Human Attention in VQA:

Do Humans and Deep Networks Look at the Same Regions?

Virginia Tech, Facebook AI Research, Georgia Institute of Technology

EMNLP 2016

2020.06.15

Hanyang univ. AILAB 정지은

Index

1. Introduction

Introduction

- **Visual Question and Answering (VQA) :**

  이미지와 그 이미지에 대한 질문(Question)이 주어졌을 때, 해당 질문에 맞는 올바른 답변(Answer)을 하는 task

Is something under the sink broken?	yes yes yes	no no no
What number do you see?	33 33 33	5 6 7

Can you park here?	no no no	no no yes
What color is the hydrant?	white and orange white and orange white and orange	red red yellow

Introduction

- Figure 1 shows **human attention maps** on the same image for two different questions.

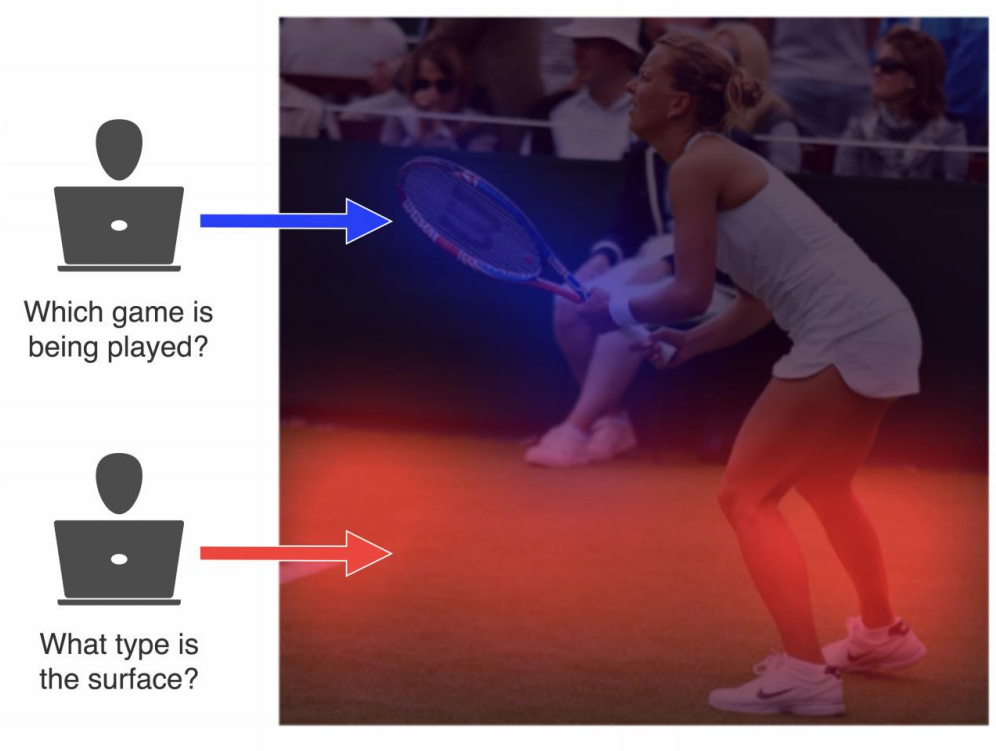


Figure 1: Different human attention regions based on question. (best viewed in color)

Contributions

1. Human attention maps를 수집하기 위한 새로운 인터페이스 설계

ㄴ 기존의 대규모 VQA dataset => VQA-HAT(Human ATtention) 데이터세트 구축 및 공개

Contributions

1. Human attention maps를 수집하기 위한 새로운 인터페이스 설계

ㄴ 기존의 대규모 VQA dataset => VQA-HAT(Human ATtention) 데이터세트 구축 및 공개

2. Vision attention mechanisms 에 대한 명확한 해석

- 사람은 VQA할 때 이미지의 어느 영역에 집중하는가?

