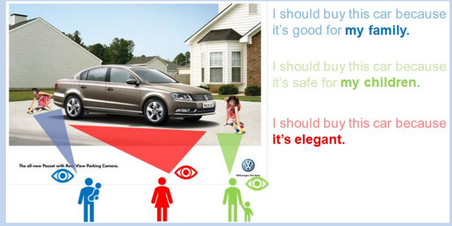
Cross-modality personalization for retrieval

Nils Murrugarra-Llerena and Adriana Kovashka
Department of Computer Science - University of Pittsburgh

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Motivation

- Existing methods annotate images using **only image pixels**. However, our perception is affected by our **personality,** experience, and bias.
- Thus, learning jointly gaze, personality and image captioning can be beneficial.



Overview

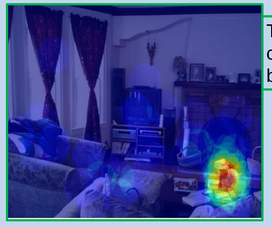
- We develop a model for **cross-modality personalized retrieval**.
- Our method combines **content** and **style** constraints. Content encourages similarity of the samples (e.g. captions) provided on the same image. In contrast, style encourages similarity of the samples from the same user.
- Our approach outperforms three different baselines on two datasets.

Related work

- We model the relationship between **two channels affected by personality** (gaze and captions). In contrast, prior work only considers relationships between captions and personality (Park et al. 2017, Veit et al. 2018) or gaze and personality/sentiment (Fan et al. 2018, Xu et al. 2017).
- A variant of our method exploits privileged information at training time.

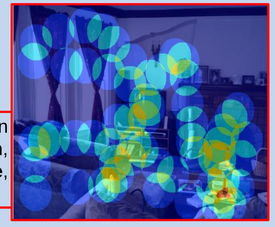
Cross-modality dataset

- We collect caption and gaze data, along with responses to personality surveys, for images in **two datasets (Ads and COCO)** using Amazon Mechanical Turk.
- We collect more than 4,000 annotations, 900 unique images and 270 tasks.



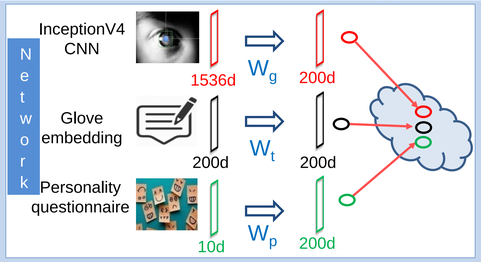
The classic den complete with fireplace but **without the family.**

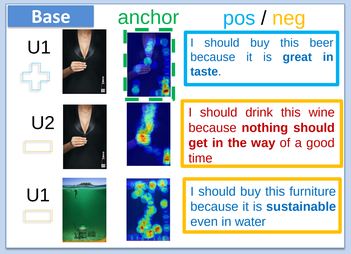
Living room with an **entertainment system**, **fireplace**, **coffee table**, and **furniture**.



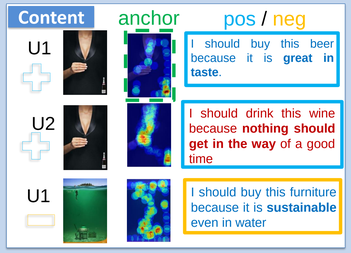
- To ensure quality, we use **validation images** where it is clear where a gaze map should reasonably be located (e.g. objects on plain background).
- Our dataset is available on: **www.cs.pitt.edu/~nineil/crossmod/**

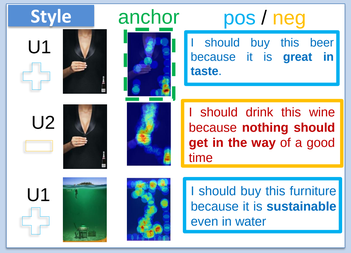
Approach





$$L = \max(d(a,p)-d(a,n)+margin,0)$$

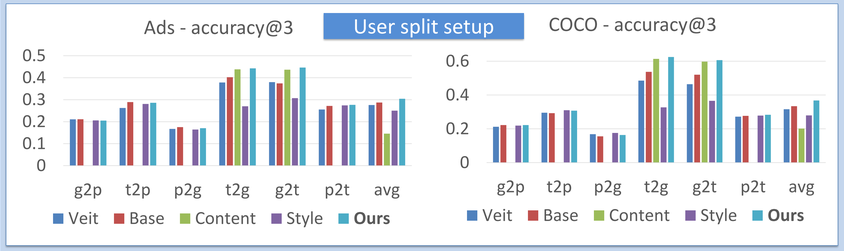




Evaluation

Our method combines base, content, and style constraints. We compare **Ours** with four different baselines using **top-3 accuracy**:

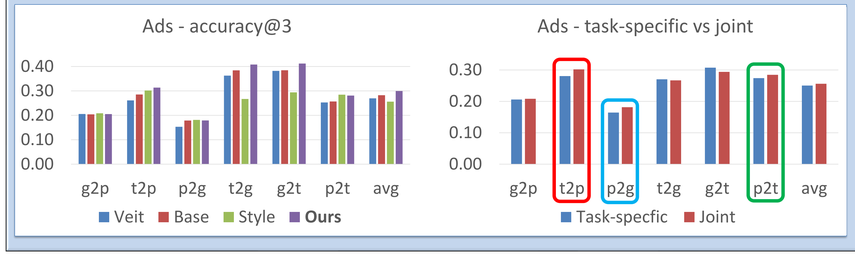
- **Base** (metric learning, Faghri et al. BMVC 2018).
- **Content** and **Style** are components of our method.
- **Veit** et al. CVPR 2018 (matrix factorization).



- Our random split setup shows higher performance. In Ads, *g2p* performance is **0.2375** compared to 0.2051; and *g2t* performance is **0.6519** vs 0.4463.

Evaluation (cont'd)

- Previously, we train three task-specific networks (e.g. one for *t2g/g2t*).
- We next train jointly for gaze, text, personality, and see benefit for most tasks.



Qualitative results

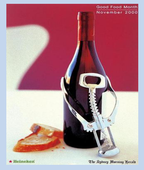
We show how distinct the samples provided by different users are, and how consistent the differences between users are with user personality from surveys.

I should consider a wide variety of travel options because **my family deserves it.**

I should bring my family to Dubai because **it is a family friendly vacation spot**

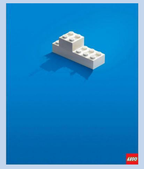
I should take more time to put myself together because it will allow me to find more **appropriate companionship.**

I should buy Gucci Guilty perfume because it will make me a **sexier person**



I should buy a bottle of wine because I **deserve** a treat today.

I should drink Heineken because it **goes along with** fine food.



I should buy this lego piece because **I enjoy** legos.

I should play with Legos because they **allow you to be creative**.

Conclusion and acknowledgements

- We developed an approach for retrieving samples that capture different user **perceptions** of the same image across modalities.
- We combine our **style** constraints with standard **content** constraints.
- Learning **jointly** gaze, captions and personality is better than learning in isolation.
- We are grateful for NSF CRII award 1566270, Google Faculty Research Awards.