

Joint understanding of vision and language for online blog posts

Gunhee Kim

Computer Science and Engineering



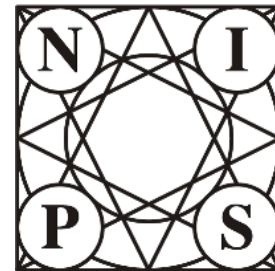
서울대학교

SEOUL NATIONAL UNIVERSITY

June 3, 2016

Outline

- Expressing an Image Stream with a Sequence of Natural Sentences. NIPS2015.
- Automatic Photo Posting and Commenting on Social Networks via Context Memory Networks (submitted)



Expressing an Image Stream with a Sequence of Natural Sentences

Cesc Chunseong Park

Gunhee Kim

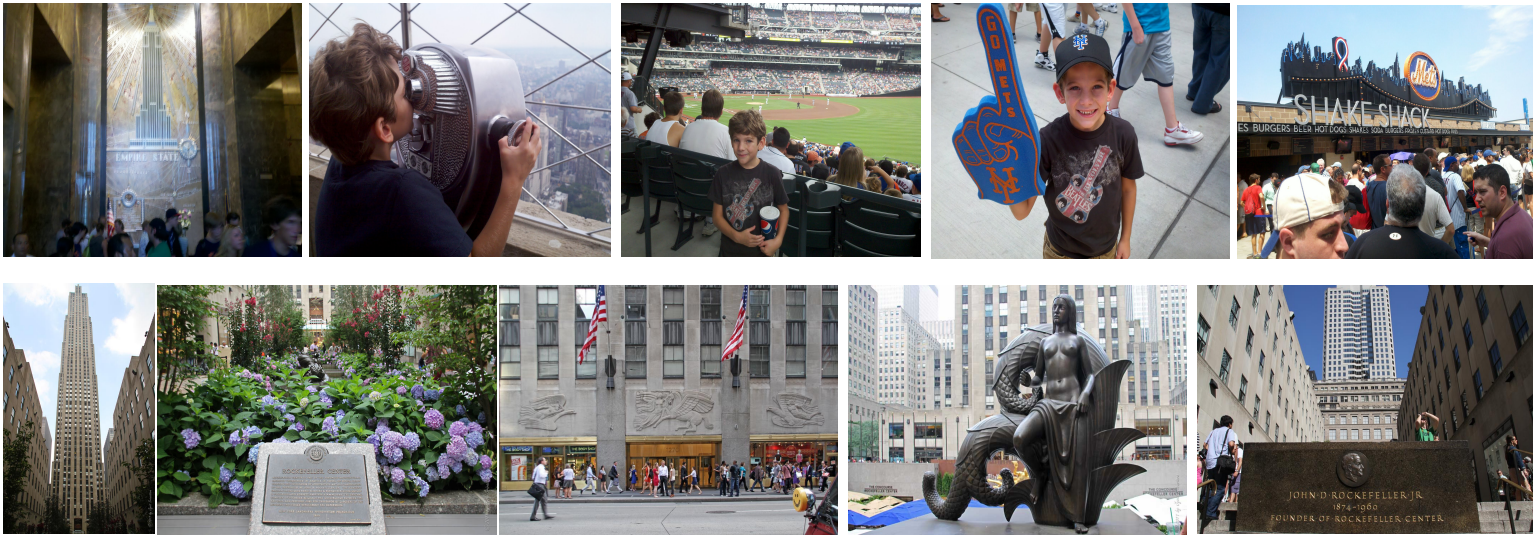


SEOUL NATIONAL UNIV.
VISION & LEARNING

(NIPS 2015)

General Users' Photo Stream

Suppose that you and your family visit NYC

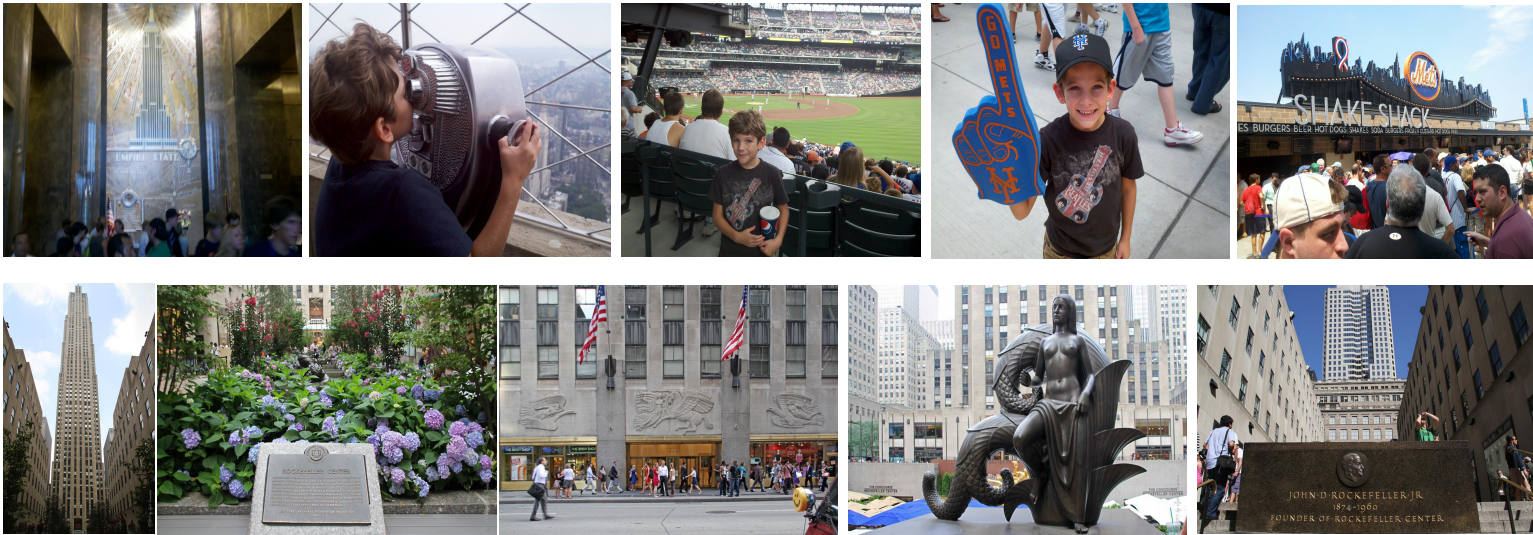


Users usually don't take only a single picture, but multiple pictures as a stream

A photo stream is a thread of the user's story

General Users' Photo Stream

Suppose that you and your family visit NYC

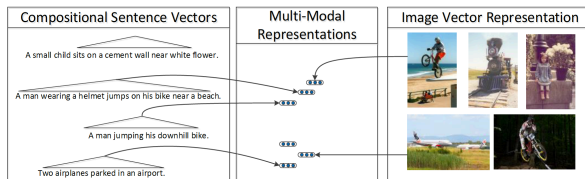


☹ Users do not organize the photo streams for later use

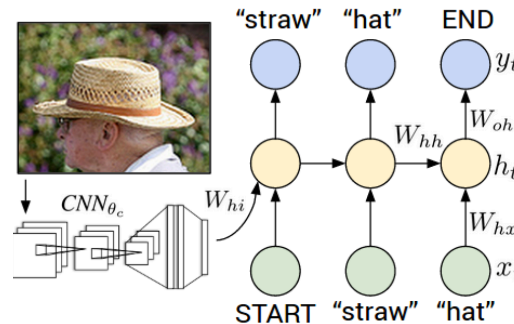
Can we write a travelogue for a given photo stream?

