MANTECH Publications
Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 1, Issue 2, July-December 2016

Self-Supervised & Unsupervised Representation Learning: Techniques, Models and Applications

Sabita Kakkar[1], Karan Dev Mehta[2], Farhan Ali[3], Raghunath Tiwari[4]
Assistant Professor, Associate Professor
Department of Computer Science
Green Valley Institute of Technology
Email ID: *Sabitakakkarr0@yahoo.com, mehta17karand@gmail.com, farhanali62@rediffmail.com*

Abstract

Representation learning has become one of the most important research directions in modern Artificial Intelligence. Instead of relying heavily on large amounts of labeled data, self-supervised and unsupervised learning techniques aim to learn meaningful data representations from unlabeled data. These approaches reduce the dependency on human annotations and allow models to learn hidden structures and patterns present in raw data. Self-supervised learning creates surrogate tasks from the data itself, while unsupervised learning extracts intrinsic patterns without labels. This paper reviews major techniques, architectures, and recent developments in self-supervised and unsupervised representation learning across vision, natural language processing, and speech domains. Methods such as autoencoders, contrastive learning, generative models, masked modeling, and clustering-based learning are discussed in detail. Applications in healthcare, robotics, recommender systems, and autonomous systems are highlighted. Challenges and future directions are also examined.

Keywords: *Self-supervised learning, Unsupervised learning, Representation learning, Contrastive learning, Autoencoders, Masked modeling, Generative models, Deep learning*

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

INTRODUCTION

Deep learning models have achieved significant success in recent years, but most of these successes depend on huge amounts of labeled data. Creating labeled datasets is time consuming, expensive, and sometimes impractical. This limitation motivated researchers to explore learning methods that can extract useful knowledge from unlabeled data. Representation learning is a method where models automatically learn features or representations from data, rather than relying on handcrafted features.

Self-supervised learning (SSL) and unsupervised learning are two key paradigms for representation learning without labels. Unsupervised learning attempts to discover patterns and structure in data, while self-supervised learning designs auxiliary tasks, called pretext tasks, where labels are automatically generated from the data itself.

These techniques have recently revolutionized computer vision and natural language processing. Models such as BERT, GPT, SimCLR, BYOL, and MoCo demonstrate how learning from unlabeled data can produce powerful representations that generalize well to downstream tasks.

2. FUNDAMENTALS OF REPRESENTATION LEARNING

Representation learning is the process of automatically discovering features or encodings from raw data that can effectively capture its underlying structure and semantics. Traditional machine learning often relied on handcrafted features (e.g., SIFT for images, TF-IDF for text), but these features may fail to capture complex patterns in large-scale or high-dimensional data. Representation learning, often powered by deep neural networks, enables the model to learn **embeddings** that summarize the essential information while being compact and useful for downstream tasks such as classification, clustering, or prediction.

Formally, given a dataset X={x1,x2,…,xn}X = \{x_1, x_2, \dots, x_n\}X={x1,x2,…,xn}, representation learning aims to learn a mapping function f:X→Zf: X \rightarrow Zf:X→Z, where ZZZ is a latent space (embedding space) capturing the important properties of the data.

2.1 What is a Representation?

A **representation** is essentially a structured vector or tensor that encodes meaningful

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

information about the input data. Instead of raw input (like pixels in an image or words in a sentence), representations aim to summarize patterns, relationships, and semantic information in a compact form.

Key Characteristics of a Good Representation

1. **Informative:**
   - Contains sufficient information to distinguish between different classes or data points.
   - Example: For handwritten digits, a good representation should distinguish "3" from "8" effectively.
2. **Compact:**
   - Reduces dimensionality while retaining essential information.
   - Example: Instead of a 28×28 pixel image (784 dimensions), a 64-dimensional vector could serve as an effective embedding.
3. **Transferable:**
   - Can be applied to multiple tasks beyond the original training objective.
   - Example: Image embeddings trained via self-supervised learning can be reused for classification, object detection, or segmentation.
4. **Robust to Noise:**
   - Small perturbations or irrelevant variations in input do not drastically affect the representation.
   - Example: In speech recognition, representations should be robust to background noise or variations in voice pitch.

2.2 Difference Between Supervised, Unsupervised, and Self-Supervised Learning

Representation learning can be achieved through various learning paradigms. Understanding the differences between these approaches is crucial to selecting the appropriate method.

