PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Adnen Abdessaied

WS 23/24

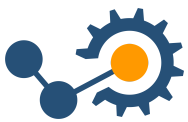
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Neuro-Symbolic Visual Dialog

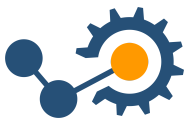
Introduction

- Many neuro-symbolic approaches rely on external executors to process some generated code Yi et al. [2018]; Mao et al. [2019]; Abdessaied et al. [2022]



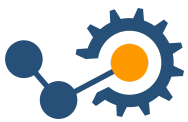
Introduction

- Many neuro-symbolic approaches rely on external executors to process some generated code Yi et al. [2018]; Mao et al. [2019]; Abdessaied et al. [2022]
- Typically, these executors are compatible with a specific Domain Specific Language (DSL)

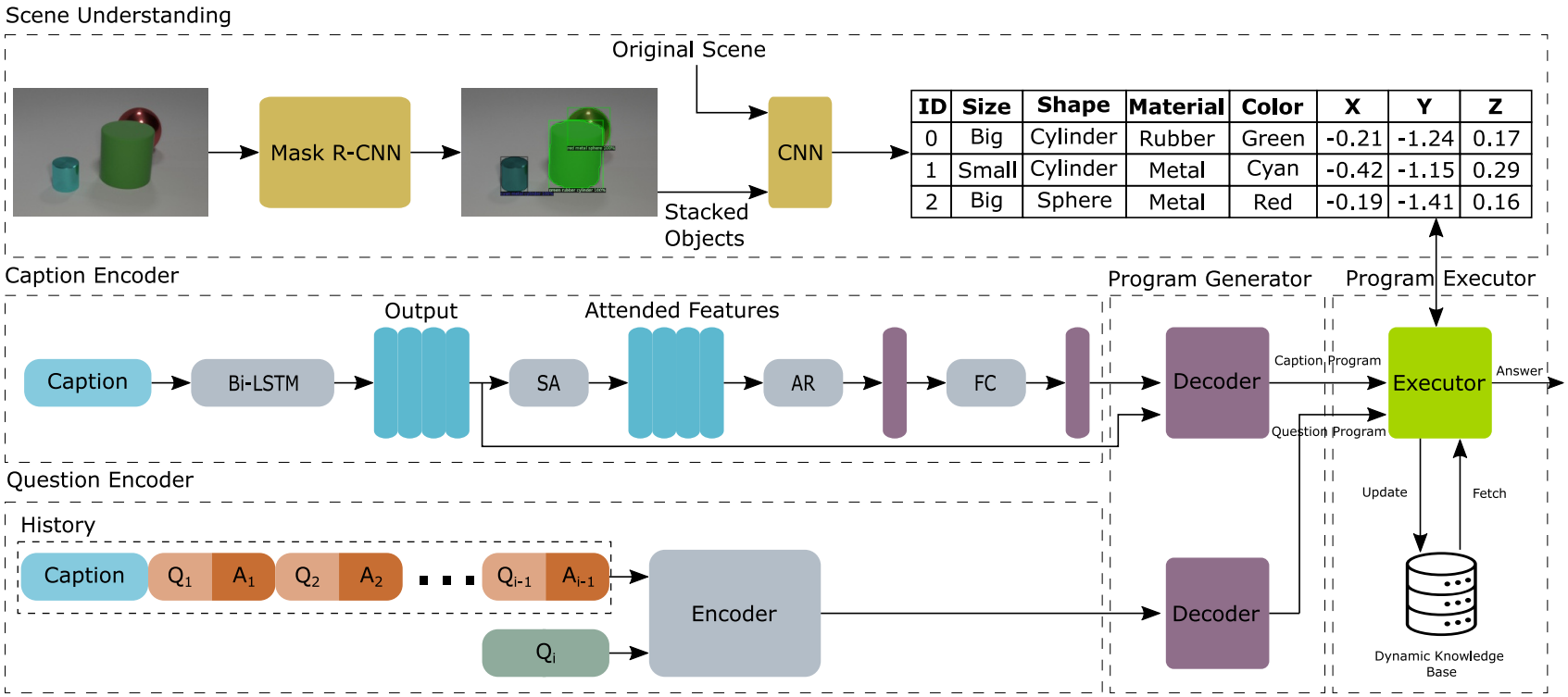


Introduction

- Many neuro-symbolic approaches rely on external executors to process some generated code Yi et al. [2018]; Mao et al. [2019]; Abdessaied et al. [2022]
- Typically, these executors are compatible with a specific Domain Specific Language (DSL)
- The DSL is task-dependent (e.g. VQA Johnson et al. [2017], VideoQA Yi et al. [2020], Visual Dialog Kottur et al. [2019], etc)



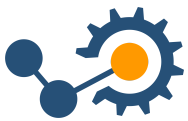
Neuro-Symbolic Visual Dialog (Abdessaied et al. [2022])



Source: Abdessaied et al. [2022]

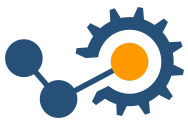
Neuro-Symbolic Visual Dialog (Abdessaied et al. [2022])

- NSVD is trained to separately generate caption and question programs



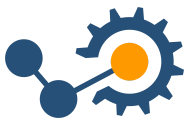
Neuro-Symbolic Visual Dialog (Abdessaied et al. [2022])

- NSVD is trained to separately generate caption and question programs
- Both types of programs are dealt with by two dedicated parsers



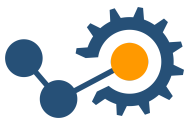
Project Goals

- The main goal is to unify the generation of the caption and question programs.



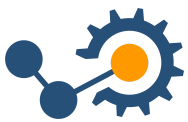
Project Goals

- The main goal is to unify the generation of the caption and question programs.
- Implement only one encoder that takes in the caption and the dialog history as inputs



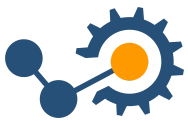
Project Goals

- The main goal is to unify the generation of the caption and question programs.
- Implement only one encoder that takes in the caption and the dialog history as inputs
- The generator outputs both the caption and question programs



Project Goals

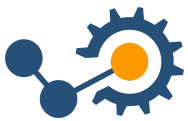
- The main goal is to unify the generation of the caption and question programs.
- Implement only one encoder that takes in the caption and the dialog history as inputs
- The generator outputs both the caption and question programs
- Re-implement some of modules of the CLEVR-Dialog DSL (Re-implementation task)



Project Goals

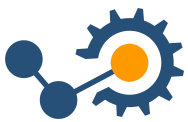
	Func. Name	Func. Args.	Func. Out.	fetch	update			
					Handle	Conv. Subj.	Seen Objs.	Groups
Caption Programs	count-att	attr	none	✗	✓	✗	✓	✓
	extreme-right	⊎ [attr_1,..,attr_4]	none	✗	✓	✓	✓	✗
	extreme-left	⊎ [attr_1,..,attr_4]	none	✗	✓	✓	✓	✗
	extreme-behind	⊎ [attr_1,..,attr_4]	none	✗	✓	✓	✓	✗
	extreme-front	⊎ [attr_1,..,attr_4]	none	✗	✓	✓	✓	✗
	extreme-centre	⊎ [attr_1,..,attr_4]	none	✗	✓	✓	✓	✗
	unique-obj	⊎ [attr_1,..,attr_4]	none	✗	✓	✓	✓	✗
	obj-relation	attr_obj_1,pos,attr_obj_2	none	✗	✓	✓	✓	✗
Question Programs	count-all	-	num	✗	✗	✗	✗	✓
	count-other	-	num	✗	✗	✓	✓	✗
	count-all-group	-	num	✗	✗	✗	✗	✗
	count-attribute	attr	num	✗	✓	✓	✓	✓
	count-attribute-group	attr	num	✗	✓	✓	✓	✓
	count-obj-rel-imm	pos	num	✗	✗	✓	✓	✓
	count-obj-rel-imm-2	pos	num	✗	✗	✓	✓	✓
	count-obj-rel-early	pos,attr	num	✓	✓	✓	✓	✓
	count-obj-exclude-imm	attr_type	num	✗	✗	✓	✓	✓
	count-obj-exclude-early	attr_type,attr	num	✓	✗	✓	✓	✓
	exist-other	-	yes/no	✗	✗	✗	✓	✓
	exist-attribute	attr	yes/no	✗	✓	✗	✓	✓
	exist-attribute-group	attr	yes/no	✗	✓	✓	✓	✓
	exist-obj-rel-imm	pos	yes/no	✗	✗	✓	✓	✓
	exist-obj-rel-imm2	pos	yes/no	✗	✗	✓	✓	✓
	exist-obj-rel-early	pos,attr	yes/no	✓	✓	✓	✓	✓
	exist-obj-exclude-imm	attr_type	yes/no	✗	✗	✓	✓	✓
	exist-obj-exclude-early	attr_type,attr	yes/no	✓	✗	✗	✗	✓
	seek-attr-imm	attr_type	attr	✗	✓	✗	✗	✗
	seek-attr-imm2	attr_type	attr	✗	✓	✗	✗	✗
	seek-attr-early	attr_type,attr	attr	✓	✓	✓	✓	✗
	seek-attr-sim-early	attr_type,attr	attr	✓	✓	✓	✓	✗
	seek-attr-rel-imm	attr_type	attr	✗	✓	✓	✓	✗
	seek-attr-rel-early	attr_type,pos,attr	attr	✓	✓	✓	✓	✗

Source: Abdessaied et al. [2022]



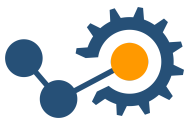
Instructions

- The data and starter code are available on Kaggle



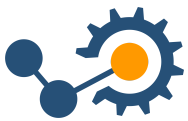
Instructions

- The data and starter code are available on Kaggle
- Adapt the model architecture to match the proposed modifications



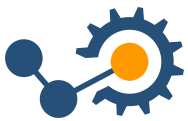
Instructions

- The data and starter code are available on Kaggle
- Adapt the model architecture to match the proposed modifications
- Train the seq2seq program generator
- Complete the CLEVR-Dialog DSL



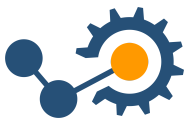
Instructions

- The data and starter code are available on Kaggle
- Adapt the model architecture to match the proposed modifications
- Train the seq2seq program generator
- Complete the CLEVR-Dialog DSL
- Compare your results with the NSVD paper (program accuracy & answer accuracy)



Instructions

- The data and starter code are available on Kaggle
- Adapt the model architecture to match the proposed modifications
- Train the seq2seq program generator
- Complete the CLEVR-Dialog DSL
- Compare your results with the NSVD paper (program accuracy & answer accuracy)
- More details are in the README file – Happy Coding!



References

A. Abdessaied, M. Bâce, and A. Bulling. Neuro-Symbolic Visual Dialog. In *COLING*, 2022.

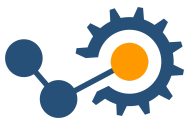
J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach. CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog. In J. Burstein, C. Doran, and T. Solorio, editors, *NAACL*, 2019.

J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*, 2019.

K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NeurIPS*, 2018.

