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

Pablo Rivas¹ Gisela Bichler^{*2} Tomas Cerny^{*1} Laurie Giddens^{*3} Stacie Petter^{*4}

Abstract

Unbiased representation learning is still an object of study under specific applications and contexts. Novel architectures are usually crafted to resolve particular problems using mixtures of fundamental pieces. This paper presents different image feature extraction mechanisms that work together with residual connections to encode perceptual image information in an autoencoder configuration. We use image data that aims to support a larger research agenda dealing with issues regarding criminal activity in consumer-to-consumer online platforms. Preliminary results suggest that the proposed architecture can learn rich spaces using ours and other image datasets resolving important challenges that are identified.

1. Introduction

Online consumer-to-consumer (C2C) transactions surged throughout the COVID-19 pandemic as individuals looked for opportunities to generate income by selling goods and services or sought to save money by purchasing secondhand items (Goddevrind et al., 2021). Unfortunately, some online C2C marketplaces are vulnerable to criminal exploitation in which stolen or trafficked goods and services are sold. To sell stolen or trafficked goods and services in online C2C marketplaces, criminals must victimize individuals by stealing physical items or forcing victims to engage in services on behalf of the trafficker. When consumers discover they have purchased illicit goods or services in online C2C marketplaces, consumers’ trust in these markets diminishes. Stakeholders and law enforcement agencies are combating

^{*}Equal contribution ¹Department of Computer Science, Baylor University, Texas, USA ²School of Criminology and Criminal Justice, California State University, San Bernardino, California, USA ³Information Technology and Decisions Sciences Department, University of North Texas, USA ⁴School of Business, Wake Forest University, North Carolina, USA. Correspondence to: Pablo Rivas <Pablo_Rivas@Baylor.edu>.

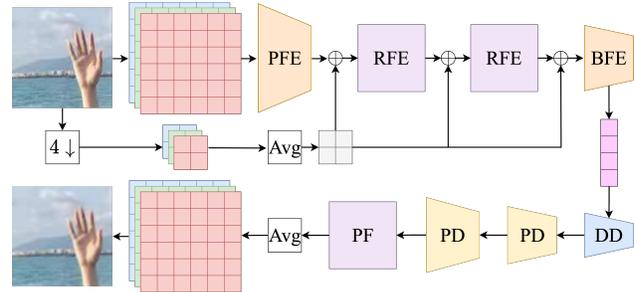


Figure 1. Proposed residual autoencoder that uses convolutional LSTMs for perceptual information extraction and compression.

the explosive growth in illicit e-commerce by conducting large-scale investigations (Ballhaus & Ramachandran, 2021; CBS, 2020). Scientists are re-calibrating explanations of criminal behavior and modifying crime prevention strategies in the face of dynamic post-COVID crime opportunity structures (Stickle & Felson, 2020).

Stakeholders and scholars are searching for commonalities in text, and image-based expressions that may signal a potential victim of sex trafficking (Dubrawski et al., 2015; Ibanez & Suthers, 2014) to develop efficient deep learning-based models, such as the one we introduce in Figure 1, to aid detection efforts for identifying online advertisements that likely involve sex trafficking victims (Kejriwal et al., 2017; Spotlight, 2021; Tong et al., 2017). Such efforts would have an even greater impact on crime reduction if machine learning algorithms could identify universal trafficking properties.

Currently, there are insufficient resources to monitor all online platforms and advertisements for suspicious activity. This leads our team to inquire how machine learning-based NLP and computer vision can be used to find markers and patterns of suspected online illicit activity and whether such models generalize to different commodities and marketplaces. While this is a large multidisciplinary project that will span two years, this paper focuses on one of the initial machine learning challenges, i.e., the encoding of C2C images that will enable independent image analysis and multimodal architecture design that integrates text as part of downstream trafficking detection tasks. Specifically, the

challenges dealing with images from C2C marketplaces related to trafficking are many, including: i) that investigators should minimize exposure to pornographic material; ii) the amount of image data posted daily on these platforms is impossible to analyze manually by humans; iii) that although the data is publicly available, it may contain biometric identifiers (e.g., face) and other personal identity markers readily available in its native image format, but should not. Therefore, our approach to these challenges is reflected in our contributions, which are as follows:

1. we design and develop an autoencoder model that will learn rich features that will produce a latent representation of an image;
2. our model uses modern machine learning techniques to learn low-dimensional representations at scale with few parameters in comparison to other similar models;
3. the proposed model hides personal identifiers from plain sight while keeping necessary information for downstream trafficking detection tasks.

