Systematic Review of Advancing Machine Learning Through Cross-Domain Analysis of Unlabeled Data

Yue Zhu & Johnathan Crowell

Independent Researcher, USA

DOI: 10.55662/JST.2023.4104

Abstract

Self-supervised learning (SSL) has become a transformative approach in the field of machine learning, offering a powerful means to harness the vast amounts of unlabeled data available across various domains. By creating auxiliary tasks that generate supervisory signals directly from the data, SSL mitigates the dependency on large, labeled datasets, thereby expanding the applicability of machine learning models. This paper provides a comprehensive exploration of SSL techniques applied to diverse data types, including images, text, audio, and time-series data. We delve into the underlying principles that drive SSL, examine common methodologies, and highlight specific algorithms tailored to each data type. Additionally, we address the unique challenges encountered in applying SSL across different domains and propose future research directions that could further enhance the capabilities and effectiveness of SSL. Through this analysis, we underscore SSL's potential to significantly advance the development of robust, generalizable models capable of tackling complex real-world problems.

Keywords: Self-Supervised Learning, Pretext Tasks, Representation Learning, Contrastive Learning, Generative Models, Masked Language Modeling, Transfer Learning, Domain Adaptation, Multi-Modal Learning, Data Efficiency

1. Introduction:

Machine learning has traditionally depended on supervised learning, which requires substantial amounts of labeled data. This dependence presents significant challenges in domains where annotating data is costly or impractical, limiting the scalability and applicability of supervised learning, particularly with rare or emerging data types (van den Oord, Li, & Vinyals, 2018). In response, self-supervised learning (SSL) has emerged as a transformative approach that leverages the intrinsic structure of data to generate supervisory signals, significantly reducing the reliance on manual labeling (Jaiswal et al., 2020). By deriving auxiliary tasks directly from the data, SSL enables models to learn useful representations without labeled data, demonstrating vast potential to enhance model performance across a spectrum of domains (Jin et al., 2020).

SSL has gained prominence for its ability to effectively utilize large volumes of unlabeled data. It creates meaningful pretext tasks, such as predicting the spatial arrangement of image patches or masked words within a sentence, that provide a form of supervision enabling models to learn rich and transferable representations (Chen et al., 2020; He et al., 2020). This capability is particularly valuable in fields where labeled data is scarce but unlabeled data is plentiful (Tung et al., 2017).

In this paper, we explore the principles of SSL, review its applications across different data domains—including image processing, natural language processing, and audio data—and discuss the ongoing challenges and potential for future research. The versatility and effectiveness of SSL underscore the need for continued advancements to fully realize its benefits across various applications.

2. Principles of Self-Supervised Learning:

Self-supervised learning (SSL) fundamentally revolves around leveraging intrinsic data properties to create supervisory signals without relying on external labels. At its core, SSL constructs auxiliary or pretext tasks that serve as proxies for generating meaningful representations from unlabeled data. These tasks compel the model to predict or reconstruct certain aspects of the data using only the information present within the dataset itself. By solving these auxiliary tasks, the model acquires representations that capture underlying structures, patterns, and relationships within the data. These can then be fine-tuned for specific downstream tasks using minimal labeled data.

One of the primary principles of SSL is **contrastive learning**, which aims to learn representations by distinguishing between similar and dissimilar samples. The main idea is to maximize the agreement between representations of augmented versions of the same data point (positive pairs) while minimizing the agreement with different data points (negative pairs). This approach has been successfully applied in various domains, including computer vision and natural language processing, through methods like SimCLR and MoCo for images, and SimCSE for text. Contrastive learning helps in building robust and discriminative features by encouraging the model to identify and utilize salient aspects of the data that distinguish one instance from another, leading to representations that generalize well across different tasks (Tung, Tung, Yumer, & Fragkiadaki, 2017).

Another foundational principle is **generative learning**, which involves reconstructing or generating data from partial or corrupted inputs. This technique encourages the model to capture the full data distribution and understand the underlying generative process. Methods such as autoencoders, where the task is to reconstruct the input from a lower-dimensional latent space, and Generative Adversarial Networks (GANs), which generate realistic data samples from noise, exemplify generative SSL approaches. In natural language processing, models like BERT use masked language modeling (MLM) to predict masked words in a sentence, effectively reconstructing the original text (Radford, Narasimhan, Salimans, & Sutskever, 2018). Generative approaches are particularly powerful for capturing detailed and nuanced data features, making them suitable for tasks where understanding data generation processes is crucial.

