phi-LSTM: A Phrase-based Hierarchical LSTM Model for Image Captioning

Ying Hua Tan and Chee Seng Chan
Faculty of Computer Science & Information Technology,
University of Malaya, MALAYSIA

Motivations

◊ Conventional treats sentence as sequence of words, and disregard all other linguistic syntax and structure a sentence should have.

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

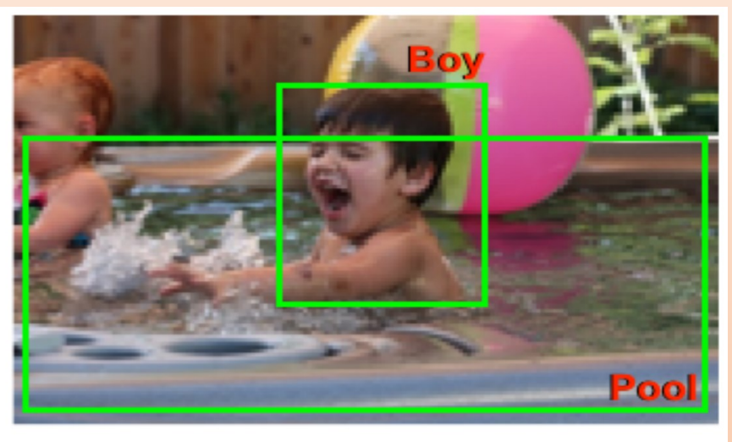
— Prof. Victor Yngve

◊ **Question:** Given the importance of sentence structure, how would it affect a language model that generates image caption if the sentence is encoded in a structural manner?

Objectives

1) Design a phrase-based model for image captioning.
2) Investigate on its performance as compared to a pure sequence model.

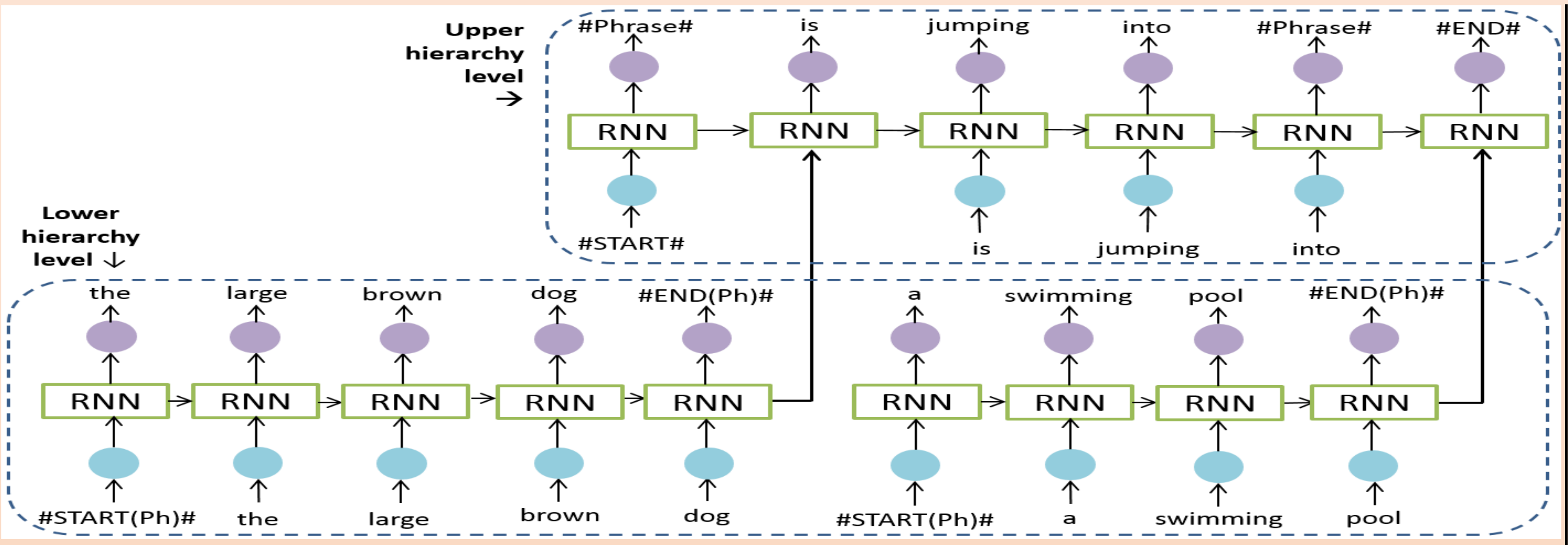
Image:



Our proposed: *A little boy is playing in the pool.*

Groundtruth: A boy with a beach ball behind him playing in a pool.

