Bottleneck-based Encoder-decoder ARchitecture (BEAR) for Learning Unbiased Consumer-to-Consumer Image Representations

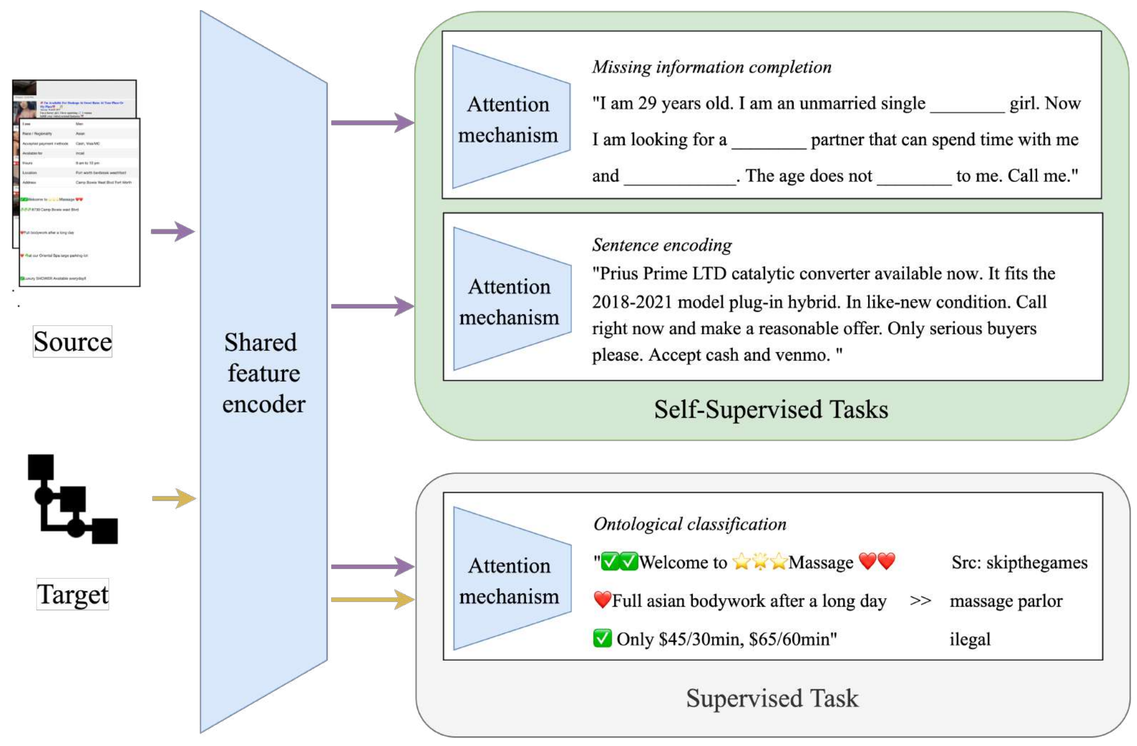
Pablo Rivas,[1] Gisela Bichler,[2*] Tomas Cerny,[1*] Laurie Giddens,[3*] and Stacie Petter[4*]

*Equal Contribution. [1]Department of Computer Science, School of Engineering and Computer Science, Baylor University, Texas, USA. [2]School of Criminology and Criminal Justice, California State University, San Bernardino, California, USA. [3]Information Technology and Decisions Sciences Department, University of North Texas, USA. [4]School of Business, Wake Forest University, North Carolina, USA

Contributions

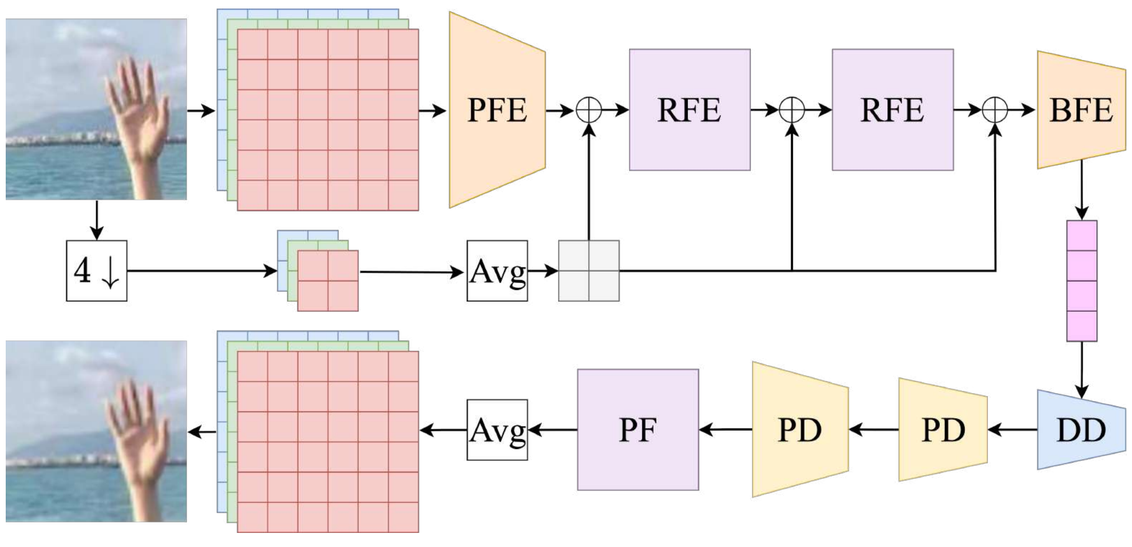
- We design an autoencoder model that learns rich features that produce latent representations of images.
- Our model uses modern machine learning techniques to learn low-dimensional representations at scale with few parameters in comparison to other similar models.
- The proposed model hides personal identifiers from plain sight while keeping necessary information for downstream trafficking detection tasks.