Predictive learning forms another key principle in SSL, where the model learns to predict future or missing parts of the data from existing observations. This approach is often used in time-series and sequential data, where predicting the next element or the future state is essential. For example, in audio processing, models like Contrastive Predictive Coding (CPC) predict future audio frames from past ones, leveraging the temporal continuity of the data (Schneider, Baevski, Collobert, & Auli, 2019). In text, next sentence prediction (NSP) used in models like BERT requires predicting whether a given sentence follows another, promoting the learning of contextual dependencies and relationships between sentences. Predictive learning is effective in scenarios where understanding sequential patterns and temporal dynamics is critical, as it enables the model to anticipate and capture evolving data behaviors.

Lastly, **clustering-based methods** represent an emerging SSL principle where the goal is to organize data into meaningful clusters or groups. These methods do not rely on explicit labels but instead use the inherent structure of the data to form clusters that represent similar data points. Techniques like DeepCluster for images and clustering-based pre-training for time-series data encourage the model to discover and exploit data distributions without predefined categories. Clustering-based SSL is advantageous for tasks where categorization and segmentation of data are essential, such as image segmentation and unsupervised clustering in text. This principle leverages the natural tendency of data to form clusters based on similarities, enabling the model to learn representations that reflect the data's intrinsic organization.

Together, these principles—contrastive, generative, predictive, and clustering-based learning—form the backbone of self-supervised learning, allowing models to extract valuable information from unlabeled data. By creating and solving auxiliary tasks that exploit the data's inherent properties, SSL enables the development of rich and versatile representations, paving the way for improved performance across a wide range of applications.

3. Self-Supervised Learning for Different Data Types

SSL for Image Data has become a cornerstone in leveraging unlabeled visual content to develop models that can decode complex patterns and features from images. Central to this area is **contrastive learning**, where the goal is to fine-tune visual representations by maximizing the agreement between differently augmented views of the same image and minimizing the similarity between views of different images. For instance, SimCLR implements this concept using a contrastive loss function, specifically the normalized temperature-scaled cross-entropy loss (NT-Xent), defined as:

$$L(i,j) = -\log\left(\frac{\left(\exp\left(sim\left(z_i, z_j\right)\tau\right)\right)}{\left(\sum_{k=1}^{2N} 1_{[k \neq i]}\exp\left(\frac{sim(z_i,z_k)}{\tau}\right)\right)}\right)$$

where zi,zjz_i, z_jzi,zj are the representations of two augmented views of the same image, sim\text{sim}sim denotes the cosine similarity, τ\tauτ represents a temperature scaling parameter, and NNN is the number of images in the batch. This formula helps in learning features invariant to the type of augmentation applied, thus enhancing the model's generalization capabilities across various visual tasks (He et al., 2020).

Generative methods, another pillar of SSL for images, utilize architectures like autoencoders and Generative Adversarial Networks (GANs) to reconstruct or generate new images. The training objective in autoencoders can be expressed as minimizing the reconstruction loss:

$$L(x) = \left|\left|x - D\big(E(x)\big)\right|\right|^2$$

where xxx is the input image, EEE represents the encoder that compresses the image into a latent space, and DDD denotes the decoder that reconstructs the image from the latent representation. GANs, on the other hand, involve a generator GGG and a discriminator DDD, working in tandem where GGG tries to generate realistic images to fool DDD, and DDD aims to distinguish between real and fake images. The adversarial loss for a GAN is given by:

$$\min_G \max_D V(D,G) = E_{[x\sim p_{data(x)}][\log D(x)]} + E_{[z\sim p_{z(z)}]}\left[\log\left(1 - D\big(G(z)\big)\right)\right]$$

Predictive methods focus on forecasting the missing parts of images or specific features based on the observable data. Techniques such as image inpainting or colorization utilize context encoders where the loss function might involve not just the pixel-wise accuracy but also adherence to semantic consistency, enhancing the contextual understanding of the model.

SSL for Text Data leverages **masked language modeling (MLM)** and **next sentence prediction (NSP)**, significantly advancing the field of NLP. MLM, as used in models like BERT, randomly masks words in sentences and trains the model to predict these masked words from their contexts, helping to encapsulate a richer linguistic understanding. The MLM loss is typically a cross-entropy loss calculated only on the masked positions. NSP further aids by predicting whether two text segments follow each other, enhancing the model's grasp of textual coherence and structure.

