AMC: Attention guided Multi-modal Correlation Learning for Image Search

Kan Chen[1], Trung Bui[2], Chen Fang[2], Zhaowen Wang[2], Ram Nevatia[1]

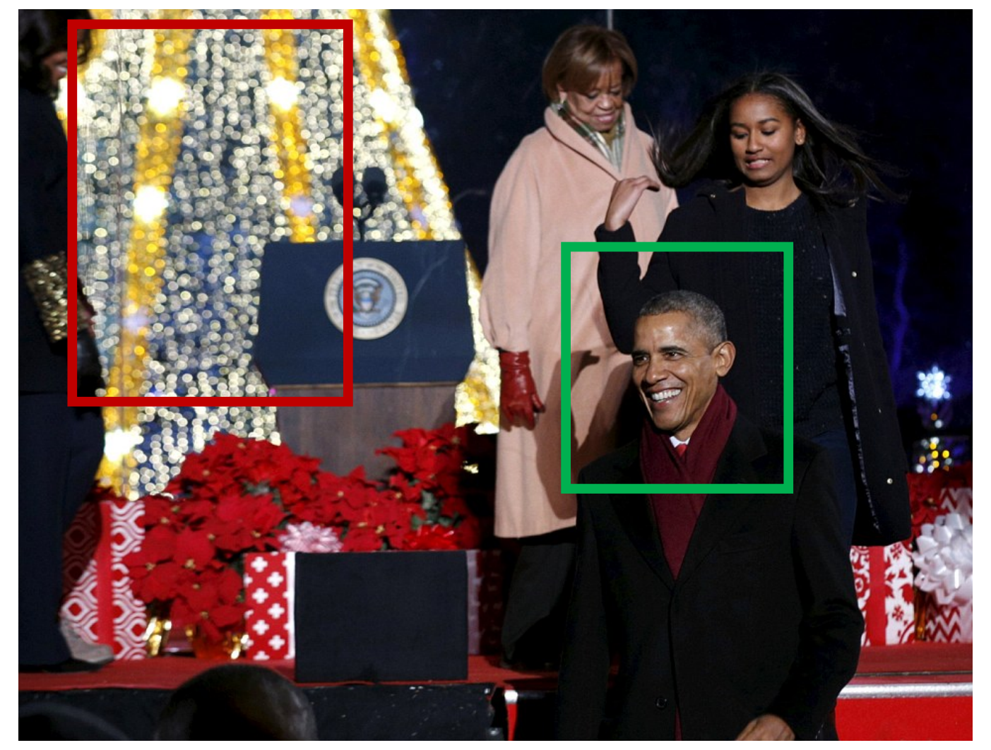
[1]University of Southern California, [2]Adobe Research