While other important methodologies such as Vision Transformers (Dosovitskiy et al., 2020), BYOL (Grill et al., 2020), or CLIP (Radford et al., 2021) represent great advantages in the field, they also bring some undesirable effects into our research, which we discuss next in Section 2 along with a brief background on autoencoders. Section 3 introduces the proposed architecture; Section 4 discusses our results, and we conclude in Section 5.

2. Background

In machine learning, researchers are particularly interested in exploring different learning paradigms, including supervised and unsupervised, as well as other recent variants that include self-supervised, contrastive learning, etc. For our particular problem, some of these paradigms are not immediately applicable; our problem statement is to find an architecture and learning paradigm for image representations with a lightweight, scalable model that hides from plain sight personal identifiers, yet it needs to embed sufficient information about image perception for downstream multimodal human trafficking tasks.

Dosovitskiy et al. (2020) recently introduced the *Vision Transformer* (ViT), which is an attention-based encoder mechanism that shows superior performance in downstream classification tasks. ViT demonstrated state-of-the-art performance in traditional image datasets for classification. Although the ViT is scalable and generalizes well, its major disadvantages are that it can be costly to train, it is data-hungry, and the baseline ViT, as is typical of most transformer architectures, has a large number of parameters; it can have between 86M and 632M of parameters.

Grill et al. (2020) introduced the *Bootstrap Your Own Latent* (BYOL) approach that uses a self-supervised approach to learn image representations based on sampling the distribution of the input image \mathbf{x} on two networks, one that yields predictions and one that yields projections. The model then is optimized to minimize the loss:

$$\mathcal{L}(\theta, \xi; \mathbf{z}_\theta, \mathbf{z}'_\xi) = 2 - 2 \cdot \frac{\langle q_\theta(\mathbf{z}_\theta), \mathbf{z}'_\xi \rangle}{\|q_\theta(\mathbf{z}_\theta)\|_2 \cdot \|\mathbf{z}'_\xi\|_2}$$

model model hyperparameters
projections of latent representations

where $q_\theta(\cdot)$ is a linear predictor. This optimization ensures that the learned representations on the target network are not trivially predicted and promotes robustness in learning the distribution of the input at the level of the representation space. However, for our context, there are also disadvantages, i.e., the methodology relies on data for classification tasks or other label-based data, e.g., ImageNet, to guarantee sufficient data for bootstrapping and distribution approximation. Further, the baseline model, i.e., based on ResNets, is significantly large, from 250M to 375M of parameters.

More recently, Radford et al. (2021) made significant progress in contrastive learning by introducing a *Contrastive Language–Image Pre-training* (CLIP) methodology. CLIP is a multimodal approach that considers language and image information to align the representation space of text and images that belong together, allowing for the navigation of the latent space either with text or images with the right prompts. Due to the nature of the language portion of CLIP, it can also be extremely costly to fully train CLIP on C2C text and image data. Therefore, in our context CLIP is more suitable for transfer learning from a pre-trained model on data similar to C2C marketplaces, e.g., social media. However, at this stage of our project, we are investigating only image representations that will obfuscate personally identifiable information, and minimize the investigator’s exposure to explicit trafficking content, all the while using a lightweight model that still produces good latent representations. With this in mind, we propose to use an autoencoder architecture that learns convolutional filters, similar to ViT’s patches, and that uses a similarity-based loss and two slightly different parallel networks as in BYOL.

A convolutional autoencoder (CAE) is a type of architecture that falls within the CNN family, and it is traditionally trained using an unsupervised learning approach (Kumar, 2020). The most common use for CAEs is to learn low-dimensional image representations by minimizing an image reconstruction error criteria through learning optimal convolutional filters at different layers (Fukushima & Miyake, 1982). CNNs offer additional intriguing characteristics and architectures that have resulted in new applications and variants (Aloysius & Geetha, 2017; Sultana et al., 2020; Zhang

et al., 2019). Although many AEs, such as the variational autoencoder (VAE) and the denoising AE, have been successfully implemented (Castellini et al., 2017; Jhamb et al., 2018; Li & She, 2017; Wetzel, 2017; Yan et al., 2020), there is still research being done in alternative ways to make them more efficient during learning, or making them optimize and generalize better.