K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Lei Shi

WS 23/24

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

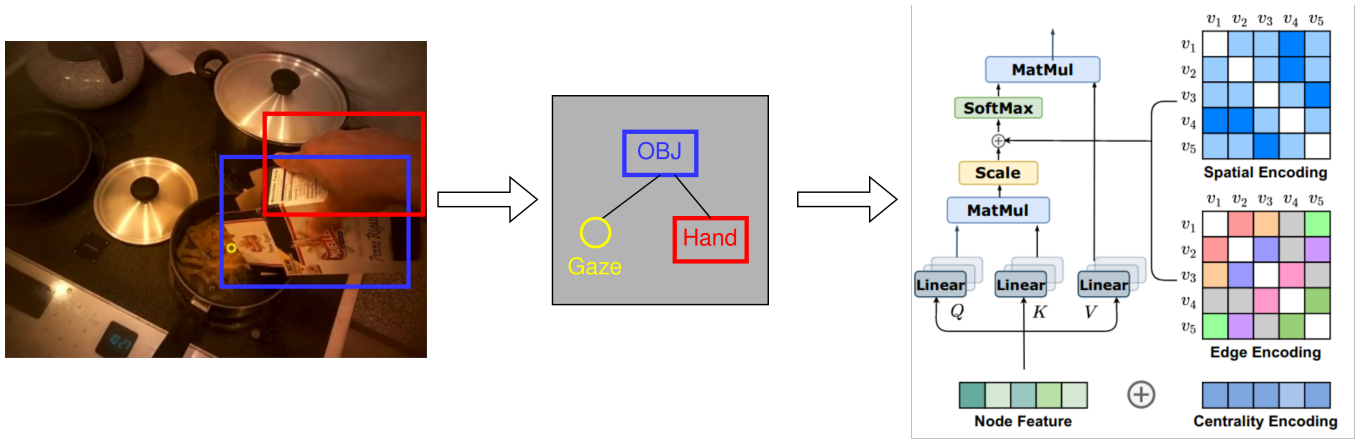
Gaze-based Action Recognition Using Graphformers

Motivation

Use gaze to improve action recognition on GTEA Gaze+ (Li et al. [2015])

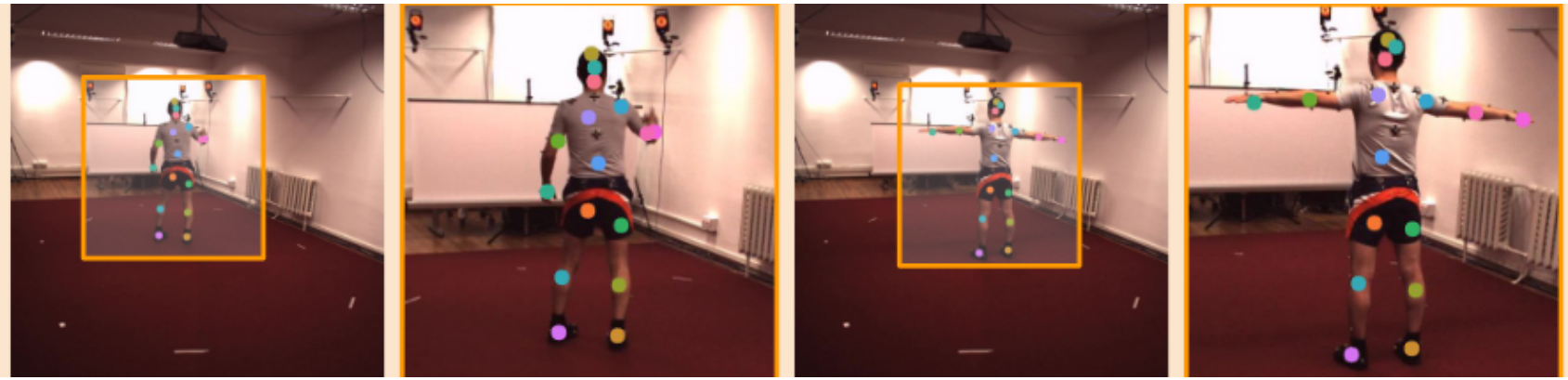
Approach

Train Graphformer (Ying et al. [2021]) to predict actions



Behavioral Keypoint Discovery for Action Recognition

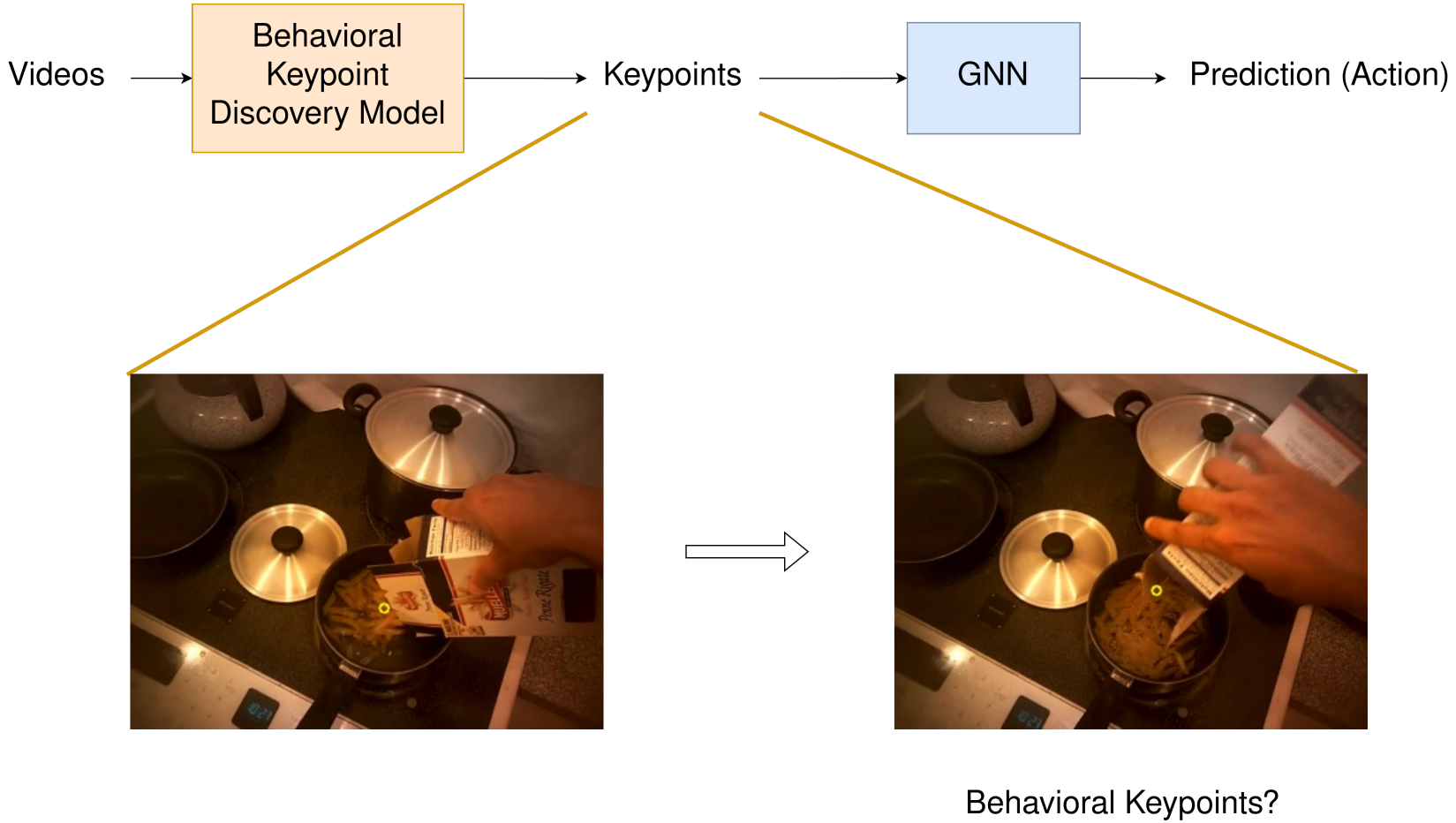
Motivation



Source: Sun et al. [2022]

Approach

GTEA Gaze+ (Li et al. [2015])



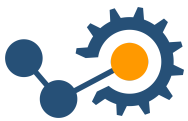
Source: Sun et al. [2022]

References

Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 287–295, 2015.

J. J. Sun, S. Ryou, R. H. Goldshmid, B. Weissbourd, J. O. Dabiri, D. J. Anderson, A. Kennedy, Y. Yue, and P. Perona. Self-supervised keypoint discovery in behavioral videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2171–2180, 2022.

C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Matteo Bortoletto

WS 23/24

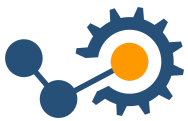
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

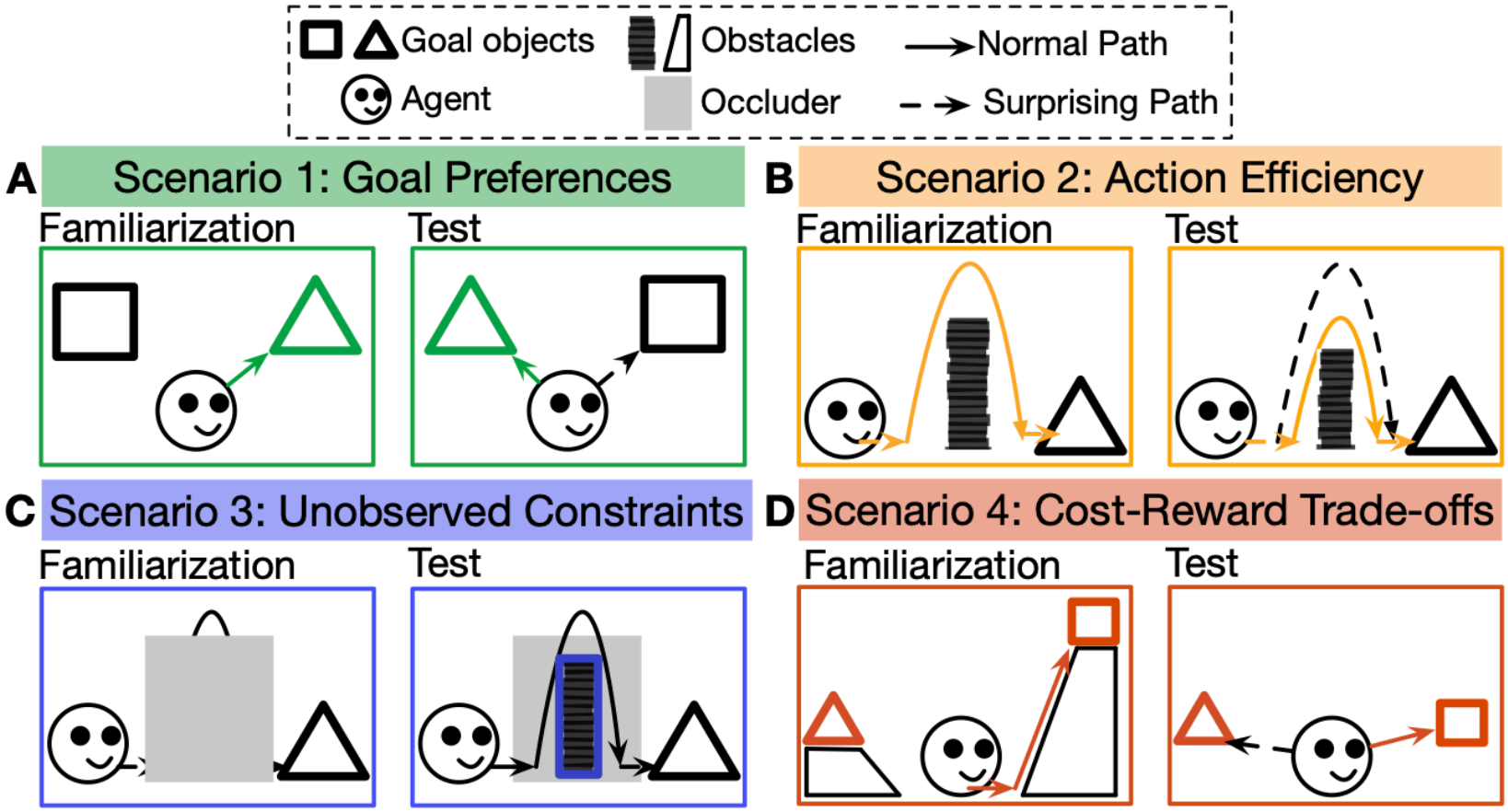
Learning core psychological reasoning

Motivation

- Core psychological reasoning comes naturally to people, but not to neural networks
- However, core psychological reasoning is crucial for human-AI interaction (virtual assistants, chatbots, robots)