Proposed phi-LSTM



Model Architecture: a) Lower Hierarchy (Phrase); b) Upper Hierarchy (Sentence) Levels

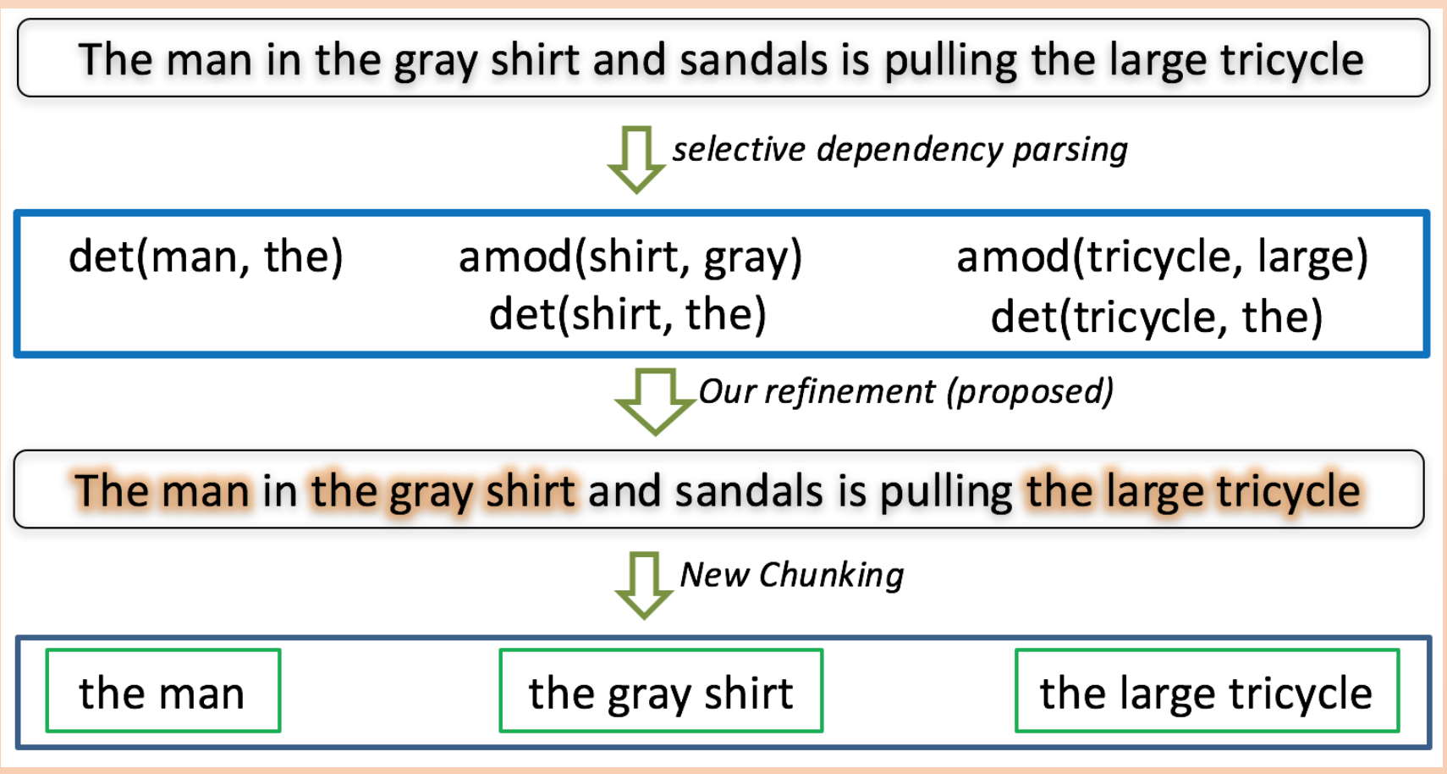
Step 1: Phrase Chunking:

◊ Characteristic of image descriptions:

- Consists of mostly noun phrases (NP), linked with verb and prepositional phrases.
- Each NP is strongly image relevant.
- Each NP has similar syntactic role.

