Power System Technology
ISSN:1000-3673

Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

Enhancing Robustness and Generalization in Deep Learning Models for Image Processing.

[*1]Dr. Nand Kumar, [2]Balakumar Muniandi, [3]Dr. Elma Sibonghanoy Groenewald, [4]Pravin Gawande, [5]Sohong Dhar, [6]Gonesh Chandra Saha

[*1]Assistant Professor, Department of Physics, Jawaharlal Nehru Memorial PG College, Barabanki, UP, India.
ORCID- 0009-0004-3139-3385.

[2]Associate Professor of Practice, Lawrence Technological University, Michigan, USA.
ORCID: - 0000-0003-2298-5093.

[3] Executive Department , CEO, SG Virtuosos International,
1501-1502 Tran Phu Street, Loc Tho Ward, Nha Trang City,
Khan Hoa Province, Vietnam 650000.
ORCID:0000-0001-7813-2773.

[4]Assistant Professor, Electronics and Telcommunication Engineering,
Vishwakarma Institute of Information Technology, Pune - 48.
ORCID:0000-0003-3342-2368.

[5]Data Scientist Associate, Azure, Microsoft, Jadavpur University.

[6]Associate Professor, Department of Computer Science & Information Technology,
Bangabandhu Sheikh Mujibur Rahman Agricultural University (BSMRAU), Gazipur 1706.

[*]*Corresponding Author: -Dr. Nand Kumar*

Abstract: - In recent years, deep learning models have demonstrated remarkable success in various image processing tasks, ranging from object recognition to medical image analysis. However, their performance often degrades in the presence of unseen data or adversarial attacks, highlighting the need for enhancing robustness and generalization. This paper explores innovative approaches to address these challenges, aiming to improve the reliability and applicability of deep learning models in real-world scenarios. The first section of the paper delves into the importance of robustness and generalization in deep learning models for image processing tasks. It discusses the implications of model vulnerabilities, such as overfitting to training data and susceptibility to adversarial perturbations, on the reliability of model predictions. [1] Through a comprehensive review of existing literature, various factors influencing robustness and generalization are identified, including dataset diversity, model architecture, regularization techniques, and adversarial training methods. The paper proposes novel methodologies to enhance the robustness

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

and generalization capabilities of deep learning models. One key approach involves the integration of diverse training data sources, including synthetic data augmentation and domain adaptation techniques, to expose the model to a wider range of scenarios and improve its ability to generalize to unseen data. Additionally, advanced regularization techniques, such as dropout and batch normalization, are explored to mitigate overfitting and improve model generalization. The paper investigates the effectiveness of adversarial training strategies in enhancing model robustness against adversarial attacks. By incorporating adversarially generated examples during training, deep learning models can learn to better resist perturbations and maintain performance under adversarial conditions. Moreover, the paper explores the potential of incorporating uncertainty estimation methods, such as Bayesian neural networks and Monte Carlo dropout, to quantify model uncertainty and improve robustness in uncertain environments. This paper presents a comprehensive investigation into enhancing robustness and generalization in deep learning models for image processing tasks. By addressing key challenges such as overfitting, dataset bias, and adversarial vulnerabilities, the proposed methodologies offer promising avenues for improving the reliability and applicability of deep learning models in real-world scenarios.

Keywords: - Deep Learning, Robustness, Generalization, Image Processing, Adversarial Attacks, Dataset Diversity, Regularization Techniques, Adversarial Training.