Learning Type	Requires Labels	Learning Strategy	Example
Supervised	Yes	Learn mapping from input to labels	Image classification
Unsupervised	No	Discover patterns	Clustering

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Learning Type	Requires Labels	Learning Strategy	Example
Self-Supervised	No manual labels	Create pretext task	Masked word prediction

3. UNSUPERVISED REPRESENTATION LEARNING TECHNIQUES

Unsupervised representation learning aims to learn meaningful data embeddings without relying on labeled datasets. The goal is to discover patterns, structures, and latent features that capture the essence of the data. These representations can then be used for downstream tasks such as clustering, anomaly detection, or initialization for supervised learning.

Several techniques have been developed in recent years, each with unique strengths and applications. The major categories include **autoencoders**, **generative models**, **clustering-based methods**, and **dimensionality reduction techniques**.

3.1 Autoencoders

Autoencoders are neural networks designed to learn compact representations of input data by reconstructing it. They consist of two main components:

1. **Encoder** fθ(x)f_\theta(x)fθ(x): Transforms the input xxx into a lower-dimensional latent vector zzz.
2. **Decoder** gϕ(z)g_\phi(z)gϕ(z): Reconstructs the input from the latent vector, producing x^=gϕ(fθ(x))\hat{x} = g_\phi(f_\theta(x))x^=gϕ(fθ(x)).

The network is trained to minimize a reconstruction loss, typically the mean squared error (MSE) for continuous data or cross-entropy for binary data:

Lrecon=||x−x^||2\mathcal{L}_{recon} = ||x - \hat{x}||^2Lrecon=||x−x^||2

Autoencoders not only compress data but also capture underlying structures that can serve as useful representations.

Types of Autoencoders

- **Sparse Autoencoders:**

Introduce sparsity constraints in the hidden layer to force the network to learn a compact, informative representation. Useful for capturing significant features while ignoring redundancy.

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- **Denoising Autoencoders:**

The network is trained to reconstruct the original input from a corrupted version. This encourages the model to learn robust features resilient to noise.

- **Variational Autoencoders (VAE):**

Introduce a probabilistic latent space where each input is represented by a distribution rather than a single vector. This allows sampling new data points from the learned distribution. The loss combines reconstruction error with Kullback-Leibler divergence to regularize the latent space.

LVAE=Eqϕ(z|x)[log⁡pθ(x|z)]−KL(qϕ(z|x)||p(z))\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x) || p(z))LVAE=Eqϕ(z|x)[logpθ(x|z)]−KL(qϕ(z|x)||p(z))

Applications: Image compression, anomaly detection, feature extraction for downstream tasks.

3.2 Generative Models

Generative models explicitly model the probability distribution of data and can generate new samples similar to the training data. These models are also useful for learning representations in an unsupervised manner.

- **Generative Adversarial Networks (GANs):**

Introduced by Goodfellow et al., GANs consist of a generator GGG and a discriminator DDD in a min-max game:

min⁡Gmax⁡DEx∼pdata[log⁡D(x)]+Ez∼pz[log⁡(1−D(G(z)))]\min_G \max_D \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log (1 - D(G(z)))]GminDmaxEx∼pdata[logD(x)]+Ez∼pz[log(1−D(G(z)))]

The generator learns to produce realistic samples while the discriminator distinguishes real from fake samples. The latent vector zzz learned by the generator can serve as a high-quality representation.

- **Variational Autoencoders (VAEs):**

Already mentioned in autoencoders, VAEs are both generative and representation learning models.

- **Flow-based Models:**

Learn invertible transformations between simple distributions (like Gaussian) and complex data distributions. These models allow exact likelihood estimation and produce meaningful

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

latent embeddings.

Applications: Data generation, anomaly detection, semi-supervised learning.

3.3 Clustering-Based Learning

Clustering-based representation learning methods aim to group similar data points together in the latent space. Learning embeddings that reflect cluster structure improves downstream task performance.

- **K-means clustering:**

Classic algorithm that partitions data into kkk clusters based on Euclidean distance. Representations are learned by iteratively assigning points to clusters and updating centroids.

- **DeepCluster:**

Uses deep neural networks to learn features while simultaneously clustering representations. The network is trained using pseudo-labels generated from clusters.