- VQA 모델의 attention mechanisms은 사람과 비슷하게 작동하는가?

Contributions

1. Human attention maps를 수집하기 위한 새로운 인터페이스 설계

ㄴ 기존의 대규모 VQA dataset => VQA-HAT(Human ATtention) 데이터세트 구축 및 공개

2. Vision attention mechanisms 에 대한 명확한 해석

- 사람은 VQA할 때 이미지의 어느 영역에 집중하는가?

- VQA 모델의 attention mechanisms은 사람과 비슷하게 작동하는가?

ㄴ VQA SOTA 모델들에 대한 정량적, 정성적 평가

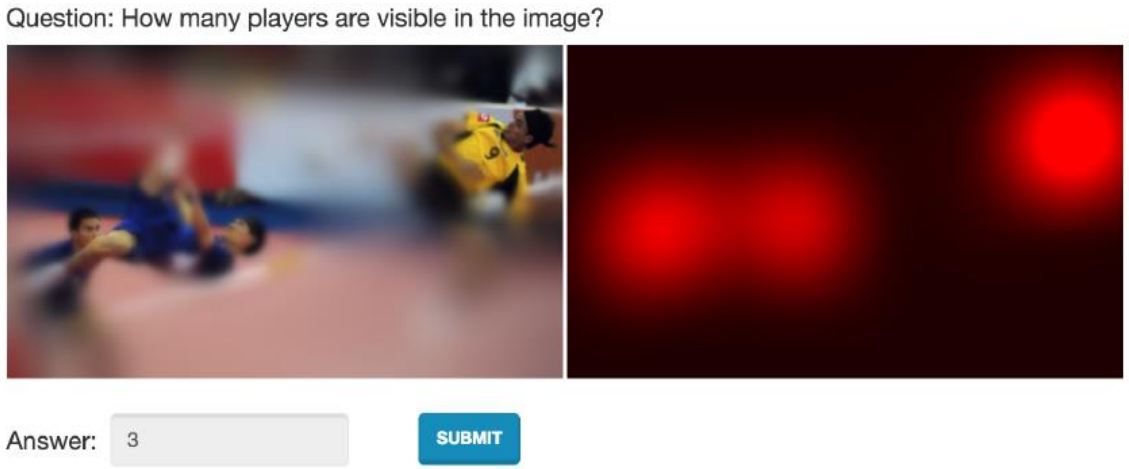
ㄴ 정량평가 : Rank-order correlation

ㄴ 정성평가 : Visualizations

2. VQA-HAT (Human ATtention) Dataset

VQA-HAT (Human ATtention) Dataset

- **AMT**
- **Deblurring** exercise for answering visual questions
  - **Click and drag**
  - **Sharpening is gradual**
  - Question-image pairs from the VQA dataset : 58,475 train / 1,374 val
  - Conducted approximately 20000 Human Intelligence Tasks (HITs) on AMT, 800 unique workers.



VQA-HAT (Human ATtention) Dataset

- Figure 2 shows examples of **collected human attention maps**

(a) (b) (c)

Figure 2: (a-c): Column 1 shows deblurred image, and column 2 shows human attention map.

VQA-HAT (Human ATtention) Dataset

- Human attention maps 의 위치를 6가지로 클러스터링 했을 때, 각 위치에 주로 나타났던 질문들의 예시.

