IMAGE CAPTIONING

LSTM APPLICATIONS

MOTIVATIONS

Globally, it is estimated that approximately 1.3 billion people live with some form of distance or near vision impairment. With regards to distance vision, 188.5 million have mild vision impairment, 217 million have moderate to severe vision impairment, and 36 million people are blind.

IMAGE CAPTIONING

Perception with respect to machines

SHOW, ATTEND AND TELL
NEURAL IMAGE CAPTION GENERATION WITH VISUAL ATTENTION

Kelvin Xu - Jimmy Lei Ba - Ryan Kiros - Kyunghyun Cho - Aaron Courville - Ruslan Salakhutdinov - Richard S. Zemel - Yoshua Bengio - 2016

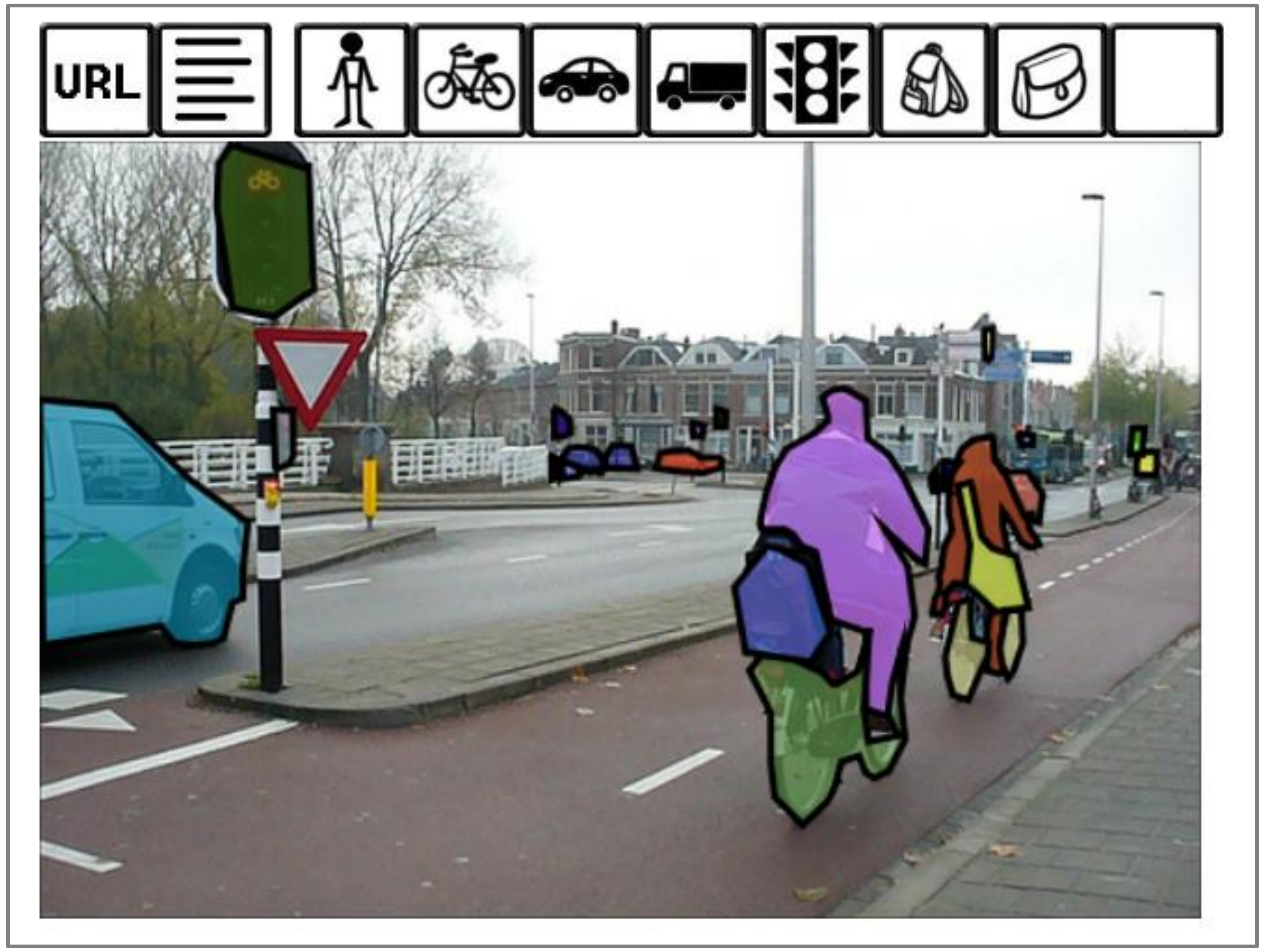
Attention Based Model learns to describe image contents

