A Survey on AI Sustainability: Emerging Trends on Learning Algorithms and Research Challenges

Zhenghua Chen [1,2], Min Wu[1], Alvin Chan[3], Xiaoli Li[1,2,3], and Yew-Soon Ong[2,3]

[1]Institute for Infocomm Research, A*STAR, Singapore
[2]Centre for Frontier AI Research, A*STAR, Singapore
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore

***Abstract*—Artificial Intelligence (AI) is a fast-growing research and development (R&D) discipline which is attracting increasing attention because it promises to bring vast benefits for consumers and businesses, with considerable benefits promised in productivity growth and innovation. To date, significant accomplishments have been reported in many areas that have been deemed challenging for machines, ranging from computer vision, natural language processing, audio analysis to smart sensing and many others. The technology trend in realizing success has developed towards increasingly complex and large-size AI models to solve more complex problems at superior performance and robustness. This rapid progress, however, has taken place at the expense of substantial environmental costs and resources. In addition, debates on the societal impacts of AI, such as fairness, safety, and privacy, have continued to grow in intensity. These issues have reflected major concerns pertaining to the sustainable development of AI. In this work, major trends in machine learning approaches that can address the sustainability problem of AI have been reviewed. Specifically, the emerging AI methodologies and algorithms are examined for addressing the sustainability issue of AI in two major aspects, *i.e.*, environmental sustainability and social sustainability of AI. Then, the major limitations of the existing studies are highlighted, and potential research challenges and directions are proposed for the development of the next generation of sustainable AI techniques. It is believed that this technical review can help promote a sustainable development of AI R&D activities for the research community.**

***Index Terms*—Artificial Intelligence, Environmental Sustainability of AI, Social Sustainability of AI**

I. Introduction

Artificial Intelligence (AI) is intelligence demonstrated by computers or machines as opposed to the natural intelligence presented by humans[1]. It is capable of performing tasks that conventionally require human intelligence, and even surpasses human capabilities in many challenging tasks that involve a large amount of data to analyze and learn from. In recent years, it has achieved remarkable success in many areas, such as computer vision [1], natural language processing [2], recommendation systems [3], intelligent sensing [4], condition monitoring systems [5], advanced Web search engines [6], speech recognition [7], automated decision-making [8], and competing at the highest level in strategic game systems [9], etc. AI plays an important role in enhancing human work productivity and efficiency, reducing operational costs in many time-consuming and labour-intensive applications across different industry verticals and government agencies. Notable examples of AI applications include assisting doctors in diagnosis, developing self-driving cars, monitoring the operating conditions of expensive equipment, recommending suitable products at the right time and location, and facilitating human communication and interaction via machine translation tools. From Fig. 1, it can observe an increasing trend of AI in not only the traditional discipline of computer science, but also other novel application domains, including material science, telecommunications, physics, environmental science, and others. In this light, it is no doubt that AI has significant impacts on and changes even traditional fields of studies and industries. All these will bring about a new era of interactions and augmentations between AI and human activities.

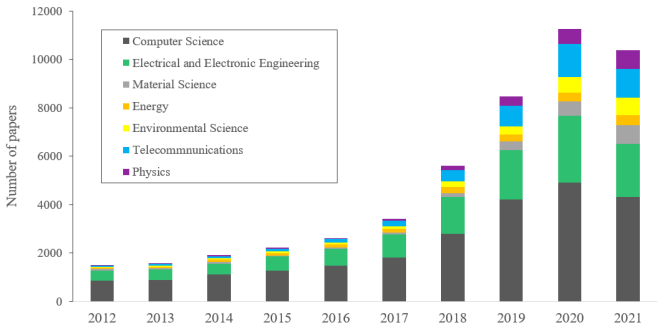

Fig. 1. The trend of AI-related papers published in different disciplines for the past 10 years. Data are obtained from Web of Science on Nov 22, 2021.

Apart from the technical achievements of AI, there have been increasing concerns about the sustainable development of AI R&D activities [10], [11]. For instance, researchers have reported that the carbon footprint of training a Transformer model is as high as 626,155 pounds of CO2 emission [12]. Other concerns in AI ethics include fairness and privacy, which could significantly impede its progress and usage in safety-critical real-world applications if not appropriately taken into consideration [13]. In order to support continual progress and development of AI, the sustainability challenges must be taken into serious consideration and more efforts should be dedicated to this critical field.

In order to clearly define what is sustainable AI, the

Corresponding author: Zhenghua Chen (Email: chen0832@e.ntu.edu.sg)
[1]https://en.wikipedia.org/wiki/Artificial_intelligence

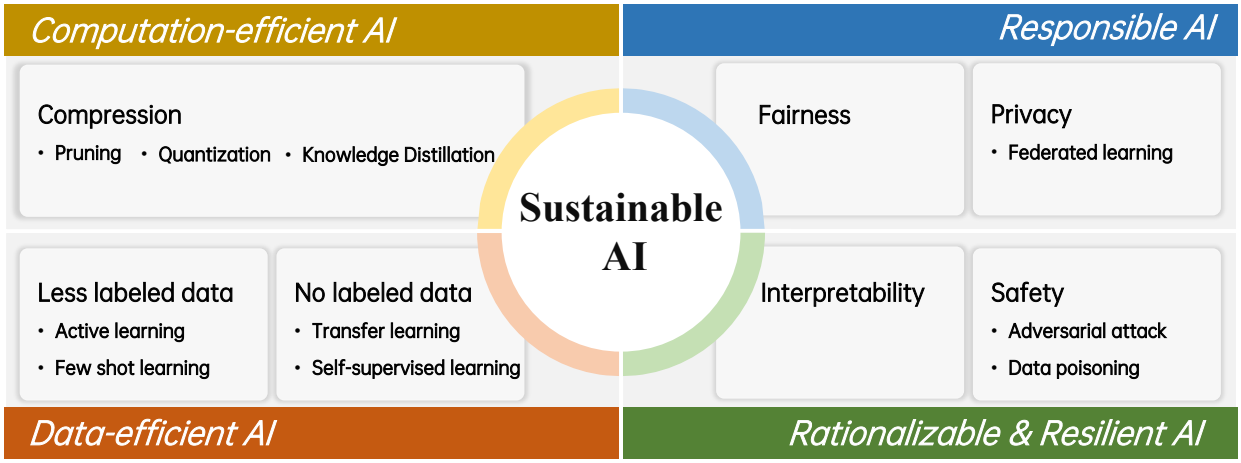

Fig. 2. Overview of sustainable AI.

definition of 'sustainable development' is first investigated. Mensah explained the term 'sustainable development' as "the entire issue of sustainable development centres around inter- and intragenerational equity anchored essentially on three-dimensional distinct but interconnected pillars, namely the environment, economy, and society" [14]. The sustainable development of AI should also consider these three key factors: environment, economy, and society. As discussed above, AI has significantly improved our productivity and efficiency, and leads to great economic impacts. Despite of its achievements in economy, it recently faces great risks in environmental and societal issues, for example, the significant carbon emissions and AI ethics of fairness and privacy. Thus, sustainable AI could be defined as developing the innovation of AI to promote economic development, while simultaneously addressing its negative impacts on the environment and ethical concerns to the society, which refer to environmental sustainability and social sustainability of AI, respectively. Under this definition, various sustainability challenges must be addressed.

In this survey, our aim is to review the sustainability issue of AI from two major aspects, namely, examining the recent and emerging technical research progresses made in the environmental sustainability, and secondly, focusing on the social sustainability of AI. In particular, we consider the following sustainability challenges in terms of the two aforementioned aspects while conducting the review of this work.

- Significant carbon emissions for AI model training and inferences.
- Huge efforts on data collection and annotation.
- Fairness and privacy issues of AI models
- Understanding the decision making process of AI models
- Ensuring the safety of AI in terms of adversarial attack and data poisoning.

To date, several surveys that focused on sustainable AI [10], [11], [15], [16] have been reported. Sustainable AI was defined in two perspectives [11], *i.e.*, AI for sustainability (*e.g.*, using AI to address climate change) and the sustainability of AI (*e.g.*, reducing carbon emissions from AI models). In addition to the concepts above, the author in [11] also discussed possible high-level strategies for sustainable AI. Larsson *et al.* discussed sustainable AI in terms of the ethical, social and legal challenges of AI [10]. They also summarized some recommendations for the development of sustainable AI, including AI governance, multi-disciplinary collaboration, transparency and accountability of AI. Kindylidi *et al.* presented sustainable AI from the standpoint of customer protection, where the main focus has been on AI-related policies [15]. It is worth noting that the existing works have focused on the strategic views for sustainable AI. None of the reviews to date, on the other hand, has considered the technical and algorithmic progresses towards achieving AI sustainability. Considering that more and more attention has been devoted to these key technologies for achieving sustainable AI, this technical review attempts to fill this gap by reviewing the emerging AI technologies (as shown in Fig. 2), highlighting how they contribute to the sustainable development of AI, discussing key challenges ahead, and presenting potential future challenges and notable research directions worthy of investing time in.

The rest of the paper is organized as follows. The detailed technical reviews from two perspectives of environmental sustainability and social sustainability of AI are first introduced in Section II and III, respectively. Then, detailed discussions and analyses are provided in Section IV, and the future research challenges and directions are presented in Section V. Finally, the conclusions of this paper are made in Section VI.

II. Environmental Sustainability of AI

With increasing size and complexity of AI models, one major concern is their significant carbon footprint. For example, a well-known study by [11] reported that the process of training a GPT-3 model, *i.e.*, a natural language processing model, can lead to 1,212,172 pounds of CO2 emissions. This is roughly equivalent to the CO2 emissions produced by ten cars over their entire lifespans. Fig. 3 shows a clear trend of substantial CO2 carbon footprint of advanced AI models in comparison with daily CO2 emissions of human activities. The huge models, like Transformer and GPT-3, are also called Red AI that pursues better performance (*e.g.*,

accuracy) while ignoring their huge carbon footprint [17]. An alternative concept is Green AI, which considers efficiency as a primary evaluation metric alongside model performance [17]. The efficiency of AI models in Green AI research is directly related to environmental impacts. Thus, the research efforts of Green AI also reveal the critical need to take environmental impacts into major consideration when developing AI.