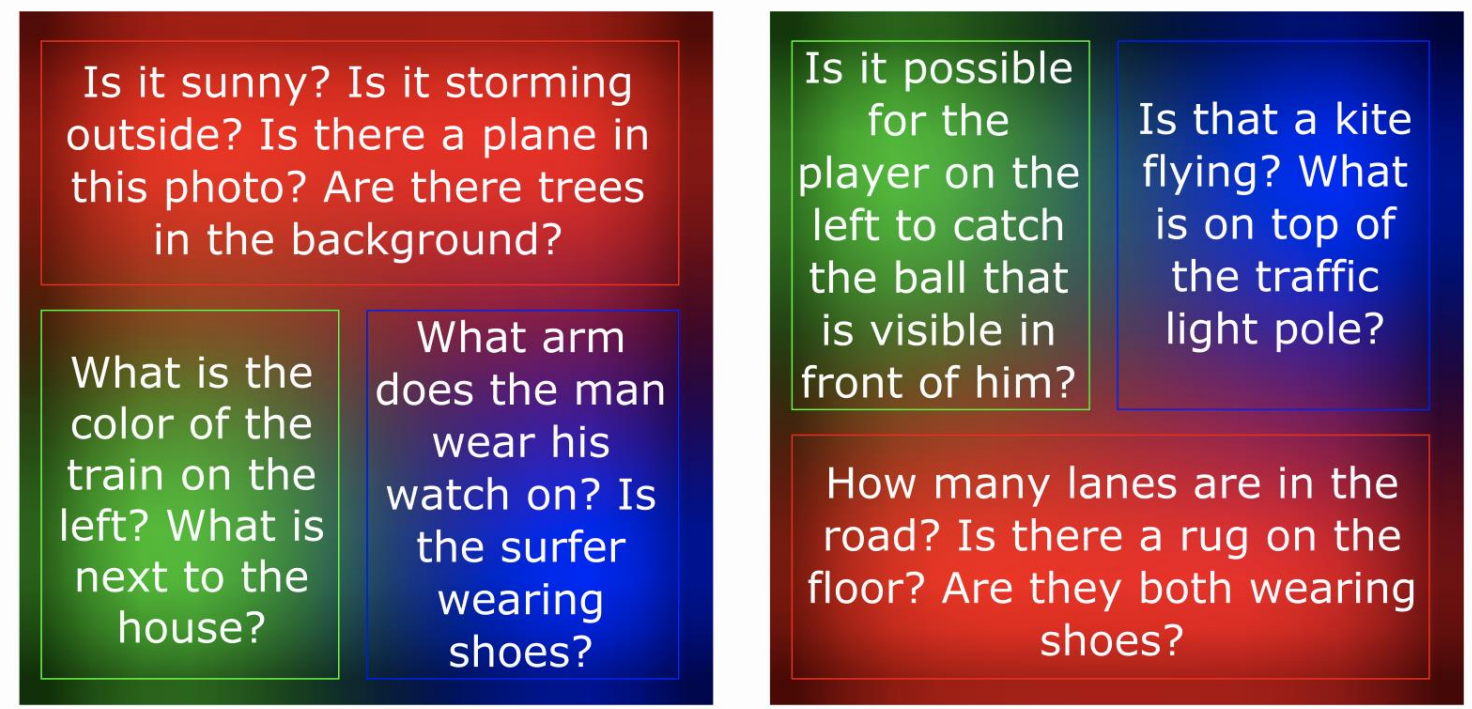


Figure 3

3. Experiments

1. Main questions

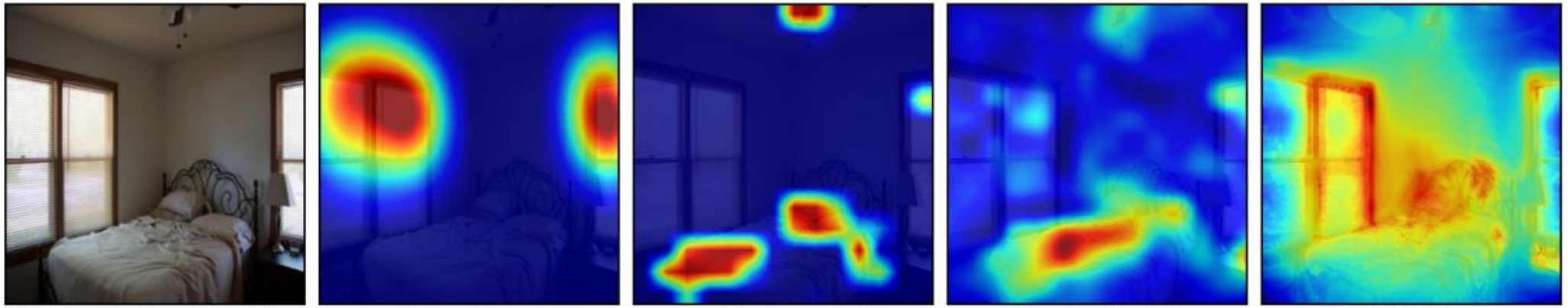
Do neural networks look at the same regions as humans to answer a visual question?

