CVPR 2017

Semantic Compositional Networks for Visual Captioning

Presenter: Zhe Gan

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu,

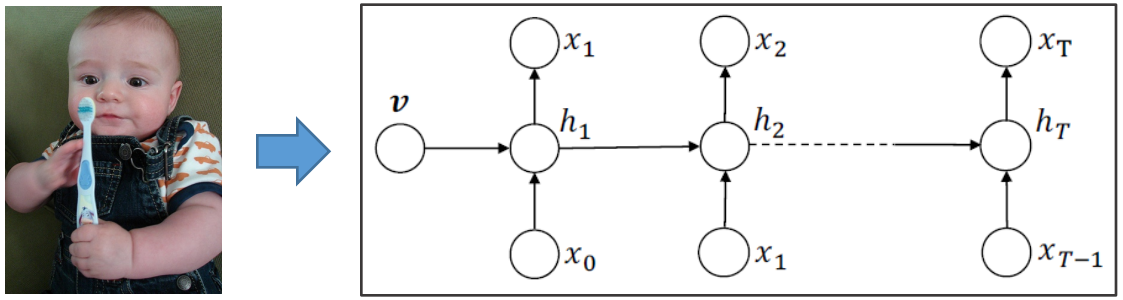
Kenneth Tran, Jianfeng Gao, Lawrence Carin, Li Deng

Microsoft Research & Duke University & Tsinghua University

Traditional Image Captioning

Baseline:

- Suboptimal quality
- Not interpretable; not easy to control the caption



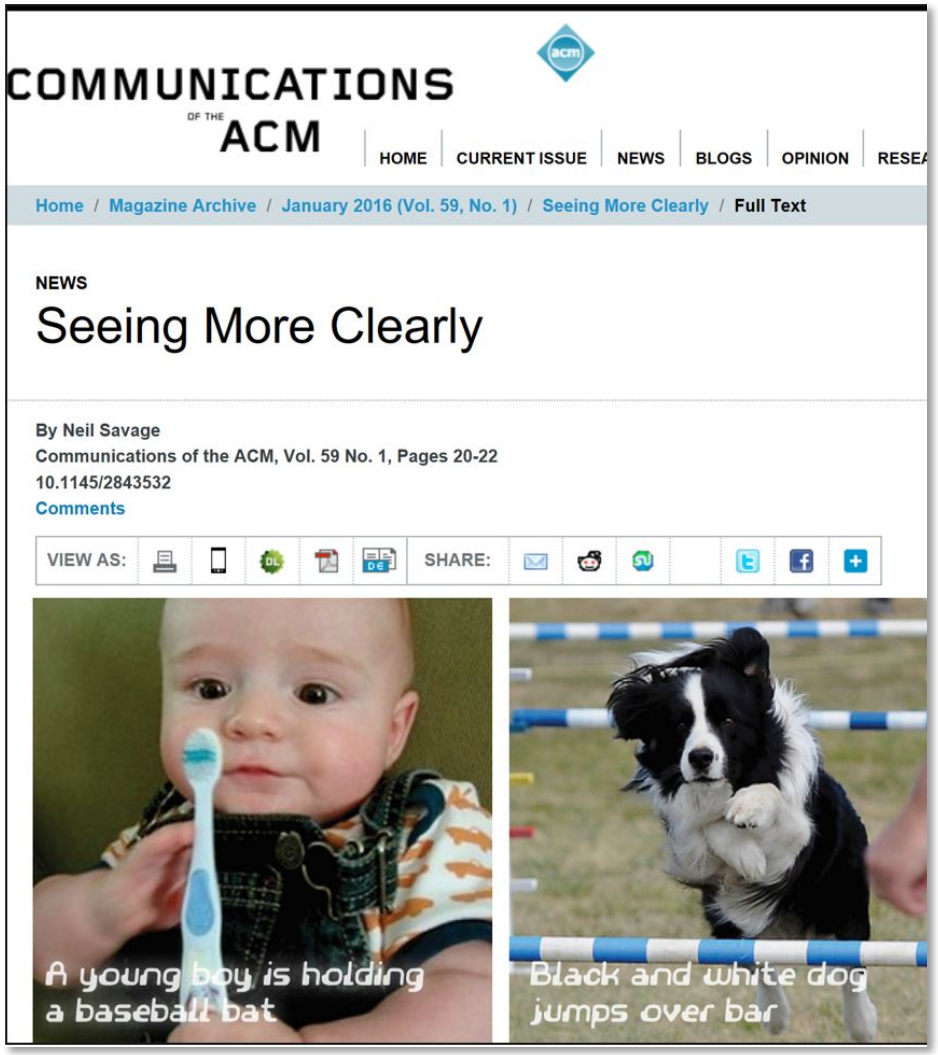


Image Captioning with Control

Conceptually, learn 1000 LSTMs, one for each semantic attribute.
Combine these 1000 LSTMs, weighted by the attributes' likelihood.
Run tensor decomposition to reduce # parameters to fit GPU.

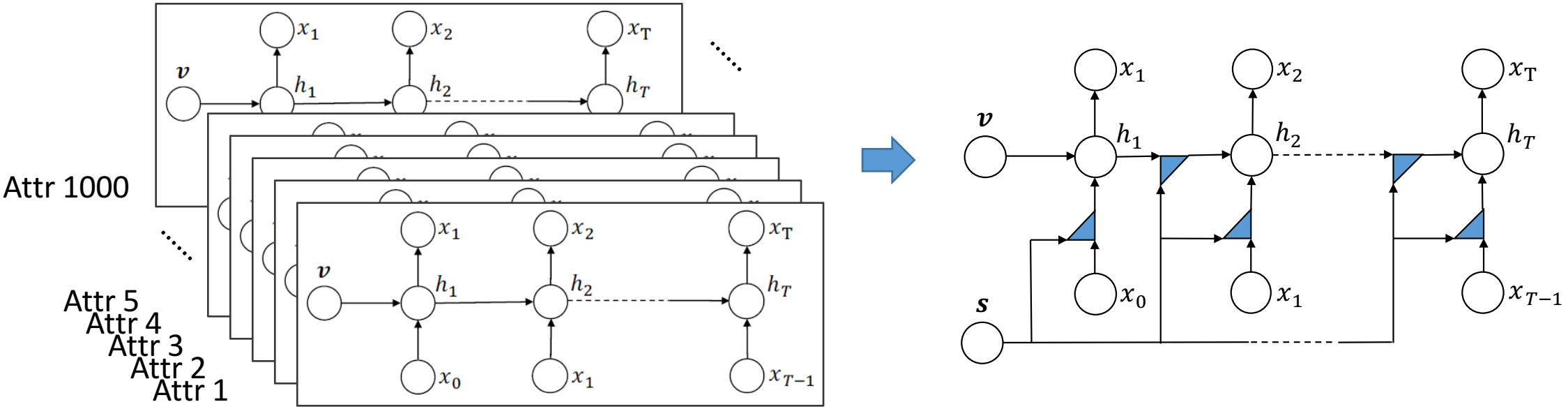


Image Captioning with Control

Detected semantic concepts:

person (0.998), baby (0.983), holding (0.952), small (0.697), sitting (0.638), toothbrush (0.538), child (0.502), mouth (0.438)

Overall caption generated by the SCN:

a baby holding a toothbrush in its mouth

Influence the caption by changing the tag:

1. Replace "**baby**" with "**girl**": *a little girl holding a toothbrush in her mouth*
2. Replace "**toothbrush**" with "**baseball**": *a baby holding a baseball bat in his hand*
3. Replace "**toothbrush**" with "**pizza**": *a baby holding a piece of pizza in his mouth*

Quantitative results

State-of-the-art results on both image and video captioning

COCO

	BLEU-4	METEOR	CIDEr-D
Best in CVPR'16	0.310	0.260	0.940
SCN (ours)	0.341	0.261	1.041

Youtube2Text

	BLEU-4	METEOR	CIDEr-D
Best in CVPR'16	0.499	0.326	0.658
SCN (ours)	0.511	0.335	0.777

Summary

- Our SCN can be considered as efficiently learning an ensemble of 1000 LSTMs, one for each semantic concept.
- Our SCN provides an interpretable way to control the generation of captions.

Come to our poster for details

Semantic Compositional Networks for Visual Captioning GitHub

https://github.com/zhegan27/Semantic_Compositional_Nets

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu,
Kenneth Tran, Jianfeng Gao, Lawrence Carin, Li Deng