Keyword-driven Image Captioning via Context-dependent Bilateral LSTM

Xiaodan Zhang[1,2],

Shuqiang Jiang[2], Qixiang Ye[2], Jianbin Jiao[2], Rynson W.H. Lau[1]

[1] City University of Hong Kong

[2] University of Chinese Academy of Sciences

Image captioning

Image

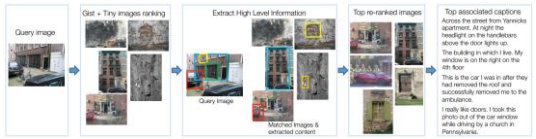
Language description

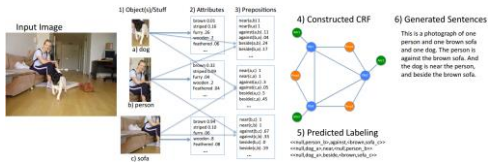
a group of children playing baseball out side.

a young boy in a striped shirt is leaning against a tree while another child sits at a picnic table

a tan and white dog swimming towards a waterfall

History and related works

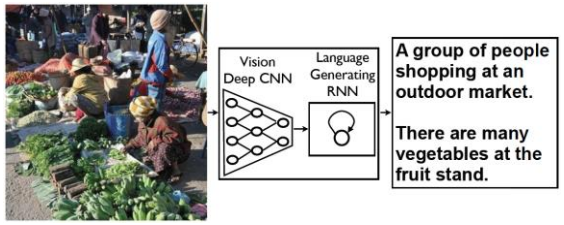




Search-based | Template-based

2011

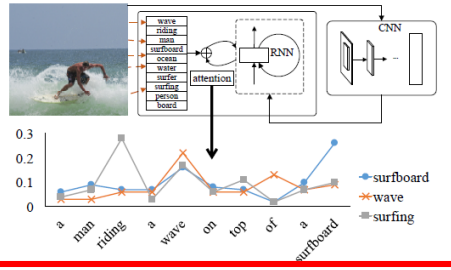
1. V. Ordonez et al. Im2text: Describing images using 1 million captioned photographs. NIPS 2011.
2. G. Kulkarni *et al.* Baby talk: Understanding and generating simple image descriptions. CVPR 2011.
3. H. Fang et al. From Captions to Visual Concepts and Back. CVPR 2015.



CNN-RNN based

2015

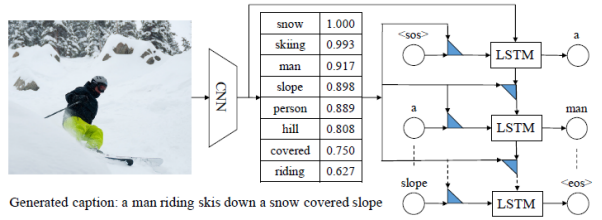
1. O. Vinyals *et al*. Show and tell: a neural image caption generator. CVPR 2015.
2. A. Karpathy et al. Deep visual-semantic alignments for generating image descriptions. CVPR 2015.
3. X. Jia *et al*. Guiding Long Short Term Memory for image caption generation. ICCV 2015.
4. K. Xu *et al*. Show, attend and tell: neural image caption generation with visual attention. ICML 2015.



CNN-RNN based

2016

1. Q. You *et al*. Image captioning with semantic attention. CVPR 2016.
2. Q. Wu *et al*. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? CVPR 2016.
3. J. Johnson *et al.* DenseCap: Fully Convolutional Localization Networks for Dense Captioning**.** CVPR 2016.



CNN-RNN-based

2017

1. G. Zhe *et al*. Semantic Compositional Networks for Visual Captioning. CVPR 2017.
2. Z. Ren *et al*. Deep Reinforcement Learning-based Image Captioning with Embedding Reward**.** CVPR 2017.

CNN-RNN based Image captioning model

Vision
Deep CNN

Language
Generating
RNN

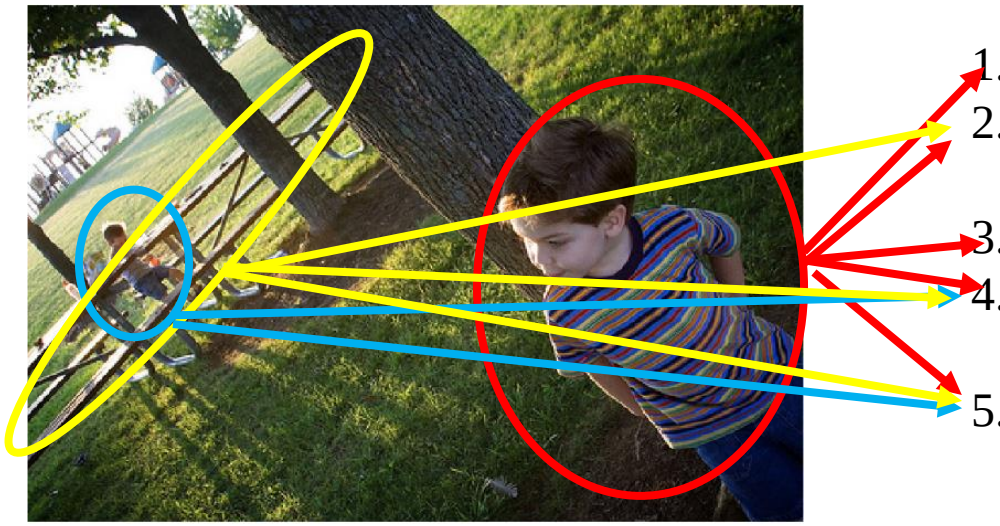
a tan and white dog swimming towards a waterfall

Deep CNN:
- GoogLeNet
- VGGNet

RNN:
- RNN
- LSTM

[1] O Vinyals, A Toshev, S Bengio, D Erhan, Show and tell: A neural image caption generator, CVPR 2015.

