Semantic Compositional Networks for Visual Captioning

Zhe Gan[†], Chuang Gan[*], Xiaodong He[‡], Yunchen Pu[†], Kenneth Tran[‡], Jianfeng Gao[‡], Lawrence Carin[†], Li Deng[‡]

[†]Duke University, [*]Tsinghua University, [‡]Microsoft Research Redmond

Introduction

Main Contributions:

- 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.

Key ideas:

- Semantic concepts (*i.e.*, tags) are first detected from the image.
- The SCN then extends each weight matrix of the long short-term memory (LSTM) network to an ensemble of tag-dependent weight matrices.
- The degree to which each member of the ensemble is used to generate a caption is tied to the image-dependent probability of the corresponding tag.

Semantic Compositional Networks

Semantic concept detection:

- First select a set of tags from the captions in the training set
- Then treat image tagging as a multi-label classification task
- Let $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{iK}] \in \{0,1\}^K$ be the label vector
  - $y_{ik} = 1$ if image i is annotated with tag k; $y_{ik} = 0$ otherwise.
- Let $\boldsymbol{v}_i$ represent the image feature vector, the cost function to be minimized is

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \Big(y_{ik} \log s_{ik} + (1 - y_{ik}) \log(1 - s_{ik}) \Big) \tag{1}$$

- $\boldsymbol{s}_i = \sigma\big(f(\boldsymbol{v}_i)\big)$ is the semantic feature vector.

Review of RNN for image captioning:

- The probability of caption $\mathbf{X}$ given image feature vector $\boldsymbol{v}$ is

$$p(\mathbf{X}|\mathbf{I}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}, \boldsymbol{v}) \tag{2}$$

- Each conditional $p(\boldsymbol{x}_t | \boldsymbol{x}_{<t}, \boldsymbol{v})$ is specified as softmax$(\mathbf{V}\boldsymbol{h}_t)$.
- Consider an RNN with a simple transition function

$$\boldsymbol{h}_t = \sigma(\mathbf{W}\boldsymbol{x}_{t-1} + \mathbf{U}\boldsymbol{h}_{t-1} + I(t=1) \cdot \mathbf{C}\boldsymbol{v}) \tag{3}$$

SCN: extending each weight matrix of the conventional RNN to be an ensemble of a set of tag-dependent weight matrices

$$\boldsymbol{h}_t = \sigma(\mathbf{W}(\boldsymbol{s})\boldsymbol{x}_{t-1} + \mathbf{U}(\boldsymbol{s})\boldsymbol{h}_{t-1} + I(t=1) \cdot \mathbf{C}\boldsymbol{v}) \tag{4}$$

- Given $\boldsymbol{s} \in \mathbb{R}^K$, we define tensors $\mathbf{W}_{\mathcal{T}} \in \mathbb{R}^{n_h \times n_x \times K}$ and $\mathbf{U}_{\mathcal{T}} \in \mathbb{R}^{n_h \times n_h \times K}$.
- $\mathbf{W}(\boldsymbol{s}) \in \mathbb{R}^{n_h \times n_x}$ and $\mathbf{U}(\boldsymbol{s}) \in \mathbb{R}^{n_h \times n_h}$ can be specified as

$$\mathbf{W}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k \mathbf{W}_{\mathcal{T}}[k], \; \mathbf{U}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k \mathbf{U}_{\mathcal{T}}[k] \tag{5}$$

- Can be interpreted as *jointly* training an ensemble of K RNNs in total.
- Though appealing, the number of parameters is proportional to K, which is prohibitive for large K (*e.g.*, $K = 1000$ for COCO).
- In order to remedy this problem, we factorize $\mathbf{W}(\boldsymbol{s})$ and $\mathbf{U}(\boldsymbol{s})$ as

$$\mathbf{W}(\boldsymbol{s}) = \mathbf{W}_a \cdot \mathrm{diag}(\mathbf{W}_b \boldsymbol{s}) \cdot \mathbf{W}_c \,, \tag{6}$$

$$\mathbf{U}(\boldsymbol{s}) = \mathbf{U}_a \cdot \mathrm{diag}(\mathbf{U}_b \boldsymbol{s}) \cdot \mathbf{U}_c \tag{7}$$

Model architecture

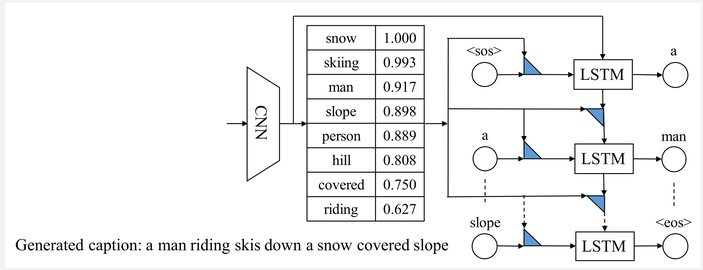

Figure: Overview of the proposed model.