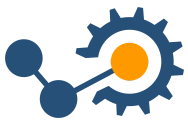
AGENT benchmark



Source: Shu et al. [2021]

Tasks

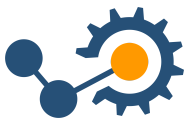
- Implement the ToMnet-G baseline and evaluate it on AGENT
- Compare the results with a model provided by me
- Perform some ablation studies and/or analyses on the results



Reverse engineering Theory of Mind

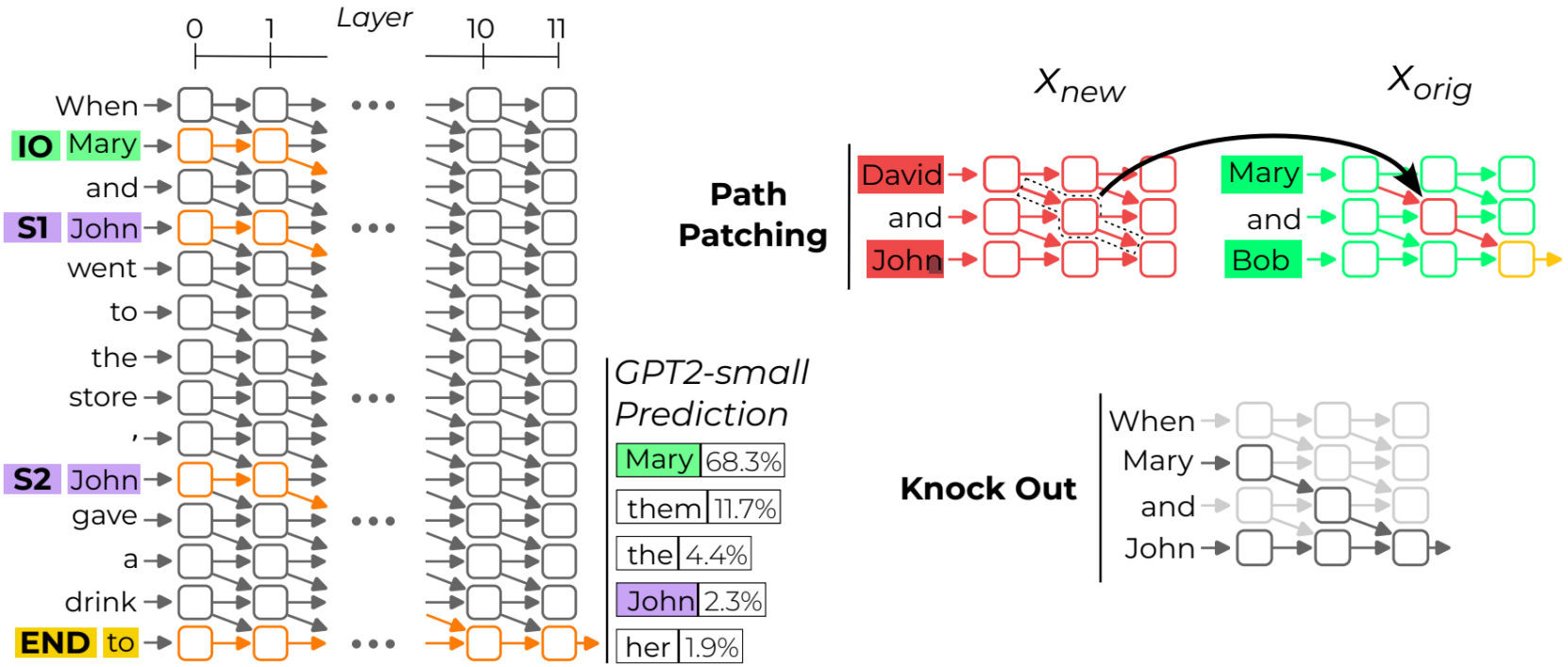
Motivation

- Theory of Mind refers to our capability to understand others' (or our own's) mental states
- Large language models have proved to have some Theory of Mind capabilities but it is not clear why or how they display these capabilities



Mechanistic interpretability

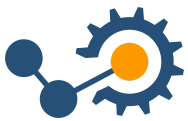
The goal of mechanistic interpretability is to take a trained model and reverse engineer the algorithms the model learned during training from its weights.



Source: Wang et al. [2022]

Tasks

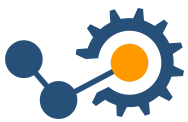
- Perform an exploratory analysis on a dataset consisting of natural language tasks that evaluate Theory of Mind
- Take a pre-trained language model (e.g. GPT-2)
- Apply mechanistic interpretability techniques to study how transformers solve Theory of Mind tasks
  - Logit attribution
  - Head attribution
  - Attention analysis
  - Activation patching



References

T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spelke, J. Tenenbaum, and T. Ullman. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pages 9614–9625. PMLR, 2021.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Zhiming Hu

WS 23/24

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Motion In-betweening

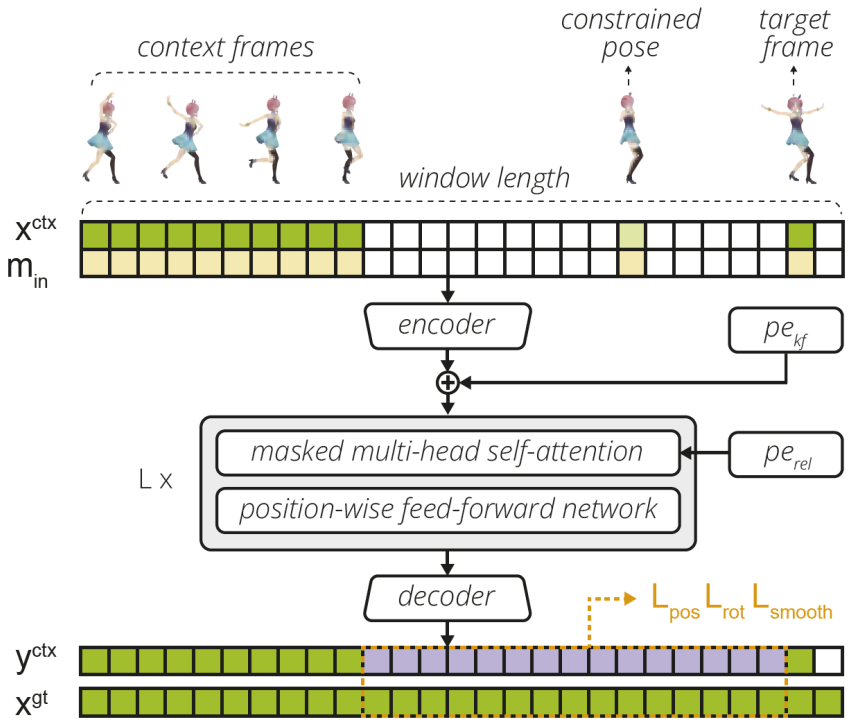
Problem definition

- Input: observed body poses + target pose
- Output: poses between the observed poses and target pose

Source: Qin TOG'22

Approach

- Input: observed body poses + target pose
- Output: poses between the observed poses and target pose



Source: Qin TOG'22

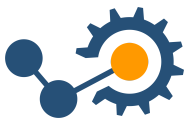
Future Frames-guided Motion In-betweening

Problem definition

- Input: observed body poses + target future poses
- Output: body poses between the observed and target poses

Novelty

- Learn effective features from future frames
- Fuse future features to generate motion in-betweening



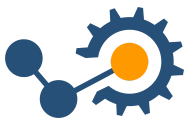
Gaze-guided Motion In-betweening

Problem definition

- Input: observed body poses and eye gaze + future poses and gaze
- Output: body poses between the observed and target poses

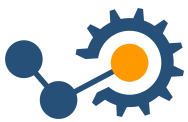
Novelty

- Learn effective features from eye gaze
- Fuse eye gaze features to generate motion in-betweening



References

Qin TOG'22. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics*, 41(6):1–16, 2022.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Constantin Ruhdorfer

WS 23/24

Perceptual User Interfaces Group, University of Stuttgart

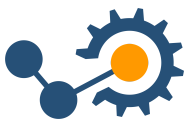
www.perceptualui.org

Overview

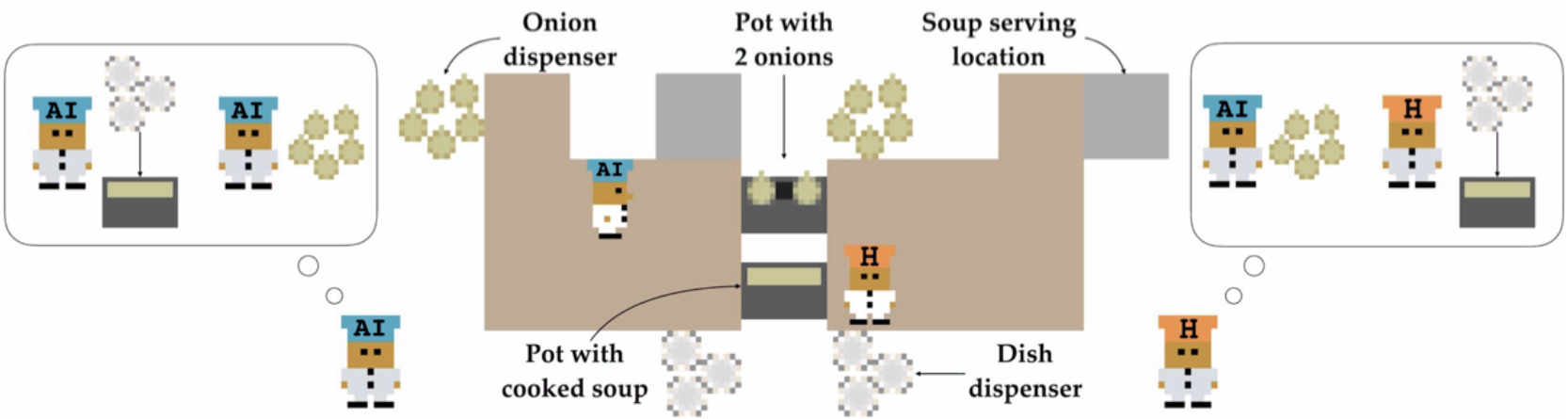
I offer two projects at the intersection of **RL**, **Human-AI Cooperation** and **Machine Theory of Mind**:

1. Comparing Imitation Learning Methods for building Human Models

2. Discovering Diverse Behaviour for Zero-Shot Cooperation



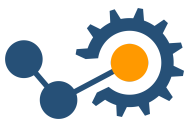
Overcooked



Source: [Carroll et al., 2019]

1) Comparing Imitation Learning Methods for building Human Models

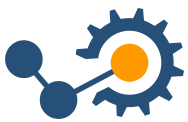
- Learned human models via imitation learning (IL) frequently are used to evaluate cooperative RL agents, e. g. Carroll et al. [2019]
- The solution space in IL is big (see [Ziebart et al., 2008])
- IL method probably affects the human model and thus the evaluation



1) Comparing Imitation Learning Methods for building Human Models

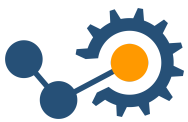
Your tasks:

- Retrain the behaviour cloning baselines (BC) from Carroll et al. [2019] in PyTorch using existing works like Gleave et al. [2022]
- Replace BC with two more SOTA approaches (pick from [Ho and Ermon, 2016; Wang et al., 2022] or [Gleave et al., 2022])
- Exhibit if they omit visually different strategies ...
- ... and if you can distinguish them during game-play among other agents I will provide



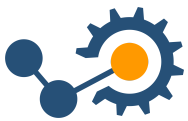
Example Self-Play Agent Performs Great with Itself … (SP with SP)

Self-Play – Self-Play



… but not with Others (Biased with SP)

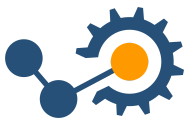
Biased-P. – Self-Play



2) Discovering Diverse Behaviour for Zero-Shot Cooperation

I propose to try:

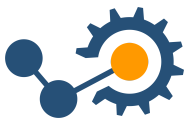
- Reward-Switching Policy Optimisation (RSPO) [Zhou et al., 2022]
- Finds new strategies by finding locally optimal but sufficently different policies
- Not yet evaluated in zero-shot cooperation



2) Discovering Diverse Behaviour for Zero-Shot Cooperation

Your tasks:

- Implement (Multi-Agent) RSPO for Overcooked and test on the layout Asymmetric Advantages
- Compare to Self-Play trained with MAPPO [Yu et al., 2022] (Mutli-Agent PPO [Schulman et al., 2017])
- Verify against strongly biased policies (see Yu et al. [2023])



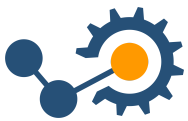
References

M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-ai coordination. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A. Gleave, M. Taufeeque, J. Rocamonde, E. Jenner, S. H. Wang, S. Toyer, M. Ernestus, N. Belrose, S. Emmons, and S. Russell. imitation: Clean imitation learning implementations. *arXiv preprint arXiv:2211.11972*, 2022.