Limitations



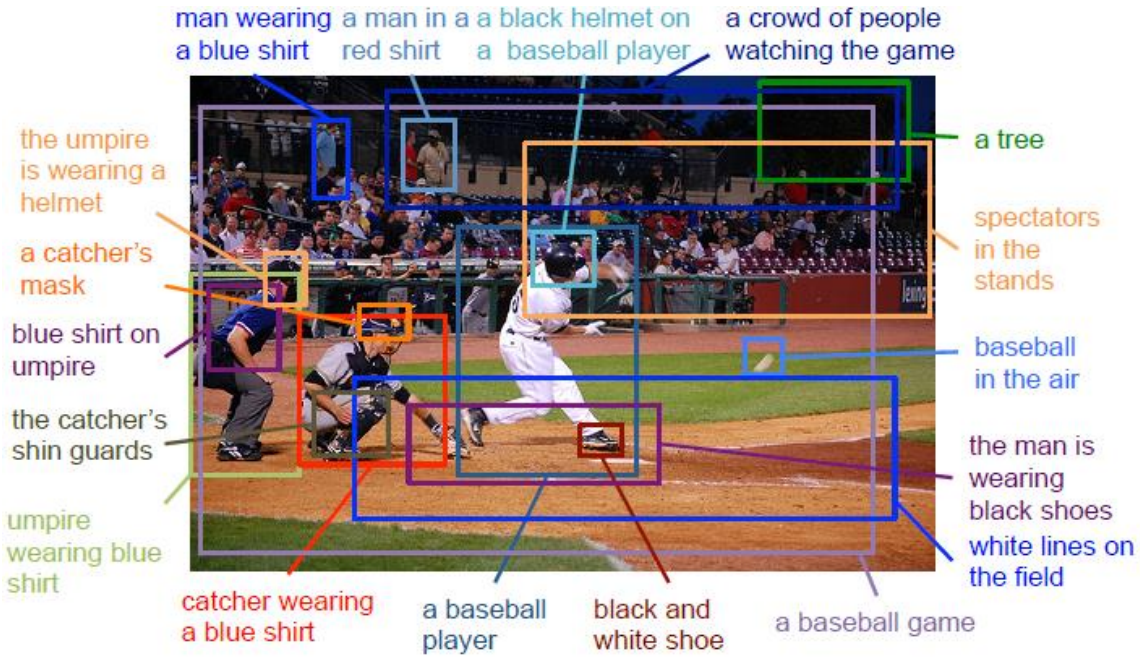
1. A boy hiding behind a tree
2. A boy in a striped t-shirt is standing by a tree in front of the picnic tables
3. A little boy in a striped shirt is standing behind a tree
4. A young boy in a striped shirt is leaning against a tree while another child sits at a picnic table
5. Two boys in a park , one standing near a tree and one sitting at a picnic table with the playground behind them .

Ambiguity

- An image contains too much information to be precisely described in one sentence.
- Image captioning suppose to be a highly customized task, and the user may have different focus for one image.

Prior arts

Dense captioning [1, 2]



Paragraph description [3]

Sentences

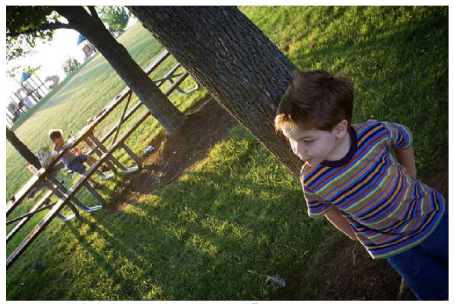
1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

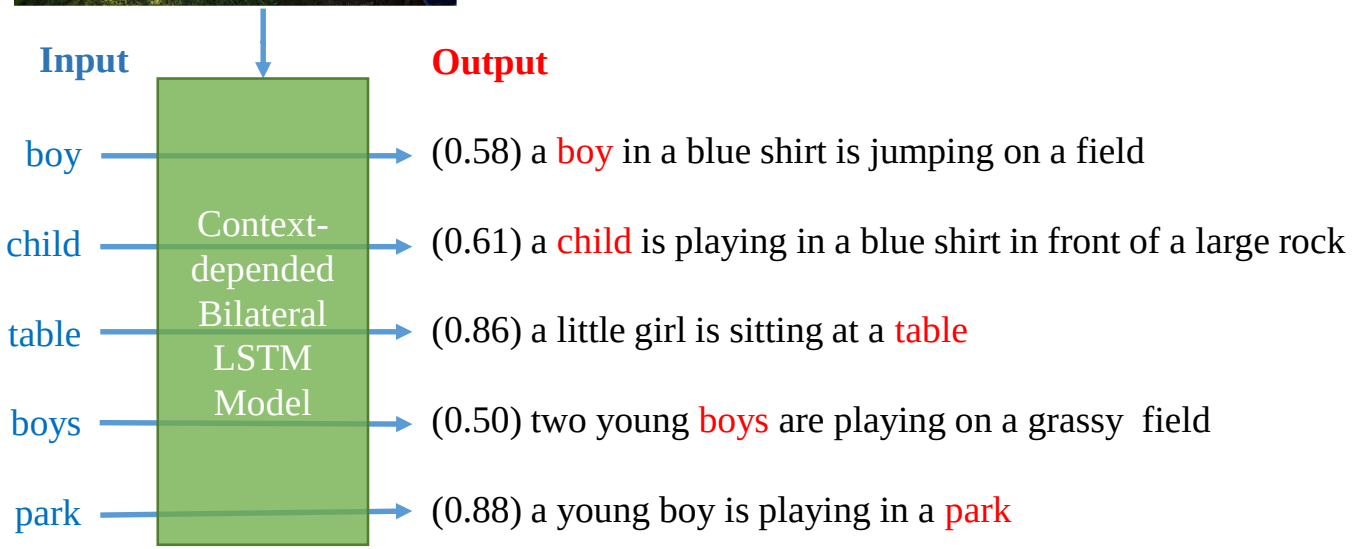
1. Justin Johnson, Andrej Karpathy, Li Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. CVPR 2016 .
2. Linjie Yang, Kevin Tang, Jianchao Yang, Li-Jia Li. Dense Captioning with Joint Inference and Visual Context. CVPR 2017.
3. Jonathan Krause, Justin Johnson, Ranjay Krishna, Li Fei-Fei. A Hierarchical Approach for Generating Descriptive Image Paragraphs. CVPR 2017.

