An Empirical Study of Language CNN for Image Captioning

Jiuxiang Gu, Gang Wang, Jianfei Cai, Tsuhan Chen

Yuhuan Xiu

51164500032

Outline

1 Introduction

2 Motivation

3 Model Architecture

4 Experiments

5 Conclusion

Image Caption

"*translating*" an **image** to proper **sentences**.

Input an image

Output a sentence description

"A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop"

Methods

Classical approach: **Retrieval and ranking**

Weakness: can not generate proper captions for **a new combination of objection**

Neural networks: Encoder-decoder framework based on RNN / LSTM. Given the ground truth words **S** = { S[0],S[1],…S[t] } and the corresponding image **I**, the loss can be written as:

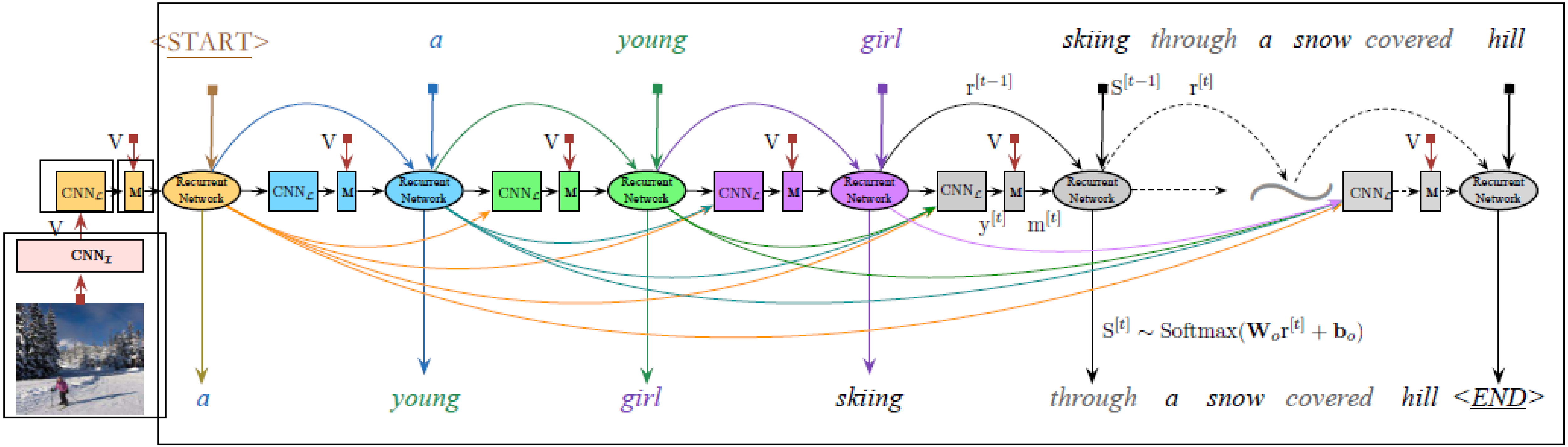
$$L(S,I) = -\sum_{t=0}^{N-1} \log P\left(S^{[t]} \middle| S^{[0]}, S^{[1]}, \dots, S^{[t-1]}, I\right)$$