J. Ho and S. Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL *http://arxiv.org/abs/1707.06347*.



References ii

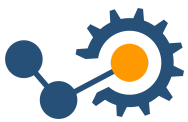
C. Wang, C. Pérez-D'Arpino, D. Xu, L. Fei-Fei, K. Liu, and S. Savarese. Co-GAIL: Learning Diverse Strategies for Human-Robot Collaboration. In *Proceedings of the 5th Conference on Robot Learning*, pages 1279–1290. PMLR, Jan. 2022.

C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

C. Yu, J. Gao, W. Liu, B. Xu, H. Tang, J. Yang, Y. Wang, and Y. Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations*, 2023.

Z. Zhou, W. Fu, B. Zhang, and Y. Wu. Continuously discovering novel strategies via reward-switching policy optimization. In *International Conference on Learning Representations*, 2022.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum Entropy Inverse Reinforcement Learning. 2008.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Chuhan Jiao

WS 23/24

Perceptual User Interfaces Group, University of Stuttgart

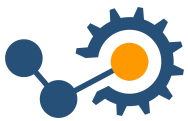
www.perceptualui.org

Gaze and Head Joint Representation Learning

Missing values in eye tracking

In eye tracking systems, missing values in gaze data due to blinks, pupil detection failures, make the data unsuitable for analysis and downstream tasks.

1.299718	(-0.22416(	(0.001733(	(-0.974313	(-0.332351	(-0.386718	(569.7052	569.7052	377.8484	(445.9551(	222.9776	184.082	(463.3526(	231.6763	184.0106
1.308049	(-0.22441(	(0.000418(	(-0.973631	(-0.322403	(-0.377975	(568.6189(	568.619	377.2063	(447.6668(	223.8334	184.0589	(465.4500(	232.725	183.8268
1.316381	(-0.22466(	(-0.00089(	(-0.972934	(-0.31726(	(-0.36824(	(567.5347(	567.5348	376.5613	(449.6978(	224.8489	184.1484	(466.0113(	233.0057	183.4837
1.324712	(-0.22491(	(-0.00221(	(-0.972225	(-0.309127	(-0.357772	(566.4530(	566.453	375.9136	(451.9657(	225.9829	184.5291	(467.5321(	233.7661	183.8273
1.333044	(-0.22534(	(-0.00392(	(-0.971011	(-0.314651	(-0.346267	(564.6904,	564.6904	375.0654	(453.8847(	226.9424	182.0093	(464.0019(	232.001	181.416
1.341375	(-0.22579(	(-0.00567(	(-0.969725	(-0.315185	(-0.342041	(562.8822(	562.8822	374.1964	(453.4291(	226.7146	178.1416	(462.0228(	231.0114	181.112
1.349707	(-0.22586(	(-0.00563(	(-0.969452	(NaN, NaN	(NaN, NaN	(562.5314,	562.5314	374.1944	(nan,nan)			(nan,nan)		
1.358038	(-0.22592(	(-0.00548(	(-0.96924(	(NaN, NaN	(NaN, NaN	(562.2677(	562.2677	374.2444	(nan,nan)			(nan,nan)		
1.36637	(-0.22597(	(-0.00533(	(-0.969028	(NaN, NaN	(NaN, NaN	(562.0040(	562.004	374.2946	(nan,nan)			(nan,nan)		
1.374701	(-0.22608(	(-0.00579(	(-0.967974	(NaN, NaN	(NaN, NaN	(560.6412,	560.6412	374.0486	(nan,nan)			(nan,nan)		
1.383033	(-0.22622(	(-0.00680(	(-0.966321	(NaN, NaN	(NaN, NaN	(558.5387(	558.5387	373.5579	(nan,nan)			(nan,nan)		
1.391364	(-0.22621(	(-0.00794(	(-0.965295	(NaN, NaN	(NaN, NaN	(557.2477,	557.2477	373.095	(nan,nan)			(nan,nan)		
1.399696	(-0.22603(	(-0.00907(	(-0.96467(	(NaN, NaN	(NaN, NaN	(556.4594(	556.4595	372.7628	(nan,nan)			(nan,nan)		
1.408027	(-0.22588(	(-0.01005(	(-0.964055	(NaN, NaN	(NaN, NaN	(555.7039(	555.704	372.4918	(nan,nan)			(nan,nan)		
1.416359	(-0.22575(	(-0.01078(	(-0.963465	(NaN, NaN	(NaN, NaN	(554.9970(	554.9971	372.3236	(nan,nan)			(nan,nan)		
1.42469	(-0.22552(	(-0.01169(	(-0.962719	(-0.038604	(-0.063308	(554.1149(	554.1149	372.1482	(533.8561(	266.9281	172.0006	(541.7614(	270.8807	178.9577
1.433022	(-0.22515(	(-0.01289(	(-0.961713	(-0.028289	(-0.050582	(552.9493(	552.9494	371.9656	(536.7628(	268.3814	176.4055	(543.8966(	271.9483	180.4416
1.441354	(-0.22459(	(-0.01417(	(-0.960497	(-0.038131	(-0.055282	(551.5731(	551.5732	371.8548	(533.8827(	266.9414	176.5245	(539.3710(	269.6855	180.5839
1.449685	(-0.22398(	(-0.01497(	(-0.959536	(-0.033248	(-0.054901	(550.5111(	550.5112	371.8554	(532.9428(	266.4714	177.5561	(539.8715(	269.9358	179.7863
1.458017	(-0.22351(	(-0.01549(	(-0.958875	(-0.055603	(-0.056793	(549.7847(	549.7847	371.8166	(531.6108(	265.8054	179.1654	(531.9917(	265.9959	180.2847
1.466348	(-0.22296(	(-0.01605(	(-0.958132	(-0.070770	(-0.051834	(548.9674(	548.9674	371.6919	(532.3804(	266.1902	177.9591	(526.3209(	263.1605	179.2665
1.47468	(-0.22242(	(-0.01657(	(-0.957607	(-0.065444	(-0.056259	(548.3970(	548.397	371.6265	(530.3941(	265.1971	179.5551	(527.4546(	263.7273	179.2407
1.483011	(-0.22188(	(-0.01709(	(-0.957167	(-0.067367	(-0.048416	(547.9228,	547.9228	371.5836	(532.4296(	266.2148	180.7656	(526.3651(	263.1826	179.7425



