Image Captioning using Phrase-based Hierarchical LSTM Model

Chee Seng Chan PhD SMIEEE

23 October 2017

Nvidia AI Conference, Singapore

email: cs.chan@um.edu.my

Introduction

- **Aim:** *Automatic generate a full sentence describing an image.*
- Motivated by the significant progress of image classification and statistical language model.
- Applications:
  - Early childhood educations
  - Scene understanding for the visual impairments
  - Image retrievals

Two children are playing on a swing made out of a tire.

Backgrounds

- Processing of Image, *I*:
  - Represented as a vector using feature learning algorithm such as convolutional neural network (CNN)
- Processing of Language:
  - Each sentence is equivalent to a sequence of words.
  - A statistical model is trained to predict the conditional probability of next word given all previous words

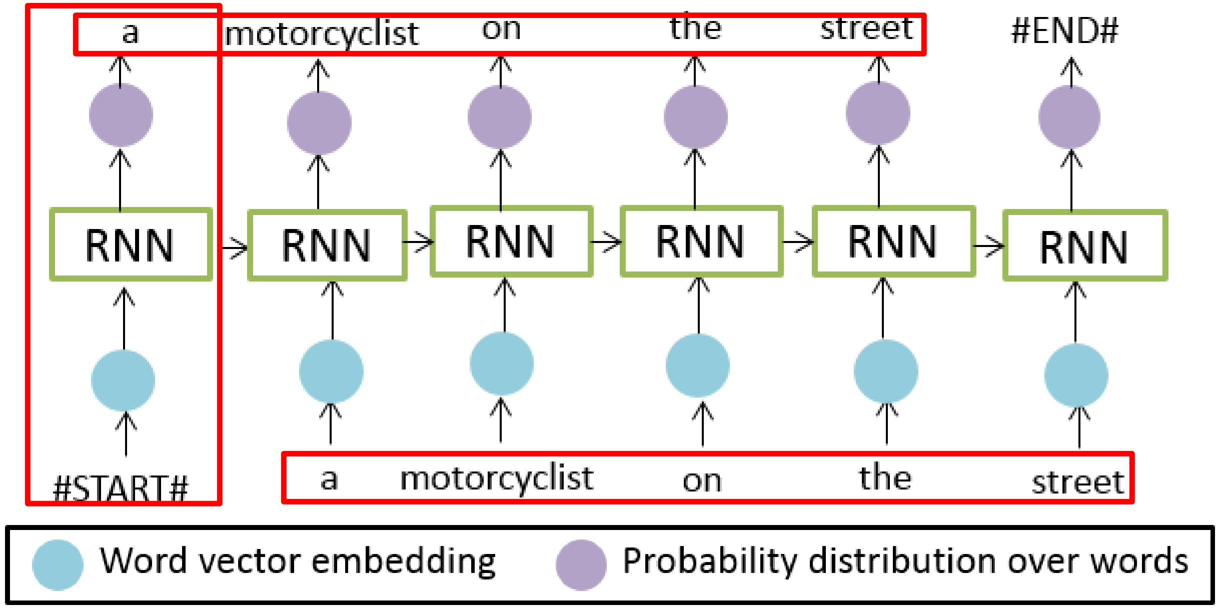
$$P(w_T) = \prod_{t=1}^{T} P(w_t|w_{t-1})$$

- Multimodal Embedding
  - Prediction of next word also conditioned on image

$$P(w_T) = \prod_{t=1}^{T} P(w_t|w_{t-1}, I)$$

Backgrounds

- Sequence is learned with Recurrent Neural Network (RNN).



- The most popular variant of RNN is Long Short-Term Memory (LSTM).

Problem Statement

- Conventional models treat a sentence as a sequence of words.
- All other linguistic syntax and structure are disregarded.
- Sentence structure is one of the most prominent characteristic of sentence!

Two dogs are running in the snow.
NP *VP* *PP* *NP*

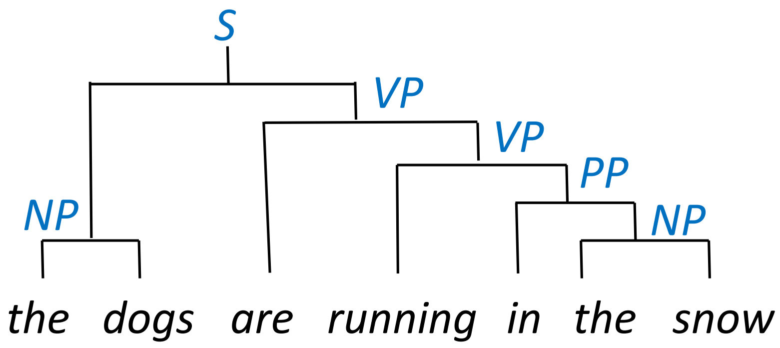
NP = noun phrase
VP = verb phrase
PP = prepositional phrase

PROBLEM STATEMENT

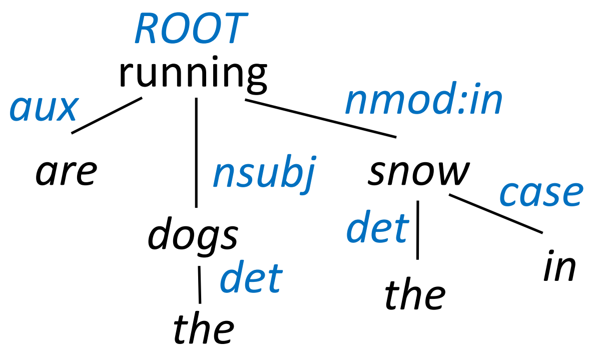
- Quoted on Victor Yngve [14] (an influential contributor in linguistic theory):

  "language structure involving, in some form or other, a phrase structure hierarchy, or immediate constituent organization"

- Example:



Phrase structure grammar



Dependency grammar

RESEARCH INTEREST & OBJECTIVE

*Is it really okay to treat sentence as only sequence of words, while **disregarding any other important characteristic of sentence** such as structure?*

1. Design of phrase-based model for image captioning. This is one of the most earliest work after **PbIC[13].**
2. Investigate on its performance as compared to a pure sequence model.