Ours: Keyword-driven Image Captioning



Ground truth sentence:

Boys kicking soccer ball in the grass under a tree



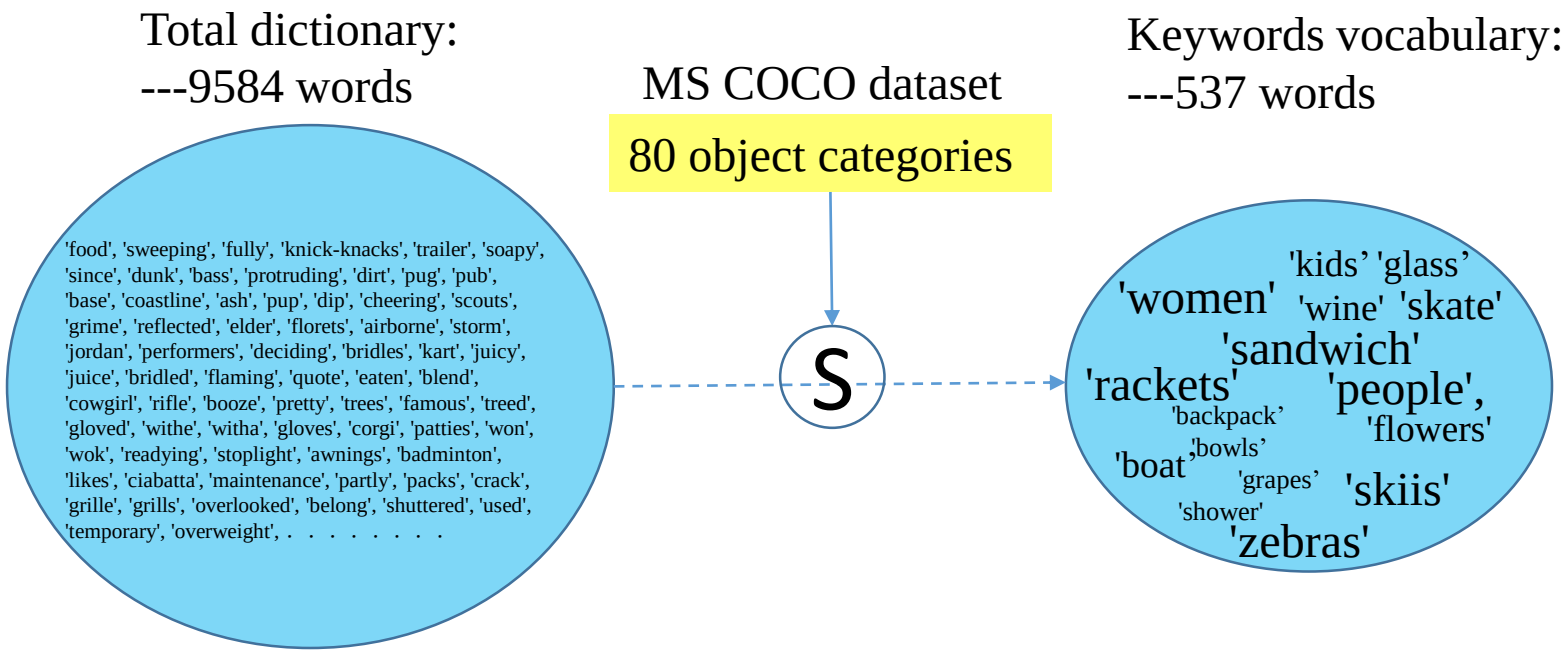
Input: An Image

Output: customized captions

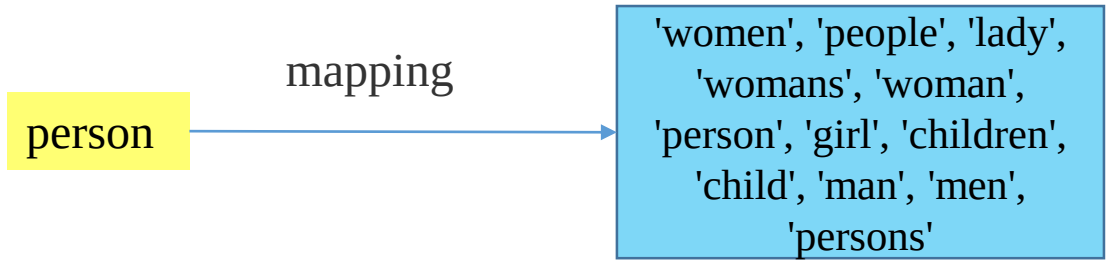
Two steps:

- Keywords generation
- Sentence generation

Keywords Generation



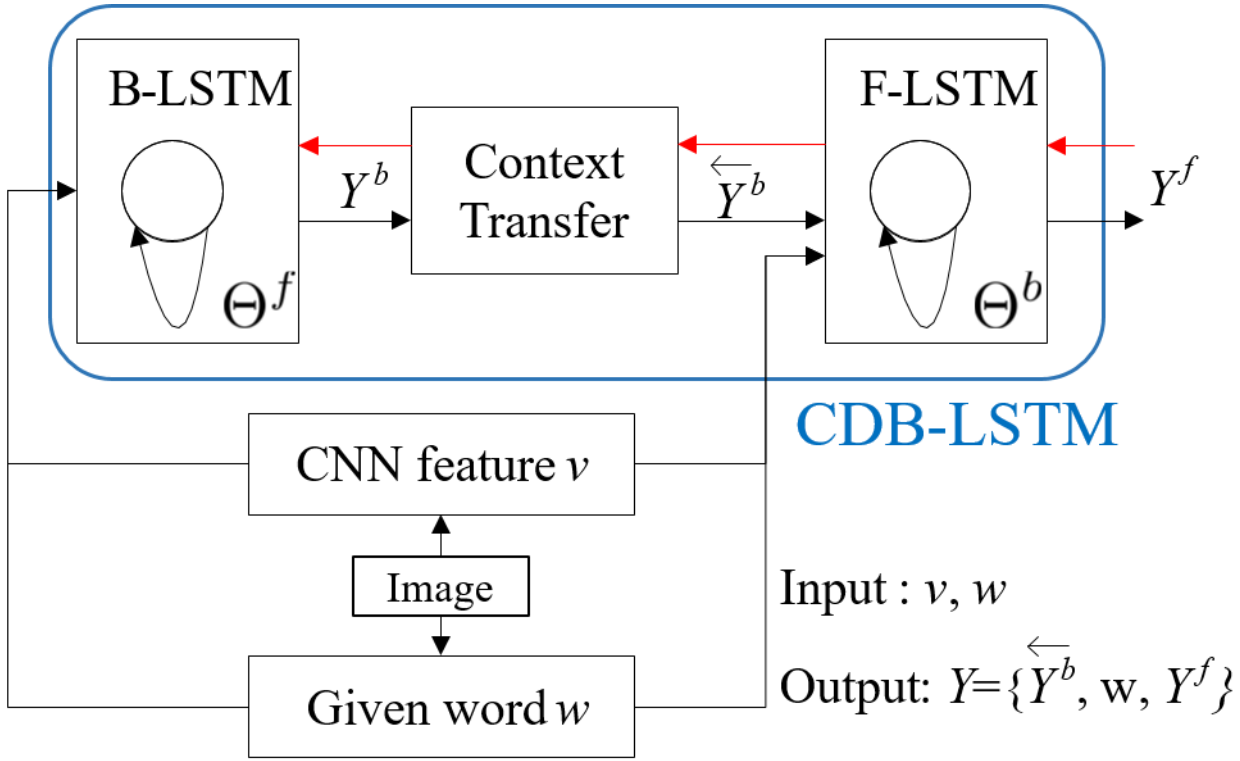
- Keyword sources
  - Word dictionary refinement
  - Object Detection



[1] R. Rehurek and P. Sojka. Software framework for topic modelling with large corpora. LREC Workshop on New Challenges for NLP Frameworks, 2010.

Sentence generation

- Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM)



CDB-LSTM is an end-to-end model that contains two cascaded sub-models:

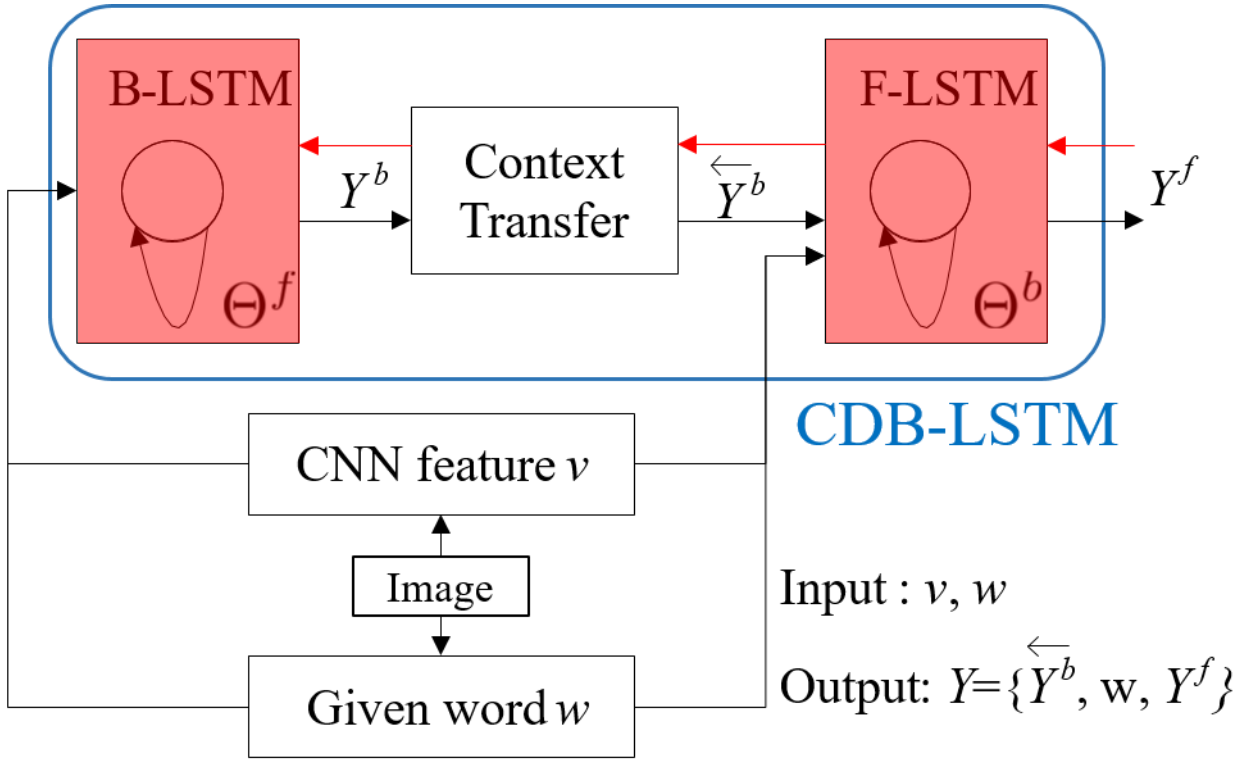
B-LSTM **F-LSTM**

GT: a brown and black dog swimming in a river

$\overleftarrow{Y}^b$ w Y^f

Sentence generation

- Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM)