Previous Work – Image Captioning

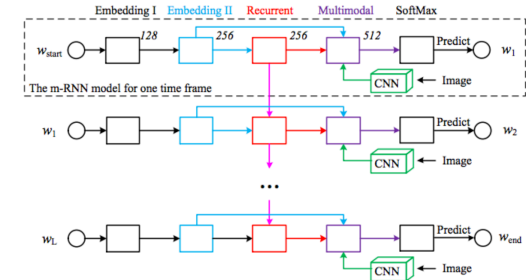
Retrieve or generate a descriptive natural language sentence for a given image



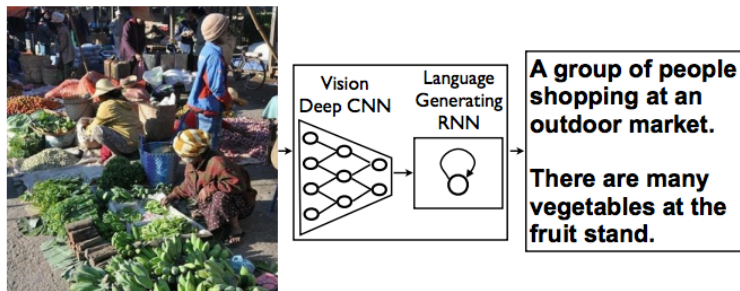
[Socher et al, TACL2013]



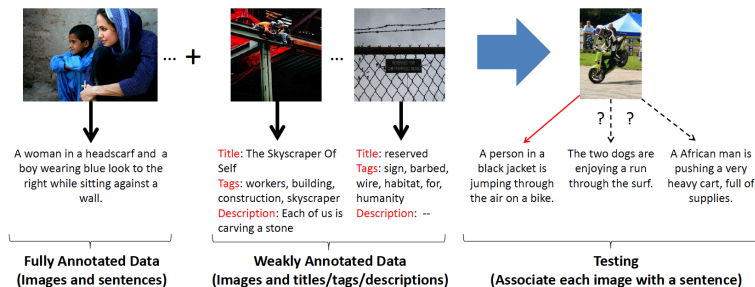
[Karpathy et al. CVPR2015]



[Mao et al, ICLR2015]



[Vinyal et al, CVPR2015]

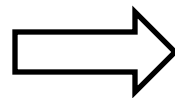


[Gong et al. ECCV2014]

Many more!

Limitation of Previous Work

Much of previous work mainly discuss the relation between a single image and a single sentence



A kid is smiling ...

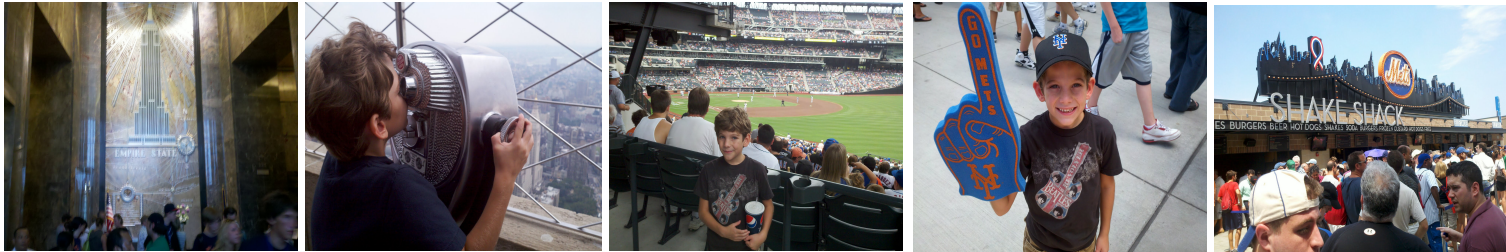
☹️ Absence of correlation, coherence and story for a stream of images

Extend both input and output dimension to a *sequence* of images and a *sequence* of sentences

Problem Statement

Objective: express an image stream to a coherent sequence of sentences

A query of image stream



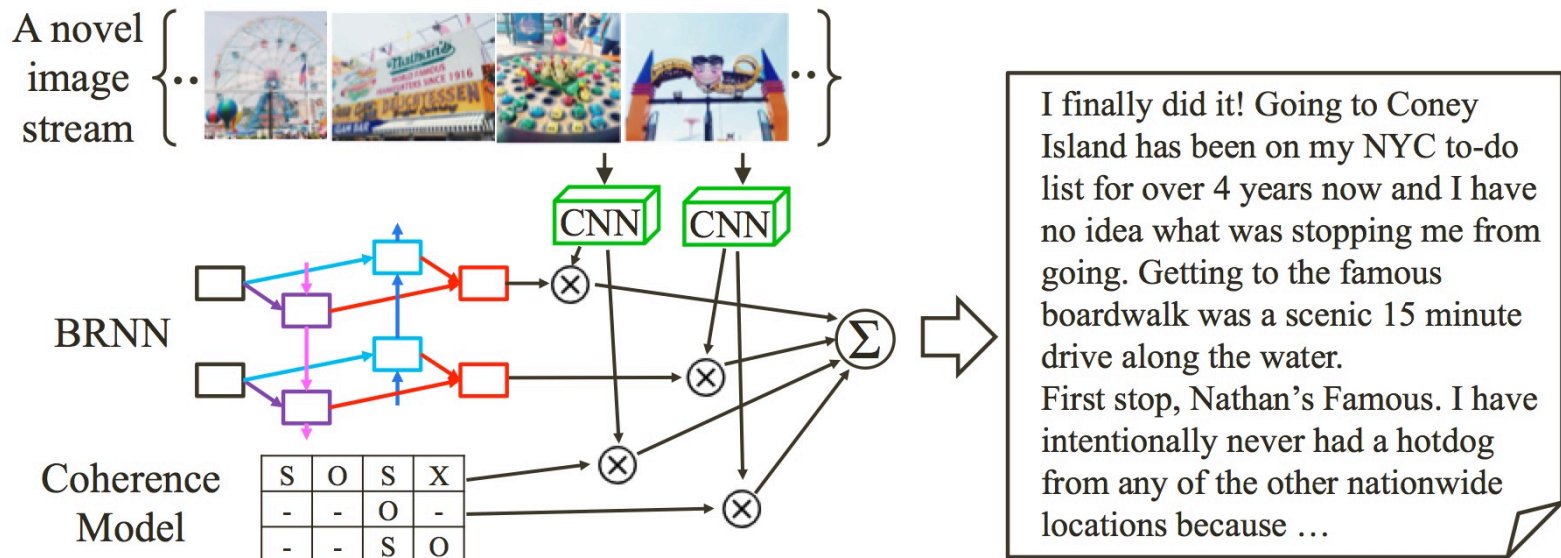
A coherent sequence of sentences

We took a couple days for family vacation in NYC to get away...
Empire state building right off the bat. Caeden is checking out the view.
Caeden's first MLB game and my first in a while MLB game...
He might be a mets fan...
Shake Shack...

Our Solution – CRCN

Coherence Recurrent Convolutional Network

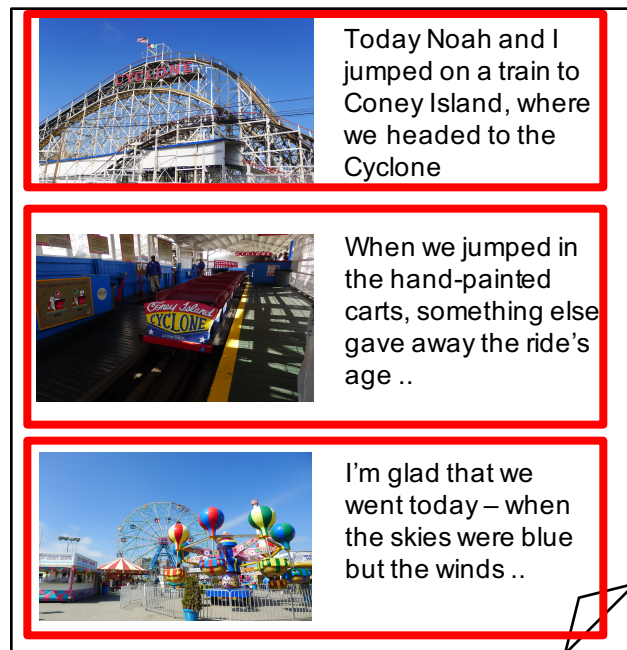
- (1) **Convolutional neural networks** for image description
- (2) **Bidirectional recurrent neural networks** for language model
- (3) **Coherence model** for a smooth flow of multi-sentences



How to Learn the Relations?

How to learn the relation between a image stream and a sequence of sentences?

A) A set of blog posts as text-image parallel training data





Blogs are written in a way of ***storytelling***

- Blog pictures are selected as the most canonical ones out of photo albums
- Informative sentences associated with pictures about location, sentiments, actors, ...

