Attend to You: Personalized Image Captioning with Context Sequence Memory Networks

Cesc Chunseong Park — Seoul National University
Byeongchang Kim — Seoul National University
Gunhee Kim — Seoul National University

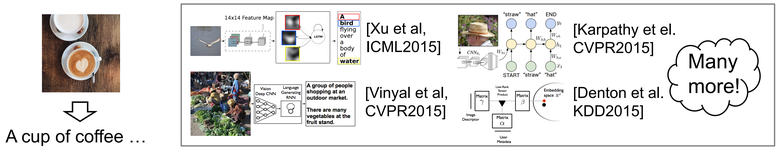
SEOUL NATIONAL UNIV. VISION & LEARNING

Code and **dataset** are available at https://github.com/cesc-park/attend2u

CVPR July 21-26 2017 HONOLULU

Motivation

- Previous image captioning creates a general description of an image



- Users craft sentences based on their experiences using their own words

User 1 *Beautiful solitude in the morning*
User 2 *The beautiful Melbourne, I love spring*
User 3 *Beautiful day for a wedding*

Extend **image captioning** to reflect user's personality

Objective

Generate **captions** from image and user's context

Query Image

User's Active Vocabulary

#eeeeeats, #londonfood #brunchfix, #smashedavo #freshbrunch, #yumyum

#eeeeeats, london, fresh salad, fruits, #yumyum sandwich, foodie,

Task1. Hashtag prediction

#freshbrunch #smashedavo #avocado #bacon #londonfood #eeeeeats

Task2. Post generation

#eeeeeats smashed avocado, bacon sandwich and fresh fruits

Our Solution: CSMN

Context **S**equence **M**emory **N**etwork

(1) **Multi-type memory cell** to condition different types of context information

(2) **CNN memory I/O structure** to jointly represent nearby ordered memory slots

(3) **Sequence generation w/o RNN** to capture long-term info without vanishing gradient

(1) : Image feature, active vocabulary, and previous words
(2) : Adopting CNN memory structure for better context understanding
(3) : Appending generated words for state-based sequence generation

InstaPIC-1.1M Dataset

- Collected from Instagram
- 1,124,815 unique posts and 6,315 unique users

Dataset	# posts	# users
caption	721,176	4,820
hashtag	518,116	3,633

Collect Instagram Posts
⇩
Preprocess Posts/Hashtags
⇩
Extract User's Active Vocabulary

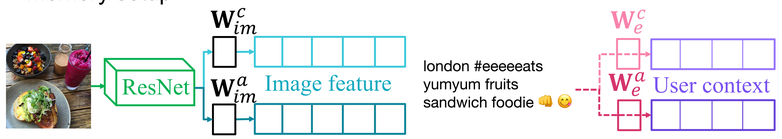
Goal: Collect refined posts from Instagram
- 27 general categories from Pinterest
- 5 < *caption length* < 15, 50 < *# posts per user* < 1,000

Goal: Build a vocabulary dictionary
- 40K for caption, 60K for hashtag

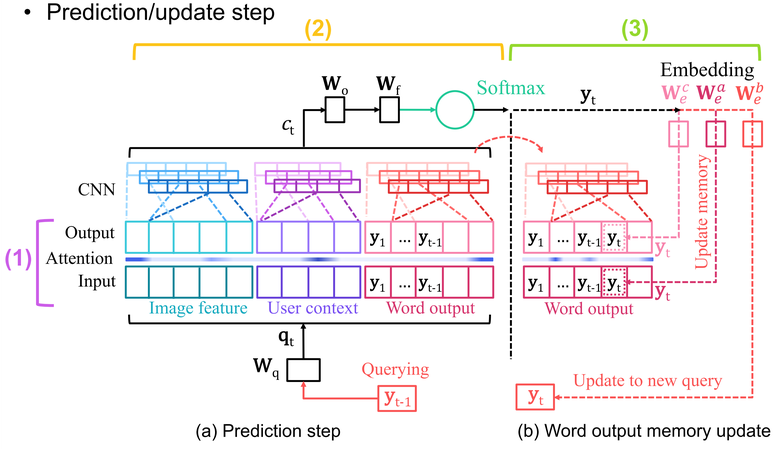
Goal: Build user's active vocabulary set
- TF-IDF weighted top-D frequent words from the user's previous posts

CSMN Architecture

- Memory setup



- Prediction/update step



(a) Prediction step (b) Word output memory update

⚠ Please see the paper for more details !

Quantitative Results

- Measured by both language and retrieval metrics

(CSMN-*): Ours and variants
(seq2seq): [Vinyals et al. NIPS15]
(ShowTell): [Vinyals et al. TPAMI16]
(AttendTell): [Xu et al. ICML15]
(1NN): 1 nearest neighbor

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

(a) Post generation

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

(b) Hashtag prediction

User Studies via Amazon Mechanical Turk

- General users' preferences over the captions created by different methods for a query image

vs. Baselines	(1NN-UsrIm)	(Showtell)	(seq2seq)
Hashtag Prediction	**67.0** (201/300)	**88.0** (264/300)	**81.3** (244/300)
Post Generation	**73.0** (219/300)	**78.0** (234/300)	**81.3** (244/300)

Qualitative Results

- Post generation examples

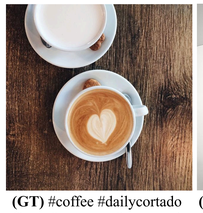
(GT) pool pass for the summer ✔
(Ours) the pool was absolutely perfect ☀
(NoCNN) the beach

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

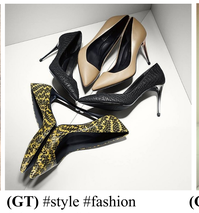
(GT) this speaks to me literarily
(Ours) I love this #quote
(Showtell) is the only thing that matters _UNK

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

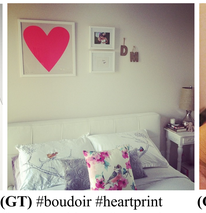
- Hashtag prediction examples



(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