UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

SRI International

A Study on
Multimodal and Interactive Explanations for Visual Question Answering

Kamran Alipour,[1] Jurgen P. Schulze,[1] Yi Yao,[2] Avi Ziskind,[2] and Giedrius Burachas [2]

[1]UC San Diego

[2]SRI International

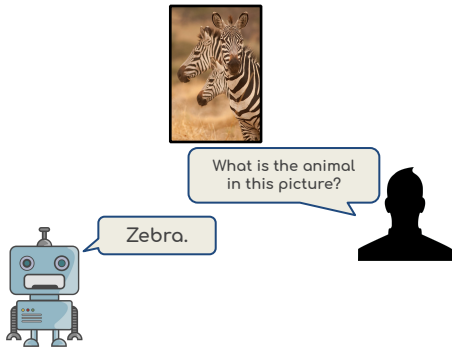
DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

XAI EXPLAINABLE ARTIFICIAL INTELLIGENCE

Introduction

Visual Question Answering

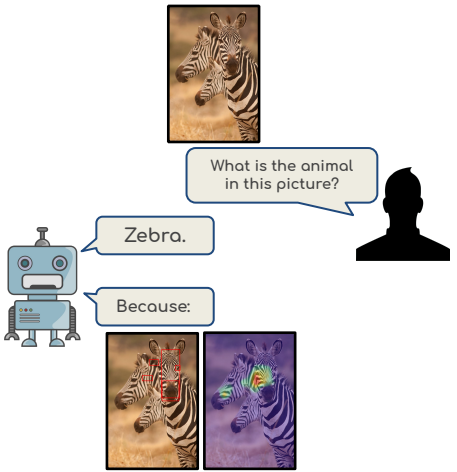
- Answering a question in natural language regarding a given image
- Attention-based models on image features influenced by the question to produce answers



Introduction

eXplainable AI (XAI)

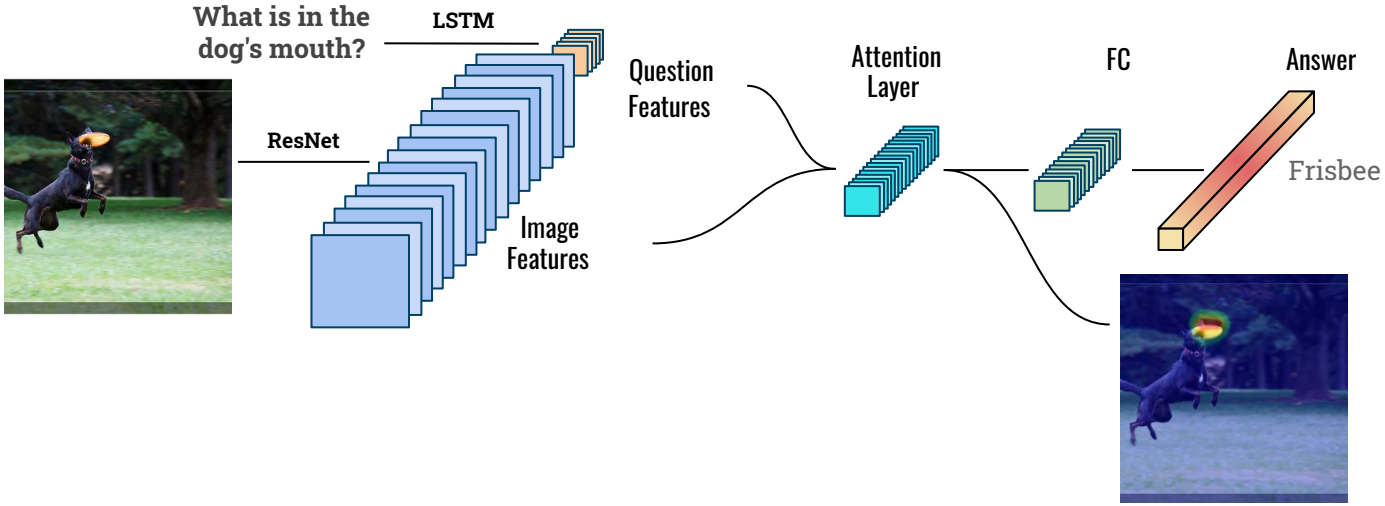
- Used for years in medicine, robotics, ...
- Mainly attention-based
- Various formats: visual, textual, ...



