PATRIK REIZINGER

The Falsificationist View of Machine Learning

Machine learning pushes the frontiers of algorithmic achievements, though the striving for state-of-the-art performance often obscures the fragility of enforcing decisions amid uncertainty. This paper interprets machine learning within Karl Popper's epistemology. We assess machine learning paradigms' fit for falsificationism and argue that the new interpretation can improve robustness. Though the price is to accept unambiguous decisions, the restriction of the hypothesis space still adds value. The context for our work is established by comparison with similar techniques and highlighting its limitations.

Keywords: *machine learning, epistemology, artificial intelligence, falsificationism, Popper, robustness*

Acknowledgment

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and ELLIS for supporting Patrik Reizinger.

Author Information

Patrik Reizinger, University of Tübingen, IMPRS-IS, ELLIS
https://orcid.org/0000-0001-9861-0293

How to cite this article:

Reizinger, Patrik. "The Falsificationist View of Machine Learning".
Információs Társadalom XXIII, no. 2 (2023): 117–127.
https://dx.doi.org/10.22503/inftars.XXIII.2023.2.7

1. Introduction

Falsificationism (Popper 2002) provides a sobering view of scientific progress—an insight generally neglected by engineers. Applying modern scientific advancements requires making decisions in complex environments, but optimizing performance often sacrifices robustness.

Nassim Nicolas Taleb (2020, 2007) argues that the 21st century challenges humanity with Black Swans—highly improbable events with considerable losses. Such events are the reality of solutions in medicine (Monti, Zhang and Hyvärinen 2020) or autonomous driving (Szemenyei and Reizinger 2019) – often powered by artificial intelligence (AI). Although researchers made notable progress in protecting neural networks against adversarial examples (Schott et al. 2018) and quantifying uncertainty (Gawlikowski et al. 2021), the authors argue that the field could benefit from adopting Popper's philosophy. That is, the process of falsification: a careful evaluation of neural networks' predictions.

The belief of obtaining reliable, task-specific models with a limited amount of data and the ever-increasing pressure to achieve state-of-the-art performance obscure the fragility of the quest for perfection in a noisy setting: the need for a decision disregards whether the best solution is reliable and superior compared to the alternatives, resulting in notorious failures.

The Popperian flavor of mathematical methods is not unknown: statistical hypothesis testing. (Gretton et al. 2007) provides conclusions based on falsifying hypotheses, and is applied, e.g., in software testing (Tóth et al. 2017). This paper examines modern machine learning methods in the falsificationist context, arguing that constraining the hypothesis space would improve decision quality and reliability, as opposed to striving for an unambiguous decision.

Popper's philosophy inspired several researchers in the sciences, even in the broad context of learning systems. Berkson and Wettersten (1984) showed that falsificationism can be seen as learning theory—nonetheless, differences were also highlighted, e.g., by comparing Popper's degree of falsifiability and the Vapnik-Chervonenkis (VC) dimension of statistical learning theory (Corfield, Schölkopf and Vapnik 2009). Vasconcelos, Cardonha and Gonçalves (2018) utilized Popper's philosophy for fair hiring decisions.

Similar to Vasconcelos, Cardonha and Gonçalves (2018), we rely on falsificationism and utilize it in a broad sense. Namely, Popper's claim—that probabilistic statements are neither verifiable nor falsifiable (Popper 2010)—would render his arguments invalid for probabilistic machine learning. However, defining a decision threshold for probabilities admits a falsificationst view of probabilistic systems. By drawing parallels to Popper's philosophy, our goal is to motivate an epistemological view of machine learning.

Although we start from a theoretical assessment, our conclusions focus on the practical. We claim that falsifying hypotheses can potentially reduce false predictions and express uncertainty while providing additional information compared to traditional machine learning approaches. According to our best knowledge, this is the first work that interprets modern machine learning in a Popperian way.