IEEE 2017 Conference on Computer Vision and Pattern Recognition

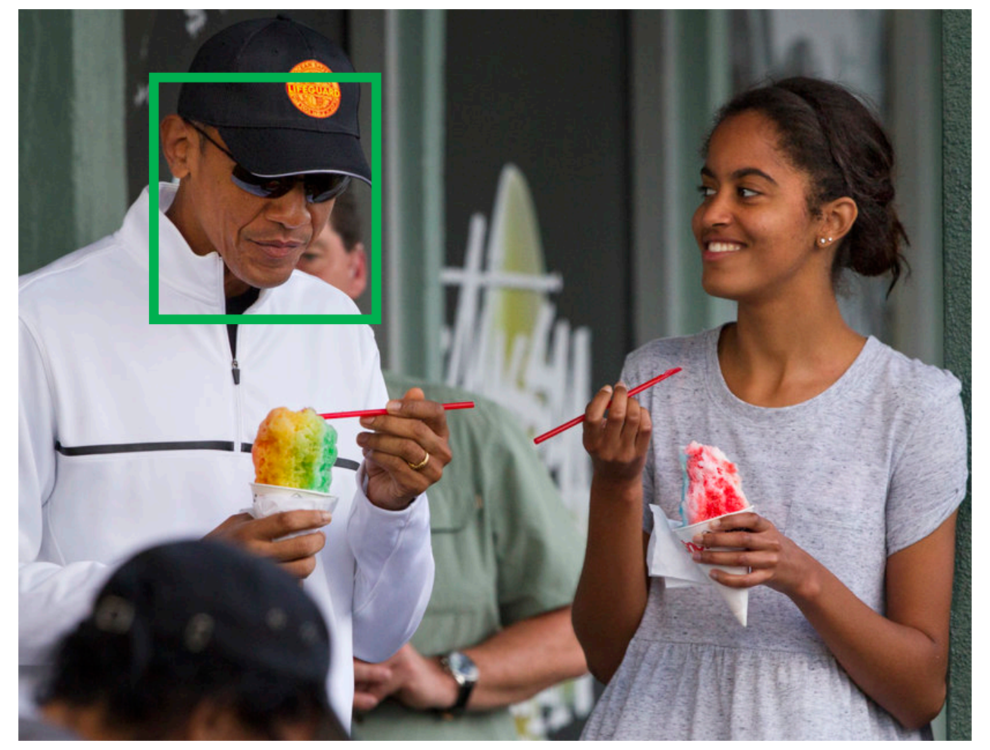
Introduction

Query1: Barack Obama

Query2: Christmas



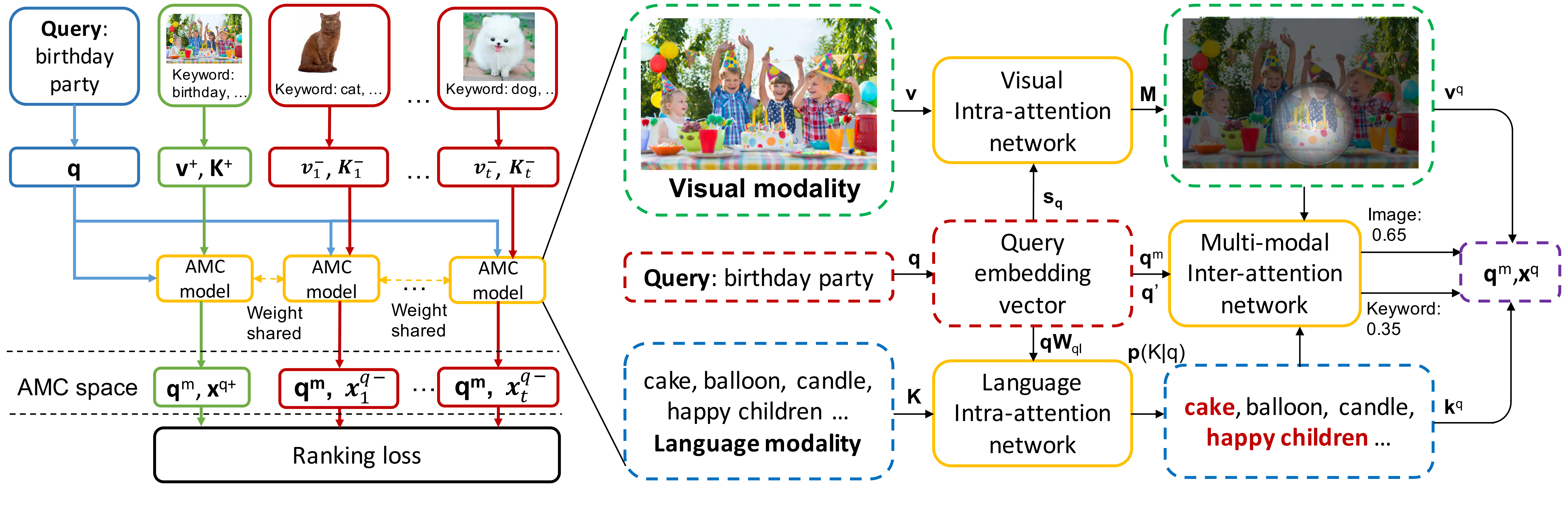
Keyword: US president, **Christmas Tree**, ceremony, family …



