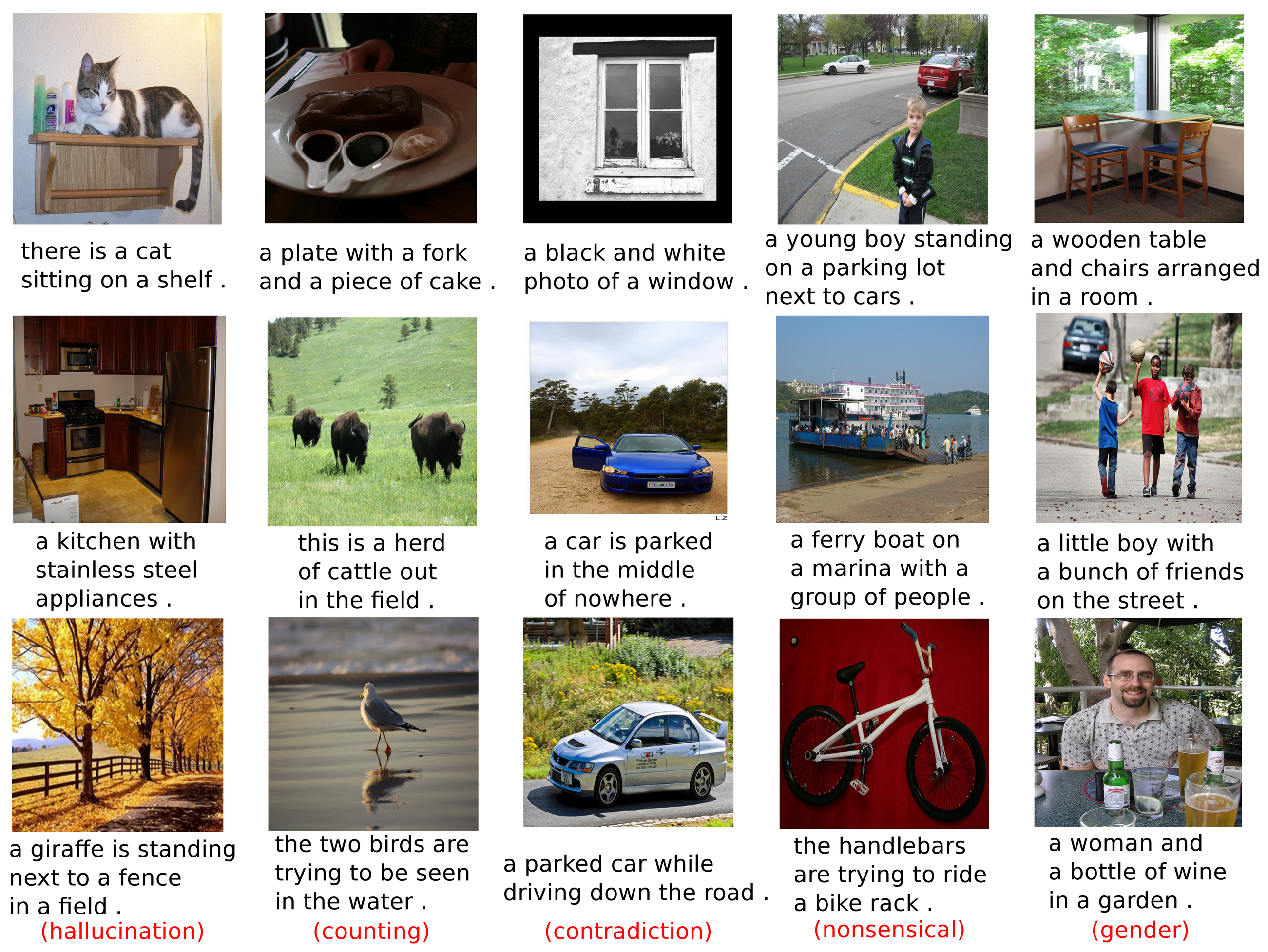


Figure: Sample generated captions from our proposed model. The bottom row shows some representative error cases, along with our description (in red) of the type of error. None of these generated descriptions appear word for word in the training set.