Deterministic Training using Standard Propagation

Learn to gaze on Salient Object

DATASET

COCO – Common Objects in Context



COCO Image Previewing

OVERVIEW

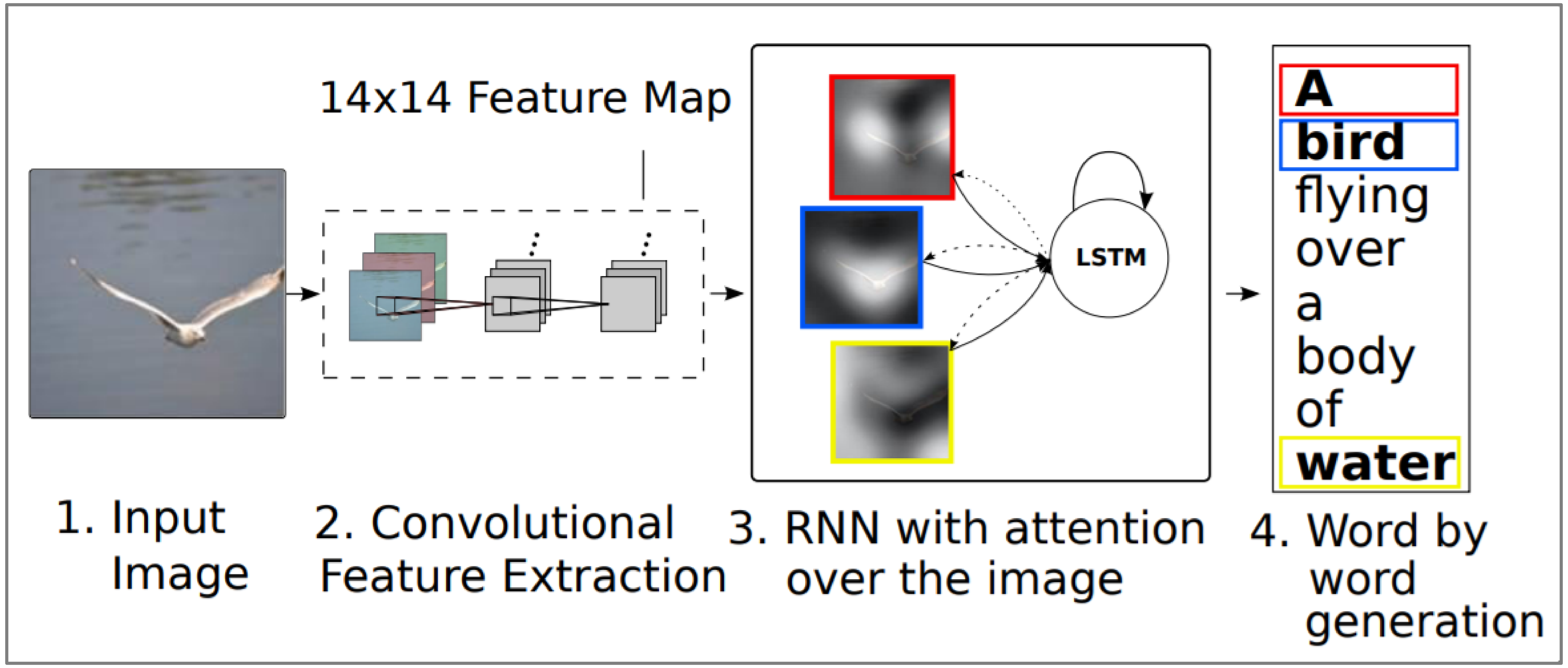


Figure 1 – Model learning word/image alignment

CONTRIBUTIONS

Two attention-based image caption generators under a common framework.

A **"soft" deterministic** attention mechanism trainable by standard back-propagation methods.

A **"hard" stochastic** attention mechanism trainable by maximizing an approximate variational lower bound

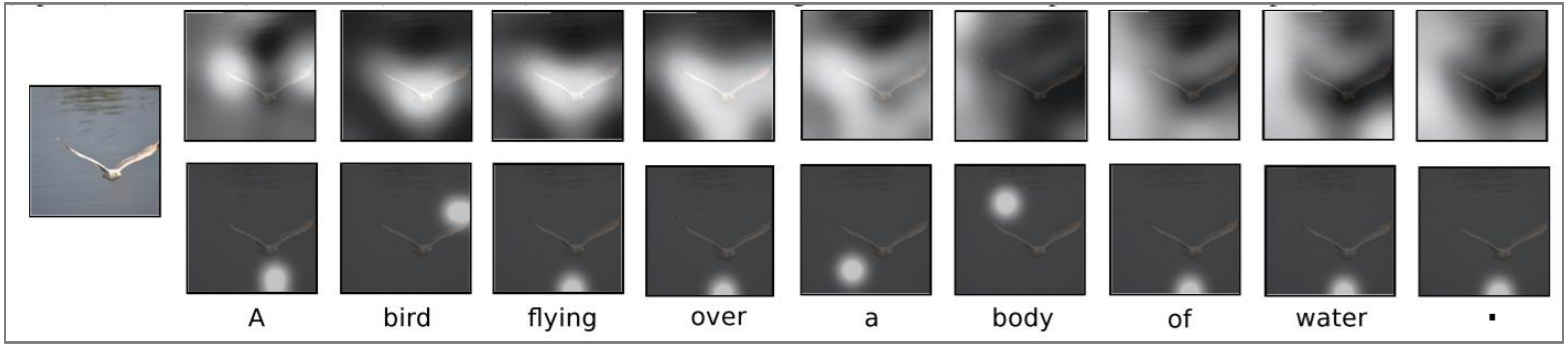


Figure 2 – Soft (top) & Hard (bottom) Attention models

IMPLEMENTATION

ENCODER - CONVOLUTIONAL FEATURES

Model takes a single raw image and generates a caption y encoded as a sequence of 1-of-K encoded words

$$y = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}, \ \mathbf{y}_i \in \mathbb{R}^K$$

Extract a set of feature vectors (annotations vectors)