DESIGN MOTIVATION

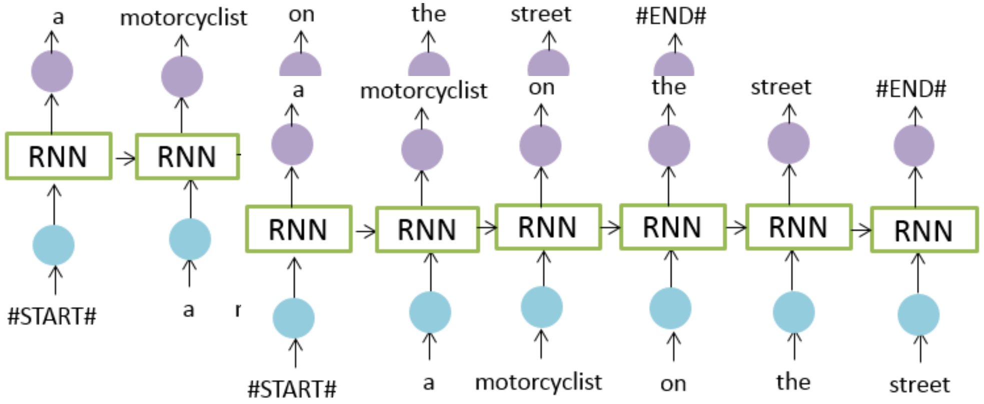
A young girl wearing a yellow shirt with a blue backpack is walking next to a fence covered with a blue plastic cover .

- Noun phrases form most of an image caption.
- They have similar syntactic role
- They have strong relation with the image.

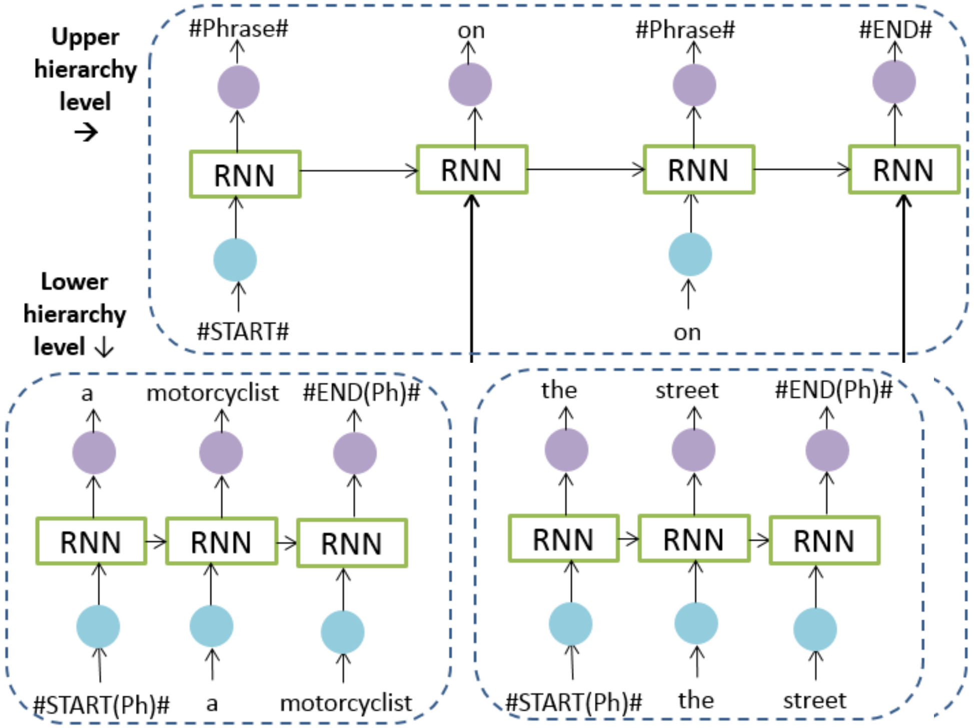
Conventional vs. Proposal

Sentence:

A motorcyclist on the street.