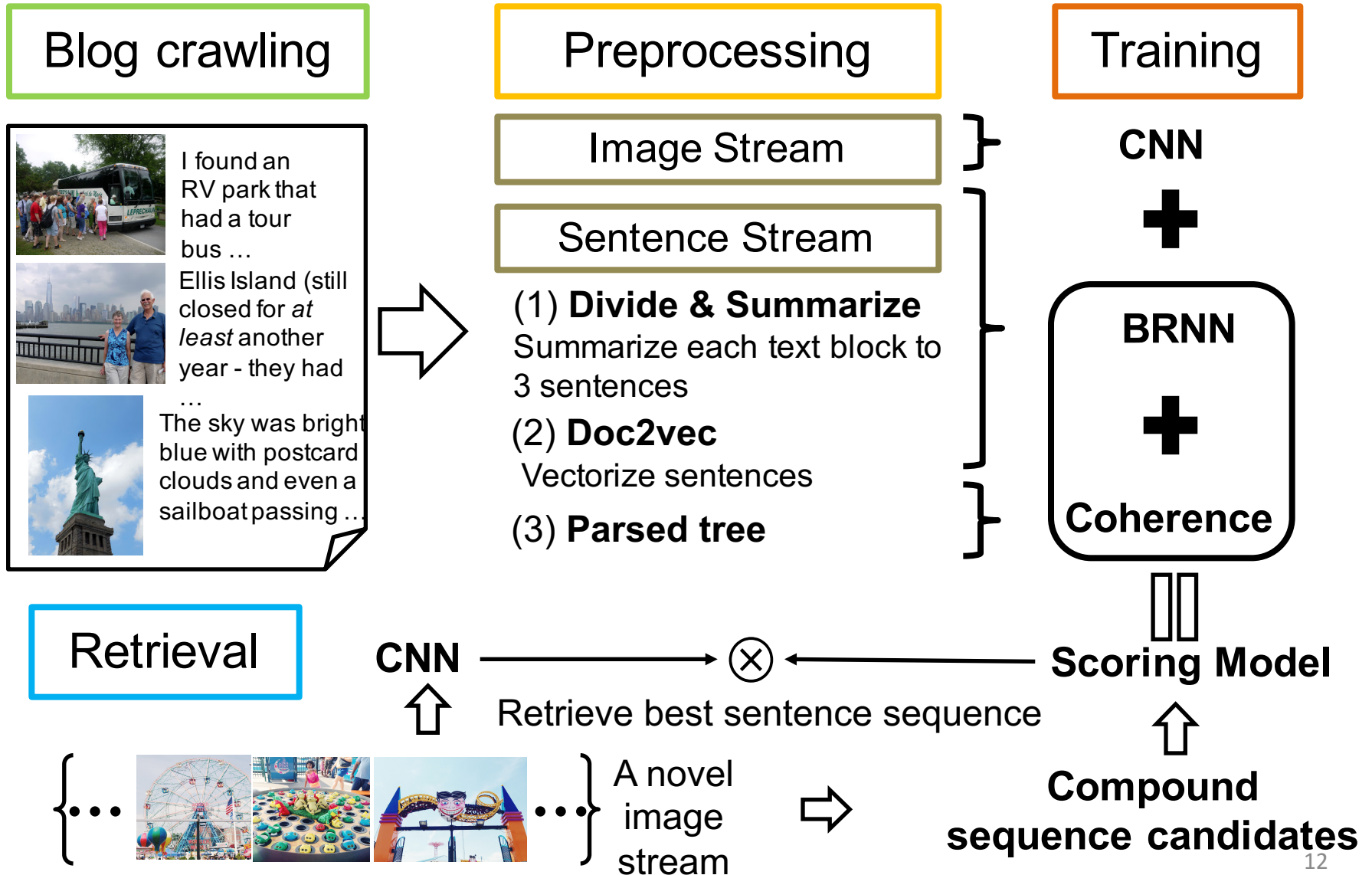
Dataset for Blog Posts

Collect from  **Blogger**™  **WORDPRESS**

Two topics	# of blog posts	# of images
	7,717	60,545
	11,863	78,467

☺ Our approach is unsupervised, and thus applicable to any domains with sufficient data

Overview of Algorithm



Preprocessing Blog Posts

Input: A set of training blogs $\mathcal{B} = \{B^1, \dots, B^L\}$ ($L = |\mathcal{B}|$)






1. Blog segmentation: $B^n \rightarrow \{(I_1^n, T_1^n), \dots, (I_m^n, T_m^n)\}$

- By distance btw text blocks and images

2. Text block summarization: $T_i^n \rightarrow P_i^n$

- Select top-3 ranked summary sentences using latent semantic analysis method

We... main street ...	 <p>We can see street sign for Time square. It is easy to find main street ...</p>
This week... photo...	 <p>There is always something interesting going on in Times Square. This week, a Parisian street artist and photographer named J R is...</p>
We... in night ...	 <p>We take selfie in the middle of Time Square. Neon signs are ... in night...</p>

Preprocessing Blog Posts

Input: A set of training blogs $\mathcal{B} = \{B^1, \dots, B^L\}$ ($L = |\mathcal{B}|$)



1. Text description: P_l^n 🗣️ p_l^n

- Extract the paragraph vector using doc2vec model

(ROOT (NP (NP (PRP ... We... main street ...

(ROOT (S (PP (IN after) This week... photo...

2. Parse tree extraction: P_l^n 🗣️ z_l^n

- Extract a parse tree from Stanford-NLP tool

(ROOT (S (NP (NP ... We... in night ...

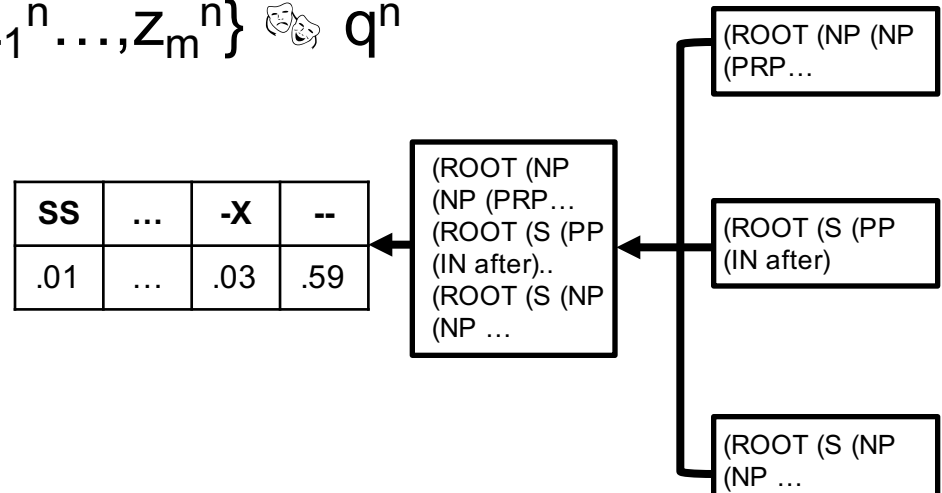
Preprocessing Blog Posts

Input: A set of training blogs $\mathcal{B} = \{B^1, \dots, B^L\}$ ($L = |\mathcal{B}|$)