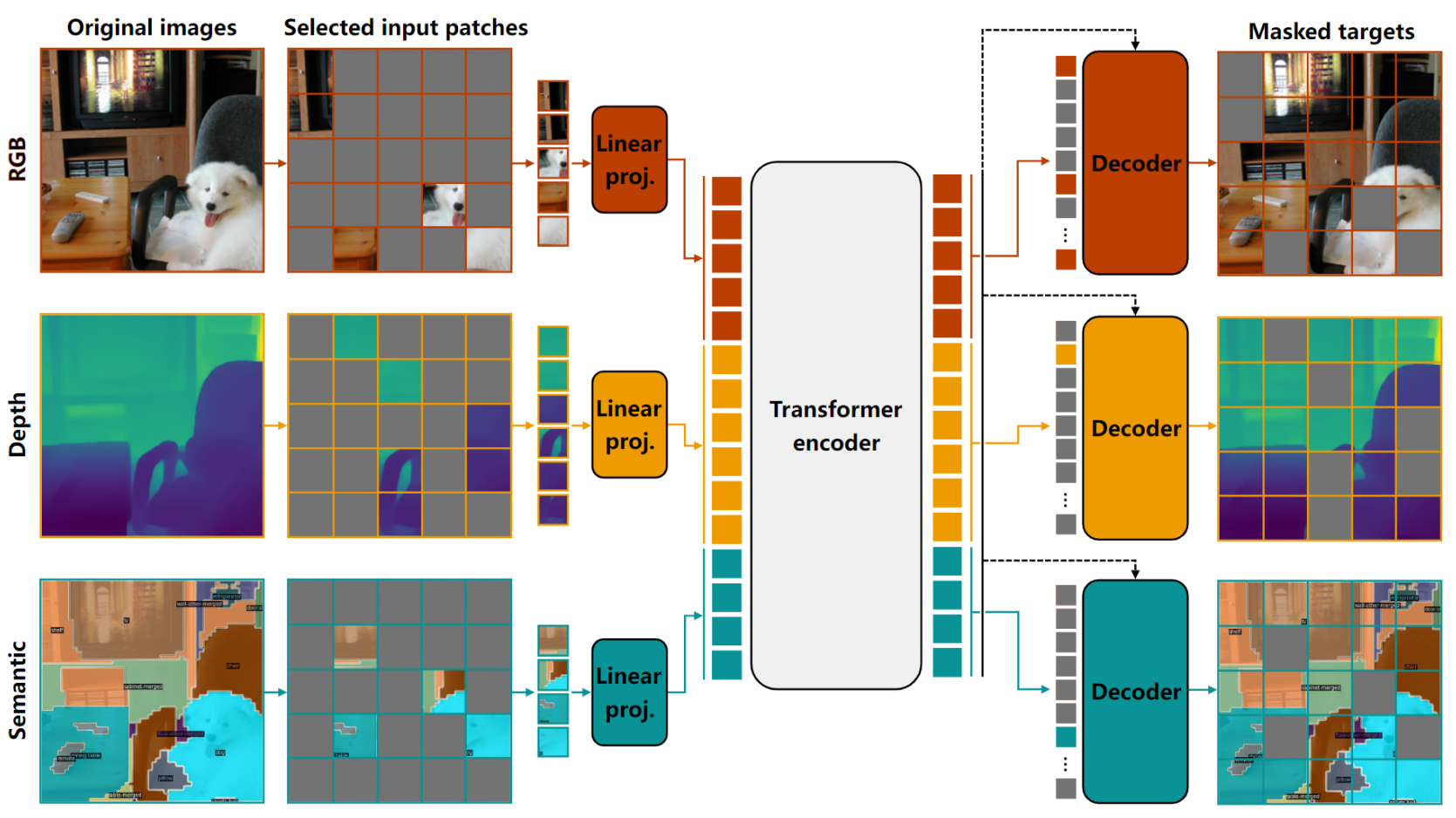
Head-eye coordination

1. Head movements and eye movements are correlated.

2. No missing values from head tracking sensors.

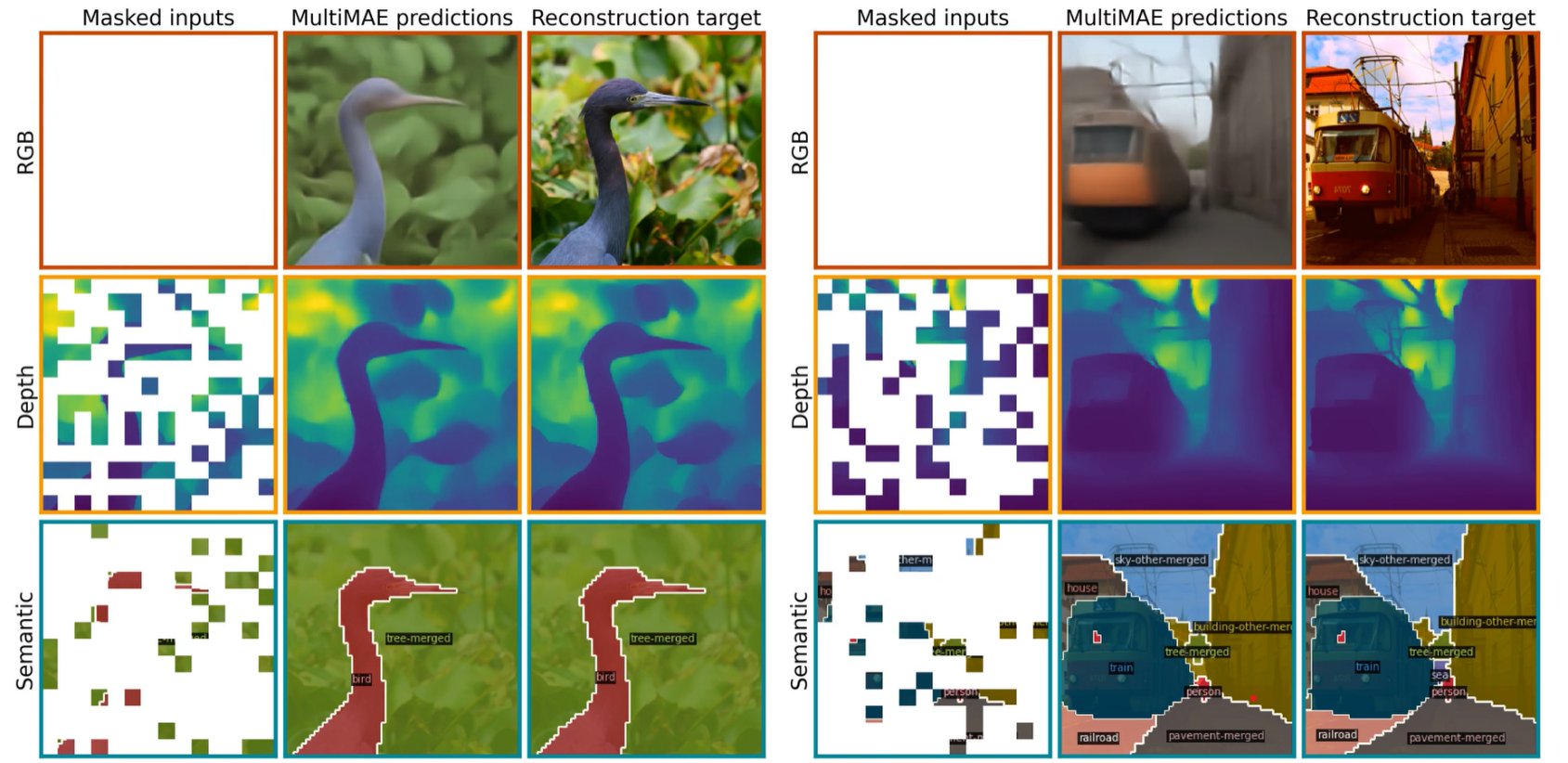
3. Can we use the head information in filling the missing gaze?

MultiMAE



Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." ECCV, 2022.

MultiMAE



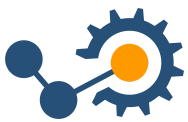
Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." ECCV, 2022.

Tasks

1. Build a multi-modal MAE for gaze-head joint representation learning following state-of-the-art methods

2. Evaluate the method of filling the missing gaze samples, increasing data efficiency, and increasing the performance of down-stream tasks.

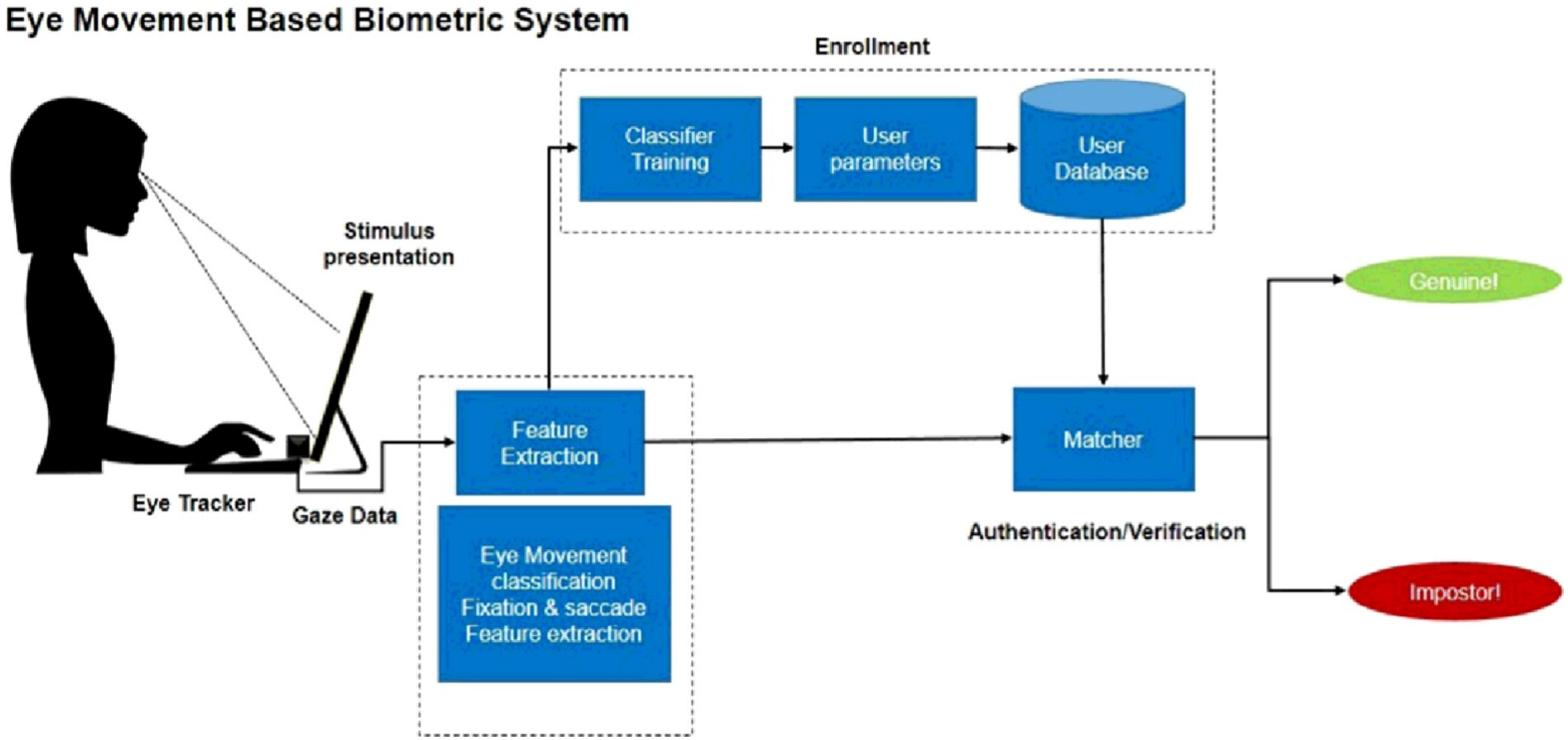
Dataset: *https://cuhksz-inml.github.io/head_gaze_dataset/*



Towards better understanding of gaze-based user authenticators

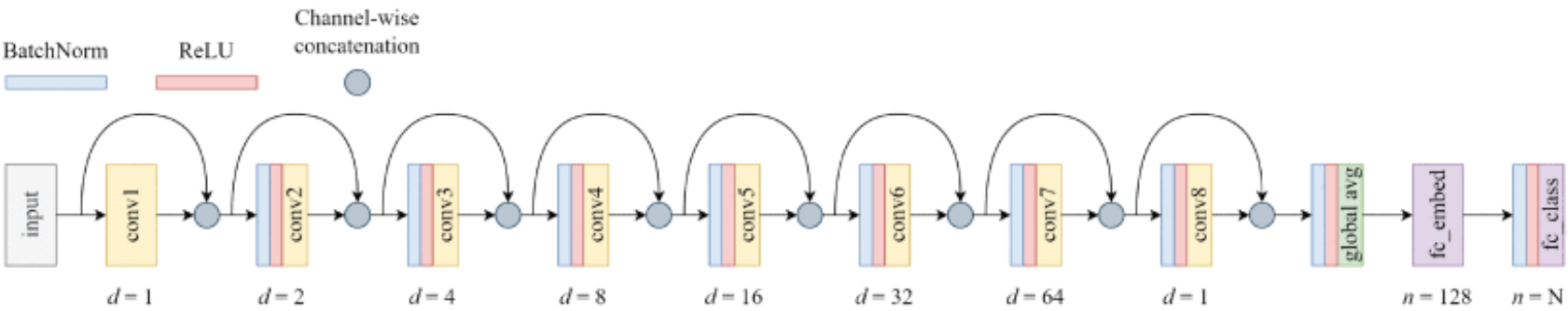
Eye movement biometrics

- Eye movement biometrics is a relatively recent behavioral biometric modality that may have the potential to become the primary authentication method in VR and AR



Eye Know You Too (EKYT)

EKYT is a CNN-based gaze-based user authentication approach that satisfies the FIDO Biometrics Requirements' recommendation of 5% false rejection rate at 1-in-10,000 false acceptance rate.



D. Lohr and O. V. Komogortsev, "Eye Know You Too: Toward Viable End-to-End Eye Movement Biometrics for User Authentication," in IEEE Transactions on Information Forensics and Security, 2022

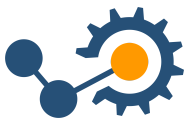
Tasks

1. Apply gradient-based attribution methods to a pre-trained EKYT model to analyze which parts in human eye movements are considered important towards identity. Compare the results of with hand-crafted features, e.g. fixations and saccades

2. Choose one from the following two
   - Implement a new user authentication model and compare it with EKYT
   - Implement a generative model, e.g. GAN, diffusion models, to attack EKYT.

EKYT code and pre-trained models:

https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/61ZGZN



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Stefan Geyer

WS 23/24

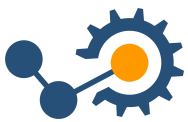
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Enhancing Computer Vision Models with Uncertainty Estimations

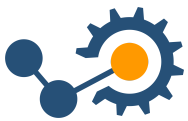
Motivation

- Exploration of uncertainty estimation in neural networks for computer vision.
- Importance of understanding uncertainty for real-world applications: How certain is my model?
- Utilization of pre-trained models (CLIP (Radford et al. [2021]) and\or DINOv2 (Oquab et al. [2023])) to extend with uncertainty estimation.



