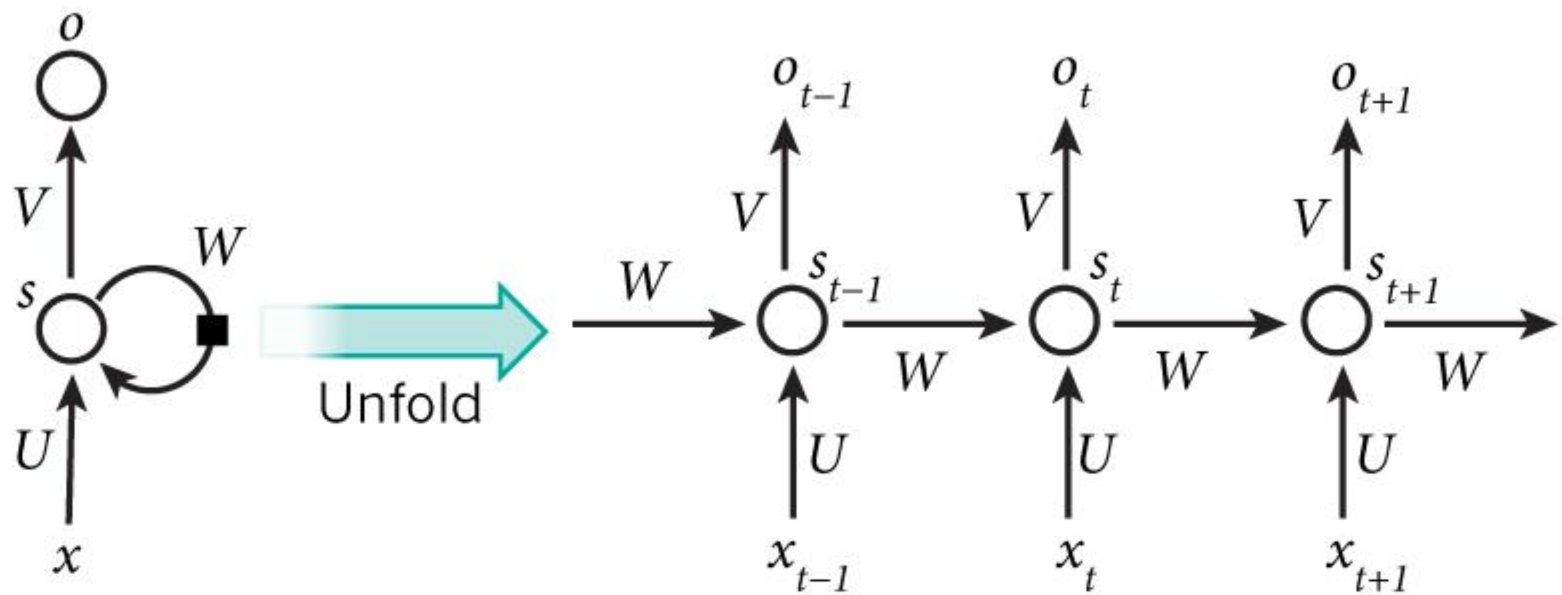
Learning Sequences: image caption with region-based attention and scene factorization

Changshui Zhang， Junqi Jin, Kun Fu
and Fei Sha
zcs@mail.Tsinghua.edu.cn
Shenzhen, 2016,8

Modeling Sequences

- Autoregressive models(AR)
- Linear Dynamical Systems
- Hidden Markov Models(HMM)
- Recurrent Neural Networks(RNN)
- Long Short Term Memory (LSTM)

Recurrent Neural Networks(RNN)



Long Short Term Memory (LSTM)

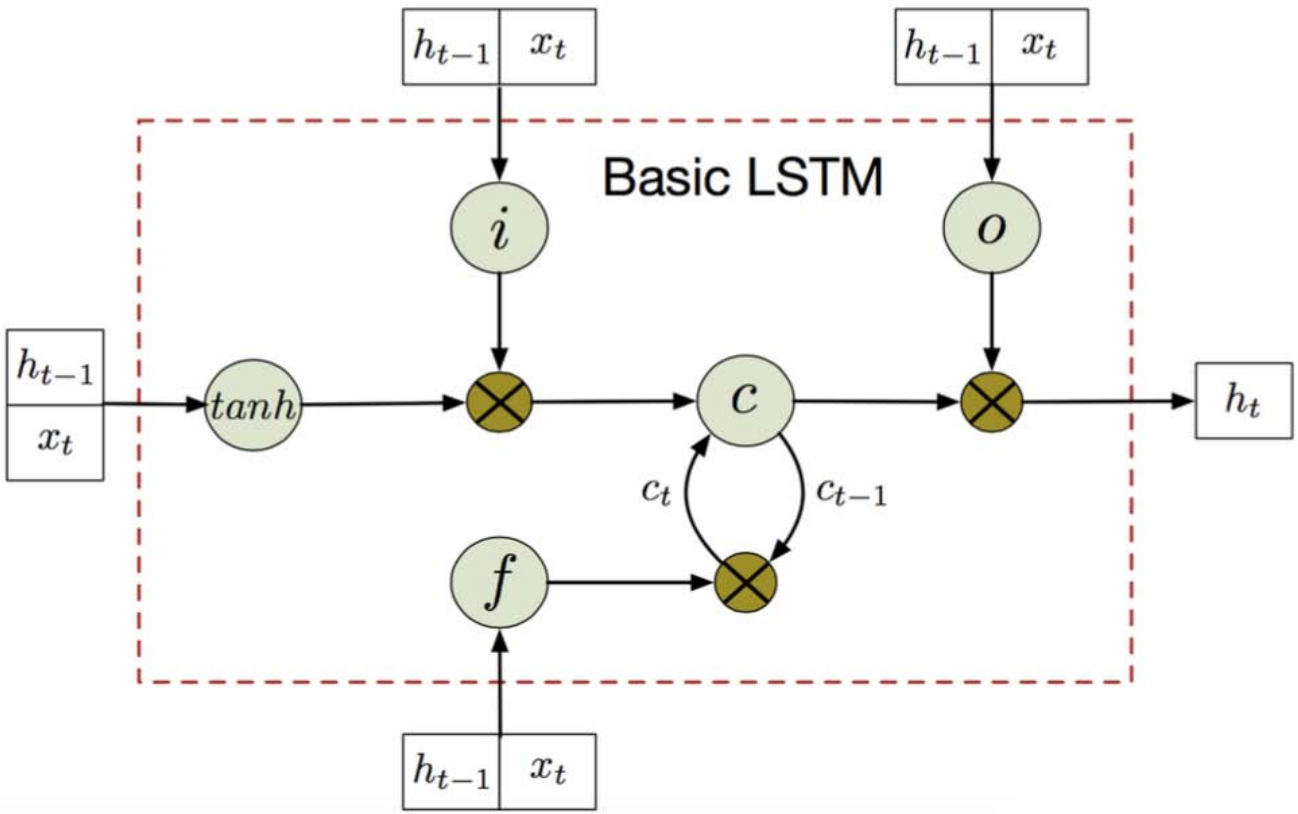


Image Caption

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Short History

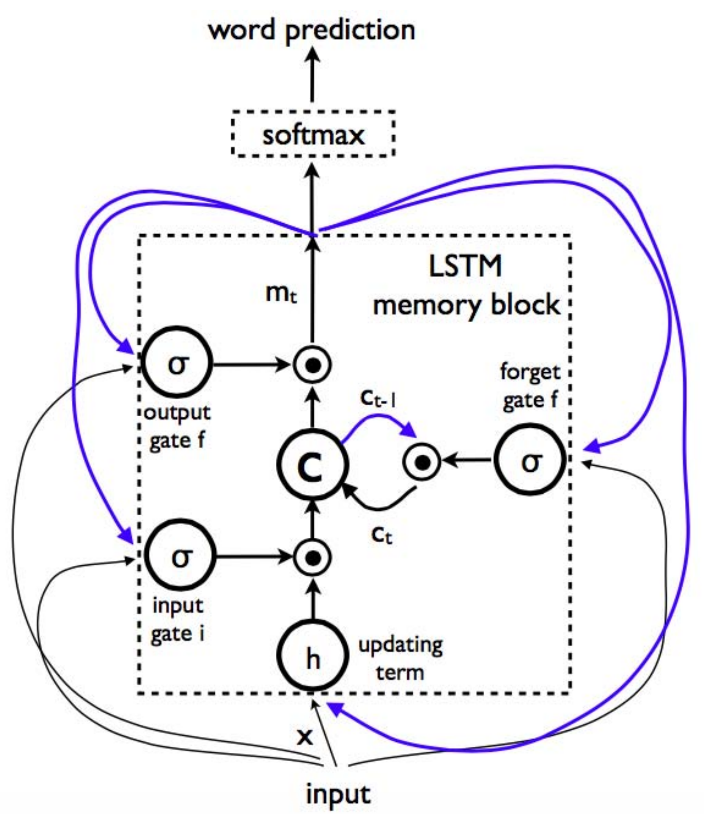
Farhadi et al., Every picture tells a story: Generating sentences from images. ECCV, 2010.
Ordonez et al., Im2text: Describing images using 1 million captioned photographs. NIPS, 2011.
Yang et al., Corpus-guided sentence generation of natural images. EMNLP, 2011.
Kulkarni et al., Baby talk: Understanding and generating simple image descriptions. CVPR, 2011.
Mitchell et al., Midge: Generating Image Descriptions From Computer Vision Detections. EACL, 2012.

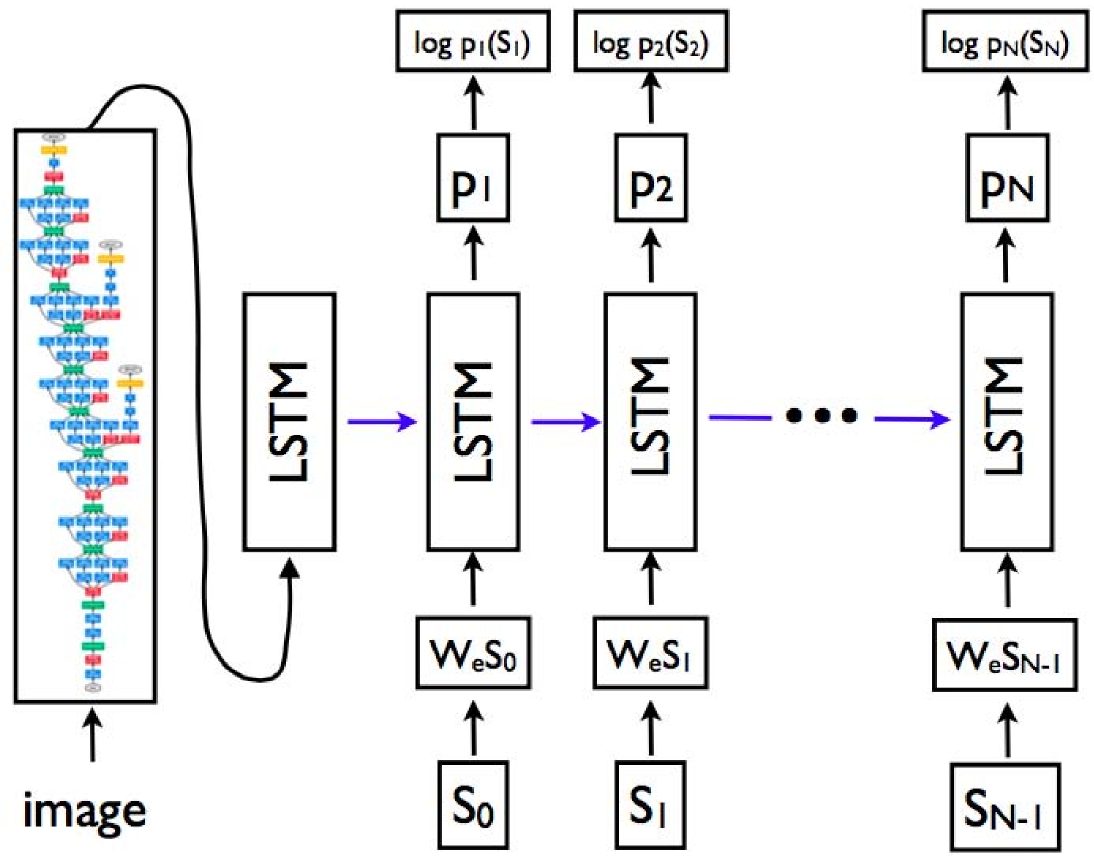
Related Systems

- Samy Bengio in Google
- Junhua Mao, Alan L. Yuille in UCLA
- Mitchell in Microsoft
- Li FeiFei in Stanford
- Yoshua Bengio in Montreal
- Trevor Darrell in UC Berkeley

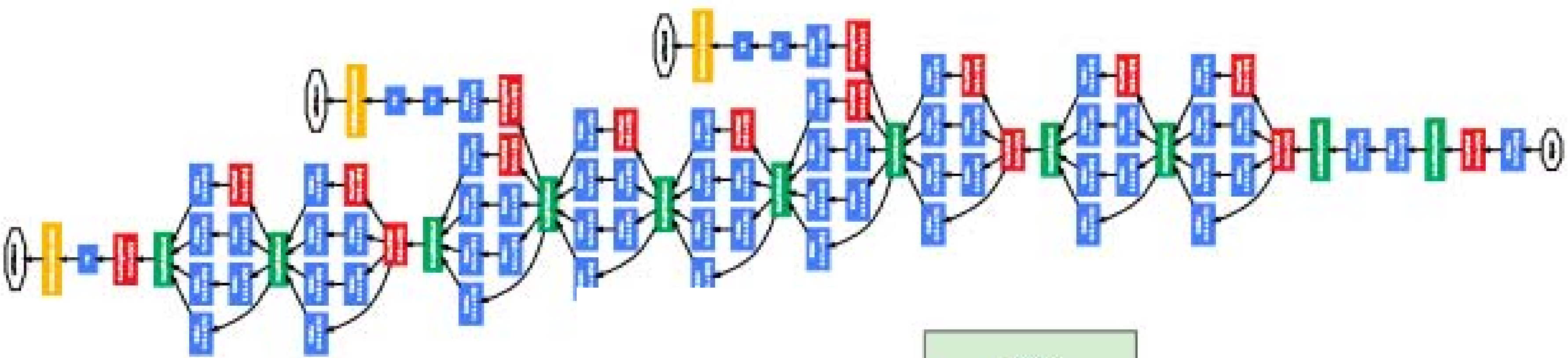
- From 2014.11, most papers from arxiv.org