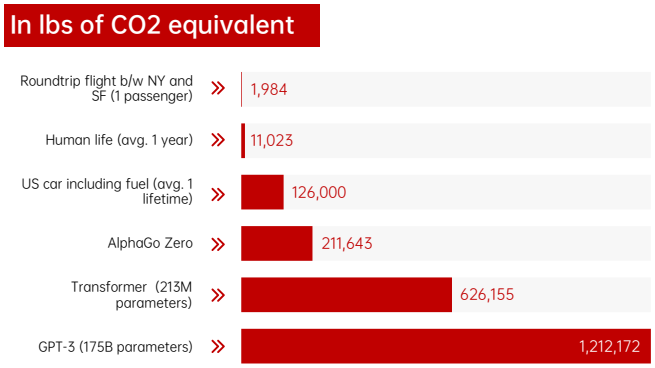

Fig. 3. CO2 carbon footprint benchmarks. Source data from [11], [18], [19].

This section intends to address environmental sustainability impacts on AI by reviewing related AI techniques, gaining deep insights into their function, and analyzing their potential contributions and limitations to environmental sustainability. Due to the limited space for the manuscript, only major and emerging AI algorithms are covered in the two core categories, namely, computation-efficient AI and data-efficient AI, which can address the sustainability challenges of significant carbon emissions for model training and inference, and huge efforts for data collection and annotation. Note that both basic and advanced papers for a specific AI technique are searched and reviewed, but excluding papers that violate the initial objective (*e.g.*, reducing carbon emissions).

A. Computation-efficient AI

AI models have become more and more powerful along with the dramatic increase in their model sizes. According to Fig. 4, the computation requirements of AI models will double every 3.4 months[2], which is much faster than Moore's law claiming to double every two years. The substantial size of AI models has a noteworthy impact on environmental sustainability, as more powerful platforms are required for their deployment and more energy will be consumed for real-time implementation.

To reduce the environmental impact of AI, an effective way is to develop computation-efficient AI that requires less computational resources and at the same time generates smaller yet accurate models [22]. This has generally been established as compressed AI models. The compressed models could significantly reduce storage memory and energy for deployment, hence directly leading to lower carbon footprint. Many advanced techniques have been developed to build compressed AI models which can be divided into three categories, *i.e.*, pruning, quantization, and knowledge distillation. The details of each category are discussed in the following subsections.

[2]https://openai.com/blog/ai-and-compute/

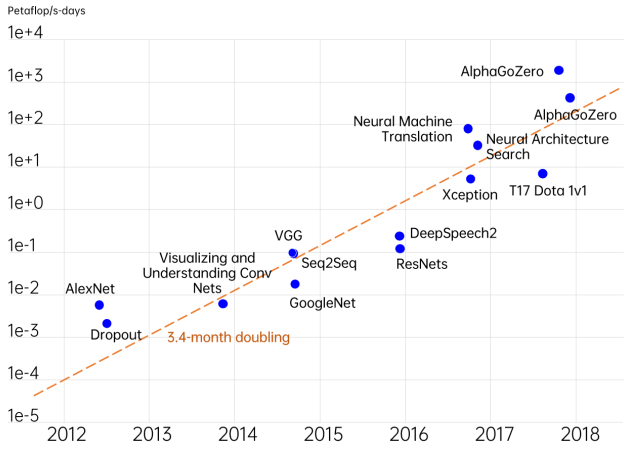

Fig. 4. The computing requirements increased by 300,000 times from AlexNet [20] to AlphaGo Zero [21].

1) Pruning: With the consensus that AI models are typically over-parameterized [23]–[25], network pruning serves to remove redundant weights or neurons that have the least impact on accuracy, resulting in significant reduction of storage memory and computational cost. It is one of the most effective methods to compress AI models. Fig. 5 shows the process of pruning for a typical neural network based AI model.

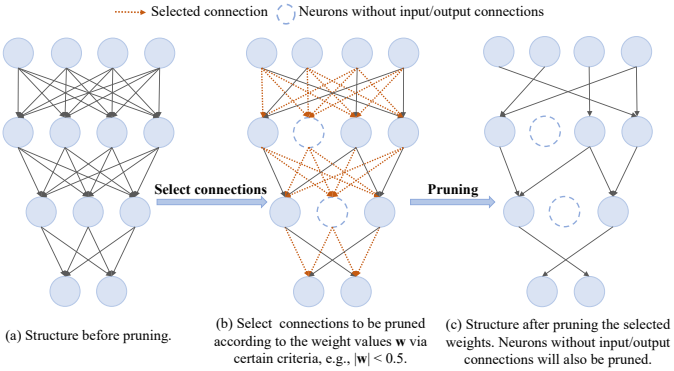

Fig. 5. The process of weight/neuron pruning. The first step is to determine which connections to be pruned. Taking the magnitude-based pruning [26] as an example, the connection whose weight amplitude $|\mathbf{w}|$ is lower than a preset threshold will be pruned (red dotted arrow). If all the input or output connections of a neuron are pruned, then the neuron can also be removed entirely (blue dotted circle). When a connection is pruned, the corresponding value in the weight matrix will be set to zero.

Pruning criteria are essential for all pruning approaches. One of the simplest heuristics is to estimate importance in terms of the absolute value of a weight, called magnitude-based pruning. Zero-valued weights or weights within certain threshold are removed in [26], [27]. Moreover, to make the weights zero, regularization term is often adopted during training to increase weight sparsity [28]. However, these element-wise magnitude-based pruning methods may result in unstructured network organizations which are difficult to be compressed or accelerated without specialized hardware support [29].

Structured pruning approaches, on the other hand, alleviate the above issue by pruning the network on the filter or layer

level. For instance, Wen *et al.* proposed a Structured Sparsity Learning (SSL) approach to regularize network structures (*i.e.*, filters, channels, and layers). Group Lasso is then used to zero out the grouped weights [30]. In addition to the weights, some other indicators have also been explored for pruning. Liu *et al.* considered the scaling factor from batch normalization layer and pruned those channels of small scaling factor values [31]. Hu *et al.* argued that certain activation functions such as Rectified Linear Unit (ReLU) could generate numerous zeros, and these zero activation neurons are redundant. Hence, they can be removed without affecting the overall performance of the network [32]. Moreover, some other researchers evaluated the importance (or saliency) of a neuron via its impact on the objective function. For instance, in Optimal Brain Damage [23] and Optimal Brain Surgeon [33], the second derivative (Hessian matrix) of the objective function with respect to the parameters is used to determine the redundant weights. Similarly, Lee *et al.* introduced a saliency criterion based on the normalized magnitude of the derivatives in the objective function with respect to each weight [34].

With advanced pruning techniques, it is able to significantly reduce network size, which leads to reduced memory requirements and less computational cost in inference during model deployment stage [35]. However, some pruning techniques may result in more training efforts during model building stage, such as methods that are based on the procedures of training, pruning, and fine-tuning/re-training [36], [37]. A more efficient way is to perform pruning at initialization [38] or using sparse training, *i.e.*, training under sparsity constraints [39].

2) Quantization: Another approach for computation-efficient AI is via quantization, where 32-bit floating-point parameters are quantized into lower numerical precision of 16-bit [40], 8-bit [41] or even 1-bit [42], [43]. Using low-bit numerical representations not only reduces model storage cost but also results in significant speed-up during inference, as the most time-consuming process, floating point Multiply-Accumulate (MAC) operations, could be avoided. Jacob *et al.* quantized both weights and activations into 8-bit integers for integer-arithmetic-only hardware [44]. Incremental Network Quantization (INQ) proposed in [45] converted full-precision weights into power-of-two values with lower precision. It allows the replacement of multiplication operations with bit-shift operations which are more efficient for hardware like Field-Programmable Gate Array (FPGA).

Although quantization is generally applied to weights and activations, it can also be extended to gradient computations. DoReFa-Net [46] quantized gradients to numbers with bit-width of less than 8 before back-propagation. It allows quantization of the network during training or fine-tuning. Instead of quantizing a single weight, vector quantization approaches focused on factorizing the weight matrix by weight clustering and sharing [47], [48].

Instead of only using pruning or quantization for network compression, a combination of these two have generally been reported to lead to higher compression rate. Han *et al.* presented a three-stage pipeline which consists of pruning, weight sharing by vector quantization, and Huffman encoding to compress a pre-trained model, and it achieved state-of-the-art performance at that time [49]. Both pruning and quantization are effective ways to reduce model size for efficient learning. One limitation is that they can only deal with models that share a common network architecture, which means compressing a cumbersome network to an efficient one with less weights (or neurons) and/or lower-precision. They cannot be used across heterogeneous networks, which can otherwise be addressed with the knowledge distillation technique.

3) Knowledge Distillation: Knowledge distillation (KD) is a flexible yet efficient approach to compress AI models. It aims to transfer the knowledge learnt from a cumbersome network to a compact one, which is also known as 'Teacher-Student' learning framework [50]. With the knowledge from a teacher, a lightweight student network can achieve competitive performance as the cumbersome (teacher) network yet at a higher computational-efficiency. The success of KD relies mainly on two key factors: the knowledge and distilling schemes.

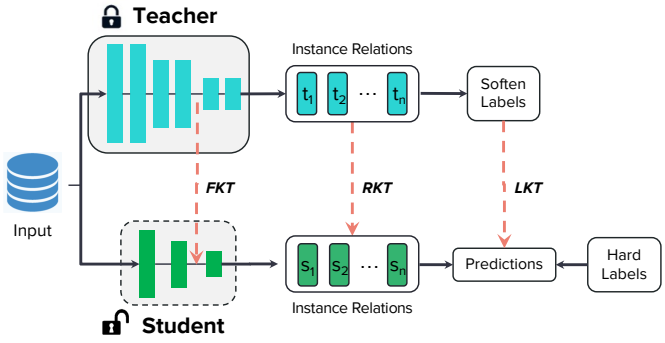

Fig. 6. An illustration of knowledge distillation. FKT refers to Feature-based Knowledge Transfer, RKT refers to Relation-based Knowledge Transfer, and LKT refers to Logits-based Knowledge Transfer.

The knowledge information can be categorized into three different types: logits-based, feature-based, and relation-based knowledge [51], which are depicted in Fig. 6. The vanilla KD approach leverages softened logits from a cumbersome teacher as the knowledge to guide the training process of a student [52]. Romero *et al.* extended it by introducing the 'hint' knowledge from intermediate layers of a teacher network, which can be termed feature-based knowledge [53]. Instead of directly minimizing the discrepancy of feature maps between the teacher and the student, other advanced feature-based knowledge has also been explored. For instance, Zagoruyko and Komodakis regarded the teacher's attention maps derived from feature maps as the knowledge to be transferred [54]. Different from the aforementioned two knowledge types, relation-based KD methods emphasize the transfer of intra-relationships between data samples [55]–[57] or correlations between feature maps from multiple intermediate layers [58], [59].