2. VQA Attention Models

- Stacked Attention Network (SAN) (Yang et al., 2016) with two attention layers (SAN-2)3
- Hierarchical Co-Attention Network (HieCoAtt) (Lu et al., 2016)
  - with word-level (HieCoAtt-W),
  - with phrase-level (HieCoAtt-P)
  - with question-level (HieCoAtt-Q) attention maps

3. Comparison Metric : Rank Correlation

1. Machin-generated, human-generated Attention map을 14 x 14 로 re-scaling
2. Spatial attention에 따라 픽셀의 순위(rank)를 정한 ranked list를 만든다.
3. 두개의 ranked lists 사이의 correlation을 구한다.



Machin-generated = [1, 2, **5**, 3, 4]

Human-generated = [1, **5**, 2, 3, 4]

Human Attention Maps vs Unsupervised Attention Models

Model	Rank-correlation
SAN-2 (Yang et al., 2016)	0.249 ± 0.004
HieCoAtt-W (Lu et al., 2016)	0.246 ± 0.004
HieCoAtt-P (Lu et al., 2016)	0.256 ± 0.004
HieCoAtt-Q (Lu et al., 2016)	0.264 ± 0.004
Random	0.000 ± 0.001
Judd et al. (Judd et al., 2009)	0.497 ± 0.004
Human	0.623 ± 0.003