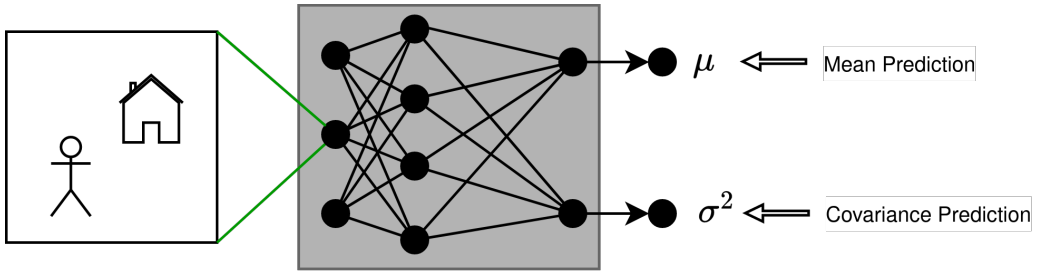
Problem

- We want to quantify the uncertainties in predictions made by computer vision models.
- For this we want to distinguish between epistemic and aleatoric uncertainty. (See Hüllermeier and Waegeman [2019]

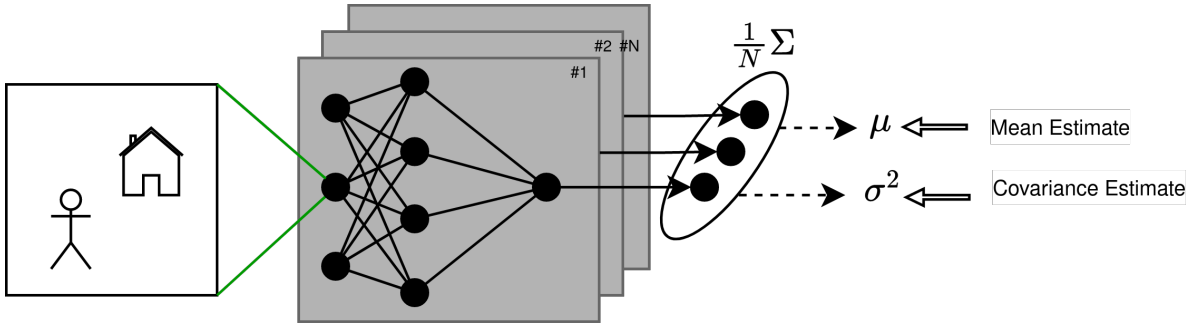


Approaches

- Both uncertainties can be estimated:



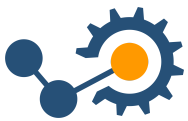
Aleatoric uncertainty: Inherent uncertainty of the dataset. Can not be removed but can be estimated by a model. Might occur if similar inputs have different ground truth targets without causal explainability.



Epistemic uncertainty: Uncertainty of learned model parameter. Can be removed by infinite amount of data. Very hard to predict by a model itself, but can be estimated by multiple model-samples (aka. Ensemble of models)

Datasets

- Training on ImageNet dataset for classification.
- Evaluation on COCO dataset and Cholec80 dataset.
- Qualitative evaluation: How do the uncertainties vary with small or high domain shift?

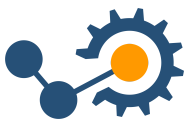


References

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *CoRR*, abs/1910.09457, 2019. URL *http://arxiv.org/abs/1910.09457*.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Malte Sönnichsen

WS 23/24

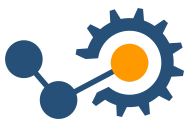
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Visual Foundation Models Insights: Exploring Dimensionality Reduction of Patch Embeddings

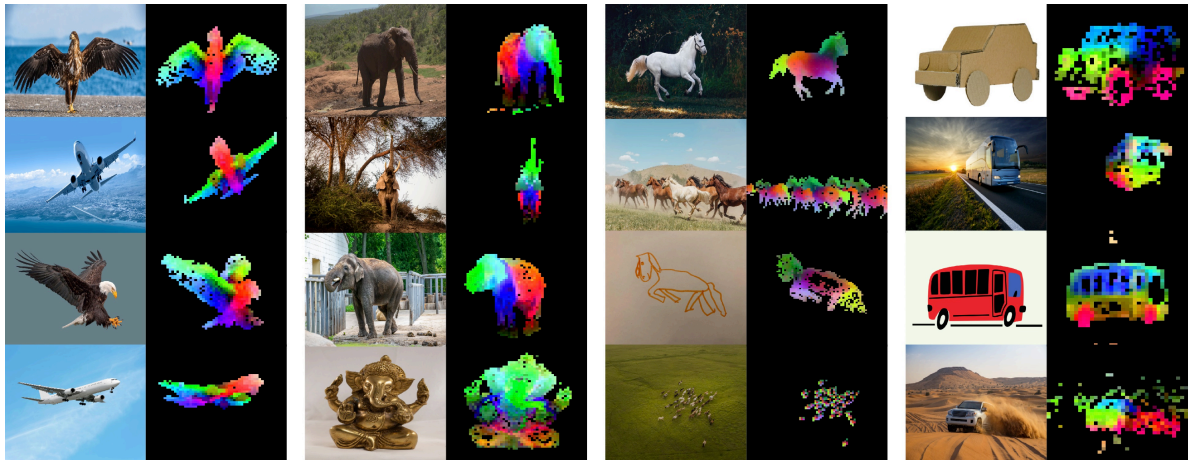
Motivation - Visual Foundation Models Insights

- Embedding space of Foundation Models contains rich information
- Multiple interesting use cases:
  - Zero shot image clustering/classification
  - Object localisation within an image
  - Zero shot segmentation



Segmentation

- Analyze the embedding space of foundation models, i.e., DINOv2 (Oquab et al. [2023]), GLIP (Li* et al. [2022]).
- Clustering of similar images within the embedding space.
- Investigate relationships of image parts within clusters

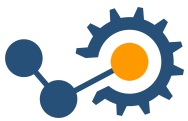


Source: DINOv2 Paper (*https://arxiv.org/pdf/2304.07193.pdf*)

References

L. H. Li*, P. Zhang*, H. Zhang*, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao. Grounded language-image pre-training. In *CVPR*, 2022.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Alina Roitberg

WS 23/24

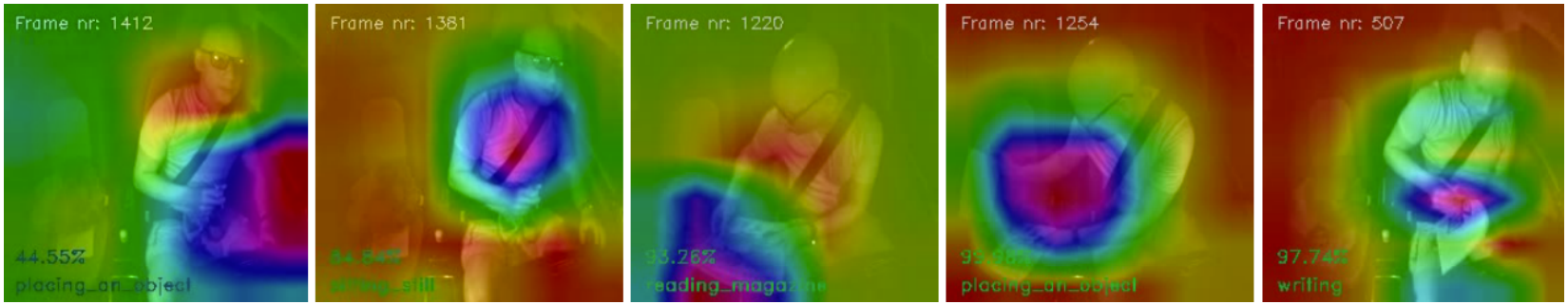
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Interpretable Driver Activity Recognition

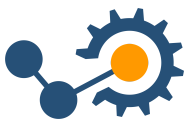
Introduction

- Investigating **interpretability** in machine learning models for driver activity recognition is essential for safety and trust.
- **Attribution methods** can uncover feature influence on model predictions, offering insights into model behavior.

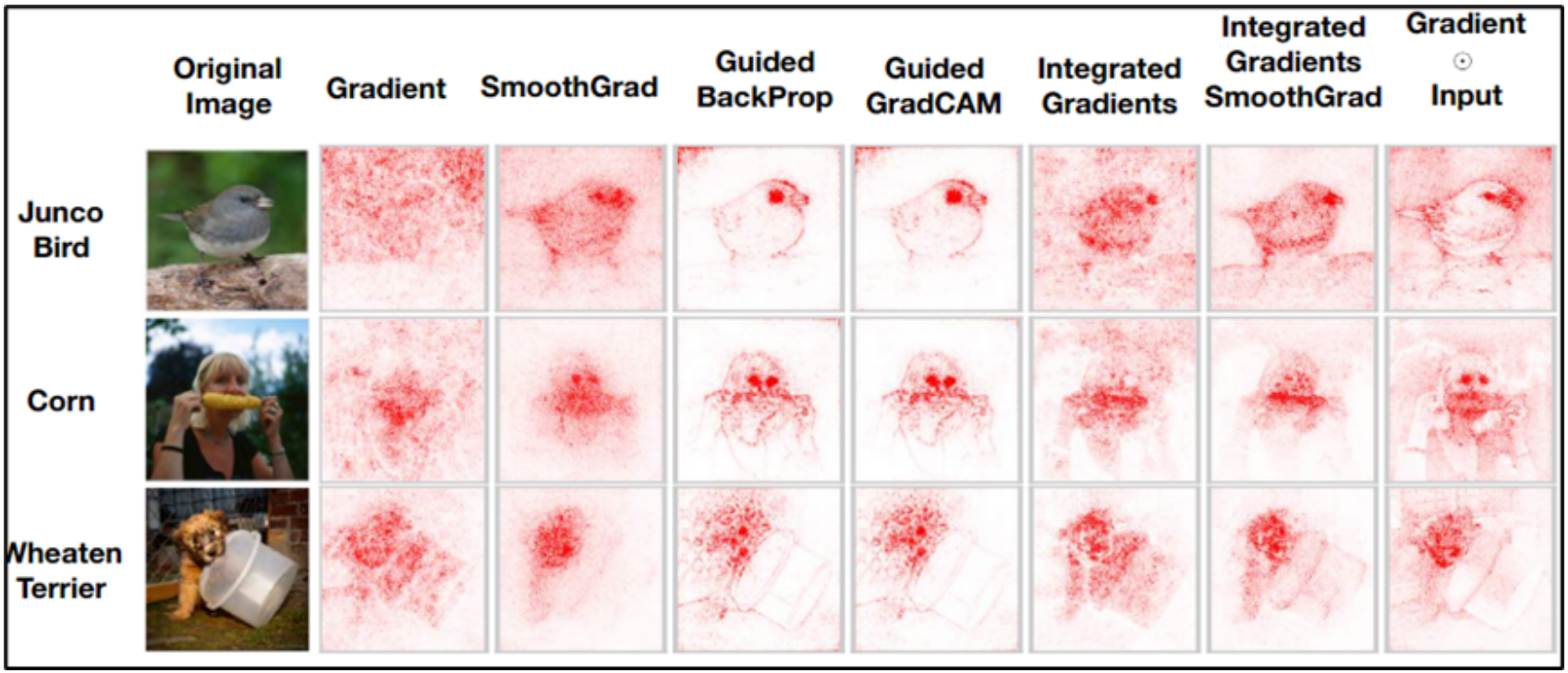


Objectives

- Get familiar with the Drive&Act dataset (Martin et al. [2019]) and at least one CNN-based recognition model used for driver activity recognition.
- Implement and compare two different attribution methods in the task of driver activity recognition (nice method overview: Adebayo et al. [2018]).
- Apply and evaluate attribution methods within a temporal context using the sliding window approach.