First, we discuss supervised, self-supervised, unsupervised, and reinforcement learning. We conclude that the Popperian approach is generally applicable. Since unsupervised learning lacks hypotheses, falsificationism can only apply if we have a priori assumptions about the representation. Our analysis contrasts classification and regression methods—pointing out that falsificationism naturally fits only the former, though techniques exist for the latter as well.

Second, we compare the merits of falsificationism to other robustness-improving strategies, such as predicting confidence or using ensembles. Third, we address the case of adversarial examples to point out the limitations of our proposal. By providing an epistemological context for machine learning algorithms, we hope to inspire a discussion that improves the robustness of AI.

2. The falsificationist machine learning taxonomy

This section analyzes machine learning concepts from a falsificationist point of view, assesses their suitability for the paradigm, and also addresses practical implications.

2.1. Terminology

Machine learning algorithms constitute a mapping between different domains: the data fed to the network is called the input sample or observation, and the output of the network is the prediction (this can be discrete or continuous). When discrete, it is generally called a label/class, whereas prediction or action can be both. In the discrete case, we speak of classification, in the continuous, of regression. The network learns and uses a (latent) representation, whereas the goal specified by the designer is prescribed by the objective/loss.

2.2. Supervised Learning

Supervised learning learns a mapping given pairs of inputs and desired outputs. Examples include image classification (the image is the input and the label is the desired output) or stock price prediction (a time series is the input; the next element in the future is the desired output). The difference compared to rule-based algorithms is that, although we know the desired output, we cannot specify how to produce it from the inputs.

Classification

Standard algorithms maximize the correct class's probability—incorrect labels can have arbitrarily close probabilities unless they are less than the correct one. The falsificationist interpretation requires that all incorrect labels' probabilities are pushed toward zero—this is equivalent to maximizing the correct label's probability. To contrast the differences, consider classification with three labels. Technically, the correct class is predicted as soon as it has the highest probability. However, this approach is agnostic to the difference to the second-highest probability: it considers the solution correct even if the three probabilities are 0.331, 0.33, and 0.299 (the correct class having the highest probability). Minimizing the incorrect classes' probabilities requires the highest possible difference, implying a notion of robustness.

This is similar to the loss (called hinge loss) of support vector machines (SVMs) (Schölkopf et al. 1999), where the model is incentivized to increase the correct label's probability above the others plus a margin. As the margin goes toward one, we recover the falsificationist approach.

Pushing the label distribution to a Dirac delta is not novel—compare it with the review of Gawlikowski et al. (2021)—though its falsificationist interpretation is. We advocate for the falsificationist approach for a more reliable prediction and a safer failure. Namely, hard-to-classify samples could potentially have a predicted label distribution with multiple large entries. Driving any of those to zero—even if not leading to a definitive conclusion—can help to reduce the hypothesis class yielding (partial) suspension of judgment, since the output is a label set, but it is reduced.

Does this mean that the model does not provide added value? On the contrary, as the hypothesis space is reduced, the false certainty of a crisp decision is alleviated. Acknowledging the practical limitations, we should not deceive ourselves that algorithms can always make a good decision.

A similar approach to the falsificationist view in classification is set-valued prediction, where a model generally outputs the labels with the k highest probabilities (Lapin, Hein and Schiele 2016; Mortier, Hüllermeier and Waegeman 2022). This approach reduces the hypothesis class in the same way as the falsificationist approach—one could also predict the labels above a specific probability threshold.

Regression

Regression differs from classification in having an infinite label set. Theoretically, this is no burden to adapt the falsificationist view on the predictions (possible theories in the sciences are also uncountable), but it restricts practical applicability, as excluding infeasible options will not lead to the correct solution. To overcome this problem, one could divide the values into mutually exclusive categories—turning the regression problem into classification—but that would sacrifice resolution.

2.3. Unsupervised Learning

Unsupervised algorithms—e.g., clustering algorithms such as k-Means or t-SNE (Van der Maaten and Hinton 2008)—utilize unlabeled information to extract a "good" representation that can be used to solve multiple downstream tasks—their competitive advantage is avoiding the expensive labeling process. Since the desired output is unknown, the only feedback about performance is via the objective function.