Motivation

1. The Vanishing gradient problem.
2. Hierarchical structure of word sequence.

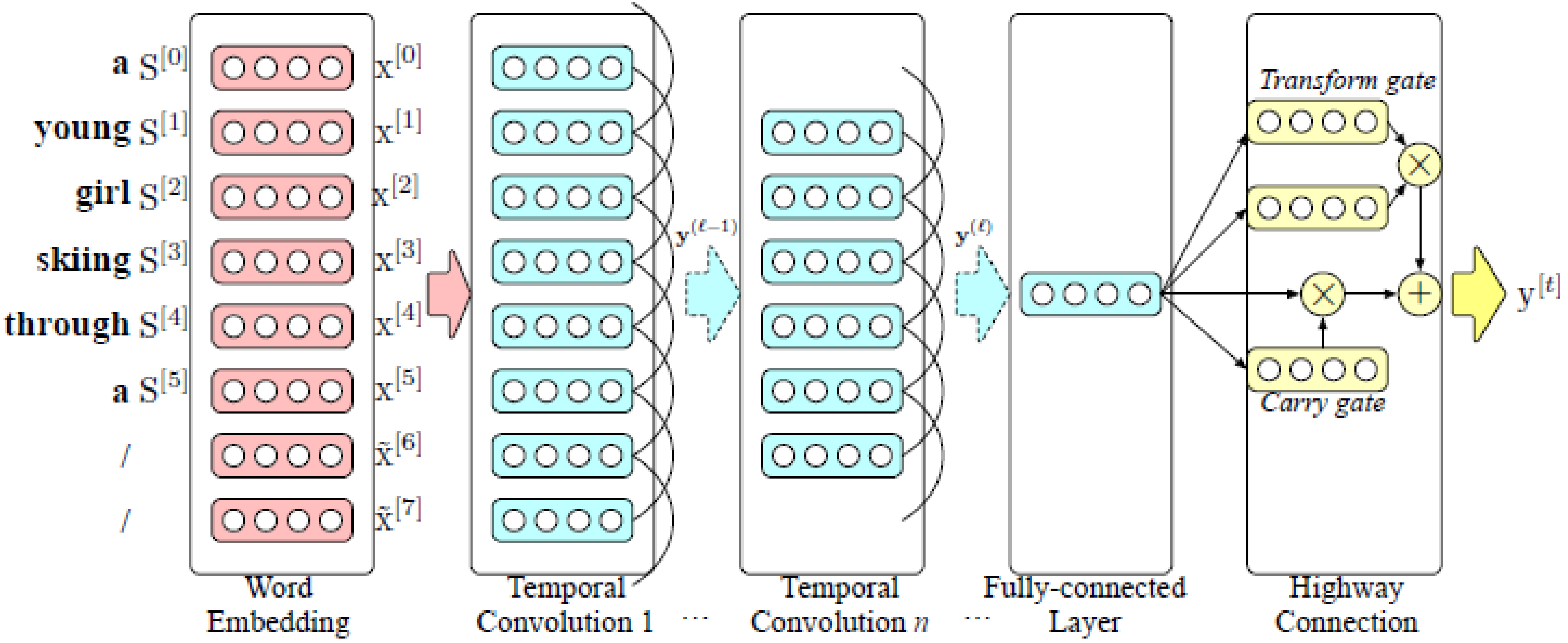
Language CNN is feed with all previous predicted words and can model the hierarchical structure and long-range dependences in word sequence.
.+ **recurrent networks (RNN, LSTM, RHN)** to model the dynamic temporal behavior.

Model Architecture



$$
\begin{aligned}
\mathrm{V} &= \mathrm{CNN}_{\mathcal{I}}(\mathbf{I}) \quad \text{（ VGGNet ）} && (1)\\
\mathrm{y}^{[t]} &= \mathrm{CNN}_{\mathcal{L}}(\mathrm{S}^{[0]}, \mathrm{S}^{[1]}, \cdots, \mathrm{S}^{[t-1]}) && (2)\\
\mathrm{m}^{[t]} &= f_{\text{multimodal}}(\mathrm{y}^{[t]}, \mathrm{V}) && (3)\\
\mathrm{r}^{[t]} &= f_{\text{recurrent}}(\mathrm{r}^{[t-1]}, \mathrm{x}^{[t-1]}, \mathrm{m}^{[t]}) && (4)\\
\mathrm{S}^{[t]} &\sim \arg\max_{\mathcal{S}} \mathrm{Softmax}(\mathbf{W}_o \mathrm{r}^{[t]} + \mathbf{b}_o) && (5)
\end{aligned}
$$