The proposed architecture can be similar to the work by Zhou et al. (2020) which uses residual connections in an image autoencoder; and the work of Zhao et al. (2021) that takes advantage of temporal data. However, our work is unique in using several elements together to produce rich representations that solve the problem defined earlier.

3. Experimental Architecture

As established earlier, our BEAR autoencoder is composed of different light-weight feature extraction mechanisms that take advantage of spatial relationships and residual information across layers. A square input image $\mathbf{x} \in \mathbb{R}^{n \times n \times d}$ with n rows, n columns, and an arbitrary depth d , in our context is a classic three-channel color image. The input image is also downsampled by a factor of 4 in all dimensions except depth, $\tilde{\mathbf{x}} \in \mathbb{R}^{n/4 \times n/4 \times d}$. While \mathbf{x} feeds the main feature extraction pipeline, $\tilde{\mathbf{x}}$ feeds the residual connections.

3.1. Encoder

From Figure 1 we can observe that the first piece of the BEAR model goes through a *Perceptual Feature Encoder*, $\mathbf{z}_{\text{PFE}} = \text{PFE}(\mathbf{x}; \theta)$, which is composed of two blocks of regularized convolutional long-short term memory models (LSTMs), first introduced by Shi et al. (2015), which yield features \mathbf{z}_{PFE} , see Figure 2. Then the data goes through a *Residual Feature Entanglement* piece, $\mathbf{z}_{\text{RFE}} = \text{RFE}(\mathbf{z}_{\text{PFE}}, \tilde{\mathbf{x}}; \theta)$, which uses perceptual features and the residual from the original downsampled image; this process is done twice back to back, see Figure 3. Note that the RFE segment ought to be removed if anyone fine-tunes the model for downstream object recognition tasks, as feature entanglement is known to have detrimental effects in fitting labels (Zheng et al., 2021); however, feature entanglement is naturally beneficial for unsupervised image representation learning such as in our context; both of these go into the encoder, as we discuss next.

Next the model goes into a *Bottleneck Feature Encoder*, $\mathbf{z}_{\text{BFE}} = \text{BFE}(\mathbf{z}_{\text{RFE}}, \tilde{\mathbf{x}}; \theta)$, which is comprised of a convolutional LSTM and a dense layer, see Figure 4. Both PFE and BFE reduce dimensions as part of the process, and BFE is the last piece of the encoding process. Thus, $\mathbf{z}_{\text{BFE}} \in \mathbb{R}^m$ is the learned representation of the input in a lower m -dimensional space, i.e., \mathbb{R}^{256} .

3.2. Decoder

The decoder in the architecture is relatively different in the sense that it has no residual connections and starts with a *Dense Decoder*, $\mathbf{z}_{\text{DD}} = \text{DD}(\mathbf{z}_{\text{BFE}}; \theta)$, which begins with a dense layer that produces features to facilitate reconstruction, see Figure 5. Next, we reconstruct increasing dimensions through back-to-back *Perceptual Decoders*, $\mathbf{z}_{\text{PD}} = \text{PD}(\mathbf{z}_{\text{DD}}; \theta)$, and $\mathbf{z}_{\text{PD}} = \text{PD}(\mathbf{z}_{\text{PD}}; \theta)$, both of which are based on parallel convolutional networks, see Figure 6. Finally, we have trainable *Perceptual Features*, $\mathbf{z}_{\text{PF}} = \text{PF}(\mathbf{z}_{\text{PD}}; \theta)$, consisting on parallel convolutional networks for larger feature re-composition, see Figure 7. The reconstructed image is obtained by averaging the channels of the PF, $\hat{\mathbf{x}} = \text{Avg}(\mathbf{z}_{\text{PF}})$.