Local coherence feature: $\{z_1^n, \dots, z_m^n\}$ 🧐 q^n

- Extract the local coherence feature vector from concatenated parse trees



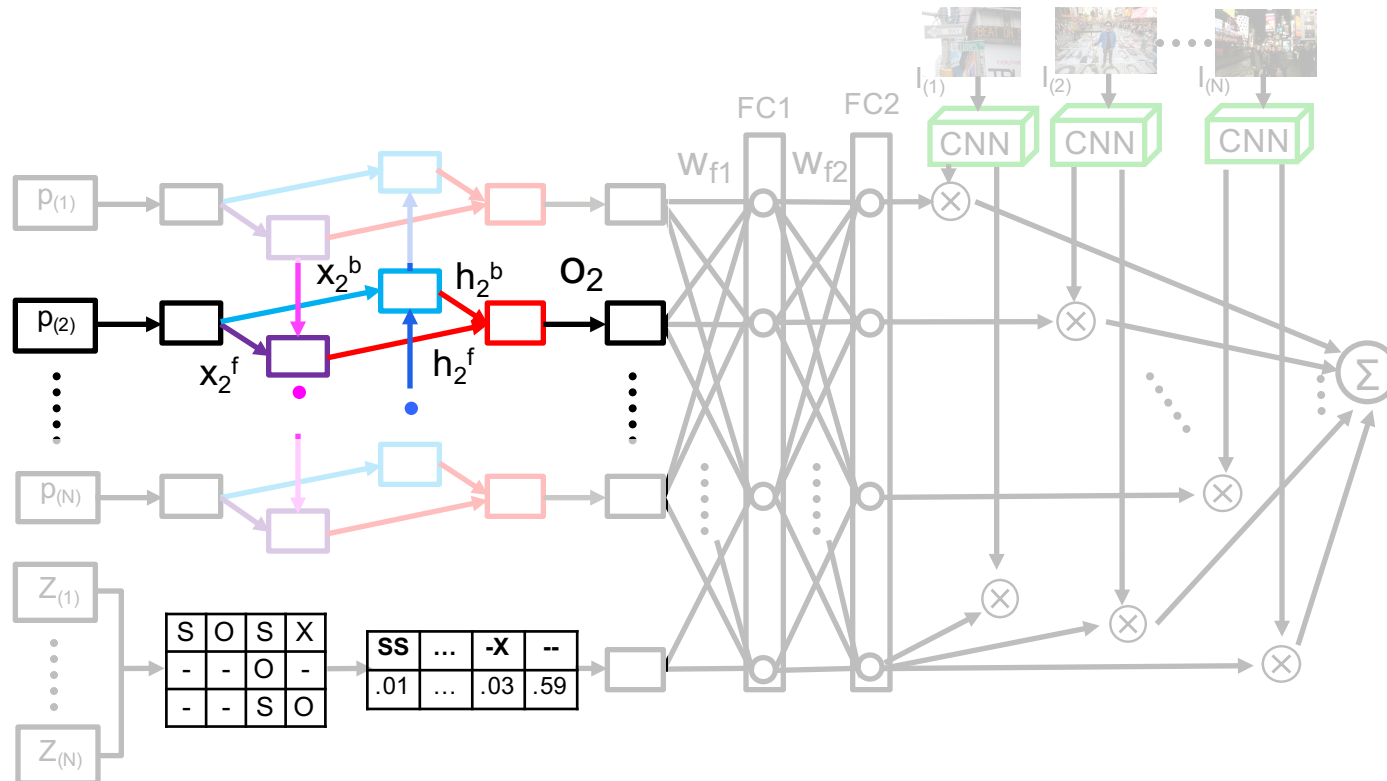
CRCN Architecture

The Bi-directional RNN part

- Represent a content flow of text sequences

$$x_t^f = f(W_i^f p_t + b_i^f); \quad x_t^b = f(W_i^b p_t + b_i^b);$$

$$h_t^f = f(x_t^f + W_f h_{t-1}^f + b_f); \quad h_t^b = f(x_t^b + W_b h_{t+1}^b + b_b); \quad o_t = W_o(h_t^f + h_t^b) + b_o$$

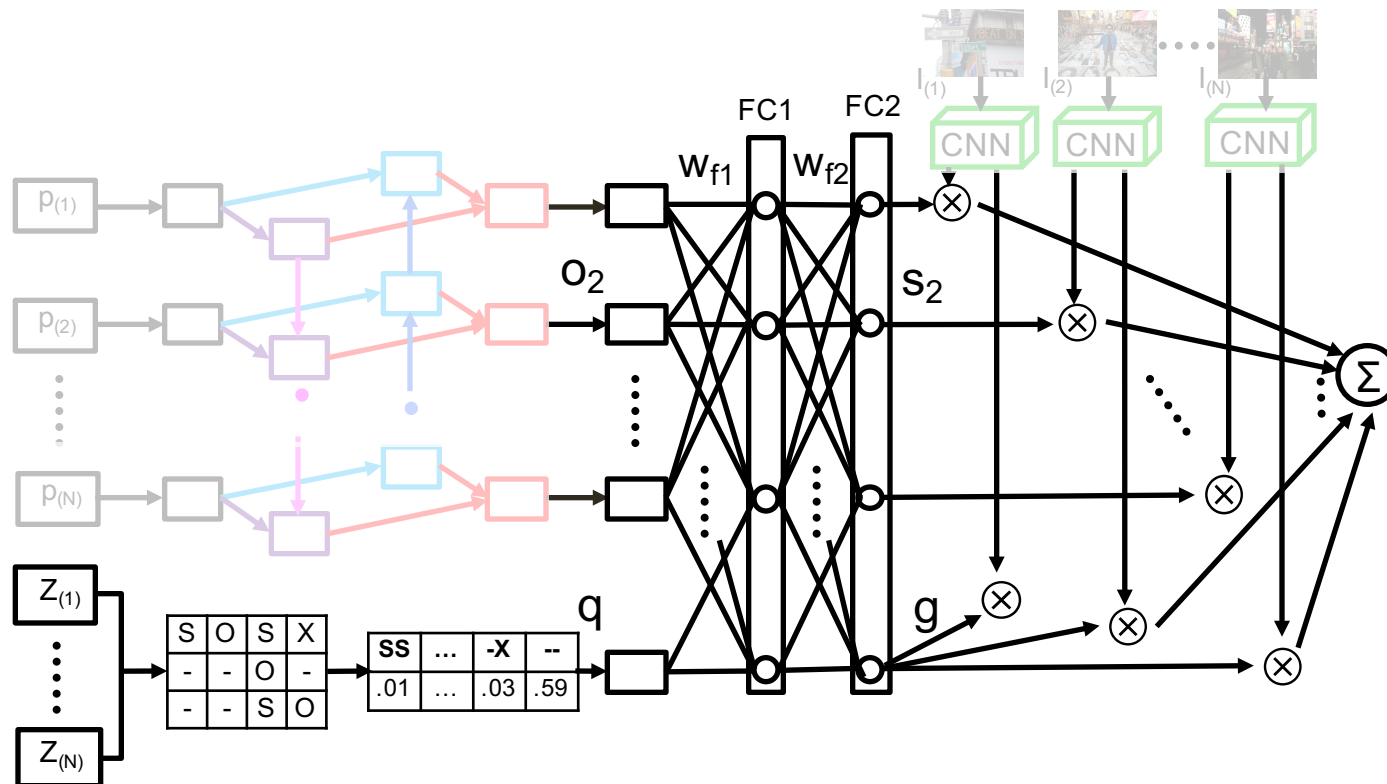


CRCN Architecture

Mixing the output of BRNN and coherence model

- Goes through two fully connected (FC) layers

$$O = [o_1|o_2|..|o_N]; \quad S = [s_1|s_2|..|s_N]; \quad W_{f2}W_{f1}[O|q] = [S|g]$$