1.Introduction: - The advent of deep learning has revolutionized the field of image processing, enabling unprecedented achievements in tasks ranging from object recognition to image synthesis. However, the deployment of deep learning models in real-world applications is often hindered by their susceptibility to adversarial attacks, domain shifts, and overfitting. Consequently, ensuring the robustness and generalization of these models has become a paramount concern for researchers and practitioners alike. [2] Robustness in deep learning refers to a model's ability to maintain high performance even in the face of perturbations, noise, or adversarial inputs. Achieving robustness is essential for ensuring the reliability and safety of deep learning systems deployed in critical applications, such as medical imaging or autonomous driving. However, traditional deep learning architectures often exhibit vulnerabilities to adversarial attacks, where carefully crafted perturbations can lead to misclassification or erroneous behavior. Moreover, variations in input data distribution between the training and deployment environments can undermine the generalization capabilities of deep learning models, leading to performance degradation in real-world scenarios.

Addressing these challenges requires a multifaceted approach that encompasses algorithmic innovations, robust training methodologies, and rigorous evaluation protocols. In recent years, researchers have proposed a plethora of techniques to enhance the robustness and generalization of deep learning models for image processing. [3] Adversarial training, for instance, involves

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

augmenting the training data with adversarial examples to improve the model's resilience against adversarial attacks. Similarly, domain adaptation methods aim to bridge the gap between different data distributions by learning domain-invariant representations, thereby facilitating better generalization to unseen environments.

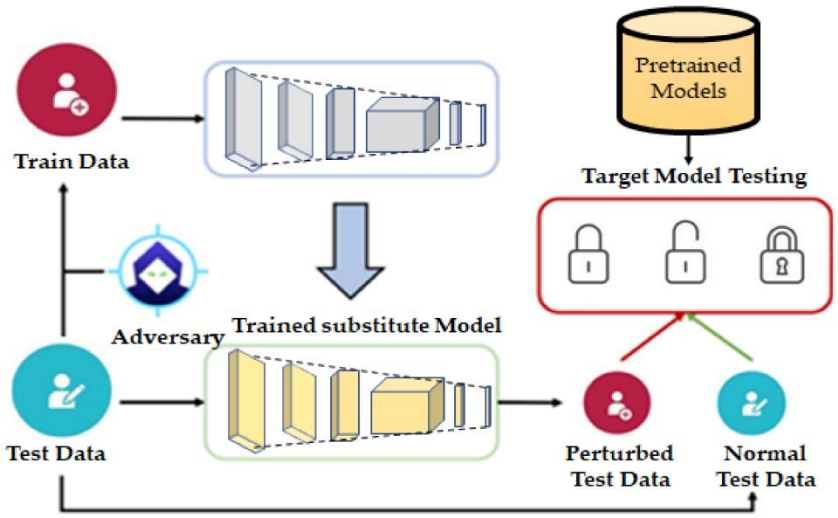

Figure 1 Robustness for Deep Learning Models for Image Processing.

Furthermore, regularization techniques, such as dropout or weight decay, play a crucial role in mitigating overfitting and improving the generalization performance of deep learning models. Additionally, data augmentation strategies, [4] such as random rotations or translations, can help expose the model to a diverse range of input variations during training, thereby enhancing its robustness to unforeseen conditions. In this paper, we provide a comprehensive overview of the latest advancements and methodologies aimed at enhancing the robustness and generalization of deep learning models for image processing.

2. Literature Review: - The literature on enhancing the robustness and generalization of deep learning models for image processing reflects a growing recognition of the critical importance of addressing these challenges for real-world deployment. [5] Adversarial attacks, domain shifts, and overfitting remain significant hurdles, prompting researchers to explore a variety of strategies and techniques to bolster model resilience and adaptability.

Power System Technology
ISSN:1000-3673

Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

Adversarial training has emerged as a prominent approach to fortifying deep learning models against adversarial attacks. Szegedy et al. (2013) introduced the concept of adversarial examples, demonstrating that imperceptible perturbations can lead to misclassification. Subsequent work by Goodfellow et al. (2015) proposed adversarial training, where models are trained on a combination of clean and adversarial examples, thereby improving their robustness. This approach has been further refined with techniques like adversarial training with ensemble methods (Tramèr et al., 2018) and adversarial training with label smoothing (Warde-Farley & Goodfellow, 2016), achieving notable success in defending against adversarial attacks.

