Uncovering the Power of Semi-Supervised Learning

VAJRATIYA VAJROBOL 1

[1]International Center for AI and Cyber Security Research and Innovations. Asia University, Taiwan.

(e-mail: vvajratiya@gmail.com).

ABSTRACT Semi-supervised learning has been adopted in various sectors, offering improved machine learning models that maximize the use of labeled and unlabeled data. The label is adopted as an attempt to solve problems in various industries. Semi-supervised learning: a new paradigm addresses the traditional limitations of traditional supervised methods. It addresses theoretical foundations, advantages such as effective layerwise training of deep networks, and applications to NLP, computer vision, and healthcare. We discuss success stories in industry, challenges, and major techniques, including pseudo-labeling. Lastly, the article concludes with a firm indication of how semi-supervised learning could transform various machine learning paradigms.

KEYWORDS Semi-supervised learning, supervised learning, Machine Learning

I. Introduction

A. Brief overview of supervised learning

A key component of machine learning has always been supervised learning, in which models are trained on labeled datasets and assigned labels based on the input data. This paradigm has shown to be successful in a number of applications, including natural language processing and picture recognition. However, classic supervised learning algorithms encounter scaling issues and resource-intensive labeling operations when the amount of available data grows rapidly [1].

B. Limitations of traditional supervised learning

Often, traditional supervised learning approaches are overwhelmed by the enormous amount of unlabeled data that is available. Large volumes of data become too costly and impossible to manually label, creating a bottleneck in the training pipeline. Furthermore, labeled datasets might not adequately represent the variety and intricacy of real-world situations, which would hinder the model's capacity for successful generalization [2].

C. Introduction to semi-supervised learning

Semi-supervised learning is an alternative since it acknowledges the shortcomings of supervised techniques. This hybrid paradigm enhances model performance by integrating labeled and unlabeled data [3]. Semi-supervised learning, as opposed to classical supervised learning, acknowledges that not all data need explicit labels in order to train models efficiently. This approach leverages the hidden information included in the massive volumes of unlabeled data to provide a deeper and more comprehensive understanding of the underlying patterns.

In semi-supervised learning, the model makes use of the inherent relationships and patterns found in unlabeled data in addition to learning from the labeled instances [4]. In addition to addressing the difficulties presented by large datasets, this also makes use of latent features, producing models that are more reliable and broadly applicable. Through the use of sophisticated algorithms, the combination of labeled and unlabeled data allows semi-supervised learning to overcome the constraints of supervised techniques.

In the parts that follow, We will look more closely at internal functioning of semi-supervised learning in the sections that follow. We will also examine how to extract meaningful information from unlabeled data and talk about the advantages in terms of model scalability, practicality, and accuracy. We will go into the core concepts, algorithmic foundations, and noteworthy uses of semi-supervised learning as we reveal its power and show how it can revolutionize machine learning as a whole.

II. Understanding of semi-supervised learning

At the intersection of supervised and unsupervised techniques, semi-supervised learning provides a distinct method for machine learning. Instead of going to the extremes of only using explicit labels or looking at datasets that have never been seen before, the model in this hybrid paradigm gains insights from both labeled and unlabeled data. Semi-supervised learning is a strategy that combines the best features of both traditional supervised learning and unsupervised learning. Traditional supervised learning relies on labeled examples, while unsupervised learning explores unlabeled data.

To completely appreciate the significance of semi-supervised learning, one must compare it with other options. In supervised learning, models that are trained on labeled data map inputs to suitable outputs. Unsupervised learning,

on the other hand, searches datasets without predefined labels for inherent patterns and relationships [4]. Semi-supervised learning encourages a more comprehensive and adaptable learning process by making use of both the vast pool of unlabeled data to uncover hidden patterns and the clarity and direction that labeled data offers.

III. Theoretical Foundations

A. Overview of the self-training method

A key component of semi-supervised learning's theoretical foundations is the self-training approach [5]. Self-training essentially entails using the available labeled data to iteratively train a model, which is then used to predict labels for the instances that are not labeled. The training set is then updated with these projected labels, thereby increasing the amount of labeled data. This iterative process goes on, improving the model's performance over time by improving its understanding and leveraging the growing labeled dataset.