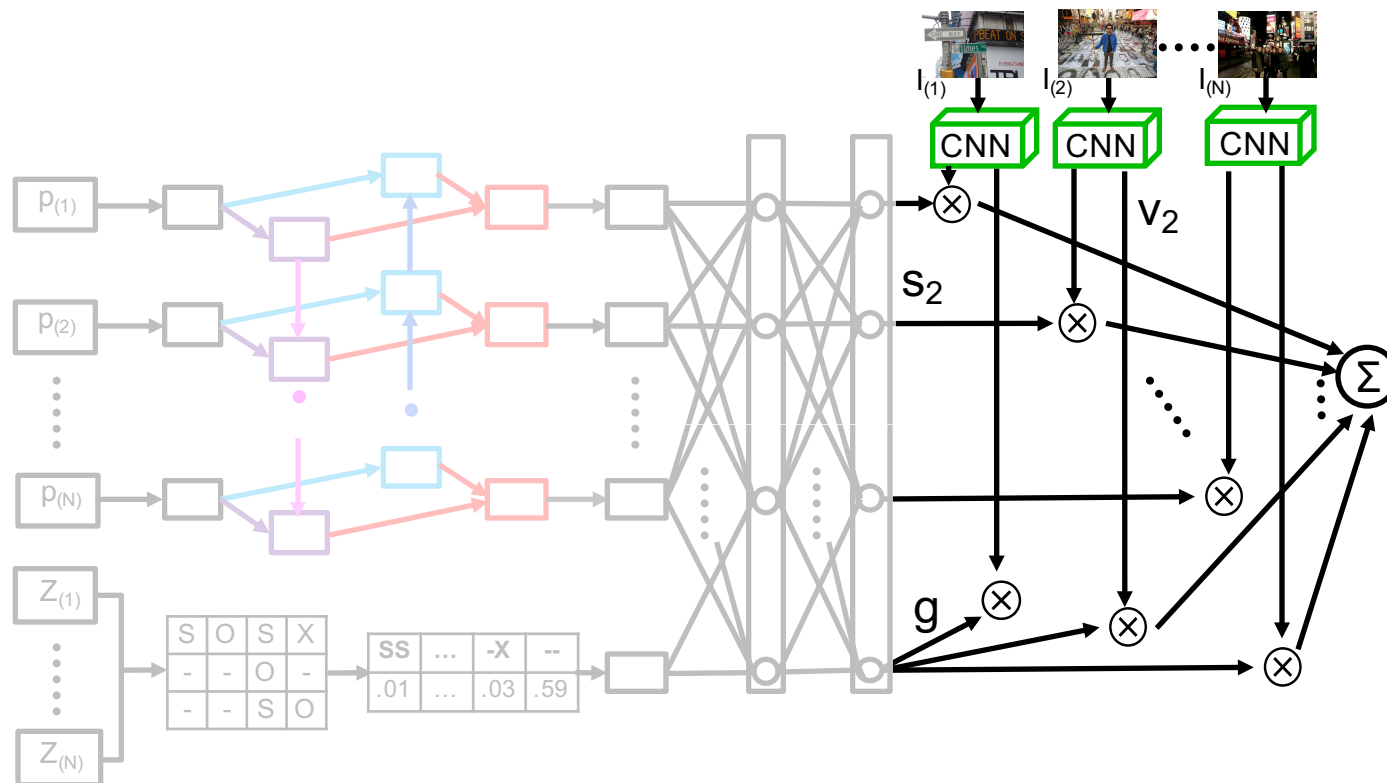
CRCN Architecture

The compatibility score btw image and sentence sequence

$$S_{kl} = \sum_{t=1 \dots N} \underbrace{s_t^k \cdot v_t^l}_{\text{Compatibility btw an image and corresponding sentence}} + \underbrace{g^k \cdot v_t^l}_{\text{Coherence term btw an image stream and a sentence sequence}}$$

Compatibility btw an image and corresponding sentence

Coherence term btw an image stream and a sentence sequence



CRCN Architecture

The cost function to train the CRCN model

$$C(\theta) = \sum_k \left[\sum_l \max(0, 1 + \underbrace{S_{kl} - S_{kk}}_{\text{Misaligned Pairs}}) + \sum_l \max(0, 1 + \underbrace{S_{lk} - S_{kk}}_{\text{Aligned Pairs}}) \right]$$

Misaligned Pairs should be few

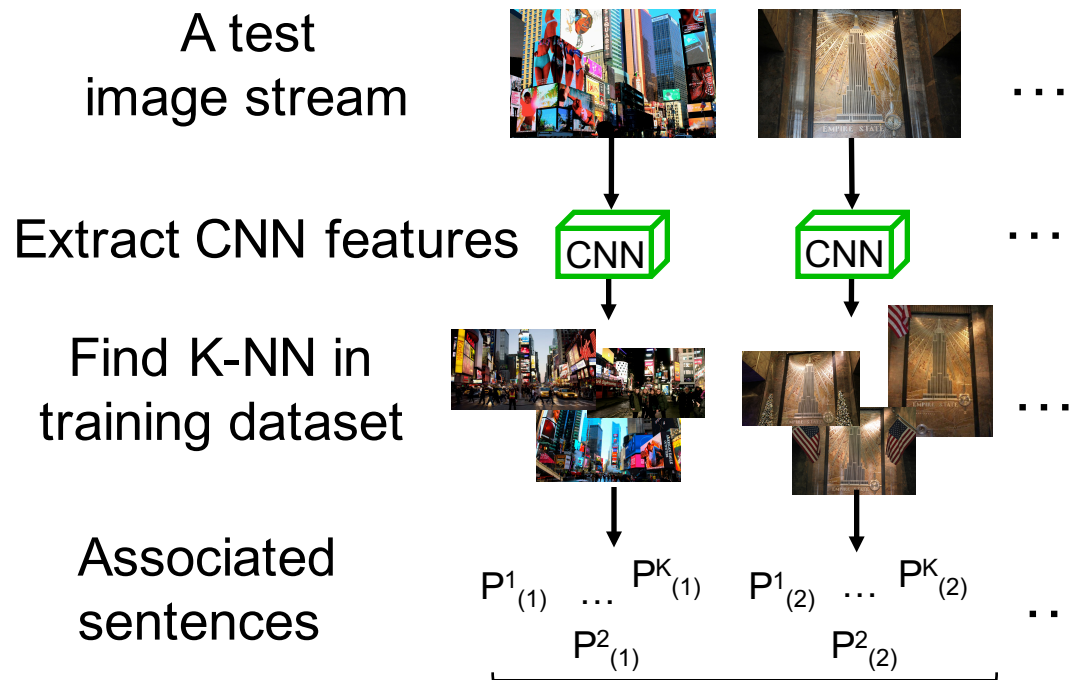
Aligned Pairs should be many

Optimization

- Use the back propagation through time (BPTT) algorithm to train our model
- Apply the stochastic gradient descent (SGD) with mini-batches of 100 data streams
- Use RMSprop optimizer among many SGD techniques

Retrieval Sentence Sequences

Retrieve *best* sentence sequences for a query image stream



Divide-conquer search strategy!

- Almost optimal!
- The **local** fluency and coherence is *required* for the **global** one



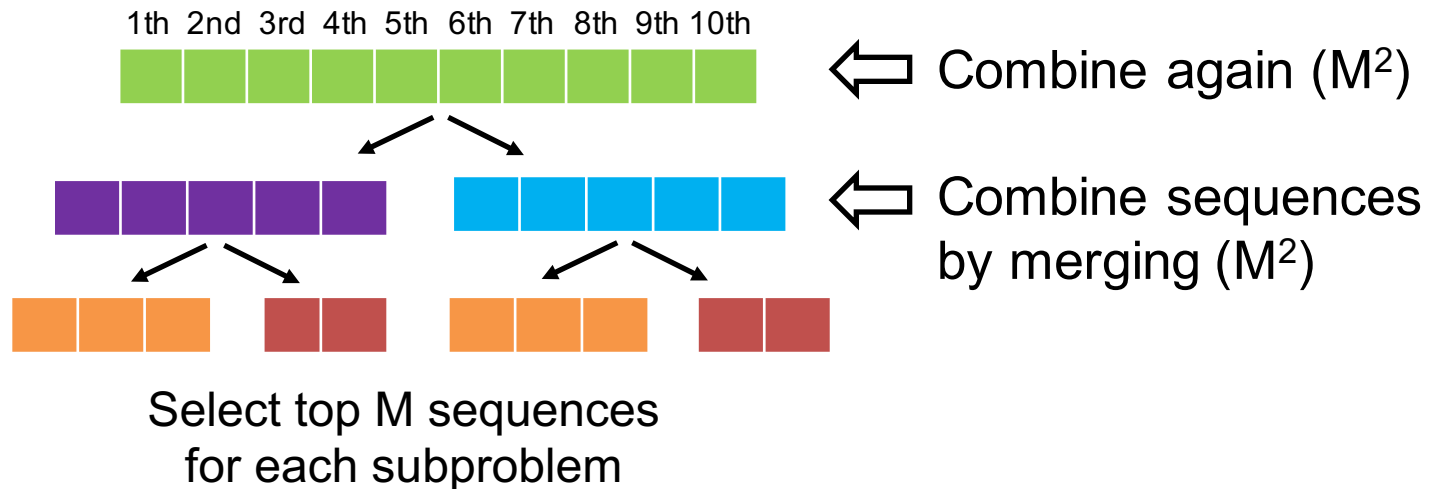
Too many combinations (K^L)

$P^1_{(1)}$	$P^1_{(2)}$	$P^1_{(3)}$...
$P^1_{(1)}$	$P^2_{(2)}$	$P^1_{(3)}$...