◊ Partitioning the learning of NP and sentence structure

◊ Dependency parsing (**Stanford CoreNLP tool**)



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

selective dependency parsing

det(man, the) amod(shirt, gray) det(shirt, the) amod(tricycle, large) det(tricycle, the)

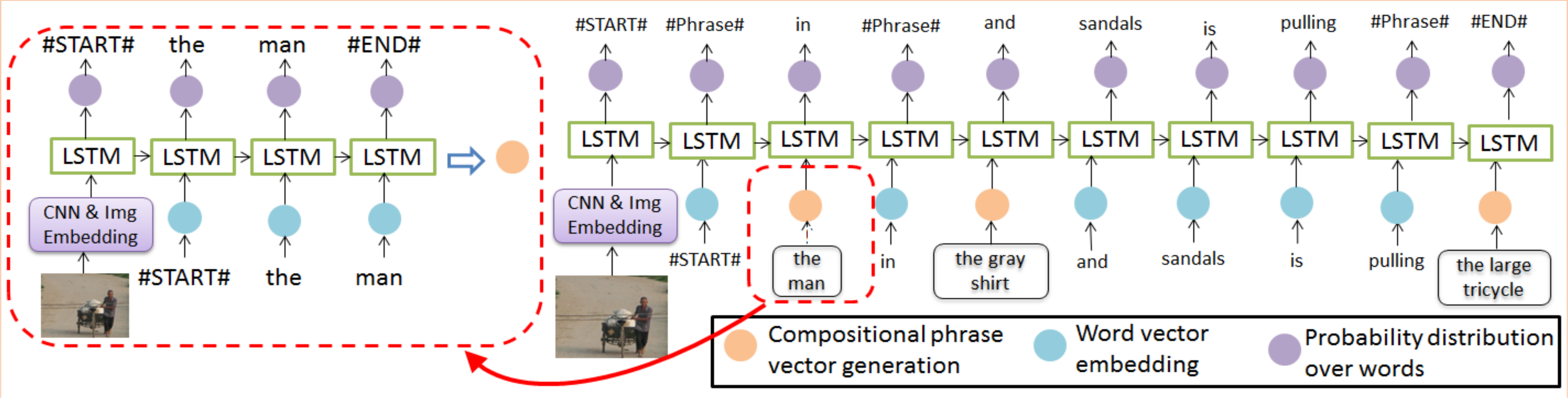
Our refinement (proposed)

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

New Chunking

the man | the gray shirt | the large tricycle

Step 2: Encoding of Phrase and Sentence:



◊ Sentence = sequence of noun phrases and words.

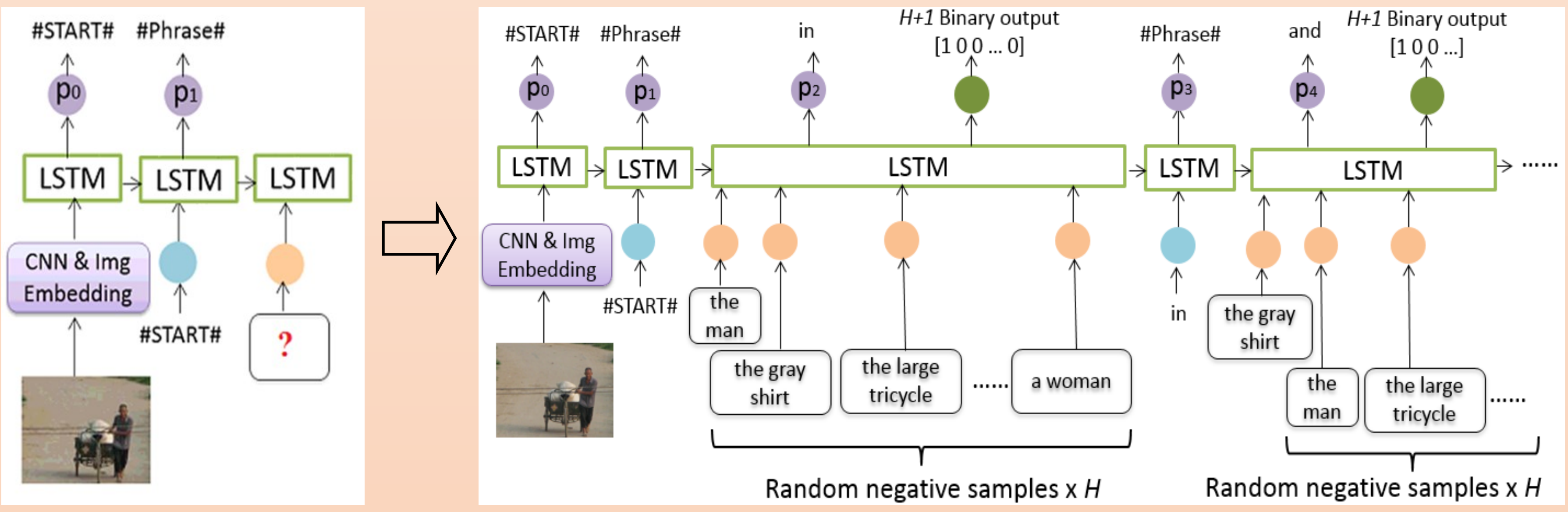
◊ A '*phrase*' token is added into the corpus