Architectures such as Variational AutoEncoders (VAEs) (Kingma and Welling 2013) learn a representation and a generative model for, e.g., generating realistic images—i.e., there are no labels but a continuous reconstruction loss with a Kullback-Leibler divergence as inductive bias. Thus, a falsificationist interpretation does not apply to the predictions. We could treat models with an error above a threshold as falsified, but this would not bring us closer to the solution—it would reason about the reconstruction, not the representation.

Moreover, reconstruction quality is only a necessary indicator of a high-quality representation. As Alemi et al. (2018) point out, even a meaningless latent representation can produce good samples.

Thus, falsificationism is not practically applicable on the predictions, for we cannot restrict the hypothesis space, similar to regression—lacking the desired output, we do not even know the hypothesis space. However, if we assume that the underlying representation has particular (measurable and testable) properties such as independence, then evaluating such properties can be a basis for falsification. But the clear difference compared to any other paradigms is that the falsificationist view concerns the properties of the representation, not the predictions.

2.4. Self-Supervised Learning

Self-supervised learning (Lil'Log 2019) resembles supervised learning in the sense that labels are predicted from observations, except that the labels are generated by the model itself. This auxiliary labeling process can exploit unlabeled data (similar to unsupervised learning) and makes falsification feasible in some scenarios.

2.4.1. Generative Adversarial Networks (GANs)

GANs (Goodfellow et al. 2014) consist of a discriminator and a generator. The discriminator distinguishes between real and generated ("fake") images, whereas the generator's role is to deceive the discriminator by producing realistic samples. The architecture casts generative modeling into binary classification, so falsificationism is applicable. Nonetheless, this bears no practical advantage as excluding one label means the same as accepting the other.

2.4.2. Contrastive Learning

Contrastive learning (Lil'Log 2021a; Chen et al. 2020) is the perfect practical example of falsificationism. It learns a representation via an auxiliary classification task by combining samples with different (negative pairs) and same labels (positive pairs). Its objective incentivizes alignment, uniformity, and separability (Wang and

Isola 2020). By alignment, samples of the same class are forced to have a similar representation, whereas negative pairs are mapped to different latents, ensuring separability. Uniformity incentivizes evenly distributed representations in the latent space.

Positive and negative pairs reflect how scientists verify hypotheses. If the representation (“the hypothesis”) is not able to match specific observations (the positive pairs), then it is necessarily wrong. Nevertheless, this is not sufficient; we can imagine that if all samples get mapped to the same representation (known as mode collapse), then the alignment is perfect, but the representation is meaningless. By enforcing different representations for negative pairs, they are “repelled” from each other, rendering them distinguishable.

The similarity is more evident via hard negative mining (Robinson et al. 2020), which collects negative samples with similar representations but different labels. Hard negative samples contain more information, so they help refine the decision boundary between classes and improve separability. Thus, hard negatives are more informative and have a higher degree of falsifiability.

2.5. Semi-supervised Learning

Semi-supervised learning combines supervised and unsupervised strategies (Lil’Log 2021b) and (Lil’Log 2021a), relying on a large unlabeled and a much smaller labeled dataset. Some methods utilize both datasets simultaneously, whereas others rely on unsupervised data for representation learning (pre-training) and then deploy the labeled samples for fine-tuning.

Being a mixture of two paradigms, we can rely on the conclusions of the parts. Generally, supervised learning admits the falsificationist approach on the prediction, whereas unsupervised learning does not—since in this case the representation is used for a supervised task, it might not be meaningful to discuss the falsifiability of the representation’s properties. Depending on classification/regression, the same considerations apply as above.

2.6. Reinforcement Learning

Reinforcement learning resembles how humans learn: a decision-making agent interacts with its environment via its actions and receives a reward (feedback). As the agent does not have access to the optimal policy, which it aims to learn, it can compare two actions only relatively, based on the received reward.