Contributions

- Generated multiple explanation modes from attention features and annotations
- Introduced an interactive explanation mode
- Prediction task: A novel assessment of the efficacy of explanations
- A user study to evaluate our XAI system

eXplainable Visual Question Answering (XVQA)



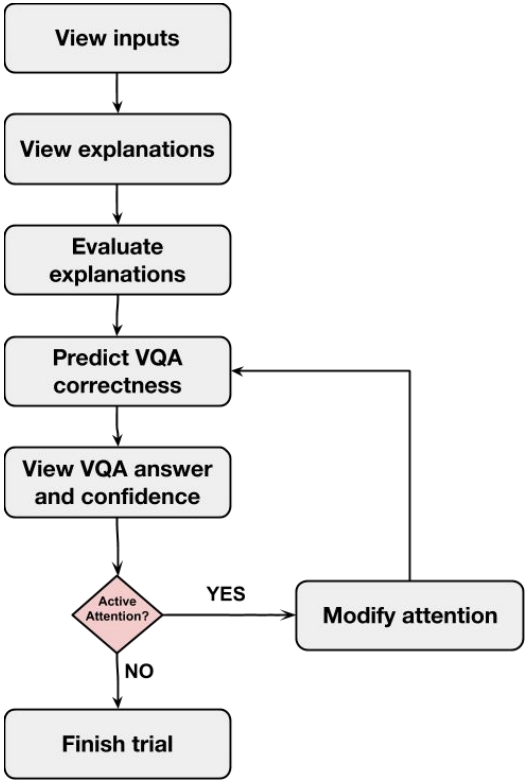
User Study

Input

- Images and Questions
- Explanations
- VQA correctness and confidence

TASK:

- Rank explanations
- Predict if VQA answers correctly based on explanations.



User Study

Subjects:	90
Trials:	10513

Group	
NE	Control group
SP	Spatial attention
SA	Active (steerable) attention
SE	Semantic
OA	Object attention
AL	All explanations

Explanation modes

Spatial attention

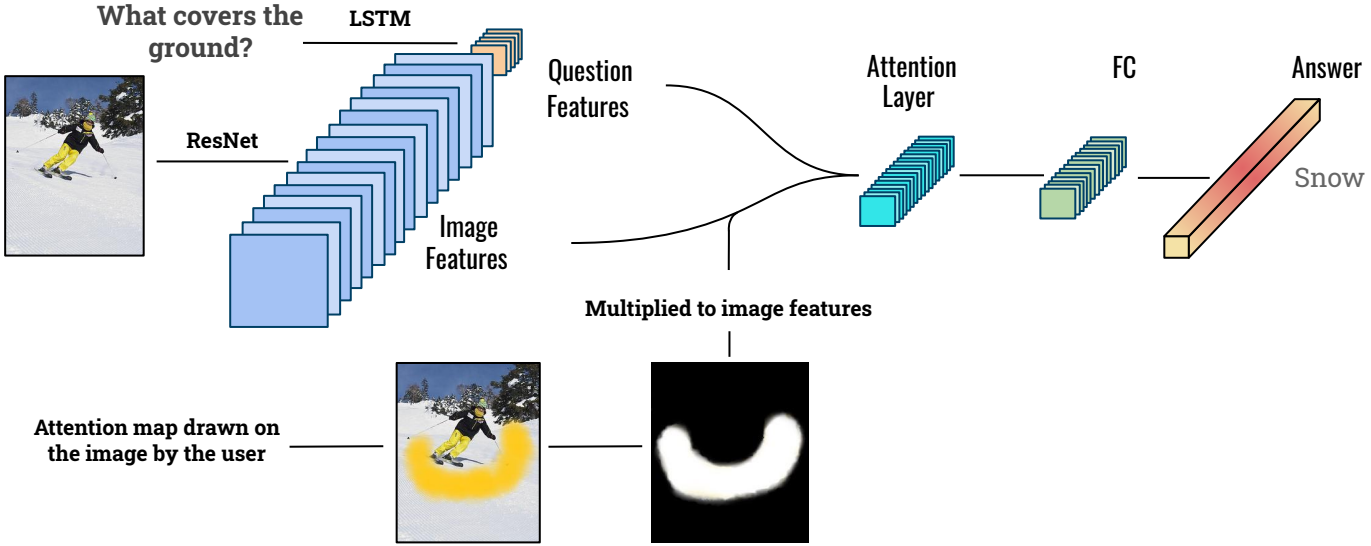
Shows the parts of the image the model is focused on while preparing the answer.