Domain adaptation has also garnered attention as a means to enhance model generalization across different data distributions. Ganin et al. (2016) introduced domain-adversarial neural networks, which learn domain-invariant representations by simultaneously minimizing domain discrepancy and maximizing task performance. Similar approaches, such as adversarial discriminative domain adaptation (Tzeng et al., 2017) and cycle-consistent adversarial networks (Zhu et al., 2017), have been proposed to address domain shifts in image processing tasks, enabling models to generalize effectively to unseen domains.

Regularization techniques play a pivotal role in mitigating overfitting and improving generalization performance. Dropout, introduced by Srivastava et al. (2014), randomly drops neurons during training, effectively preventing co-adaptation of features and promoting model robustness. Weight decay, another widely used regularization method, imposes penalties on large weights, encouraging simpler and more generalizable models (Krogh & Hertz, 1992).

Data augmentation strategies have also proven instrumental in enhancing model robustness by exposing models to diverse input variations during training. Techniques such as random rotations, translations, and flips augment the training data, facilitating better generalization to unseen conditions (Shorten & Khoshgoftaar, 2019).

3.Robustness in Deep Learning Models for Image processing: Challenges and Vulnerabilities: Robustness in deep learning models for image processing is a critical aspect that ensures their reliability and effectiveness in real-world applications. It refers to the ability of a model to maintain high performance even in the presence of perturbations, noise, or adversarial inputs. [6] Achieving robustness is essential because images encountered in practical scenarios may vary significantly from those seen during training, leading to performance degradation if the model cannot generalize well.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

One of the key challenges in achieving robustness is the vulnerability of deep learning models to adversarial attacks. Adversarial examples are crafted by making imperceptible perturbations to input images, leading to misclassification by the model. [7] These attacks can have serious consequences, particularly in safety-critical applications like autonomous vehicles or medical diagnosis. Various techniques have been proposed to enhance the robustness of deep learning models against adversarial attacks, including adversarial training, defensive distillation, and gradient masking.

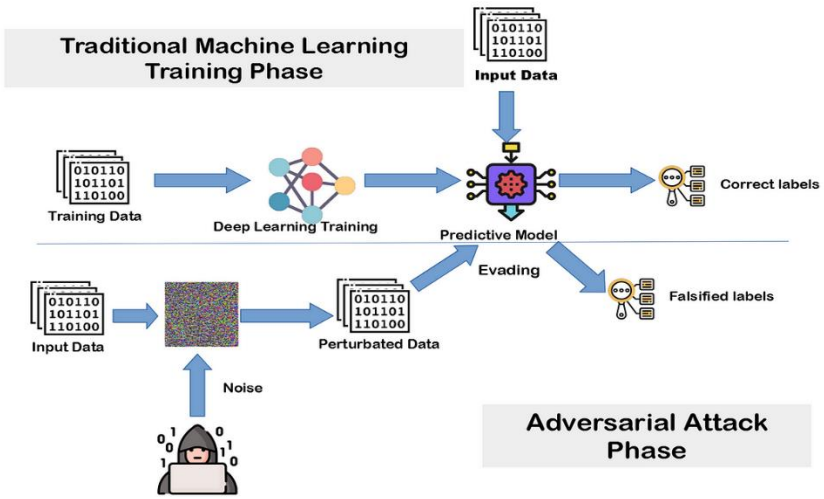

Figure 2 Adversarial Attacks

Adversarial training is a widely used approach where the model is trained on a combination of clean and adversarial examples. By exposing the model to adversarial perturbations during training, it learns to be more robust against such attacks. [8] Defensive distillation involves training a robust model using the softened outputs of another model trained on the same task, effectively smoothing out the decision boundaries and making the model more resistant to adversarial perturbations. Gradient masking techniques aim to obscure gradient information to prevent adversaries from crafting effective perturbations, although they may come with trade-offs in terms of model performance and interpretability.

