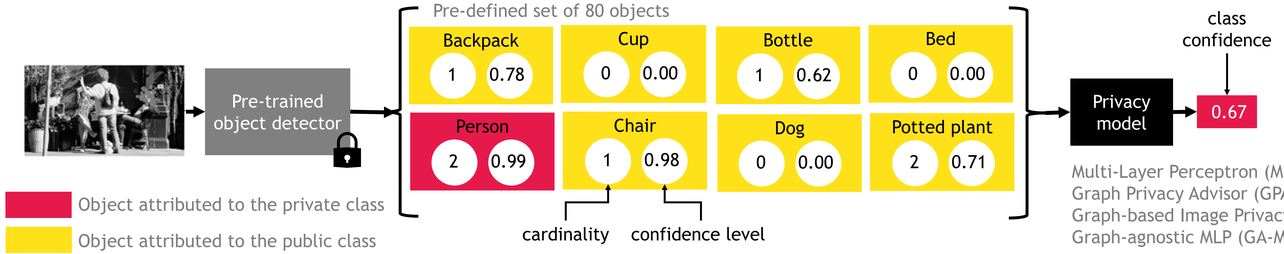
# Explaining models relating objects and privacy

Alessio Xompero, Myriam Bontonou, Jean-Michel Arbona, Emmanouil Benetos, Andrea Cavallaro
a.xompero@qmul.ac.uk, myriam.bontonou@ens-lyon.fr

## 1. Introduction

- Automatic warning for privacy protection
- Lack of privacy awareness associated with online media sharing
- Challenges of predicting images as private:
  - vast variety of content
  - *subjectivity* of privacy
- Considered methods for image privacy
  - **two-stage pipeline**
  - **no end-to-end training**

*Why a trained model predicts an image as private?*




- Post-hoc explainability method: Integrated Gradients (IG) [5]
- *Completeness axiom:* quantifying the contribution of the features of all objects towards the model's decision

Reference input $r^c = \mathbf{0}, \forall c : f_\theta(R) = 0$ ← public class

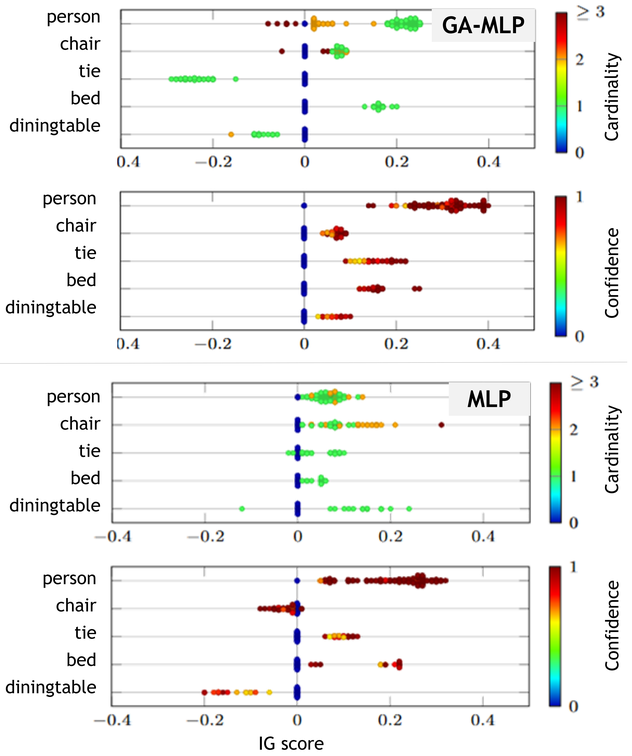
Privacy model $f_\theta(X) - f_\theta(R) = \sum_{c=0}^{C-1} \sum_{j=0}^{F-1} \phi(x_j^c)$ — Explainability score: $\in [-1,1]$

Input objects and features — object feature

$C$: number of objects (80), $F$: number of features (2)

## 3. Relevant objects and features for privacy

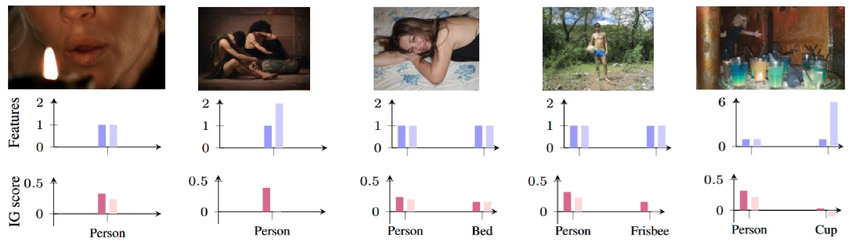
- Positive scores supporting privacy decisions by the models
- Multiple objects contributing to the decision