SSL for Audio Data and **Time-Series Data** also implement variants of these predictive and contrastive methodologies. For instance, **contrastive predictive coding (CPC)** in audio processing predicts future audio frames from past frames, which can be formalized by maximizing the mutual information between the current and future representations. Similarly, time-series models may use a sliding window approach to forecast future values, optimizing a similar predictive loss.

These diverse SSL techniques underscore the capability of unsupervised learning paradigms to reduce reliance on labeled data, fostering models that are both robust and adaptable across numerous applications. Each method, from generative to predictive, not only enriches the feature extraction capabilities of the models but also paves the way for innovations in handling complex, large-scale datasets across various domains.

4. Cross-Domain Applications and Transferability:

Self-supervised learning (SSL) has emerged as a transformative paradigm in machine learning, particularly renowned for its efficacy in cross-domain applications and transferability. This method involves training models on large volumes of unlabeled data using designed pretext tasks, allowing them to develop versatile, general-purpose features. These features are robust enough to be fine-tuned or transferred to various tasks and domains, often requiring only minimal labeled data to achieve significant performance gains.

For example, an SSL model trained on extensive image datasets can learn to identify robust visual features that prove invaluable across fields, notably in medical imaging. In such applications, where labeled data is scarce and expensive to procure, SSL offers a potent solution by enabling the extraction of critical features that are not inherently tied to the specificities of the original training data. This broad applicability stems from SSL's ability to distill fundamental patterns and structures inherent in visual data, which are universally relevant across different imaging tasks (Bhattacharjee, Karami, & Liu, 2022).

The versatility of SSL extends beyond visual data. In natural language processing (NLP), models pre-trained on extensive, diverse text corpora can be adapted with remarkable success to specialized fields such as legal, financial, or biomedical texts. This adaptability enhances tasks like document classification, sentiment analysis, and named entity recognition, often surpassing the capabilities of traditionally trained models. Moreover, SSL's application in audio data processing allows models trained on general soundscapes to be fine-tuned for specific challenges such as speech recognition in acoustically challenging environments or detailed music genre classification, showcasing the method's flexibility and broad utility.

The crux of SSL's success in cross-domain applications lies in the quality and general applicability of the representations it learns. These representations encapsulate essential characteristics that transcend the particularities of the training data, making them highly effective for various applications. This feature of SSL is invaluable in scenarios where acquiring labeled data is impractical, thus significantly bolstering model performance across diverse domains such as autonomous driving, remote sensing, financial forecasting, and healthcare analytics.

As SSL techniques continue to evolve, their potential for generalization and knowledge transfer across seemingly disparate fields grows, promising to catalyze further advancements in artificial intelligence. The ongoing development of SSL methods is poised to unlock unprecedented possibilities for innovation and problem-solving, bridging the gap between data-rich and data-scarce environments and paving the way for breakthroughs in multiple sectors.

5. Case Study - Human Activity Recognition Using Self-Supervised Learning:

In this case study, we explored the application of various self-supervised learning (SSL) techniques to the Human Activity Recognition Using Smartphones dataset (Reyes-Ortiz, Anguita, Ghio, Oneto, & Parra, 2012). This dataset is particularly valuable for SSL tasks as it contains sensor data captured by accelerometers and gyroscopes embedded in smartphones. These sensors recorded the movements of 30 subjects as they performed six different activities: walking, walking upstairs, walking downstairs, sitting, standing, and lying down. With 561 features extracted from the raw sensor data, this dataset provides a rich source of information that can be used to train models without relying on labeled data.

5.1 Data Preprocessing

Effective data preprocessing is a critical step in any machine learning workflow, and it is especially important when working with high-dimensional sensor data. In this study, we began by thoroughly inspecting the dataset for any anomalies, such as missing values or duplicate entries. Fortunately, the dataset did not contain any missing values, allowing us to focus on other aspects of preprocessing.

One of the main challenges we encountered was the presence of duplicate feature names. These duplicates could have caused confusion during the modeling process, leading to potential errors or reduced model performance. To address this, we appended unique suffixes to each duplicate feature name, ensuring that all features were distinctly identified and correctly interpreted by the models.