Another challenge to robustness in deep learning models for image processing is the presence of noisy or corrupted input data. [9] Traditional deep learning models may struggle to perform well when faced with images that contain artifacts, occlusions, or variations in lighting conditions. Robustness against such noise can be improved through data augmentation techniques, such as random rotations, translations, or adding Gaussian noise to the input images during training.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

Additionally, robust architectures like convolutional neural networks (CNNs) with skip connections or attention mechanisms can better handle noisy input data by capturing relevant features while ignoring irrelevant noise.

4. Generalization in Deep Learning Models for Image processing: Challenges and Vulnerabilities: Generalization in deep learning models for image processing is a fundamental aspect that determines their ability to perform well on unseen data samples or tasks beyond their training distribution. [10] It encompasses the model's capacity to extract meaningful features and patterns from the training data and apply them effectively to novel inputs. Achieving strong generalization is essential for ensuring the reliability, adaptability, and scalability of deep learning models in real-world applications.

One of the key challenges in generalization is overfitting, where the model learns to memorize the training data rather than capturing underlying patterns. Overfitting occurs when the model becomes overly complex, capturing noise or irrelevant variations in the training data. Regularization techniques play a crucial role in mitigating overfitting by imposing constraints on the model's parameters, thereby encouraging simpler and more generalizable representations. Techniques such as dropout, weight decay, and early stopping are commonly used to prevent overfitting and improve generalization performance.

Another challenge to generalization in deep learning models for image processing is domain shifts, where the distribution of the input data differs between the training and deployment environments. For example, a model trained on images captured under controlled conditions in a laboratory may struggle to generalize to images captured in real-world scenarios with varying lighting conditions, backgrounds, or camera perspectives. Domain adaptation techniques aim to address this challenge by learning domain-invariant representations that are robust to changes in the input data distribution. By aligning the feature spaces of different domains, domain adaptation enables the model to generalize effectively across diverse environments.

Data augmentation is another effective strategy for improving generalization in deep learning models for image processing. By augmenting the training data with diverse transformations such as random rotations, translations, or flips, data augmentation exposes the model to a broader range of input variations, thereby improving its ability to generalize to unseen conditions. Furthermore, data augmentation can help mitigate the effects of imbalanced or limited training data by generating synthetic examples that augment the available training samples.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

Transfer learning is a powerful technique for leveraging pre-trained models to improve generalization performance, especially when labeled training data is scarce or costly to obtain. By fine-tuning a pre-trained model on a new task or domain, transfer learning allows the model to transfer knowledge learned from the source domain to the target domain, thereby accelerating the learning process and improving generalization performance.

5. Enhancing the robustness and generalization of deep learning models for Image Processing: - is essential for their effective deployment in real-world scenarios. Robustness ensures that models can maintain high performance even in the face of adversarial attacks, noise, or distribution shifts, while generalization enables models to generalize well to unseen data samples or tasks beyond their training distribution. In this article, we explore various approaches and techniques aimed at enhancing the robustness and generalization of deep learning models, encompassing adversarial training, domain adaptation, regularization techniques, and data augmentation strategies.