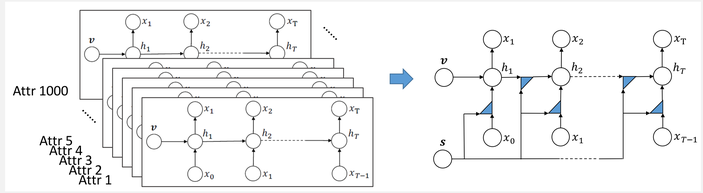

Figure: SCN learns an ensemble of 1000 LSTMs, one for each semantic concept.

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*

Figure: Examples of SCN-based image captioning.

SCN: we obtain SCN with an RNN as

$$\tilde{\boldsymbol{x}}_{t-1} = \mathbf{W}_b \boldsymbol{s} \odot \mathbf{W}_c \boldsymbol{x}_{t-1} \,, \qquad \tilde{\boldsymbol{h}}_{t-1} = \mathbf{U}_b \boldsymbol{s} \odot \mathbf{U}_c \boldsymbol{h}_{t-1} \,, \tag{8}$$

$$\boldsymbol{z} = I(t=1) \cdot \mathbf{C}\boldsymbol{v} \,, \qquad \boldsymbol{h}_t = \sigma(\mathbf{W}_a \tilde{\boldsymbol{x}}_{t-1} + \mathbf{U}_a \tilde{\boldsymbol{h}}_{t-1} + \boldsymbol{z}) \,. \tag{9}$$

Let $\boldsymbol{w}_{bk}$ represent the kth column of $\mathbf{W}_b$, then

$$\mathbf{W}(\boldsymbol{s}) = \sum_{k=1}^{K} s_k [\mathbf{W}_a \cdot \mathrm{diag}(\boldsymbol{w}_{bk}) \cdot \mathbf{W}_c] \tag{10}$$

- The RNN weight matrices that correspond to each tag share "structure".
- We introduce LSTM units and generalize SCN-RNN to SCN-LSTM.

Experiments

Code: `https://github.com/zhegan27/Semantic_Compositional_Nets`

COCO results on Karpathy's split (small 5k test)

Methods	COCO					
	B-1	B-2	B-3	B-4	M	C
Best in CVPR'16	0.74	0.56	0.42	0.31	0.26	0.94
LSTM-R	0.698	0.525	0.390	0.292	0.238	0.889
LSTM-T	0.716	0.546	0.411	0.312	0.250	0.952
LSTM-RT	0.724	0.555	0.419	0.316	0.252	0.970
LSTM-RT$_2$	0.730	0.568	0.430	0.322	0.249	0.977
SCN-LSTM	0.728	0.566	0.433	0.330	0.257	1.012
SCN-LSTM Ens. of 5	**0.741**	**0.578**	**0.444**	**0.341**	**0.261**	**1.041**

COCO results on test server

Model	B-1	B-2	B-3	B-4	M	R	C
SCN-LSTM	**0.740**	**0.575**	**0.436**	**0.331**	**0.257**	**0.543**	**1.003**
ATT	0.731	0.565	0.424	0.316	0.250	0.535	0.943
OV	0.713	0.542	0.407	0.309	0.254	0.530	0.943
MSR Cap	0.715	0.543	0.407	0.308	0.248	0.526	0.931

Youtube2Text for video captioning

Model	B-4	M	C
Best in CVPR'16	0.499	0.326	0.658
LSTM-CR	0.469	0.317	0.688
LSTM-T	0.473	0.324	0.699
LSTM-CRT	0.475	0.316	0.647
LSTM-CRT$_2$	0.469	0.326	0.706
SCN-LSTM	0.502	0.334	0.770
SCN-LSTM Ens. of 5	**0.511**	**0.335**	**0.777**

Importance of using detected tags

Detected Tags:
book (1), shelf (1), table (0.965), sitting (0.955), person (0.955), library (0.908), room (0.829), front (0.464)

Generated captions:
LSTM-R: a young girl is playing a video game
LSTM-RT$_2$: a group of people sitting at a table
SCN-LSTM: two women sitting at a table in a library

Importance of using visual features

Detected Tags:
indoor (0.952), dog (0.828), sitting (0.647), stuffed (0.602), white (0.544), next (0.527), laying (0.509), cat (0.402)

Generated captions:
SCN-LSTM-T: a dog laying on top of a stuffed animal
SCN-LSTM: a teddy bear laying on top of a stuffed animal