- **SwAV (Swapping Assignments between Views):**

A modern contrastive-clustering approach where images are clustered, and the model predicts cluster assignments from different augmented views of the same image. SwAV avoids the need for negative samples, improving scalability.

Benefits: Automatically groups data into meaningful structures; works well for unlabeled data.

3.4 Dimensionality Reduction

Dimensionality reduction techniques aim to transform high-dimensional data into a lower-dimensional latent space while preserving essential structure. These techniques are foundational to unsupervised representation learning.

- **Principal Component Analysis (PCA):**

Linear method that projects data onto the directions of maximum variance. Useful for compressing features and removing redundancy.

- **t-Distributed Stochastic Neighbor Embedding (t-SNE):**

Non-linear technique preserving local neighborhood structure. Widely used for visualization of high-dimensional embeddings.

- **Uniform Manifold Approximation and Projection (UMAP):**

Similar to t-SNE but faster and preserves both local and global structure. Often used for visualizing complex datasets.

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Applications: Visualization, noise reduction, initialization of embeddings for neural networks.

4. SELF-SUPERVISED REPRESENTATION LEARNING (SSL)

Self-supervised learning (SSL) is a paradigm that enables models to learn meaningful representations from **unlabeled data** by creating artificial labels, known as **pseudo-labels**, from the data itself. Unlike traditional supervised learning, SSL does not rely on human-provided labels, making it highly scalable for large datasets.

The key idea is to design a **pretext task** — an auxiliary problem where the input provides the supervision. Solving the pretext task encourages the model to learn features that are transferable to downstream tasks such as classification, detection, or segmentation.

4.1 Pretext Tasks

Pretext tasks are artificially constructed tasks used to supervise the model without external labels. They exploit inherent structures in the data to generate targets automatically.

Common Pretext Tasks in Computer Vision

1. **Image Rotation Prediction:**
   - The model is given an image rotated by 0°, 90°, 180°, or 270° and must predict the rotation angle.
   - This encourages the model to understand the **spatial structure and object orientation**.
   - Reference: Gidaris et al., 2018.
2. **Colorization:**
   - The network is trained to predict the color channels of a grayscale image.
   - By learning to restore color, the model captures **semantic understanding of objects**.
   - Example: A cat is recognized as gray/black patterns in grayscale but colored realistically.
3. **Jigsaw Puzzle Solving:**
   - The image is split into patches and shuffled. The model must predict the correct arrangement.
   - Encourages understanding of **spatial relationships** and global structure.
4. **Masked Token Prediction (NLP & Vision):**
   - Portions of the input (words, image patches) are masked, and the model predicts the missing parts.

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- Example in NLP: **BERT** predicts masked words using context.
- Example in Vision: **MAE (Masked Autoencoders)** predicts missing image patches.

4.2 Contrastive Learning

Contrastive learning is a powerful approach in SSL, where the model **compares pairs of data points** to learn representations. The key principle is:

- **Pull together embeddings of positive pairs** (different views of the same sample).
- **Push apart embeddings of negative pairs** (different samples).

Mechanism

1. Generate two augmented views xi,xjx_i, x_jxi,xj of the same sample.
2. Pass them through an encoder fθf_\thetafθ to obtain embeddings zi,zjz_i, z_jzi,zj.
3. Use a **contrastive loss**, e.g., **InfoNCE**:

Li,j=−logexp(sim(zi,zj)/τ)∑k=1Nexp(sim(zi,zk)/τ)\mathcal{L}_{i,j} = - \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(z_i, z_k)/\tau)}Li,j =−log∑k=1Nexp(sim(zi,zk)/τ)exp(sim(zi,zj)/τ)

Where sim()\text{sim}()sim() is cosine similarity, τ\tauτ is a temperature parameter, and NNN is the batch size.

Popular Contrastive Methods

- **SimCLR:**

Uses strong data augmentations and large batch sizes to generate positive pairs.

- **MoCo (Momentum Contrast):**

Maintains a dynamic dictionary with a momentum-updated encoder for negative pairs.

- **BYOL (Bootstrap Your Own Latent):**

Does not require negative samples; uses two networks (online & target) to predict consistent embeddings.

- **CPC (Contrastive Predictive Coding):**

Predicts future latent representations in sequential data using contrastive loss.