Language CNN Layer



Language CNN Layer

Word Embedding :

$$\mathbf{x} = \left[\mathrm{x}^{[0]}, \mathrm{x}^{[1]}, \cdots, \mathrm{x}^{[t-1]}\right]^T, \mathbf{x} \in \mathbb{R}^{t \times K} \tag{6}$$

Convolution:

$$\mathrm{y}_i^{(\ell)}(\mathbf{x}) = \sigma(\mathbf{w}_L^{(l)} \mathbf{y}_i^{(\ell-1)} + \mathrm{b}_L^{(\ell)}) \tag{7}$$

$$\mathbf{y}^{(0)} \stackrel{\text{def}}{=} \begin{cases} \left[\mathrm{x}^{[t-L_{\mathcal{L}}]}, \cdots, \mathrm{x}^{[t-1]}\right]^T, & \text{if } t \geq L_{\mathcal{L}} \\ \left[\mathrm{x}^{[0]}, \cdots, \mathrm{x}^{[t-1]}, \tilde{\mathrm{x}}^{[t]}, \cdots, \tilde{\mathrm{x}}^{[L_{\mathcal{L}}-1]}\right]^T & \text{otherwise} \end{cases} \tag{8}$$

Multimodal Fusion Layer

Fuse sentence representation y[t] and **image features V**.

$$\mathrm{m}^{[t]} = f_{\mathrm{multimodal}}(\mathrm{y}^{[t]}, \mathrm{V}) \tag{9}$$

$$= \sigma\left(f_{\mathrm{y}}(\mathrm{y}^{[t]}; \mathbf{W}_{\mathrm{Y}}, \mathrm{b}_{\mathrm{Y}}) + g_{\mathrm{v}}(\mathrm{V}; \mathbf{W}_{\mathrm{V}}, \mathrm{b}_{\mathrm{V}})\right) \tag{10}$$

Recurrent Networks

Transition equations

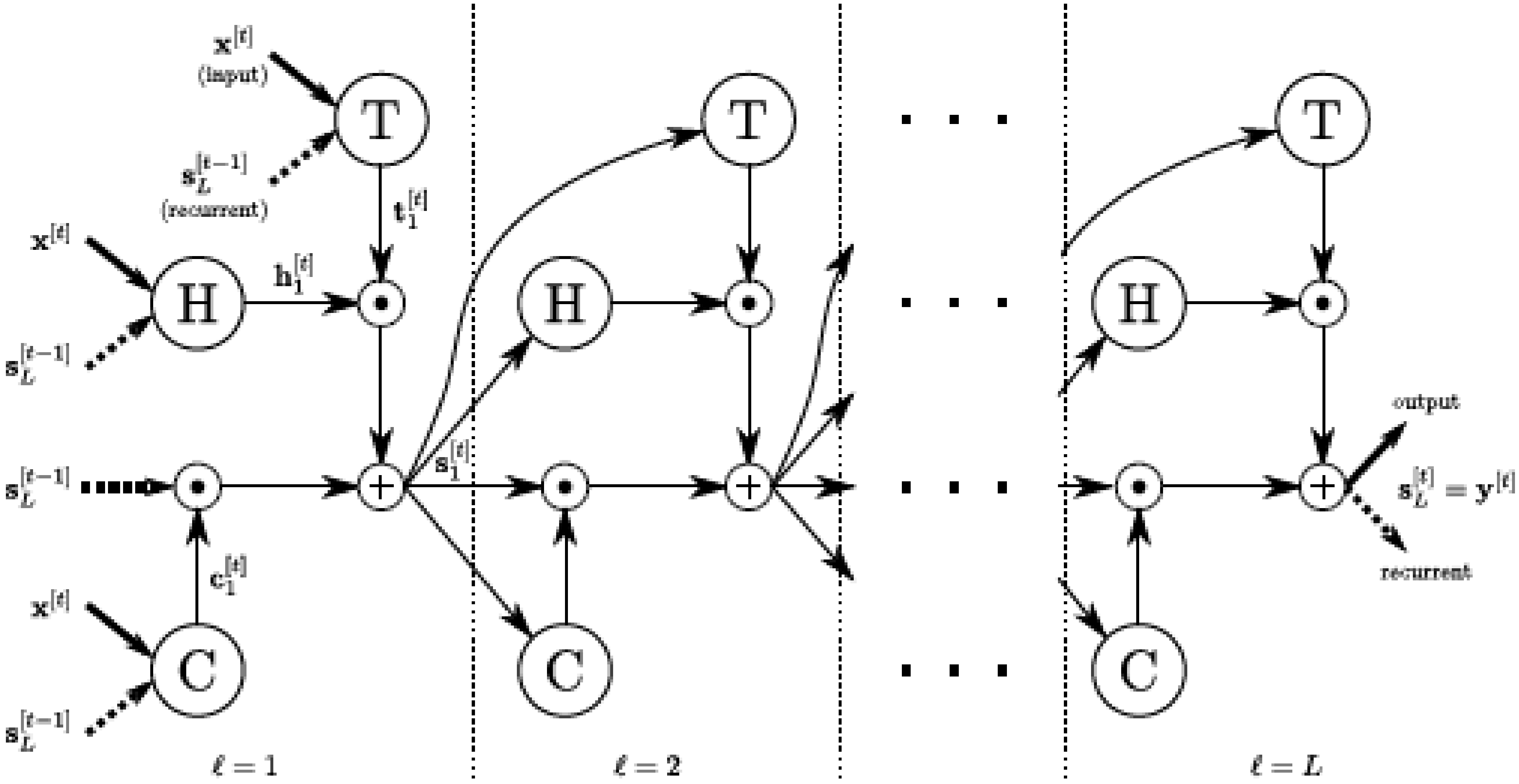
$$\mathrm{r}^{[t]} = f_{\mathrm{recurrent}}(\mathrm{r}^{[t-1]}, \mathrm{x}^{[t-1]}, \mathrm{m}^{[t]}) \tag{11}$$