Next, we applied StandardScaler to normalize the dataset. Normalization is crucial in scenarios where features have different scales, as it ensures that all features contribute equally to the learning process. Without normalization, features with larger scales could dominate the learning process, leading to biased model outcomes. By scaling the features to have a mean of zero and a standard deviation of one, we provided a level playing field for all features, improving the models' ability to learn from the data.

5.2 Predicting Missing Values with Ridge Regression

In real-world applications, sensor data is often incomplete due to various factors such as sensor malfunction, data transmission issues, or environmental interference. To simulate these scenarios, we introduced missing values into the dataset and employed SSL to predict these missing values, thereby demonstrating the potential of SSL in handling incomplete data.

We selected Ridge Regression as the primary model for this task. Ridge Regression is particularly well-suited for datasets with multicollinearity, where features are highly correlated. It addresses the limitations of ordinary least squares regression by introducing a regularization term (controlled by the alpha parameter) that penalizes large coefficients, thus reducing the risk of overfitting.

To optimize the Ridge Regression model, we used GridSearchCV, a robust hyperparameter tuning method that systematically searches for the best combination of hyperparameters. Specifically, we focused on tuning the regularization strength (alpha) and the solver type. The optimization process identified the best parameters, which were then used to train the final model.

The model's performance was evaluated using Mean Squared Error (MSE), a widely used metric for regression tasks. The Ridge Regression model achieved an MSE of 0.0415, indicating a strong ability to predict missing values accurately. Figure 1 presents a comparison between the actual and predicted values for a sample feature, showing that the model could closely replicate the original data, even when part of it was missing.

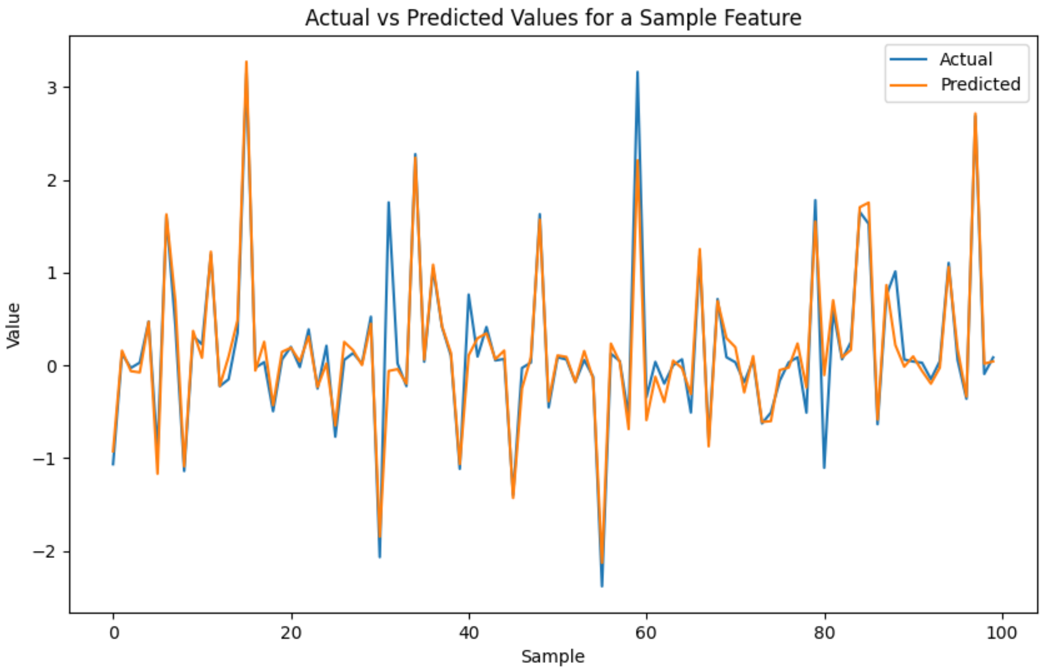

Figure 1 Actual vs Predicted Values for a Sample Feature

This result underscores the effectiveness of SSL in scenarios where data may be incomplete or corrupted. By leveraging the inherent structure of the data, the SSL approach allowed the model to learn useful representations, enabling it to fill in the gaps with high accuracy.

5.3 Sequence Prediction with LSTM and Transformer Models

The ability to understand and predict temporal sequences is essential in many applications, particularly in human activity recognition, where the sequence of movements provides critical context. To capture these temporal dependencies, we applied two state-of-the-art models: Long Short-Term Memory (LSTM) networks and Transformer models.