CDB-LSTM is an end-to-end model that contains two cascaded sub-models:

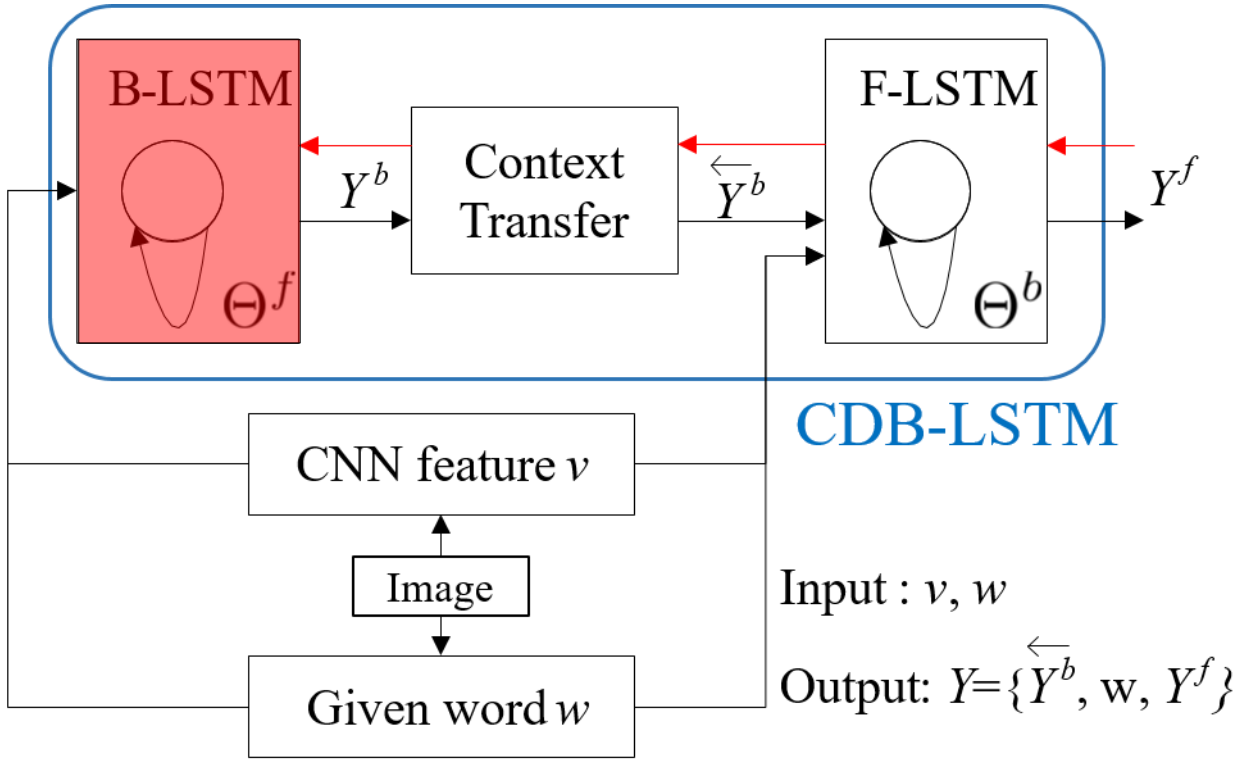
B-LSTM **F-LSTM**

GT: a brown and black dog swimming in a river

$\overleftarrow{Y}^b$ w Y^f

Sentence generation

- Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM)