Show and Tell: A Neural Image Caption Generator - Google

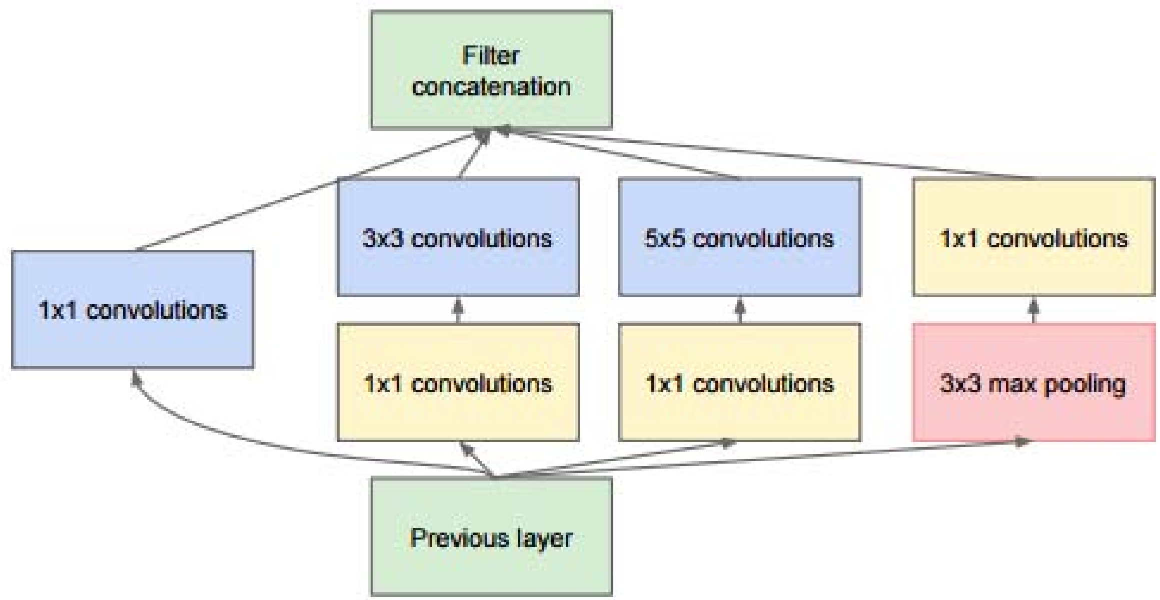




Team name	Entry description	**Classification error**	Localization error
GoogLeNet	No localization. Top5 val score is 6.66% error.	0.06656	0.606257
VGG	a combination of multiple ConvNets, including a net trained on images of different size (fusion weights learnt on the validation set); detected boxes were not updated	0.07325	0.256167

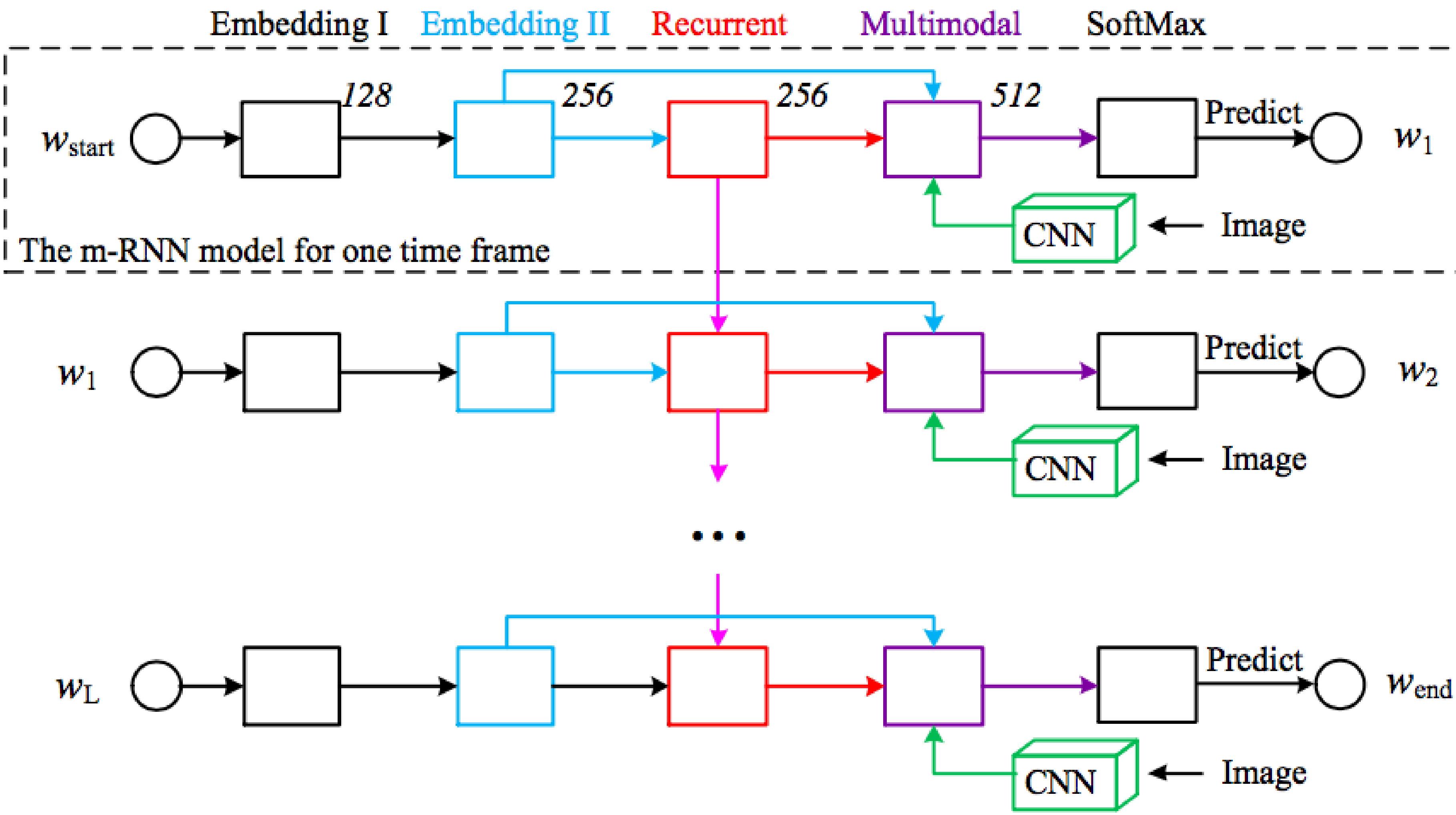


ILSVRC2014



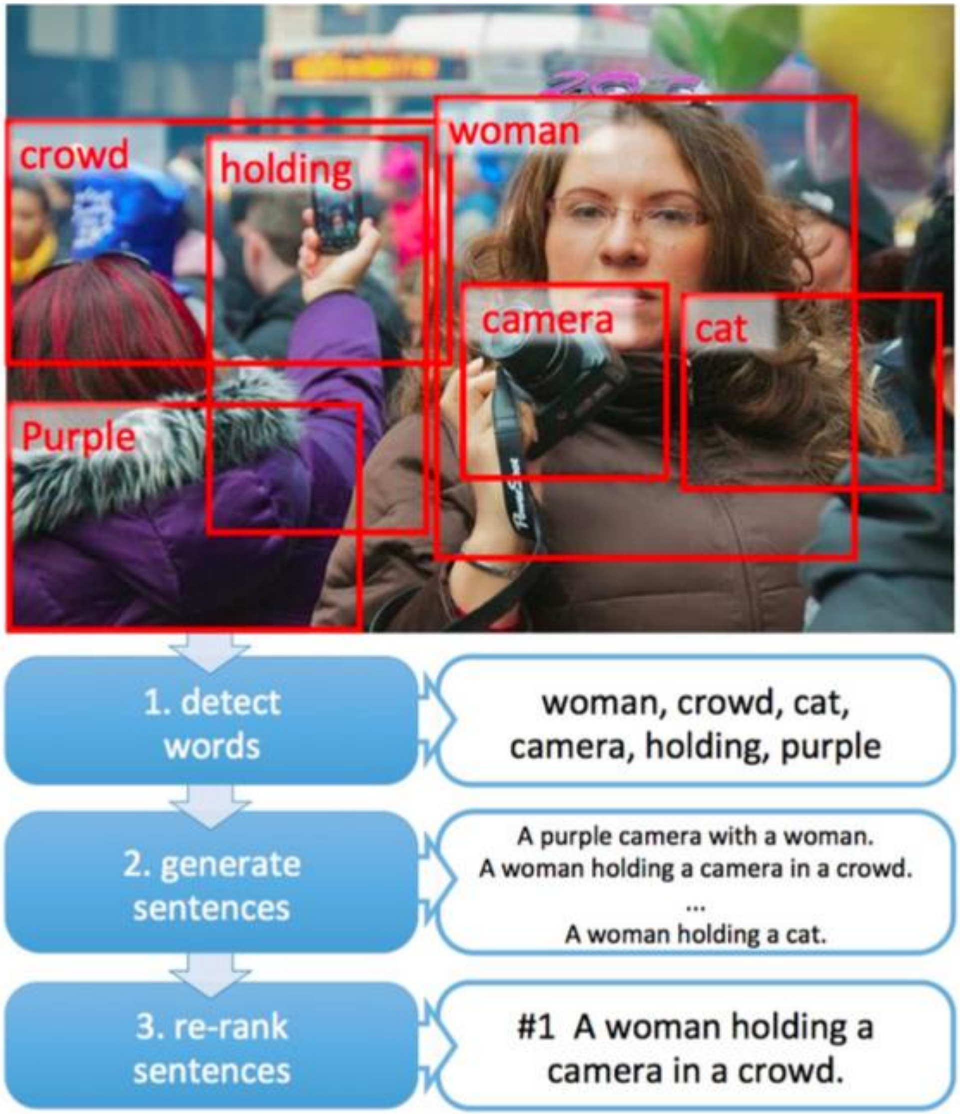
(b) Inception module with dimension reductions

DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)-Baidu



(b). The m-RNN model

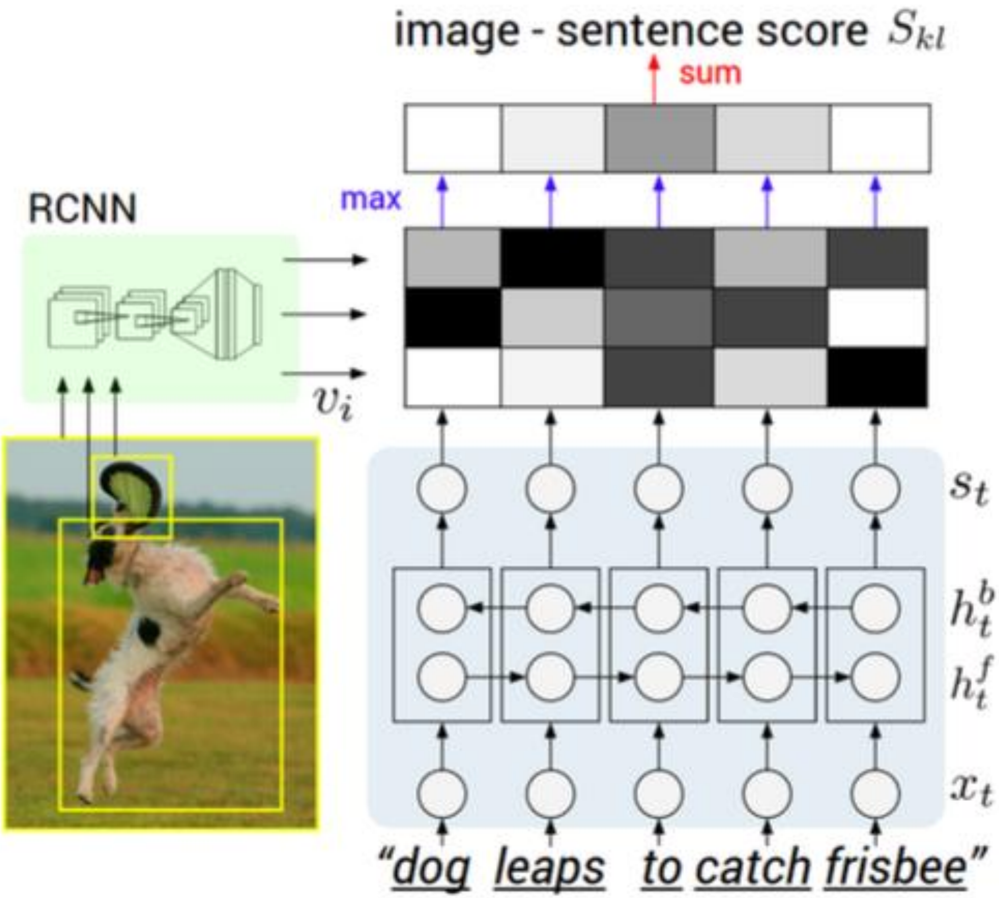
$$\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, \langle s \rangle, \tilde{\mathcal{V}}_{l-1}) = \frac{\exp\left[\sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, \langle s \rangle, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup \langle /s \rangle} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1}, \cdots, \bar{w}_1, \langle s \rangle, \tilde{\mathcal{V}}_{l-1})\right]}$$