Actions can be discrete or continuous, so the classification and regression arguments—discussed for supervised learning—apply. The only difference is that here we do not have the correct output. Nonetheless, we can utilize the same practical strategies to restrict the possible solutions (e.g., when predicting an action from a discrete set); thus, bearing the benefits of the falsificationist approach.

3. Extensions

This section discusses strategies to increase the robustness of machine learning models. We assess how using “unknown” labels, confidence scores, uncertainty estimates, and ensembles relate to falsificationist machine learning.

3.1. “Unknown” labels

Including a label for objects from unknown classes (usually denoted as UNK) in classification can signal uncertain decisions (Radford et al. 2019)—and makes the hypothesis space complete. This strategy is crucial in out-of-distribution (OOD) data (Yang et al. 2021), or the extreme case of changing environments, where not only the class probabilities change but also new classes are introduced.

One could argue that the UNK label expresses suspension of judgment. We believe this is partially true, since it can express that with a single prediction the output is inconclusive. However, it can fail for hard-to-categorize data. Imagine an animal classification task with cows, -cats, dogs, and UNK. Assume the prediction assigns 0.501 to the “dog” and 0.499 to the “cat” label. Despite including UNK, a standard algorithm would predict “dog,” although UNK expresses the uncertainty better. However, predicting UNK would neglect the information that the hypothesis space is restricted to dogs and cats (based on the probabilities of this example). The falsificationist view would suspend judgment and give the reduced label set “dog” and “cat.“

Still, it is beneficial to include the UNK label. Assume the same probability distribution as before, but the object is a lion. Not including UNK would imply the object is a dog or a cat. Suspension of judgment is still the correct answer, but the hypothesis space of “dog” and “cat” would be incorrect. This example highlights that falsificationist machine learning and the UNK label are orthogonal: falsificationism makes the decisions more robust (even with sacrificing unambiguity) for a *given* hypothesis space, whereas the UNK label extends the hypothesis space to account for distributional changes.

3.2. Confidence scores

Confidence scores express prediction feasibility with a value in [0;1]. In object detection (Szemenyei and Reizinger 2019), when the number of objects can vary within scenes, confidence represents whether an object is present; thus, filtering out false positives is a compensation for the architectural bias that fixes the number of objects. In natural language applications—such as in Microsoft (2021)—their purpose is to estimate chatbot answer quality. We focus on the latter case.

Is a high confidence score necessary or sufficient for correct predictions? Assume an image of an animal resembling both a dog and a cat (which is not unrealistic). Our classifier predicts the label and its confidence and assigns a probability of 0.501 to “dog” and 0.499 to “cat.” Assume that the prediction is correct; the image contains a dog. Low confidence would be consistent with the probabilities (the same holds when the image is of a cat), though it would not contain additional value compared to the label distribution. Although high confidence would not contradict the correctness of the prediction, it could not express the label distribution's uncertainty.

Thus, we conclude that confidence scores can be redundant to the label distribution for prediction quality assessment. Moreover, predicting confidence—similar to the UNK label—neglects information: a "dog" label with a confidence of 0.6 is less informative than outputting the label set "dog" and "cat" and requires an additional mechanism (to predict the confidence scores), whereas the falsificationist approach does not.

On the other hand, when confidence describes another property (e.g., the presence of an object as in the object detection case), it remains useful, as such information is not contained in the labels. That is, if no object is present, predicting zero confidence scores for all classes could be a potential solution; however, including a separate class label of no object might also be sufficient.

3.3. Ensembles

Ensembles (also known as expert systems) pool different models to increase prediction performance and reliability (Ţifrea, Stavarache and Yang 2021; Pathak, Gandhi and Gupta 2019; Masegosa 2020) with multiple strategies, e.g., consensus or majority vote for classification and weighting for regression.

If the models agree, their prediction is accepted. In the case of classification, this means that all/most models predict the same label, whereas regression needs a tolerance (e.g., to assess whether the predictions are "close enough"). On the other hand, the prediction is falsified when one model draws a different conclusion. Even in that case, the set/range of predictions for classification/regression could be used as the reduced hypothesis space—a clear practical advantage.

