Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention


Sounak Mondal


Zhibo Yang


Seoyoung Ahn


Dimitris Samaras


Gregory Zelinsky


Minh Hoai

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Stony Brook University

TUE-AM-137

Motivation

- We focus on gaze prediction for *visual search*
- Previous models have needed, for *each* target category
  - *human gaze training data,* and
  - *detectors to encode target*
- Hard to scale when gaze/detection annotation is unavailable for a target

find **"bottle"**

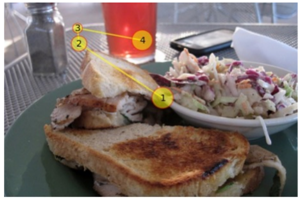
find **"stand mixer"**

Proposed Solution

- We propose a novel ***ZeroGaze*** task to evaluate scalability
- We propose a novel ***Gazeformer*** *model to solve ZeroGaze*
    - *Gazeformer* is more scalable, more effective and faster than previous methods

Training Dataset:
search target in *N* categories

find **"fork"**



find **"cup"**

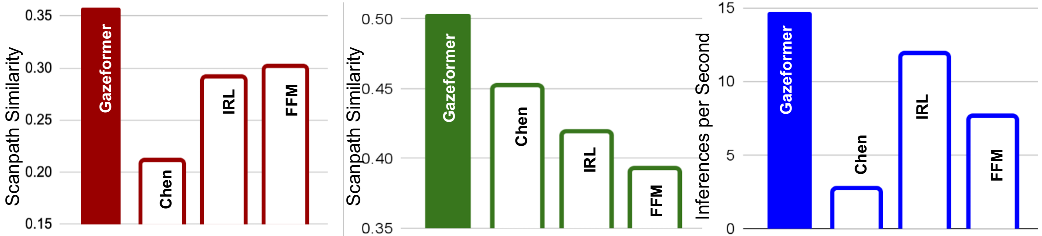
...

Training

Search target
***outside* N categories**
(**ZeroGaze** setting)
e.g. find "knife"

Performance for
search target
***in* N categories**
(**GazeTrain** setting)
e.g. find "fork"

Inference
Throughput



"Scalable" ***"Effective"*** ***"Fast"***

Inference

Gaze Prediction for HCI applications

- Recently, *gaze prediction* models *for visual search* have been used in several *HCI* applications
  - AR/VR
  - Robotics

- Besides being accurate and fast, these models must be *scalable*
  - Must extend to new targets in the real world
    - Collecting annotation for all possible targets is impractical!

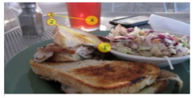
Image courtesy: TOBII VR

ZeroGaze

- We introduce the ***ZeroGaze*** task
    - Tests *scalability* of gaze prediction models for visual search
    - Extends *zero-shot learning* to gaze prediction

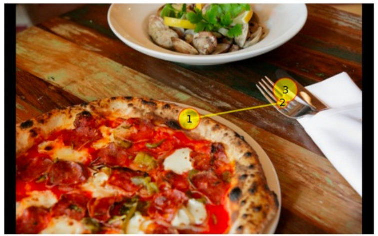
Training Dataset:
Gaze data for
N categories