QUESTION: Where is this place?

ANSWER: Airport.

Explanation modes

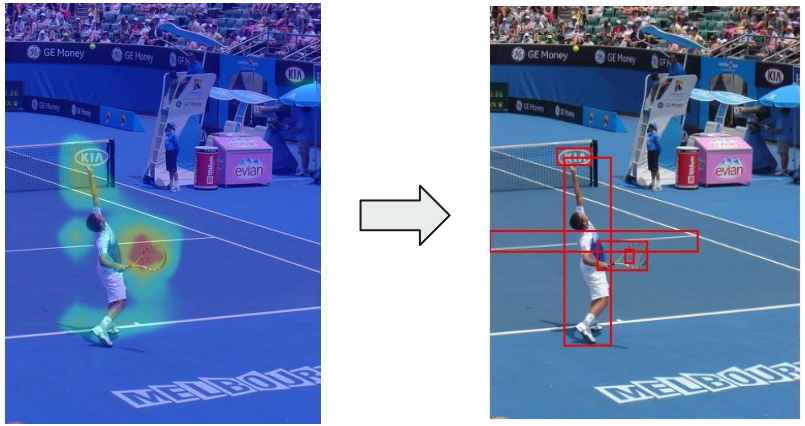
Active (streerable) attention



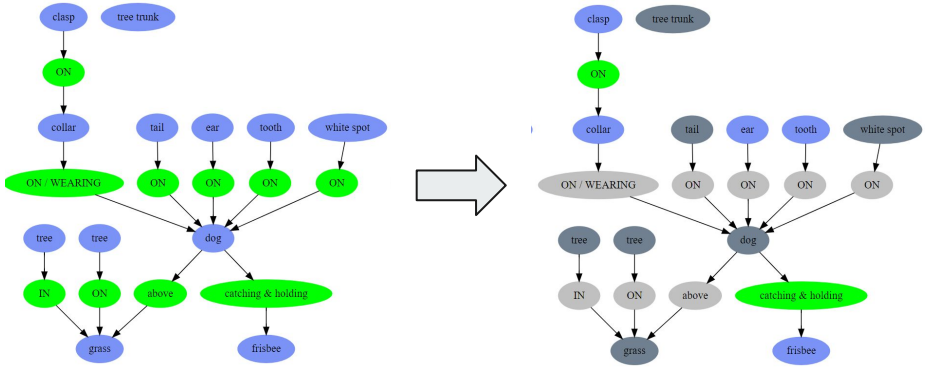
Explanation modes

Semantic:

Bounding box



Scene graph



Explanation modes

Semantic:

Textual Explanation [Ghosh et al. 2019]

QUESTION: What is this sport?

ANSWER: Soccer.

Explanations:

Because the image contains: **white line** on **soccer field**

Because the image contains: **man** on **soccer field**

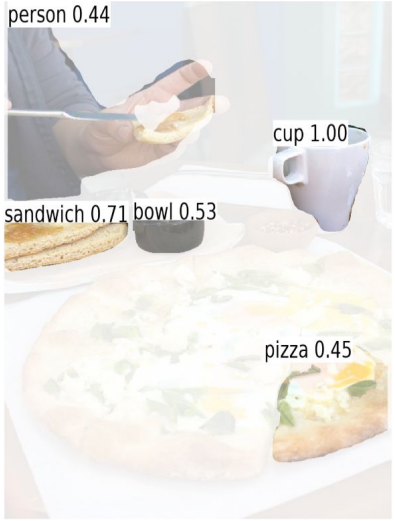
Because the image contains: **soccer shorts** with **numbers**