conventional



proposal

Word vector embedding — Probability distribution over words

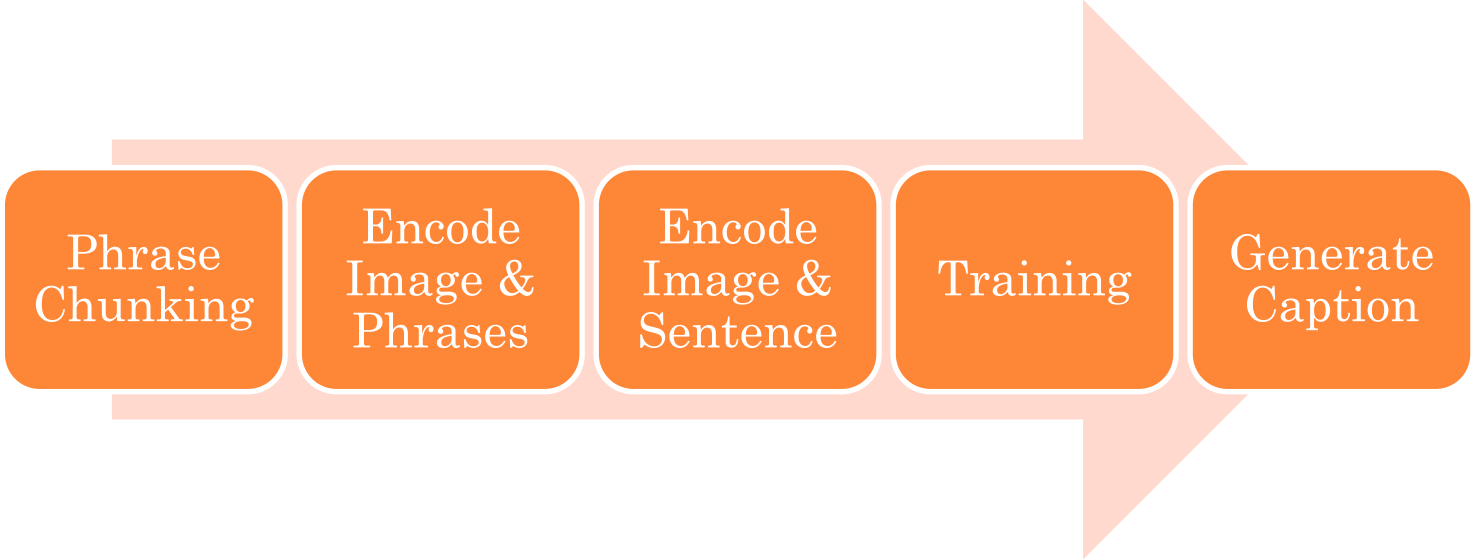
Related Works

Methods	Details (Red words are their cons)	References
Template based	• Generate sentence from a fix template. • Sentence generated is rigid.	1-4
Composition Method	• Stitch up image relevant phrases to form a sentence. • Computational cost is high.	5-7
Neural Network	• Trained to predict sequence. • Only model words sequence.	**mRNN** [8], **NIC** [9], **DeepVS** [10], **LCRNN** [12]

- The closest work is "Phrase based Image Captioning" **PbIC[13]** proposed by Lebret et al.
- They encode each sentence as <u>phrase sequence only</u> while my proposal is to encode as <u>sequence of phrase and words</u>.
- They use simpler model.

Proposed model

- Training Data: image sentence pair



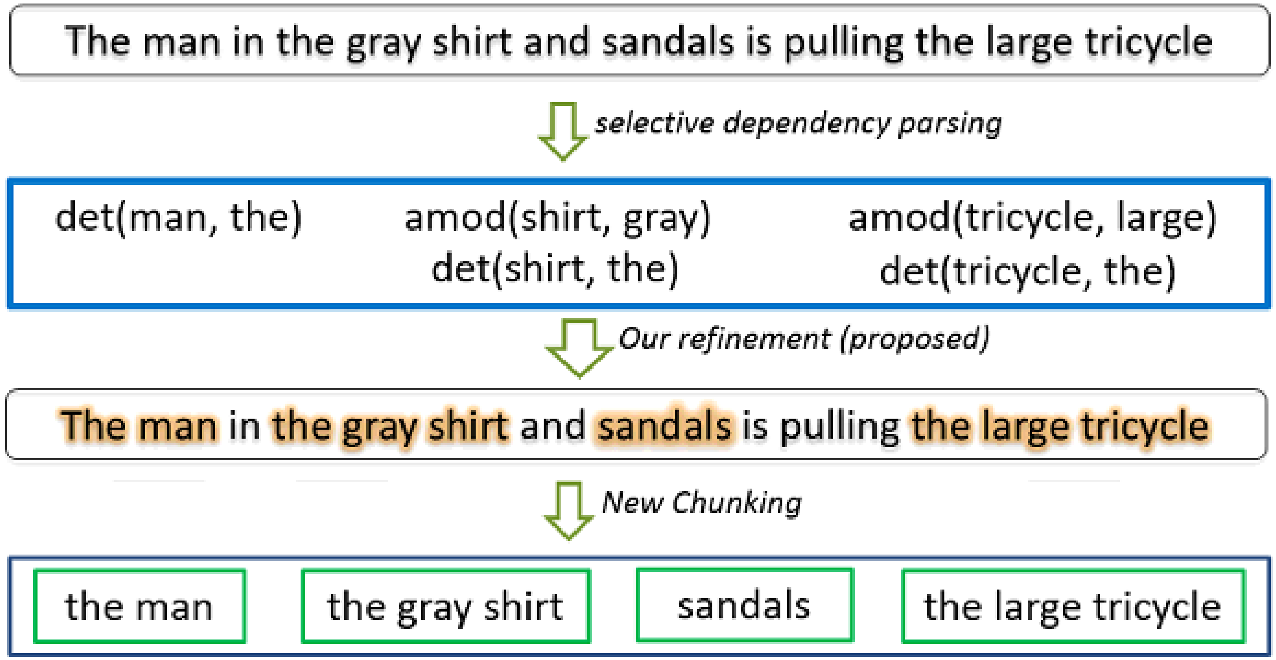
Proposed model:
1) Phrase Chunking

- Approach to identify the constituents of a sentence.
- Extract only noun phrase – prominent in image description
- **Dependency parse*** with selected relations:
  - **det** – determiner (*e.g.: "a man"*)
  - **amod** - adjective modifier (*e.g.: "green shirt"*)
  - **nummod** - numeric modifier (*e.g.: "two dogs"*)
  - **compound** - compound (*e.g.: "basketball court"*)
  - **advmod** - adverbial modifier, when modifying meaning of adjective (*e.g.: "dimly lit room"*)
  - **nmod:of** & **nmod:poss** - nominal modifier for possessive alteration (*e.g.: "his hand"*)