Background and Motivation

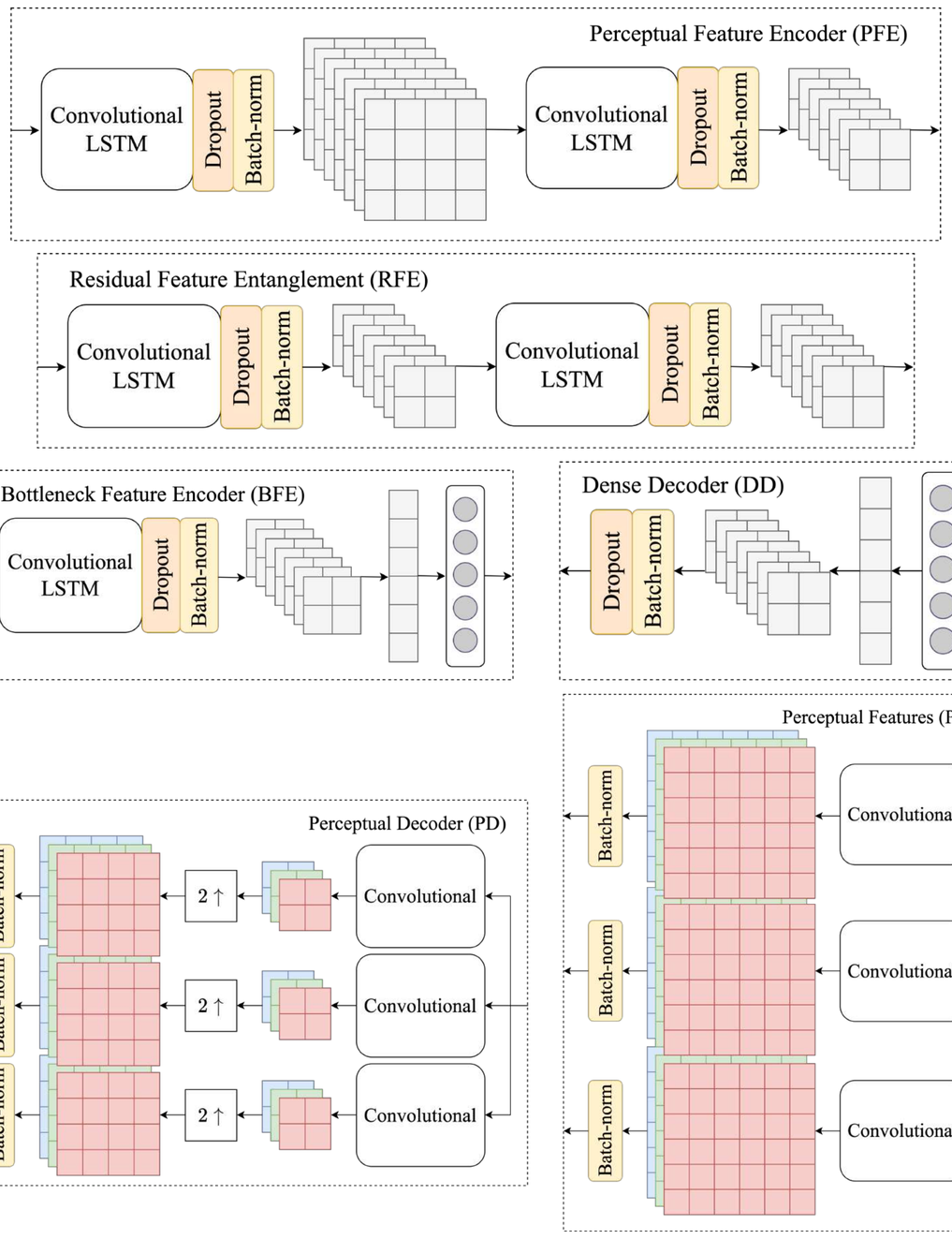


- **Consumer-to-Consumer Marketplaces**. With the COVID-19 global pandemic, C2C marketplaces became heavily used, fostering an increase in illegal activities such as human trafficking and trafficking of stolen goods.
- **Big Picture**. We want to extract multimodal features from C2C data using attention mechanisms on self-supervised, contrastive, and supervised learning tasks.
- **SOTA**. Most transformer-based approaches such as BYOL or ViT, are data-hungry and too large to train from scratch or fine-tune. Other big models such as CLIP or GPT can cost hundreds of millions of dollars to train. We need lightweight, scalable, image representations that hide personally identifiable information from plain sight.