Ensembles can potentially assess representation quality in unsupervised learning (what we want to extract) instead of proxies, such as reconstruction quality. Even beyond testing properties of the representation (such as independence), it might also be possible that an ensemble can identify models with good representation but poor reconstruction through raising a flag if all ensemble members report equivalent representations but possibly suboptimal reconstruction quality—such as VAEs with a good encoder but a poor decoder (Alemi et al. 2018). However, this requires a unique representation (or a correction for invariances) since generally it is not guaranteed that even the same neural network learns the same representation when trained multiple times.

3.4. Uncertainty estimates

Probabilistic machine learning quantifies the prediction's uncertainty via probability distributions' variance/entropy instead of single-point estimates (Gawlikowski et al. 2021). We acknowledge that these methods improve robustness and argue for their use in the falsificationist framework. Instead of “only” assessing uncertainty, the available information can also restrict the hypothesis space. For example, the variance in regression can describe a feasible interval.

4. Limitations

Although falsificationism can improve the robustness of machine learning by interpreting the predictions differently (perhaps with a change of objective function but the same architecture), it is not the Holy Grail.

We see falsificationism as a means of improving predictions’ reliability near the decision boundary—where multiple options are possible. On the other hand, adversarial examples—samples malevolently modified by exploiting the networks’ computational properties to trick the model into believing that the sample belongs to a different class—will have a negative effect, unless a defense is put into place. This is irrespective of using the falsificationist approach or not. Namely, such examples exploit the model’s properties, and as falsificationism does not change the architecture but mainly the interpretation of the predictions, it has no means to defend against adversarial attacks automatically.

5. Conclusion

This paper draws inspiration from Karl Popper’s falsificationist philosophy and investigates its applicability to the predictions of machine learning models. We contrasted theoretical and practical arguments and reinterpreted existing machine learning methods. Our main conclusion is that—mainly by interpreting the predictions differently—it is generally possible to adapt the falsificationist view.

Moreover, we voiced our support for doing so, as the deployment of machine learning in critical systems requires robustness—we believe that starting to think with the falsificationist mindset could help achieve this goal.

References

Alemi, Alexander, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. “Fixing a broken ELBO.“ In *35th International Conference on Machine Learning vol. 80*, 159–168. Stockholmsmässan, Stockholm Sweden: PMLR, 2018.

Berkson, William, and John Wettersten. “Learning from Error, Karl Popper’s Psychology of Learning. ” Synthese 78, no. 3, 1989.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations.” In *37th International Conference on Machine Learning vol. 119*, 1597–1607. Stockholmsmässan, Stockholm Sweden: PMLR, 2020.

Corfield, David, Bernhard Schölkopf, and Vladimir Vapnik. “Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions.” Journal for General Philosophy of Science 40, 51-58, 2009.

Gawlikowski, Jakob, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. “A survey of uncertainty in deep neural networks.” arXiv preprint arXiv:2107.03342, 2021.
https://doi.org/10.48550/arXiv.2107.03342

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets.” Advances in neural information processing systems 27, 2014.

Gretton, Arthur, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. “A kernel statistical test of independence.” Advances in neural information processing systems 20, 2007.

Kingma, Diederik P., and Max Welling. “Auto-encoding variational bayes.” arXiv preprint arXiv:1312.6114, 2013.
https://doi.org/10.48550/arXiv.1312.6114

Lapin, Maksim, Matthias Hein, and Bernt Schiele. “Loss functions for top-k error: Analysis and insights.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 1468-1477. Las Vegas, NV, USA: CVPR, 2016.