3.3. Training

The BEAR model is trained using a classic binary cross entropy loss given by:

$$\mathcal{L}(\theta; \mathbf{x}) = -\frac{1}{d} \sum_{i=1}^d \mathbf{x}_i \cdot \log \hat{\mathbf{x}}_i + (1 - \mathbf{x}_i) \cdot \log (1 - \hat{\mathbf{x}}_i) \quad (1)$$

↑ **autoencoder model hyperparameters**
↑ **input image in $\mathbb{R}^{n \times n \times d}$**

where $\mathbf{x} \in \mathbb{R}^{n \times n \times d}$ is the input, $\hat{\mathbf{x}}$ is the reconstruction, and θ are the model parameters. This loss can also be replaced with an MSE loss achieving similar results. Minimizing the loss in (1) repeatedly, using random starts, yields consistent results across the board using the Adam optimizer with an initial learning rate of 0.0001 and an automatic decay after a five epoch plateau. The learning stops after no improvements in the validation loss after 10 epochs.

The dataset consists of images of C2C sites publicly available. However, due to privacy concerns, our data is not publicly available to protect the potentially identifiable private information of humans. The total number of images is 1,998,680; they all vary in size originally but are resized to $128 \times 128 \times 3$ during training. The results obtained are discussed next.

4. Results

We trained the proposed architecture, as discussed in the previous section, using different three data sets as indicated in Table 1. The first dataset is C2C image data publicly available on the web; however, we do not possess the rights to distribute. This dataset contains images of people, objects, text, numbers, social media accounts, etc. These may contain images of trafficking victims although the images are not labeled, since this is a future task for a different stage of our project. Our intention on using the C2C data at this

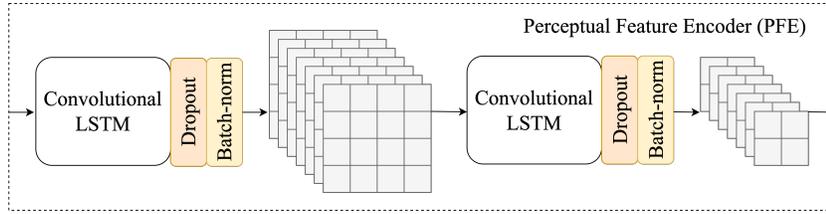


Figure 2. The *Perceptual Feature Encoder* architecture that reduces dimensions while extracting perceptual information.

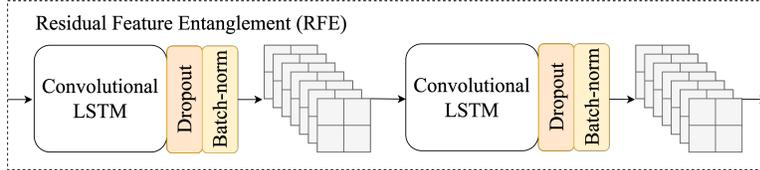


Figure 3. The *Residual Feature Entanglement* that uses as input both the previous layer information and the residual, preserving dimensions.

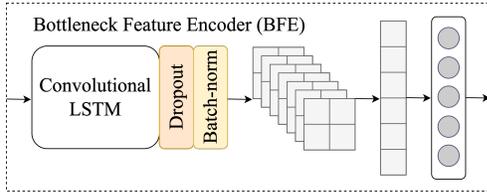


Figure 4. The BFE uses a convolutional LSTM and a dense layer.

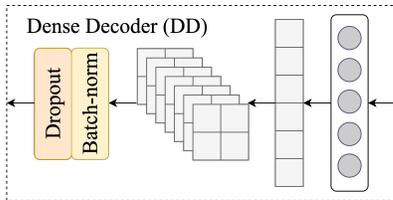


Figure 5. The DE reconstructs feature maps using a dense layer.

early stage of our entire research project is to test the ability of our model to achieve fast converge and fast throughput, while producing image representations that obfuscate personal identifiers and information that is readily available visually.

The experiment in Table 1 is to illustrate the convergence time required to fully train the BEAR architecture on C2C data and two other popular datasets. The table gives an idea of the training time on a single GPU NVIDIA 1080 card with 8GB. This suggests that the model behaves consistently with respect to the size of the dataset.