Keyword: **President Obama**, **Christmas holiday**, Ice-cream, Happy Malia …

- **Image Search:** Given a textual query, image search systems retrieve a set of related images by the rank of their relevance.
- **Motivation:** Nowadays, an increasing number of images on the Internet are available with associated meta data in rich modalities (e.g., titles, keywords, tags, etc.), which can be exploited for better similarity measure with queries.
- **Challenge:** Not all modalities are equally informative due to the variation in query's intent.
- **Approach:**
  - We introduce an attention mechanism to adaptively evaluate the relevance between a modality and query's intent. We consider two kinds of attention.
  - Intra-attention: an image search system should attend on the most informative parts for each modality
  - Inter-attention: an image search system should carefully balance the importance of each modality according to query's intent

Attention guided Multi-modal Correlation (AMC) Learning Framework



(a) AMC framework

(b) AMC Model details

- Given a query, images and related keywords are projected to a raw embedding space. AMC model then generates a query-guided multi-modal representation for each image. The correlation between query and image is measured by the cosine distance in the AMC space.
- AMC model consists of a visual intra-attention network (VAN), a language intra-attention network (LAN) and a multi-modal inter-attention network (MTN). VAN and LAN attend on informative parts within each modality and MTN balances the importance of different modalities according to the query's intent.

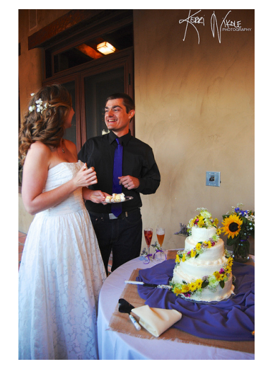
Datasets

- Two image search datasets: Clickture [1] and Adobe Stock [2]
- One caption ranking dataset: COCO Image caption dataset [3]
- We label each image with a keyword set within the above datasets (~100 keywords/image) using a keyword generation program which contains noisy tags imitating real world web image search. (left: clickture dataset, right: COCO image caption dataset)

Keyword: beautiful female, couple, woman, girl, happy, attractive, boyfriend, smiling, beauty, friends, women, people, young adult, fun, caucasian, man, male, pretty, background …

Keyword: man, people, couple, business, woman, young, office, male, smile, happy, caucasian, team, listening person, female, businessperson…

Keyword: beautiful, people, friends, women, group, young adult, shopping, fun, female, happy, attractive, men, woman, party, male, smiling …



Keyword: wedding, bride, woman, beautiful, table, couple, flower, celebration, food, white, flowers, happy, caucasian, setting, groom, home, bouquet, plate, cake, girl, adult, fun, bridal, female, love, party, vase, day, fork, breakfast …

Keyword: food, woman, breakfast, restaurant, meal, female, diet, young, tomato, hands, background, dinner, salad, orange …

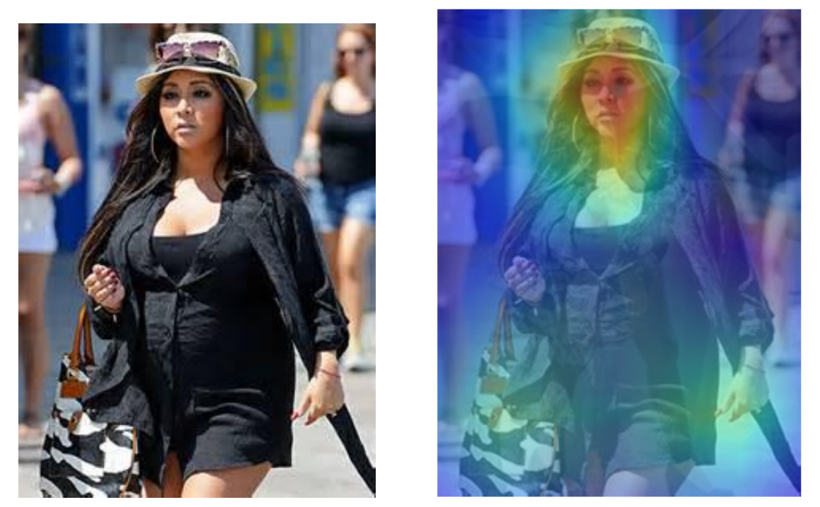
Keyword: bathroom, toilet, shower, interior, white sink, bath, modern, WC, clean, bathtub, home design, house, contemporary …

