Politecnico
di Torino
1859

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Artificial Intelligence for Industry 4.0 (37thcycle)

Generative modelling for improved decision making in data-limited applications

By

Alessio Mascolini
******

Supervisor(s):
Prof. Macii Enrico, Supervisor
Prof. Di Cataldo Santa, Co-Supervisor

Dissertation Referees:
Prof. Baraldi Lorenzo, University of Modena and Reggio Emilia
Prof. Bolelli Federico, University of Modena and Reggio Emilia

Politecnico di Torino
2025

Summary

The field of deep learning has experienced incredible progress in recent years. A major driving force behind this advancement is a new breed of models known as generative models, which have the ability to learn from vast quantities of raw, unlabeled data, like images, text, or sensor readings, without any specific instructions on what to look for.

This thesis explores how these powerful generative models can be applied to solve real-world problems, particularly in industrial settings. We'll look at how they can be used in situations where data is scarce, where computers have limited power, and where it's crucial to get results quickly and understand why a model is making a certain prediction.

Motivation and Challenges

The increasing digitization of industrial processes, often referred to as Industry 4.0, has led to an explosion of data generated by sensors, machines, and what are called cyber-physical systems (CPS). CPS are systems that tightly integrate computing, communication, and control technologies with physical processes, such as a modern factory, a power grid, or even a self-driving car. This data presents both opportunities and challenges, as on the one hand it offers the potential to gain deeper insights into operational processes, optimize performance, predict failures, and improve overall efficiency, yet on the other hand, extracting meaningful information from this raw, often noisy, and high-dimensional data requires sophisticated analytical techniques.

Traditional machine learning approaches, particularly supervised learning, have shown promise in various industrial applications. However, their reliance on large, labeled datasets poses a significant hurdle. In many industrial settings, obtaining

accurately labeled data is a costly and time-consuming process, often requiring expert knowledge and manual annotation. Furthermore, the dynamic nature of industrial environments, with constantly evolving processes and potential unforeseen anomalies, makes it difficult to create a static labeled dataset that encompasses all possible scenarios. This is where the power of unsupervised learning, and specifically generative models, becomes apparent.

Generative models, by their very nature, learn the underlying structure of the data without requiring explicit labels. This capability is particularly valuable in industrial contexts where unlabeled data is abundant, but labeled data is scarce. By pretraining on vast amounts of unlabeled data, these models can learn rich representations that capture the normal operating patterns and inherent variability of the system. These learned representations can then be leveraged for various downstream tasks, such as anomaly detection (finding unusual patterns that might indicate a problem), predictive maintenance (predicting when a machine is likely to fail), and process optimization (finding ways to make the process more efficient), with minimal or no additional labeled data.

The application of generative models in industrial settings also presents unique challenges. Industrial data often exhibits complex temporal dependencies. This means that what happens at one point in time is strongly influenced by what happened in the past. The data is also high dimensional, and often contains a lot of noise, caused by random fluctuations or errors in the data that can make it harder to identify the underlying patterns. Real-time performance is also often a critical requirement, particularly in applications like anomaly detection, where timely intervention can prevent costly equipment failures or safety hazards. This necessitates the development of models that are not only accurate but also computationally efficient and capable of operating on resource-constrained edge devices, located close to the source of the data, rather than in a remote data center or in the cloud. In addition to these requirements black box models, even if highly accurate, may not be readily accepted in industrial environments where operators and engineers need to understand the reasoning behind the model's predictions. So if a model flags a machine as being about to fail, the engineers need to know why it's making that prediction so they can take appropriate action.

Research Objectives

This thesis aims to address these challenges by exploring and developing novel generative modeling techniques tailored to the specific requirements of industrial applications. The primary research objectives are:

1. **To investigate how well generative models can learn useful representations from unlabeled industrial data.** This involves exploring different types of generative models, and evaluating their ability to capture the underlying structure and patterns of diverse industrial datasets.
2. **To develop lightweight and efficient generative models that can run in real-time on devices with limited computing power.** This requires designing models that use less memory and can make predictions quickly.
3. **To make generative models used for industrial anomaly detection more understandable.** This involves finding ways to explain why a model is making a particular prediction, so that human operators can trust and use the model effectively.
4. **To demonstrate that these methods work in real-world industrial situations.** This involves testing the models on both publicly available datasets and data from actual industrial settings, showing that they can solve specific problems in manufacturing and other areas.

Key Contributions

Addressing the research objectives requires grounding the work in practical application. Therefore, this PhD investigates multiple diverse industrial case studies, differing in context, data, and technology, which are detailed later. The core contributions of this thesis, emerge directly from the insights and results obtained through these real-world investigations:

- **GAN-DL:** A new method for analyzing biological images that uses a generative adversarial network to learn from images without needing any labels.

- **VARADE and VARADE++:** New methods for detecting unusual patterns (anomalies) in real-time from multivariate time series data acquired by industrial sensors. These models are designed to be both accurate and efficient, so they can run on devices with limited computing power.
- **Robotic Arm Dataset (RoAD):** A new, comprehensive annotated dataset of sensor data collected from a robotic arm, designed to help researchers develop and test anomaly detection algorithms for industrial robots.
- **A Foundational Time Series Embedding:** A foundation model for zero-shot anomaly detection in time series, pretrained on vast, diverse datasets and requiring no in-domain fine-tuning.
- **High-Resolution Class Activation Mapping:** A way to make deep learning models more explainable by showing which parts of an input (like an image) are most important for the model's prediction.

These contributions demonstrate the power of generative pretraining in various scenarios, going from data-limited contexts like biological image analysis, to strictly resource-constrained settings as is the case for edge computing, to contexts where pre-existing data may be almost completely absent.