B. Co-training and its application in semi-supervised learning

Another key theoretical component of semi-supervised learning that suggests a cooperative learning approach is co-training. The model is trained utilizing different data viewpoints or subsets when co-training [6]. This assumes that these subsets are conditionally independent given the output labels. As one perspective validates the insights gained from the other, cooperation happens as the model iteratively enhances its understanding. This approach performs exceptionally well when there is sufficient data in one perspective but insufficient labeled data in another, leading to a comprehensive and well-rounded learning process.

C. Multi-view learning as a mechanism for exploiting diverse data perspectives.

Multi-view learning is a prominent technique in the field of semi-supervised learning that allows for the utilization of several viewpoints found in the data [7]. This method takes into account various data representations , each providing a distinct understanding perspective. The model gets a stronger and more thorough understanding of the underlying patterns by concurrently learning from these several points of view. This is beneficial, particularly in situations where a single viewpoint might not be adequate to fully convey the nuances of the information. The effectiveness of semi-supervised learning frameworks is boosted overall by multi-view learning, which turns into a potent tool for utilizing the richness and diversity found in both labeled and unlabeled datasets.

IV. Benefits of Learning Under Semi-Supervision

A. Efficient utilization of limited labeled data

The capacity of semi-supervised learning to maximize the use of a small pool of labeled data is one of its main advantages. Obtaining labeled data in typical supervised algorithms can be expensive and resource-intensive. On the other hand, semi-supervised models effectively utilize the available labeled examples, optimizing the utility of each labeled sample, by including unlabeled input during the learning process [8]. This becomes especially useful in situations when getting large-scale annotated datasets is difficult or not feasible.

B. Improved generalization and model robustness

Machine learning models that use semi-supervised learning perform better overall and in terms of generalization [9]. Learning from both labeled and unlabeled data exposes the model to a wider range of patterns and scenarios found in the data. This exposure makes it easier to comprehend the underlying structures more thoroughly, which improves the model's ability to generalize to new examples

C. Cost-effective model training and scalability

A major benefit of semi-supervised learning is its scalability and cost-effectiveness [6]. Large data sets may be laboriously and financially expensive to manually label. By utilizing the abundance of unlabeled data, semi-supervised techniques lessen this load by eliminating the requirement for intensive labeling efforts. This improves scalability and reduces the cost of training, enabling models to process large datasets without hitting the bottlenecks linked to conventional supervised learning. Semi-supervised learning is a desirable alternative in applications where accuracy and cost are critical factors because of its effective utilization of resources.

V. Practical Applications

Semi-supervised learning has useful applications in the field of Natural Language Processing (NLP). Sentiment analysis is one prominent area where the model may be used to identify and comprehend sentiments expressed in vast amounts of text by utilizing both labeled and unlabeled textual data [10]. Furthermore, semi-supervised methods are advantageous for Named Entity Recognition (NER), an important information extraction job, since they allow the model to better recognize and categorize things in unstructured data by using unlabeled text [11].

Entering the field of computer vision, semi-supervised learning demonstrates effectiveness across a range of tasks. Models can use both labeled and unlabeled images in image classification to increase robustness and accuracy [12]. Another important use is object detection, which makes use of the viewpoints provided by semi-supervised learning [13]. By using a combination of labeled and unlabeled visual input, the model is able to identify and categorize things in intricate scenarios.

Semi-supervised learning plays a major role in the healthcare industry in activities like medical picture analysis and disease detection [14]. These models can improve their diagnostic performance and expand to a wider spectrum of patient cases by integrating unlabeled medical data. Semi-supervised learning is a promising strategy for enhancing

medical decision-making and developing healthcare technology since it can be used to large unlabeled datasets as well as labeled patient data.

VI. Challenge

It presents issues that require careful attention when using the semi-supervised learning . While labeled datasets have clear standards, one significant challenge is maintaining the quality of unlabeled data [15]. Strategies like data cleaning and preprocessing are necessary to sort through the inconsistent relevance and accuracy of unlabeled samples and preserve the integrity of the learning process.

Considering any biases that may exist in the unlabeled dataset is another important factor to examine. Skewed and unjust predictions might result from unintentional biases in these datasets being reinforced during model training [16]. Ensuring that the model's outputs are not unintentionally influenced by unfair or erroneous patterns contained in the unlabeled data requires careful examination and remedial actions in order to detect and mitigate biases.