Along with this knowledge information, various distilling schemes have also been proposed. Although most KD methods follow the conventional single-teacher single-student distilling scheme, some other promising schemes have also been explored in the literature. The scheme of distilling informative

knowledge from an ensemble of teachers instead of a single teacher could better enhance the student's generalization ability [60], [61]. Apart from distilling knowledge from large-scale networks, another research direction intends to boost the performance of a compact network by transferring the knowledge between networks which share identical architecture (also termed as self-knowledge distillation) [62]. Furlanello *et al.* proposed a sequential distilling scheme called Born-Again Networks (BANs), which takes the model in earlier generation as the teacher and trains a new initialized identical model [62], [63]. To better align feature representations between high-capacity and low-capacity networks, adversarial and contrastive distilling schemes are often adopted. To demonstrate the feasibility of transferring knowledge between two disparate network architectures, Xu *et al.* exploited adversarial distilling scheme to transfer knowledge from a much bigger LSTM-based teacher to a smaller CNN-based student [64].

Generally, KD methods are more flexible. The teacher and student networks in KD can be homogeneous or heterogeneous architectures, which is different from pruning and quantization. However, in most cases, KD methods require more training efforts, as they need to train both the teacher and student networks. In fact, most of the training efforts have been given to the teacher network which is often large in size and complex. To solve this issue, one way is to consider adopting a trained network as the teacher [65]. Building a shared community for sharing trained AI models can significantly reduce repeated model training efforts, which can be an interesting direction for sustainable AI.

Computation-efficient AI can lead to compressed AI models which improve environmental sustainability of AI in two main aspects. Compressed models can be deployed on resource-limited devices (*e.g.*, embedded systems) with less memory and computational power. This also contributes to the feasibility of AI for various real-world applications. On the other hand, they are light-weighted and thus more efficient in real-time implementation, which can dramatically save energy and reduce carbon footprint. Li *et al.* presented some substantiated examples for different model compression methods [66]. Selected results are shown in Table I. Note that as the carbon footprint is linearly proportional to energy cost as stated in [19], the authors choose to measure train and inference energy costs in the paper. It can be found that all the three compression methods can significantly reduce the inference energy cost (IEC). However, pruning and KD require much more energy for model training. Quantization does not consume additional energy for model training, but results in a less compression effect (higher IEC compared to pruning and KD).

Based on Table I, although computation-efficient AI has clear merits towards sustainable AI, it has the drawback of increasing training effort. For example, the training of cumbersome teacher in knowledge distillation can be tedious. To solve this problem, a possible way is to build a shared community, where researchers can upload and share their trained powerful but cumbersome models. Then, the tedious training process of teacher models can be avoided or at the very least not repeated. Nevertheless, compressed AI models are particularly useful in applications, where they will be run for routine and long-term tasks, such as in condition monitoring where the system performs real-time equipment health condition monitoring for a few years [64]. In this case, the price that we pay to turn big models into smaller yet accurate models via tedious training process will be worthwhile in the long run. In summary, computation-efficient AI can contribute to environmental sustainability of AI with possible but addressable side effects by reducing the tedious training process, which requires more attention and continuous research efforts.

TABLE I
SELECTED RESULTS FOR DIFFERENT MODEL COMPRESSION METHODS IN TERMS OF TRAIN ENERGY COST (TEC), ACCURACY (ACC.) AND INFERENCE ENERGY COST (IEC) [66]. BASELINE REFERS TO MODEL WITHOUT COMPRESSION. ACCURACY IS MEASURED ON *CIFAR100* AND ACCURACY TOLERANCE τ IS SET TO 2. THE UNIT FOR TEC AND IEC IS $\mathbf{W} \cdot \mathbf{h}$. T REFERS TO THE TEACHER MODEL AND S REFERS TO THE STUDENT MODEL IN KD.

Baseline	Model		TEC	Acc.	IEC
	VGG16		138.59	73.37	8.93E-07
Pruning	VGG16		158.36	71.09	5.67E-07
Quantization	VGG16		138.59	73.23	5.86E-07
KD	VGG13 (T)	VGG8 (S)	181.61	72.98	5.39E-07

B. Data-efficient AI

The other class of technology for environmental sustainability is to reduce the huge burden on data collection and annotation which require enormous human efforts and resources, as AI models generally require huge labeled datasets for model training. Data-efficient AI addresses sustainable development of AI models by reducing human's manual efforts and resources. ImageNet [67], for example, is said to have taken two and a half years of effort to be collected and annotated. However, most real-world applications would not have such luxuries, especially if a target use case is niche and one is unable to use the existing datasets. Compounding the problem is that the size of datasets directly affects training time and carbon footprint. In this subsection, advanced approaches that can help reduce the data dependency, especially the size of labeled data, are reviewed. These methodologies can be divided into two categories, *i.e.*, using less labeled data and no labeled data.

1) Less Labeled Data: AI is a typical data-driven method and its performance correlates highly with sufficient amount of labeled data for model training. This huge burden on data collection and annotation will significantly impact the environmental sustainability of AI. To use less labeled data, some progressive AI technologies have been developed, including active learning and few-shot learning.

a) Active Learning: To use less labeled data for model training, active learning intends to find representative samples (*e.g.*, a small portion of the entire dataset) to be annotated by the oracle (*e.g.*, human annotator), such that a good supervised learning model can be trained with intelligently selected annotated samples [68]. Generally, active learning techniques

can be divided into three different groups of membership query synthesis, stream-based selective sampling, and pool-based sampling, as shown in Fig. 7.

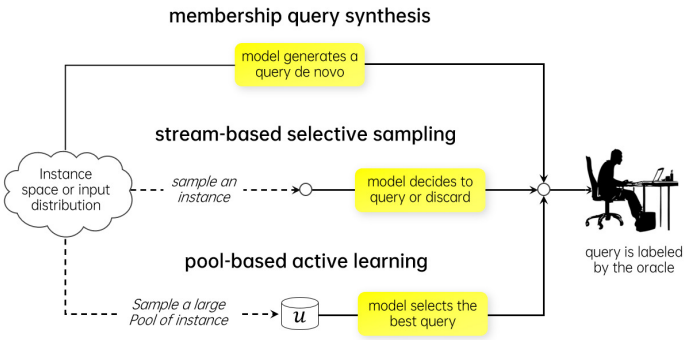

Fig. 7. The procedure for active learning [69].

For the membership query synthesis method, learners can select any unlabeled samples from instance space or generated samples by the learners for annotation [70]. It is particularly effective for small datasets. However, the generated samples may be difficult for the oracle to annotate, which limits its feasibility in many real-world applications. Stream-based selective sampling sequentially sends a data sample to an active learner, which will decide whether to annotate or discard the sample [71]. On the other hand, pool-based sampling approach aims to select samples for annotation from a pool of unlabeled samples. Hence, it is able to select more than one sample at a time, which can be more efficient.

Currently, pool-based sampling is the most widely used technique in active learning. Its key component is the sample selection strategy. The basic idea of selection strategies is to select samples that contain rich information and are highly representative and diverse [68]. A popular selection strategy is based on uncertainty sampling [72], which measures the uncertainty of samples based on techniques like least confidence and entropy [73]. Density based methods intend to query consecutive points from the instance pool to construct a core set for representing the distribution of feature space of original samples [74]. Active learning can significantly reduce the efforts for data annotation as it only annotates representative and hard/ambiguous/uncertain examples. Besides, the training can be more efficient as the selected samples represent a small portion of the entire dataset.

b) Few-shot Learning: Few-shot learning is another class of approaches that aims to learn from less labeled data for training AI models [75]. To solve few-shot learning problems, there are two main groups of methodologies. The first one is based on meta-learning, also known as learning to learn [76]. Meta-learning techniques generally consist of a teacher model and a student model, where the teacher learns how to optimize the student while the student learns how to perform downstream tasks. This can be seen in [77] where the outputs of the teacher models were used to train the student model by optimizing and selecting the parameters from a higher dimensional space due to the limited training samples. Generally, initial meta-learners are biased towards the existing tasks in parameter optimization, but Jamal and Qi [78] demonstrated the possibility of using an agnostic meta-learner to overcome those limitations. Another promising work is presented by Hou *et al.* [79], where attention modules for meta-learners were used to achieve the task-specific initialization of base models for fast adaptation in meta-learning.

Another popular approach for few-shot learning is metric learning. The general direction is to compare the embeddings of the support (few-shot training samples) and query (test samples) sets. Examples of such work include the Siamese [80] and Triplet [81] networks, which compared two embeddings at one time, and the Matching [82], Prototypical [83], and Relation [84] networks, which compared across multiple embeddings from support and query sets. A distinguishing feature across metric learning approaches is the distance measures for embeddings. The Triplet, Matching, and Prototypical networks adopt pre-defined distance measures, whereas the Siamese and Relation networks learn the distance measures by using neural networks.

Similar to active learning, few-shot learning can significantly reduce the amount of labeled data for model training, which saves a lot of resources required for data collection and annotation, thus reducing carbon footprint in these tedious processes. Substantiated examples are shown in Table II, where active learning and few-shot learning are tested on various benchmark datasets with varying percentages of labeled data for model training [84]–[87]. It can be found that these data-efficient learning algorithms can achieve good performance with a small portion of the labeled data, which significantly reduces the burden on data collection and annotation. However, few-shot learning may not be able to reduce training effort if learning with complex algorithms, like meta-learning which requires additional training efforts, or metric learning where comparing high-dimensional embeddings may be time-consuming. These negative impacts must be considered and evaluated under the perspective of environmental sustainability.

TABLE II
RESULTS OF SELECTED DATA-EFFICIENT LEARNING ALGORITHMS IN TERMS OF THE PERCENTAGE OF THE LABELED DATA FOR MODEL TRAINING AND TESTING ACCURACY (ACC.) ON VARIOUS DATASETS.