BEAR Design Paradigm

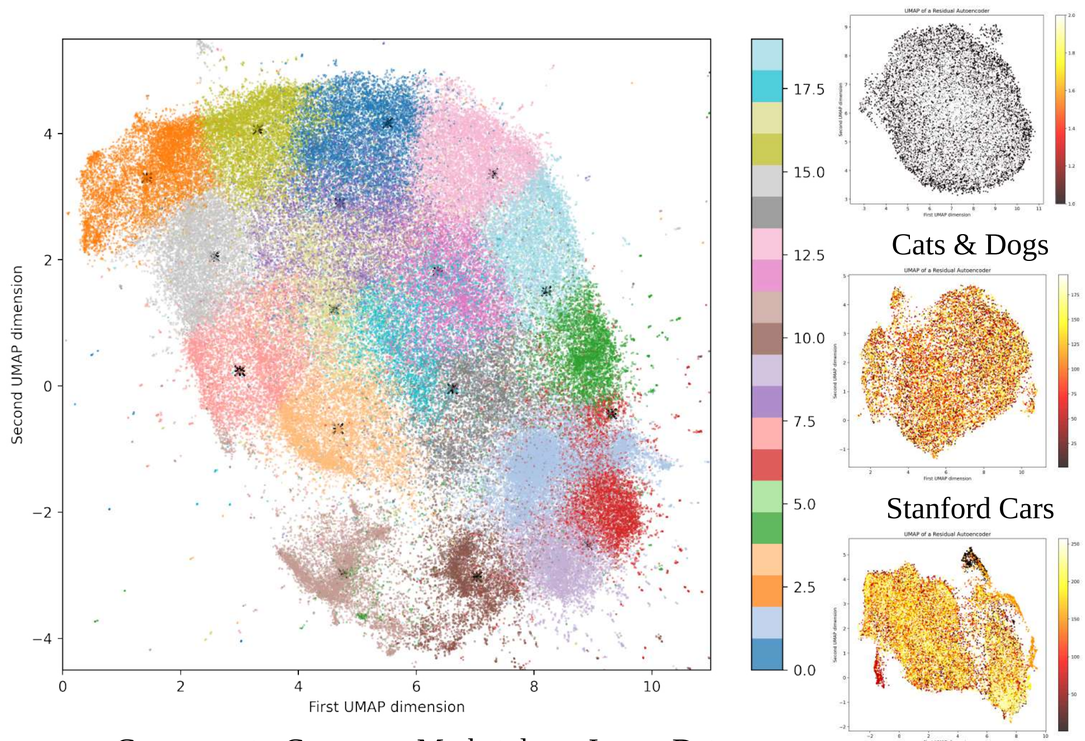


This preliminary study was funded by the National Science Foundation under grants CHE-1905043, CNS-2136961, and CNS-2210091.

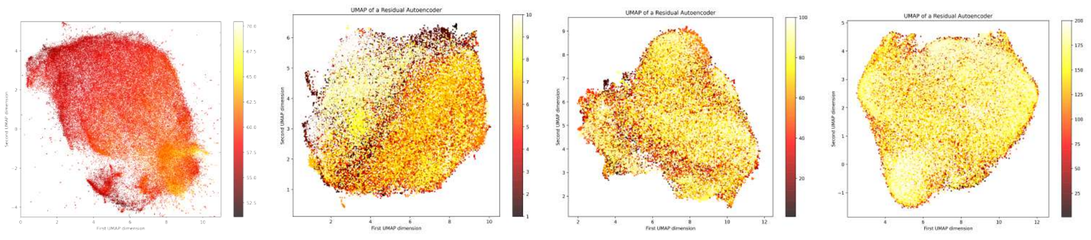


Why Does It Work?

- Data augmentation was key to successful self-supervision.
- High-quality latent space due to perceptual feature learning.
- Scalability granted by convolutional-based design.
- Baseline model is only 4M, cf. ViT 86M, BYOL 250M.
- **Next steps**: label-based self & weak supervision, fine-tuning.



Consumer-to-Consumer Marketplaces Image Data.



Code

ICML International Conference On Machine Learning

LXAI

NSF

BU ENGINEERING & COMPUTER SCIENCE