Lil’Log. Weng, Lilian. “Contrastive Representation Learning.” 2021a. Accessed May 31, 2021.
https://lilianweng.github.io/posts/2021-05-31-contrastive/

Lil’Log. Weng, Lilian. “Learning with Not Enough Data Part 1: Semi-Supervised Learning.” 2021b. Accessed December 5, 2021.
https://lilianweng.github.io/posts/2021-12-05-semi-supervised/

Lil’Log. Weng, Lilian. “Self-Supervised Representation Learning.” 2019. Accessed November 10, 2019.
https://lilianweng.github.io/posts/2019-11-10-self-supervised/

Masegosa, Andres. “Learning under model misspecification: Applications to variational and ensemble methods.” In *Advances in Neural Information Processing Systems 33*, 5479–5491. NeurIPS Virtual-Only Conference, 2020.
https://proceedings.neurips.cc/paper/2020

Microsoft. “Confidence Score – QnA Maker – Azure Cognitive Services.” 2021. Accessed January 17, 2022.
https://docs.microsoft.com/en-us/azure/cognitive-services/qnamaker/concepts/confidence-score

Monti, Ricardo Pio, Kun Zhang, and Aapo Hyvärinen. “Causal discovery with general non-linear relationships using non-linear ica.” In *35th Uncertainty in Artificial Intelligence Conference vol. 115*, 186–195. Tel Aviv, Israel: PMLR, 2020.

Mortier, Thomas, Eyke Hüllermeier, Krzysztof Dembczyński, and Willem Waegeman. “Set-valued prediction in hierarchical classification with constrained representation complexity.” arXiv preprint arXiv:2203.06676, 2022.
https://doi.org/10.48550/arXiv.2203.06676

Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. “Self-supervised exploration via disagreement.” In *36th International Conference on Machine Learning vol. 97*, 5062–5071. Long Beach, CA, USA: PMLR, 2019.

Popper, Karl. *The Logic of Scientific Discovery*. London: Routledge, 2002. https://www.routledge.com/The-Logic-of-Scientific-Discovery/Popper/p/book/9780415278447

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019.)

Robinson, Joshua, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. "Contrastive learning with hard negative samples." arXiv preprint arXiv:2010.04592, 2020. https://doi.org/10.48550/arXiv.2010.04592

Schölkopf, Bernhard, Robert C. Williamson, Alex Smola, John Shawe-Taylor, and John Platt. "Support vector method for novelty detection." *Advances in Neural Information Processing Systems*, 1999.

Schott, Lukas, Jonas Rauber, Matthias Bethge, and Wieland Brendel. "Towards the first adversarially robust neural network model on MNIST." arXiv preprint arXiv:1805.09190, 2018. https://doi.org/10.48550/arXiv.1805.09190

Szemenyei, Marton, and Patrik Reizinger. "Attention-based curiosity in multi-agent reinforcement learning environments." In *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, 176–181. Athens, Greece: IEEE, 2019. https://doi.org/10.1109/ICCAIRO47923.2019

Taleb, Nassim Nicholas. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House, 2007.

Taleb, Nassim Nicholas. "Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications." *arXiv preprint arXiv:2001.10488*, 2020. https://doi.org/10.48550/arXiv.2001.10488

Ţifrea, Alexandru, Eric Stavarache, and Fanny Yang. "Novelty detection using ensembles with regularized disagreement." In *ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning*, 2021. http://www.gatsby.ucl.ac.uk/~balaji/udl2021/accepted-papers/UDL2021-paper-100.pdf

Tóth, Tamás, Ákos Hajdu, András Vörös, Zoltán Micskei, and István Majzik. "Theta: a framework for abstraction refinement-based model checking." In 2017 Formal Methods in Computer Aided Design (FMCAD), pp. 176-179. IEEE, 2017.

Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. 11, 2008.

Vasconcelos, Marisa, Carlos Cardonha, and Bernardo Gonçalves. "Modeling epistemological principles for bias mitigation in AI systems: An illustration in hiring decisions." In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* 323-329. New York: Association for Computing Machinery, 2018. https://doi.org/10.1145/3278721.3278751

Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." In International Conference on Machine Learning, pp. 9929-9939. PMLR, 2020.

Yang, Jingkang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. "Generalized out-of-distribution detection: A survey." arXiv preprint arXiv:2110.11334, 2021. https://doi.org/10.48550/arXiv.2110.11334