find **"fork"**



find **"cup"**

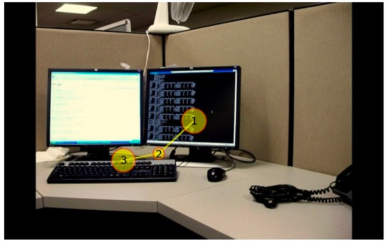
find **"laptop"**



find **"fork"**

GazeTrain

Search category *in* N categories



find **"keyboard"**

ZeroGaze

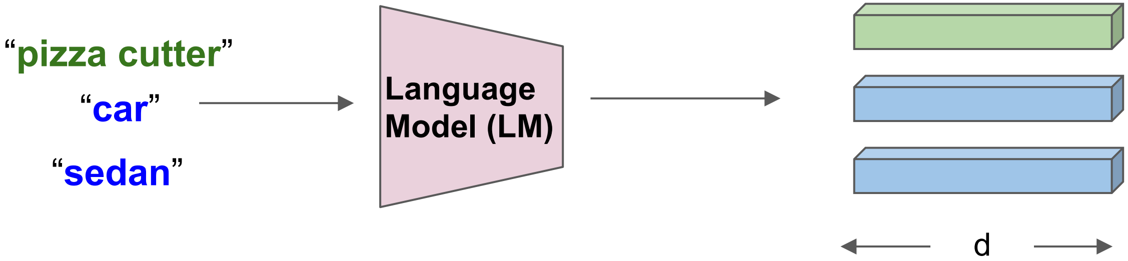
Search category *outside* N categories

Gazeformer

- We solve ZeroGaze with the novel *Gazeformer* model
  - Improves scalability, effectiveness and efficiency of gaze prediction
- Key components of *Gazeformer*
  - *Language-based Target Encoding*
  - *Transformer Encoder-Decoder Architecture*
  - *Fixation Modeling in Image Space*

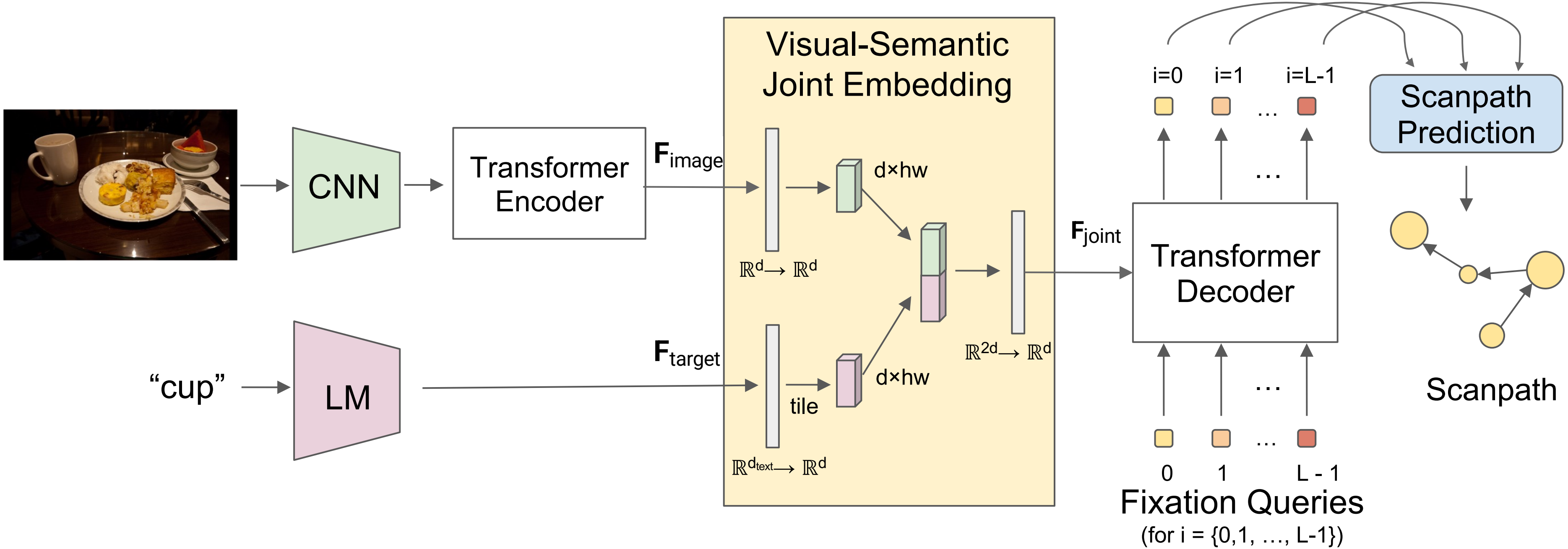
Language-based Target Encoding

- *Gazeformer* uses a pre-trained language model to encode target name
  - *Scalable* - can encode any target using its name
  - Embodied semantics helps extending to unknown targets