$$\mathrm{S}^{[t]} \sim \arg\max_{S} \mathrm{Softmax}(\mathbf{W}_o \mathrm{r}^{[t]} + \mathrm{b}_o) \tag{12}$$

Language CNN +

Simple RNN / LSTM / GRU / Recurrent Highway network (RHN)

Recurrent Highway Network



Recurrent Highway Network

Transition equations

M： $R^{2K+d} \rightarrow R^{3d}$

$$\begin{pmatrix} \mathrm{t}^{[t]} \\ \mathrm{c}^{[t]} \\ \mathrm{h}^{[t]} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(\mathrm{M} \begin{pmatrix} \mathrm{r}^{[t-1]} \\ \mathrm{z}^{[t]} \end{pmatrix} \right) \quad (14)$$

$$\mathrm{r}^{[t]} = \mathrm{h}^{[t]} \odot \mathrm{t}^{[t]} + \mathrm{c}^{[t]} \odot \mathrm{r}^{[t-1]} \quad (15)$$

$$\mathrm{z}^{[t]} = [f_{\mathrm{multimodal}}(\mathrm{CNN}_{\mathcal{L}}(\mathrm{x}^{[0,\cdots,t-1]}), \mathrm{V}); \mathrm{x}^{[t-1]}] \quad (16)$$

Implementation Details

Training:

All models are trained with **Adam**

Dropout and **early stopping** are used to avoid overfitting

Testing:

Beam Search technology (beam search size = 2)

Experiments - Datasets & Evaluation

Datasets : MS COCO and **Flickr30k**.

Metrics :

- **BLEU-n**
- **METEOR**
- **CIDEr** : Cosine similarity with TFIDF weighting.
- **SPICE** : SPICE is Calculated as an F-score over tuples, and measures how well caption models recover objects, attributes and relations.

Experiments - Models

Recurrent Network-based Models : Simple RNN, RHN, LSTM, and GRU.