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

IMPLEMENTATION

DECODER – LONG SHORT-TERM MEMORY NETWORK

Produces a caption by generating one word at every time step conditioned on a context vector

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{E}\mathbf{y}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$

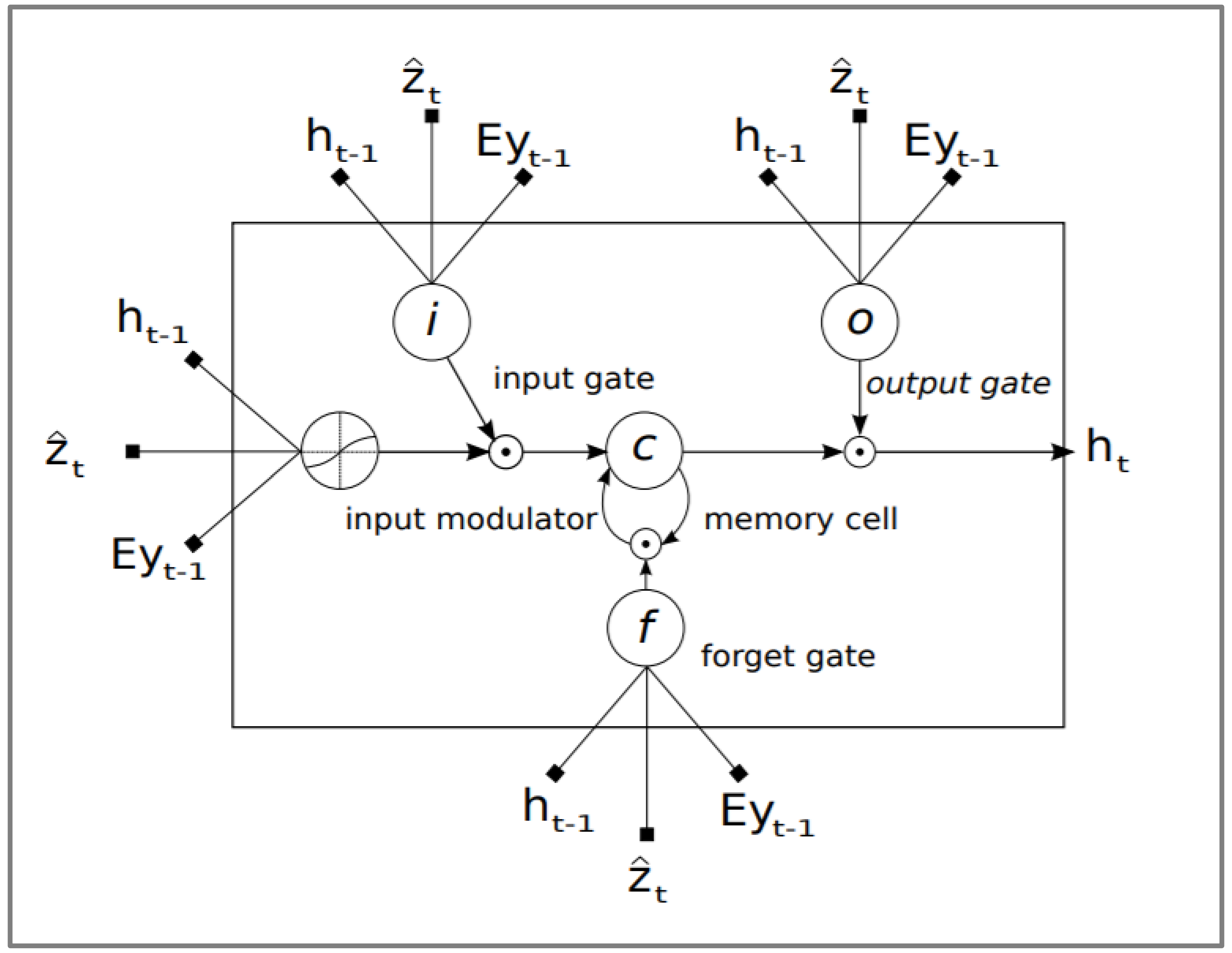


Figure 3 – LSTM Cell

RESULTS

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
		BLEU				
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Table 1 – Soft & Hard Attention models results

PERSONALIZED IMAGE CAPTIONING

Descriptive captioning with prior knowledge

ATTEND TO YOU
PERSONALIZED IMAGE CAPTIONING WITH CONTEXT SEQUENCE MEMORY NETWORKS

Cesc Chunseong Park - Byeongchang Kim -Gunhee Kim - 2017

Descriptive sentence with prior knowledge: Users vocabulary in previous documents

Context Sequence Memory Network (CSMN)