[1] T. Yao, T. Mei, and C.-W. Ngo. Learning query and image similarities with ranking canonical correlation analysis, In *ICCV*, 2015.
[2] https://stock.adobe.com [3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Mircosoft COCO: Common Objects in context. In *ECCV*, 2014.

AMC Results Visualization

Query: snooki baby bump

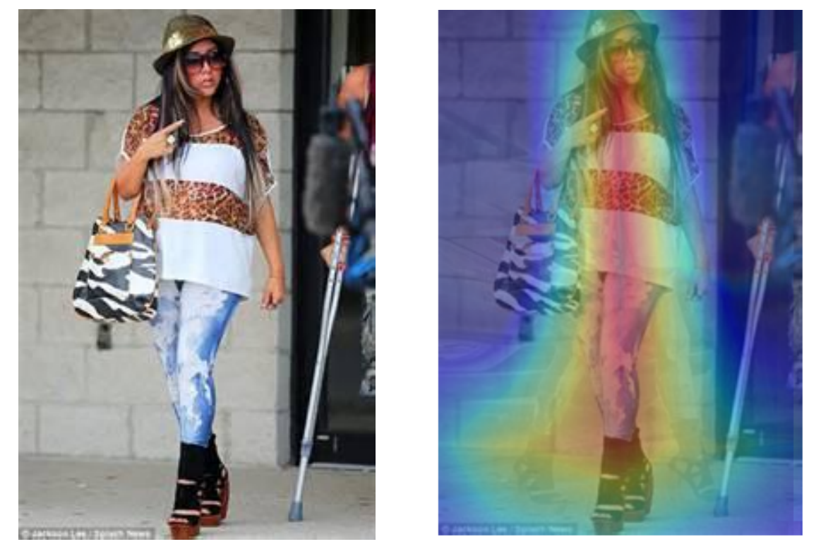
Visual: 0.6534
Language: 0.3466



transport, white, attractive, **buyer**, object, **elegance**, young, glamour, activity, arm, speaker, **woman**, shopper, **photomodel**, seated, pregnant, appearance, paint, drinking, **pretty**, smile …

Query: snooki baby bump

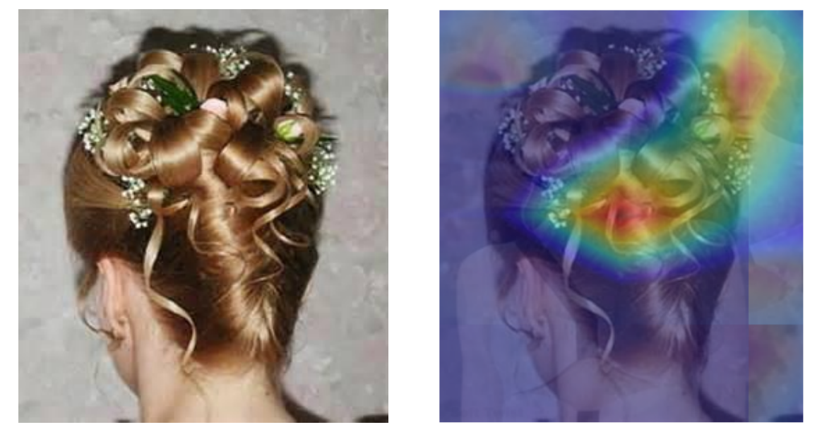
Visual: 0.7128
Language: 0.2872



attractive, art, sunglasses, breakage, **elegance**, young, industrial, computer, café, belly, **woman**, candy, **women**, camera, cars, stroll, paint, singer, american, person, tourist, arrival, people …