Language CNN-based Models : CNN_L + Simple RNN, CNN_L + RHN, CNN_L + LSTM, and CNN_L + GRU.

Experimental Results – Analysis of CNN_L on MSCOCO

Approach	Params	B@4	C	Approach	Params	B@4	C
Simple RNN	5.4M	27.0	87.0	LSTM	7.0M	29.2	92.6
$CNN_{\mathcal{L}}$	6.3M	18.4	56.8	$LSTM_2$	9.1M	**29.7**	93.2
$CNN_{\mathcal{L}}$+RNN	**11.7M**	29.5	**95.2**	$LSTM_3$	11.2M	29.3	92.9

CNN_L : "***a person on a wave***"

CNN_L+RNN : "***a young man surfing a wave***".

Experimental Results – Analysis of CNNL on MSCOCO

Approach	B@4	C	Approach	B@4	C
$\text{Avg}_{\text{history}}$+RHN	30.1	95.8	$\text{CNN}_{\mathcal{L}_{2\text{ words}}}$+RHN	29.2	93.8
$\text{CNN}_{\mathcal{L}^{*}_{16\text{ words}}}$+RHN	28.9	91.9	$\text{CNN}_{\mathcal{L}_{4\text{ words}}}$+RHN	29.5	95.8
$\text{CNN}_{\mathcal{L}}$+RHN	**30.6**	**98.9**	$\text{CNN}_{\mathcal{L}_{8\text{ words}}}$+RHN	30.0	95.9

Experimental Results - on MS COCO

Approach	B@1	B@2	B@3	B@4	M	C	S
Simple RNN	70.1	52.1	37.6	27.0	23.2	87.0	16.0
$\text{CNN}_{\mathcal{L}}$+RNN	72.2	55.0	40.7	29.5	24.5	95.2	17.6
RHN	70.5	52.7	37.8	27.0	24.0	90.6	17.2
$\text{CNN}_{\mathcal{L}}$+RHN	72.3	55.3	**41.3**	**30.6**	**25.2**	98.9	**18.3**
LSTM	70.8	53.6	39.5	29.2	24.5	92.6	17.1
$\text{CNN}_{\mathcal{L}}$+LSTM	72.1	54.6	40.9	30.4	25.1	**99.1**	18.0
GRU	71.6	54.1	39.7	28.9	24.3	93.3	17.2
$\text{CNN}_{\mathcal{L}}$+GRU	**72.6**	**55.4**	41.1	30.3	24.6	96.1	17.6

Experimental Results - on Flickr30k

Approach	B@1	B@2	B@3	B@4	M	C	S
Simple RNN	60.5	41.3	28.0	19.1	17.1	32.5	10.5
$\text{CNN}_{\mathcal{L}}$+RNN	71.3	53.8	39.6	28.7	**22.6**	**65.4**	**15.6**
RHN	62.1	43.1	29.4	20.0	17.7	38.4	11.4
$\text{CNN}_{\mathcal{L}}$+RHN	**73.8**	**56.3**	**41.9**	**30.7**	21.6	61.8	15.0
LSTM	60.9	41.8	28.3	19.3	17.6	35.0	11.1
$\text{CNN}_{\mathcal{L}}$+LSTM	64.5	45.8	32.2	22.4	19.0	45.0	12.5
GRU	61.4	42.5	29.1	20.0	18.1	39.5	11.4
$\text{CNN}_{\mathcal{L}}$+GRU	71.4	54.0	39.5	28.2	21.1	57.9	14.5