**Stanford CoreNLP Software - https://stanfordnlp.github.io/CoreNLP/*

Proposed model: 1) Phrase Chunking

- Chunking from dependency parse

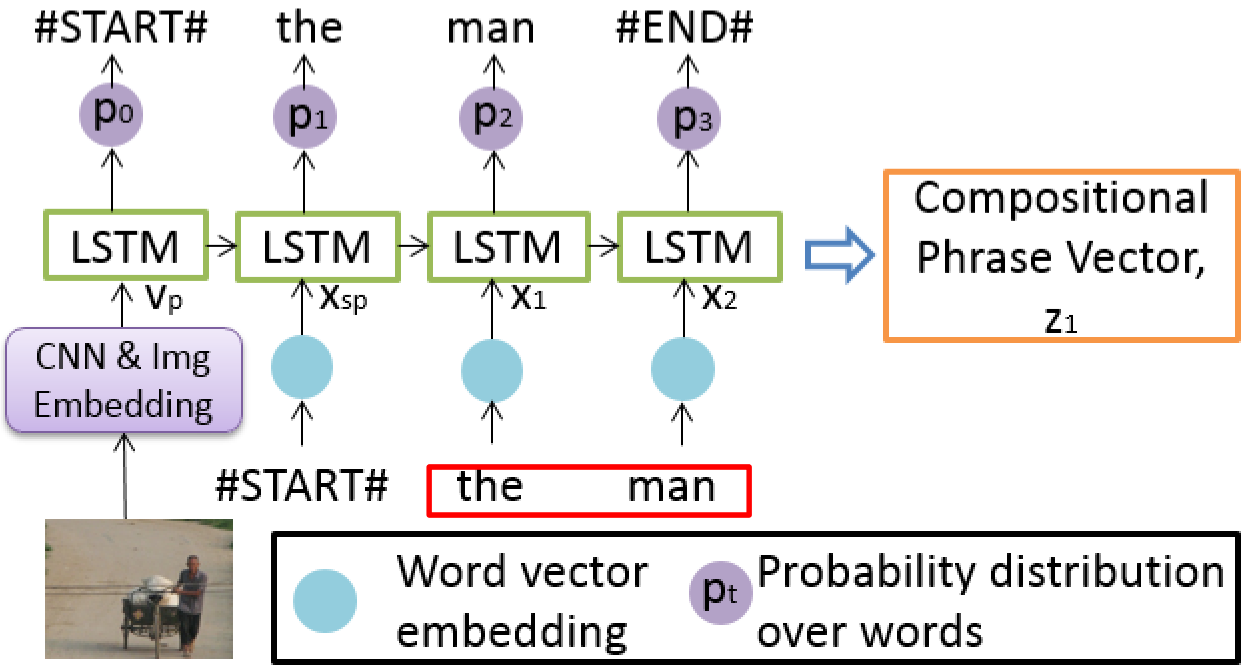


Proposed Model:
2) Compositional vector of phrase

UNIVERSITY OF MALAYA
The Leader in Research & Innovation

NVIDIA

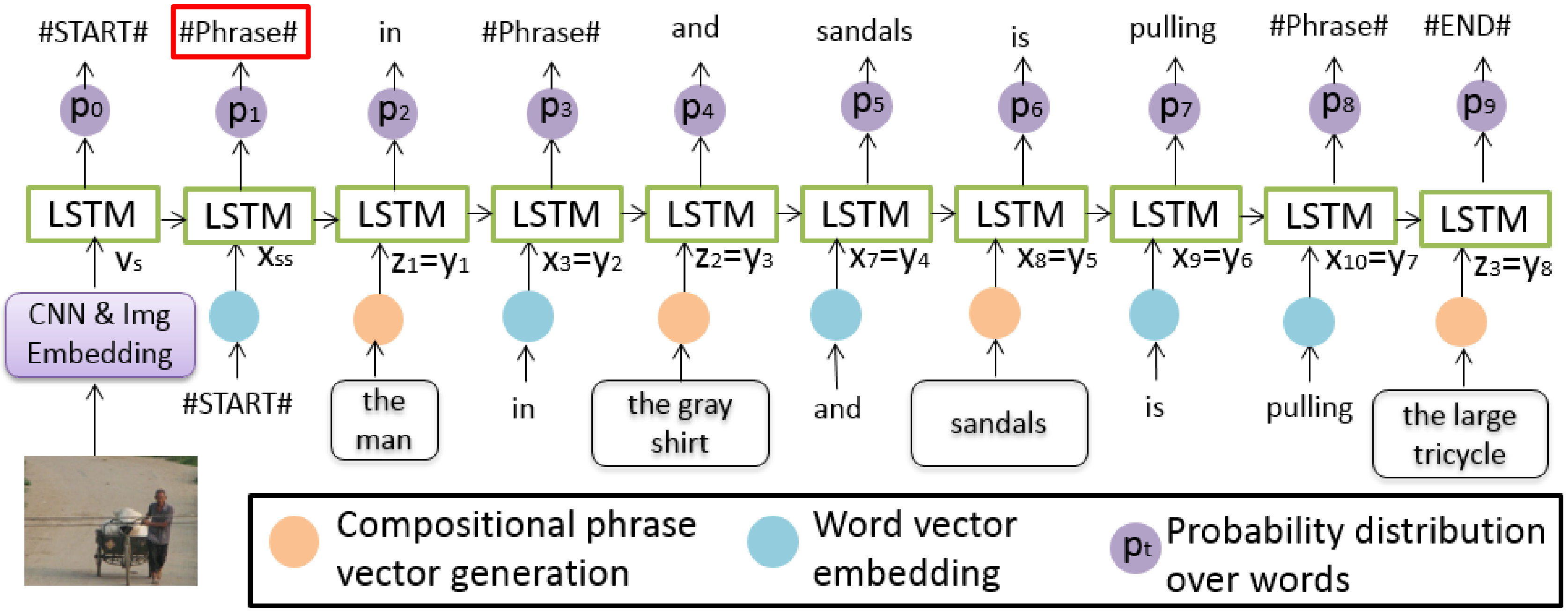
- Our proposed architecture is the hierarchical counterpart of **NIC** model proposed by Vinyals et al [9]



Phrases: *the man*, *the gray shirt*, *sandals*, *the large tricycle*

Proposed Model:
3) Sentence Encoding

Sentence:
The man in the gray shirt and sandals is pulling the large tricycle.



- A 'phrase' token is added into the corpus for prediction

TRAINING

- Objective function:

$$\mathcal{C}_F(\theta) = -\frac{1}{L}\sum_{j=1}^{M}[N_j \boxed{\log_2 \mathcal{PPL}(\mathbf{S_j}|\mathbf{I_j})} + \mathcal{C}_{PSj}] + \lambda_\theta \cdot \parallel \theta \parallel_2^2 \qquad L = M \times \sum_{j=1}^{M} N_j \ .$$

j / M = index / total no of training sentence

- Perplexity:

$$\log_2 \mathcal{PPL}(\mathbf{S}|\mathbf{I}) = -\frac{1}{S}\sum_{t_s=-1}^{S} \log_2 \mathbf{p}_{\mathbf{t}_s}$$

$$\log_2 \mathcal{PPL}(\mathbf{S}|\mathbf{I}) = -\frac{1}{N}\left[\sum_{t_s=-1}^{Q} \log_2 \mathbf{p}_{\mathbf{t}_s} + \sum_{i=1}^{R}\left[\sum_{t_p=-1}^{P_i} \log_2 \mathbf{p}_{\mathbf{t}_p}\right]\right] , \quad N = Q + \sum_{i=1}^{R} P_i \ .$$

$\mathbf{p}_{\mathbf{t}_p}$ / $\mathbf{p}_{\mathbf{t}_s}$ = probability distribution over words on the particular time step for phrase / sentence

t_p / P = time step / total no. of time step in phrase

t_s / Q = time step / total no. of time step in sentence

i / R = index / total no. of phrase in sentence **I**

Training – phrase selection objective

- Objective function:

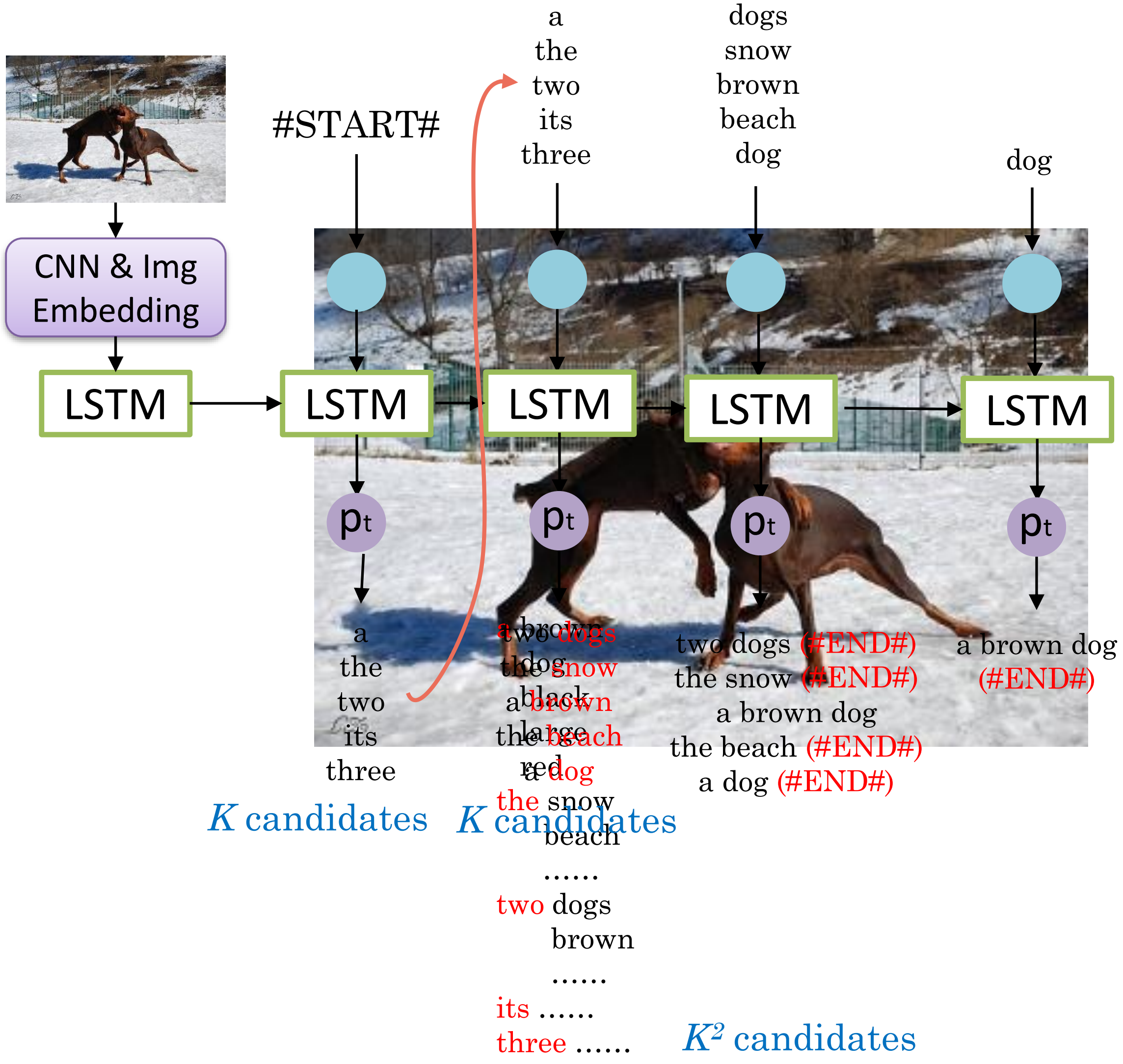
$$\mathcal{C}_F(\theta) = -\frac{1}{L}\sum_{j=1}^{M}[N_j \log_2 \mathcal{PPL}(\mathbf{S_j}|\mathbf{I_j}) + \boxed{\mathcal{C}_{PSj}]} + \lambda_\theta \cdot \parallel \theta \parallel_2^2$$

- Cost of phrase selection objective:

$$\mathcal{C}_{PS} = \sum_{t_s \in \mathcal{P}} \sum_{k=1}^{H} \kappa_{t_s k} \sigma(1 - y_{t_s k} h_{t_s k} \mathbf{W_{ps}}) \ .$$

$\mathbf{W_{ps}}$ = trainable parameters
$h_{t_s k}$ = hidden output at t_s for input k
$y_{t_s k}$ = label of input k at t_s
$\kappa_{t_s k}$ = normalizing constant based on
k / H = index / total no of inputs at t_s
$\mathcal{P}$ = set of t_s which the input is phrase

Graphical Illustration: Sentence generation (Phrase Level)



Selected Phrases:

a brown dog

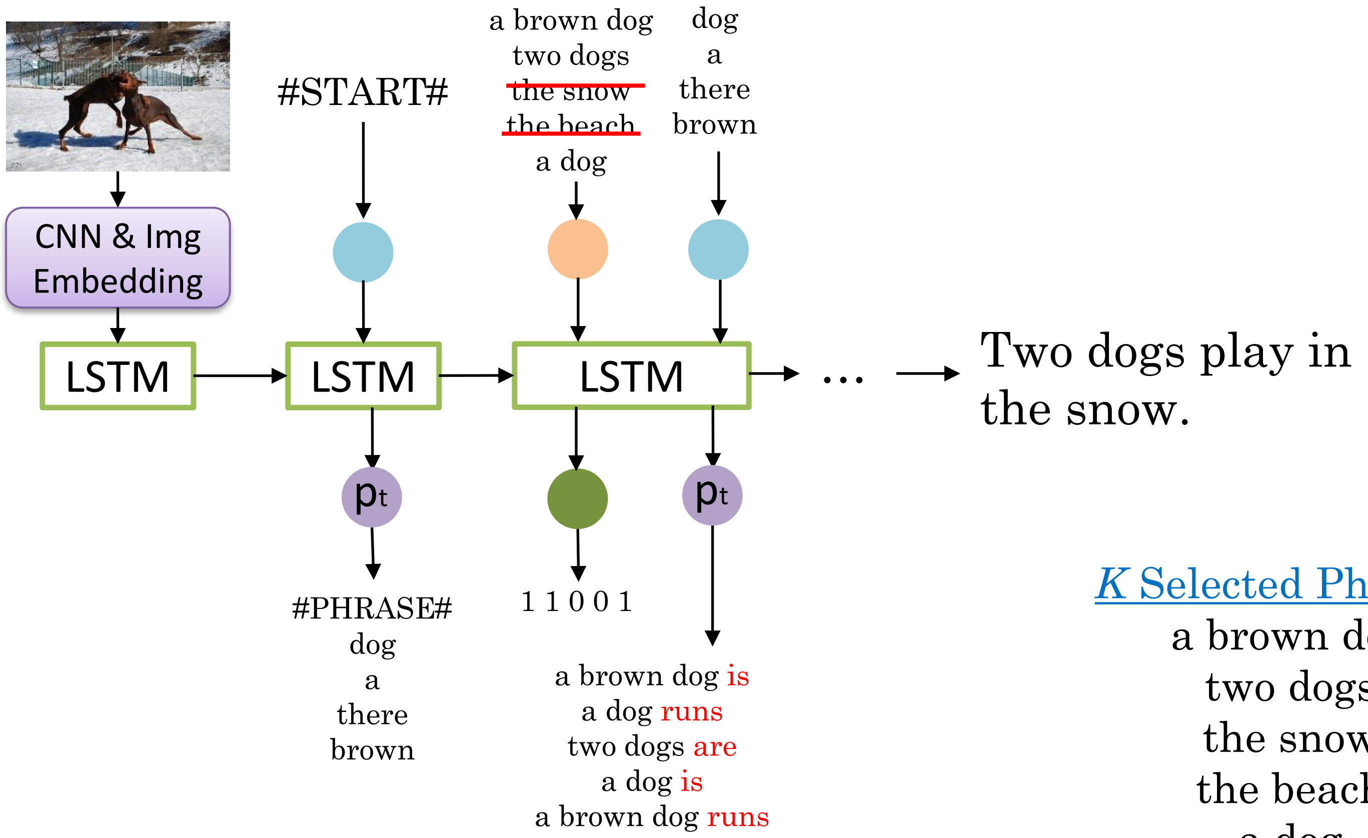
two dogs

the snow

the beach

a dog

Graphical Illustration: Sentence generation (Sentence Level)



K Selected Phrases:
a brown dog
two dogs
the snow
the beach
a dog

Experiment

- Tested on Flickr8k and Flickr30k datasets.
- Each image is annotated with five descriptions by human.
- 1k of images are used for validation and another 1k of images are used for testing, while the rest are for training (consistent with state-of-the-art).

- *A woman in a red coat with a man in a white and black coat and a black dog in the snow.*
- *Two people and a dog are in the snow.*