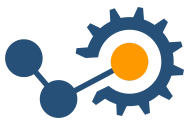
Examples of attribution methods



Source: Adebayo et al. [2018]

Results and Discussion

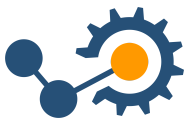
- Present the qualitative and, ideally, quantitative results of the attribution methods.
- Discuss the interpretability of the models in the context of temporal analysis.
- Evaluate the implications of the findings for real-world driver assistance systems.



References

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Alina Roitberg

WS 23/24

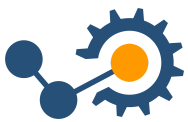
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Temporal Driver Activity Detection

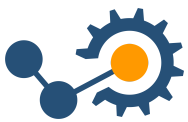
Introduction

- Driver activity detection involves identifying both the type and the timing of activities from sensor data.
- Temporal detection is crucial for understanding context and predicting driver behavior over time.
- The aim is to extend classification models to detect activities in a continuous input stream.



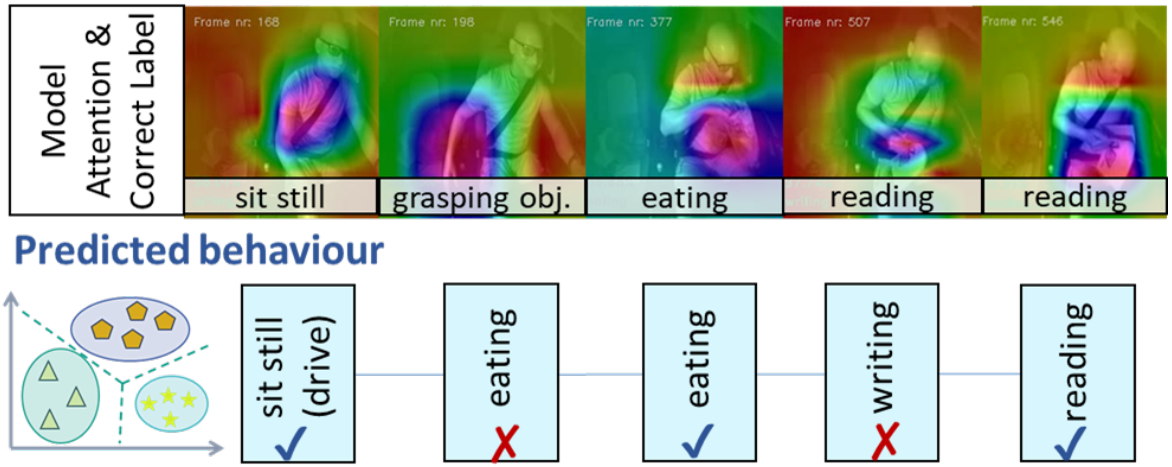
Objectives

- Formalize the task of temporal driver activity detection and establish appropriate metrics.
- Develop and evaluate a sliding window approach for activity segmentation.
- Implement a detection model that utilizes a further mechanisms to detect activity segments (e.g. temporal convolution, RNN).



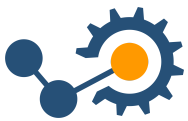
Goal 1: Task Formalization and Metrics

- Define temporal activity detection in the context of driver monitoring, e.g., using Drive&Act (Martin et al. [2019]).
- Choose metrics such as Intersection over Union (IoU) for window-based detection.



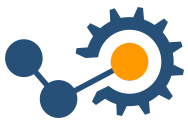
Goal 2: Sliding Window Mechanism

- Implement a standard model for driver activity classification and extend it with a sliding window mechanism.
- Explore window size and overlap to optimize performance.



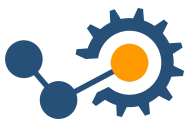
Goal 3: Neural Detection Model

- Implement an additional neural netowrk (e.g., RNN, attention-based network).
- Evaluate the approach in detecting driver activities, comparing it to sliding window.



References

M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Malte Sönnichsen

WS 23/24

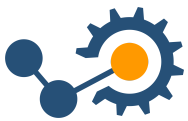
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Exploring Differential Privacy for Visual Affect Recognition

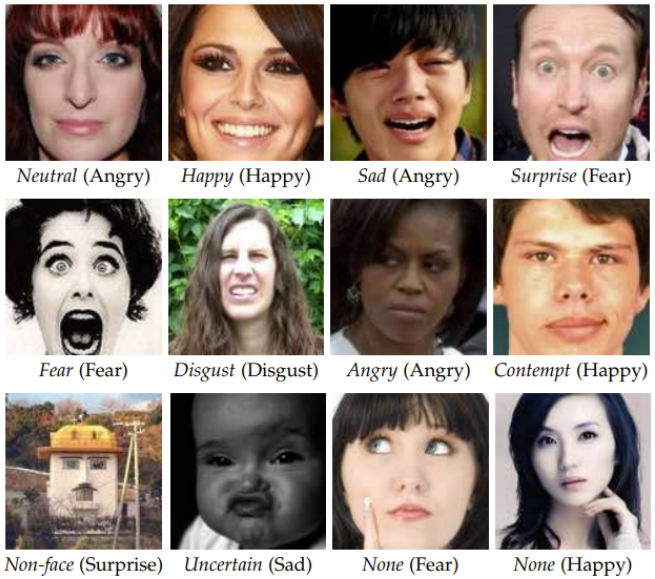
Motivation - Private Training

- Model weights may contain private information about individuals.
- Training data can be reconstructed given trained model weights.
- Training data may be private (Medical Data, Finance Data).



Differential Privacy

- Use Differential Privacy to prevent the leakage training data.
- Train a classifier on AffectNet (Mollahosseini et al. [2017]) using DP-SGD, i.e., Opacus (Yousefpour et al. [2021]).
- Investigate the trade-off between utility and privacy.

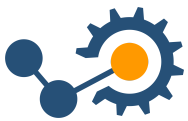


Source: Affect recognition on the AffectNet dataset.

References

A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A new database for facial expression, valence, and arousal computation in the wild. *IEEE Transactions on Affective Computing*, 2017. URL `http://mohammadmahoor.com/affectnet/`.

A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Alina Roitberg

WS 23/24

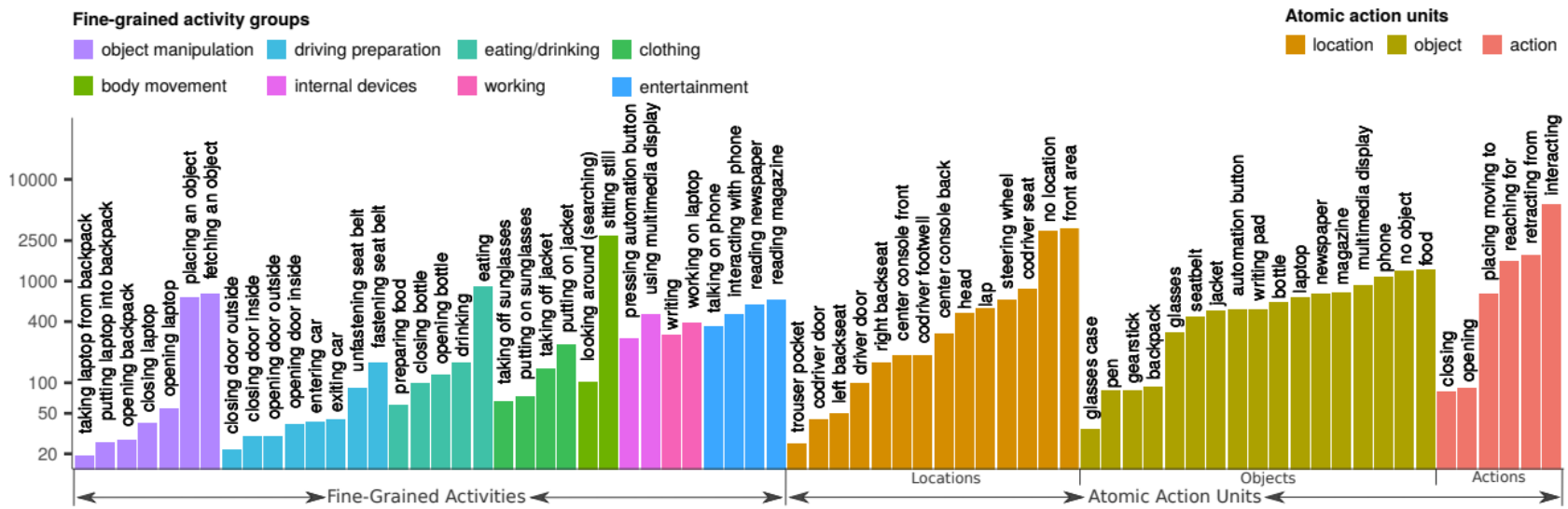
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Driver Activity Recognition with Imbalanced Data

Introduction

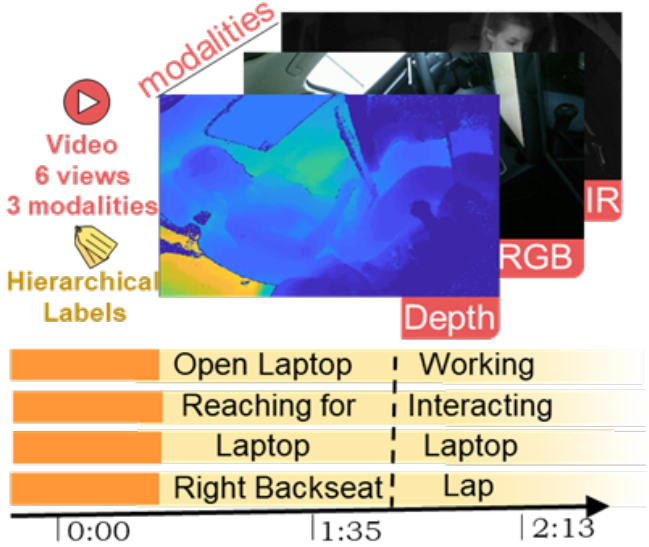
- **Imbalanced training data** is a prevalent issue in real-world datasets, including the Drive & Act dataset Martin et al. [2019] for driver activity recognition.



Source: Drive & Act dataset Martin et al. [2019]

Objectives

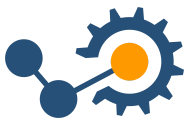
- Analyze the extent and the effect of data imbalance within the Drive & Act dataset.
- Implement and compare various techniques for handling imbalanced data.



Source: Enter Caption

Approaches

- Exploring over-sampling and under-sampling, and their effects on the dataset.
- Using SMOTE (Chawla et al. [2002]) or other generative models to balance class distribution.
- Adjusting the loss function to prioritize minority classes during training, (e.g., via the loss of Cao et al. [2019]).

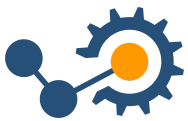


References

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Mayar Elfares

WS 23/24

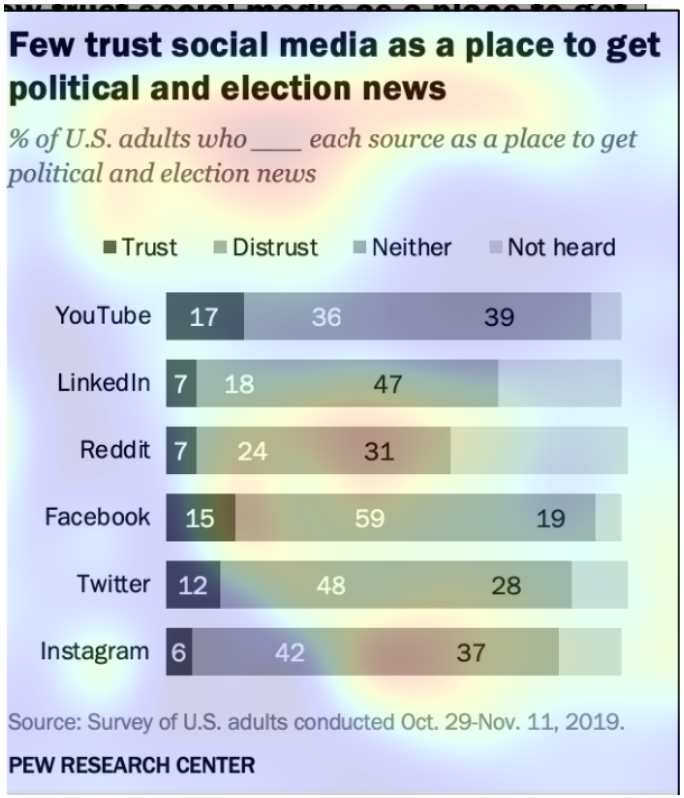
Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Saliency Inference Attack

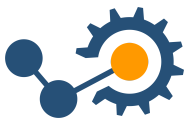
Saliency Inference Attack

- **Goal**: Perform a new side channel attack for model information stealing, i.e., infer information about the SalChartQA dataset.