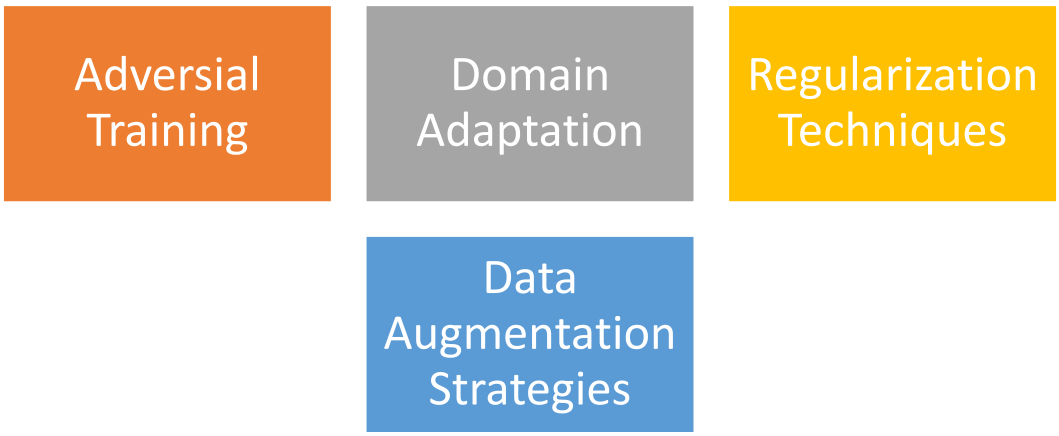

Figure 3 Techniques to Enhance Robustness and Generalisation

5.1Adversarial Training: Adversarial training is a prominent approach for enhancing the robustness of deep learning models against adversarial attacks. [11] It involves augmenting the training data with adversarial examples generated by adding imperceptible perturbations to input images. By training on a combination of clean and adversarial examples, the model learns to be more resilient to adversarial perturbations during inference.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

One common method for generating adversarial examples is the Fast Gradient Sign Method (FGSM) proposed by Goodfellow et al. (2015). FGSM computes the gradient of the loss function with respect to the input image and perturbs the image in the direction that maximizes the loss. Adversarial training using FGSM has been shown to improve the robustness of deep learning models against adversarial attacks.

The Fast Gradient Sign Method (FGSM) is a powerful technique used in adversarial training for deep learning models, particularly in the context of enhancing model robustness against adversarial attacks. Proposed by Goodfellow et al. in 2015, the FGSM leverages the gradient of the loss function with respect to the input data to generate adversarial perturbations efficiently.

The key idea behind the FGSM is to compute the sign of the gradient of the loss function with respect to the input data and then perturb the input data in the direction that maximizes the loss. [12] This perturbation is added to the original input data to generate an adversarial example. Mathematically, the FGSM can be expressed as follows:

Adversarial example: - $x+\epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$

Where: -

- x is the original input data.
- ϵ is a small scalar that controls the magnitude of the perturbation.
- $J(\theta, x, y)$ is the loss function with parameters θ, input data x, and true labels y.
- $\nabla_x J(\theta, x, y)$ is the gradient of the loss function with respect to the input data.

By adding the perturbation in the direction of the gradient sign, the FGSM effectively maximizes the loss with respect to the input data, leading to misclassification by the model. The magnitude of the perturbation is controlled by the parameter ϵ, which determines the strength of the adversarial attack. Small values of ϵ result in subtle perturbations that are less likely to be detected by humans but still effective in deceiving the model.

Adversarial training can be further enhanced by incorporating ensemble methods or incorporating adversarial examples during fine-tuning. Tramèr et al. (2018) proposed adversarial training with ensemble methods, where multiple models are trained on different subsets of adversarial examples and combined to improve robustness. Additionally, Madry et al. (2018) introduced the robust optimization framework, where models are trained using projected gradient descent to maximize the worst-case loss over a set of adversarial examples.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

5.2Domain Adaptation: Domain adaptation techniques aim to address the challenge of distribution shifts between the training and deployment environments. [13] When the distribution of the input data differs between domains, models may struggle to generalize well to unseen domains. Domain adaptation methods aim to learn domain-invariant representations that are robust to changes in the input data distribution.

One popular approach for domain adaptation is adversarial domain adaptation, where a domain discriminator is trained to distinguish between source and target domain samples, while a feature extractor is trained to confuse the domain discriminator by generating domain-invariant features. Ganin et al. (2016) introduced domain-adversarial neural networks (DANN), which simultaneously minimize the domain discrepancy and maximize the task performance, effectively aligning the feature spaces of different domains.