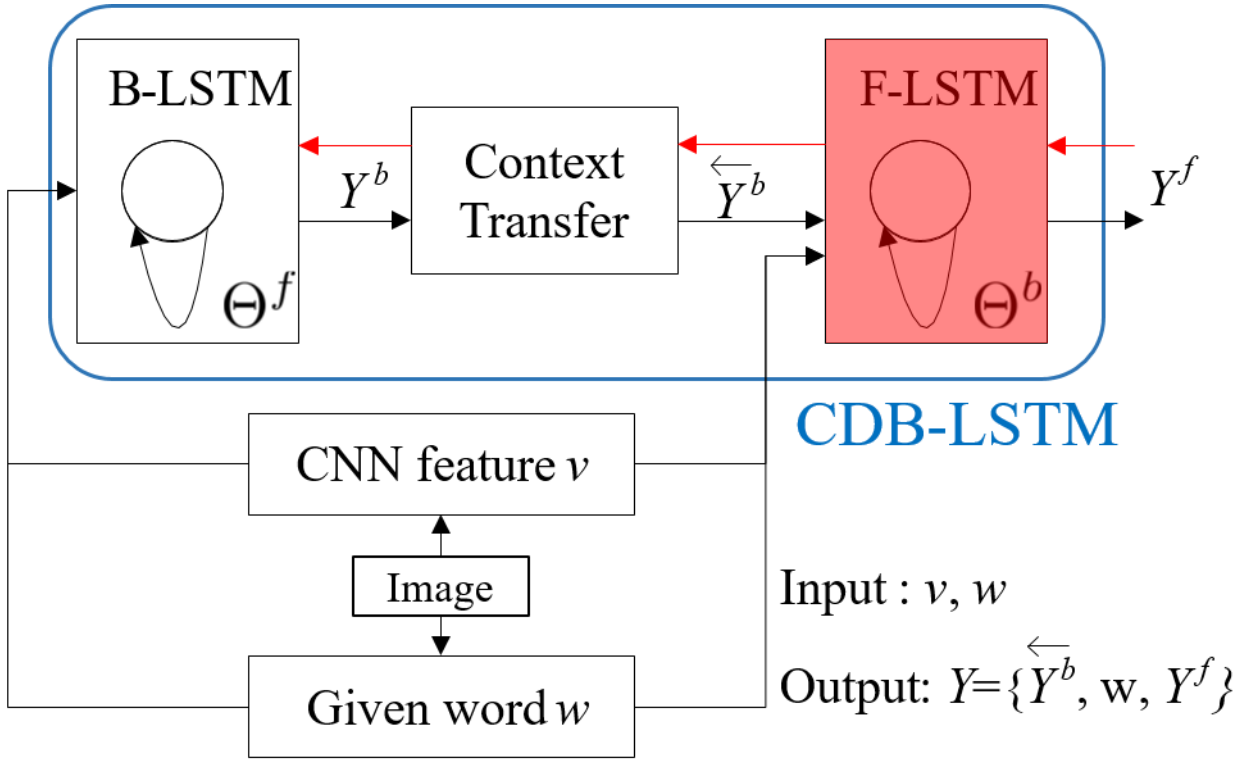


Backward Model $Y^b = LSTM^b(v, w)$

$$
\begin{aligned}
x_0 &= W_{ix}\, v, \\
x_1 &= W_{wx}\, w, \\
h_t &= RNN(h_{t-1}, x_t),\ t \in \{1, \cdots, M+1\}, \\
y_t, p_t &\propto exp(W_{hp} h_t),\ t \in \{1, \cdots, M+1\}, \\
x_{t+1} &= W_{ex} y_t,\ t \in \{1, \cdots, M\},
\end{aligned}
$$

Sentence generation

- Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM)

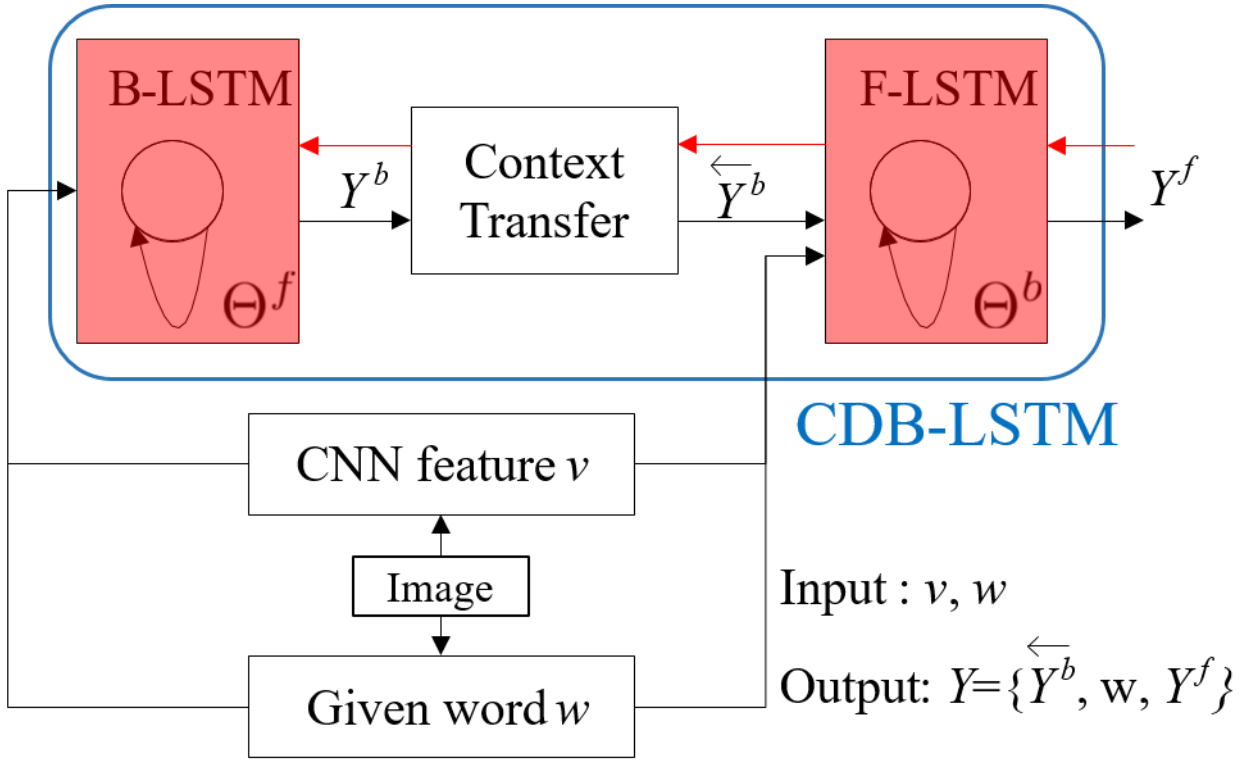


Forward Model $Y^f = LSTM^f(v, w, Y^b)$

$$
\begin{aligned}
x_0 &= W_{ix}\, v, \\
x_t &= W_{ex}\, \overleftarrow{Y}^b{}_t,\ t \in \{1, \cdots, M\}, \\
x_t &= W_{wx}\, w,\ t = M+1, \\
h_t &= RNN(h_{t-1}, x_t),\ t \in \{1, \cdots, K\}, \\
y_t, p_t &\propto exp(W_{hp}\, h_t),\ t \in \{M+1, \cdots, K\}, \\
x_{t+1} &= W_{ex} y_t,\ t \in \{M+1, \cdots, K-1\},
\end{aligned}
$$

Sentence generation

- Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM)