Especially semi-supervised scenarios and model validation, conventional measures might not be sufficient. Carefully choosing metrics that accurately assess the model's performance is necessary due to the distinct interaction between labeled and unlabeled data [17]. To evaluate the model's capacity for broad generalization and stable outputs in the face of real-world complexity, it becomes imperative to balance the evaluation across both kinds of datasets.

VII. Techniques and Algorithms

A. Pseudo-labeling

One well-known method is pseudo-labeling, in which the model creates pseudo-labels based on its predictions on unlabeled data. These pseudo-labels are used as ground truth during training, so converting unlabeled examples into de facto labeled samples. Through the iterative process, the model improves performance and generalization by absorbing insights from unlabeled data to further expand its knowledge [18].

B. Co-training algorithms

Co-training algorithms are an example of a cooperative method where the model is trained using several perspectives or data subsets [19]. Every viewpoint offers a different viewpoint, and the collaboration happens when the model improves its comprehension by taking in information from one viewpoint and then confirming it with information from another. This method ensures a more thorough learning process and works especially well when some characteristics of the data are better represented in one view but not in the other.

C. Consistency regularization

The goal of consistent regularization is to encourage reliable and consistent predictions in the face of various input data perturbations. The goal of training the model is to provide comparable results for marginally altered inputs, whether via data augmentations or other perturbation techniques. This incentivizes the model to identify stable and consistent features within the data, hence utilizing unlabeled examples to enhance generalization and overall performance [20].

These methods and strategies demonstrate the flexibility of semi-supervised learning, providing ways to efficiently integrate unlabeled data and improve the general performance of machine learning models.

VIII. Real-world Success Stories

A. Google's use of semi-supervised learning for improved search algorithms

Google has improved its search engines with the help of semi-supervised learning. Google's algorithms can better comprehend the subtleties of user searches and increase the relevance of search results by utilizing both labeled and unlabeled data [21]. Google is able to improve its algorithms and produce more precise and contextually relevant search results because of the abundance of unlabeled data on the internet, which offers insightful information.

B. Applications in recommendation systems

In recommendation systems, where user choice prediction is crucial, semi-supervised learning has achieved great success. Recommendation algorithms can offer more accurate and personalized suggestions if they use both unlabeled data (implicit user behavior) and labeled data (expressed user ratings) [22]. With the help of this strategy, websites like Netflix and Amazon may improve user experience by providing personalized recommendations that are based on a more thorough grasp of customer interests.

C. Advancements in autonomous vehicles using semi-supervised techniques.

These real-world success stories highlight how semi-supervised learning may be applied practically in a variety of fields, demonstrating how it can improve performance, flexibility, and efficiency when addressing challenging, data-intensive issues [23-28].

IX. Conclusions

In conclusion, the analysis of semi-supervised learning reveals an interplay involving its pros and cons. Our approach tackles the labeled data side constraints effectively and at less cost through the adoption of untapped datasets. Therefore, continued research and training programs are important to encourage usage and further inquiry in the industry, embracing its practical benefits. Semi-supervised learning can change the course of machine learning paradigms, leading to general approaches and creating a deep understanding of hard inorganic data. Finally, the breakthrough of semi-supervised learning appears to be a breakthrough force in AI that changes the very nature of machine learning.

References

[1] Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

[2] Lee, V. L. S., Gan, K. H., Tan, T. P., & Abdullah, R. (2019). Semi-supervised learning for sentiment classification using a small number of labeled data. Procedia Computer Science, 161, 577-584.

[3] Pise, N. N., & Kulkarni, P. (2008, December). A survey of semi-supervised learning methods. In 2008 International conference on computational intelligence and security (Vol. 2, pp. 30-34). IEEE.

[4] Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.

[5] Zhu, X., & Goldberg, A. B. (2022). Introduction to semi-supervised learning. Springer Nature.

[6] Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: A brief review. Int. J. Eng. Technol, 7(1.8), 81.

[7] Yan, X., Hu, S., Mao, Y., Ye, Y., & Yu, H. (2021). Deep multi-view learning methods: A review. Neurocomputing, 448, 106-129.

[8] Rebuffi, S. A., Ehrhardt, S., Han, K., Vedaldi, A., & Zisserman, A. (2020). Semi-supervised learning with scarce annotations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 762-763).