Active Learning	MNIST [85]		PASCAL VOC [86]	
	Percentage	Acc.	Percentage	Acc.
	3.4%	95.0%	20%	72.3%
Few-shot Learning	CUB [87]		Omniglot [84]	
	Percentage	Acc.	Percentage	Acc.
	4.2%	92.9%	25%	99.8%

2) No Labeled Data: More advanced AI techniques intend to release the burden of data annotation by learning AI models via data from related tasks or unlabeled data. Two typical techniques, *i.e.*, transfer learning and self-supervised learning, are introduced, and their contributions to environmental sustainability are analyzed.

a) Transfer Learning: Transfer learning aims to solve a given task (denoted as target domain) with labeled data from related ones (denoted as source domain) [88], as shown in Fig. 8. Note that generally the target domain only contains

unlabeled data. The main idea in transfer learning is to minimize the domain difference between the source and target domains. There are two types of approaches to minimize domain difference. One is distance-based methods, where a distance function, such as MMD [89], [90], CORAL [91], etc., is defined to measure the domain difference, and it will be minimized during training, such that the gap between the source and target domains is minimal [92]. Then, the source classifier/regressor trained using the labeled source domain data can be used for the classification/regression task in the target domain. The other one is adversarial-based methods, which are inspired by generative adversarial network (GAN) [93]–[95]. They design a discriminator to distinguish features generated by the feature encoders of source and target domains. At the same time, the feature encoders of the source and target domains will be trained to push the features from both domains to be similar such that the discriminator cannot separate them [96].

Transfer learning which assumes the availability of a labeled source domain (related to target) is a common setting in real applications. For example, when performing an image classification task with no available labeled data, an effective way is to use transfer learning with public datasets, like ImageNet [67]. However, it still requires the collection of unlabeled data in the target domain for the adaptation. Besides, the training effort will also increase as the model will be trained with both source and target domain data. In this case, there will be a trade-off between the efforts of data annotation and model training. If annotation efforts in terms of carbon footprint can be quantified, it may help us better evaluate on whether transfer learning can actually contribute to environmental sustainability.

Note that there is a special case of transfer learning, named fine-tuning [97], which assumes the availability of few labeled data in the target domain. Then, fine-tuning aims to train just few layers of the learnt network from the source domain (also known as pre-trained model) by using few labeled samples in the target domain. Suppose that the pre-trained model is only obtained from a shared community without any training efforts, fine-tuning can significantly contribute to environmental sustainability as it does not require large datasets (either labeled or unlabeled) and can dramatically reduce training costs (only parameters from few layers will be trained). Here, it is emphasized again the importance of creating a shared community for trained AI models towards the development of sustainable AI.

b) Self-supervised Learning: Another popular technique that does not use labeled data is self-supervised learning, which is able to learn representations directly from unlabeled samples. It consists of two main approaches to achieve this objective. The first one is contrastive learning, which learns features by identifying similar data samples and dissimilar ones. SimCLR [98] presented a typical contrastive learning framework. It learns representations by maximizing the similarity of two different augmented views of one sample and minimizing the similarity of augmented views of different samples. It has shown that the number of negative pairs will significantly affect the performance of representation learning.

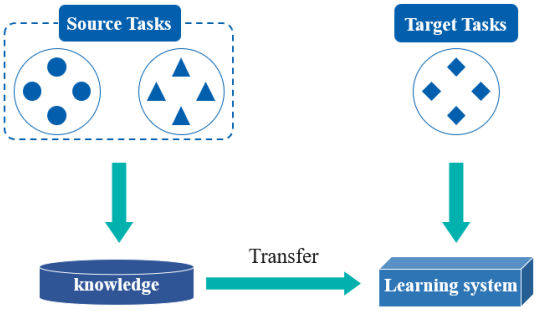

Fig. 8. Learning process of transfer learning [88].

However, a large number of negative pairs require huge memory. To address this issue, MoCo [99] designed a memory bank to store and reuse representations of data. MoCo-V2 [100] is an updated version of MoCo, which also borrowed the ideas of project head and augmentation from SimCLR. BYOL [101] claimed a new state-of-the-art performance for representation learning without using any negative samples. It is composed of two networks, *i.e.*, online and target networks, that interact and learn from each other. SimSiam [102] simplified the architecture of BYOL and developed a novel method of stop gradient to prevent it from collapsing during learning, which has been proven effective in representation learning.

Another widely used approach for self-supervised learning is based on pretext tasks. The objective is to design pretext tasks as supervised signals for learning representative features. For instance, a popular way is to predict the relative positions of image patches [103], which will promote the model to recognize objects and patches, such that good feature representations can be learnt. Alternatively, it can predict the rotation degrees of images, which define four degrees for prediction, *i.e.*, 0, 90, 180, and 270 [104]. Noroozi and Favaro proposed a pretext task by solving jigsaw puzzles of images [105]. In this challenging task, the model tends to learn a feature mapping of objects in images and their correct spatial relations. Other visual pretext tasks include colorization [106], image inpainting [107], perturbations [108], etc. In addition to visual applications, many pretext tasks have been proposed in natural language processing (NLP), such as predicting center words or neighbor words in a sentence [109]. The powerful BERT [2] has designed several pretext tasks. For example, it masks a word in a sentence and predicts it. Other pretext tasks include next sentence prediction, sentence permutation, document rotation, etc.

Self-supervised learning can extract useful information from the data by contrasting between samples with their augmentations or defining pretext tasks as supervised signals. It totally releases the effort of data annotation. However, the creation of augmented samples or pretext tasks can lead to additional training burden, which will consume more computational power. Due to the lack of ground truth labels, self-supervised learning generally requires more training iterations to converge. Overall, the current self-supervised learning avoids the cost and efforts in data annotation but dramatically increases

the training burden. Although self-supervised learning is a very promising research area in AI, it requires further research to reduce its training effort for sustainable development of AI.

III. Social Sustainability of AI

In addition to the environmental impact of AI, another key issue that requires a lot of attention is its social impact. Here, social sustainability of AI means issues pertaining to users and interactions between users and AI models.

This section intends to address social sustainability of AI by revealing on the social issues of existing AI algorithms, reviewing specific AI techniques that can potentially address these social issues, analyzing their progresses and limitations. Due to limited space, only major and emerging AI algorithms are covered in the two technical aspects of social sustainability of AI, namely, Responsible AI and Rationalizable & Resilient AI, that can address the sustainability challenges of fairness, data privacy, interpretability, and safety of AI. Specifically, in responsible AI, the key issues of AI fairness across different users and user data privacy will be explored. In rationalizable & resilient AI, the interpretability and safety of AI models are discussed, which are crucial in cultivating greater acceptance of AI models [110]. Here, basic papers for the definition of key concepts and recently developed papers to show the progress in addressing the social sustainability of AI are searched. As each topic may contain a large number of papers, only the representative ones are selected and discussed from the perspective of social impacts.

A. Responsible AI

It can be foreseen that AI systems will soon play a key role in various aspects of human life. As such, AI ethics must be considered when designing AI systems. In this technical review, the concept of responsible AI is defined as a responsibility to users from the perspective of two major aspects of fairness and user data privacy. The fairness of AI is to eliminate the biases in AI models towards different users (*e.g.*, with different gender, race, religion, etc.), where many research efforts have been directed to this critical area. Another key aspect is the data privacy of users. Without careful and secure protection of data privacy, users will not be willing to share their data for the training of AI models. Here, emerging AI techniques are reviewed for addressing AI fairness and privacy issues in the following paragraphs.

1) Fairness: As various AI models are continually making great achievements in sociotechnical systems, they may suffer from unexpected social implications, including social bias towards race, gender, poverty, disabilities, etc. A classic example is a software system named COMPAS used by American courts to judge the risk of an offender recommitting another crime [111]. In this system, black people are always assigned with higher risk than other groups, even though other inputs are similar. This is due to the imbalance issue in the training data, i.e., blacks and whites have different base rates of recidivism. Such biases or discriminations can also be found in facial recognition systems [112] and recommendation systems [113]. In some facial recognition systems, Black and Asian faces are falsely identified 10 to 100 times more often than White faces. Similarly, such bias is also due to the insufficient image data for minorities (i.e., Blacks and Asians) to train the facial recognition system. Nevertheless, it is crucial to promote AI fairness to mitigate these biases and discrimination issues for social harmony and justice.

Unfairness generated by AI models usually arises from two types of biases, namely, biases in data and the algorithmic biases [114]. The most common bias in data is that the data contains sensitive features (also called protected attributes), *e.g.*, race and gender. The use of such features in AI models will lead to direct or indirect discrimination in some specific tasks. There are some other biases in data, such as conformity bias and popularity bias, in recommendation systems [115]. Conformity bias refers to that the users tend to behave similarly to their friends, and such behaviors do not always reflect their true preferences. Popularity bias means that popular items tend to be exposed more. Popular items would thus be recommended more frequently, and less popular items tend to get limited attention. Different from data bias, algorithmic bias is induced purely by AI algorithms. Various AI algorithms have their own assumptions to better learn the models and generalize to unseen data. Such assumptions, denoted as inductive bias, usually help the models achieve desirable results. In addition, different optimization functions and regularization terms can be selected for model training, which can also lead to biased prediction results. Do refer to [114], [115] for more details about different types of biases.

To mitigate the above biases, various metrics to quantify the fairness have been developed in recent years [114], [116]. Some of the most widely used definitions of fairness include demographic parity [117], equal opportunity [118], and fairness through awareness/unawareness [119]. Assume that A is a sensitive feature, Y is the class label, $\hat{Y}$ is the predicted label, and $P(\hat{Y})$ is the probability of $\hat{Y}$. Here, both A and Y are binary. Demographic parity requires the prediction to be independent of the sensitive feature, *i.e.*, $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$. Equal opportunity means that the positive prediction is independent of the sensitive feature, given that the inputs are actual positives, *i.e.*, $P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1)$. As for fairness through awareness, it refers to that an algorithm is fair if it gives similar predictions to similar individuals. Fairness through unawareness means that an algorithm is fair as long as any sensitive feature A is not explicitly used for prediction. For different types of fairness, please refer to [114], [116], [119] for more details. In addition, fairness definitions can fall under two categories, namely, individual fairness and group fairness. Individual fairness requires the model to generate similar predictions for similar individuals, while group fairness requires the model to treat different groups equally. Therefore, demographic parity and equal opportunity are group fairness, while fairness through awareness/unawareness is individual fairness.

To achieve AI fairness, various approaches have been proposed [114], [116], [120]. These approaches can be categorized as pre-processing, in-processing, and post-processing methods as shown in Fig. 9. In particular, pre-processing

approaches focus on transforming the data to remove biases and discriminations. Common pre-processing approaches include variable blinding [121], [122], relabelling [123], and data reweighing [121]. Moreover, adversarial learning is also applied to generate high-quality fair training data [124]. In-processing methods include fairness constraints during model training and maximize both performance (*e.g.*, accuracy) and fairness. For example, fairness constraints are considered when optimizing the accuracy for classification [125] and regression tasks [126]. Finally, post-processing methods aim to improve the model fairness by applying transformations (*e.g.*, thresholding [127]) to the model outputs.

