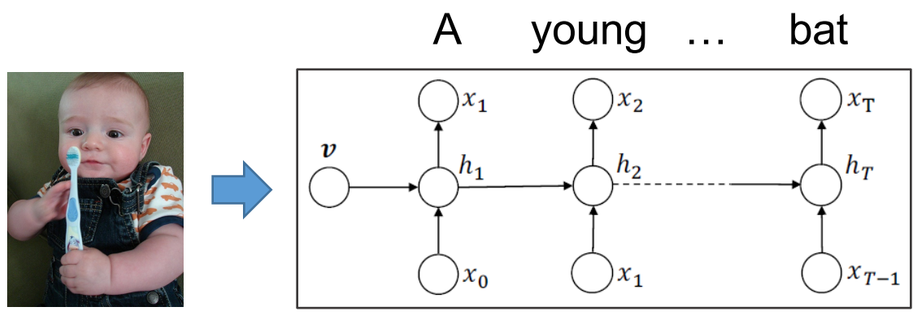
Controllable Visual Captioning

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

Contribution

- **Semantic Compositional Nets**: (i) control the content of generated captions; (ii) can be used for caption editing
- Efficiently training a Mixture of 1000 LSTM Experts, one for each tag.
- **StyleNet**: control the style of generated captions

Traditional Image Captioning



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

Semantic Compositional Nets

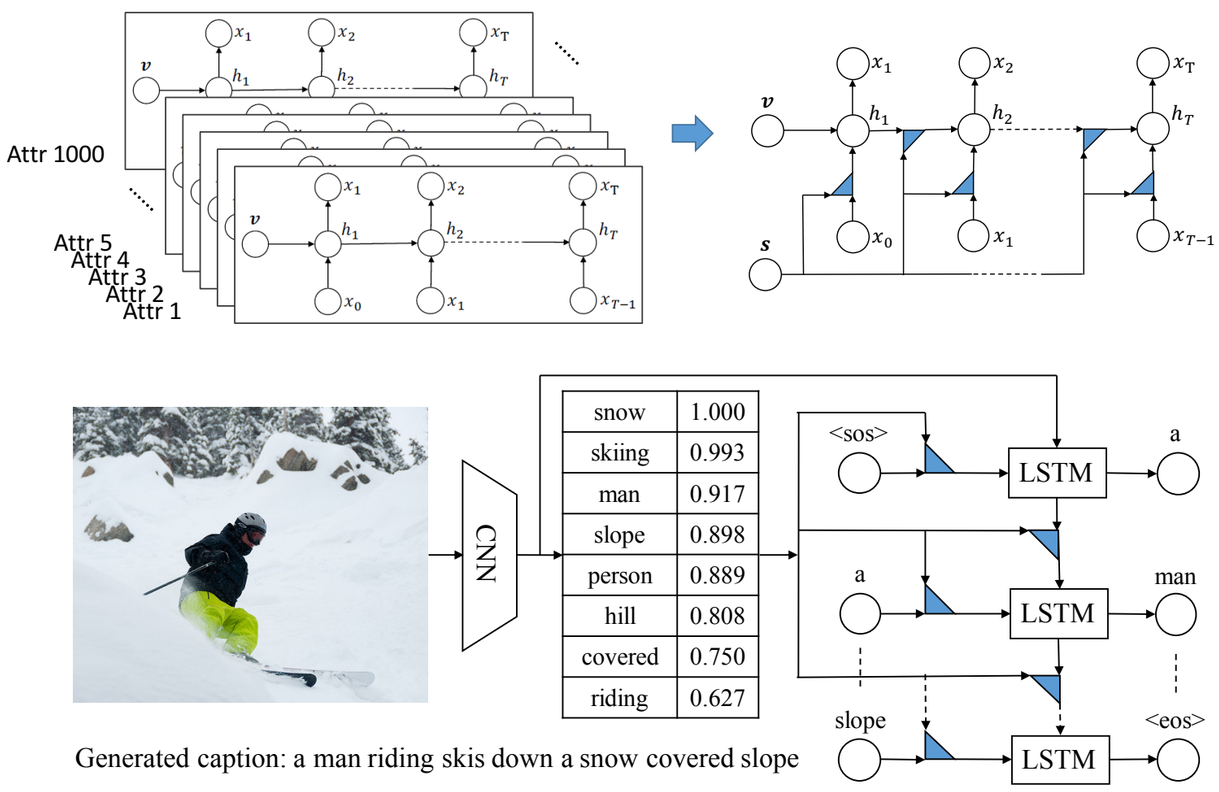
- Conceptually, we learn 1000 LSTM experts, one for each tag
- Combine these 1000 LSTMs, weighted by the tags' likelihood
- Run tensor decomposition to reduce # of parameters to fit GPU