Another approach for domain adaptation is unsupervised domain adaptation, where models are trained on labeled source domain data and unlabeled target domain data. Tzeng et al. (2017) proposed adversarial discriminative domain adaptation (ADDA), which learns domain-invariant representations by aligning the distributions of source and target domain features in a shared feature space.

5.3 Regularization Techniques: Regularization techniques play a crucial role in preventing overfitting and improving generalization performance. By imposing constraints on the model's parameters, regularization techniques encourage simpler and more generalizable representations.

In image processing tasks, deep learning models often face the challenge of overfitting, where the model learns to memorize the training data rather than capturing underlying patterns. Overfitting can occur due to the high dimensionality of image data and the complexity of deep neural network architectures. Dropout regularization mitigates overfitting by introducing randomness into the training process, effectively forcing the model to learn more robust and generalizable representations.

The dropout regularization technique works by randomly dropping a certain percentage of neurons in each layer of the network during training. At each training iteration, neurons are probabilistically deactivated with a specified dropout probability, typically set between 0.2 and 0.5. [14] This dropout process is applied independently to each neuron, meaning that different subsets of neurons are dropped at each iteration, introducing diversity into the network's learning process.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

During forward pass, the output of each neuron is scaled by the dropout probability, effectively simulating the presence of all neurons during inference. During backpropagation, gradients are only propagated through active neurons, which helps prevent overfitting by preventing the network from relying too heavily on specific neurons or features.

Dropout is a popular regularization technique introduced by Srivastava et al. (2014), where neurons are randomly dropped during training to prevent co-adaptation of features and improve model generalization. Weight decay, another commonly used regularization method, imposes penalties on large weights to encourage simpler models. Additionally, early stopping, which stops training when the validation error starts to increase, is another effective regularization technique for preventing overfitting.

5.4 Data Augmentation Strategies: Data augmentation techniques aim to improve model generalization by augmenting the training data with diverse transformations. By exposing the model to a broader range of input variations during training, data augmentation helps the model generalize better to unseen conditions.

Common data augmentation techniques include random rotations, translations, flips, and adding noise to the input images. [15] Data augmentation can also help mitigate the effects of imbalanced or limited training data by generating synthetic examples that augment the available training samples.

Here's how random rotations as a data augmentation technique can be implemented for image processing using deep learning:

Implementation:

- Given an input image, a random rotation angle is sampled from a predefined range, such as [-10°, 10°].
- The image is then rotated by the sampled angle using an appropriate image processing library or framework, such as OpenCV or TensorFlow.
- After rotation, the image may need to be resized or padded to maintain the original dimensions, depending on the specific requirements of the deep learning model.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