OVERVIEW

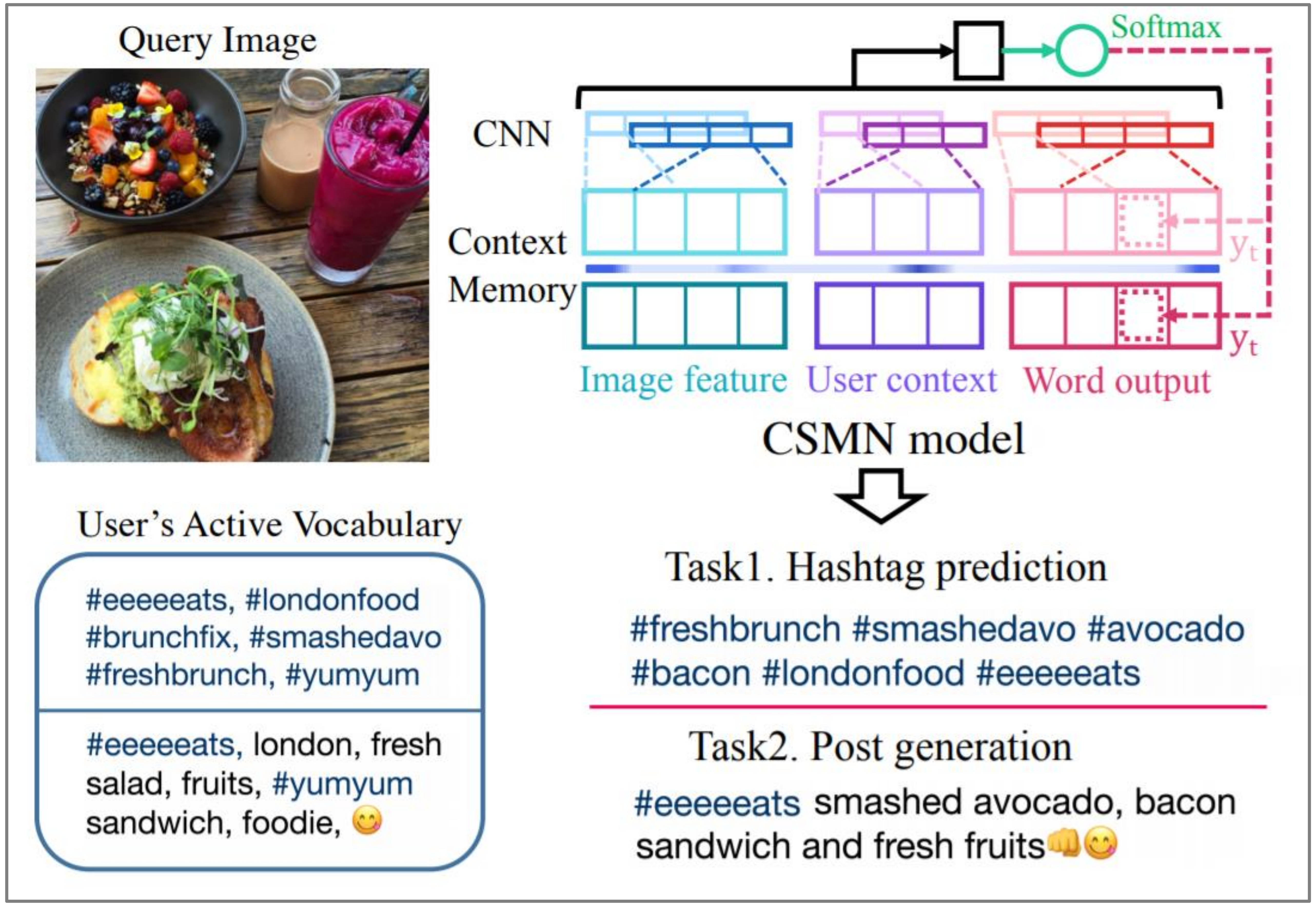


Figure 4 – Personalized image captioning with an Instagram example

DATASET

INSTAGRAM POSTS

Dataset	# posts	# users	# posts/user	# words/post
caption	721,176	4,820	149.6 (118)	8.55 (8)
hashtag	518,116	3,633	142.6 (107)	7.45 (7)

270 search keys – 10 most common hashtags over 27 categories (e.g. food, styles)

3.5M posts from 18k users

FILTERING

English only, links removal, remove user bias (> 50 posts) & limit post lengths (irrelevance)