$$\boldsymbol{h}_t = \sigma(\mathbf{W}(\boldsymbol{s})\boldsymbol{x}_{t-1} + \mathbf{U}(\boldsymbol{s})\boldsymbol{h}_{t-1} + \boldsymbol{z})$$

$$\mathbf{W}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k \mathbf{W}_{\mathcal{T}}[k],\ \mathbf{U}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k \mathbf{U}_{\mathcal{T}}[k]$$

$$\mathbf{W}(\boldsymbol{s}) = \mathbf{W}_a \cdot \mathrm{diag}(\mathbf{W}_b \boldsymbol{s}) \cdot \mathbf{W}_c\,,$$
$$\mathbf{U}(\boldsymbol{s}) = \mathbf{U}_a \cdot \mathrm{diag}(\mathbf{U}_b \boldsymbol{s}) \cdot \mathbf{U}_c\,,$$

$$\mathbf{W}(\boldsymbol{s}) = \textstyle\sum_{k=1}^{K} s_k [\mathbf{W}_a \cdot \mathrm{diag}(\boldsymbol{w}_{bk}) \cdot \mathbf{W}_c]\,.$$



Generated caption: a man riding skis down a snow covered slope

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)

Semantic composition:
1. Only using "**baby**": *a baby in a*
2. Only using "**holding**": *a person holding a hand*
3. Only using "**toothbrush**": *a pair of toothbrush*
4. Only using "**mouth**": *a man with a toothbrush*
5. Using "**baby**" and "**mouth**": *a baby brushing its teeth*

Overall caption generated by the SCN:
a baby holding a toothbrush in its mouth

Influence the caption by changing the tag:
6. Replace "**baby**" with "**girl**": *a little girl holding a toothbrush in her mouth*
7. Replace "**toothbrush**" with "**baseball**": *a baby holding a baseball bat in his hand*
8. Replace "**toothbrush**" with "**pizza**": *a baby holding a piece of pizza in his mouth*