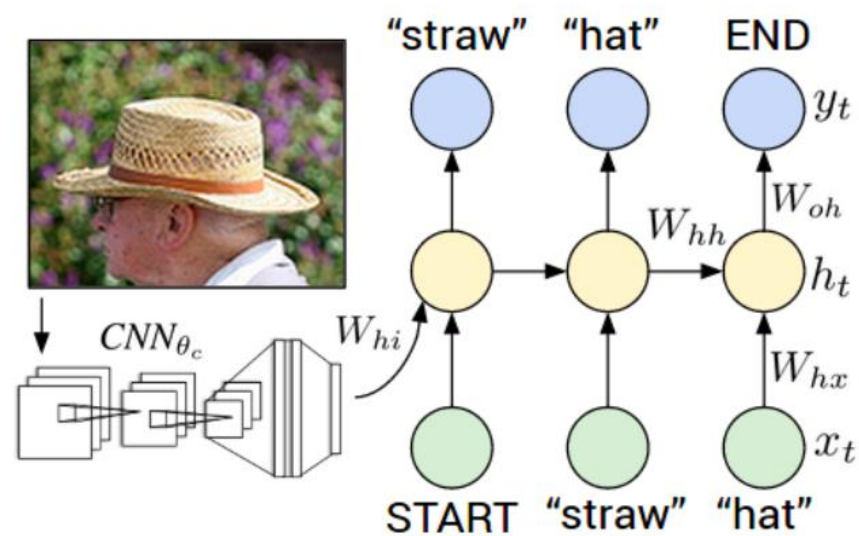


$$1 - \prod_{j \in b_i} \left(1 - p_{ij}^{w}\right)$$

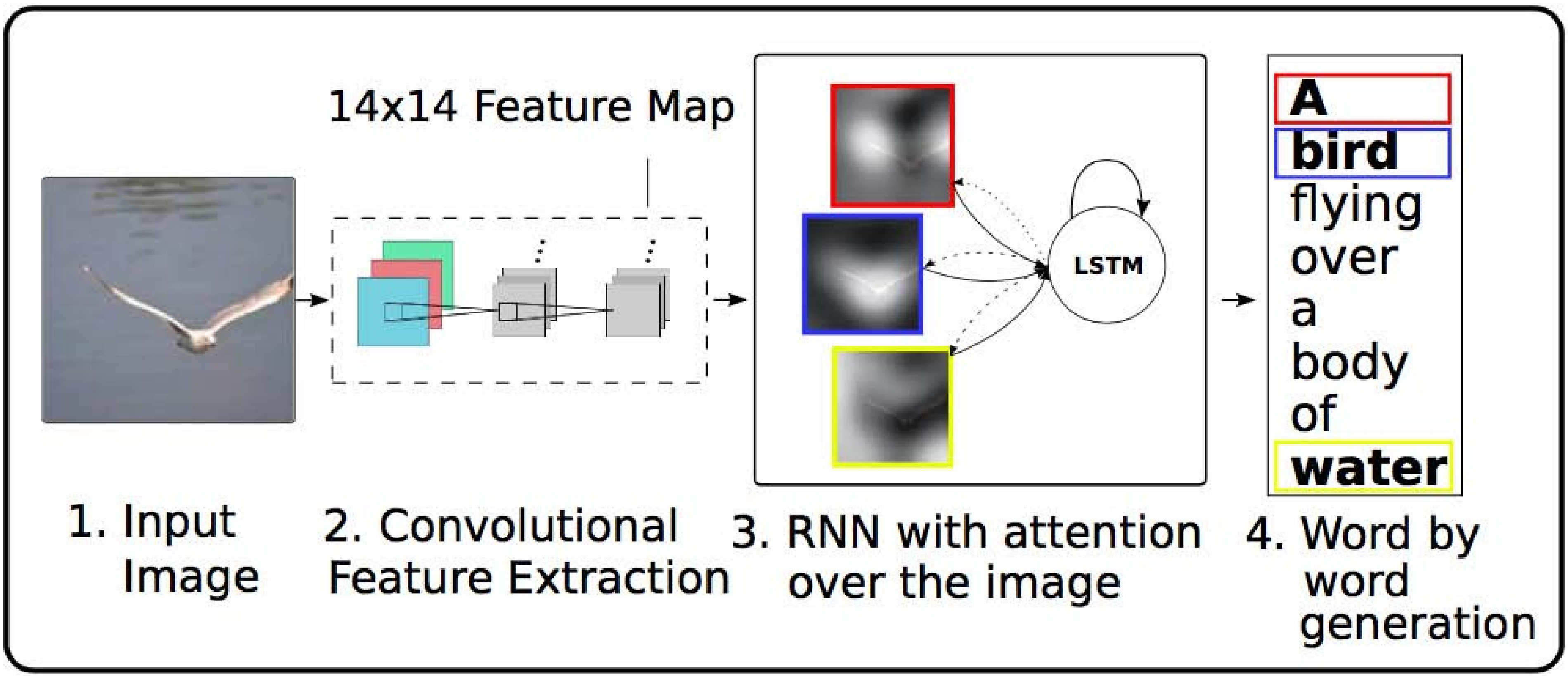
From Captions to Visual Concepts and Back – Microsoft

Deep Visual-Semantic Alignments for Generating Image Descriptions-Stanford

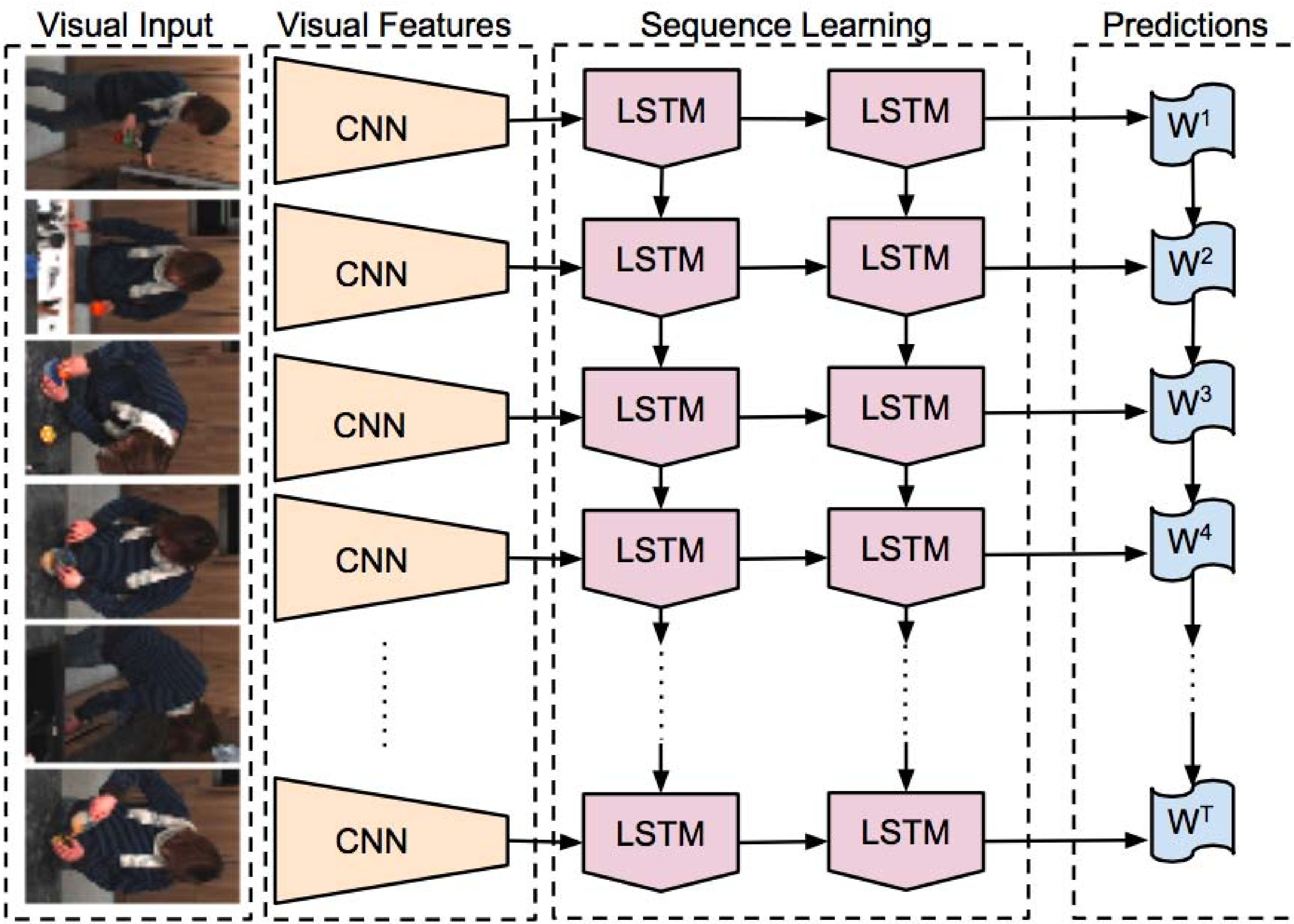




Show, Attend and Tell: Neural Image Caption Generation with Visual Attention-Montreal



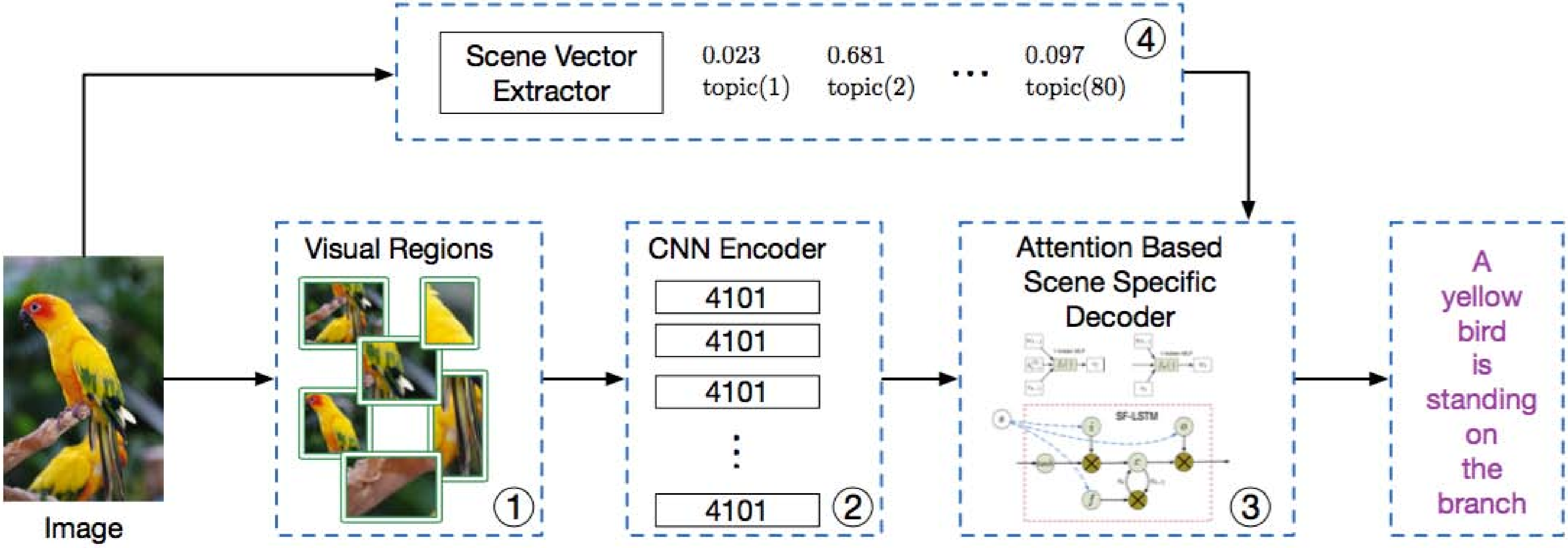
Long-term Recurrent Convolutional Networks for Visual Recognition and Description- UC Berkeley



Our work

Framework

- Image representation with localized patches at multiple scales (Region-based attention)
- Attention-based multi-modal LSTM decoder
- Scene factored LSTM



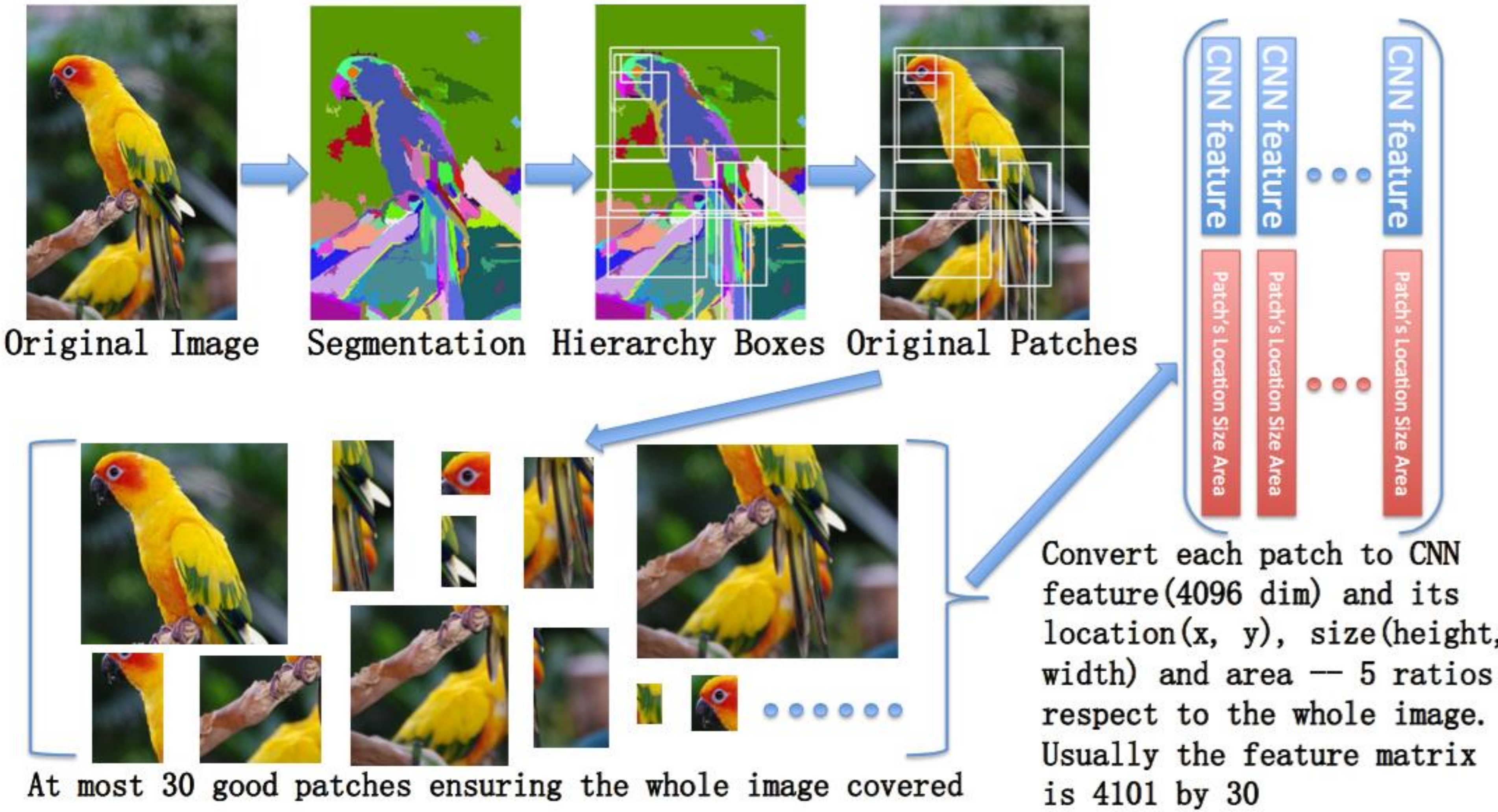
Framework 1: Region-based Attention

- Attention transits from regions to regions, being aligned with the transition of text meaning.
- A good region should be:

  1. semantically meaningful (high level concepts)
  2. primitive and non-compositional (single concept)
  3. contextually rich (interaction)
- Selective search fits the above

Framework 1: Selective Search

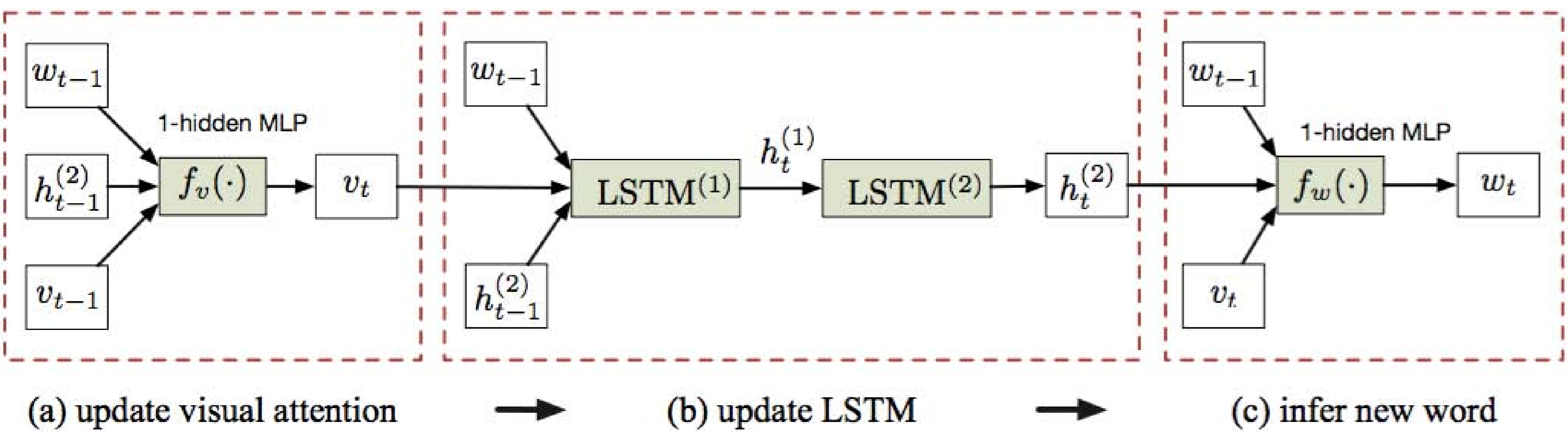
- Localized regions at multi scales



Framework 2: LSTM Decoder

- Three stages in one time step:

  1. Capture the visual attention transition
  2. Use an abstract meaning to bridge image and text
  3. Infer the new word based on attention and meaning



Framework 2-1: Attention Update

- Represent the feature vectors of regions:

$$\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{\mathsf{R}}\}$$

- Compute the score of region i in time t:

$$p_{it} \propto \exp\{f_v(\boldsymbol{r}_i, \boldsymbol{P_w w}_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{v}_{t-1})\}, \forall\ i = 1, 2, \cdots, \mathsf{R}$$

- Sum up according to the score:

$$\boldsymbol{v}_t = \sum_i p_{it} \boldsymbol{r}_i$$

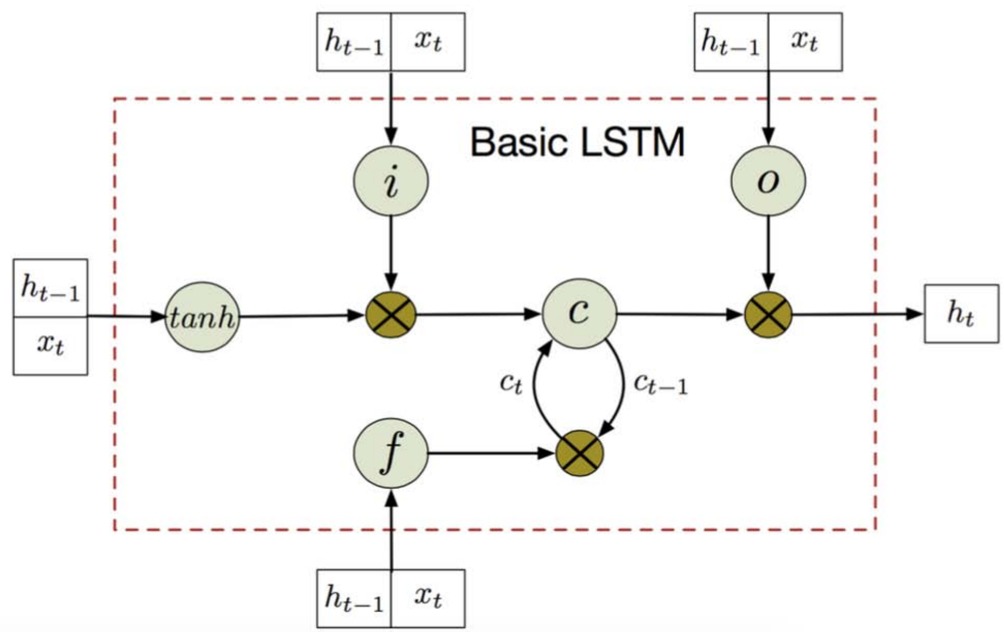
Framework 2-2: LSTM

- 2-layers LSTM is used

$$\begin{pmatrix} \boldsymbol{i}_t^{(1)} \\ \boldsymbol{f}_t^{(1)} \\ \boldsymbol{o}_t^{(1)} \\ \boldsymbol{g}_t^{(1)} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{T}^{(1)} \begin{pmatrix} \boldsymbol{P_w w}_{t-1} \\ \boldsymbol{h}_{t-1}^{(1)} \\ \boldsymbol{h}_{t-1}^{(2)} \\ \boldsymbol{v}_t \end{pmatrix} \quad \begin{pmatrix} \boldsymbol{i}_t^{(2)} \\ \boldsymbol{f}_t^{(2)} \\ \boldsymbol{o}_t^{(2)} \\ \boldsymbol{g}_t^{(2)} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{T}^{(2)} \begin{pmatrix} \boldsymbol{h}_t^{(1)} \\ \boldsymbol{h}_{t-1}^{(2)} \end{pmatrix}$$

$$\boldsymbol{c}_t^{(1)} = \boldsymbol{f}_t^{(1)} \odot \boldsymbol{c}_{t-1}^{(1)} + \boldsymbol{i}_t^{(1)} \odot \boldsymbol{g}_t^{(1)}$$
$$\boldsymbol{h}_t^{(1)} = \boldsymbol{o}_t^{(1)} \odot \tanh(\boldsymbol{c}_t^{(1)})$$

$$\boldsymbol{c}_t^{(2)} = \boldsymbol{f}_t^{(2)} \odot \boldsymbol{c}_{t-1}^{(2)} + \boldsymbol{i}_t^{(2)} \odot \boldsymbol{g}_t^{(2)}$$
$$\boldsymbol{h}_t^{(2)} = \boldsymbol{o}_t^{(2)} \odot \tanh(\boldsymbol{c}_t^{(2)})$$



Framework 2-3: Word Inference

- Predict the word distribution

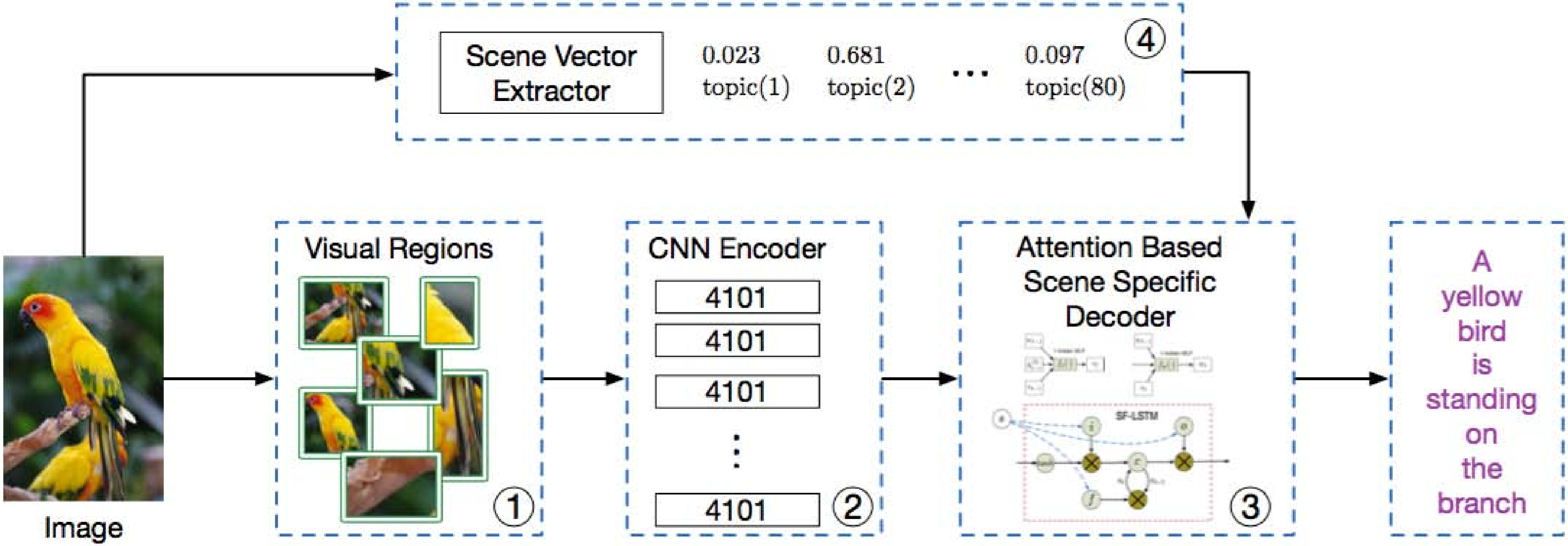
$$p_{wt} \propto \exp\{f_w(\boldsymbol{P}_{\boldsymbol{w}}\boldsymbol{w}_{t-1}, \boldsymbol{h}_t, \boldsymbol{v}_t)\}, \forall\ w = 1, 2, \cdots, \mathrm{W}$$

- Beam Search:

A pre-determined number of best-by-now sentences are computed and kept to be expanded with new words in the future.

Review of Framework

- CNN encoder + LSTM decoder
- Region-based attention + scene-factored LSTM



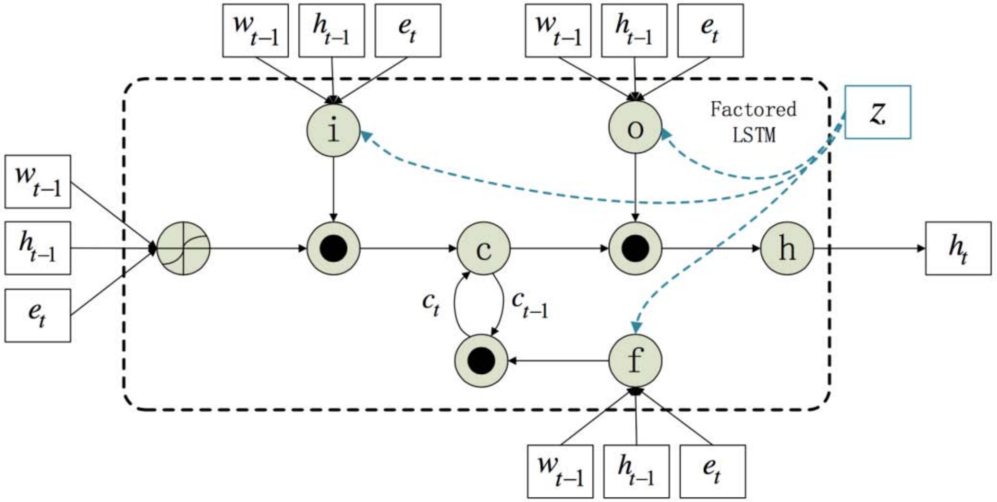
Scene-Factored LSTM

- Making LSTM adaptive to different scene, we factor the gates' weight matrix by the scene vector z
- A, F, B are shared among images, scene vector z is used to give image a specific language model.