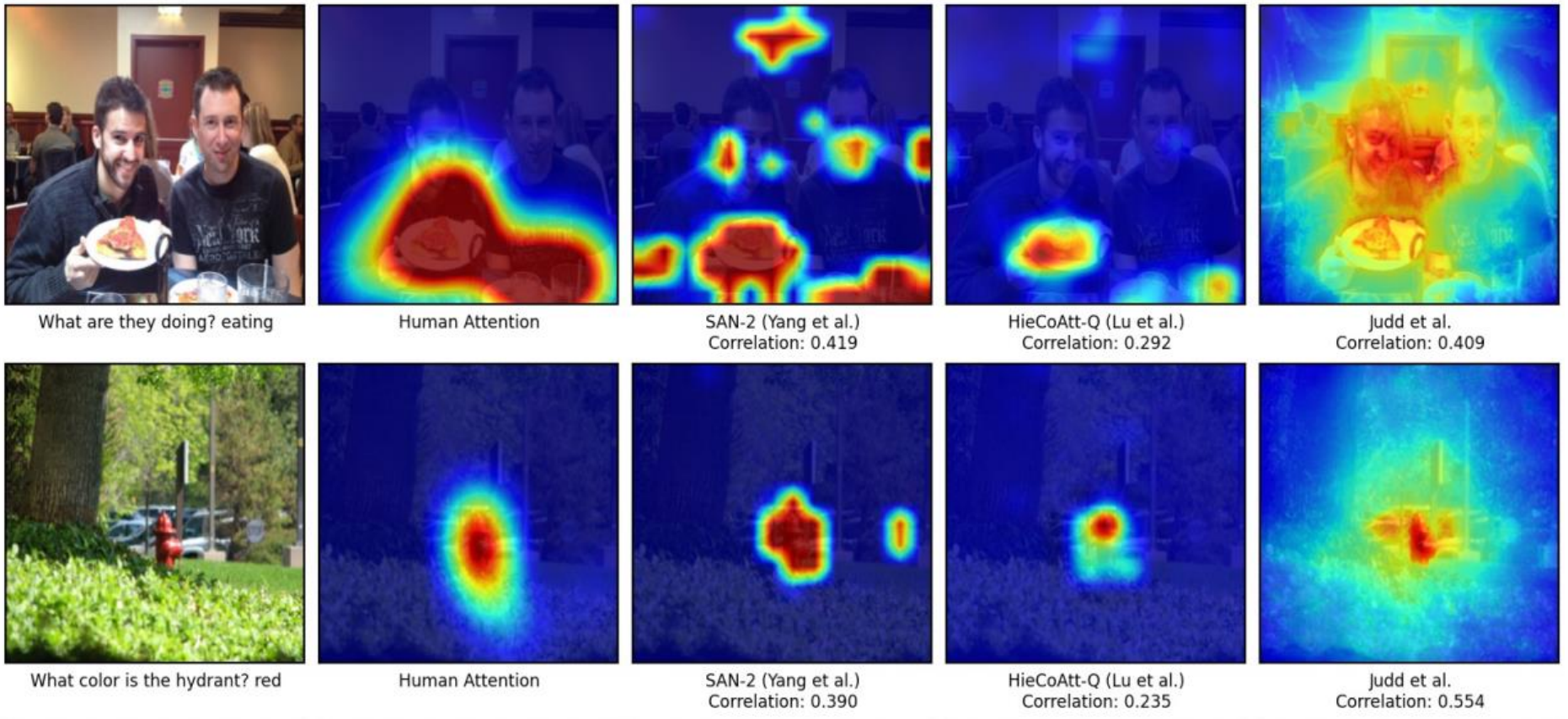
Table 1: Mean rank-correlation coefficients (higher is better); error bars show standard error of means. We can see that both SAN-2 and HieCoAtt attention maps are positively correlated with human attention maps, but not as strongly as task-independent Judd saliency maps.

- Image-question pair 당 세 명이 생성한 human attention maps 와의 평균 rank-correlation을 구했을 때, 0.623