Retrieval Sentence Sequences

Halve the search candidate length Q times ($K^L \rightarrow K^{L/2^Q}$)

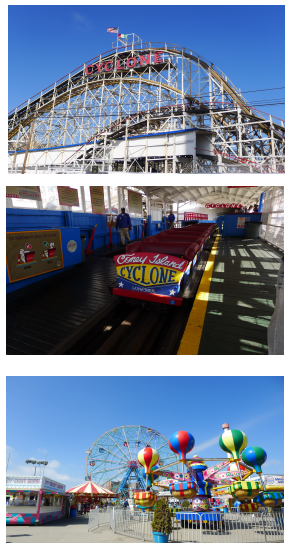
- Using the beam search idea
- Find the top-M best sequence candidates in the lowest level
- Recursively increase the candidate lengths while the max candidate size is limited to M ($M = 50$)



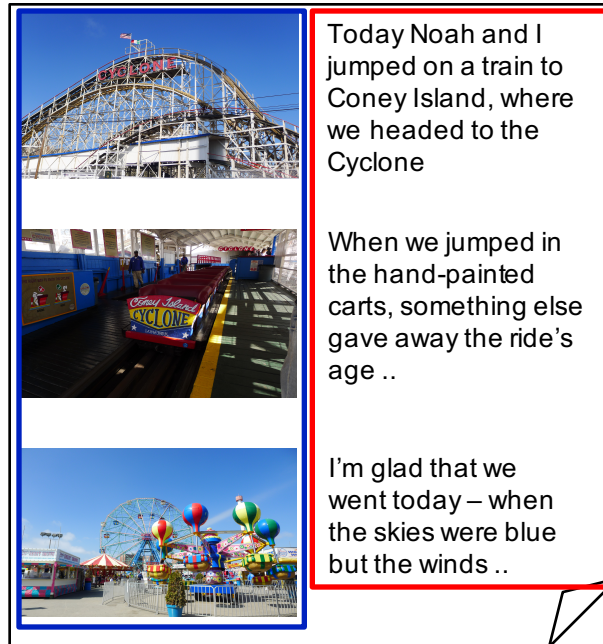
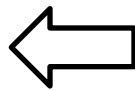
Experiment Setup

Task: for a test image stream, retrieve a sentence sequence

- 90/10% of blog posts used as the training/test set
- For a test blog,



Query
images

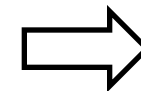


Today Noah and I
jumped on a train to
Coney Island, where
we headed to the
Cyclone

When we jumped in
the hand-painted
cars, something else
gave away the ride's
age ..

I'm glad that we
went today – when
the skies were blue
but the winds ..

Ground
Truth



Today Noah ...

When we jumped...

When the skies
were blue ...

Where we headed to...

Something else gave
away ...

When the skies
were blue ...

Today Noah ...

Something else gave
away ...

I'm glad that we
went today ...

Baselines

CNN+LSTM

- mLSTM model from [Vinyals et al. CVPR 2015]

CNN+RNN

- mRNN model from [Karpathy et al. CVPR 2015]

LBL, MLBL-B and MLBL-F

- Log-bilinear models from [Kiros et al. ICML 2014]

GloMatch

- Retrieve sentence by the nearest images on GIST and Tiny features from [Ordonez et al. NIPS 2011]

1NN

- Retrieve sentence by the nearest neighbor on CNN

Quantitative Results


Measured by both language and retrieval metrics

- BLEU, CIDEr, METEOR, R@K, and lower median rank values
- Our approach (RCN) and (CRCN) achieves the best performance

	Language metrics						Retrieval metrics			
	B-1	B-2	B-3	B-4	CIDEr	METEOR	R@1	R@5	R@10	MedRank
New York City										
(CNN+LSTM) [30]	21.31	3.65	0.57	0.14	9.1	5.73	0.95	5.24	8.57	84.5
(CNN+RNN) [9]	6.21	0.01	0.00	0.00	0.5	1.34	0.48	2.86	4.29	120.5
(MLBL-F) [12]	21.03	1.92	0.12	0.01	4.3	6.03	0.71	4.52	7.86	87.0
(MLBL-B) [12]	20.43	1.54	0.09	0.01	2.6	5.30	0.48	3.57	5.48	101.5
(LBL) [12]	20.96	1.68	0.08	0.01	2.6	5.29	1.19	4.52	7.38	100.5
(GloMatch) [21]	19.00	1.59	0.04	0.0	2.80	5.17	0.24	2.62	4.05	95.00
(1NN)	25.97	3.42	0.60	0.22	15.9	7.06	5.95	13.57	20.71	63.50
(RCN)	27.09	5.45	2.56	2.10	33.5	7.87	3.80	18.33	30.24	29.00
(CRCN)	26.83	5.37	2.57	2.08	30.9	7.69	11.67	31.19	43.57	14.00
Disneyland										
(CNN+LSTM) [30]	27.99	3.55	0.38	0.08	10.0	4.51	3.06	8.16	14.29	65.0
(CNN+RNN) [9]	6.04	0.00	0.00	0.00	0.4	1.34	1.02	3.40	5.78	88.0
(MLBL-F) [12]	15.75	1.61	0.07	0.01	4.9	7.12	0.68	4.08	10.54	63.0
(MLBL-B) [12]	15.65	1.32	0.05	0.00	3.8	5.83	0.34	2.72	6.80	69.0
(LBL) [12]	18.94	1.70	0.06	0.01	3.4	4.99	1.02	4.08	7.82	62.0
(GloMatch) [21]	11.94	0.37	0.01	0.00	2.2	4.31	2.04	5.78	7.48	73.0
(1NN)	25.92	3.34	0.71	0.38	19.5	7.46	9.18	19.05	27.21	45.0
(RCN)	28.15	6.84	4.11	3.52	51.3	8.87	5.10	20.07	28.57	29.5
(CRCN)	28.40	6.88	4.11	3.49	52.7	8.78	14.29	31.29	43.20	16.0

User Studies via AMT

Goal: Find general users' preferences between text sequences by different methods for a given photo stream

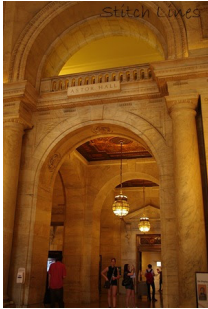
- Randomly select 100 test streams of 5 images
- Our method and one baseline predict text sequences
- Pairwise preference test via  **amazon mechanical turk**™
Artificial Artificial Intelligence

Quantitative results

- A higher number than 50% validate our approach (CRCN)
- The coherence becomes more critical as the passage is longer (4th vs 5th columns)

Baselines	(GloMatch)	(CNN+LSTM)	(MLBL-B)	(RCN)	(RCN N \geq 8)
<i>NYC</i>	92.7% (139/150)	80.0% (120/150)	69.3% (104/150)	54.0% (81/150)	57.0% (131/230)
<i>Disneyland</i>	95.3% (143/150)	82.0% (123/150)	70.7% (106/150)	56.0% (84/150)	60.1% (143/238)

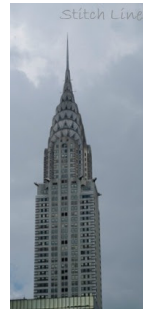
Results for NYC dataset



(1)



(2)



(3)



(4)



(5)

(CRCN) (1) One of the hallway **arches inside of the library** (2) As we **walked through the library** I noticed an exhibit called lunch hour nyc it captured my attention as I had also taken a tour of nyc food carts during my trip (3) Here is the top of the **Chrysler building** everyone's favorite skyscraper in new york. (4) After leaving the nypl we walked along 42nd st. (5) We walked down fifth avenue from **rockefeller centre checking** out the windows in saks the designer stores and eventually making our way to the **impressive new york public library**.

(RCN) (1) As you walk along in some spots it looks like the buildings are sprouting up out of the high line plants (2) Charlie and his aunt donna relax on the high line after a steamy stroll (3) However navigating the new york subway system can be like trying to find your way through the amazon jungle sans guide (4) We loved nyc! (5) Getting ready for the sunny day...putting sunscreen on.

Results for NYC dataset



(1)



(2)



(3)



(4)



(5)

(CRCN) (1) As you walk along in some spots it looks like the buildings are sprouting up out of the high line plants. (2) Getting ready for the sunny day...putting sunscreen on. (3) Well after the statue of liberty we stumbled upon the guys opening the stand and there were only about 10 people in line so we immediately jumped in. (4) We loved nyc! (5) Katherine and her dad cool off.

(CNN+LSTM) (1) Central park zoo, (2) Kai rowing trip lahnfahrt around 1999. (3) new york city. (4) new york city. (5) new york city.

Outline

- Expressing an Image Stream with a Sequence of Natural Sentences. NIPS2015.
- Automatic Photo Posting and Commenting on Social Networks via Context Memory Networks (submitted)

Automatic Photo Posting and Commenting on Social Networks via Context Memory Networks

Cesc Park

Gunhee Kim



SEOUL NATIONAL UNIV.
VISION & LEARNING

(Submitted)

Background: Automatic Writing of image Post

As photo-sharing social networks rapidly gain popularity...