720k captions & 520k hashtags

PREPROCESSING

Vocabularies building, frequency based, 40k captions, 60k hastags

IMPLEMENTATION

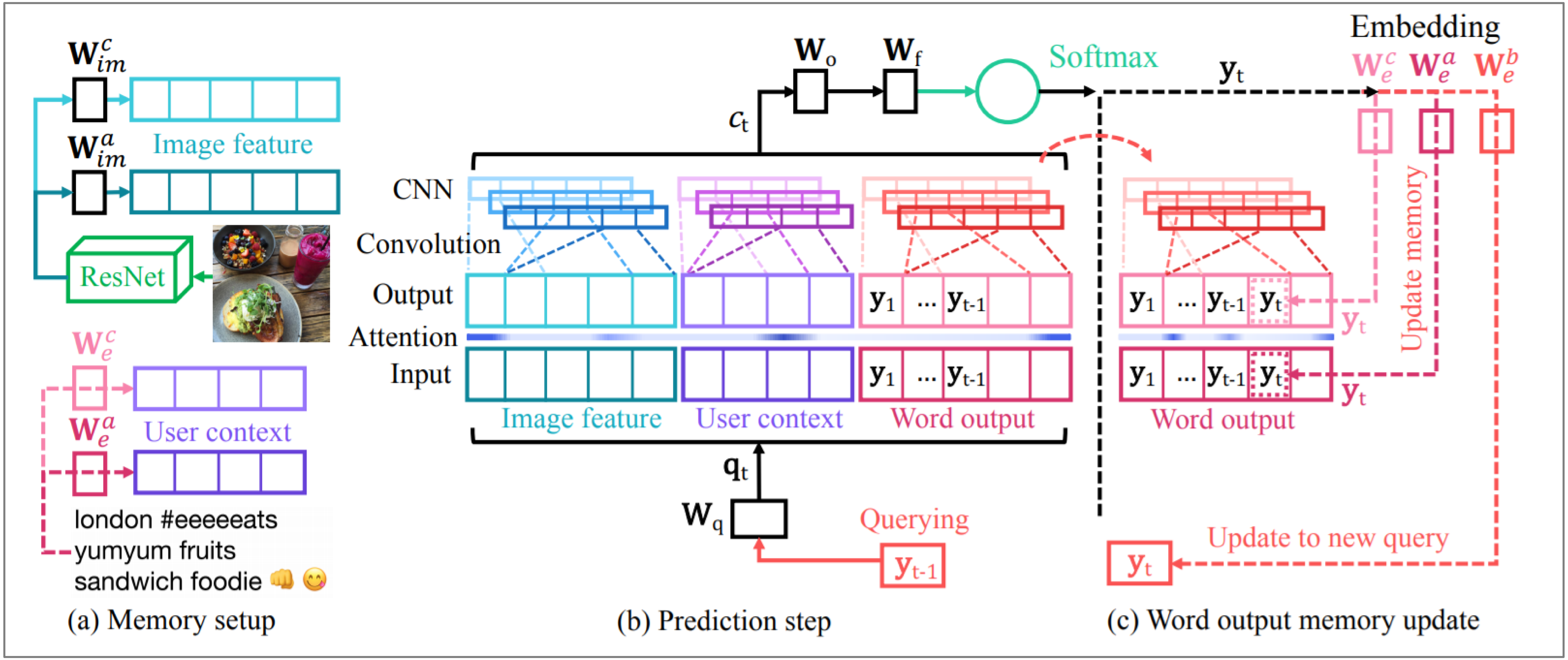


Figure 5 – Context sequence memory network (CSMN) Model

CONTEXT MEMORY

IMAGE MEMORY

ResNet 101 pretrained on the ImageNet 2012 dataset

(7 × 7) feature maps of res5c layer (49 cells) model exploits spatial attention

pool5 feature (1 cell) single memory focus

$$\mathbf{m}_{im,j}^{a/c} = \mathrm{ReLU}(\mathbf{W}_{im}^{a/c}\mathbf{I}_j^{p5} + \mathbf{b}_{im}^{a/c}).$$