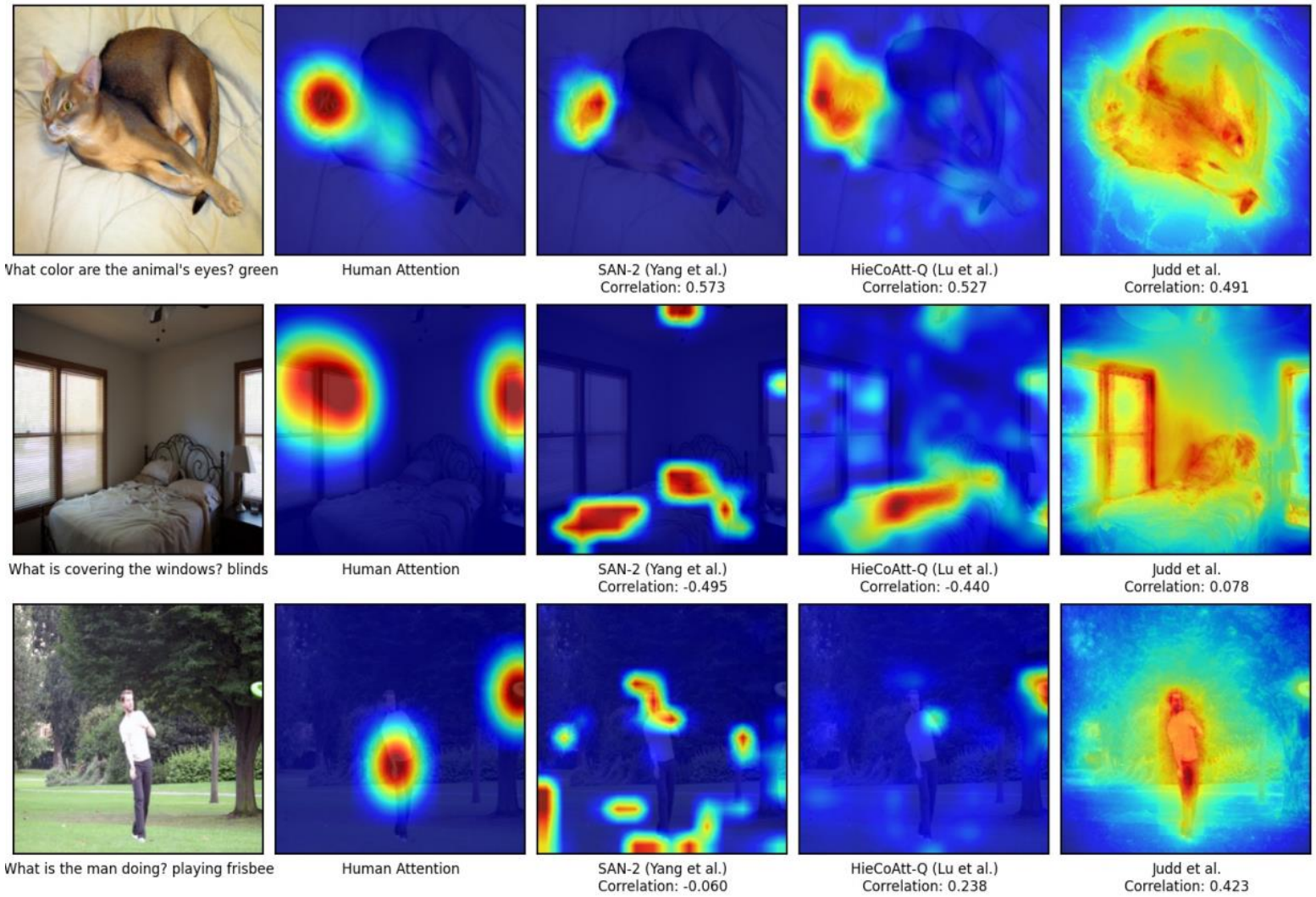
- human attention map에 순위의 variations을 주기위해 random noise를 추가한 후 3번 수행했을 때 평균 rank-correlation임

VQA와는 관련이 없지만 attention으로 사용할 수 있는 eye-tracking predicting 모델

Human Attention Maps vs Unsupervised Attention Models



Human Attention Maps vs Unsupervised Attention Models



Center biased 문제

- **이미지의 중앙에 있는 Salient objects (돌출물체)는 보통 질문의 정답이 될 가능성이 높다.**

- **근거는 ?**

  - Val set에 대해서 Central attention map을 인위적으로 생성한 후
    각각의 attention map과의 평균 rank-correlation 을 비교한 결과
    - Judd(eye-tracking) : 0.877
    - Human attention map : 0.458

    => 즉, Judd (eye-tracking) 는 중앙을 예측하는 경향이 매우 높음을 알 수 있음

Center biased image 제거 후 결과

- Center attention map과 rank-correlation이 높은 human attention map을 제거하고 재평가
  - 나머지는 비슷한데 Judd는 매우 하락

Model	Rank-correlation
SAN-2 (Yang et al., 2016)	0.038 ± 0.011
HieCoAtt-W (Lu et al., 2016)	0.062 ± 0.012
HieCoAtt-P (Lu et al., 2016)	0.048 ± 0.010
HieCoAtt-Q (Lu et al., 2016)	0.114 ± 0.012
Judd et al. (Judd et al., 2009)	-0.063 ± 0.009

Table 2: Correlation on the reduced set without center bias goes down significantly for Judd saliency since they have a strong center bias. Relative trends among SAN-2 & HieCoAtt are similar to those over the whole validation set (reported in Table 1).