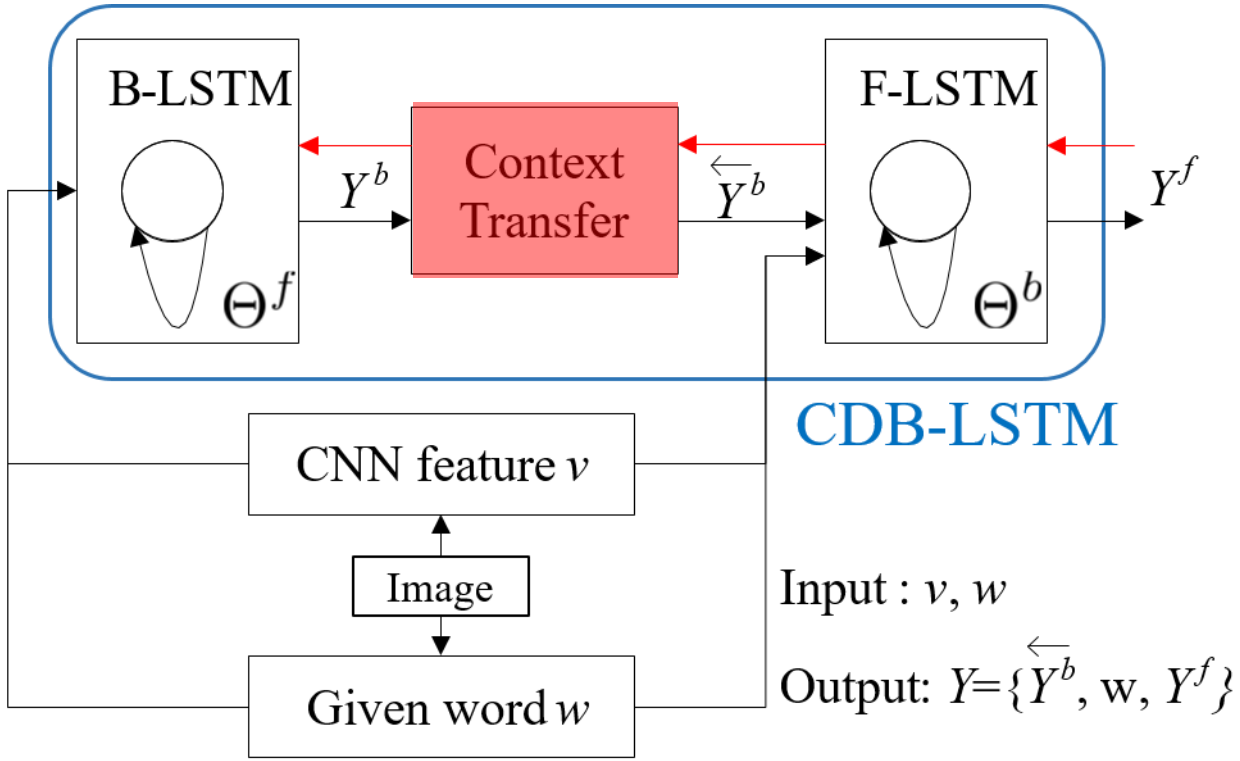


Unified Loss Function

$$
\begin{aligned}
L(B,w,F) &= L(B/w) + L(F/B,w) \\
&= -\sum_{t^b=1}^{M+1} logp\left(y_{t^b}^b | x_{t^b}^b; \Theta^b\right) \\
&\quad -\sum_{t^f=M+1}^{M+N+1} logp\left(y_{t^f}^f | x_{t^f}^f; \Theta^f\right)
\end{aligned}
$$

Sentence generation

- Pipeline of the context-dependent bilateral LSTM model (CDB-LSTM)



Context Transfer Module

Forward propagation:

$$x_t^f = W_{ex}^f \overleftarrow{Y}_t^b$$

Back propagation:

$$\frac{\partial L}{\partial y_t^b} = \frac{\partial L}{\partial L_t^b}\frac{\partial L_t^b}{\partial y_t^b} + \frac{\partial L}{\partial L_t^f}\frac{\partial L_t^f}{\partial y_t^b}$$

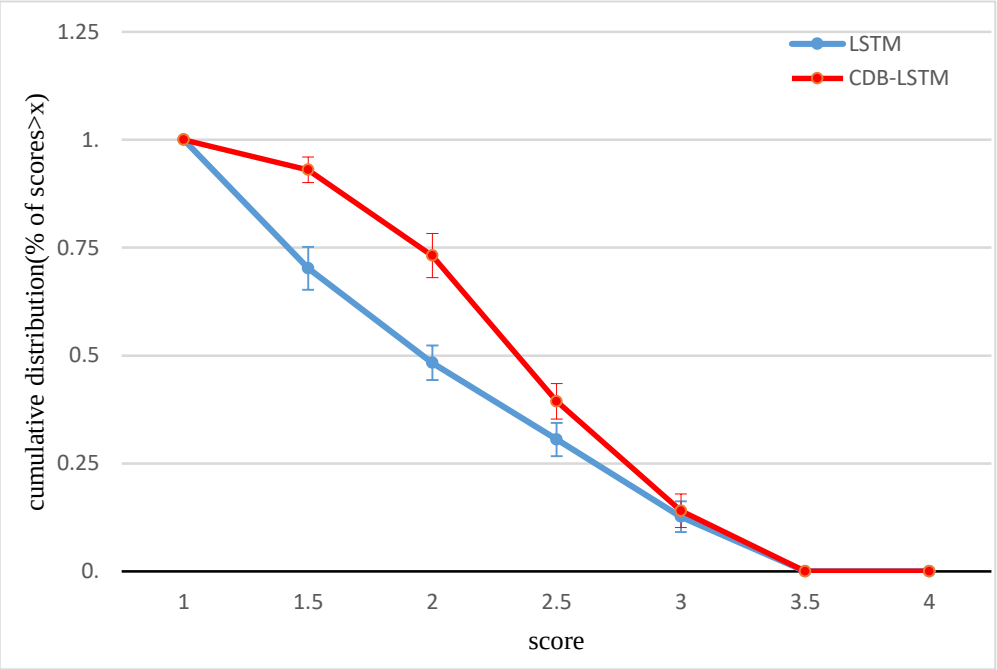
Experiments

- CNN model: VGGNet
- Word embedding: one hot vector
- RNN model: LSTM
- Dataset: MSCOCO(120K)
- Evaluation:
  - Human evaluation
  - Automatic evaluation

Results

Human evaluation

Higher cumulative distribution



100 images, 50 participants
The descriptiveness is rated on a four-point scale [1, 2], and higher is better.