Applications: Image classification, video understanding, speech embeddings.

4.3 Masked Modeling

Masked modeling is a specialized form of self-supervision, where **parts of the input are masked**, and the model learns to **predict the missing content** using context. This forces the

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

model to extract deep structural information.

Examples

1. **BERT (Masked Language Modeling):**
   - Randomly masks words in a sentence.
   - The model predicts the masked tokens using surrounding context.
   - Learns contextualized word embeddings transferable to downstream NLP tasks.
2. **MAE (Masked Autoencoders for Vision):**
   - Randomly masks image patches.
   - The model reconstructs masked patches from visible ones.
   - Encourages learning of global and local image features efficiently.
3. **BEiT (Bidirectional Encoder representation for Image Transformers):**
   - Uses discrete visual tokens as prediction targets.
   - Improves representation quality for vision transformers.

Advantages of Masked Modeling

- Reduces dependence on large datasets with labels.
- Captures **global context** and **local structure**.
- Compatible with both NLP and vision tasks.

4.4 Predictive Learning

Predictive learning trains models to **anticipate future data points or contextual information**. Unlike contrastive learning, which compares pairs, predictive learning focuses on **temporal or sequential dependencies**.

Examples

1. **Sequence Prediction in NLP:**
   - Language models predict the next word in a sentence (autoregressive models).
   - Example: GPT series learns context to generate coherent text.
2. **Video Frame Prediction:**
   - Predict future frames in a video from past frames.
   - Captures temporal structure, motion patterns, and scene dynamics.
3. **Time-Series Forecasting:**
   - Predicting future sensor readings or financial data.

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- Encoders learn features that summarize temporal trends and anomalies.

Advantages: Enables learning of dynamic patterns and long-term dependencies; often used in reinforcement learning and robotics.

5. ARCHITECTURES USED

Architecture	Usage
CNN	Vision tasks
RNN / LSTM	Sequential data
Transformers	NLP, Vision
Graph Neural Networks	Graph data

6. APPLICATIONS

6.1 Computer Vision

Image classification, segmentation, object detection.

6.2 Natural Language Processing

Language modeling, translation, summarization.

6.3 Speech Processing

Speech recognition, speaker identification.

6.4 Healthcare

Medical image analysis without labels.

6.5 Robotics

Learning environment representations for navigation.

ADVANTAGES

- Reduces labeling cost
- Learns robust features
- Transfer learning friendly
- Works with large unlabeled datasets

CHALLENGES

- Designing effective pretext tasks

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

- Computational cost
- Negative sampling issues in contrastive learning
- Evaluation difficulties

RECENT ADVANCES

- Vision Transformers with SSL
- Multimodal SSL (CLIP)
- Large scale pretraining
- Self-distillation methods

FUTURE DIRECTIONS

- Cross-modal representation learning
- Low-resource SSL
- Efficient training methods
- Better theoretical understanding

CONCLUSION

Self-supervised and unsupervised representation learning are reshaping how AI models learn from data. These techniques reduce dependency on labeled datasets and enable models to capture meaningful structures from raw inputs. From computer vision to language and healthcare, these methods have demonstrated their capability to generalize across tasks. Despite challenges, continuous advancements in architectures and learning strategies promise a future where models learn more like humans, from observation and context rather than explicit supervision.

REFERENCES

1. Hinton, G. E., & Salakhutdinov, R. (2006). Reducing dimensionality with neural networks. *Science*.
2. Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *ICLR*.
3. Goodfellow, I. et al. (2014). Generative Adversarial Nets. *NeurIPS*.
4. Chen, T. et al. (2020). A Simple Framework for Contrastive Learning of Visual Representations (SimCLR). *ICML*.
5. He, K. et al. (2020). Momentum Contrast for Unsupervised Visual Representation

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Learning (MoCo). *CVPR*.

6. Grill, J. B. et al. (2020). Bootstrap Your Own Latent (BYOL). *NeurIPS*.
7. Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.
8. Dosovitskiy, A. et al. (2021). An Image is Worth 16x16 Words: Vision Transformer (ViT). *ICLR*.
9. Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). *ICML*.
10. Caron, M. et al. (2021). Emerging Properties in Self-Supervised Vision Transformers (DINO). *ICCV*.

Page 97-108 © MANTECH PUBLICATIONS 2016. All Rights Reserved