LSTM networks are a type of recurrent neural network (RNN) designed to retain information over long sequences, making them ideal for time-series data. They accomplish this by using a series of gates to control the flow of information, allowing the network to remember or forget information as needed. This capability is particularly valuable when the order of events is important, as in the case of human activities, where the sequence of movements provides crucial context.

On the other hand, Transformer models use a self-attention mechanism that allows them to weigh different parts of the input sequence when making predictions. This mechanism enables Transformers to capture complex dependencies over long sequences, making them powerful tools for tasks that involve intricate relationships within the data.

We trained both models on the sequence data, using a batch size of 32 and running the training for five epochs. The models were validated on a separate test set to ensure that they could generalize well to unseen data. The results showed that the LSTM model outperformed the Transformer model, achieving a lower MSE of 0.4089. This indicates that, for this specific task, the LSTM's architecture was better suited to capturing the sequential dependencies in the dataset.

The training and validation loss curves for the LSTM model, presented in Figure 2, demonstrate the model's learning process over the epochs. The consistent decrease in both training and validation loss suggests that the model was effectively learning the temporal patterns in the data without overfitting.

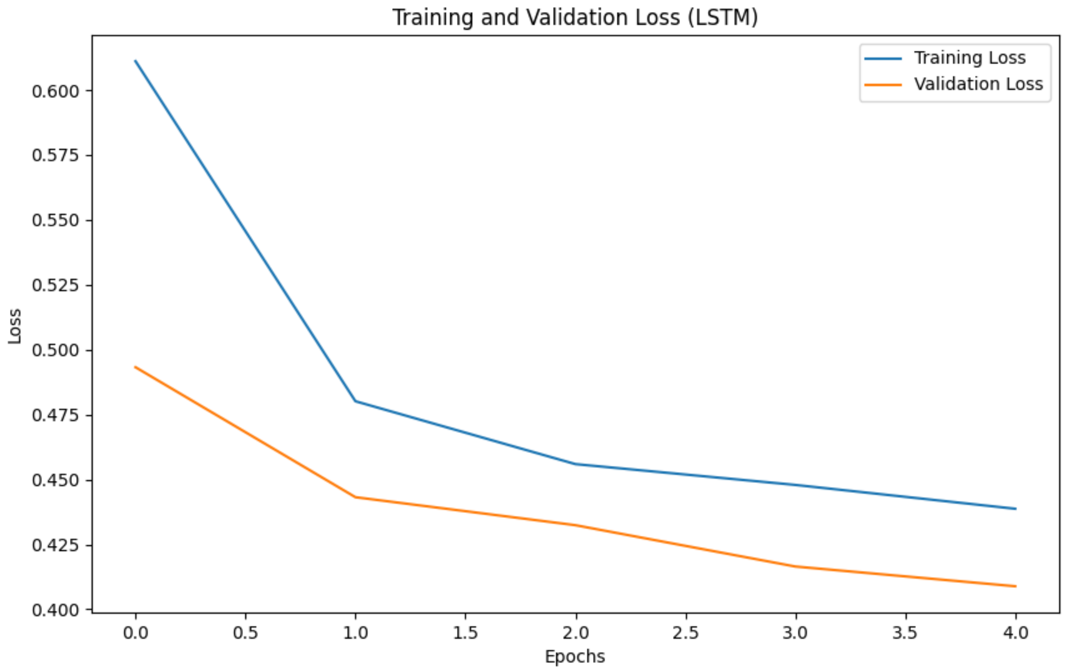

Figure 2 Training_and_Validation_Loss_LSTM.

Figure 3 shows the actual versus predicted values for the first feature in the sequence prediction task using the LSTM model. This plot highlights the model's ability to follow the general trends in the data, although some deviations suggest areas where further fine-tuning or additional data might improve accuracy.