Transformer Encoder-Decoder Architecture

- *Gazeformer* adopts a transformer encoder-decoder architecture
  - Learns interactions between image and target semantics
  - Models spatio-temporal context for scanpath generation



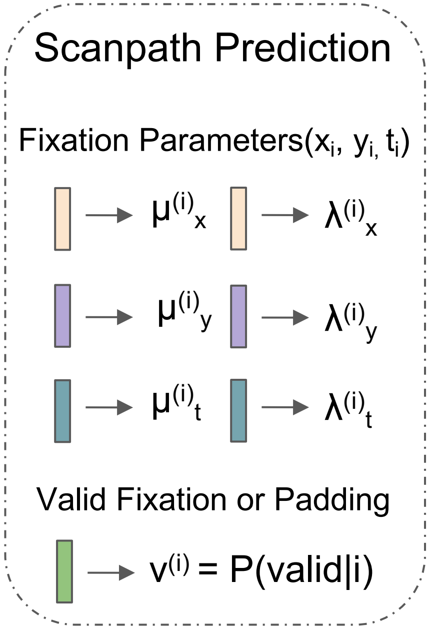
Fixation Modeling in Image Space

- Previous methods predicted fixation probabilities over patches
  - Does *not* penalize *distance* of predicted fixations from ground truth

- We *regress* fixation parameters using Gaussian distributions

$$x_i = \mu_{x_i} + \epsilon_{x_i} \cdot \exp(0.5\lambda_{x_i}), \quad y_i = \mu_{y_i} + \epsilon_{y_i} \cdot \exp(0.5\lambda_{y_i}),$$
$$t_i = \mu_{t_i} + \epsilon_{t_i} \cdot \exp(0.5\lambda_{t_i}), \quad \epsilon_{x_i}, \epsilon_{y_i}, \epsilon_{t_i} \in \mathcal{N}(0,1).$$

- Gazeformer learns *scanpath termination*
  - Separate MLP learns if a latent vector corresponds to a valid fixation or padding



Experimental Results: ZeroGaze

	SS↑		SemSS↑		FED↓		SemFED↓		MM	CC	NSS
	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	↑	↑	↑
IRL	0.290	-	0.314	-	4.606	-	4.377	-	0.774	0.241	4.018
Chen *et al.*	0.210	0.041	0.211	0.034	5.720	210.498	5.608	211.636	0.717	0.002	0.001
FFM	0.300	-	0.334	-	3.271	-	2.918	-	0.731	0.271	**5.247**
Gazeformer-noDur	**0.359**	-	**0.391**	-	2.788	-	2.474	-	**0.822**	0.316	4.671
Gazeformer	0.358	**0.312**	**0.391**	**0.348**	**2.766**	**12.505**	**2.438**	**10.391**	0.812	**0.324**	4.929

- We implement ZeroGaze setting on COCO-Search18 using Leave-One-Out scheme
- *Gazeformer* outperforms baselines in **ZeroGaze** setting on multiple metrics

Experimental Results: GazeTrain

	SS↑		SemSS↑		FED↓		SemFED↓		MM ↑	CC ↑	NSS ↑
	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur			
Human	0.490	0.409	0.548	0.456	2.531	11.526	1.637	8.086	0.857	0.472	8.129
IRL	0.418	-	0.499	-	2.722	-	2.182	-	0.833	0.434	6.895
Chen *et al.*	0.451	0.403	0.504	0.446	2.187	10.795	1.788	8.782	0.820	0.547	6.901
FFM	0.392	-	0.443	-	2.693	-	2.284	-	0.808	0.370	5.576
Gazeformer-noDur	**0.504**	-	**0.534**	-	**2.061**	-	**1.742**	-	0.849	0.559	8.356
Gazeformer	**0.504**	**0.451**	0.525	**0.485**	2.072	**9.708**	1.810	**7.688**	**0.852**	**0.561**	**8.375**