Next, we are interested in analyzing the learned space and the representations, particularly in the context of C2C data. Figure 8 depicts the UMAP 2D representation of the learned representations using k -means to identify potential clusters (Becht et al., 2019). Using the elbow method, we identi-

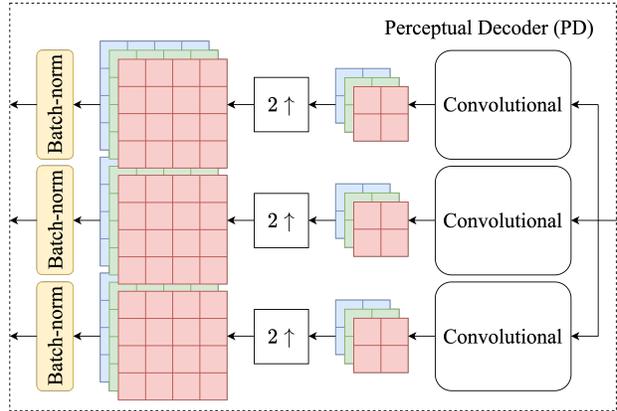


Figure 6. The PD uses parallel convolutions before up-sampling.

Table 1. Data set comparison with respect to convergence time.

DATA SET	N	SIZE	LABELLED	TIME (H)
C2C	1.9M	VARIES	×	132
CIFAR-10	60K	32×32	✓	20
IMAGENET	1.2M	VARIES	✓	99

fied k between 10 and 20 to be a good number of clusters. The figure identifies 20 clusters with different colors and displays the centroids. The figure suggests that the latent space is formed of several distinct groups that are distributed across the space and clustered according to their perceptual information. Further examination in C2C data can demonstrate that cluster prototypes are distinct among each other, with respect to their visual characteristics beyond the trivial color information.

To show further that the latent space distributes data displaying different information, we can simply look at the norm

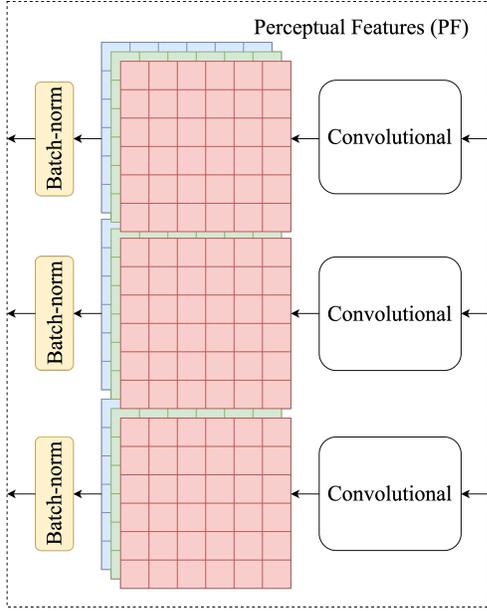


Figure 7. The PF reconstructs perceptual features with parallel convolutions.

of the representation vector $\|\mathbf{z}_{\text{BFE}}\|_2$, as shown in Figure 9. From the figure, we can appreciate that the data points are often associated and grouped along with other vectors with similar perceptual information; however, the space is not limited or distributed solely by norm similarity. The feature space has perceptual image entanglement, which is a desired property prior to our future work, which includes re-training the model on labeled C2C data for detecting criminal activity in a multimodal setting using contrastive learning.

Finally, we conducted another experiment consisting of running k -means not in a reduced 2D space but in a slightly higher-dimensional space, i.e., 20 dimensions, in comparison to the 10-dimensional space of Figure 8, where we would obtain a result with slightly fuzzier cluster limits. These effects are usually expected due to the curse of dimensionality that affects k -means. The experiments were conducted in C2C, CIFAR-10, and ImageNet data. From these experiments, we can consistently observe that the data is distributed by perceptual information and not just by color codes; further, k -means converges consistently to satisfy the cluster quality objective: $\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{z}_{\text{BFE}} \in S_i} \|\mathbf{z}_{\text{BFE}} - \boldsymbol{\mu}_i\|^2$, for partition sets $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$. This empirically shows that the model will be consistent across different image datasets.

