PAL: Intelligence Augmentation using Egocentric Visual Context Detection

Mina Khan
MIT Media Lab
75 Amherst St, Cambridge, MA, USA
minakhan01@gmail.com

Pattie Maes
MIT Media Lab
75 Amherst St, Cambridge, MA, USA
pattie@media.mit.edu

Abstract

Egocentric visual context detection can support intelligence augmentation applications. We created a wearable system, called PAL, for wearable, personalized, and privacy-preserving egocentric visual context detection. PAL has a wearable device with a camera, heart-rate sensor, on-device deep learning, and audio input/output. PAL also has a mobile/web application for personalized context labeling. We used on-device deep learning models for generic object and face detection, low-shot custom face and context recognition (e.g., activities like brushing teeth), and custom context clustering (e.g., indoor locations). The models had over 80% accuracy in in-the-wild contexts (~1000 images) and we tested PAL for intelligence augmentation applications like behavior change. We have made PAL is open-source to further support intelligence augmentation using personalized and privacy-preserving egocentric visual contexts.

1. Introduction

Egocentric visual contexts have been useful in context-aware intelligence augmentation [9], e.g., memory augmentation and assistive technology [31, 8]. However, visual context tracking, especially sending user data to the cloud for deep learning, raises privacy concerns and deep learning models also have to be personalized for each user.

We created a wearable system, called PAL (Figure 1), for personalized and privacy-preserving egocentric visual context recognition using on-device, human-in-the-loop, and low-shot deep learning. PAL uses on-device deep learning for real-time and privacy-preserving processing of user data, and includes multimodal sensing, e.g., using camera, heart-rate, physical activity, and geolocation [14]. PAL also supports user input for human-in-the-loop training of personalized visual contexts. We used on-device models for generic object and face detection, personalized and low-shot custom face and recognition, and semi-supervised custom context clustering [19]. Compared to existing wearable systems, which use at least 100 training images per custom context [21] and do not use privacy-preserving on-device deep learning, PAL’s on-device models for low-shot and continual learning use ~10 training images per context. Also, PAL uses active learning for context clustering so that the users do not have to explicitly train different contexts.

We make three contributions: i. a wearable device for privacy-preserving and personalized egocentric visual context detection using on-device and human-in-the-loop deep learning; ii. a system for recognizing custom contexts, faces, and clusters using low-shot, custom-trainable, and active learning; iii. real-world applications and evaluations.

2. Related Work

PAL is a wearable system for personalized and privacy-preserving egocentric visual context detection using on-device, human-in-the-loop, and low-shot deep learning. PAL also includes multimodal sensors and input/outputs. The existing wearable systems for egocentric visual context detection do not support multimodal sensing and on-device, human-in-the-loop, and low-shot deep learning like PAL.

2.1. Wearable Cameras and Deep Learning

Wearable cameras are commonly used for intelligence augmentation applications [31, 8] and even combined with physiological sensors [11]. Deep learning has also been used with egocentric cameras, e.g., for predicting daily activities [3], eating recognition [2], visual assistance [26, 27], visual guides [30], and face recognition [5]. However, none use on-device deep learning, especially for personalized and privacy-preserving egocentric visual contexts. On-device deep learning systems have also been used for computer vision [25], but they do not support personalized, low-shot, and human-in-the-loop visual context detection.

2.2. Privacy-preserving Deep Learning

Privacy-preserving approaches include privacy-preserving collaborative deep learning for human activity recognition [24] and image distort or modification [7, 1]. However, none of these systems use on-device deep learning to avoid sending data to the cloud for processing.
2.3. Personalized Deep Learning

Wearable deep learning-based egocentric visual context recognition has been used for personalized object and face recognition [20, 21]. However, they do not use on-device deep learning for privacy-preserving context detection and also need ~100 images for training each personalized class, whereas our custom context and face recognition models use 1 to 10 images. There are also approaches for low-shot and continual deep learning [28] and active deep learning [12], including for image clustering [4]. However, none have been deployed in in-the-wild wearable settings, especially for on-device and personalized deep learning.

3. Design and Implementation

We created a wearable system, called PAL, with multimodal sensing (camera, geolocation, heart-rate, and physical activity), on-device deep learning, and audio input/output. PAL’s wearable device, on-device deep learning models, and mobile/web app are described below.

We made 3 design decisions to enable personalized and privacy-preserving egocentric visual context detection. First, we decided to use a head-mounted egocentric camera to capture the user’s visual context, but decided to keep the camera small, light, and non-conspicuous. Second, we decided to use on-device deep learning for model training and inference so that user images were not sent to the cloud for processing and can be processed on-device for privacy-preserving sensing, e.g., hiding bystander’s faces. Third, we decided to enable human-in-the-loop and active learning to learn personalized contexts for intelligence augmentation.

3.1. Wearable Device

PAL’s wearable device has an ear-hook and an on-body component (Figure 2). The ear-hook has a camera, heart-rate sensor, and open-ear audio output. The on-body component has a microprocessor and a deep learning accelerator. The device takes periodic pictures, has a 5-hour battery life (0.3A at 5.25V, 2500mAh battery). The device supports audio input/output and the users can press a button for starting/stopping custom face and context training sessions.