Our encoder-decoder model

- Encoder-decoder model

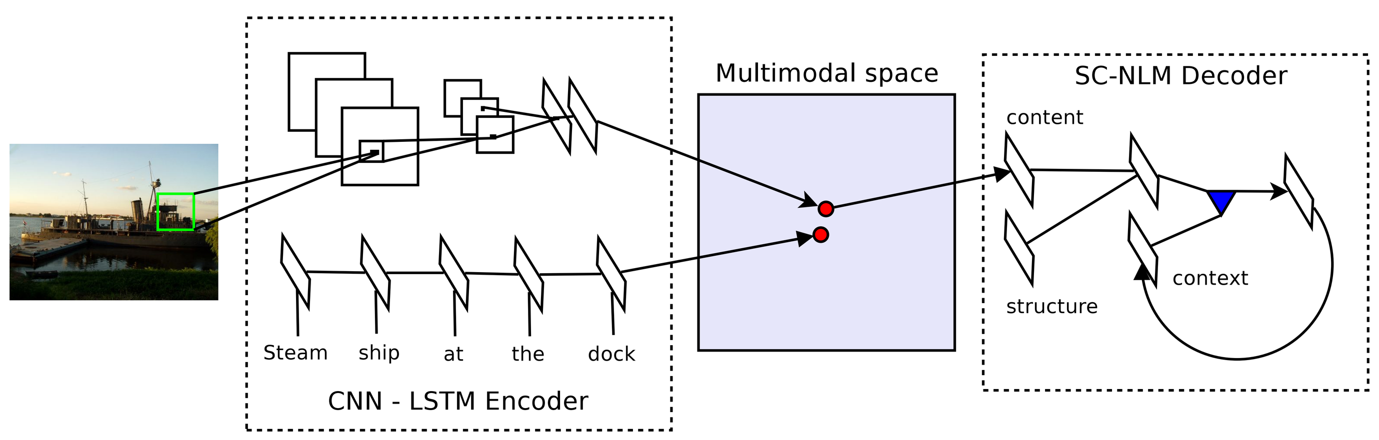


Figure: **Encoder:** A deep convolutional network (CNN) and long short-term memory recurrent network (LSTM) for learning a joint image-sentence embedding. **Decoder:** A new neural language model that combines structure and content vectors for generating words one at a time in sequence.

Encoder (ConvNet-LSTM)

- Given: (image, sentence) training pairs

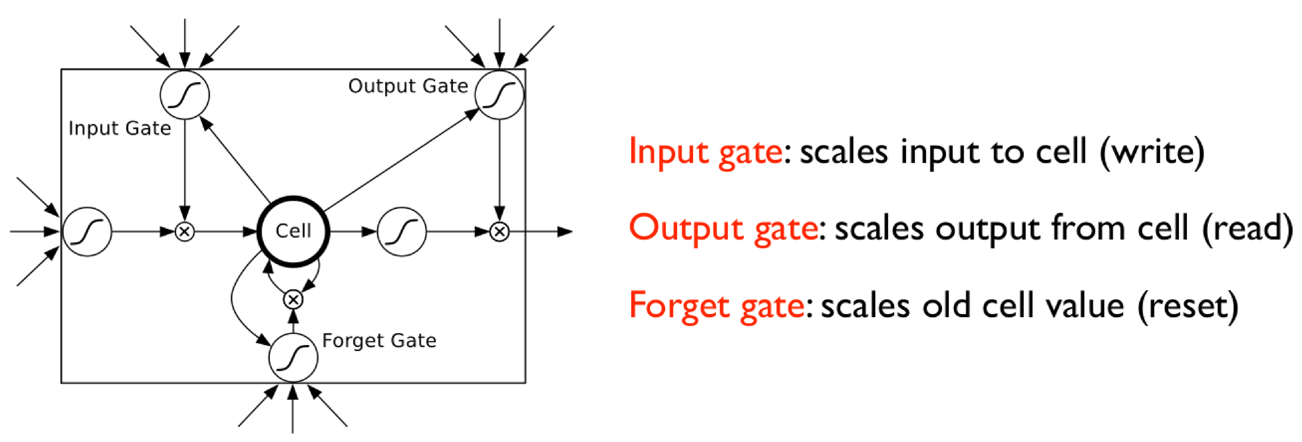


Figure: LSTM. Inputs are word vectors, hidden states gives sentence vectors.