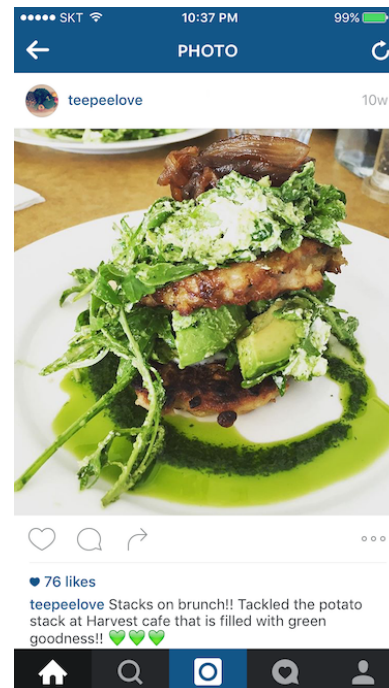
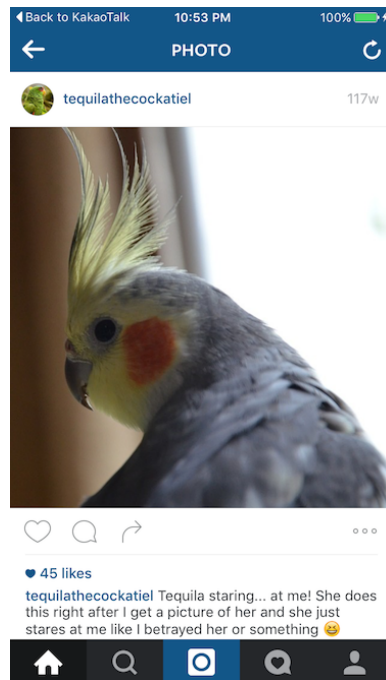
facebook



Instagram

Pinterest

General users increasingly create posts with images

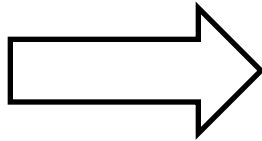


Objective 1: Photo Completion

Can we automatically complete a post for a given image?

- Crafting text is a much more cumbersome than taking a picture

Query image



Task1. Hashtag prediction

#brunch#maple#bacon#smashedavo
#fruit#granola#dragonfruit#smoothie#food

Task2. Post generation

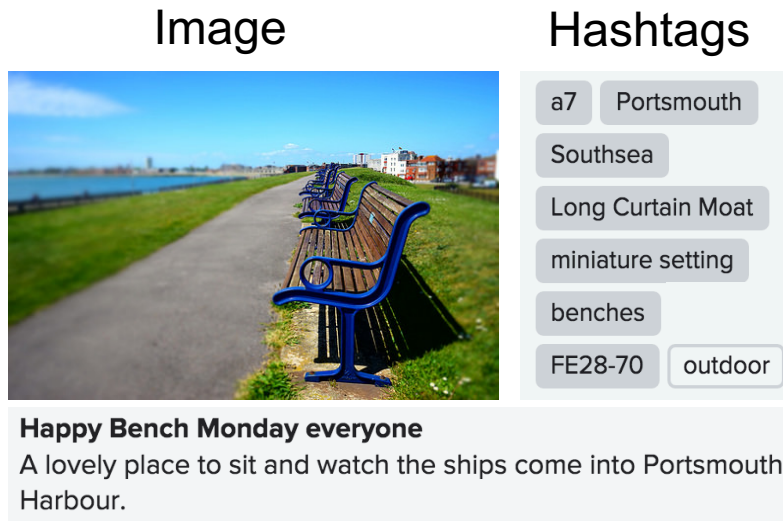
Delicious brunch! Smashed avocado bacon sandwich and fresh fruits 🥰

- Hashtag prediction: Predicting a list of appropriate hashtags
- Post generation: Generating a natural language sentence (hashtags, emoticons and normal words)

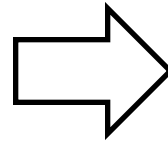
Objective 2: Automatic Commenting

Can we automatically comment a photo post?

- Assist prompt communication between users



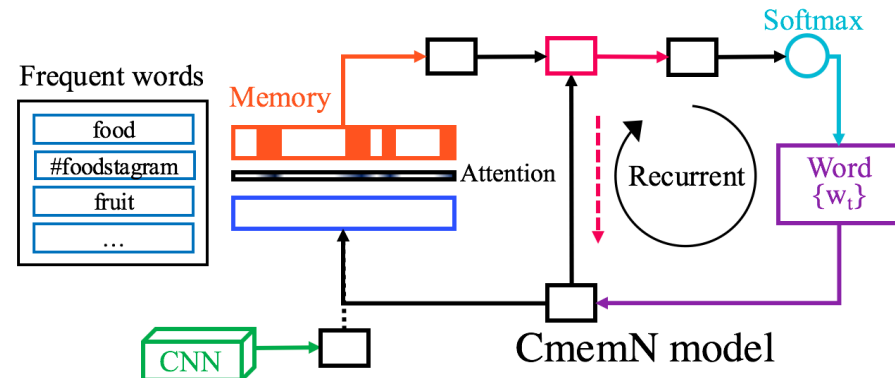
Task3. Comment generation



i like your photo it seems a placer where worth going

Context memory network (CmemN)

- Comment generation: Creating a suitable natural comments



Problem Formulation

Newly collected datasets of  Instagram 

- English post only

	# posts	# users
Instagram	11,590K	47,075

	# posts	# comments	# hashtags	# words in text	# words in comments
Flickr	42,393	72,269	14.23 (11)	107.58 (65)	10.77 (9)

Text preprocessing

- Build a dictionary of 20K most frequent words
- Separate dictionaries for the three tasks

Context Memory Network (CMemN) Model

Based on Memory networks [Graves14, Sukhbaata15]

- A multimodal neural network + A long-term memory to access

Two novelties

- Novel memory usage as a context repository for a target task
- Integration of the memory network with a recurrent model to generate a sequence of output.

Construction of context memory

- K-NN context: K-nearest training posts to the query image
- User context: Previous posts that are written by author of query

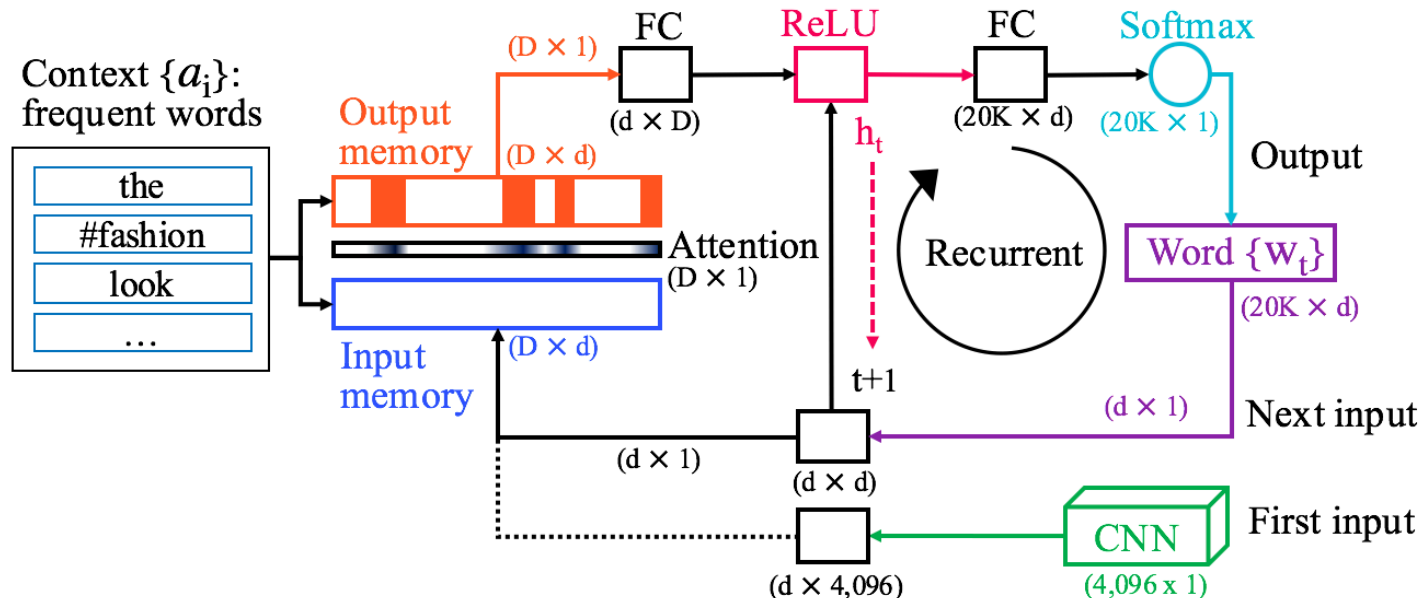
A. Graves et al. Neural Turing Machines. Arxiv 2014.