## 4. Examples of explaining privacy decisions

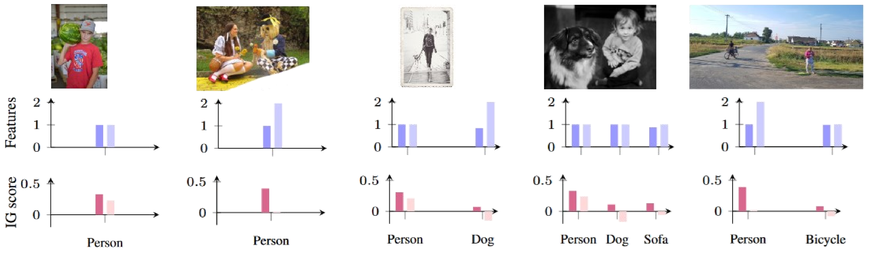
- Positive IG scores support privacy decisions by graph-agnostic MLP
- Images from PrivacyAlert training set [6]
- Confidence features (left bar), cardinality features (right bar)

Images **correctly** classified as private




high confidence in detecting a *person* significantly influences the model's decision

Images **incorrectly** classified as private




images labelled as *public* often **misclassified** due to the detection of *person*

## 5. Explainable-by-design person-centric classification

Two strategies
- S1: image classified as private if *at least one person* is detected
- S2: S1 + *maximum person cardinality of 2*

| Method | Public | | Private | | Overall | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| All private | 0.00 | 0.00 | 25.06 | 100.00 | 12.53 | 50.00 |
| All public | 74.94 | 100.00 | 0.00 | 0.00 | 37.47 | 50.00 |
| MLP | 86.29 | 82.32 | 53.52 | 60.89 | 69.90 | 71.60 |
| GPA* | 75.30 | 97.62 | 37.25 | 4.22 | 56.28 | 50.92 |
| GPA*^ | 74.94 | 100.00 | 0.00 | 0.00 | 37.47 | 50.00 |
| GIP* | 74.94 | 100.00 | 0.00 | 0.00 | 37.47 | 50.00 |
| Graph-agnostic MLP | 88.87 | 77.71 | 51.53 | 70.89 | 70.20 | 74.30 |
| Person-centric (S1) | 94.76 | 55.05 | 40.34 | 90.89 | 67.55 | 72.97 |
| Person-centric (S2) | 89.67 | 73.55 | 48.55 | 74.67 | 69.11 | 74.11 |

* adapted for fair comparison: *Degenerate to (almost) all public*
^ corrected implementation

Images with people not necessarily private

Most of private images has people (high recall)

## 6. Conclusions

https://github.com/graphnex/ig-privacy

- Identified and quantified *relevant objects features* for privacy models' decision
- Person-centric strategies as *reference baselines* for future comparisons
- MLP and Graph-agnostic *biased* towards the presence of the object *person*

## Acknowledgment

graphnex.eecs.qmul.ac.uk

This work was supported by the CHIST-ERA programme through the project GraphNEx, under UK EPSRC grant EP/V062107/1, France ANR grant ANR-21-CHR4-0009, and Swiss NSF grant 195579.

## References

[1] A. Tonge and C. Caragea, "Image privacy prediction using deep features", *AAAI*, 2016
[2] D. Stoidis and A. Cavallaro. Content-based Graph Privacy Advisor", *IEEE BigMM*, 2022
[3] G. Yang, J. Cao, *et al.*, "Graph-based Neural Networks for Explainable Image Privacy Inference", *Patt. Rec.*, 2020
[4] V. P. Dwivedi, C. K. Joshi, *et al.*, "Benchmarking Graph Neural Networks", *JMLR*, 2023
[5] M. Sundararajan, A. Taly, Q. Yan, "Axiomatic attribution for deep networks", *ICML*, 2017
[6] C. Zhao, J. Mangat, *et al.*, "PrivacyAlert: A Dataset for Image Privacy Pre-diction", AAAI ICWSM, 2022