Explanation modes

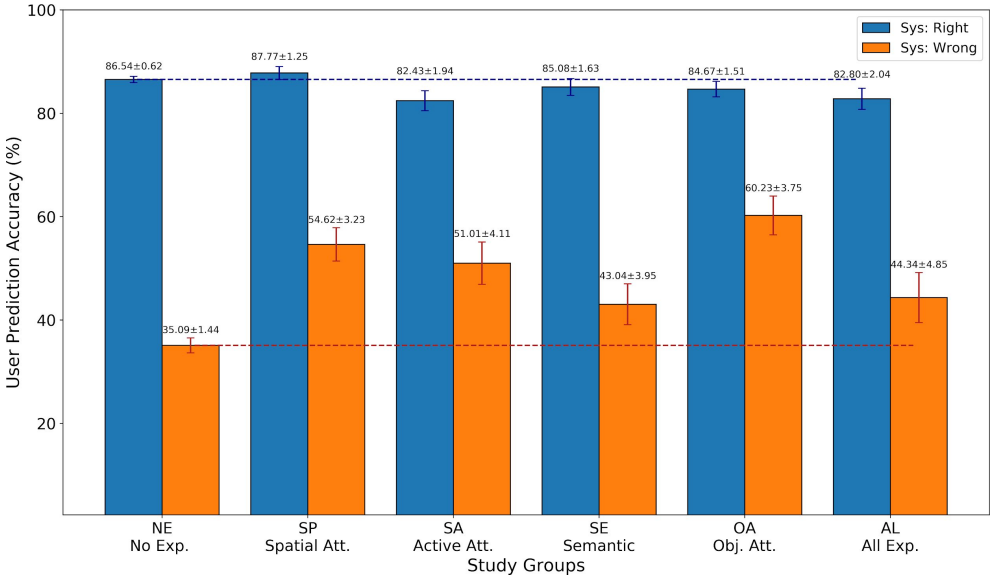
Object attention [Ray et al. 2019]

QUESTION: What food is he eating?

ANSWER: Sandwich.