$W_i^{(scene)}$ under scene vector z is denoted $W_i^{(z)}$

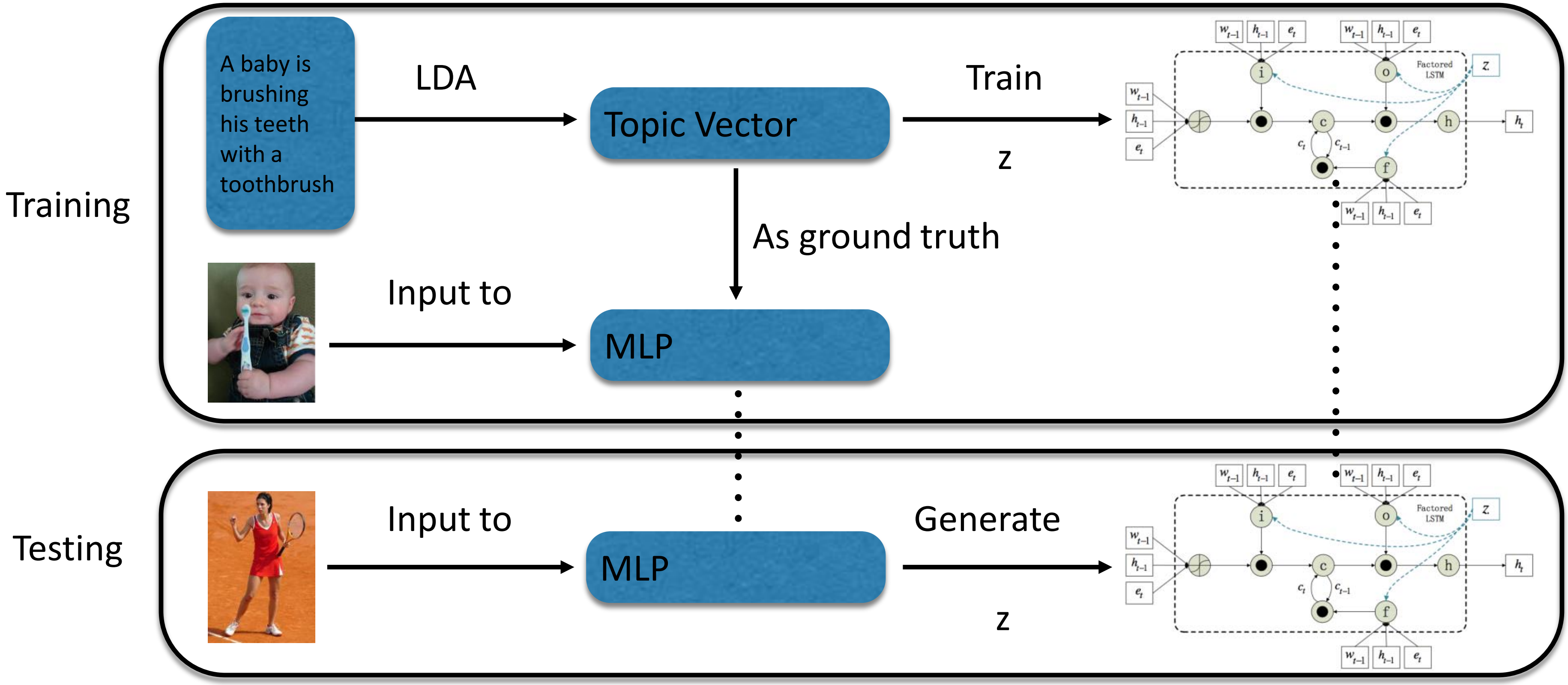
$$i_t = sigmoid(W_i^{(scene)}[w_{t-1}^T, h_{t-1}^T, e_t^T]^T)$$

$$W_i^{(z)} = W_i z \approx A_i diag(F_i z) B_i$$

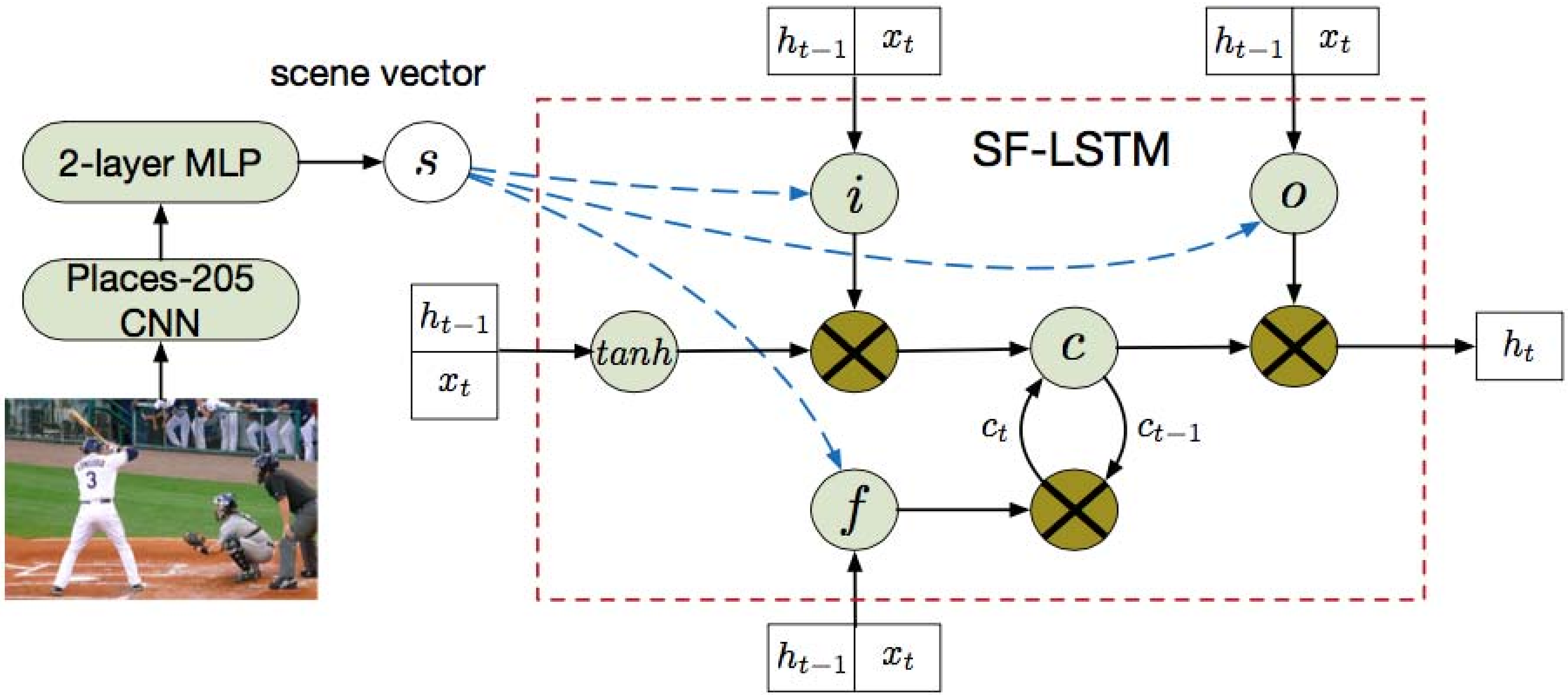


Obtaining Scene Vector

- For training images, use Latent Dirichlet Allocation (LDA).
- For testing images, using an MLP to predict scene vector.



Scene-Factored LSTM



$$\boldsymbol{W} = \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{F}\boldsymbol{s})\,\boldsymbol{B}$$

Scene-Factored LSTM

Training: Objective Function

- Negative log-likelihood of sentence given image

$$L = \frac{1}{N} \sum_{n} \log p \left(w_1^{(n)}, w_2^{(n)}, \ldots, w_{S_n}^{(n)} | I^{(n)} \right)$$

- Decomposed version:

$$L = \frac{1}{N} \sum_{n} \sum_{t=1}^{S_n} \log p \left(w_t^{(n)} | w_{0:t-1}^{(n)}, I^{(n)} \right)$$

Training: Optimization

- Optimizer: Adam (ICLR 2015)
- Adaptive effective step for each parameter, varying according to the variance
- Use the recommended hyper-parameters and they always work.
- One minibatch consists of sentences with the same length. Obtain 3x - 5x acceleration.

Training: Related Dataset

- MSCOCO: 82783, 40503, 40775 (5 captions)
- Flickr30K: 29000, 1000, 1000
- Flickr8K: 6000, 1000, 1000
- Places Database: 2.4M images, 205 categories

Data Set

- Describe all the important parts of the scene.
- Do not start the sentences with “There is”.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentences should contain at least 8 words.

MSCOCO

- Training 82783, validation 40504, test 40775 images. Each image has 5 captions.
- Evaluation, MSCOCO API
- Evaluation metrics: BLEU- 1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR and CIDEr-D

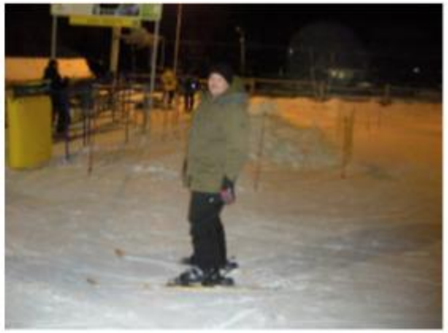
BLEU - k

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

$$P = \frac{m}{w_t} = \frac{7}{7} = 1 \qquad P = \frac{2}{7}$$

$$b(C,S) = \begin{cases} 1 & \text{if } l_C > l_S \\ e^{1-l_S/l_C} & \text{if } l_C \leq l_S \end{cases},$$

Generated Captions



a man riding skis down a snow covered slope

a train on a track near a train station

a bathroom with a toilet sink and mirror

a herd of zebra standing next to each other

a hot dog and french fries on a plate

a woman holding a nintendo wii game controller

a public transit bus on a city street

a black dog holding a frisbee in its mouth

a polar bear standing on top of a rock

Original Image
A
brown
cow
is
standing
in
the
grass

A bunch of

A herd of

Black Cat

Filled with

Fire Hydrant

Flying

Fries

Laying

Red

Sign

Evaluation

Table 1: Evaluation of various systems on the task of image captioning, on MSCOCO dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L	CIDEr-D
DeepVS[8]	62.5	45.0	32.1	23.0	19.5	–	66.0
LRCN[4]	62.8	44.2	30.4	21.0	–	–	–
Google NIC [24]	66.6	46.1	32.9	24.6	–	–	–
mRNN[16]	67	49	35	25	–	–	–
OUR-BASE-GREEDY	64.0	46.6	32.6	22.6	20.0	47.4	70.7
OUR-SF-GREEDY	67.8	49.4	34.8	24.2	21.8	49.1	74.3
OUR-RA-GREEDY	67.7	49.5	34.7	23.5	22.2	49.1	75.1
OUR-(RA+SF)-GREEDY	69.1	50.4	35.7	24.6	22.1	50.1	78.3
OUR-(RA+SF)-BEAM	**69.7**	**51.9**	**38.1**	**28.2**	**23.5**	**50.9**	**83.8**

Table 1: Evaluation of various systems on the task of image captioning

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L	CIDEr-D
Flickr8K							
DeepVS [3]	51	31	12	–	–	–	–
mRNN [8]	58	28	23	–	–	–	–
Google NIC [9]	63	41	27	–	–	–	–
OUR-(SF+RA)-BEAM	**66.5**	**47.8**	**33.2**	**22.4**	**20.8**	**48.6**	**56.5**
Flickr30K							
DeepVS [3]	50	30	15	–	–	–	–
LRCN [2]	59	39	25	16	–	–	–
mRNN [8]	60	41	28	19	–	–	–
Google NIC [9]	**67**	45	30	–	–	–	–
OUR-(SF+RA)-BEAM	**67.0**	**47.5**	**33.0**	**24.3**	**19.4**	**47.0**	**53.1**

Retrieval Task

Table 2: Evaluation with the tasks of image and captions retrieval on the MSCOCO dataset

	Caption − > Image				Image − > Caption			
	R@1	**R@5**	**R@10**	**Med** r	**R@1**	**R@5**	**R@10**	**Med** r
DeepVS	20.9	52.8	69.2	4.0	29.4	62.0	75.9	2.5
mRNN	29.0	42.2	77.0	3.0	**41.0**	**73.0**	**83.5**	**2.0**
OUR-RA+SF-BEAM	**29.3**	**62.8**	**77.2**	**2.0**	36.9	67.0	78.6	**2.0**