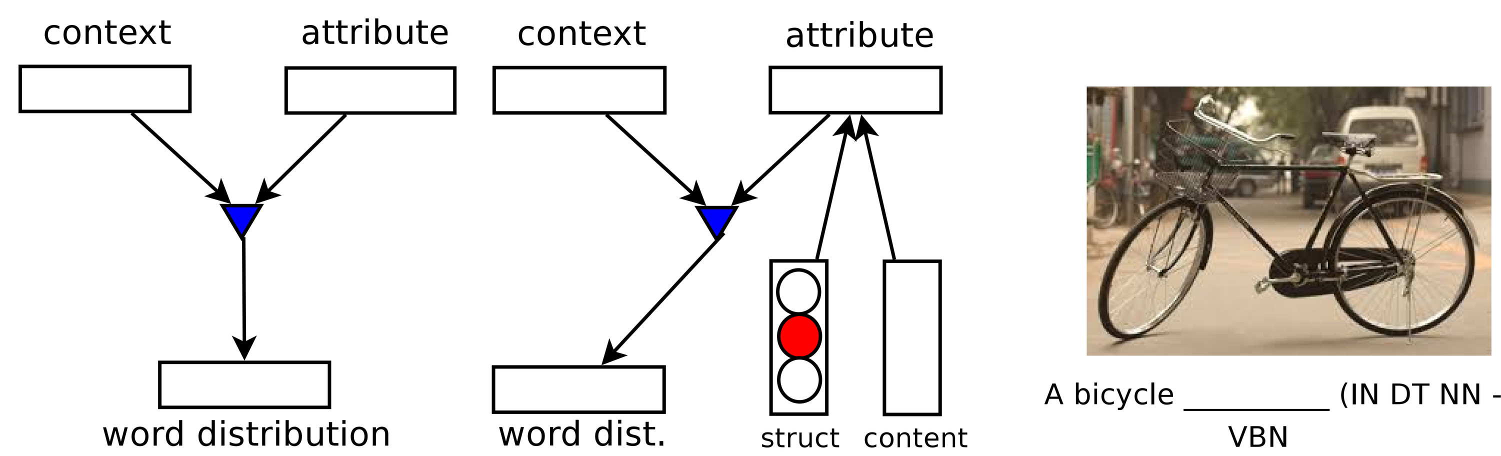
- $\mathbf{x}$: linear transformation of ConvNet 4096 dim output, unit norm
- $\mathbf{v}$: Sentence embedding (hidden state of the LSTM), unit norm
- Minimize the following pairwise ranking loss:

$$\min_{\theta} \sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} + \sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\},$$

- Where $\mathbf{v}_k, \mathbf{x}_k$ are contrastive terms

Decoder (SC-NLM)

- Let $\mathbf{v}$ denote the image embedding from the multimodal space
- Given a sentence $w_1, \ldots, w_N$, with corresponding POS tags $t_1, \ldots, t_N$
- Model the distribution $P(w_n = i | w_{1:n-1}, t_{n:n+k}, \mathbf{v})$ for previous word context $w_{1:n-1}$ and forward POS context $t_{n:n+k}$



(a) Multiplicative NLM (b) Structure-content NLM (c) SC-NLM prediction

Figure: **Left:** Multiplicative neural language model. **Middle:** Structure-content neural language model (SC-NLM). **Right:** The prediction problem of an SC-NLM: estimate $P(w_n = i | w_{1:n-1}, t_{n:n+k}, \mathbf{v})$, where "A bicycle" is the context $w_{1:n-1}$, "VBN IN DT NN" is the forward structure context $t_{n:n+k}$, and the content $\mathbf{v}$ is the image embedding.

How to generate descriptions

- Given an image, map it into the multimodal space to get $\mathbf{v}$
- Sample a POS sequence from the training set
- Generate a description $\mathbf{x}$ from the SC-NLM
- Score the description with a **translation model** ($s(\mathbf{x}, \mathbf{v})$) and an n-gram **language model**.