- *Two people are interacting with a dog that has bitten an object one of them is holding.*
- *Two people are walking up a snowy hill with a dog.*
- *Two people playing on a snowy hill.*

Qualitative results (Phrase)

- Phrase generation:

a person
a man
the air
a dirt bike
a bike
a motorcycle
his bike
a bicycle
a helmet
the dirt

a little girl
a girl
a young girl
a child
a woman
the camera
a boy
the girl
a baby
a small child

the water
two dogs
the ocean
a dog
the beach
a man
a brown dog
three dogs
two people
a black dog

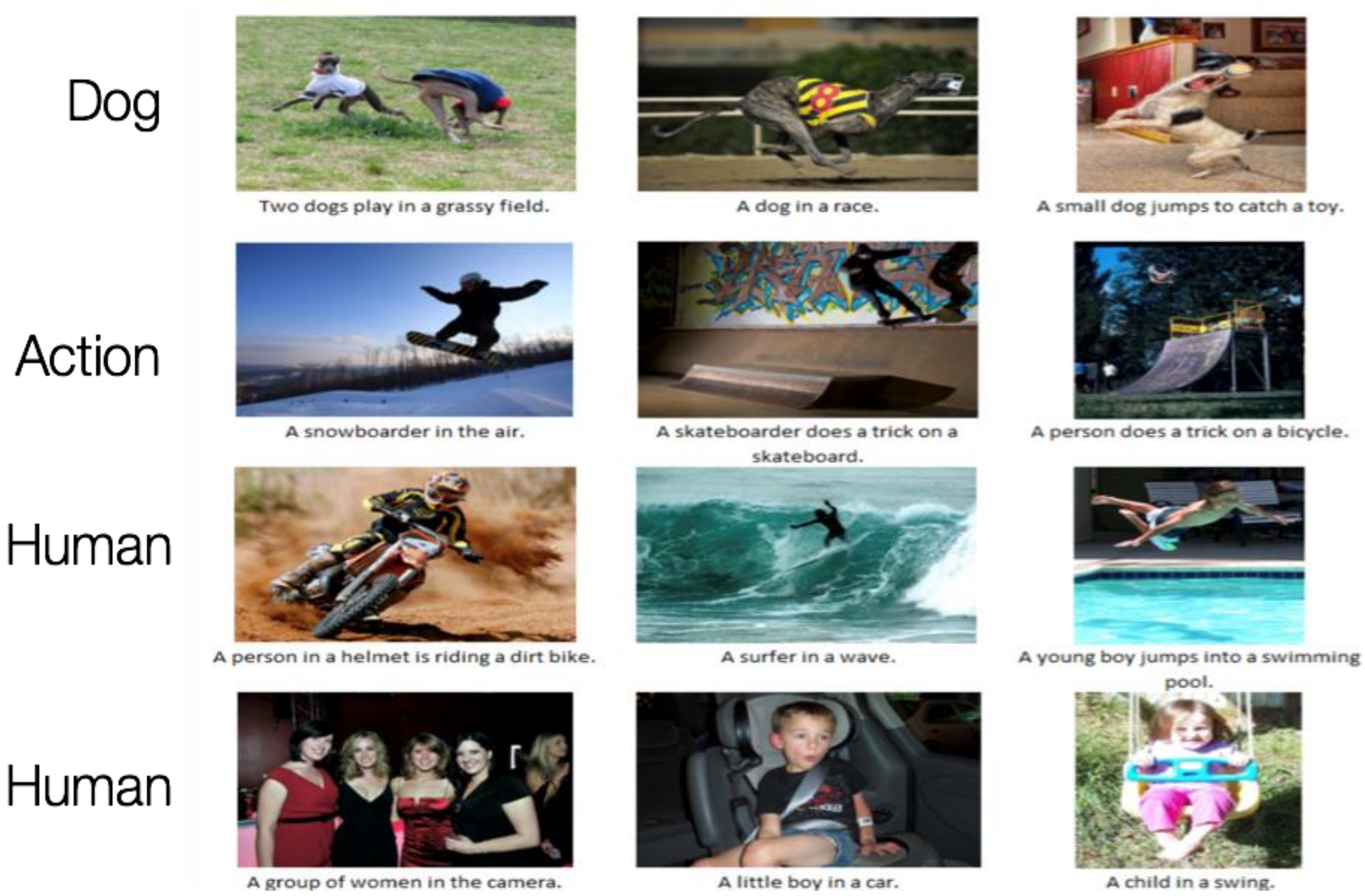
a group of people
a group of children
a crowd
a man
the air
the background
a building
several people
three people
the street