arXiv Paper

- http://arxiv.org/abs/1506.06272

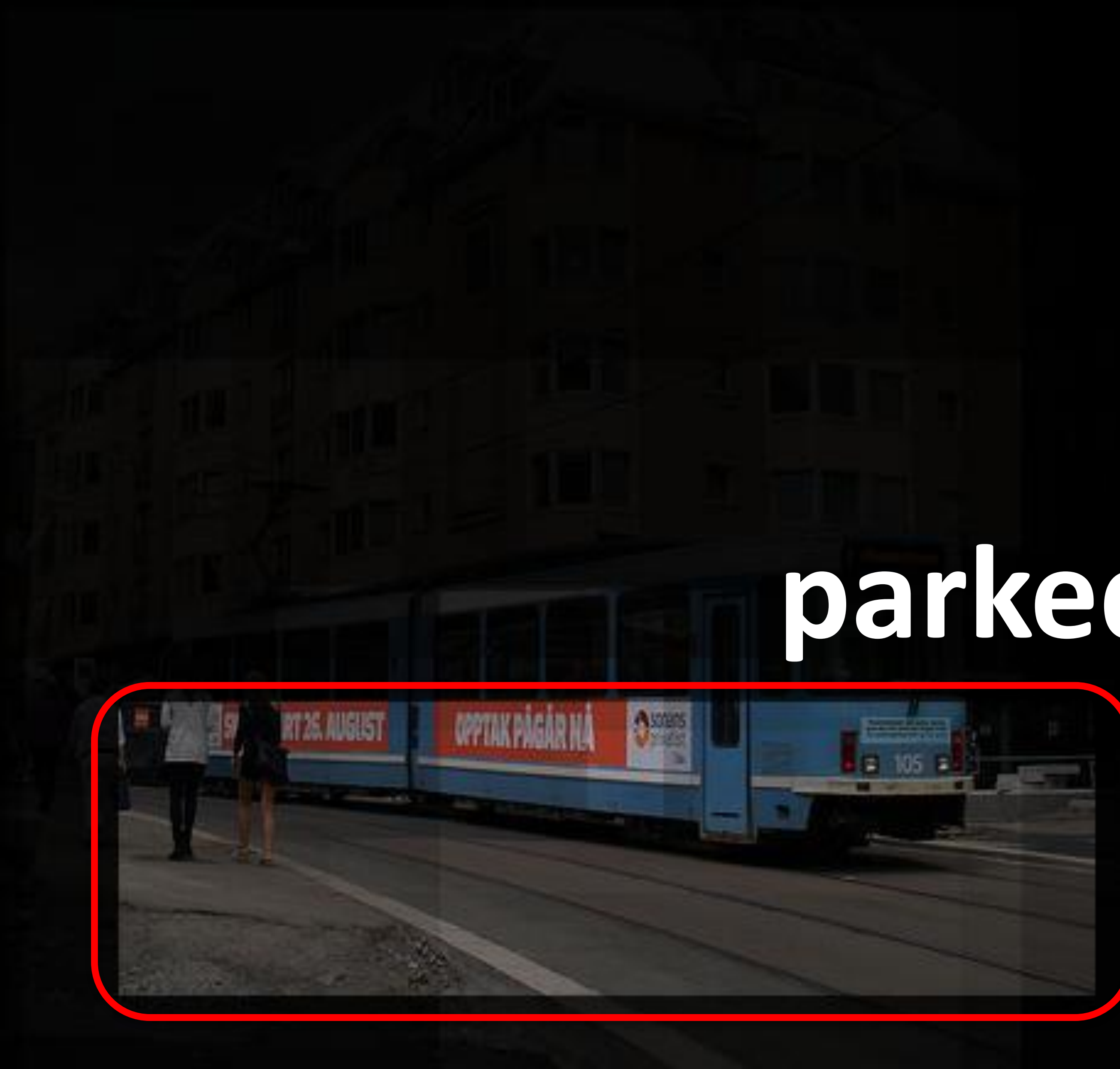
Demo


OPPTAK PÅGÅR NÅ
105
Original Image

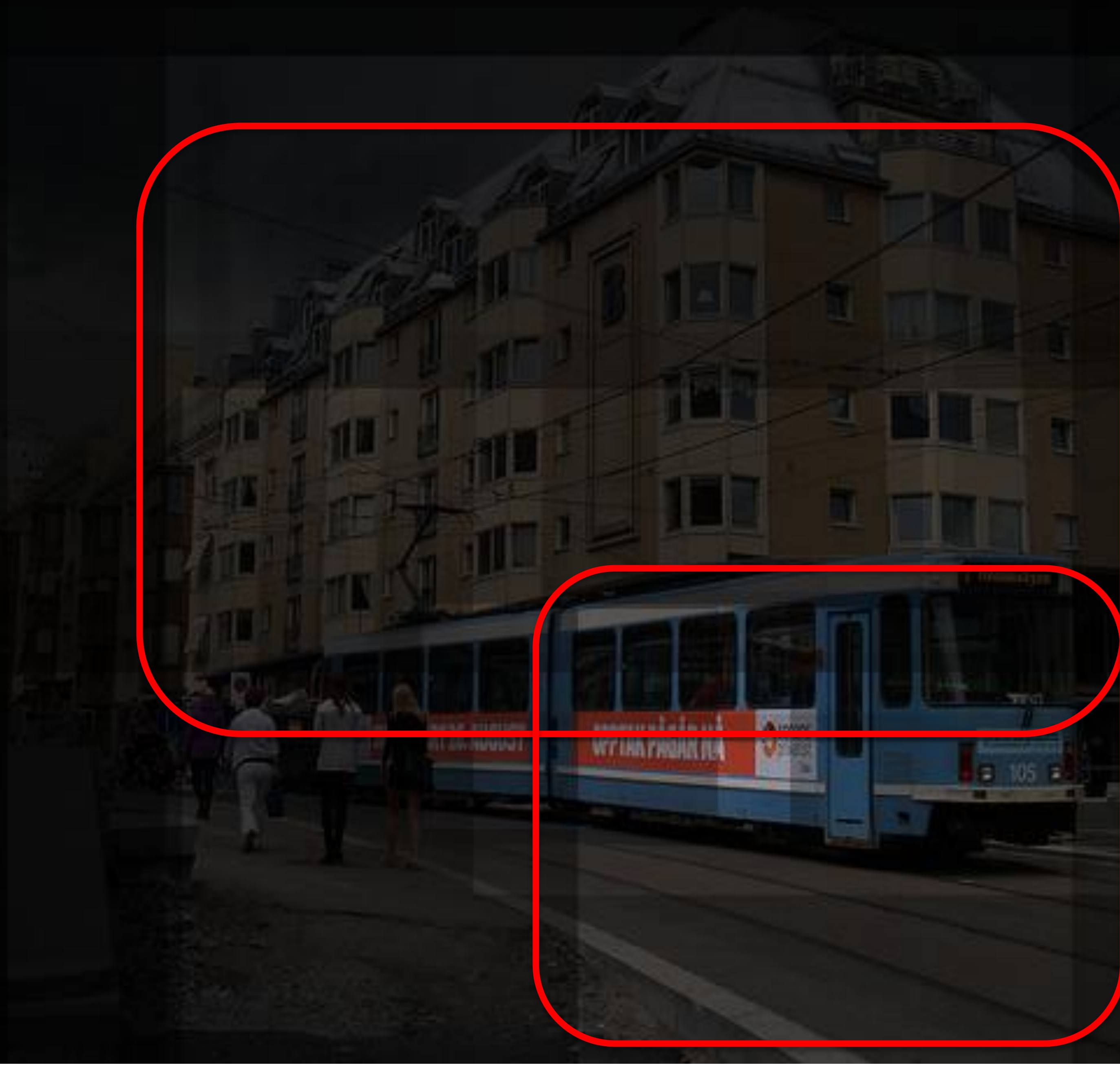
A

bus

is



parked



on

a

city

street

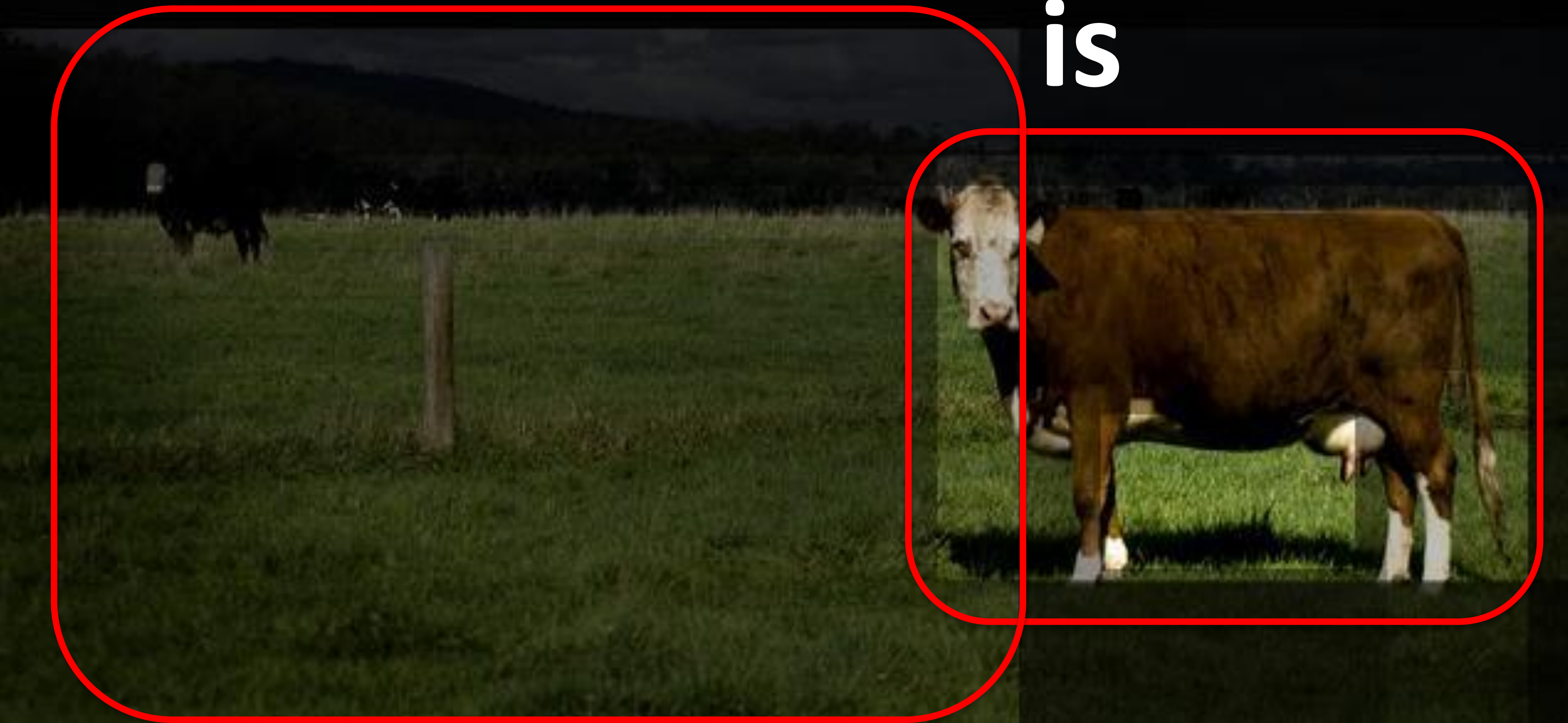

Original Image


A


brown

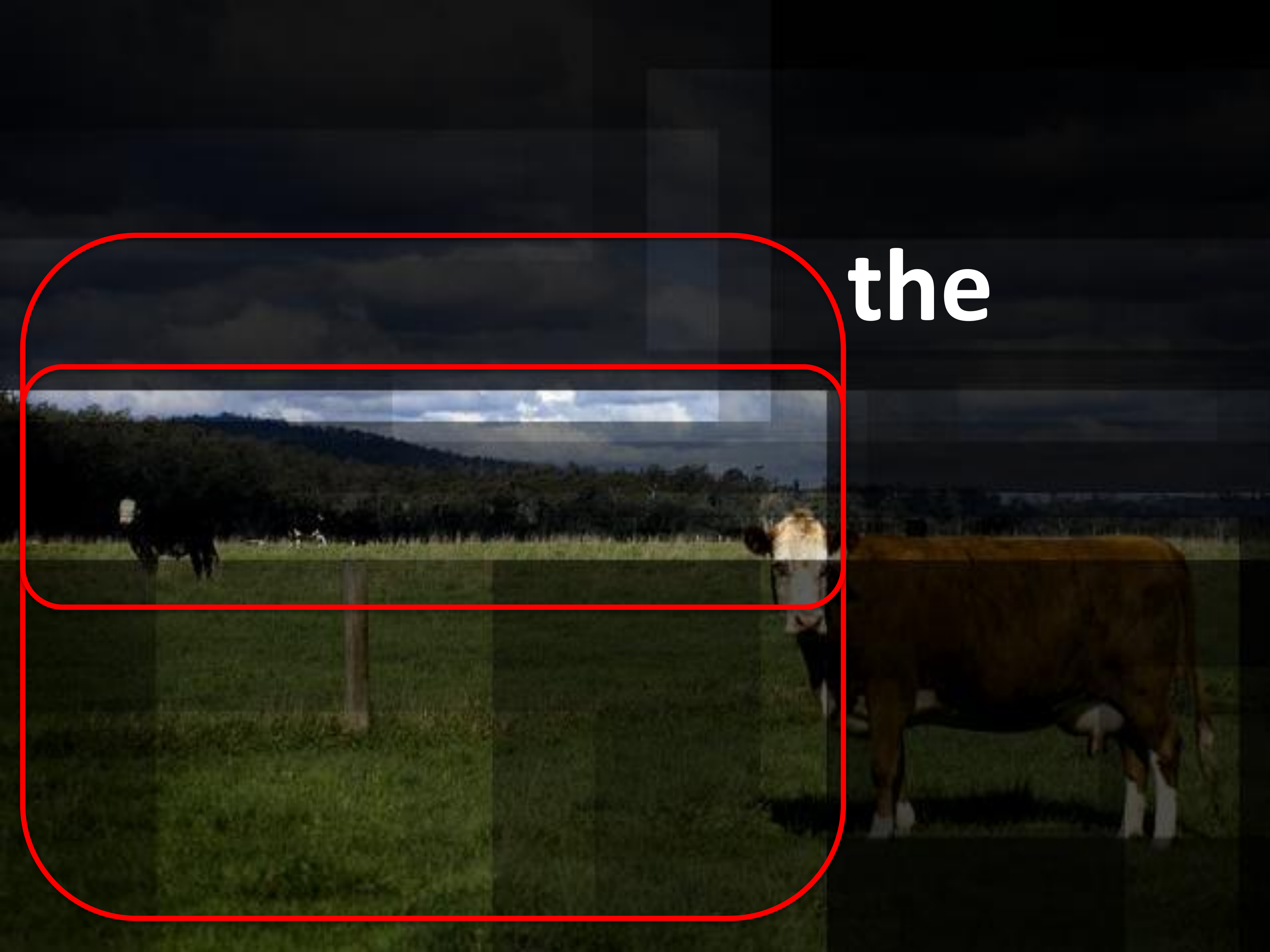
cow

is



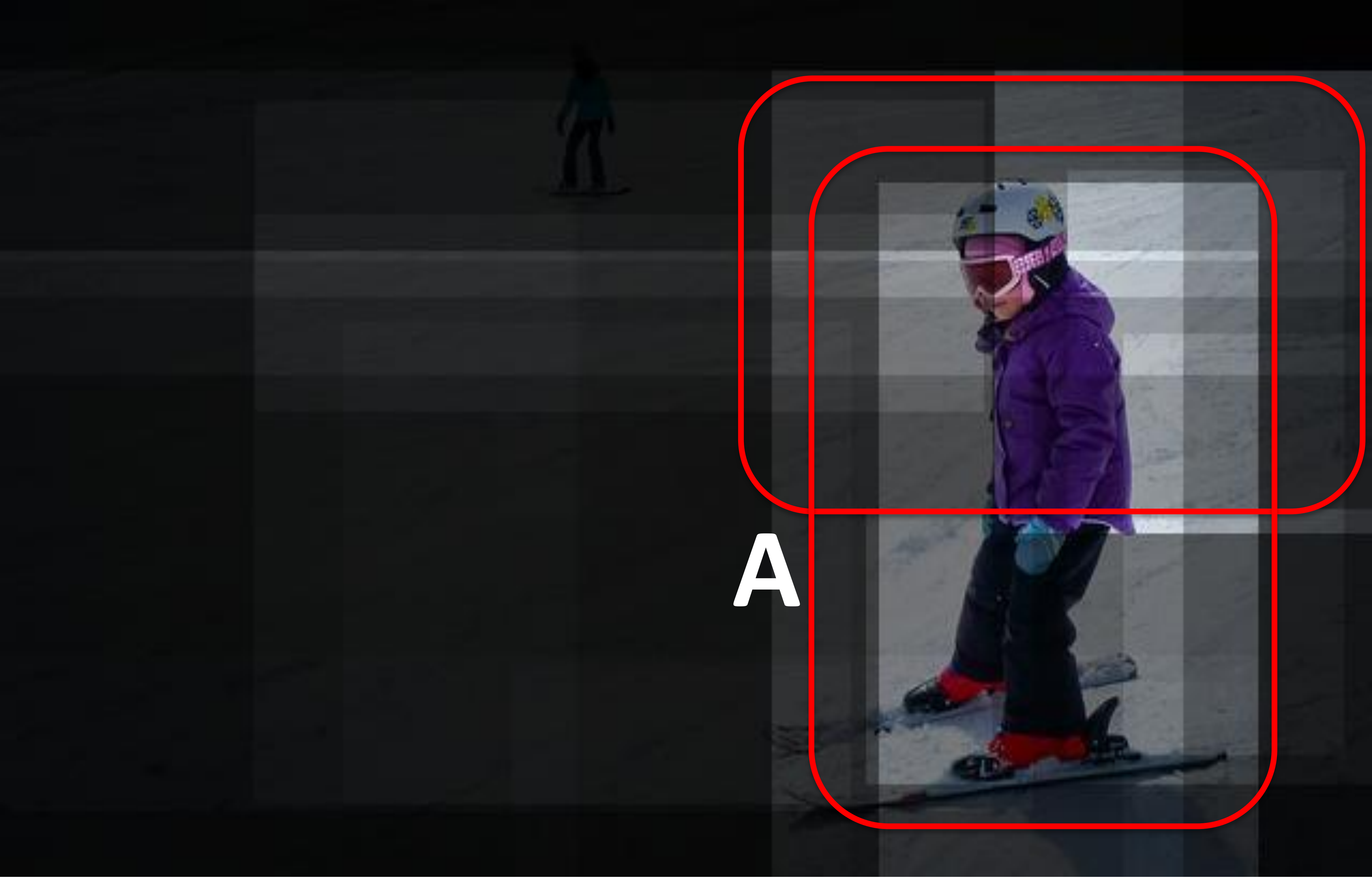

standing


in


the

grass

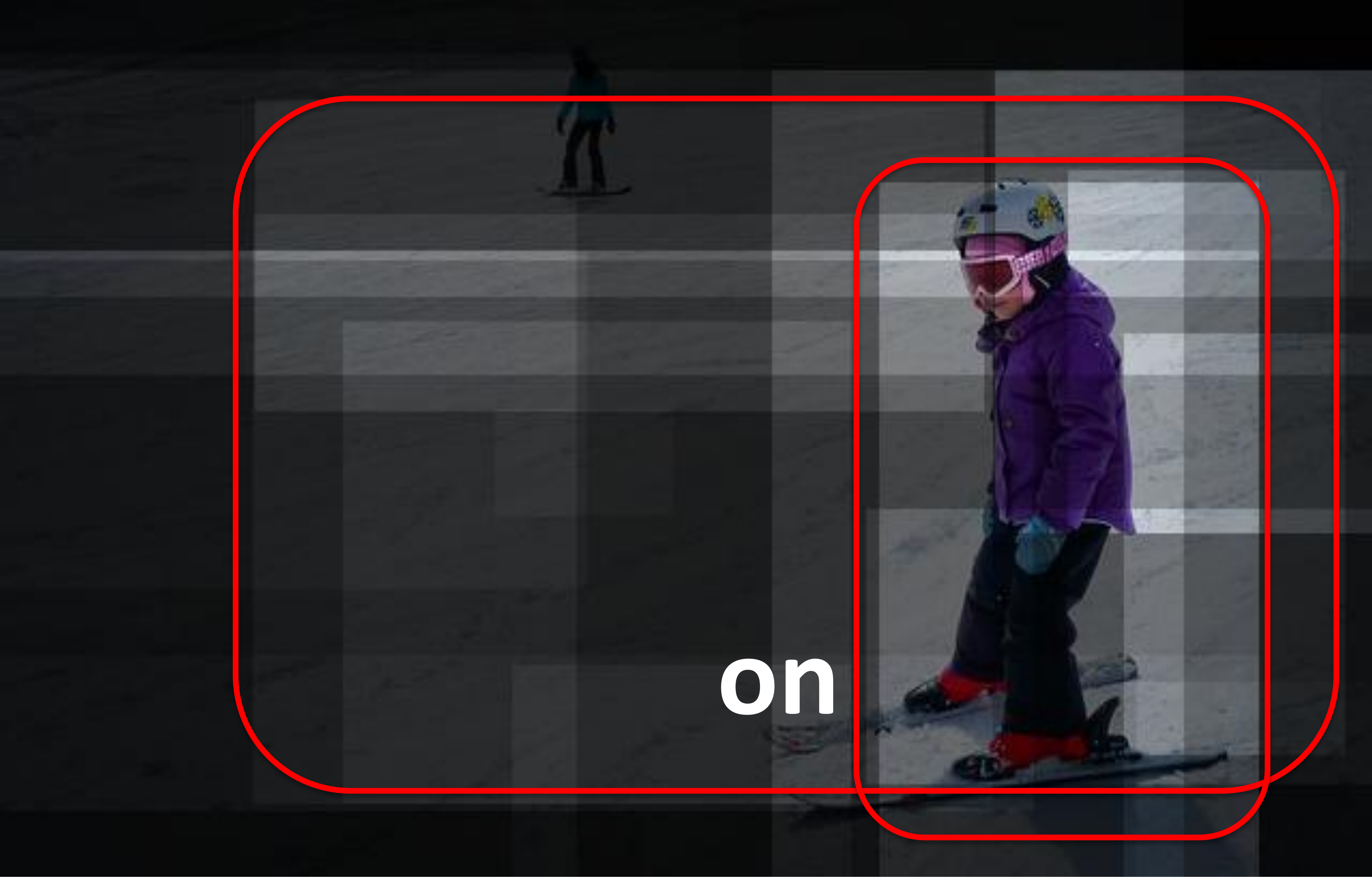

Original Image


A


person