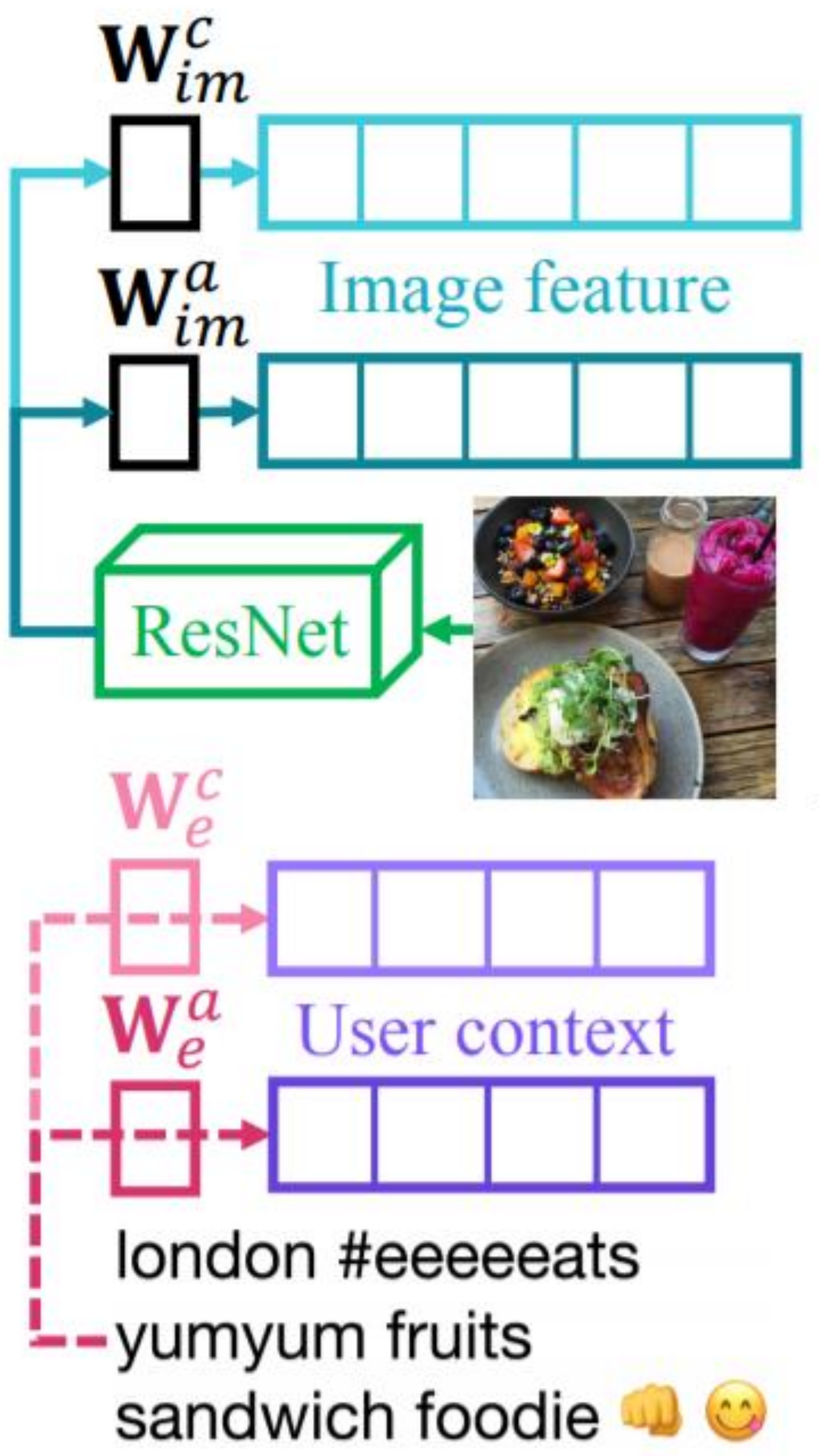
USER CONTEXT MEMORY

User's most frequent words from previous posts (D)

Uses frequency-inverse document frequency (TF-IDF)

Results in ignoring too general terms that many users commonly use

$$\mathbf{u}_j^a = \mathbf{W}_e^a\mathbf{u}_j, \mathbf{u}_j^c = \mathbf{W}_e^c\mathbf{u}_j; \mathbf{y}_j; \;\; j \in 1, \ldots, D$$
$$\mathbf{m}_{us,j}^{a/c} = \mathrm{ReLU}(\mathbf{W}_h[\mathbf{u}_j^{a/c}] + \mathbf{b}_h),$$



(a) Memory setup

STATE-BASED SEQUENCE GENERATION

Approach does not involve any RNN module

Sequentially store all previously generated words into the memory

=> Enables to predict each output word by selectively attending on the combinations of all previous words, image regions, and user context

$$\mathbf{q}_t = \text{ReLU}(\mathbf{W}_q\mathbf{x}_t + \mathbf{b}_q), \text{ where } \mathbf{x}_t = \mathbf{W}_e^b\mathbf{y}_{t-1}.$$

$$\mathbf{p}_t = \text{softmax}(\mathbf{M}_t^a\mathbf{q}_t), \ \mathbf{M}o_t(*,i) = \mathbf{p}_t \circ \mathbf{M}_t^c(*,i).$$

$$\mathbf{M}o_t = [\mathbf{m}_{im,1:49}^o \oplus \mathbf{m}_{us,1:D}^{a/c} \oplus \mathbf{m}_{ot,1:t-1}^{a/c}].$$

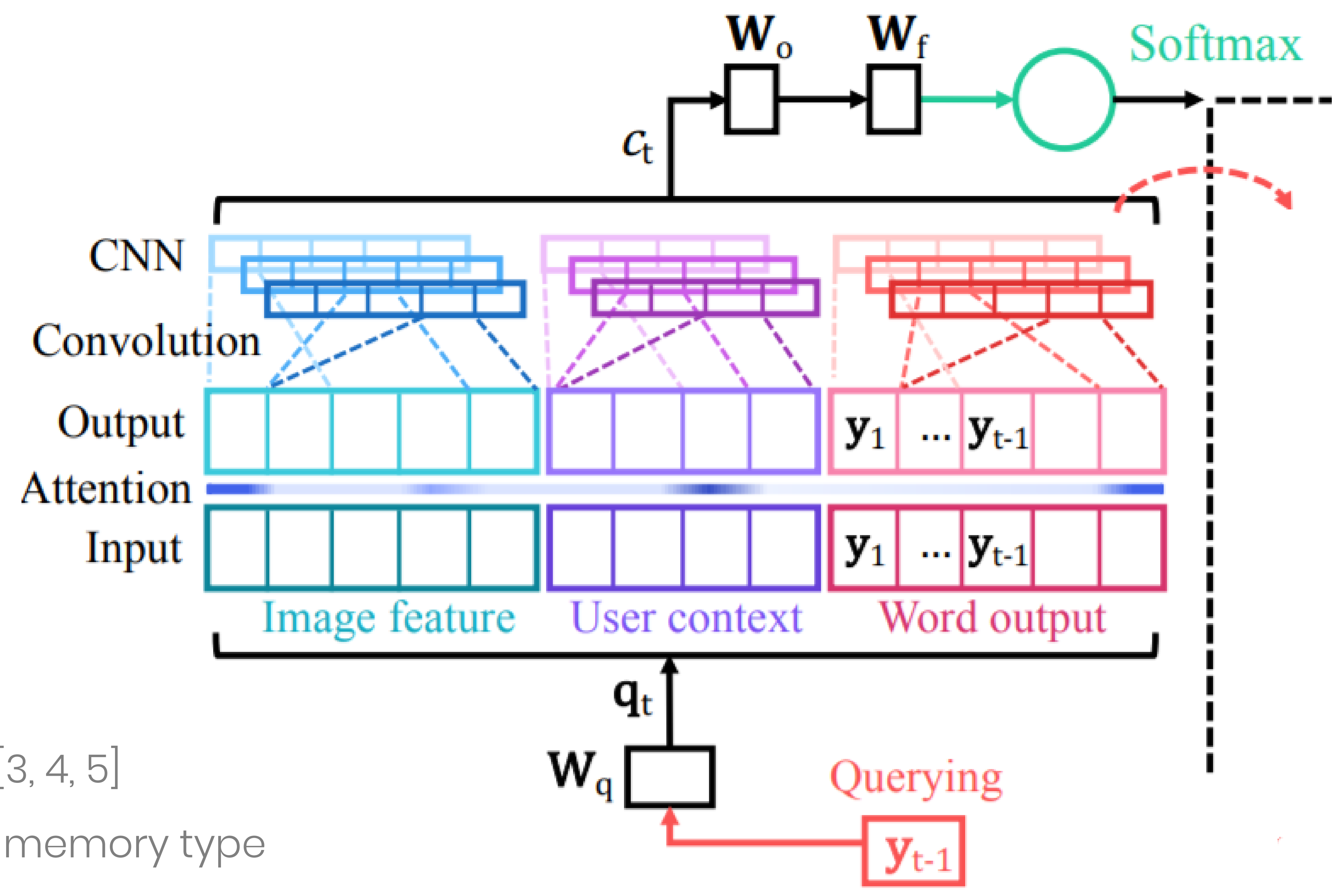
Define a set of three filters whose depth is 300 by changing window sizes h = [3, 4, 5]