Step 3: Phrase Selection Objective:

◊ **Decoding stage**: generate phrases ⟶ generate full sentence

◊ All NP = a '*phrase*' token (decoding sentence)

◊ Which NP = the input of next time step?



◊ **Phrase selection objective** ⟶ train the model for recognizing probable NP inputs

Objective Function:

◊ Overall objective function:

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

$$L = M \times \sum_{j=1}^{M} N_j$$

◊ Perplexity of each sentence:

$$\log_2 \mathcal{PPL}(\mathbf{S}|\mathbf{I}) = -\frac{1}{N}\left[\underbrace{\sum_{t_s=-1}^{Q} \log_2 \mathbf{p}_{\mathbf{t}_s}}_{\text{each word in a sentence}} + \overbrace{\sum_{i=1}^{R}\left[\underbrace{\sum_{t_p=-1}^{P_i} \log_2 \mathbf{p}_{\mathbf{t}_p}}_{\text{each word in a phrase}}\right]}^{\text{each phrase in a sentence}}\right]$$

$$N = Q + \sum_{i=1}^{R} P_i$$

◊ Phrase selection objective:

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

Other Settings:

◊ **CNN model:** VGG-16 pre-trained on ImageNet

◊ **LSTM parameters:** different for phrase and sentence level, with dropout

◊ **Word embedding parameters:** same for both levels

◊ **Words discarded:** occurrence < 5 times (Flickr8k) / 8 times (Flickr30k)

◊ **Optimizer**: RMSprop (minibatch size = 100)