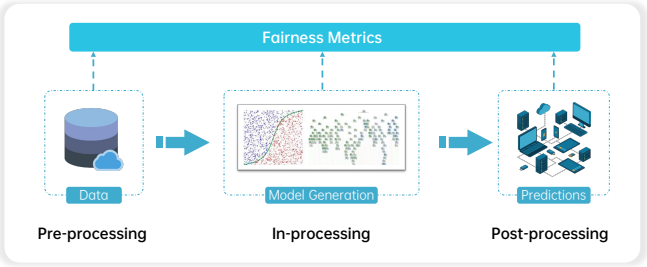

Fig. 9. Three main categories of approaches to improve fairness.

AI fairness, covering both fairness metrics and fair AI models, is a very active research topic. However, there are still many open issues and future opportunities in this area. First, as fairness is a multi-faceted concept, various fairness metrics have been proposed [128], and some of them may be incompatible with others. To address this incompatibility issue, it would be necessary to choose suitable fairness metrics based on the context and characteristics of current AI applications. Second, when improving fairness, other aspects of AI models may be affected, *i.e.*, fairness-utility trade-off. It is thus important to achieve a balance between fairness and utility factors (*e.g.*, accuracy [129], and privacy [130]).

2) Privacy: Another key aspect of responsible AI is data privacy. Generally, the training of AI models requires a large dataset which may be held by different parties/organizations. Data sharing across parties/organizations will lead to privacy concerns which would limit the cooperation for sustainable development of AI. To address this issue, federated learning (FL), which trains an algorithm across multiple decentralized edge devices or servers holding local data samples without exchanging their data samples, can effectively address these security and privacy issues [131], [132]. The authors in [131] presented a comprehensive introduction to FL, including 1) sample-based and 2) feature-based federated learning algorithms. They also discussed the differences of FL with distributed learning and edge computing, as well as recent works/applications of FL. Next, sample-based and feature-based federated learning algorithms are briefly reviewed.

Sample-based federated learning, also called horizontal federated learning, is proposed for scenarios where different parties have different data samples with the same features. Fig. 10 shows a typical framework for sample-based FL, where three hospitals have different patients with the same health

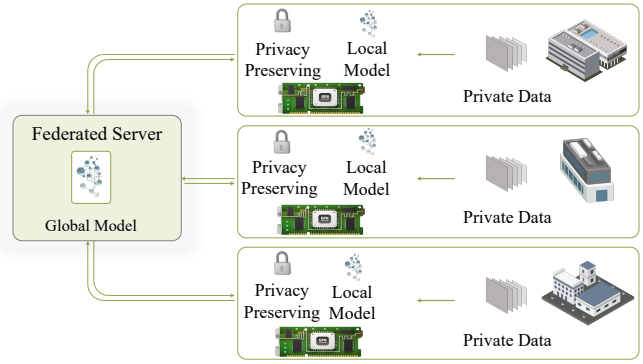

Fig. 10. An illustration of sample-based federated learning. Each party learns a local model based on its own data and then sends the local model to a cloud server. The server aggregates the model parameters from all the parties and then shares the aggregated global model to each party for updating their respective local models.

and medical features. A federated learning framework called MOCHA [133] was proposed to allow multiple sites/parties to complete their individual tasks while sharing knowledge and preserving data privacy. Meanwhile, MOCHA addressed the issues of high communication cost and fault tolerance in federated learning. The authors in [134] have proposed federated optimization methods, *e.g.*, federated stochastic variance reduced gradients (federated SVRG). These optimization methods can improve communication efficiency and facilitate the training of high-quality centralized models based on the data from multiple mobile clients. The authors presented a framework for federated learning of deep neural networks based on iterative model averaging called FedAvg [135]. The researchers in [136] further derived a bound for the convergence rate of FedAvg, showing a trade-off between communication efficiency and convergence rate. Konecny *et al.* [134] proposed two ways to reduce uplink transmission costs, *i.e.*, structured and sketched updates. The main difference between the two methods is in learning updates from variables or models. Tran *et al.* [137] investigated two trade-offs of using FL over wireless communications, *i.e.*, computation versus communication latencies and FL time versus user equipment energy consumption. The authors derived the optimal solution by characterizing the closed-form solutions. Wang *et al.* [138] proposed a CMFL framework to identify irrelevant client updates. The global tendency was used as the overall information for all the clients to check the alignment for updates.

Feature-based federated learning, also called vertical federated learning, is proposed for scenarios where the data are vertically split (*i.e.*, by features) over multiple parties. Assume that there are two parties (*i.e.*, data providers A and B) and a third-party collaborator (*i.e.*, the cloud server C). The training process of feature-based federated learning can be described in the following four steps [131]. First, cloud server C creates encryption pairs and sends a public key to A and B. A and B then encrypt and exchange the intermediate results for gradient and loss calculations. Subsequently, A and B compute encrypted gradients and send the encrypted

values to C. Lastly, C decrypts and sends the decrypted gradients and loss back to A and B, so that A and B can update the model parameters respectively. The authors presented a logistic regression classifier for feature-based federated learning where the data are vertically split [139]. Their solution includes privacy-preserving entity resolution and federated logistic regression over messages encrypted with an additively homomorphic scheme. The authors in [140] further proposed a solution for parallel distributed logistic regression in feature-based federated learning without using a third-party coordinator. Hence, the proposed solution can reduce the complexity of the system and allows any two parties to train a joint model without the help of a trusted coordinator. A lossless privacy-preserving tree-boosting system named SecureBoost [141] was proposed for feature-based federated learning. SecureBoost first conducts entity alignment under a privacy-preserving protocol and then constructs boosting trees across multiple parties through vertical federation. A privacy-preserving feature-based federated learning for tree-based models named Pivot was proposed in [142]. Similar to the solution in [140], Pivot also does not rely on any trusted third-party coordinator and thus provides better privacy protection and lower system complexity.

More recently, several advanced federated learning algorithms have been proposed and applied for more challenging scenarios (*e.g.*, cross-domain scenarios to address both the domain shift issue and the data privacy issue.) [131], [143]. Liu *et al.* [144] proposed a secure federated transfer learning (FTL) system which enables complementary knowledge to be transferred across domains without compromising data privacy. In particular, both homomorphic encryption and secret sharing were used in federated learning for privacy preservation. Federated domain adaptation [145] and federated domain generalization [146] were proposed to address the domain shift issue and generalize the federated model to the target domains (*e.g.*, new parties or sites). However, there are still many challenging issues and open questions in federated learning [147] to be addressed in the future. For example, federated learning has primarily considered supervised learning tasks where labels are available for each data provider. Extending federated learning to other machine learning paradigms, including reinforcement learning, semi-supervised, and unsupervised learning, is interesting and challenging. In addition, it is also promising to consider other ethical values (e.g., fairness, safety, interpretability, etc.) in federated learning systems.

B. Rationalizable & Resilient AI

To promote the acceptance of modern AI systems for social sustainability, the mechanism of AI models is required to be more rationalizable, in other words, possessing the ability to be rationalized (interpretable) [110]. This is also referred to as the interpretability of AI models. Another aspect of encouraging acceptance is the resilience of AI models, *i.e.*, their ability to preserve adequate performance even under extreme cases, such as adversarial attacks (input perturbations) [110], which is also known as the safety of AI. In the following, technical details for achieving AI interpretability and safety will be reviewed.

1) Interpretability: It is well known that most of the AI systems are black boxes, where it is difficult to explain how they work, especially to the general public and users. Users may not be able to trust AI models if they cannot understand their decision-making process. Especially for critical domains, like finance, medication, healthcare, and self-driving cars, it is unacceptable not to have the interpretability of AI models. Recently, more and more attention has been focused on how to explain the decisions of AI, also known as explainable AI. Many innovative explainable AI techniques have been developed in the past few years. They can be roughly divided into three categories, namely, 1) model-based, 2) model-agnostic, and 3) example-based methods.

Model-based methods mainly explore AI models that can be explained. Typical explainable AI models include linear models, *e.g.*, linear regression and logistic regression, and tree-based methods, *e.g.*, decision tree and decision rule [148], [149]. For instance, the weight of a feature in linear regression represents the importance of the feature, which can be useful in explanation. The if-then rules in a decision rule algorithm clearly indicate the relationship between inputs and outputs of the model. Domain experts, even without deep AI knowledge, can help check whether the rules that are automatically learnt from data really make sense or can be further refined.

Model-agnostic solutions are more general in explainable AI, which can be used in any AI models. A typical method is based on partial dependence plot (PDP) which shows the marginal effect of features on prediction results of an AI model [150]. LIME [151] intended to perturb inputs and observe the changes in model outputs, which can be used to explain the relationship between inputs and outputs. Another popular method named Anchors [152] tried to emphasize a part of inputs that are sufficient for an AI model to give predictions.

Example-based methods use a set of examples from a dataset to explain the underlying behaviors of AI models. Counterfactual explanations describe how an instance changes to significantly influence its prediction [153]. By designing counterfactual samples, it is possible to know how the model produces certain predictions and how to explain individual predictions. Another example-based method leverages prototypes and criticisms in which prototypes select representative samples from data and criticisms select samples that are not well represented by prototypes [154]. The way of selecting influential samples is to identify training instances that are most critical to the model or predictions [155]. Through analyzing these influential samples, one can better understand the model's behavior, which leads to better explainability.

Although various techniques have been developed for explaining AI models, the existing solutions can only deal with simple models or simply explain the relationships between inputs and outputs. The underline behaviour of complex deep AI models, *e.g.*, what has actually been learnt in each layer of a deep neural network, is still a mystery, and more research efforts should be given to this crucial area.

2) Safety: Deep learning-based AI systems have witnessed growing adoption due to their impressive performance and have recently even been expanded into critical applications where stakes are high, such as self-driving vehicles where the

safety of such systems is paramount. Notably, some motor accidents have already been connected to self-driving systems, highlighting the immense importance of AI safety. Apart from implications on human life, safe AI builds trust in society that is pivotal to its sustainable adoption by humans. Hence, it is key to consider possible threats where the system fails to perform according to its users' expectations. Recent research has shown that an adversary can undermine the safety of the system by targeting either the AI's inference or training phase, by means of adversarial examples and data poisoning respectively.