Experiment: Localization

- How well can we localize objects after training?

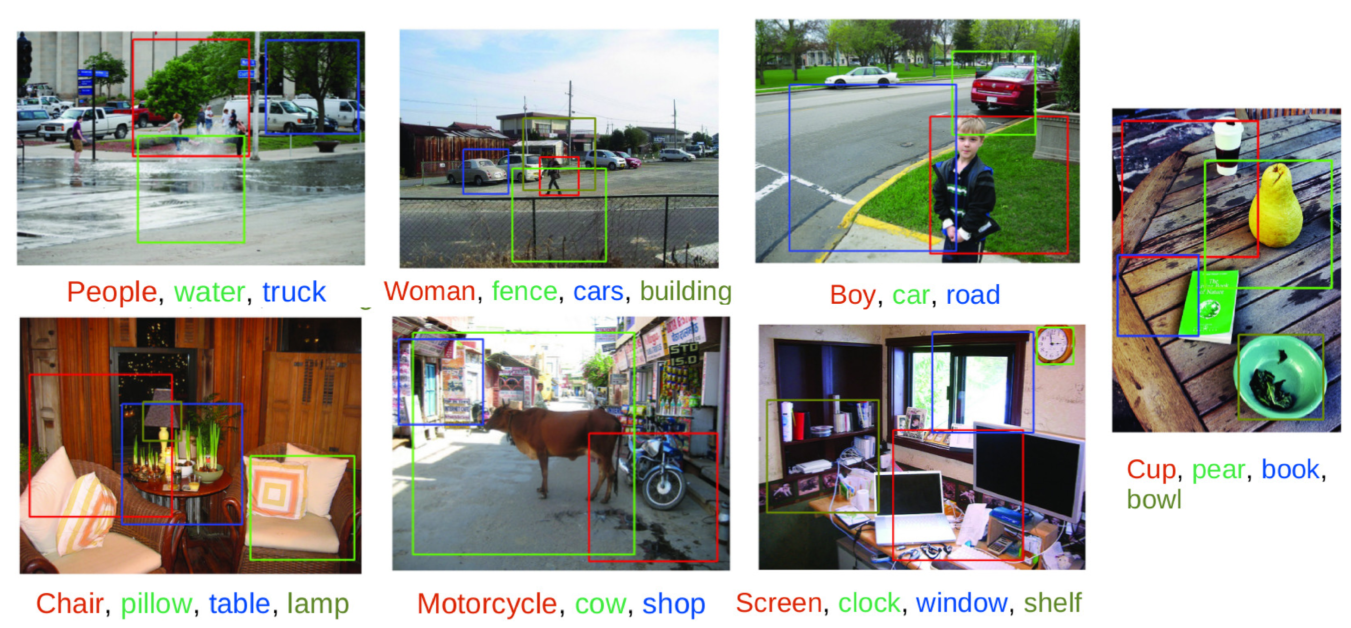
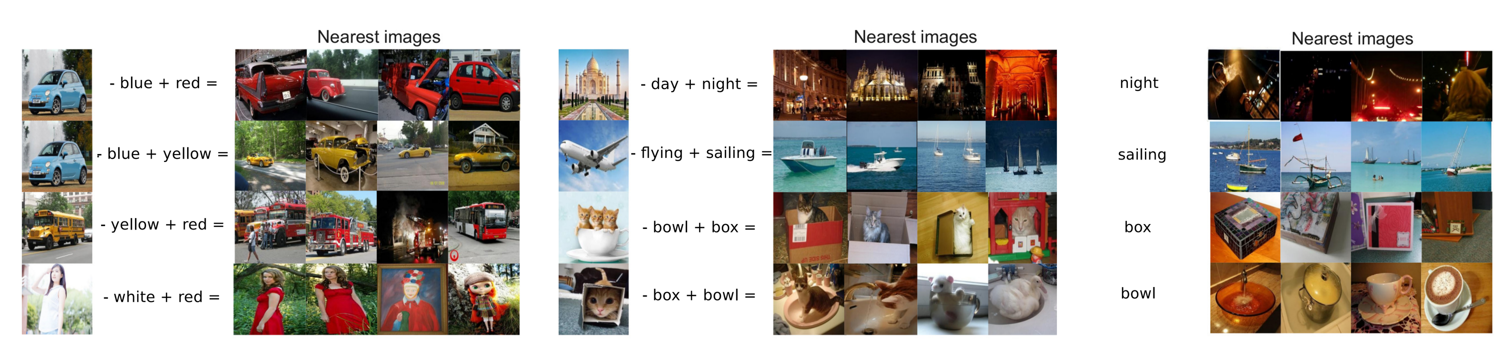