Results

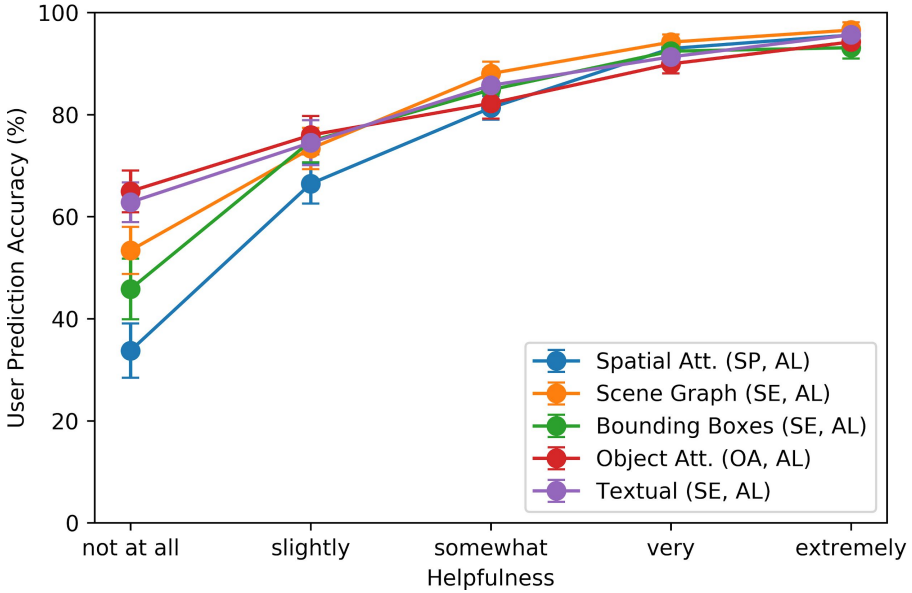


Prediction accuracies in different study groups

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

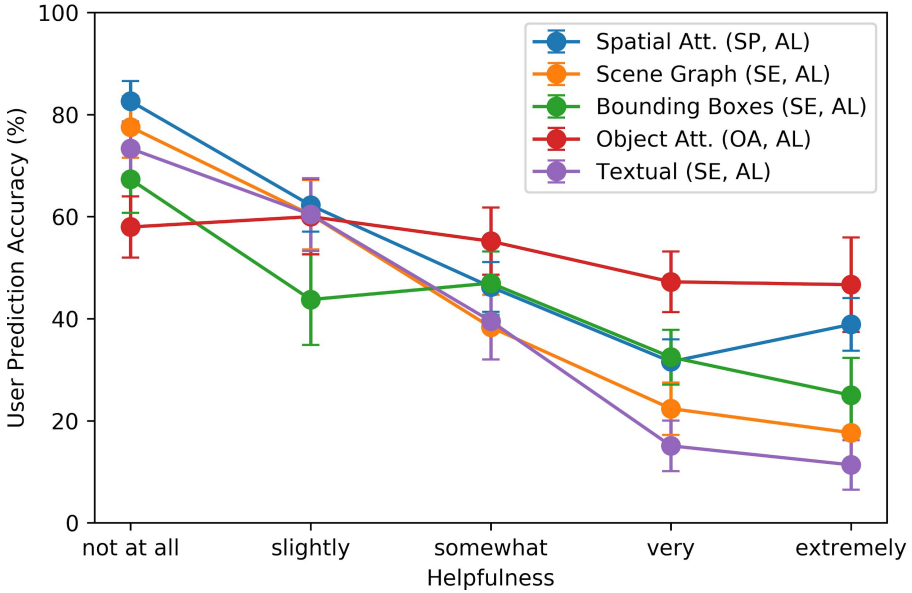
SRI International

Results



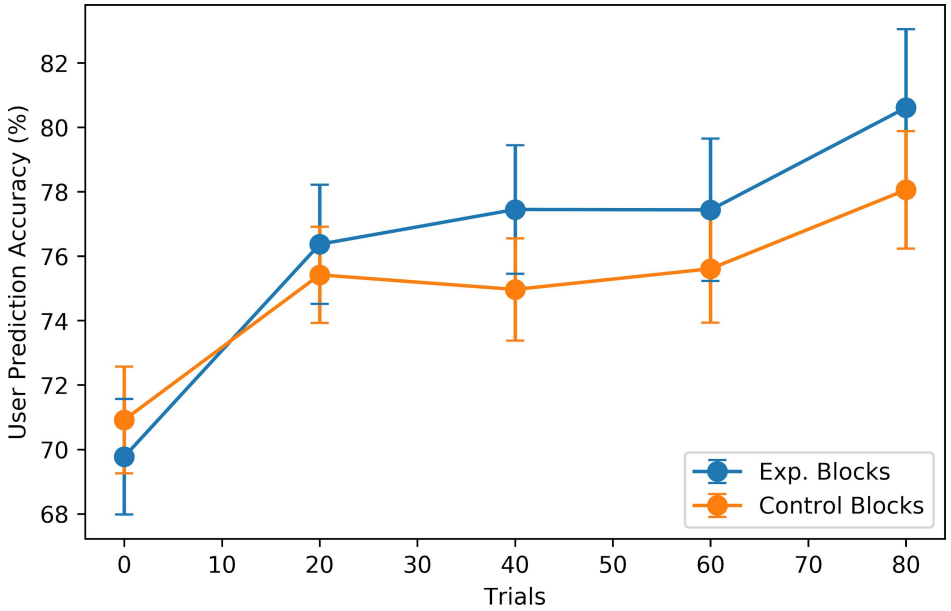
Prediction accuracy vs. explanation ratings when system accurate