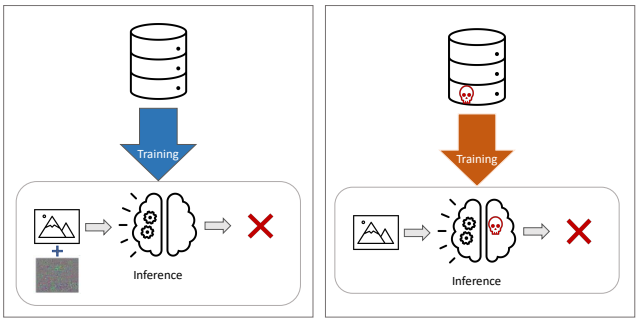

Fig. 11. Illustrations of how (left) adversarial examples and (right) data poisoning can fool an image classifier by targeting the training and inference phases, respectively.

Adversarial Examples: Much like visual blind spots in human vision, visual deep learning systems are vulnerable to imperceptible perturbations called adversarial examples [156]. In the self-driving car example, small perturbations can be added to an image of a road sign to steer the car's image classifier towards classifying the sign as a different object (*e.g.*, a high-speed sign when it is a stop sign) with high confidence, which could result in a catastrophic outcome. Such perturbations can be crafted as the input image is fed into the AI model, through the access of the model's prediction probability and gradient information [157], [158]. This information could be exploited by an adversary to compute the subtle corruption to the input image that can steer the model's prediction towards a target class. Apart from the image domain, the threat of adversarial examples exists in the audio [159]–[161] and natural language domains [162]–[164].

As mentioned above, defenses against adversarial examples have emerged. Among them, the most effective defenses are based on a simple concept of adversarial training (AT) where adversarial examples are generated and incorporated as training samples so that the AI models will learn to be robust against them during test time [165]–[168]. Initially, AT-based defenses were hindered by some limitations, such as high computational costs, due to the expensive steps of crafting strong adversarial training samples [169], [170]. To reduce the cost of AT, recent works that can reduce the number of optimization steps required for generating adversarial training examples have been proposed [167], [171], [172], while maintaining high adversarial accuracy.

In addition to the diversity of defenses, several works improved the robustness of models by minimizing the effects of small perturbations performed on the models' prediction [173]–[175] or training models to rely on less-superficial features that are more aligned with human vision [176]–[178]. Another class of defenses is provable defenses which seek to provide a performance guarantee for an AI model's performance in the face of adversarial examples [179]–[183]. Provable defenses can offer assurance to users by providing a theoretically proven worst-case scenario of the AI's performance when under attack. This can help decision-makers weigh the reward-risk trade-off of deploying the AI system with more certainty.

Data Poisoning: Modern deep-learning-based AI models rely heavily on large amounts of training data, typically mined from public sources, such as the Internet or crowdsourcing platforms. This exposes the models to the threat of data poisoning where an attacker can degrade a model's performance by corrupting a small subset of its training data as a data contributor [184]–[186]. The corruption typically involves a combination of altering the data's labels and making edits to the original training samples. A sophisticated variant of data poisoning, called backdoor poisoning, allows an adversary to control a model's prediction through a poison signature in the model's input while eluding detection as the model performs seemingly well on clean inputs [187]–[190].

Several defenses have effectively countered this threat under certain conditions. The first type of defense works by filtering poisoned samples that contain spectral signatures where their hidden states would have different statistics compared to the majority of uncorrupted samples [191]. Initial iterations of such defense either require prior knowledge, such as the ratio of poisoned samples [191], size of the poison patterns [192] or only work for small poison patterns [192], [193]. Newer variants are able to overcome these limitations to generalize a wider scope of poisoning scenarios [194], [195]. Other defense approaches include pruning away 'suspicious' neurons that lie dormant in the presence of clean validation data [196] or using differential privacy to counter the poison [197]. More recently, provable defense approaches have been studied in [198], [199] that can provide guarantees to a model's performance when the information of data poisoning is known.

The defenses and attacks on AI systems will likely see a protracted arms race as long as AI is adopted, especially in high-stake applications. It is, hence, vital to invest resources on the development of more advanced defenses to stay ahead of this race in securing AI models. On top of safety against adversaries, much work has also been studied on improving models' reliability in the face of high signal-to-noise environments [200], [201]. Exposing AI models to training samples augmented with a wide diversity of corruptions has shown to improve the performance of models during such challenging scenarios [202]–[204].

IV. Discussion

AI is changing our lives in more ways than one could imagine, such as in economy, entertainment, games, sociality,

healthcare, etc., due to the extremely fast pace of development and adoption of AI technologies. With the powerful capabilities of AI, its functionalities have been extended to be almost ubiquitous, such as the breakthrough of AlphaFold [205] in the biological domain. It is foreseeable that AI will permeate many other important areas in the near future. However, without the deep consideration of sustainable development of AI, it will violate our original intentions of developing AI to create a better world, benefiting our society and improving the quality of our lives. It is believed that now is the right moment to think beyond these huge benefits from AI and invest sufficient efforts in addressing its impacts on social, environmental and resource utilization issues. Furthermore, more and more researchers have realized the critical sustainable development issue when designing AI techniques. In this survey, AI techniques for sustainable development from two major perspectives, *i.e.*, environmental and social perspectives, have been reviewed and summarized.

Carbon footprint of AI is a major concern from the environmental perspective as AI techniques, especially deep learning, are resource-intensive. From data collection and model training to real-time inference and adaptation, huge amount of energy, memory and human efforts are required. It is clear that AI models are generating huge carbon emissions and consuming intensive computing power, exacerbated by the ever-expanding real-world applications created by technology companies, universities, research institutions, hospitals, factories, government agencies, etc.

In tackling this critical issue, the main technical challenges are to reduce the sizes of AI models and the amount of training data via emerging techniques in computation-efficient and data-efficient learning. For instance, compressing AI models into smaller sizes without significantly compromising their performance is a popular way to consume much less energy during their deployment stage. It includes techniques of network pruning, quantization and knowledge distillation, where many advanced works have been developed. For data-efficient learning, the main objective is to reduce/alleviate the use of large labeled datasets for model training, as the collection of large labeled datasets will be time-consuming, utilize intensive human efforts, and lead to a huge carbon footprint.

Even though a lot of efforts have been made in addressing the concern of AI models' carbon footprint, it is far from satisfactory in regards to true environmental sustainability of AI. Currently, most AI research is driven by performance (*e.g.*, accuracy) without taking resource constraints and carbon footprint into consideration. For example, the OpenAI GPT-3 that achieves leading performance on NLP tasks contains 175 billion parameters, which requires approximately 288 years to train on a single powerful V100 GPU [206]. Most institutes/organizations do not have sufficient resources to conduct research on GPT-3. This equipment "competition" is totally unsustainable. Pursuing better performance without considering negative impacts on the environment is a dangerous trajectory and will result in wastage of various resources and efforts. This is the right moment to critically think about our evaluation criterion for sustainable development of AI — what AI models one prefers or are qualified to be used in real-life?

In addition to the environmental concerns, the social impact of AI models is another key issue to be addressed. Social sustainability of AI is considered through the Responsible AI and Rationalizable & Resilient AI perspectives. For responsible AI, two vital aspects of fairness and data privacy are key ethics and the main focus. The existing methods for AI fairness try to solve the problem from a data perspective, which is not sufficient. Additional emerging techniques to improve model design may help to eventually solve the fairness issue. Data privacy is also a big challenge, as AI models need to be collaboratively trained with data from multiple sources. Federated learning can effectively improve data privacy, but at the expense of additional training and communication costs. These side effects may ruin the benefits created by data privacy techniques.

Another key issue in social sustainability is to achieve rationalizable & resilient AI for promoting the acceptance of AI systems. Concurrently, most AI systems are black-box, and users do not know how systems make decisions. Such black-box is not acceptable, especially for critical domains like finance, transportation and healthcare. The existing solutions for explainable AI are still in the early stage which can only explain simple models or instances. The true mechanism of advanced models, especially the powerful deep learning models, remains a mystery. The final problem is the resilience of AI models when encountering unknown perturbations, such as adversarial attacks or data poisoning, also known as the safety issue of AI. For example, there are some threats to federated learning via different types of attacks, such as inference attacks, poisoning attacks, and Sybil attacks [207], [208].

V. Challenges and Future Research Directions

To promote further development of AI, its sustainability issue must be well addressed. Various advanced methods have been designed and achieved considerable progress from the technical perspective. However, it is still far from the true or even satisfactory sustainability of AI. In this section, some potential future research directions in sustainable AI are discussed and presented.

Specifically, to achieve environmental sustainability, a fundamental way is to improve the intelligence of AI such that it does not require huge training datasets that are time-consuming and labor-intensive to prepare for model training and can easily adapt to various applications. Another way is to change our evaluation metrics to not only pursue performance in terms of accuracy but also take energy consumption and resource utilization into consideration. In this regard, a safe-sharing community for the trained AI models can be vital in promoting fast development while reducing repeated energy consumption and carbon emissions.

For social sustainability, on the other hand, unfolding complex AI models in an understandable perspective can be an efficient strategy to fully explain how AI models process data to yield explainable outcomes. It can also enhance robustness, transparency and reduce bias, as well as protection of data

privacy across different AI development stages, including data collection, processing, model design and execution, and evaluation.

a) Co-designed AI: Human intelligence is efficient and environmental-friendly, as it can learn from just a few samples and generalize the learned knowledge to different applications. However, the existing AI systems are specifically designed for particular applications and are difficult to adapt to other applications. Even though emerging techniques have been developed for different use cases, such as few-shot learning, transfer learning, self-supervised learning, etc., they can only address problems under a specific assumption, *e.g.*, availability of a large source domain dataset. Besides, a single technique may lead to conflicts when contributing to sustainability. For example, an efficient inference model via knowledge distillation may require more training efforts.