In summary, the proposed model demonstrates the following. First, BEAR is able to converge to a local minima that satisfies a reconstruction loss, suggesting that the perceptual features extracted using light-weight network segments com-

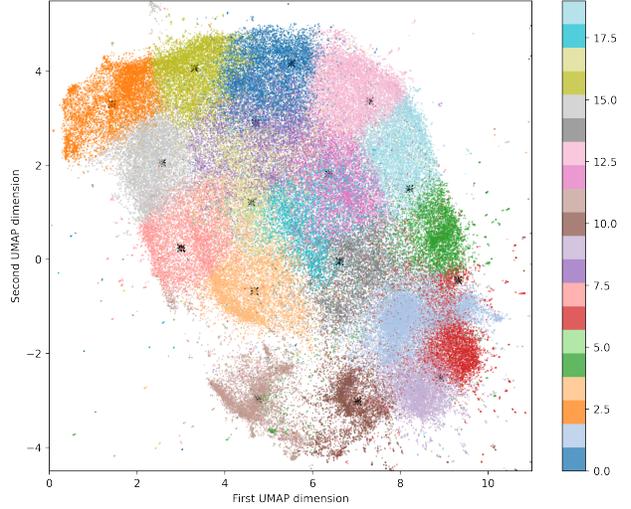


Figure 8. Learned representations on C2C data visualized using a 2D UMAP and k -means with $k = 20$.

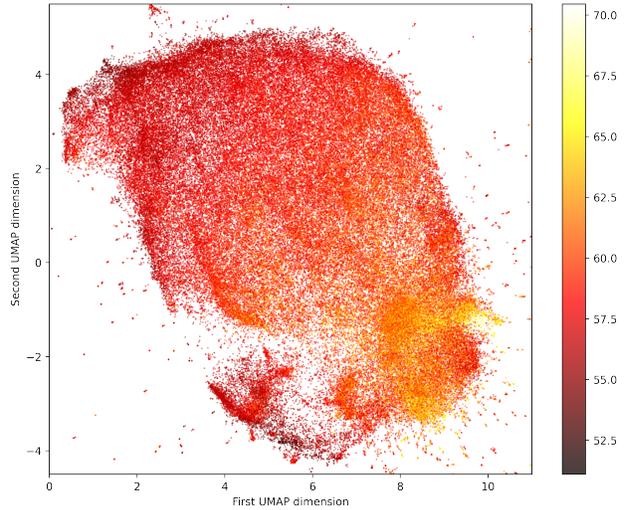


Figure 9. Learned representations on C2C data visualized using $\|\mathbf{z}_{\text{BFE}}\|_2$ as color reference.

presses sufficient information for other downstream tasks. Second, the model consistently performs in a stable manner during unsupervised learning for other popular image datasets. Third, BEAR can consistently satisfy a k -means cluster quality objective, suggesting that the latent space can induce regular groups naturally. Fourth, the model can hide from plain sight private personal information due to the natural abilities of autoencoders to compress data using unique transformations, as opposed to fixed compression methodologies based, for example, in basis functions from the discrete cosine transform (DCT), that can be used to recover DCT compressed data.

5. Conclusions

This paper presents BEAR, an autoencoder-based architecture that treats the channel information in images as sequential data in convolutional LSTMs and using residual connections to preserve features across different encoding pathways. The encoding-decoding architecture called BEAR was trained on C2C data aiming to model content that is typical of these platforms with the ultimate goal of learning robust image representations with a light-weight model.

While other state-of-the-art image feature extraction methods such as VGG and Inception-based are powerful, they are usually biased toward specific classes through label-based supervised learning or semi-supervised for transfer learning; the inductive bias of these architectures favors a limited number of classification tasks. The proposed BEAR methodology is not particularly biased toward specific labels, but remains optimal in preserving perceptual features.

In the future, we aim to use this architecture in connection with textual information and build a larger multimodal data pipeline that will aid in detecting online crime on C2C platforms using contrastive learning.