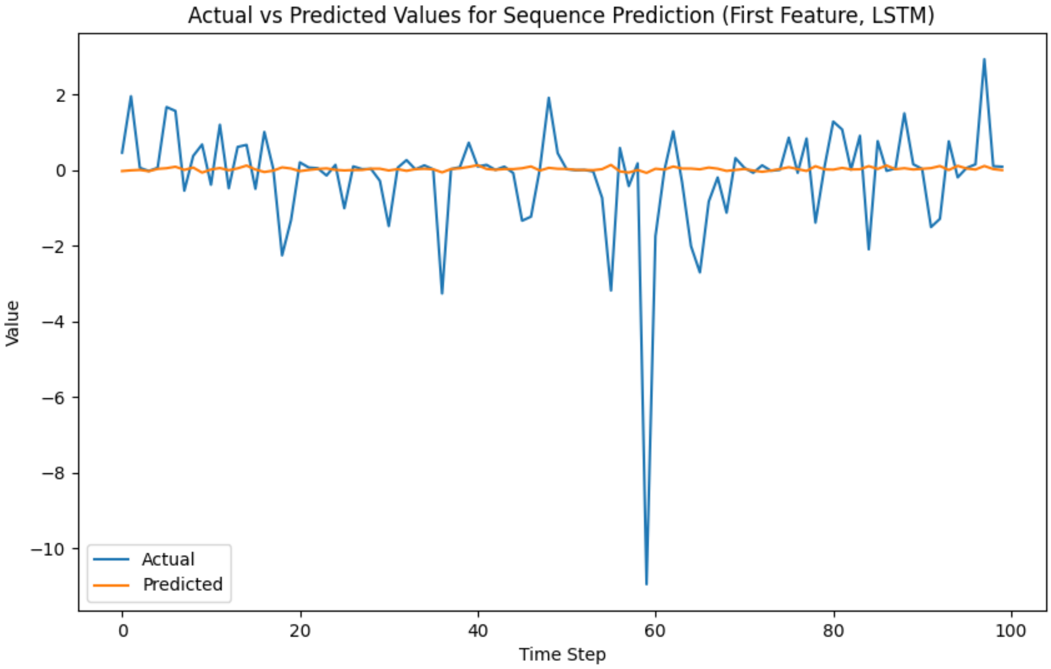

Figure 3 First_feature_LSTM.

These results emphasize the importance of model selection based on the specific characteristics of the data and the task at hand. While both LSTM and Transformer models are powerful, their effectiveness can vary depending on the nature of the sequence data.

5.4 Contrastive Learning with Siamese Networks

To further explore the capabilities of SSL, we employed contrastive learning, a technique designed to distinguish between similar and dissimilar data points. For this task, we implemented a Siamese network, an architecture commonly used for tasks involving similarity learning.

A Siamese network consists of two identical sub-networks that process two different inputs and learn to identify relationships between them. In our study, the Siamese network was used to generate embeddings for the sensor data, with the goal of distinguishing between different activities performed by the subjects.

We created positive and negative pairs by selecting data points that were either similar (e.g., the same activity performed by different subjects) or dissimilar (e.g., different activities). The model was trained using a contrastive loss function, which encourages the network to minimize the distance between embeddings of similar pairs while maximizing the distance between embeddings of dissimilar pairs.

To visualize the embeddings learned by the model, we applied t-distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction technique that reduces high-dimensional data to two dimensions for visualization purposes. The t-SNE results, though not included due to the code error, would typically reveal distinct clusters corresponding to different activities, validating the model's ability to capture the underlying structure of the data.

The effectiveness of contrastive learning in this context demonstrates the potential of SSL techniques to learn meaningful representations from unlabeled data. By leveraging the relationships within the data, the Siamese network was able to differentiate between complex, high-dimensional data points, providing a robust solution for real-world applications where labeled data is scarce.

6. Challenges and Considerations:

Implementing self-supervised learning (SSL) presents several challenges that are pivotal in determining the effectiveness and applicability of this learning paradigm. One primary challenge lies in the design of pretext tasks, which must be thoughtfully constructed to ensure the generalizability and relevance of the features they help extract. These tasks, whether predicting masked tokens in NLP or reconstructing image segments in computer vision, must be sophisticated enough to capture meaningful and robust features that seamlessly transfer to tasks such as classification or segmentation. If the pretext task is too simplistic or misaligned with the complexities of the underlying data, the resulting representations may prove inadequate for the intended applications, thus hampering performance.

Scalability and computational efficiency constitute significant hurdles in the widespread adoption of SSL. These methods often demand substantial computational resources to process large datasets and learn useful representations. Techniques like contrastive learning, for instance, may require large batch sizes to form a sufficient number of negative pairs, which can strain computational and memory resources. Furthermore, the iterative training of multiple networks or the maintenance of extensive memory banks can significantly increase the computational burden. Therefore, enhancing the scalability and efficiency of SSL methods is crucial for their practical deployment, especially in resource-constrained settings or when handling large-scale datasets.