S. Sukhbaatar et al. End-to-End Memory Networks. NIPS, 2015.

Context Memory Network (CMemN) Model

How it works

- At $t=0$, the query image goes through the CNN as input vector
- The memory part decides at which parts of memory the attention turns for an input vector
- The first word is out after a series of activation over recurrent part
- The output word becomes an input at $t=1$ until $\langle \text{EOS} \rangle$ is out



Results of Hashtag Prediction

Task: Predict a list of hashtags for a given image

- 80/10/10% split for training/validation/test
- Groundtruth (GT) and prediction by our method (Pred)



(GT) #greensmoothie #plantbased #glutenfree
#smoothie #vegetarian #vegansofig #rawvegan
#vegan #healthy #fruit #rawfood #veganfoodshare
#breakfast #eatclean

(Pred) #beautiful #love #nature #vegan
#vegansofinstagram #vegan #delicious #organic
#plantbased #vegansofig #healthy #vegetarian
#crueltyfree #whatveganeat #veganfoodshare
#healthyfoodshare



(GT) #beautiful #fashion #love #todayimwearing #ootdshare
#wiwt #fashionista #linkinbio #lookbook #fashiongram #outfit
#wiw #whatiwore #mylook #ootd #fashionpost #whatiworeto
#style #fashiondiaries #currentlywearing #instastyle #lookofthe
#todaysoutfit #outfitoftheday #outfitpost #clothes #instafashion

(Pred) #beautiful #lookbook #fashion #love #ootdshare
#todayimwearing #wiwt #fashionista #fashiongram #outfit
#wiw #whatiwore #mylook #ootd #fashionpost
#whatiworetoday

Results of Post Generation

Task: Generate a natural language description for an image

- Consist of normal words + Emoji + calling user
- Groundtruth (GT) and prediction by our method (Pred)



(GT) for today s pancake saturday I made the blueberry banana bread pancakes from @username betty goes vegan delicious along with a cara cara orange

(Pred) I m going to eat this cake 🍌 it s my last giveaway n @username n hi 🙌 please enjoy the weekend



(GT) blue lagoon nphoto by @username nphoto location blue lagoon nusa ceningan island nwhile you are at jungut batu area you will come across mangrove swamps villages and this area called blue lagoon

(Pred) I love to visit the lake reflection lake at a beach a great time of year son @username n for us look at exactly

Results of Comment Generation

Task: Generate a natural language comments

- Consist of normal words + Emoji + calling user
- Groundtruth (GT) and prediction by our method (Pred)



Text: My seeds finally came in the mail! 🍌 I prepped our milk cartons to make mini greenhouses and planted seeds for garden ...

Tags: #gardening #starts #gardeningwithkids #seeds #peppers #heirloomseeds #flowers ...

(GT) @user another thing to note is that most seeds can be planted right in the garden peas lettuce spinach and other early season

(Ours) @user I m so happy you have a great day

(INN) @user took the guys outside today

(seq2seq) @user love you guys of love in hair



Text: This summer I was in Canada for the first time and it was so absolutely amazing! I did a lot of hiking in the beautiful nature...

Tags: #kathrinpreiss #lake #canada #bc# british #columbia #hiking #wandern #see #berge #forest...

(GT) beautiful scene and amazing photo

(Ours) beautiful shot of the best mountain

(INN) have a great trip this is a beautiful shot cole that reflection is awesome a beautiful capture n

(seq2seq) excellent shot love the water background

Quantitative Results – Post Generation

Methods	B-1	B-2	B-3	B-4	METEOR	CIDEr
Post Generation Task (Instagram)						
(CNN+LSTM) (Vinyals et al., 2015)	0.027	0.009	0.003	0.001	0.024	0.014
(SHOW) (Xu et al., 2015)	0.000	0.000	0.000	0.000	0.007	0.002
(1NN)	0.000	0.000	0.000	0.000	0.008	0.005
(NNthres5)	0.027	0.006	0.002	0.001	0.020	0.009
No Memory						
(CMemN-NO-500d)	0.069	0.023	0.008	0.003	0.025	0.014
User Context						
(CMemN-UC-500d400s)	0.085	0.028	0.009	0.003	0.034	0.015
(CMemN-UC-500d500s)	0.084	0.025	0.008	0.002	0.033	0.013
(CMemN-UC-600d400s)	0.082	0.026	0.009	0.003	0.035	0.017
(CMemN-UC-600d500s)	0.109	0.031	0.009	0.002	0.046	0.027
50 NN						
(CMemN-50NN-500d200s)	0.069	0.022	0.008	0.003	0.029	0.020
(CMemN-50NN-500d300s)	0.064	0.022	0.008	0.003	0.029	0.013
(CMemN-50NN-600d300s)	0.062	0.015	0.004	0.001	0.032	0.012
100 NN						
(CMemN-100NN-500d200s)	0.069	0.023	0.007	0.002	0.029	0.011
(CMemN-100NN-500d300s)	0.067	0.022	0.007	0.002	0.028	0.012
(CMemN-100NN-500d400s)	0.065	0.022	0.007	0.002	0.028	0.014

Quantitative Results – Post Generation

Methods	B-1	B-2	B-3	B-4	METEOR	CIDEr
Comment Generation Task (Flickr)						
(1NN)	0.084	0.024	0.008	0.003	0.041	0.081
(seq2seq-1) (Sutskever et al., 2014)	0.121	0.048	0.023	0.012	0.051	0.116
(seq2seq-2) (Sutskever et al., 2014)	0.118	0.045	0.020	0.010	0.050	0.109
(seq2seq-3) (Sutskever et al., 2014)	0.117	0.043	0.019	0.009	0.049	0.111
(Show) (Xu et al., 2015)	0.065	0.029	0.013	0.000	0.027	0.087
(CMemN-NO-500d)	0.108	0.039	0.015	0.006	0.047	0.109
(CMemN-PC-500d130s (Text))	0.161	0.063	0.024	0.010	0.062	0.132
(CMemN-PC-500d80s (Tag))	0.145	0.055	0.020	0.008	0.053	0.134
(CMemN-PC-500d150s (Text+Tag))	0.153	0.059	0.023	0.009	0.060	0.136

Quantitative Results – Hashtag Prediction

Methods	Jaccard
Instagram	
(CNN+LSTM) (Vinyals et al., 2015)	0.0259212
(SHOW) (Xu et al., 2015)	0.00549232
(1NN)	0.012473
(NNthres5)	0.0156202
No Memory	
(CMemN-NO-500d)	0.0119484
User Context	
(CMemN-UC-500d200s)	0.0522096
(CMemN-UC-500d300s)	0.036189
50 NN	
(CMemN-50NN-500d200s)	0.0178078
(CMemN-50NN-500d300s)	0.0125781
100 NN	
(CMemN-100NN-500d200s)	0.0225437
(CMemN-100NN-500d300s)	0.0224681

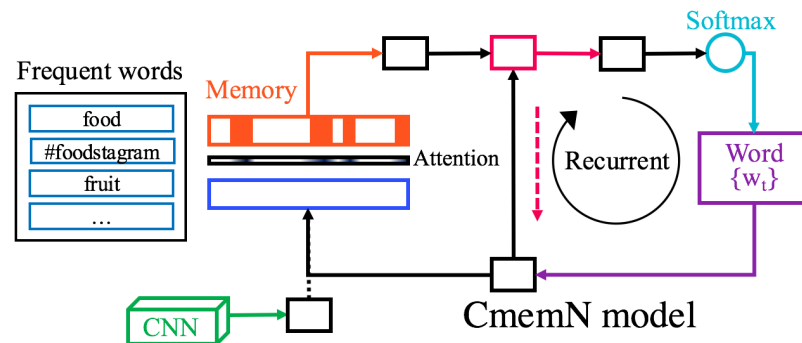
Conclusion

Achieve three post automation tasks using a single model

- Hashtag prediction + Post generation + Comment generation

Context Memory Network (CMemN) Model

- Novel memory usage as a context repository for a target task
- The memory network as a component of RNN to generate a sequence of output



Empirical verification via quantitative language metric and user studies on Instagram and Flickr dataset