Query: silk twist hair styles

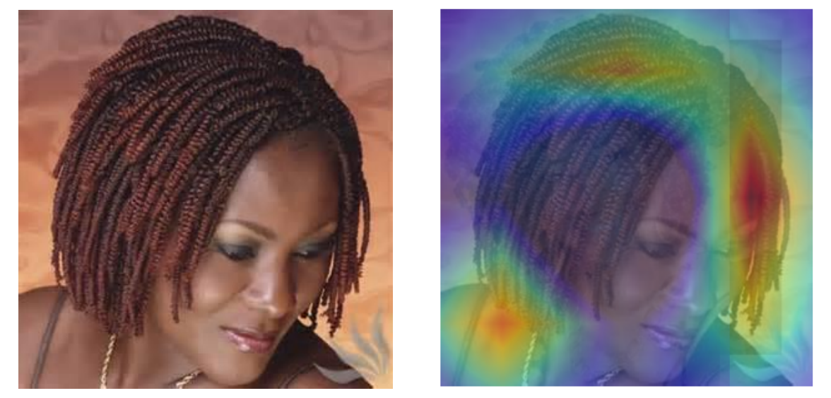
Visual: 0.5028
Language: 0.4972



nature, white, art, guard, color, rodent, event, attractive, little, heritage, dance, glamour, long, god, young, veil, **hair**, **haircut**, **woman**, eye, cut, **hairstyle** …

Query: silk twist hair styles

Visual: 0.5631
Language: 0.4369



white, **hair**, lips, shawl, human, **attractive**, expression, glamour, lovely, american, young, woman, **woman**, eye, makeup, **hairstyle** …

Quantitative Results

Approach	5	10	15	20	25
MB	0.5643	0.5755	0.5873	0.5918	0.5991
DSSM-Key	0.5715	0.5745	0.5797	0.5807	0.5823
DSSM-Img	0.6005	0.6081	0.6189	0.6192	0.6239
RCCA	0.6076	0.6190	0.6293	0.6300	0.6324
Key_{ATT}	0.5960	0.6054	0.6168	0.6204	0.6241
Img_{ATT}	0.6168	0.6233	0.6308	0.6350	0.6401
Img_{ATT}-Key_{ATT}-LF	0.6232	0.6254	0.6344	0.6376	0.6444
AMC Full	**0.6325**	**0.6353**	**0.6431**	**0.6427**	**0.6467**

Table 1: Image Search under NDCG@k metric

Approach	P@5	P@k	MAP	MRR	AUC
MB	0.5615	0.6372	0.7185	0.7564	0.6275
DSSM-Key	0.5431	0.6756	0.6969	0.7884	0.5508
DSSM-Img	0.5835	0.6705	0.7308	0.7773	0.6455
RCCA	0.5856	0.6778	0.7332	0.7894	0.6384
AMC Full	**0.6050**	**0.7069**	**0.7407**	**0.8067**	**0.6727**

Table 2: Image Search under various metrics

Approach	R@1	R@5	R@10
Random	0.1	0.5	1.0
DVSA [14]	38.4	69.9	80.5
FV [18]	39.4	67.9	80.5
m-RNN-vgg [26]	41.0	73.0	83.5
m-CNN_{ENS} [25]	42.8	73.1	84.1
Kiros *et al.* [16]	**43.4**	**75.7**	85.8
Skip-Vgg [17]	33.5	68.6	81.5
Skip-Vgg-Key-LF	34.2	69.3	82.0
AMC-Vgg	37.0	70.5	83.0
Skip-Res	39.5	73.6	86.1
Skip-Res-Key-LF	40.1	74.2	86.5
AMC-Res	41.4	75.1	**87.8**

Table 3: Caption ranking under R@k metric, AMC achieves competitive results on COCO Image Caption Ranking dataset