3.2. On-device Deep Learning Models

We added on-device visual context detection models for three tasks: i. generic object and face detection; ii. low-shot custom face and context recognition; iii. context clustering for efficient active learning of visually-similar contexts. All models are trained and inferred on the wearable device.

Object and Face Detection: The Object Detection model is trained on 90-item Common Objects in Context (COCO) dataset [22] and Face Detection on Open Images v4. Both models use the MobileNet SSD v2 architecture [13, 23].

Low-shot Custom Face and Context Recognition: We use MobileNetv1 architectures [13] for both. Custom Face Recognition models using FaceNet [29] and needs 1-2 custom training images per face. Custom Context Recognition model recognizes custom activities, e.g., brushing teeth, using weight imprinting [28]. Weight imprinting adds new classes to the existing list of classes for continual learning and needs ~10 training images per class. We used weight imprinting on embeddings from an Image Classification model pre-trained on the 1000-class ImageNet dataset [6].

Context Clustering for Active Learning: We cluster visually-similar contexts, separated by geolocations, using an image embedding generator combined with clustering. We generated embeddings using an Image Classification model (MobileNet v1), pre-trained on 1000-class ImageNet dataset. We use Density-based Spatial Clustering of Applications with Noise (DBSCAN) [10] for clustering as DBSCAN does not require a predetermined number of clusters.

3.3. Mobile/Web Application

The mobile/web application supports geolocation and physical activity tracking, personalized data labeling and visualization, and custom intervention-setting, e.g., just-in-time reminders for habit-formation support [18]. Custom contexts, context clusters, and faces are shown on the mobile app for labeling. Figure 3 shows the labeling interface.
4. Evaluations, Applications, and Future Work

We performed technical evaluations of PAL. We tested PAL’s deep learning models with 4 participants for 2 days each (~1000 in-the-wild images) [19]. Each model had over 80% accuracy – Object Detection: 98.8% (F1 = 0.79, ~1000 instances); Face Detection: 88.8% (F1 = 0.9, ~180 instances); Custom Face Recognition: 86.9% (4 faces, 120 instances); Custom Contexts: 87.2% (7 activities, ~350 images); Custom Clusters: 82% (19 contexts, ~300 images). Even images partially occluded by cheeks or hair were accurately predicted. Inference time was ~3 sec per model and ~15 sec for 5 models. Moreover, the camera captured ~70% of the user’s visual contexts and the heart-rate sensor accuracy was also comparable to state-of-the-art sensors [14].

We used PAL for real-world habit formation support and showed that habit support in egocentric visual contexts can double habit-formation compared to state-of-the-art techniques [18, 15]. We also built applications for language learning and memory support using PAL [17].

Users want behavior change support in multimodal contexts [16], and we envision that PAL can be used for context-aware behavior change support. PAL’s egocentric visual context detection can also be deployed for other intelligence augmentation applications, e.g., memory support for Alzheimer’s patients and caretakers. More sensors can be also added to PAL’s open-source platform and further work on on-device deep learning will enable better context detection using PAL, e.g., audio-visual context detection.

5. Conclusion

Context-awareness is key for intelligence augmentation and egocentric visual contexts can provide rich information about a user’s environmental and behavioral contexts. We created a wearable system, called PAL, for personalized and privacy-preserving egocentric visual context detection. PAL uses on-device deep learning for privacy-preserving processing, and we also added multimodal sensing (heart-rate, geolocation, and physical activity) and seamless audio input/output to PAL. We deployed on-device, low-shot, and human-in-the-loop deep learning models on PAL to recognize generic faces and objects as well as custom faces, contexts (e.g., activities like brushing teeth), and context clusters (e.g. indoor locations). We tested PAL in in-the-wild wearable settings and also used it for real-world language learning, behavior change, and memory augmentation applications. Thus, using on-device, low-shot, and human-in-the-loop deep learning, PAL paves the way for intelligence augmentation applications using personalized and privacy-preserving egocentric visual contexts.

References

[1] Rawan Alharbi, Mariam Tolba, Lucia C Petito, Josiah Hester, and Nabil Alshurafa. To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 3(3):1–29, 2019.

[2] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[3] Daniel Castro, Steven Hickson, Vinay Bettadapura, Edison Thomaz, Gregory Abowd, Henrik Christensen, and Irfan Essa. Predicting daily activities from egocentric images using deep learning. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ISWC ’15, pages 75–82, New York, NY, USA, Sept. 2015. Association for Computing Machinery.

[4] Luiz F. S. Coletta, Moacir Ponti, Eduardo R. Hruschka, Ayan Acharya, and Joydeep Ghosh. Combining clustering and active learning for the detection and learning of new image classes. *Neurocomputing*, 358:150–165, Sept. 2019.

[5] Ovidiu Daescu, Hongyao Huang, and Maxwell Weinzierl. Deep learning based face recognition system with smart glasses. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA ’19, pages 218–226, New York, NY, USA, June 2019. Association for Computing Machinery.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Mariella Dimiccoli, Juan Marin, and Edison Thomaz. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degra-