[1] M. Hodosh *et al*. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR 2013.
[2] O. Vinyals *et al*. Show and tell: a neural image caption generator. CVPR 2015.

Results

Automatic evaluation:

--- BLEU, Meteor, CIDEr, ROUGE. (precision, recall grammaticality)

Method	B-1	B-2	B-3	B-4	M	C	R
Google NIC[5]	66.6	46.1	32.9	24.6	–	–	–
Hard-Attention[21]	71.8	50.4	35.7	25.0	23.04	–	–
gLSTM[14]	67.0	49.1	35.8	26.4	22.74	–	–
m-RNN[22]	67.-	49.-	35.-	25.-	–	–	–
ATT[8]	70.9	53.7	40.2	30.4	24.3	–	–
LSTM	69.8	52.2	38.5	28.7	23.9	53.4	42.9
I-LSTM(GR)	45.3	34.8	24.9	17.3	18.5	64.9	45.0
I-LSTM(GM)	66.1	50.6	35.0	23.4	20.9	77.2	48.0
CDB-LSTM(GR)	73.1	53.2	35.8	23.6	21.6	78.5	49.9
CDB-LSTM(GM)	78.8	58.3	40.4	27.5	23.4	83.6	51.8
CDB-LSTM(DR)	62.9	42.5	27.9	18.4	17.2	47.1	43.2
CDB-LSTM(DM)	76.3	56.1	38.9	26.5	22.5	77.3	51.4

CDB-LSTM *vs* LSTM

GT: Two girls in soccer strips are playing on a sports field .

LSTM: a young boy is running through a field

CDB-LSTM

(field) -- two men are playing on a field
(uniforms) -- two girls in uniforms are playing with a ball
(player) -- a soccer player in a white shirt is running on the grass
(girls) -- two girls are playing in a field

GT: A boy in a striped t-shirt is standing by a tree in front of the picnic tables .

LSTM: a boy in a red shirt is jumping over a tree

CDB-LSTM

(boy) -- a boy in a blue shirt is jumping on a field
(table) -- a little girl is sitting at a table
(boys) -- two young boys are playing on a grassy field
(park) -- a young boy is playing in a park

GT: a group of children playing baseball out side.

LSTM: a group of people playing a game of frisbee

CDB-LSTM

(children) -- a group of children playing a game of baseball
(baseball) -- a baseball player is playing baseball on a field
(gloves) -- a group of people are with gloves on a field
(kids) -- a group of kids standing on a field
(grass) -- a group of people are standing in the grass

GT: A man riding skis on top of a snow covered slope.

LSTM: a man riding skis down a snow covered slope

CDB-LSTM

(man) -- a man is holding a woman on a skateboard
(skis) -- two people on skis in the snow
(child) -- a child is on a snow covered slope
(girl) -- a girl is standing on a snow covered slope
(person) -- a person on a snowy hill with a large snow covered hill

CDB-LSTM *vs* I-LSTM

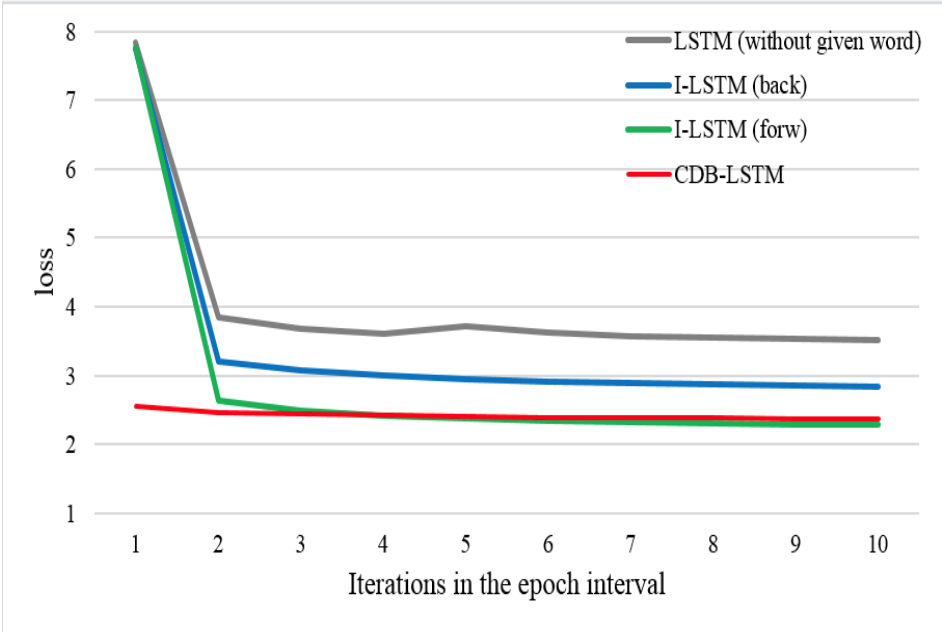
I-LSTM: a boy is doing a trick
CDB-LSTM: a boy is doing a trick on a skateboard
I-LSTM: a boy in a red jacket
CDB-LSTM: a boy in a red jacket is jumping on a skateboard

I-LSTM: a young boy is playing with a ball
CDB-LSTM: a young boy is playing with a ball in a field
I-LSTM: a boy in an orange
CDB-LSTM : a boy in an orange shirt is playing with a ball

Given word: person

I-LSTM: a person in the snow
CDB-LSTM: a person in a blue shirt and a black dog is in the air

The independent LSTM have two directions blind to each other and more likely to predict incoherent , inaccurate, or incomplete sentence.



Conclusion

- Keyword-driven image captioning
- CDB-LSTM
- Superiority in evaluation

Thanks

Ground truth sentence:

Boys kicking soccer ball in the grass under a tree