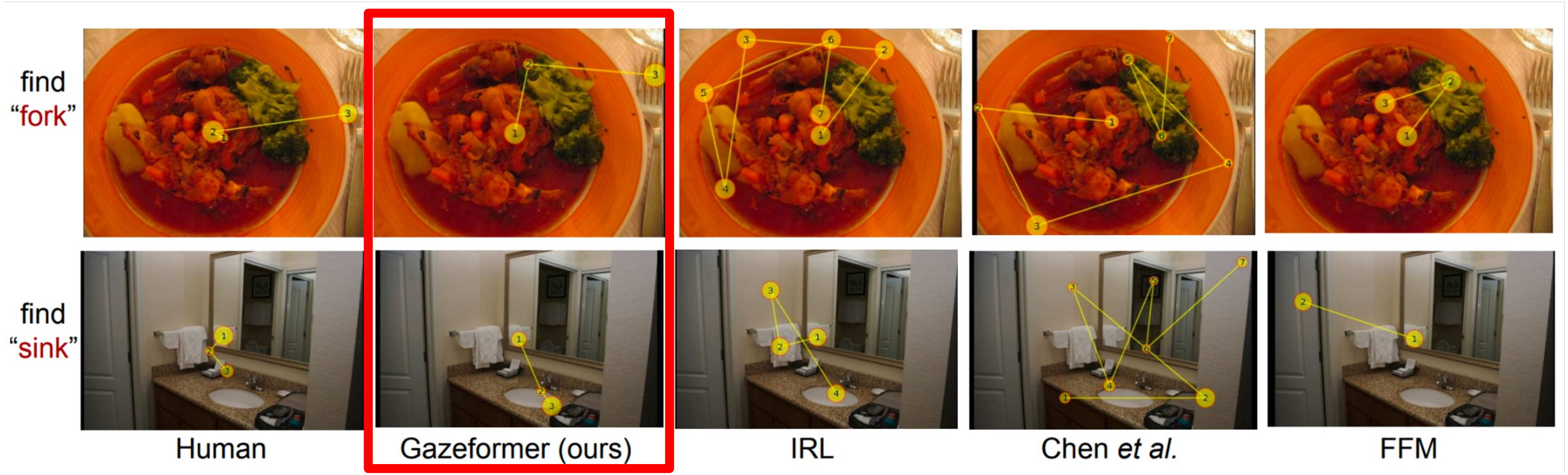
- *Gazeformer* outperforms baselines under the **GazeTrain** setting on COCO-Search18

Experimental Results: Inference Time

	Time (in ms)↓	Inferences/s ↑	Speedup ↑
Chen *et al.*	386	2.59	1X
FFM	133	7.52	2.9X
IRL	85	11.77	4.5X
Gazeformer	**68**	**14.71**	**5.7X**

- *Gazeformer* is several times **faster** than baselines

Experimental Results: Qualitative



- *Gazeformer* extends to new categories in **ZeroGaze** setting

Extensibility to Uncommon Categories

Hyponyms or synonyms of target names

find “hatchback”

find “sedan”

find “mug”

No annotation in COCO dataset

find “trash can”

find “pizza cutter”

find “soda can”

- *Gazeformer* extends to **unknown and uncommon targets**

Conclusion and Future Work

- We introduced the *ZeroGaze* task
- We proposed the novel *Gazeformer* model
- *Gazeformer* is more *scalable, effective and efficient* than previous approaches
- We hope *Gazeformer* will be extended to other visual tasks such as VQA and real-world HCI applications
- Code available at https://github.com/cvlab-stonybrook/Gazeformer

SCAN ME!