4. Conclusion & Discussion

Conclusion & Discussion

1. Task-independent saliency map (eye-tracking)과 비교했을 때
   Rank-correlation이 매우 높아 보였지만 center bias 때문이었음

Conclusion & Discussion

1. Task-independent saliency map (eye-tracking)과 비교했을 때
   Rank-correlation이 매우 높아 보였지만 center bias 때문이었음

2. 실제 VQA모델들 (2016) 의 Rank-correlation은 0.062 수준으로 낮으므로
   Visual attention mechanisms 에서는 사람이 집중하는 영역에
   attention 하지 않음을 알 수 있음

Conclusion & Discussion

1. Conducted large-scale studies on ‘human attention’ in Visual Question Answering (VQA)

2. Designed and test multiple game-inspired novel attention-annotation interfaces that require the subject to sharpen regions of a blurred image to answer a question.

3. Introduced the VQA-HAT (Human ATtention) dataset.

4. Evaluated attention maps generated by state-of-the-art VQA models against human attention both qualitatively (via visualizations) and quantitatively (via rank-order correlation).

5. Overall, our experiments show that **current VQA attention models do not seem to be looking at the same regions as humans.**

Questions

1. 지금도 VQA-HAT 데이터세트로 평가하나?

2. VQA-HAT 로 지도 학습한 사례가 있나?

- Exploring Human-Like Attention Supervision in Visual Question Answering (AAAI-18)

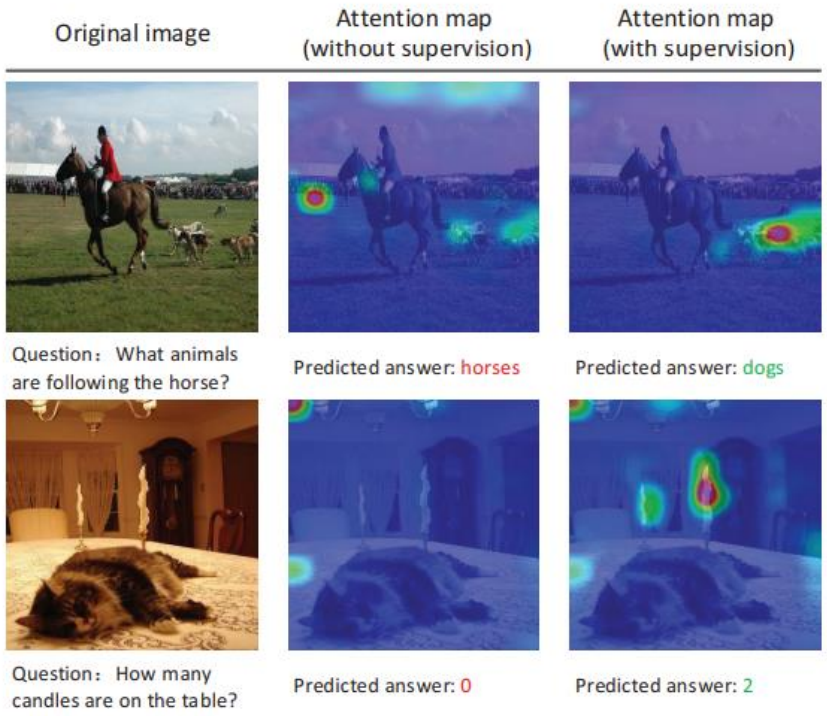


Figure 1: Visualization of images, attention maps and predicted answers. As it is possible to see, through explicit attention supervision, the attention maps are more accurate and the predicted answers yield better results. Best viewed in color.

Thank You

Reference

Paper : https://arxiv.org/pdf/1606.03556.pdf

SAN networks

https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Yang_Stacked_Attention_Networks_CVPR_2016_paper.pdf

Hierarchical Co-Attention Network

https://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering.pdf