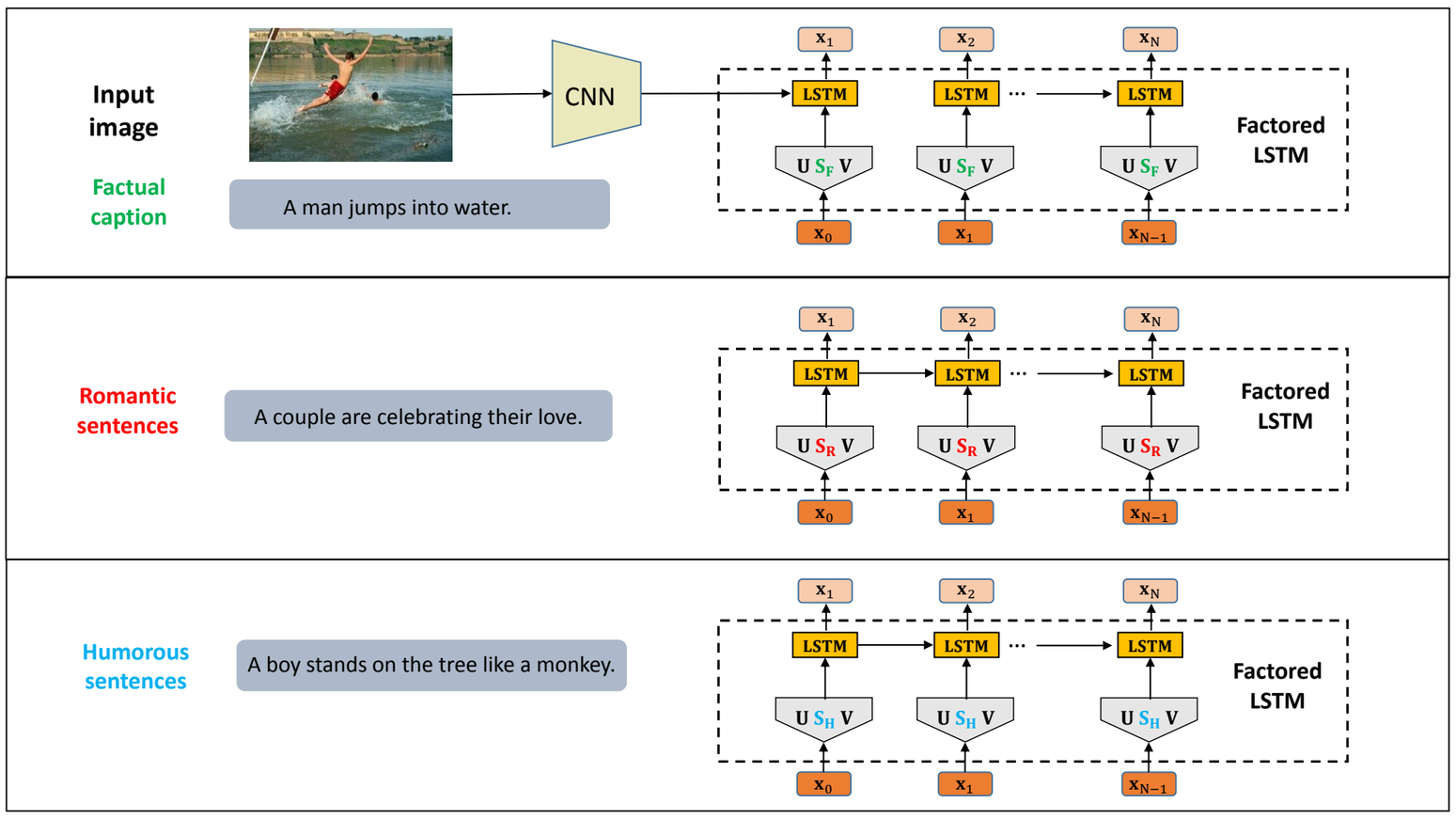
COCO

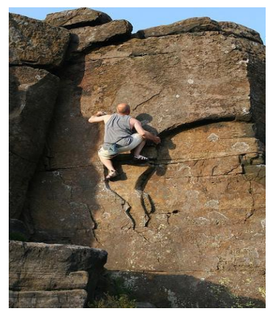
	BLEU-4	METEOR	CIDEr-D
Best in CVPR'16	0.310	0.260	0.940
SCN (ours)	0.341	0.261	1.041

Youtube2Text

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

StyleNet: Generating Attractive Visual Captions with Styles





CaptionBot: A man on a rocky hillside next to a stone wall.

Romantic: A man uses rock climbing to conquer the high.

Humorous: A man is climbing the rock like a lizard.

CaptionBot: A dog runs in the grass.

Romantic: A dog runs through the grass to meet his lover.

Humorous: A dog runs through the grass in search of the missing bones.

Standard: A man is playing guitar.
Romantic: A man practices the guitar, dream of being a rock star.
Humorous: A man is playing guitar but runs way.