Results



Prediction accuracy vs. explanation ratings when system inaccurate

Results

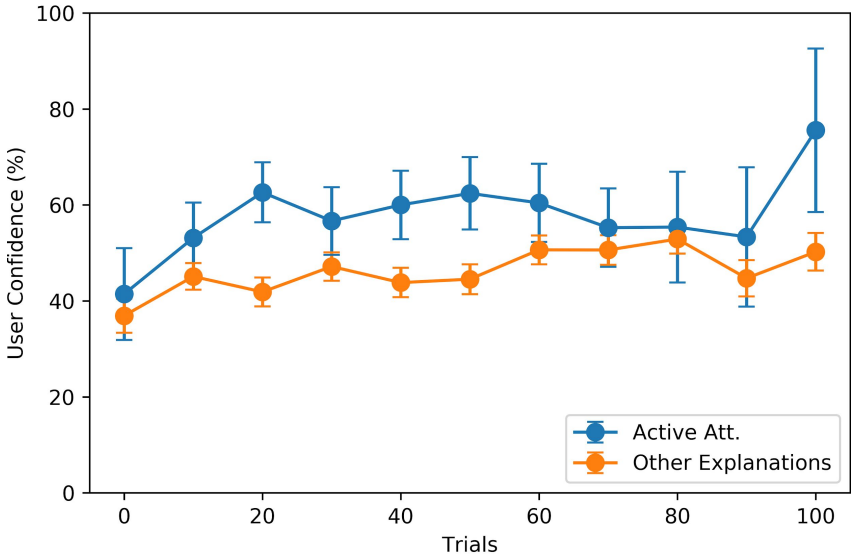


Prediction accuracy progress

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

SRI International

Results



User confidence progress: Active attention vs. other explanations

Discussion

- Interactive explanations improved users confidence compared to other explanations
- When AI is inaccurate, explanations are more helpful
- Getting exposed to multiple explanation modes can be conflicting/overwhelming and reduce prediction accuracy

Conclusion and Future Work

- Interactive experiment to probe explanation effectiveness
- Explanations help accuracy prediction
- Users confidence improved when exposed to explanations

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

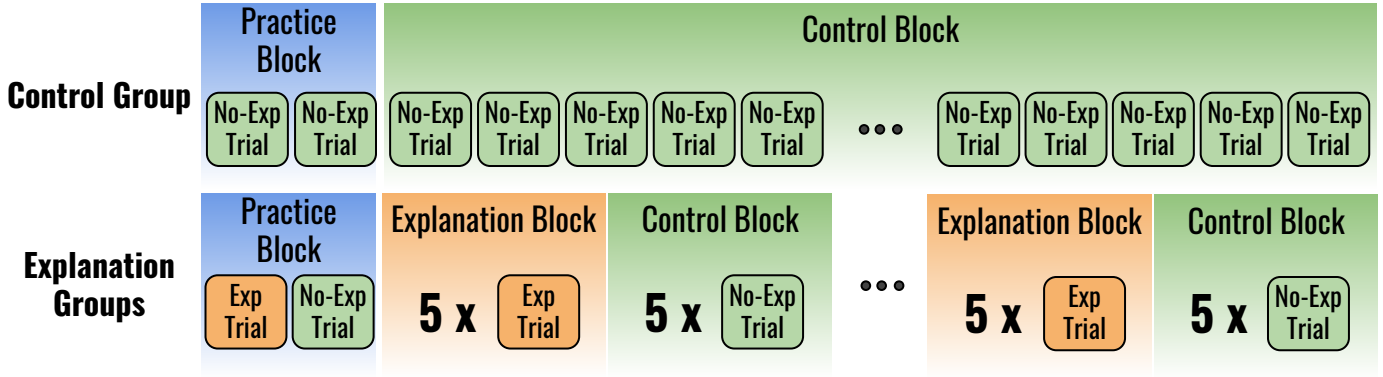
SRI International

Thank You!

User Study Statistics

Group		Subjects	Trials
NE	Control group	15	4124
SP	Spatial attention	15	1826
SA	Active (steerable) attention	15	1021
SE	Semantic	15	1261
OA	Object attention	15	1435
AL	All explanations	15	846
	Total :	90	10513

User Study Structure



User Study Structure

Input

- Images and Questions
- Explanations
- VQA correctness and confidence

TASK:

- Rank explanations
- Predict if VQA answers correctly based on explanations.