[9] Mey, A., & Loog, M. (2022). Improved Generalization in Semi-Supervised Learning: A Survey of Theoretical Results. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4747-4767.

[10] Silva, N. F. F. D., Coletta, L. F., & Hruschka, E. R. (2016). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Computing Surveys (CSUR), 49(1), 1-26.

[11] Aryoyudanta, B., Adji, T. B., & Hidayah, I. (2016, July). Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm. In 2016 International seminar on intelligent technology and its applications (ISITIA) (pp. 7-12). IEEE.

[12] Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. In Proceedings of the European conference on computer vision (eccv) (pp. 135-152).

[13] Melacci, S., Maggini, M., & Gori, M. (2009). Semi–supervised learning with constraints for multi–view object recognition. In Artificial Neural Networks–ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II 19 (pp. 653-662). Springer Berlin Heidelberg.

[14] Mohanasundaram, R., Malhotra, A. S., Arun, R., & Periasamy, P. S. (2019). Deep learning and semi-supervised and transfer learning algorithms for medical imaging. In Deep learning and parallel computing environment for bioengineering systems (pp. 139-151). Academic Press.

[15] Simmler, N., Sager, P., Andermatt, P., Chavarriaga, R., Schilling, F. P., Rosenthal, M., & Stadelmann, T. (2021, June). A survey of un-, weakly-, and semi-supervised learning methods for noisy, missing and partial labels in industrial vision applications. In 2021 8th Swiss Conference on Data Science (SDS) (pp. 26-31). IEEE.

[16] Zhang, T., Zhu, T., Li, J., Han, M., Zhou, W., & Philip, S. Y. (2020). Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. IEEE Transactions on Knowledge and Data Engineering, 34(4), 1763-1774.

[17] Chouldechova, A., Deng, S., Wang, Y., Xia, W., & Perona, P. (2022, October). Unsupervised and semi-supervised bias benchmarking in face recognition. In the European Conference on Computer Vision (pp. 289-306). Cham: Springer Nature Switzerland.

[18] Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020, July). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[19] Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. In Proceedings of the European conference on computer vision (eccv) (pp. 135-152).

[20] Wang, X., Chen, H., Xiang, H., Lin, H., Lin, X., & Heng, P. A. (2021). Deep virtual adversarial self-training with consistent regularization for semi-supervised medical image classification. Medical image analysis, 70, 102010.

[21] Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr, E. R., & Mitchell, T. M. (2010, February). Coupled semi-supervised learning for information extraction. In Proceedings of the third ACM international conference on Web search and data mining (pp. 101-110).

[22] Khan, A. H., Siddqui, J., & Sohail, S. S. (2022). A survey of recommender systems based on semi-supervised learning. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 3 (pp. 319-327). Springer Singapore.

[23] Zhang, J., Liu, H., & Lu, J. (2022). A semi-supervised 3D object detection method for autonomous driving. Displays, 71, 102117.

[24] Casillo, M., Colace, F., Gupta, B. B., Lorusso, A., Marongiu, F., Santaniello, D., & Valentino, C. (2022, January). A situation awareness approach for smart home management. In *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)* (pp. 260-265). IEEE.

[25] Ahmad, I., Qayyum, A., Gupta, B. B., Alassafi, M. O., & AlGhamdi, R. A. (2022). Ensemble of 2D residual neural networks integrated with atrous spatial pyramid pooling module for myocardium segmentation of left ventricle cardiac MRI. *Mathematics*, *10*(4), 627.

[26] Quamara, M., Gupta, B. B., & Yamaguchi, S. (2021, January). An end-to-end security framework for smart healthcare information sharing against botnet-based cyber-attacks. In *2021 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 1-4). IEEE.

[27] Gupta, B. B., & Quamara, M. (2018). A dynamic security policies generation model for access control in smart card based applications. In *Cyberspace Safety and Security: 10th International Symposium, CSS 2018, Amalfi, Italy, October 29–31, 2018, Proceedings 10* (pp. 132-143). Springer International Publishing.

[28] Akhtar, T., & Gupta, B. B. (2021). Analysing smart power grid against different cyber attacks on SCADA system. *International Journal of Innovative Computing and Applications*, *12*(4), 195-205.