Co-designed AI aims to address the above issues via combining advanced techniques using fusion frameworks. For instance, one can combine self-supervised learning with continuous learning [209] to not only achieve sample-efficacy, but also adapt to various applications (*i.e.*, no need to learn multiple models for different yet related applications). Cognitive reasoning [210] can also be incorporated for better adaptation and learning from past knowledge instead of additional training samples, resulting in both performance improvement and reduction of training efforts.

b) Energy-aware Design: Currently, most AI systems are solely evaluated by accuracy without considering the costs of resources and human efforts. This leads to larger datasets and bigger AI models, and significantly affects environments in terms of carbon footprint and energy consumption. As the natural resources are limited, this unsustainable development will be forced to terminate in the near future. To achieve sustainable development of AI, our evaluation schemes should be changed to consider not only the accuracy but also the consumed resources in data collection, model training, and inference. For instance, the energy used for training a specific AI model should be quantified. Then it is possible to jointly optimize energy consumption and the performance of AI models. Besides, the other efforts on data collection and annotation should also be transferred to energy or a new resource utilization criterion for a more comprehensive evaluation. This framework is referred as energy-aware AI design which will encourage researchers and engineers to pay more attention to the energy and resource consumption at different stages, such as data acquisition, representation learning, model training, and deployment, when designing AI systems.

c) Model Sharing Community: Training AI models requires huge datasets and consumes a large amount of energy, especially for some big models like Transformer [211] and GPT-3 [212]. In fact, these big and powerful AI models can be reused for various applications due to their universal representation learning capability. Hence, if these models are shared across our AI community, there is no need for everyone to train these big models from scratch, which can save much time and budget. Even for different, but related applications, the technique of fine-tuning can be adopted to only train few layers of the trained big models with less data from specific applications. Besides, a smaller student model can always be trained for a specific application from the big teacher models by utilizing the big teacher models through the technique of knowledge distillation. Note that the training of the smaller student model with knowledge distillation can be efficient without the training of the big models. Such sharing community will play a key role in saving training efforts of AI models for environmental sustainability and can also promote the innovation and development of AI techniques.

d) Unfolding Complex AI models: Existing explainable AI techniques can mainly explain simple models (such as the classification models of K-Nearest Neighbors and Decision Trees, and regression models of Linear Regression) or the relationship between inputs and outputs based on data attribution and feature attribution, which are far too insufficient from the model explainability perspective. With recent AI breakthrough in deep learning, the actual interpretability is the clear expression on how it works for each single layer of neural networks. CNN Explainer [213] is a good start in shedding light on deep convolutional neural networks by visualizing all convolutional operations. However, it does not explore on how these convolutional operations have been combined to make the final prediction. This issue of model-level interpretability is extremely challenging but of great importance. With the fully explainable models, it may also be able to perform self-diagnosis on whether fairness issues exist, which can thoroughly avoid biases of AI models. Besides, the robustness and data privacy aspects can also be enhanced when the operations of AI models are fully understood.

e) Preemptive Approach for AI Safety: Given how wide and diverse the possible AI applications are, it is critical to scrutinize each stage of an AI system's training and deployment processes for possible entry point where a malicious actor can attack. Privileged access management [214], a widely adopted cybersecurity practice, can help reduce the attack surface (*e.g.*, training data, model weights, etc.) where a malicious attacker has on the AI system. As the AI workflow advances with novel development, such as federated learning [131], it is also important to consider and develop defenses for new safety goals, such as privacy protection, despite the model performance that an attacker might seek to undermine. With AI systems deployed in increasingly important applications, defenses must be developed proactively to prevent catastrophic safety incidents from happening.

VI. Conclusion

In this paper, a comprehensive technical survey on challenges confronting sustainable AI has been presented, with focused discussions on the sustainability of AI through two major perspectives of environmental sustainability and social sustainability of AI. In particular, computation-efficient and data-efficient AI techniques have been first examined for addressing environmental sustainability issue. Then, social sustainability of AI has been discussed from the angles of responsible AI and rationalizable & resilient AI. Meanwhile, limitations and challenges, as well as potential research directions in sustainable AI, have been discussed and presented

in an effort to highlight the importance of this topic, attract researchers' attention, and rally more research efforts in this critical area. Finally, it is believed that AI has great potential to be sustainable and at the same time create huge economic and social impacts on our beautiful Mother Nature, provided that our researchers can continuously factor in environmental, social, resource utilization, and performance costs when investigating and improving their critical research problems and applications.

ACKNOWLEDGEMENT

This research is supported in part by the A*STAR Center for Frontier AI Research, and the School of Computer Science and Engineering at Nanyang Technological University.

REFERENCES

[1] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] S. S. Khanal, P. Prasad, A. Alsadoon, and A. Maag, "A systematic review: machine learning based recommendation systems for e-learning," *Education and Information Technologies*, vol. 25, no. 4, pp. 2635–2664, 2020.

[4] Z. Chen, M. Wu, and X. Li, *Generalization with Deep Learning: For Improvement on Sensing Capability*. World Scientific, 2021.

[5] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019.

[6] M. Chau and H. Chen, "A machine learning approach to web page filtering using content and structure analysis," *Decision Support Systems*, vol. 44, no. 2, pp. 482–494, 2008.

[7] A. Amberkar, P. Awasarmol, G. Deshmukh, and P. Dave, "Speech recognition using recurrent neural networks," in *2018 international conference on current trends towards converging technologies (ICCTCT)*. IEEE, 2018, pp. 1–4.

[8] M. Leonetti, L. Iocchi, and P. Stone, "A synthesis of automated planning and reinforcement learning for efficient, robust decision-making," *Artificial Intelligence*, vol. 241, pp. 103–130, 2016.

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[10] S. Larsson, M. Anneroth, A. Felländer, L. Felländer-Tsai, F. Heintz, and R. C. Ångström, "Sustainable ai: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence," 2019.

[11] A. van Wynsberghe, "Sustainable ai: Ai for sustainability and the sustainability of ai," *AI and Ethics*, pp. 1–6, 2021.

[12] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, 2020, pp. 13 693–13 696.

[13] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.

[14] J. Mensah, "Sustainable development: Meaning, history, principles, pillars, and implications for human action: Literature review," *Cogent social sciences*, vol. 5, no. 1, p. 1653531, 2019.

[15] I. Kindylidi and T. S. Cabral, "Sustainability of ai: The case of provision of information to consumers," *Sustainability*, vol. 13, no. 21, p. 12064, 2021.

[16] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, and F. F. Nerini, "The role of artificial intelligence in achieving the sustainable development goals," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.

[17] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[18] K. Hao, "Training a single ai model can emit as much carbon as five cars in their lifetimes," *MIT Technology Review*, 2019.

[19] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[21] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[22] H. Alemdar, V. Leroy, A. Prost-Boucle, and F. Pétrot, "Ternary neural networks for resource-efficient ai applications," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2547–2554.

[23] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in neural information processing systems*, 1990, pp. 598–605.

[24] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas, "Predicting parameters in deep learning," in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, 2013, pp. 2148–2156.

[25] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artificial Intelligence Review*, vol. 53, no. 7, pp. 5113–5155, 2020.

[26] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural network," in *NIPS*, 2015.

[27] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[28] S. Hanson and L. Pratt, "Comparing biases for minimal network construction with back-propagation," *Advances in neural information processing systems*, vol. 1, pp. 177–185, 1988.

[29] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.

[30] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," *Advances in neural information processing systems*, vol. 29, pp. 2074–2082, 2016.

[31] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2736–2744.

[32] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang, "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures," *arXiv preprint arXiv:1607.03250*, 2016.

[33] B. Hassibi and D. G. Stork, *Second order derivatives for network pruning: Optimal brain surgeon*. Morgan Kaufmann, 1993.

[34] N. Lee, T. Ajanthan, and P. H. Torr, "Snip: Single-shot network pruning based on connection sensitivity," *arXiv preprint arXiv:1810.02340*, 2018.

[35] H. X. Choong, Y.-S. Ong, A. Gupta, and R. Lim, "Jack and masters of all trades: One-pass learning of a set of model sets from foundation models," 2022.

[36] Q. Huang, K. Zhou, S. You, and U. Neumann, "Learning to prune filters in convolutional neural networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 709–718.

[37] A. Renda, J. Frankle, and M. Carbin, "Comparing rewinding and fine-tuning in neural network pruning," *arXiv preprint arXiv:2003.02389*, 2020.

[38] C. Wang, G. Zhang, and R. Grosse, "Picking winning tickets before training by preserving gradient flow," *arXiv preprint arXiv:2002.07376*, 2020.

[39] D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta, "Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science," *Nature communications*, vol. 9, no. 1, pp. 1–12, 2018.

[40] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *International conference on machine learning*. PMLR, 2015, pp. 1737–1746.

[41] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on cpus," *Deep Learning and Unsupervised Feature Learning Workshop, NIPS*, 2011.

[42] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[43] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 525–542.

[44] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[45] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," *arXiv preprint arXiv:1702.03044*, 2017.

[46] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *arXiv preprint arXiv:1606.06160*, 2016.

[47] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv:1412.6115*, 2014.

[48] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4820–4828.

[49] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[50] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[51] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[52] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS Deep Learning and Representation Learning Workshop*, vol. 1050, p. 9, 2014.

[53] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[54] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[55] S. Lee and B. C. Song, "Graph-based knowledge distillation by multi-head attention network," *arXiv preprint arXiv:1907.02226*, 2019.

[56] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[57] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[58] S. H. Lee, D. H. Kim, and B. C. Song, "Self-supervised knowledge distillation using singular value decomposition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 335–350.

[59] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[60] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[61] F. Yuan, L. Shou, J. Pei, W. Lin, M. Gong, Y. Fu, and D. Jiang, "Reinforced multi-teacher selection for knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 284–14 291.

[62] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1607–1616.

[63] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[64] Q. Xu, Z. Chen, K. Wu, C. Wang, M. Wu, and X. Li, "Kdnet-rul: A knowledge distillation framework to compress deep neural networks for machine remaining useful life prediction," *IEEE Transactions on Industrial Electronics*, 2021.

[65] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," *arXiv preprint arXiv:2106.05237*, 2021.

[66] B. Li, X. Jiang, D. Bai, Y. Zhang, N. Zheng, X. Dong, L. Liu, Y. Yang, and D. Li, "Full-cycle energy consumption benchmark for low-carbon computer vision," *arXiv preprint arXiv:2108.13465*, 2021.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[68] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang, "A survey of deep active learning," *arXiv preprint arXiv:2009.00236*, 2020.

[69] B. Settles, "Active learning literature survey," 2009.

[70] R. Schumann and I. Rehbein, "Active learning via membership query synthesis for semi-supervised sentence classification," in *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 2019, pp. 472–481.

[71] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain," *Information sciences*, vol. 285, pp. 181–203, 2014.

[72] T. He, X. Jin, G. Ding, L. Yi, and C. Yan, "Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1360–1365.

[73] Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.

[74] Y. Geifman and R. El-Yaniv, "Deep active learning over the long tail," *arXiv preprint arXiv:1711.00941*, 2017.

[75] Y. Wang and Q. Yao, "Few-shot learning: A survey," *arXiv preprint arXiv:1904.05046*, 2019.

[76] J. Vanschoren, "Meta-learning: A survey," *arXiv preprint arXiv:1810.03548*, 2018.