Experimental Results

Approach	*Flickr30k*					*MS COCO*					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
BRNN [19]	57.3	36.9	24.0	15.7	—	62.5	45.0	32.1	23.0	19.5	66.0
Google NIC [46]	—	—	—	—	—	—	—	—	27.7	23.7	85.5
LRCN [6]	58.8	39.1	25.1	16.5	—	66.9	48.9	34.9	24.9	—	—
MSR [7]	—	—	—	—	—	—	—	—	25.7	23.6	—
m-RNN [35]	60.0	41.0	28.0	19.0	—	67.0	49.0	35.0	25.0	—	—
Hard-Attention [51]	66.9	43.9	29.6	19.9	18.5	70.7	49.2	34.4	24.3	23.9	—
Soft-Attention [51]	66.7	43.4	28.8	19.1	18.5	71.8	50.4	35.7	25.0	23.0	—
ATT-FCN [53]	64.7	46.0	32.4	23.0	18.9	70.9	53.7	40.2	30.4	24.3	—
ERD+GoogLeNet [52]	—	—	—	—	—	—	—	—	29.8	24.0	88.6
emb-gLSTM [15]	64.6	44.6	30.5	20.6	17.9	67.0	49.1	35.8	26.4	22.7	81.3
VAE [40]	72.0	53.0	38.0	25.0	—	72.0	52.0	37.0	28.0	24.0	90.0
	State-of-the-art results using model assembling or extra information										
Google NICv2 [47]	—	—	—	—	—	—	—	—	32.1	25.7	99.8
Attributes-CNN+RNN [50]	73.0	55.0	40.0	28.0	—	74.0	56.0	42.0	31.0	26.0	94.0
	Our results										
$\text{CNN}_{\mathcal{L}}$+RNN	71.3	53.8	39.6	28.7	22.6	72.2	55.0	40.7	29.5	24.5	95.2
$\text{CNN}_{\mathcal{L}}$+RHN	73.8	56.3	41.9	30.7	21.6	72.3	55.3	41.3	30.6	25.2	98.9
$\text{CNN}_{\mathcal{L}}$+LSTM	64.5	45.8	32.2	22.4	19.0	72.1	54.6	40.9	30.4	25.1	99.1
$\text{CNN}_{\mathcal{L}}$+GRU	71.4	54.0	39.5	28.2	21.1	72.6	55.4	41.1	30.3	24.6	96.1

Experimental Results

CNN$_L$+RHN : a black and white cat looking at itself in a mirror
CNN$_L$+RNN : a black and white cat sitting in front of a mirror
GRU : a black and white cat standing next to a mirror
LSTM : a black and white cat sitting in a bathroom sink
RNN : a cat sitting on the floor in a bathroom

- *there is a black tuxedo cat looking in the mirror*
- *two cats sitting on top of a wooden floor*
- *a cat looking at itself in the mirror next to a tripod*
- *a cat and a tripod sitting in front of a mirror*
- *a close up of a cat in a mirror*

CNN$_L$+RHN : a man standing next to a child on a snow covered slope
CNN$_L$+RNN : a man and a woman standing on a snow covered slope
GRU : a man and a child standing on a snow covered slope
LSTM : a man and a child are standing in the snow
RNN : a man and a woman are skiing on the snow

- *a woman and child in ski gear next to a lodge*
- *a man and a child are smiling while standing on skiis*
- *a young man poses with a little kid in the snow*
- *an adult and a small child dressed for skiing*
- *a man and a little girl in skis stand in front of a mountain lodge*

Experimental Results

CNNL+RHN : a large bird perched on top of a tree

- *a bear that is hanging in a tree*
- *a young bear holding onto a pine tree*
- *a bear cub in the branches of a pine tree*
- *a black bear cub climbing a pine tree*
- *the bear cub UNK high up into the tree*

CNNL+RNN : a black and white dog standing on a sidewalk

- *a tan dog standing on a sidewalk next to a UNK and grass*
- *the dog is standing outside all alone in the backyard*
- *a dog standing on a brick walk way*
- *a brown dog is standing on the side of a walk way*
- *a brown dog standing on a brick path*

Conclusion

- In this work, we present an image captioning model with language CNN to explore both hierarchical and temporal information in sequence for image caption generation.

- Future research directions will go towards integrating extra attributes learning into image captioning, and how to apply a single language CNN for image caption generation is worth trying.

Thank You !