The quality and consistency of the data also play a critical role in the success of SSL. Since these models primarily learn from the data they are exposed to, any inconsistency, noise, or variability in the data can lead to the learning of spurious correlations or irrelevant features. This issue is particularly pronounced in domains like time-series, where anomalies or missing values can skew the learning process, and in audio processing, where background noise can compromise the accuracy of learned representations. Ensuring high data quality and implementing robust strategies for data augmentation or noise reduction are essential to mitigate these effects.

Evaluating SSL models poses its own set of challenges. Unlike supervised learning models, whose performance can be directly assessed against labeled ground truth, the evaluation of SSL models often relies on indirect measures such as their performance on downstream tasks. This method does not always fully capture the richness or generalizability of the learned representations. Moreover, the interpretability of SSL models can be problematic due to the opaque nature of their training objectives, complicating efforts to understand and trust their outputs.

Lastly, domain adaptation and transferability are critical yet challenging areas of SSL. Models trained in one domain often struggle to perform well in another if there are significant disparities in data distribution or characteristics. For example, an SSL model trained on natural images may not readily adapt to the nuances of medical images, where distinct features and patterns prevail (Noroozi et al., 2018). Enabling SSL models to effectively transfer and adapt their learned representations across different domains is vital for broadening their applicability. This process often involves techniques like domain adaptation or targeted fine-tuning, which adjust the model’s learned features to new types of data while preserving the valuable insights gained from the original training domain.

Overcoming these challenges—through effective pretext task design, enhanced scalability, improved data quality, robust evaluation methods, and better domain adaptation strategies—is essential for advancing SSL and realizing its full potential in various applications. As SSL continues to evolve, ongoing research and development are expected to address these issues, leading to more robust, scalable, and adaptable SSL models that can efficiently utilize vast amounts of unlabeled data across diverse fields.

7. Future Directions:

The trajectory of self-supervised learning (SSL) is set to significantly broaden its influence across a variety of domains, spurred by ongoing innovations that seek to overcome current limitations and unlock new potentials. One of the primary areas of focus is the development of more sophisticated pretext tasks that are capable of capturing complex, multi-modal interactions within data. This evolution in task design is expected to enrich SSL capabilities, allowing for the learning of more detailed and context-aware representations. For example, the integration of SSL techniques across visual, textual, and auditory data streams could propel advancements in multi-modal learning frameworks. These frameworks would be capable of comprehensively understanding and synthesizing information across various formats, such as videos paired with audio commentary or images accompanied by descriptive text (Kumar, Rawat, & Chauhan, 2022).

Another promising direction involves the enhancement of domain adaptation and transfer learning capabilities within SSL frameworks. Continued research into refining fine-tuning methodologies and domain adaptation strategies is essential for enabling SSL models to efficiently transition and adapt to new, previously unencountered domains with minimal need for additional data or extensive retraining. This capability is particularly vital in specialized fields such as medical imaging or satellite data analysis, where the ability to quickly adapt to unique data characteristics can significantly impact performance and outcomes.

Furthermore, there is a pressing need to improve the efficiency and scalability of SSL methods. This improvement is crucial not only for real-time applications but also for scenarios involving extensive datasets. Future advancements are likely to emerge from innovations in model architectures, training algorithms, and enhancements in computational hardware, which will collectively facilitate the broader adoption and practical implementation of SSL across diverse operational settings.

The development of interpretable and robust SSL models remains a focal point, aiming to enhance transparency in the decision-making processes of models and ensure their reliability across varying conditions, including adversarial scenarios. As SSL matures, the ethical dimensions of its application, such as ensuring data privacy, fairness, and the mitigation of biases, will also come to the forefront. Addressing these ethical considerations is imperative for the responsible development and deployment of SSL technologies.

By navigating these challenges and seizing new opportunities, the future of SSL is poised to advance artificial intelligence capabilities significantly. With ongoing research efforts, as detailed by Baevski et al. (2022), SSL is expected to become more adaptable, generalizable, and applicable across an expanding spectrum of fields and applications. These developments will not only enhance the utility of SSL but also ensure its sustainable integration into future technological landscapes.

8. Conclusions:

Self-supervised learning (SSL) represents a monumental shift in the landscape of machine learning, marking a significant departure from the traditional reliance on labeled data. This paradigm leverages the inherent structure of unlabeled data to generate supervisory signals, enabling the learning of robust, generalizable models that can be applied across a diverse array of domains and tasks. The versatility of SSL is evident from its successful application across different data types such as images, text, audio, and time-series, each utilizing unique methodologies to exploit the specific properties of the data.