[77] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[78] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 719–11 727.

[79] Z. Hou, A. Walid, and S.-Y. Kung, "Meta-learning with attention for improved few-shot learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2725–2729.

[80] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[81] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*. Springer, 2015, pp. 84–92.

[82] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[83] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[84] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[85] P. Hemmer, N. Kühl, and J. Schöffer, "Deal: deep evidential active learning for image classification," in *Deep Learning Applications, Volume 3*. Springer, 2022, pp. 171–192.

[86] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, "Multiple instance active learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5330–5339.

[87] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8012–8021.

[88] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[89] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[90] C. Chen, Z. Fu, Z. Chen, S. Jin, Z. Cheng, X. Jin, and X.-S. Hua, "Homm: Higher-order moment matching for unsupervised domain adaptation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3422–3429, 04 2020.

[91] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*. Springer, 2017, pp. 153–171.

[92] M. M. Rahman, C. Fookes, M. Baktashmotlagh, and S. Sridharan, "On minimum discrepancy estimation for deep domain adaptation," in *Domain Adaptation for Visual Understanding*. Springer, 2020, pp. 81–94.

[93] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[94] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–35, 2016.

[95] K. Wu, M. Wu, J. Yang, Z. Chen, Z. Li, and X. Li, "Deep reinforcement learning boosted partial domain adaptation," *IJCAI*, 2021.

[96] S. Wang and L. Zhang, "Self-adaptive re-weighted adversarial domain adaptation," *arXiv preprint arXiv:2006.00223*, 2020.

[97] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: transfer learning through adaptive fine-tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4805–4814.

[98] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[99] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[100] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[101] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[102] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

[103] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[104] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations (ICLR)*, 2018.

[105] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[106] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European conference on computer vision*. Springer, 2016, pp. 577–593.

[107] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[108] A. N. Carr, Q. Berthet, M. Blondel, O. Teboul, and N. Zeghidour, "Shuffle to learn: Self-supervised learning from permutations via differentiable ranking," 2020.

[109] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[110] Y.-S. Ong and A. Gupta, "Air 5: Five pillars of artificial intelligence research," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 411–415, 2019.

[111] B. Green and Y. Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 90–99.

[112] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 429–435.

[113] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *international conference on machine learning*. PMLR, 2016, pp. 1670–1679.

[114] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[115] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *arXiv preprint arXiv:2010.03240*, 2020.

[116] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[117] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[118] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[119] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[120] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[121] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[122] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[123] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 702–712.

[124] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu, "Achieving causal fairness through generative adversarial networks," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[125] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 962–970.

[126] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," in *International conference on machine learning*. PMLR, 2018, pp. 2737–2746.

[127] I. Valera, A. Singla, and M. Gomez Rodriguez, "Enhancing the accuracy and fairness of human decision making," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[128] A. Narayanan, "Translation tutorial: 21 fairness definitions and their politics," in *Proc. Conf. Fairness Accountability Transp., New York, USA*, vol. 1170, 2018, p. 3.

[129] Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi, "Understanding and improving fairness-accuracy trade-offs in multi-task learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1748–1757.

[130] M. M. Khalili, X. Zhang, M. Abroshan, and S. Sojoudi, "Improving fairness and privacy in selection problems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 8092–8100.

[131] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[132] H. Chen, B. An, D. Niyato, Y. C. Soh, and C. Miao, "Workload factoring and resource sharing via joint vertical and horizontal cloud federation networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 3, pp. 557–570, 2017.

[133] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4427–4437.

[134] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[135] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[136] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2019.

[137] N. H. Tran, W. Bao, A. Zomaya, M. N. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1387–1395.

[138] W. Luping, W. Wei, and L. Bo, "Cmfl: Mitigating communication overhead for federated learning," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 954–964.

[139] S. Hardy, W. Henecka, H. Ivey-Law, R. Nock, G. Patrini, G. Smith, and B. Thorne, "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption," *arXiv preprint arXiv:1711.10677*, 2017.

[140] S. Yang, B. Ren, X. Zhou, and L. Liu, "Parallel distributed logistic regression for vertical federated learning without third-party coordinator," *arXiv preprint arXiv:1911.09824*, 2019.

[141] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos, and Q. Yang, "Secureboost: A lossless federated learning framework," *IEEE Intelligent Systems*, 2021.

[142] Y. Wu, S. Cai, X. Xiao, G. Chen, and B. C. Ooi, "Privacy preserving vertical federated learning for tree-based models," *Proceedings of the VLDB Endowment*, vol. 13, no. 11, 2020.

[143] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[144] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A secure federated transfer learning framework," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 70–82, 2020.

[145] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," in *International Conference on Learning Representations*, 2020.

[146] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1023.

[147] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[148] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[149] X. Li and B. Liu, "Rule-based classification," in *Data Classification: Algorithms and Applications*, 2013.

[150] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[151] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[152] ——, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[153] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[154] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[155] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1885–1894.

[156] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[157] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.

[158] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.

[159] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[160] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, "" hello, it's me": Deep learning-based speech synthesis attacks in the real world," *arXiv preprint arXiv:2109.09598*, 2021.

[161] P. Żelasko, S. Joshi, Y. Shao, J. Villalba, J. Trmal, N. Dehak, and S. Khudanpur, "Adversarial attacks and defenses for speech recognition systems," *arXiv preprint arXiv:2103.17122*, 2021.

[162] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," *arXiv preprint arXiv:1707.07328*, 2017.

[163] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[164] S. Tan, S. Joty, M.-Y. Kan, and R. Socher, "It's morphin'time! combating linguistic discrimination with inflectional perturbations," *arXiv preprint arXiv:2005.04364*, 2020.

[165] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[166] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," *arXiv preprint arXiv:1901.08573*, 2019.

[167] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *arXiv preprint arXiv:1904.12843*, 2019.

[168] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 278–11 287.

[169] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018.

[170] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 501–509.

[171] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang, "Robust overfitting may be mitigated by properly learned smoothening," in *International Conference on Learning Representations*, 2020.

[172] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," *arXiv preprint arXiv:2007.02617*, 2020.

[173] H. Drucker and Y. Le Cun, "Double backpropagation increasing generalization performance," in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, vol. 2. IEEE, 1991, pp. 145–150.

[174] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[175] D. Jakubovitz and R. Giryes, "Improving dnn robustness to adversarial attacks using jacobian regularization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 514–529.

[176] C. Etmann, S. Lunz, P. Maass, and C.-B. Schönlieb, "On the connection between adversarial robustness and saliency map interpretability," *arXiv preprint arXiv:1905.04172*, 2019.

[177] A. Chan, Y. Tay, Y. S. Ong, and J. Fu, "Jacobian adversarially regularized networks for robustness," *arXiv preprint arXiv:1912.10185*, 2019.

[178] A. Chan, Y. Tay, and Y.-S. Ong, "What it thinks is important is important: Robustness transfers through input gradients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 332–341.

[179] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Advances in Neural Information Processing Systems*, 2017, pp. 2266–2276.

[180] A. Raghunathan, J. Steinhardt, and P. S. Liang, "Semidefinite relaxations for certifying robustness to adversarial examples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 877–10 887.

[181] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[182] M. Balunovic and M. Vechev, "Adversarial training and provable defenses: Bridging the gap," in *International Conference on Learning Representations*, 2019.

[183] A. Blum, T. Dick, N. Manoj, and H. Zhang, "Random smoothing might be unable to certify l_∞ robustness for high-dimensional images," *Journal of Machine Learning Research*, vol. 21, pp. 1–21, 2020.

[184] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. A. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter." *LEET*, vol. 8, pp. 1–9, 2008.

[185] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, "Support vector machines under adversarial label contamination," *Neurocomputing*, vol. 160, pp. 53–62, 2015.

[186] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in neural information processing systems*, 2017, pp. 3517–3529.

[187] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[188] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25nd Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018*. The Internet Society, 2018.

[189] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 1615–1631.

[190] A. Chan, Y. Tay, Y.-S. Ong, and A. Zhang, "Poison attacks against text datasets with conditional adversarially regularized autoencoder," *arXiv preprint arXiv:2010.02684*, 2020.

[191] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Advances in Neural Information Processing Systems*, 2018, pp. 8011–8021.

[192] X. Qiao, Y. Yang, and H. Li, "Defending neural backdoors via generative distribution modeling," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 004–14 013.

[193] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks*. IEEE, 2019, p. 0.

[194] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," *arXiv preprint arXiv:1811.03728*, 2018.

[195] A. Chan and Y.-S. Ong, "Poison as a cure: Detecting & neutralizing variable-sized backdoor attacks in deep neural networks," *arXiv preprint arXiv:1911.08040*, 2019.

[196] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 273–294.

[197] Y. Ma, X. Zhu, and J. Hsu, "Data poisoning against differentially-private learners: Attacks and defenses," *arXiv preprint arXiv:1903.09860*, 2019.

[198] E. Rosenfeld, E. Winston, P. Ravikumar, and Z. Kolter, "Certified robustness to label-flipping attacks via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8230–8241.

[199] M. Weber, X. Xu, B. Karlaš, C. Zhang, and B. Li, "Rab: Provable robustness against backdoor attacks," *arXiv preprint arXiv:2003.08904*, 2020.

[207] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *arXiv preprint arXiv:2012.06337*, 2020.

[200] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE, 2017, pp. 1–7.

[201] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," *arXiv preprint arXiv:1907.07174*, 2019.

[202] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[203] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," *arXiv preprint arXiv:2006.16241*, 2020.

[204] K. Kireev, M. Andriushchenko, and N. Flammarion, "On the effectiveness of adversarial training against common corruptions," *arXiv preprint arXiv:2103.02325*, 2021.

[205] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.

[206] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro *et al.*, "Efficient large-scale language model training on gpu clusters using megatron-lm," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021, pp. 1–15.

[208] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning." Morgan & Claypool Publishers, 2019.

[209] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios," *Neural Networks*, vol. 116, pp. 56–73, 2019.

[210] H. L. H. de Penning, A. S. d. Garcez, L. C. Lamb, and J.-J. C. Meyer, "A neural-symbolic cognitive agent for online learning and reasoning," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[211] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[212] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[213] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. Chau, "CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization," *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020.

[214] K. S. Tep, B. Martini, R. Hunt, and K.-K. R. Choo, "A taxonomy of cloud attack consequences and mitigation strategies: The role of access control and privileged access management," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1. IEEE, 2015, pp. 1073–1080.