QUALITATIVE RESULTS (SENTENCE)

Image:			
NIC: (baseline)	A skateboarder does a trick on a ramp.	A man on a snowy mountain.	A surfer rides a wave.
phi-LSTM (proposed)	A man doing a trick on a bike.	A person in the snow.	A person in the water.
Reference: (human)	A skateboarder on a ramp.	A man crouched on a snowy peak.	A surfer does a flip on a wave.

Image:			
NIC: (baseline)	A group of people are standing in front of a building.	A man is doing a trick on a skateboard.	Two dogs play in the grass.
phi-LSTM (proposed)	Three people are standing in front of three men.	A skateboarder does a trick on a ramp.	Three dogs play in a grassy field.
Reference: (human)	A group of tourists stand around as a lady puts her hand near the mouth of a statue.	A skateboarder in the air at a big outdoor ramp.	The three dogs ran in the yard.

More results (Sentences with same object(s))



NVIDIA

More results (Sentences with same scene)

Places


A group of people in the snow.


A woman in the snow.


A woman in the street.

Places


A group of people in a field.


A person is riding a dirt bike.


A man is riding a bike.

Places


A girl in the water.


A man in the water.


A surfer in the water.

Qualitative Results (Poor examples)

A man in a boat in the water.

A child in a slide.

A woman is holding a young boy.

A woman and a child are sitting in a baby.

A woman in a man in a kitchen.

A man is holding a woman.

UNIVERSITY OF MALAYA
The Leader in Research & Innovation

NVIDIA

Quantitative Results

- Evaluation metric: BLEU
- Measure n-grams precision quality between generated caption and reference sentences (human).

Flickr8k				
Models	B-1	B-2	B-3	B-4
DeepVs [4]	57.9	38.3	24.5	16.0
NIC [3] [3]	60.2(63)	40.4	25.9	16.5
phi-LSTM	**63.6**	**43.6**	**27.6**	**16.6**

Our proposed model

Flickr30k				
Models	B-1	B-2	B-3	B-4
DeepVS [4]	57.3	36.9	24.0	15.7
mRNN [2]	60	41	28	**19**
NIC [3] [4]	66.3(66)	42.3	27.7	18.3
LRCNN [6]	58.7	39.1	25.1	16.5
PbIC [30]	59	35	20	12
phi-LSTM	**66.6**	**45.8**	**28.2**	17.0

Our proposed model

More Analysis by comparing with baseline

- Given same amount of training data, and same set of test image, and same set of setting in training:
  - Our model can generate sentence formed with more variety of words in the training corpus.

- What is the minimum time a word should appears in training data, so the model can generate sentence using that word?
  - Our model (phi-LSTM) = 81
  - Baseline (NIC) = 93

Conclusion

- Proposed of hierarchical phrase-based LSTM model to generate image description.
- Hierarchical model vs pure sequential model:
  - Able to generate better description
  - Can learn with less data
- Published in ACCV 2016, extension to journal.
- Future works
  - Experiments on MSCOCO dataset
  - Evaluation on more types of automatic evaluation metrics such as ROUGE, METEOR, CIDEr
  - Apply on image sentence bi-directional retrieval
  - Tackle problem in poor results

References

1. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: ECCV 2010.
2. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: CVPR 2011.
3. Yang, Y., Teo, C.L., Daum´e III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: EMNLP 2011.
4. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daum´e III, H.: Midge: Generating image descriptions from computer vision detections. In: EACL 2012.
5. Kuznetsova, P., Ordonez, V., Berg, T.L., Choi, Y.: Treetalk: Composition and compression of trees for image descriptions. TACL 2014.

References

6. Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: CONLL 2011.
7. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: ACL 2012.
8. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: ICLR 2015.
9. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR 2015.
10. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR 2015.

References

11. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
12. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR 2015.
13. Lebret, R., Pinheiro, P.O., Collobert, R.: Phrase-based image captioning. ICML 2015
14. Yngve, V.: A model and an hypothesis for language structure. Proceedings of the American Philosophical Society **104** (1960) 444–466

THE END
Q & A?

Chee Seng Chan PhD SMIEEE
University of Malaya, Malaysia
www.cs-chan.com

Full Paper: Tan, Y. H., & Chan, C. S. (2016, November). phi-lstm: A phrase-based hierarchical LSTM model for image captioning. In *Asian Conference on Computer Vision* (ACCV) , pp. 101-117.