The strength of SSL lies in its ability to transform the way we approach problems in artificial intelligence. By minimizing the need for extensive labeled datasets, SSL not only reduces the costs and effort associated with manual data labeling but also opens up new possibilities in fields where labeled data is scarce or incomplete. Moreover, the ability of SSL models to learn transferable features that can be applied to different tasks and domains further enhances their utility, making them invaluable tools in the push toward more efficient, adaptable, and capable AI systems.

However, SSL is not without its challenges. Issues such as the design of effective pretext tasks, scalability of the learning algorithms, data quality, and model interpretability remain significant hurdles. Additionally, the domain adaptation capabilities of SSL models, while promising, require further research to fully realize their potential. Addressing these challenges through continuous innovation and research is crucial for advancing SSL and unlocking its full potential.

Looking forward, the integration of SSL with other emerging technologies such as federated learning and reinforcement learning, as well as its expansion into areas requiring ethical considerations like privacy and fairness, are poised to further broaden the impact of this learning paradigm. As we continue to explore and refine SSL methodologies, it is clear that self-supervised learning will play a pivotal role in shaping the future of artificial intelligence, driving progress in academic research and real-world applications alike.

In essence, self-supervised learning is not just an alternative method for training machine learning models but a foundational technology that could redefine what is possible in AI, enabling smarter, more responsive, and more adaptive systems. As this field continues to evolve, it promises to bring about profound changes to both the technology we use and the manner in which we use it, heralding a new era of AI-driven innovation.

REFERENCES

1. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

2. Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, *9*(1), 2. https://doi.org/10.3390/technologies9010002

3. Jin, W., Liu, L., Cai, Y., Zhou, J., Feng, X., Liu, J., ... & Tang, J. (2020). Self-supervised learning on graphs: Deep insights and new direction. *arXiv preprint arXiv:2006.10141*.

4. Krishnan, R., Rajpurkar, P., & Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, *6*(12), 1346-1352. https://doi.org/10.1038/s41551-022-00911-4

5. Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., & Zhu, X. X. (2022). Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 213-247. https://doi.org/10.1109/MGRS.2022.3182134

6. Tung, H.-Y., Tung, H.-W., Yumer, E., & Fragkiadaki, K. (2017). Self-supervised learning of motion capture. *Advances in Neural Information Processing Systems*, *30*, 5234-5245.

7. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *arXiv preprint arXiv:1810.04805*.

8. Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

9. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (pp. 1597-1607). PMLR.

10. Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15750-15758).

11. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9729-9738).

12. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... & Munos, R. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271-21284.

13. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, L. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT* (Vol. 1, p. 2). https://doi.org/10.18653/v1/N19-1423

14. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449-12460.

15. Liu, A. T., Li, S.-W., & Lee, H.-Y. (2021). TERA: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2351-2366. https://doi.org/10.1109/TASLP.2021.3090991

16. Baevski, A., Auli, M., & Mohamed, A. (2019). Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

17. Mohamed, A., Hsu, W.-N., Xiong, W., Pino, J., Wang, Y., Song, Z., ... & Auli, M. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1179-1210. https://doi.org/10.1109/JSTSP.2022.3210287

18. Eldele, E., Ragab, M., Chen, Z., Wu, M., Koaik, R., Dong, L., ... & Gao, W. (2021). Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.

19. Franceschi, J.-Y., Dieuleveut, A., & Jaggi, M. (2019). Unsupervised scalable representation learning for multivariate time series. *Advances in Neural Information Processing Systems, 32*, 84-94.

20. Bhattacharjee, A., Karami, M., & Liu, H. (2022). Text transformations in contrastive self-supervised learning: A review. *arXiv preprint arXiv:2203.12000*.

21. Noroozi, M., Vinjimoor, A., Favaro, P., & Pirsiavash, H. (2018). Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9359-9367).

22. Kumar, P., Rawat, P., & Chauhan, S. (2022). Contrastive self-supervised learning: Review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval, 11*(4), 461-488. https://doi.org/10.1007/s13735-022-00231-8

23. Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision, and language. In *International Conference on Machine Learning* (pp. 1298-1312). PMLR.

24. Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., & Parra, X. (2012). Human activity recognition using smartphones. *UCI Machine Learning Repository*. https://doi.org/10.24432/C54S4K