Acknowledgements

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

References

- Aloysius, N. and Geetha, M. A review on deep convolutional neural networks. *2017 International Conference on Communication and Signal Processing (ICCSP)*, 2017. doi: 10.1109/ICCSP.2017.8286426.
- Ballhaus, R. and Ramachandran, S. Ben dugan works for cvs., Business. *The Wall Street Journal*. September 2021. URL <https://www.wsj.com/articles/cvs-home-depot-theft-organized-crime>.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.
- Castellini, J., Poggioni, V., and Sorbi, G. Fake twitter followers detection by denoising autoencoder. In *Proceedings of the International Conference on Web Intelligence, WI '17*, pp. 195–202, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349512. doi: 10.1145/3106426.3106489. URL <https://doi.org/10.1145/3106426.3106489>.
- CBS, S. Major san francisco bay area retail theft ring busted, crime section. san francisco, ca., October 2020. URL <https://sanfrancisco.cbslocal.com/tag/stolen-property/>.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.
- Dubrawski, A., Miller, K., Barnes, M., Boecking, B., and Kennedy, E. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1(1):65–85, 2015.
- Fukushima, K. and Miyake, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.
- Goddevrind, V., Schumacher, T., Seetharaman, R., and Spillecke, D. C2c e-commerce: Could a new business model sell more old goods?, McKinsey & Company. Sep. 2021. URL <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/c2c-ecommerce-could-a->.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.
- Ibanez, M. and Suthers, D. D. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. In *2014 47th Hawaii international conference on system sciences*, pp. 1556–1565. IEEE, 2014.
- Jhamb, Y., Ebesu, T., and Fang, Y. Attentive contextual denoising autoencoder for recommendation. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '18*, pp. 27–34, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356565. doi: 10.1145/3234944.3234956. URL <https://doi.org/10.1145/3234944.3234956>.
- Kejrival, M., Ding, J., Shao, R., Kumar, A., and Szekely, P. Flagit: A system for minimally supervised human trafficking indicator mining. *arXiv preprint arXiv:1712.03086*, 2017.

- Kumar, D. V. How to Implement Convolutional Autoencoder in PyTorch with CUDA, July 2020.
- Li, X. and She, J. Relational variational autoencoder for link prediction with multimedia data. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Thematic Workshops '17, pp. 93–100, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450354165. doi: 10.1145/3126686.3126774. URL <https://doi.org/10.1145/3126686.3126774>.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
- Spotlight. Thorn, 2021. URL <https://www.thorn.org/spotlight/>.
- Stickle, B. and Felson, M. Crime rates in a pandemic: The largest criminological experiment in history. *American Journal of Criminal Justice*, 45(4):525–536, 2020.
- Sultana, F., Sufian, A., and Dutta, P. A Review of Object Detection Models based on Convolutional Neural Network. *arXiv:1905.01614 [cs]*, 1157:1–16, 2020. doi: 10.1007/978-981-15-4288-6_1. URL <http://arxiv.org/abs/1905.01614>. arXiv: 1905.01614.
- Tong, E., Zadeh, A., Jones, C., and Morency, L.-P. Combating human trafficking with deep multimodal models. *arXiv preprint arXiv:1705.02735*, 2017.
- Wetzel, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Physical Review E*, 96(2): 022140, August 2017. doi: 10.1103/PhysRevE.96.022140. URL <https://link.aps.org/doi/10.1103/PhysRevE.96.022140>.
- Yan, C., Wang, S., Yang, J., Xu, T., and Huang, J. Re-balancing variational autoencoder loss for molecule sequence generation. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB '20*, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379649. doi: 10.1145/3388440.3412458. URL <https://doi.org/10.1145/3388440.3412458>.
- Zhang, S., Tong, H., Xu, J., and Maciejewski, R. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):11, November 2019. ISSN 2197-4314. doi: 10.1186/s40649-019-0069-y. URL <https://doi.org/10.1186/s40649-019-0069-y>.
- Zhao, J., Hu, L., Dong, Y., Huang, L., Weng, S., and Zhang, D. A combination method of stacked autoencoder and 3d deep residual network for hyperspectral image classification. *International Journal of Applied Earth Observation and Geoinformation*, 102:102459, 2021.
- Zheng, Z., Sadhu, A., and Nevatia, R. Improving object detection and attribute recognition by feature entanglement reduction. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2214–2218. IEEE, 2021.
- Zhou, W., Wu, J., Lei, J., Hwang, J.-N., and Yu, L. Salient object detection in stereoscopic 3d images using a deep convolutional residual autoencoder. *IEEE Transactions on Multimedia*, 23:3388–3399, 2020.