Separately apply a single convolutional layer and max-pooling layer to each memory type

$$\mathbf{c}_{im,t}^h = \text{maxpool}(\text{ReLU}(\mathbf{w}_{im}^h * \mathbf{m}_{im,1:49}^o + \mathbf{b}_{im}^h))$$

We obtain c(im,t) by concatenating c(h,im,t) from h = 3 to 5

$$\mathbf{c}_t = [\mathbf{c}_{im,t} \oplus \mathbf{c}_{us,t} \oplus \mathbf{c}_{ot,t}]$$



(b) Prediction step

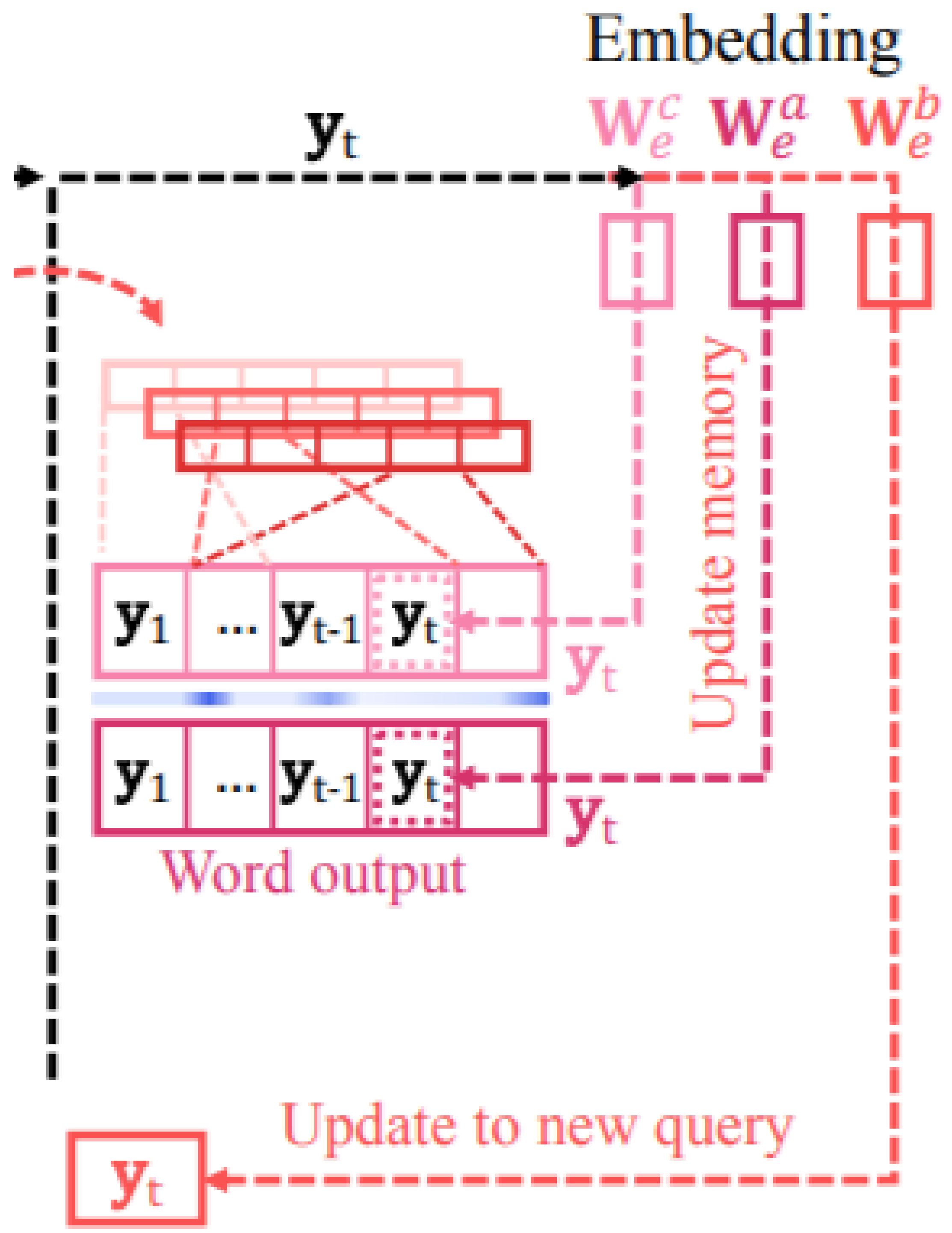
WORD PREDICTION

The hidden state ht with a weight matrix Wo, compute the output probability st over vocabularies V by a softmax layer

$$\mathbf{h}_t = \mathrm{ReLU}(\mathbf{W}_o \mathbf{c}_t + \mathbf{b}_o),$$
$$\mathbf{s}_t = \mathrm{softmax}(\mathbf{W}_f \mathbf{h}_t).$$

