Learning under Evolving, Unstable, and Modality Diverse Input Feature Spaces

Heng Lian
Ph.D. Candidate @ Data Science
William & Mary
Williamsburg, VA, USA
hlian01@wm.edu

Yi He
Advisor
William & Mary
Williamsburg, VA, USA
yihe@wm.edu

Abstract

Many existing learning methods rely on a static input feature space and require substantial training to remain effective when input conditions change. However, real-world data often violate these assumptions, presenting significant challenges for learning systems in dynamic environments. Our Study focuses on three ways in which real-world inputs often depart from standard modeling assumptions: (1) evolving feature spaces in data streams, (2) unstable input quality caused by data collection, and (3) structurally diverse multi-modal inputs. We have developed strategies that project inputs with varying feature dimensions into a shared latent space with fixed dimensionality to handle evolving feature spaces. Our goal is to develop learning systems that can adapt to complex and evolving input conditions, while minimizing computational cost relative to more direct yet costly solutions such as retraining.

Keywords

Online Learning, Open Feature Spaces, Data Streams, Multimodality

ACM Reference Format:

Heng Lian and Yi He. 2025. Learning under Evolving, Unstable, and Modality Diverse Input Feature Spaces. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Most current methods assume that the input feature space is static and fully observed during training and deployment. However, the feature space evolves over time in many real-world scenarios [7]. For example, sensor failures and hardware upgrades in soft-sensing manufacturing systems lead to the removal of old features and the introduction of new ones [9]. This dynamic nature makes it difficult for models to accommodate the changing feature space without frequent retraining, which incurs high computation and time costs.

Existing approaches [8, 11, 13] reuse the learned knowledge from the old space to assist in building models for the new space by building explicit connections between the two spaces. However, these approaches remain limited to constructing mappings between specific pairs of fixed-dimensional feature spaces, rather than enabling learning within a dynamic input environment. To address this limitation, we propose projecting all incoming instances, regardless of their original feature dimensionality, into a shared representation space with aligned dimensions [12]. Each feature that has ever appeared is assigned a learnable embedding vector, and the representation of an instance is computed by aggregating the embeddings of its active features. In addition to changes in feature dimensionality, real-world inputs often introduce two more problems. First, the quality of the input data can vary significantly due to differences in collection time, environmental conditions, or equipment. Second, inputs are often diverse and come from different sources such as images, natural language descriptions, time-series signals, or categorical metadata. These sources describe the same entity but differ in format, level of detail, and semantic emphasis, which makes it difficult to jointly utilize them to support representation learning.

These observations suggest that conventional assumptions of fixed-dimensional, fully aligned, and semantically uniform inputs are increasingly inadequate for many real-world learning tasks. Instead, there is a growing need for learning systems that can operate under structural variability, distributional instability, and representational heterogeneity. This work takes a step in that direction by developing a unified modeling approach that adapts to evolving feature spaces, tolerates input inconsistencies, and integrates semantically diverse signals without strong supervision. The goal is to support robust and generalizable representation learning under open and dynamic input conditions.

2 Method and Challenge

2.1 Learning under Evolving Feature Spaces

We first consider this problem under an online learning setting where the evolving of the feature space only happens between two spaces S_1 and S_2, and during a short overlap period, both features from both spaces are available for each instance [11]. Specifically, we assume the feature space transitions from S_1 to S_2 over time, with a short overlapping period T_b during which both $x_t^{S_1} \in \mathbb{R}^{d_1}$ and $x_t^{S_2} \in \mathbb{R}^{d_2}$ are available. To enable the transfer of information between S_1 and S_2, we employ two variational autoencoders (VAE) [10], one for each feature space. Each VAE encodes its input into a latent representation $z_t \in \mathbb{R}^z$, and reconstructs the input through a decoder. The variational loss for each stream is defined:

$$\mathcal{L}_{\text{VI}} = -\mathbb{E}_{Q(z_t|x_t)}[\log P(x_t \mid z_t)] + \text{KL}(Q(z_t \mid x_t) \,\|\, P(z_t)), \quad (1)$$

where $P(z_t) = \mathcal{N}(0, I)$ is the prior distribution and $Q(z_t \mid x_t)$ is the encoder's posterior over latent variables.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

To bridge the two feature spaces, we further introduce a reconstruction loss during the overlapping period T_b. Specifically, the decoder for S_1 is trained to reconstruct $x_t^{S_1}$ from the latent code $z_t^{S_2}$ generated by the encoder of S_2, yielding:

$$\mathcal{L}_{\text{REC}} = \ell\left(x_t^{S_1}, \text{Dec}_{2\to 1}(z_t^{S_2})\right) + \text{KL}\left(Q(z_t^{S_1} \mid x_t^{S_1}) \,\|\, Q(z_t^{S_2} \mid x_t^{S_2})\right). \tag{2}$$

This objective encourages the distributions learned from S_1 and S_2 to become similar, enabling samples from both feature spaces to be mapped into a shared latent subspace. After the S_1 features become unavailable, its corresponding decoder can still reconstruct x^{S_1} from the latent representation encoded by S_2. This allows the model to exploit the knowledge previously captured by S_1, even when only S_2 is observed.