on


skis


on


a


snowy

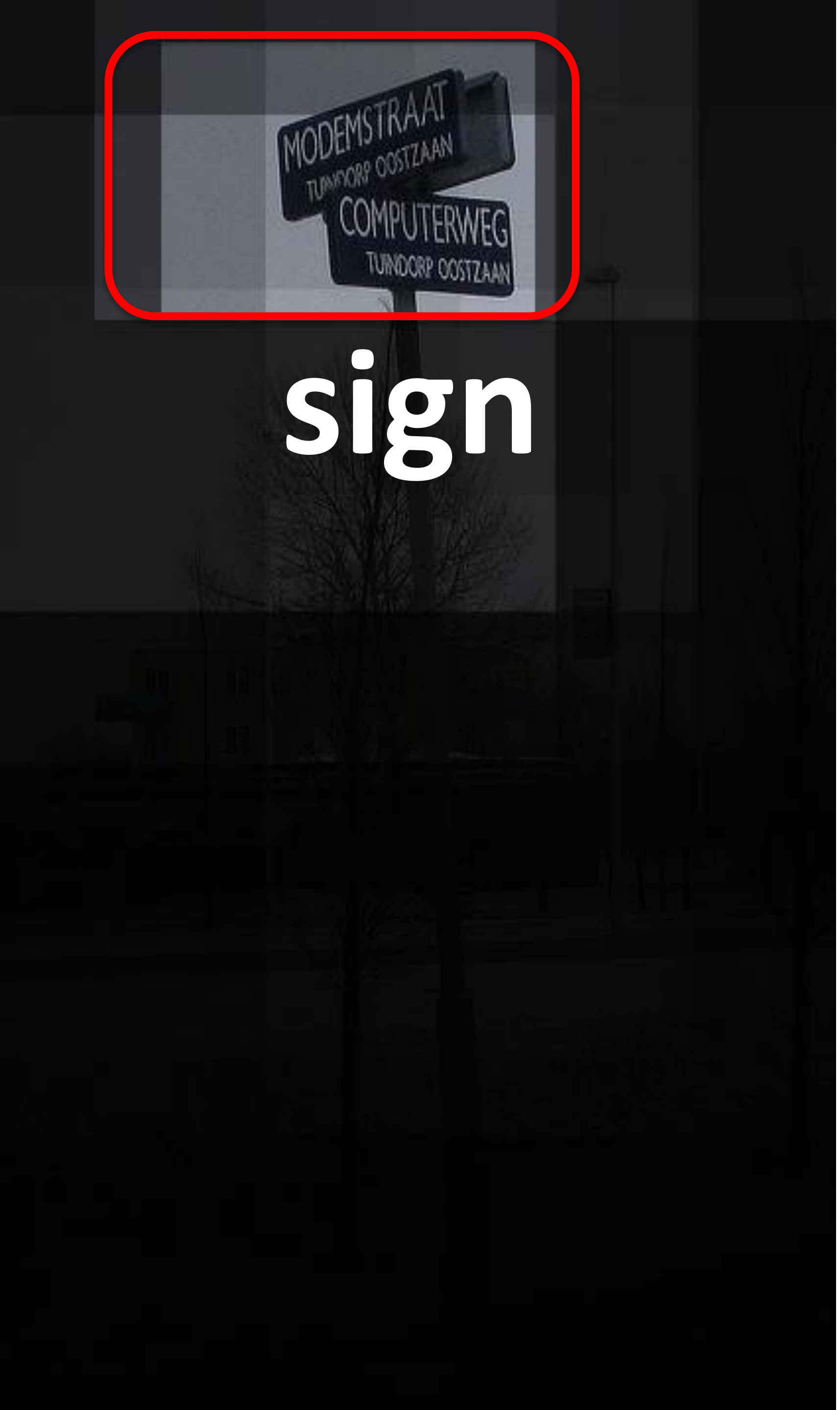
slope

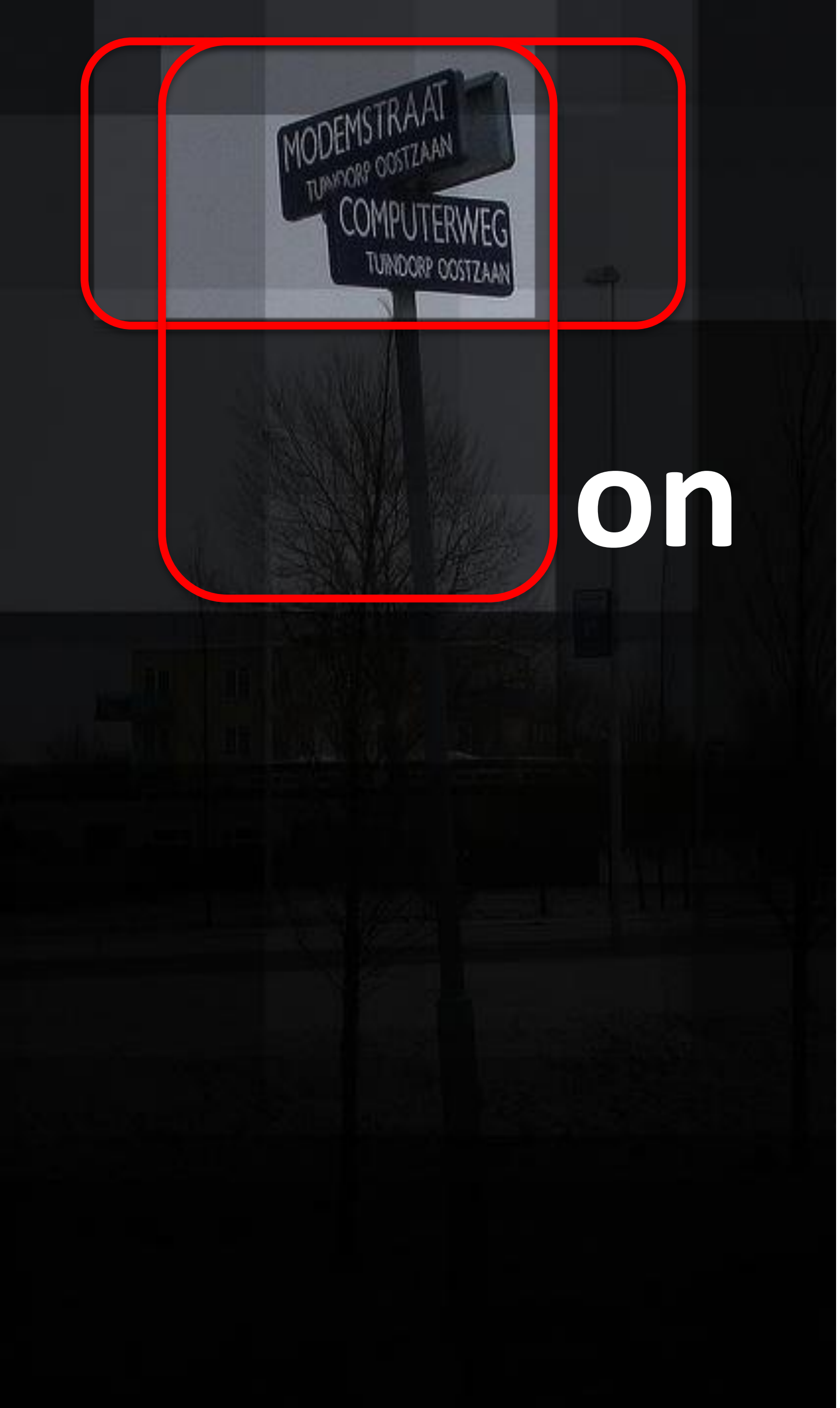

MODEMSTRAAT
COMPUTERWEG
TUINDORP OOSTZAAN
Original Image

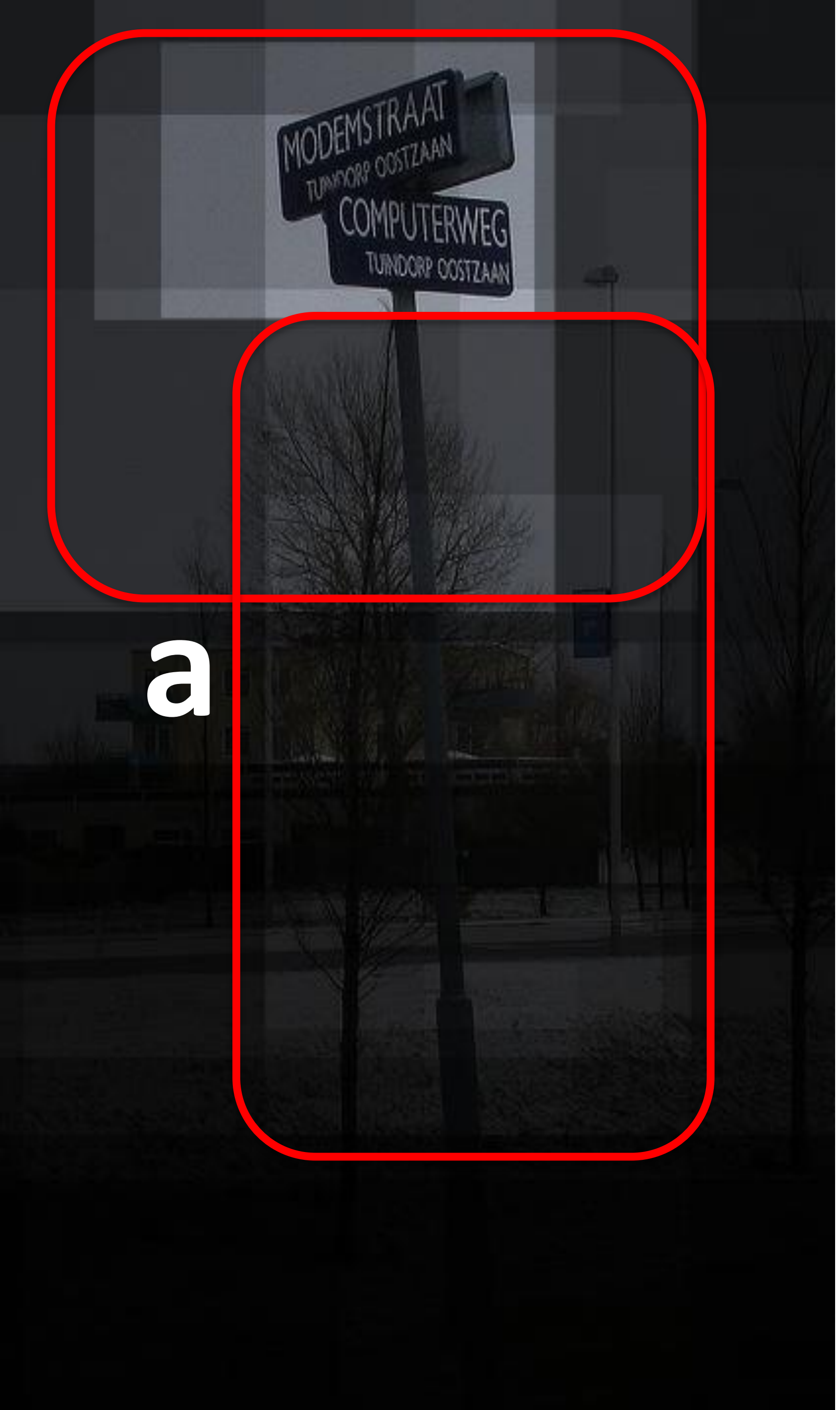

MODEMSTRAAT
COMPUTERWEG
TUINDORP OOSTZAAN
A

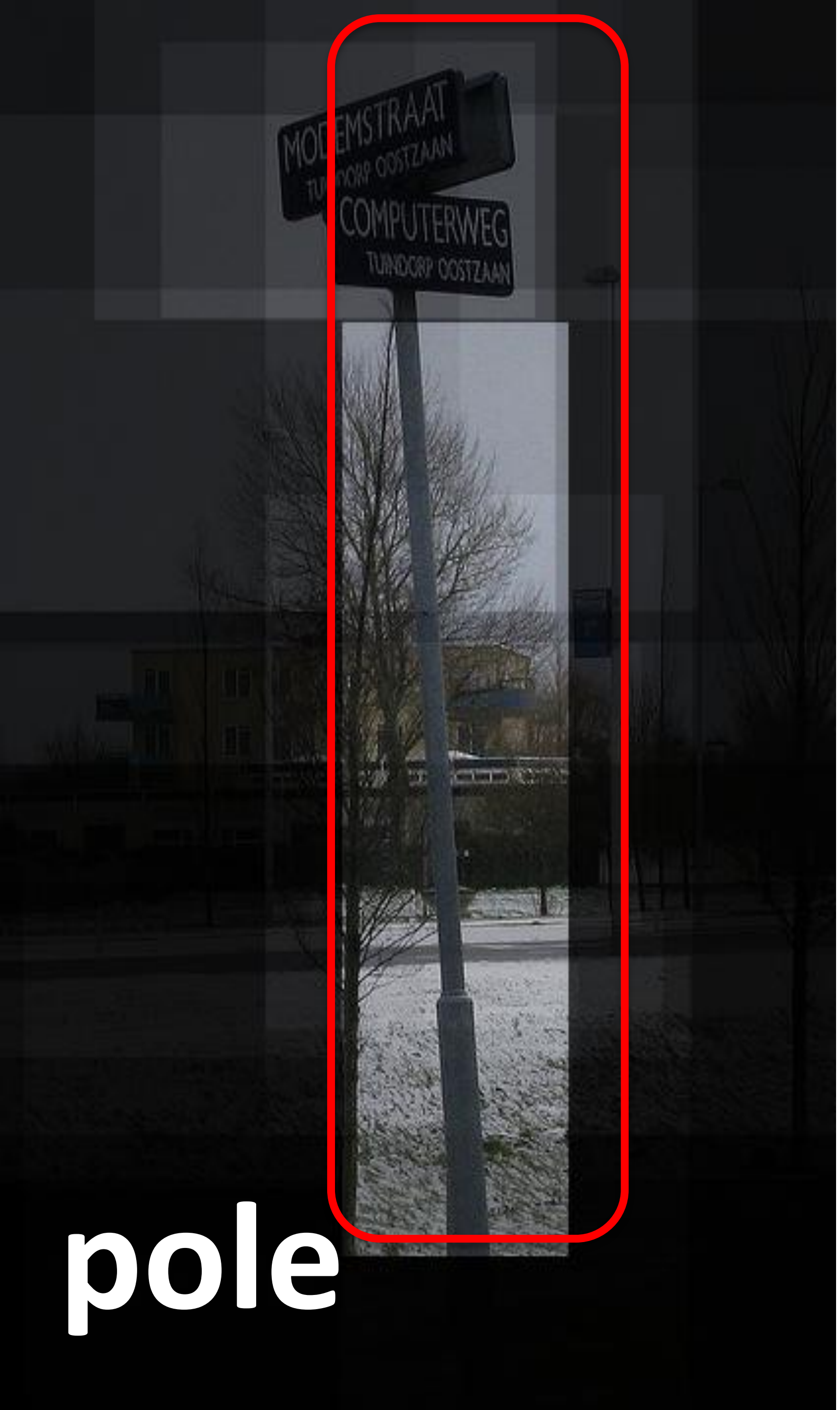

MODEMSTRAAT
TUINDORP OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN

street

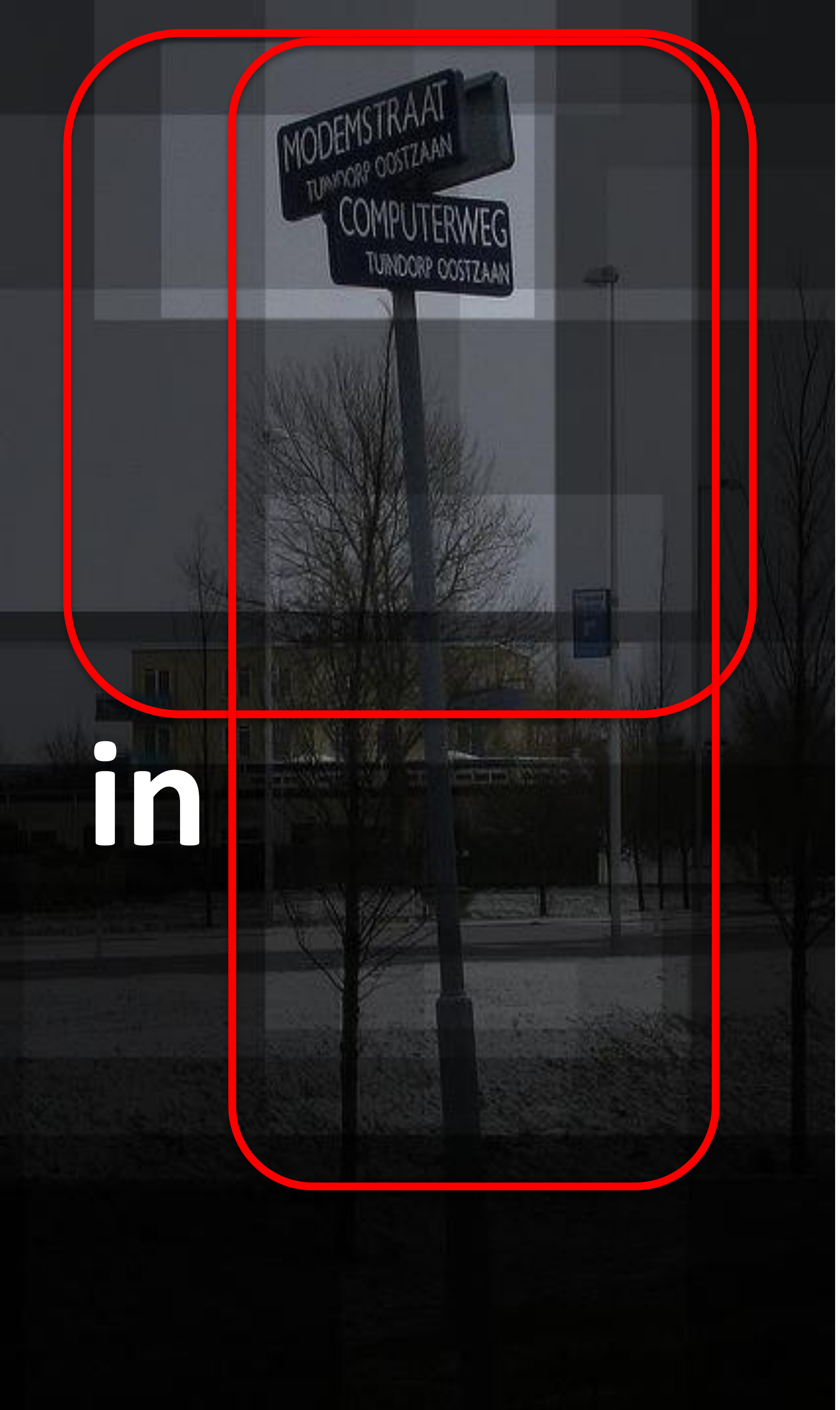

MODEMSTRAAT
OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN
sign


MODEMSTRAAT
TUINDORP OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN
on


MODEMSTRAAT
TUINDORP OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN
a



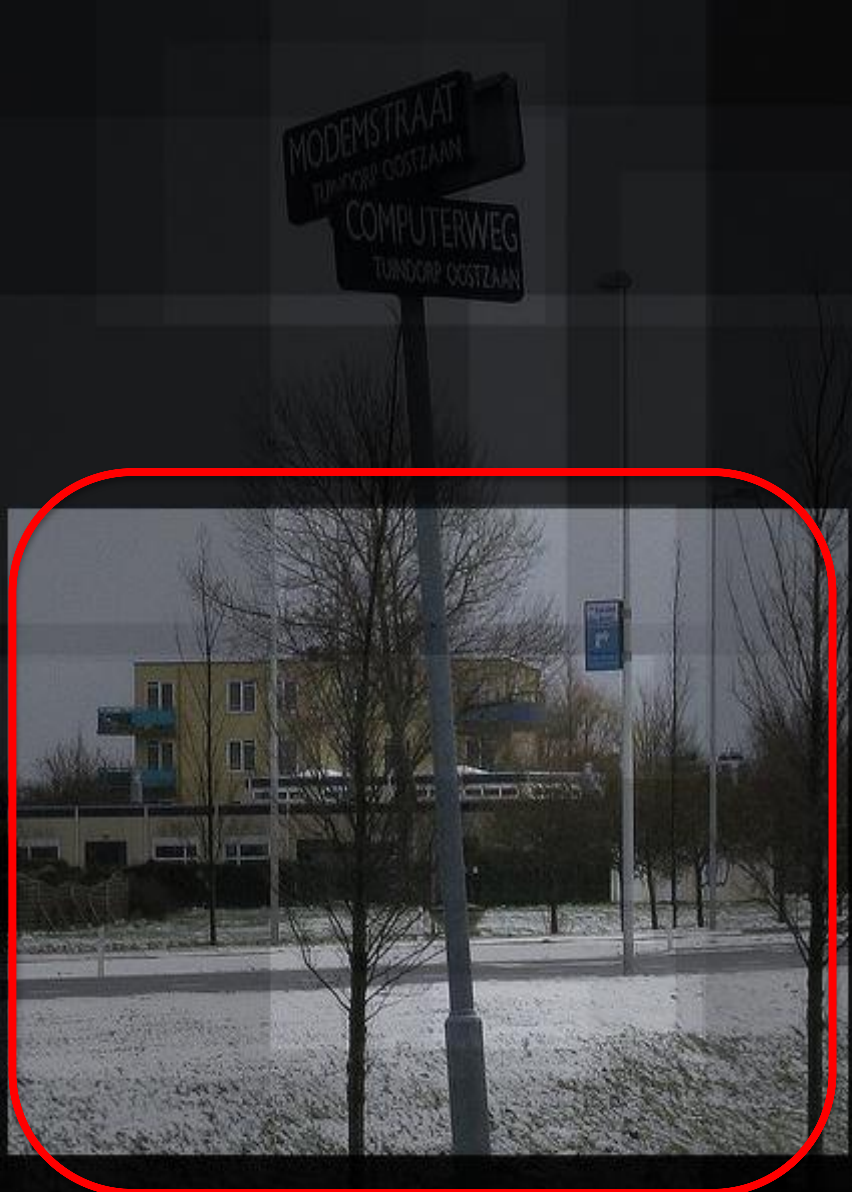
pole


MODEMSTRAAT
TUINDORP OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN

in


MODEMSTRAAT
TUINDORP OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN

the


MODEMSTRAAT
TUINDORP OOSTZAAN
COMPUTERWEG
TUINDORP OOSTZAAN

street