Select the word that attains the highest probability

$$y_t = \mathrm{argmax}_{\mathbf{s} \in \mathcal{V}}(\mathbf{s}_t)$$



(c) Word output memory update

TRAINING

Teacher forced learning, provide the correct memory state to predict next words

Softmax cross-entropy loss as the cost function for every time step predictions

$\Rightarrow$ minimizes the negative log likelihood from the estimated yt

Randomly initialize all the parameters with a uniform unit scaling of 1.0 factor

Apply mini-batch stochastic gradient descent

Use Adam optimizer with $\beta 2 = 0.9$, $\beta 2 = 0.999$ and $\epsilon = 1e - 08$

To speed up training, using four GPUs for data parallelism, and setting a batch size as 200 for each GPU

Initial learning rate is set as 0.001

Every 5 epochs, divide the learning rate by 1.2 to gradually decrease it

Training over 20 epochs

RESULTS
CAPTIONS

(GT) pool pass for the summer ✔
(Ours) the pool was absolutely perfect ☀
(NoCNN) the beach

(GT) the face in the woods
(Ours) my first painting of the day
(Usr) no enhancements needed

(GT) awesome view of the city
(Ours) the city of cincinnati is so pretty
(UsrIm) there are no words

(GT) this speaks to me literarily
(Ours) I love this #quote
(Showtell) is the only thing that matters _UNK

(GT) dinner and drinks with @username
(Ours) wine and movie night with @username
(Im) my afternoon is sorted

(GT) air is in the fall
(Ours) fall is in the air
(Usr) tis the holiday season

(GT) pretty flowers from the hubby 🌸💐
(Ours) my beautiful flowers from my hubby
(NoFB) I love

RESULTS
HASHTAGS

(GT) #fashionkids #stylish-cubs #kidzfashion …
(Ours) #pink #babygirl **#fashionkids** #cutekidsclub …

(GT) #connecticut #books #bookbarn
(Ours) **#books** #reading

(GT) #coffee #dailycortado #love #vscocam #vscogood #vscophile #coffeebreak …
(Ours) **#coffee** #coffeetime #coffeeart #latte #latteart **#coffeebreak** **#vsco**

(GT) #style #fashion #shopping #shoes #kennethcole…
(Ours) #newclothes **#fashion** **#shoes** #brogues

(GT) #boudoir #heartprint #love #weddings #potterybarn
(Ours) #decor #homedecor #interiors #interiordesign #rustic #bride #pretty **#wedding** #home #white

(GT) #greensmoothie #dairyfree #lifewithatoddler #glutenfree #vegetarian …
(Ours) **#greensmoothie** #greenjuice #smoothie #vegan #raw #juicing #eatclean #detox #cleanse

RESULTS
COMPARISONS

Methods	B-1	B-2	B-3	B-4	METEOR	CIDEr	ROUGE-L
(seq2seq)	0.050	0.012	0.003	0.000	0.024	0.034	0.065
(ShowTell)*	0.055	0.019	0.007	0.003	0.038	0.004	0.081
(AttendTell)*	0.106	0.015	0.000	0.000	0.026	0.049	0.140
(1NN-Im)*	0.071	0.020	0.007	0.004	0.032	0.059	0.069
(1NN-Usr)	0.063	0.014	0.002	0.000	0.028	0.025	0.059
(1NN-UsrIm)	0.106	0.032	0.011	0.005	0.046	0.084	0.104
(CSMN-NoCNN-P5)	0.086	0.037	0.015	0.000	0.037	0.103	0.122
(CSMN-NoUC-P5)*	0.079	0.032	0.015	0.008	0.037	0.133	0.120
(CSMN-NoWO-P5)	0.090	0.040	0.016	0.006	0.037	0.119	0.116
(CSMN-R5C)	0.097	0.034	0.013	0.006	0.040	0.107	0.110
(CSMN-P5)	**0.171**	**0.068**	**0.029**	**0.013**	**0.064**	**0.214**	**0.177**
(CSMN-W20-P5)	0.116	0.041	0.018	0.007	0.044	0.119	0.123
(CSMN-W100-P5)	0.109	0.037	0.015	0.007	0.042	0.109	0.112

Methods	F1 score	
(seq2seq)	0.132	0.085
(ShowTell)*	0.028	0.011
(AttendTell)*	0.020	0.014
(1NN-Im)*	0.049	0.110
(1NN-Usr)	0.054	0.173
(1NN-UsrIm)	0.109	0.380
(CSMN-NoCNN-P5)	0.135	0.310
(CSMN-NoUC-P5)*	0.111	0.076
(CSMN-NoWO-P5)	0.117	0.244
(CSMN-R5C)	0.192	0.340
(CSMN-P5)	**0.230**	**0.390**
(CSMN-W20-P5)	0.147	0.349
(CSMN-W80-P5)	0.135	0.341

THANK YOU