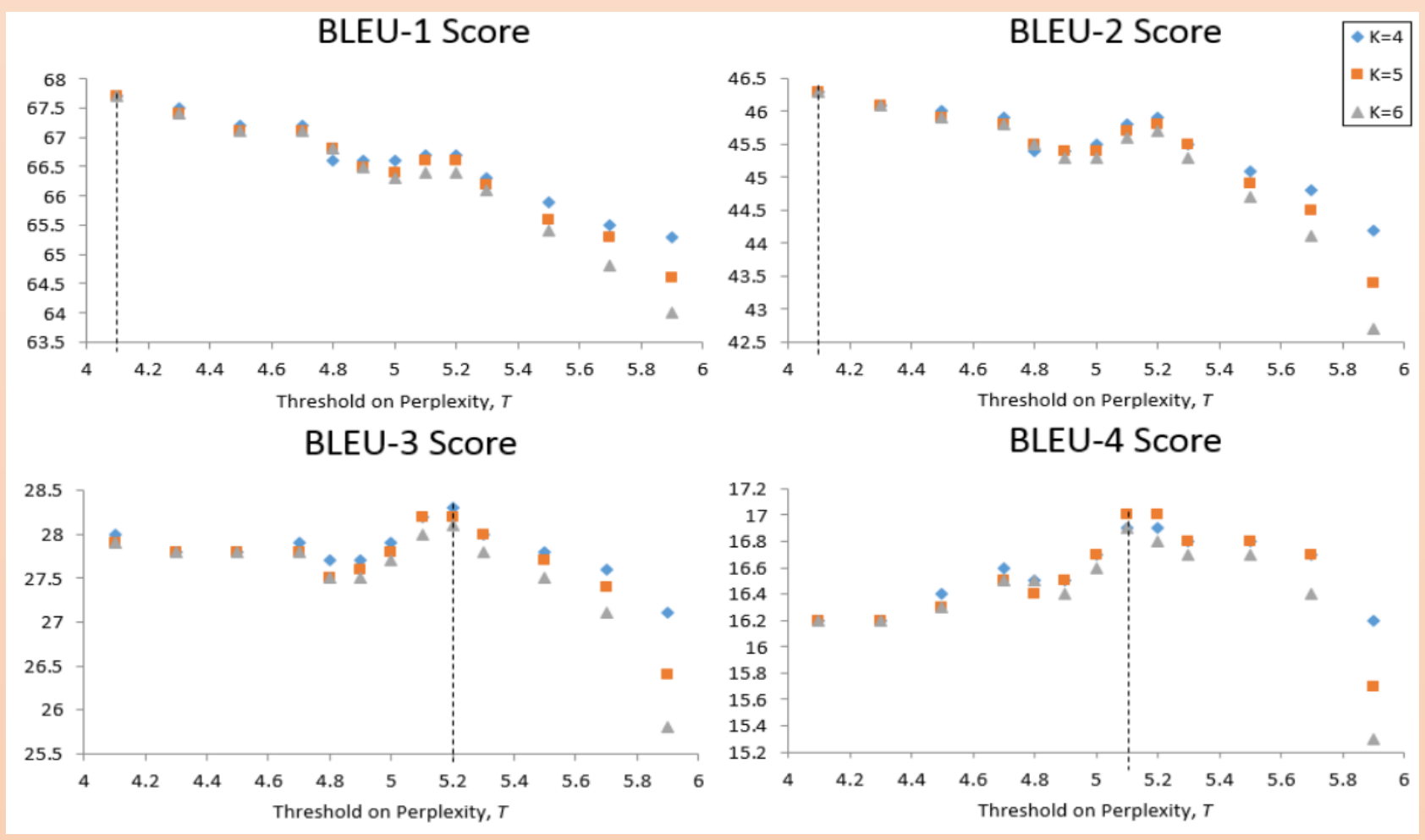
Results

Quantitative results (BLEU):

Flickr8k				
Models	B-1	B-2	B-3	B-4
NIC (CVPR'15)	60.2(63)	40.4	25.9	16.5
DeepVs (CVPR'15)	57.9	38.3	24.5	16.0
phi-LSTM	**63.6**	**43.6**	**27.6**	**16.6**

Flickr30k				
Models	B-1	B-2	B-3	B-4
mRNN (ICLR'15)	60	41	28	**19**
NIC (CVPR'15)	66.3(66)	42.3	27.7	18.3
DeepVs (CVPR'15)	57.3	36.9	24.0	15.7
LRCNN (CVPR'15)	58.7	39.1	25.1	16.5
PbIC (ICML'15)	59	35	20	12
phi-LSTM	**66.6**	**45.8**	**28.2**	17.0

BLEU score variation



◊ *T* = Perplexity threshold of a phrase

◊ *K* = Maximum number of phrases per sentence

Analysis on corpus (Flickr8k):

	Train Data		Test Data				Gen. Caption	
Number of sentence	30000		5000		1000		1000	
	Actual	Trained	Actual	Trained	Actual	Trained	NIC	phi-LSTM
Size of vocab	7371	2538	3147	1919	1507	1187	128	154
Number of words	324481	316423	54335	52683	11139	10806	8275	6750
Avg. caption length	10.8	10.5	10.9	10.5	11.1	10.8	8.3	6.8

◊ phi-LSTM is able to generate sentence formed with more variety of words.

NIC (CVPR'15)		phi-LSTM	
Word	Occurrence	Word	Occurrence
obstacle	93	*overlooking*	81
surfer	127	*obstacle*	93
bird	148	*climber*	96
woods	155	*course*	106
snowboarder	166	*surfer*	127

Top 5 least trained words inferred

NIC (CVPR'15)		phi-LSTM	
Word	Occurrence	Word	Occurrence
to	2306	*while*	1443
his	1711	*green*	931
while	1443	*by*	904
three	1052	*one*	876
small	940	*another*	713

Top 5 most trained words absent

Phrases generated:

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

◊ Red fonts = phrase with perplexity value <*T*

Sentence generated:

Images:								
phi-LSTM:	**Three people are standing in front of three men.**	**A skateboarder does a trick on a ramp.**	**Three dogs play in a grassy field.**	**A cowboy is riding a horse.**	**A person in a helmet is riding a dirt bike.**	**A man doing a trick on a bike.**	**A person in the snow.**	**A little girl in a red jacket is standing in the snow.**
NIC:	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.	A man is riding a horse.	A man on a dirt bike.	A skateboarder does a trick on a ramp.	A man on a snowy mountain.	A little boy in a red jacket is in the snow.
Groundtruth:	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.	A blond cowboy is riding a bucking bronco at the rodeo.	A dirt biker turns across the dirt.	A skateboarder on a ramp.	A man crouched on a snowy peak.	A child dressed for the cold sits in the snow.