However, this work only considers the change between two fixed-dimension feature spaces and requires an overlapping period. This assumption limits its ability to handle environments where the feature space evolves continuously and unpredictably. We consider a more flexible setting where Each data instance may be observed over a different set of features due to the appearance of new features or the disappearance of existing ones. Specifically, given the data stream consisting of instances ${x_t}_{t=1}^{T}$, where each $x_t \in \mathbb{R}^{d_t}$ is a feature vector with potentially time-varying dimensionality d_t. Let $\mathcal{U} = \mathbb{R}^{d_1} \cup \cdots \cup \mathbb{R}^{d_t}$ denote the instance space observed up to round t. We construct a mapping $\phi : \mathbb{R}^{d_t} \to \mathbb{R}^k$ that produces fixed-dimensional representations in a universal latent space, allowing instances with varying input dimensions to be embedded consistently. Each feature $f_i \in \mathcal{U}$ that has appeared up to time t is associated with a learnable embedding vector $\mathbf{f}_i \in \mathbb{R}^k$. The representation of $\mathbf{x}_t$ is computed by:

$$\mathbf{z}_t := \phi(\mathbf{x}_t) = \bigoplus_{i=1}^{d_t} (f_i \odot \mathbf{f}_i), \tag{3}$$

where $\odot$ denotes element-wise multiplication and $\bigoplus$ denotes element-wise addition. As new features are encountered, their embeddings are initialized and incorporated into the universal space without requiring retraining on previous inputs.

2.2 Challenges of Unstable Input Spaces

In one of our projects, we study the spatial population distribution of benthic species using a collection of underwater images. By identifying the species type and count in each image, we aim to construct a map of biological populations in a specific marine region. In this problem, the change of input dimensions is not a major issue. Although the images may have different pixel sizes due to camera resolution, this can be handled by standard resizing. While resizing may cause some information loss [5, 16], it is still practical and acceptable compared to training separate models for each resolution or using our proposed methods, which also require additional training costs. The main difficulty lies in the variability of the input. All images are collected from the same region, but they differ significantly due to variations in imaging devices, collection time, and environmental conditions. Some images are degraded due to poor water visibility or improper exposure, including underexposure and overexposure, which obscure visual details. In addition, many benthic organisms are difficult to distinguish because they may be partially buried in sediment, have colors similar to the surrounding seafloor, appear too small relative to the image resolution, or are only partially visible. These challenges differ significantly from those typically encountered in standard training and testing datasets, and introduce substantial inconsistency in the input space, leading to large fluctuations in model performance across different images. Existing methods like data augmentation try to address input variability by applying predefined transformations (*e.g.,* adding noise and distortion) to improve model robustness against variations in image quality and appearance. However, the variations in our setting stem from real-world conditions that are far more diverse and irregular than what can be simulated by predefined transformations, limiting the effectiveness of standard augmentation techniques. As a result, these methods often fail to provide considerable improvements when applied across images with diverse quality and content variations.

2.3 Challenge of Input Diversity

Our other project focuses on urban functional zoning which aims to classify different regions of a city into categories such as residential, commercial, or industrial. Traditional approaches typically rely on training supervised models using satellite imagery paired with manually labeled functional categories [6, 17]. However, collecting high-quality annotations at city scale is labor-intensive and costly, which limits the scalability of these methods. To address this issue, we explore the use of multiple forms of input that naturally co-describe the same geographic region. For example, a satellite image accompanied by a textual description can provide a clearer understanding of a region's function than the image alone. This motivates us to introduce input diversity into the learning process, leveraging complementary modalities to reduce supervision requirements. We proposed a method that first uses a large multi-modal model to generate textual descriptions for each satellite image, and then encodes both the image and the generated text using a shared vision-language model (e.g., CLIP [14]). A contrastive clustering objective [1] is applied to encourage alignment between the two embeddings of the same region, allowing the model to group similar regions without requiring manual labels.

Despite manually verifying that the generated texts provide detailed and accurate descriptions of the corresponding images, we observed that the embeddings produced by the same vision-language model still differ significantly in distribution. Even with contrastive learning, the alignment between the two modalities remains weak, making it difficult to fully exploit the inherently complementary information contained in each modality. Unlike scenarios where data are collected using different but related sensors (*e.g.,* multiple camera models with similar characteristics), the gap between input spaces becomes substantially larger when incorporating heterogeneous modalities such as images and text. These modalities differ not only in dimensionality but also in structure and semantics, making it a significant challenge to align them effectively for joint representation learning. Although introducing more diverse inputs can be viewed as a special case of learning under evolving feature spaces, the substantial difference between modalities makes effective alignment much more challenging than in cases involving homogeneous feature shifts.