Figure: Self-taught object localization. Even though our model does not incorporate object detections during training, it can still learn to localize.

Experiment: Multimodal linguistic regularities

- Multimodal vector space arithmetic with a linear encoder



(a) Colors (b) Image structure (c) Sanity check

Figure: Multimodal vector space arithmetic, for (a) Colors and (b) Image structure. The sanity check shows that the retrieved images are not simply the ones associated with the added word.

Image-sentence ranking

- Image-sentence ranking experiments: Flickr8K and Flickr30K
- Image annotation: for each image, rank all descriptions
- Image search: for each sentence, rank all images
- Retrieval is done within the development/test sets (1000 images)

Flickr8K

	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med r**	**R@1**	**R@5**	**R@10**	**Med r**
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
SDT-RNN [1]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
† DeViSE [2]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
† SDT-RNN [1]	6.0	22.7	34.0	23	6.6	21.6	31.7	25
DeFrag [3]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
† DeFrag [3]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
m-RNN [4]	*14.5*	*37.2*	*48.5*	*11*	11.5	*31.0*	42.4	15
† BRNN [5]	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4
‡ NIC (GoogLeNet) [6]	**(20)**		**(61)**	**6**	**(19)**		**(64)**	**5**
Our model	13.5	36.2	45.7	13	10.4	*31.0*	*43.7*	*14*
Our model (OxfordNet)	18.0	40.9	55.0	8	12.5	37.0	51.5	10

Flickr30K

	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med r**	**R@1**	**R@5**	**R@10**	**Med r**
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
† DeViSE [2]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
† SDT-RNN [1]	9.6	29.8	41.1	16	8.9	29.8	41.1	16
† DeFrag [3]	14.2	37.7	51.3	10	10.2	30.8	44.2	14
† DeFrag + Finetune CNN	16.4	40.2	54.7	8	10.3	31.4	44.5	13
m-RNN [4]	*18.4*	*40.2*	*50.9*	*10*	12.6	31.2	41.5	16
LRCN [7]					*14.0*	*34.9*	*47.0*	*11*
† BRNN [5]	22.2	48.2	61.4	**4.8**	15.2	37.7	50.5	9.2
‡ NIC (GoogLeNet) [6]	(17)		(56)	7	**(17)**		**(57)**	**7**
Our model	14.8	39.2	*50.9*	*10*	11.8	34.0	46.3	13
Our model (OxfordNet)	**23.0**	**50.7**	**62.9**	**5**	**16.8**	**42.0**	**56.5**	8

Table: Flickr8K and Flickr30K experiments. **R@K** is Recall@K (high is good). **Med r** is the median rank (low is good). Best results overall are **bold**, best results without OxfordNet or GoogLeNet features are underlined and best results that only use single frame features (without OxfordNet or GoogLeNet) are *italicized*. A † in front of the method indicates that object detections were used along with single frame features. A ‡ indicates that ensembles were used.

Additional results

- **DEMO:** See deeplearning.cs.toronto.edu/i2t

References

[1] Richard Socher, Q Le, C Manning, and A Ng. Grounded compositional semantics for finding and describing images with sentences. *In TACL*, 2014.

[2] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, and Tomas Mikolov MarcAurelio Ranzato. Devise: A deep visual-semantic embedding model. *NIPS*, 2013.

[3] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *NIPS*, 2014.

[4] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[5] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *Stanford technical report*, 2014.

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[7] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.