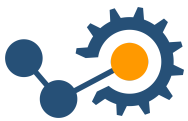
Tasks

- Given a visualisation, our goal is to infer information about the saliency prediction model.
- **Attack Model:**
  - Create a shadow training technique to generate a diverse set of data samples.
  - Train a classifier on the generated data.
  - Infer the information about the saliency prediction model.



References

- Zhang et al., 'A Plot is Worth a Thousand Words: Model Information Stealing Attacks via Scientific Plots', 32nd USENIX Security Symposium, 2023.



PUI Perceptual User Interfaces

ellis European Laboratory for Learning and Intelligent Systems

University of Stuttgart Germany

Machine Perception and Learning

for Interactive Intelligent Systems

Guanhua Zhang

WS 23/24

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

User Authentication and Identification in Mid-Air Typing

Mid-Air Typing in AR

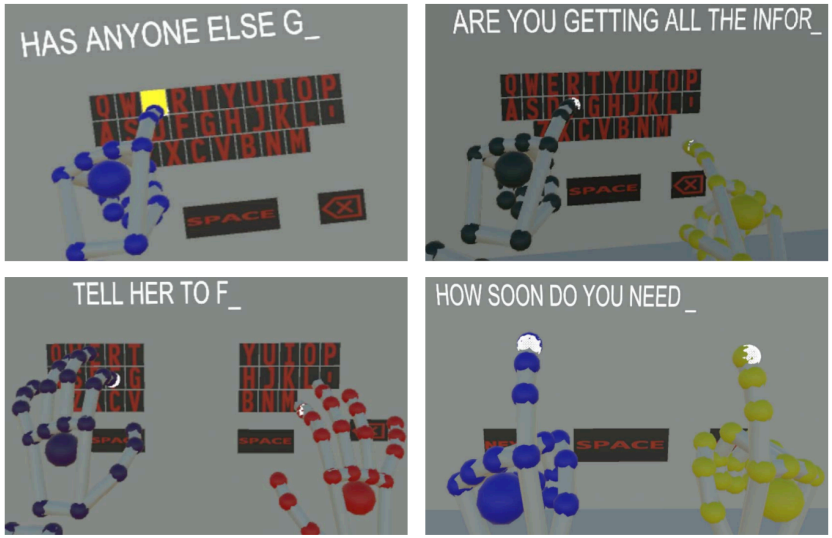


Fig. 1. Entering text using the normal keyboard with one hand (top left), the normal keyboard with two hands (top right), the split keyboard with two hands (bottom left), and the invisible keyboard (bottom right).

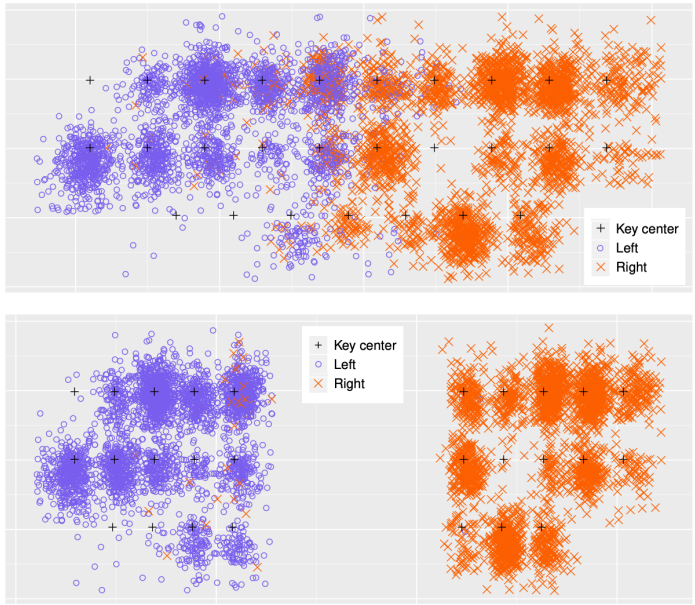


Fig. 4. Taps with the left and the right index finger in BIMANUAL (top) and SPLIT (bottom). Center of the keys are shown for better visualization.

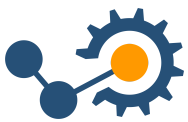
Adhikary and Vertanen, Typing on Midair Virtual Keyboards: Exploring Visual Designs and Interaction Styles,

INTERACT'21

Tasks

1. Build a user authenticator & identifier following state-of-the-art methods
    - Authentication: User A or not A
    - Identification: User A or C
2. Compare across keyboards and words

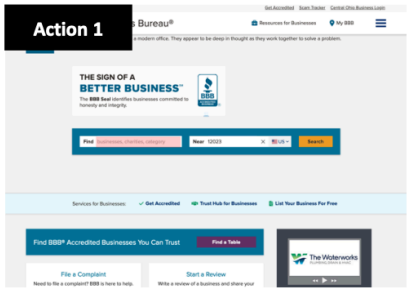
Dataset: *https://osf.io/5xwng*



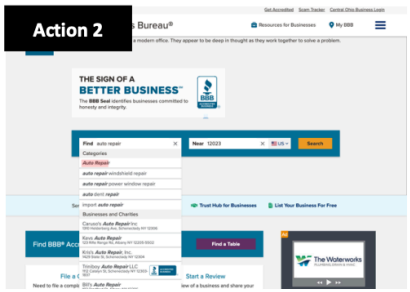
Retrieving Similar Desktop User Interfaces Using Screen2Vec

Desktop UIs

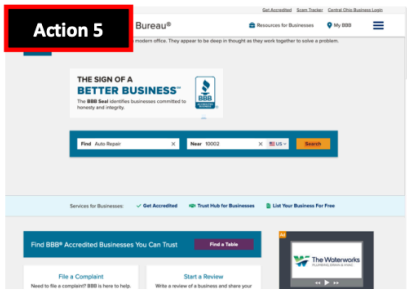
Webpage Snapshots:



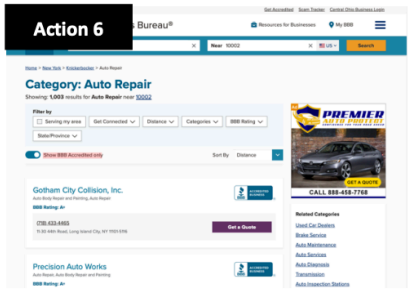
<input name="find_text" type="search">



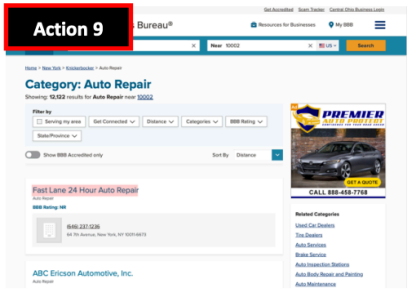
Auto Repair



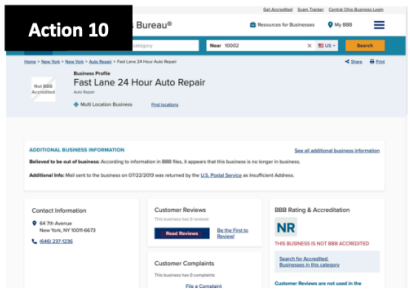
<button>Search</button>



<button>Show BBB Accredited only</button>



Fast Lane 24 Hour Auto Repair

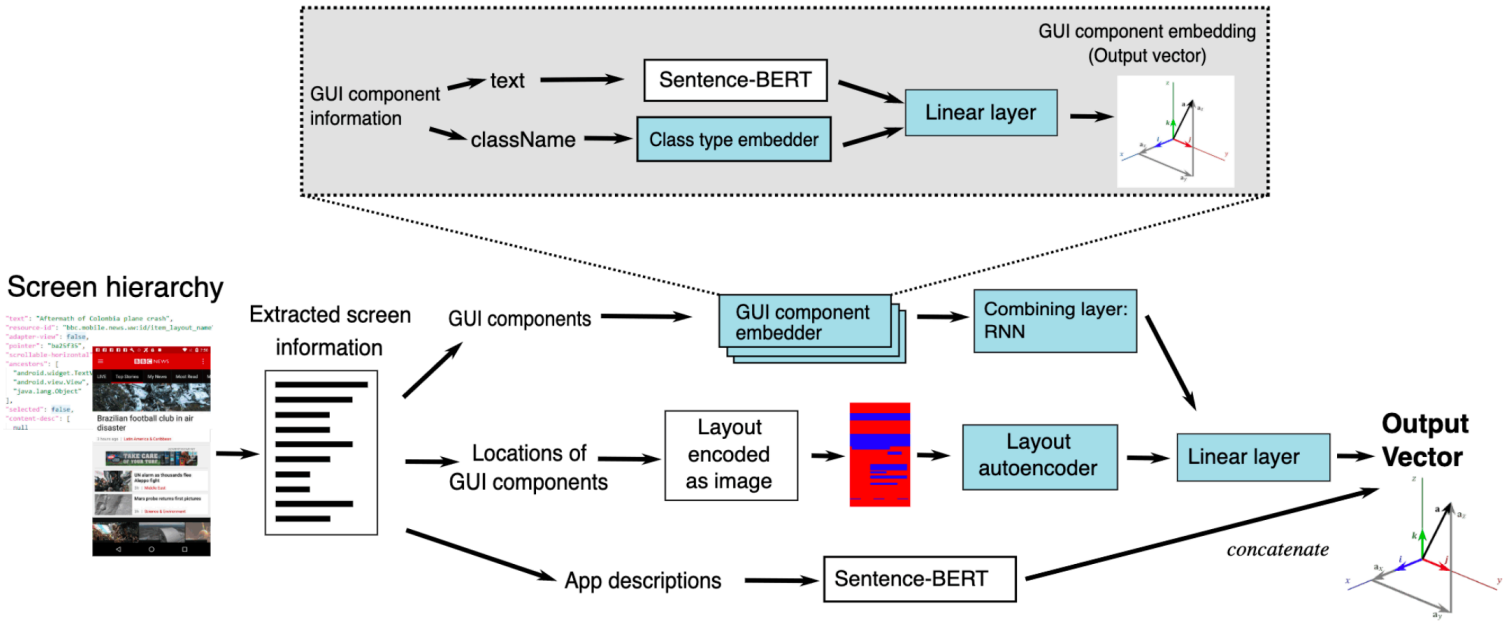


Read Reviews

Deng et al., MIND2WEB: Towards a Generalist Agent for the Web, 2023

Screen2Vec

- Learning an embedding of each **mobile** UI
- Retrieving similar UIs based on the cosine similarity between embeddings



Li et al, Screen2Vec: Semantic Embedding of GUI Screens and GUI Components, CHI'21

Tasks

1. Adapt Screen2Vec to desktop user interfaces
2. Get embeddings on UIs offered by the Mind2Web dataset
3. Retrieve similar UIs and analyse the results
    - Which features are captured and considered important? E.g., colors, layouts, websites and actions