3 Future Work and Application Study

There have been plenty of works [2–4, 12, 15] which aim at solving the above problems, such as the evolving feature space. However, most existing approaches address each problem independently. In practical learning settings, these challenges frequently co-occur and exhibit mutual influence. This phenomenon can be well illustrated by how humans learn new concepts. This phenomenon can be intuitively illustrated by how humans learn to recognize a new species. When encountering an unfamiliar animal, a person often sees its appearance, hears its vocalizations, and learns to describe it using language, all at the same time. These sources of information are processed together, forming a rich and robust concept. After even a single learning experience, a person can still recognize the same species in low-quality images with blurred or incomplete features. Humans can also use brief textual descriptions to distinguish between visually similar species. For example, a short note such as "a small dog with short legs" can help distinguish a Dachshund from a Golden Retriever, even if the person has never seen them before. In addition, humans are able to tell the difference between dogs and wolves using auditory cues such as their vocalizations.

Motivated by the observation, our future work aims to design a learning framework that explicitly leverages their interactions during training. Instead of treating each challenge as an isolated problem, our framework integrates additional inputs over time to support the interpretation of difficult input data. One typical scenario we consider involves an image that is difficult to interpret on its own, such as a high-resolution satellite photo with complex urban structures or an underwater image affected by low visibility. To support learning in such cases, we can incrementally introduce textual information as evolving features to help the model refine its understanding. For example, the keyword *warehouse* can help increase the likelihood that a satellite image containing diverse buildings is classified as industrial. The keyword *fan-shape like* can help identify a scallop that is buried under sand and lacks color cues. This direction reflects a broader principle: effective learning should not rely on any single input source, as each alone is insufficient for robust understanding. Instead, input modalities should be introduced progressively throughout the learning process, allowing the input space to be dynamic and evolve in response to task demands. Through this dynamic integration, the input space becomes increasingly enriched, offering a more comprehensive basis for representation learning. As a result, models can better capture the complexity of real-world scenarios and build more robust representations.

4 Conclusion

This work presents a perspective on learning under dynamic and diverse input conditions. We have addressed the challenge of evolving feature spaces by developing methods that project variable-dimensional inputs into a shared latent space. Besides, we are currently investigating how to improve robustness under unstable visual input conditions, and how to leverage complementary diverse modalities to reduce supervision in tasks such as urban functional zoning. We aim to develop a general learning framework that incrementally integrates diverse input modalities, enabling them to support and reinforce each other during training, in order to achieve robust understanding under dynamic input conditions.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*.

[2] Zhong Chen, Yi He, Di Wu, Huixin Zhan, Victor Sheng, and Kun Zhang. 2024. Robust sparse online learning for data streams with streaming features. In *SDM*. 181–189.

[3] Zhong Chen, Yi He, Di Wu, Wenbin Zhang, and Zhiqiang Deng. 2025. $\ell_{1,\infty}$ Mixed Norm Promoted Row Sparsity for Fast Online CUR Decomposition Learning in Varying Feature Spaces. In *SDM*. 124–133.

[4] Zhong Chen, Yi He, Di Wu, Chen Zhao, and Meikang Qiu. 2025. A Novel Sparse Active Online Learning Framework for Streaming Anomaly Detection in Open Feature Spaces. In *IJCAI*.

[5] Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In *QoMEX*. IEEE, 1–6.

[6] Zhou Guo, Jiangtian Wen, and Rui Xu. 2023. A shape and size free-CNN for urban functional zone mapping with high-resolution satellite images and POI data. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–17.

[7] Yi He, Christian Schreckenberger, Heiner Stuckenschmidt, and Xindong Wu. 2023. Towards Utilitarian Online Learning – A Review of Online Algorithms in Open Feature Space. In *IJCAI*.

[8] Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. 2017. Learning with Feature Evolvable Streams. In *NeurIPS*, Vol. 30.

[9] Yuchen Jiang, Shen Yin, Jingwei Dong, and Okyay Kaynak. 2020. A review on soft sensors for monitoring, control, and optimization of industrial processes. *IEEE Sensors Journal* 21, 11 (2020), 12868–12881.

[10] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

[11] Heng Lian, John Scovi Atwood, Bo-Jian Hou, Jian Wu, and Yi He. 2022. Online deep learning from doubly-streaming data. In *ACMMM*. 3185–3194.

[12] Heng Lian, Yu Huang, Xingquan Zhu, and Yi He. 2024. Utilitarian Online Learning from Open-World Soft Sensing. In *ICDM*. IEEE, 241–250.

[13] Heng Lian, Di Wu, Bo-Jian Hou, Jian Wu, and Yi He. 2023. Online Learning From Evolving Feature Spaces With Deep Variational Models. *IEEE Transactions on Knowledge and Data Engineering* 36, 8 (2023), 4144–4162.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PmLR, 8748–8763.

[15] Christian Schreckenberger, Yi He, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. 2023. Online random feature forests for learning in varying feature spaces. In *AAAI*, Vol. 37. 4587–4595.

[16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2022. Understanding deep learning requires rethinking generalization. In *ICLR*.

[17] Wen Zhou, Dongping Ming, Xianwei Lv, Keqi Zhou, Hanqing Bao, and Zhaoli Hong. 2020. SO–CNN based urban functional zone fine division with VHR remote sensing image. *Remote Sensing of Environment* 236 (2020).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009