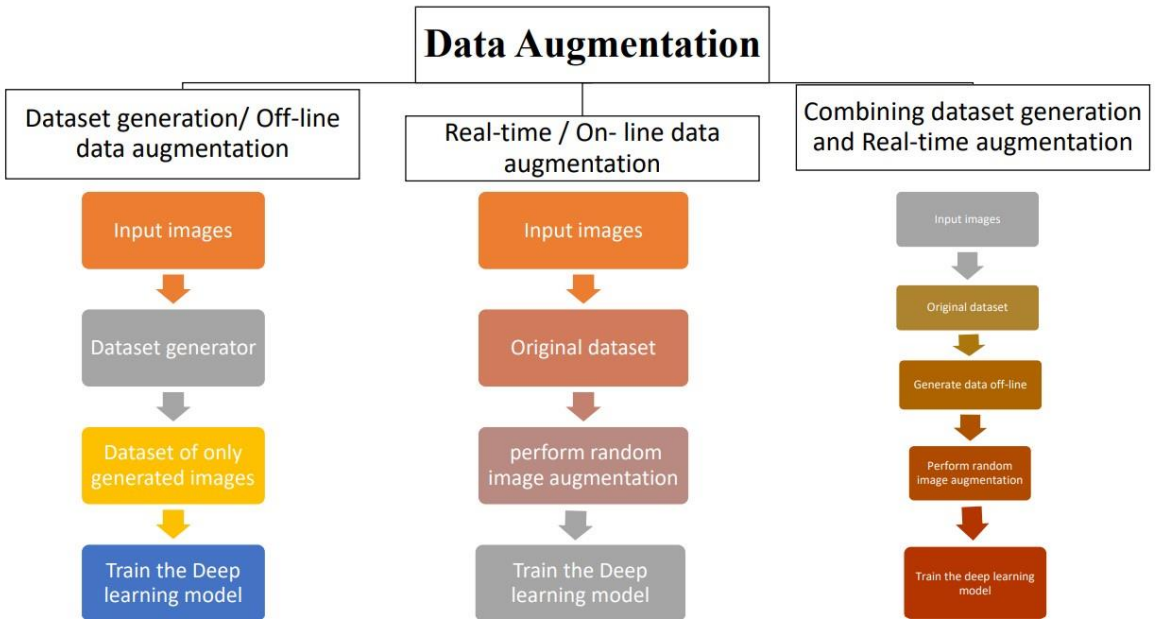

Figure 4 Data Augmentation Technique for Deep learning.

Application:

- Random rotations can be applied to the training data during the data loading or preprocessing phase before feeding the images into the deep learning model.
- By applying random rotations to the training data, the model learns to be invariant to variations in the orientation of objects, thereby improving its robustness to rotation invariance.

Benefits:

- **Introduces Variability:** Random rotations increase the diversity of the training data by introducing variations in the orientation of objects, making the model more robust to changes in rotation.
- **Mitigates Overfitting:** By exposing the model to a broader range of input variations during training, random rotations help prevent overfitting and improve generalization performance. [16]
- **Mimics Real-world Scenarios:** In real-world scenarios, images may be captured from different angles or orientations. By augmenting the training data with random rotations, the model learns to better handle such variations.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

Considerations:

- **Range of Rotation:** The range of rotation angles should be carefully chosen based on the specific characteristics of the dataset and the task at hand. Larger ranges may introduce more variability but can also lead to unrealistic transformations.
- **Interpolation Method:** When rotating images, an appropriate interpolation method should be used to ensure high-quality transformations. Common interpolation methods include bilinear or bicubic interpolation.
- **Performance Impact:** Applying random rotations during training may increase the computational overhead, particularly for large datasets or complex models. Therefore, it's important to balance the benefits of data augmentation with the computational resources available.

Overall, random rotations as a data augmentation technique provide a simple yet effective way to increase the diversity of the training data and enhance the robustness of deep learning models for image processing tasks. [17] By incorporating random rotations into the training pipeline, researchers and practitioners can improve the generalization performance of their models and better handle variations in the orientation of objects in real-world scenarios.

Enhancing the robustness and generalization of deep learning models requires a combination of approaches and techniques, including adversarial training, domain adaptation, regularization techniques, and data augmentation strategies. By leveraging these techniques, researchers aim to develop more reliable, adaptable, and generalizable models capable of handling diverse and challenging scenarios effectively in real-world applications.

6. Future Directions and Open Challenges: - As deep learning continues to advance, there are several future directions and open challenges for enhancing the robustness and generalization of models in image processing. Addressing these challenges will be crucial for deploying deep learning systems in real-world applications effectively. Here are some key areas of focus:

6.1 Adversarial Robustness: Despite significant progress in adversarial training and defense mechanisms, deep learning models remain vulnerable to sophisticated adversarial attacks. Future research should focus on developing more robust defense mechanisms that can withstand increasingly sophisticated attacks. [18] This may involve exploring new training strategies, designing more resilient architectures, and investigating the theoretical underpinnings of adversarial robustness.

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

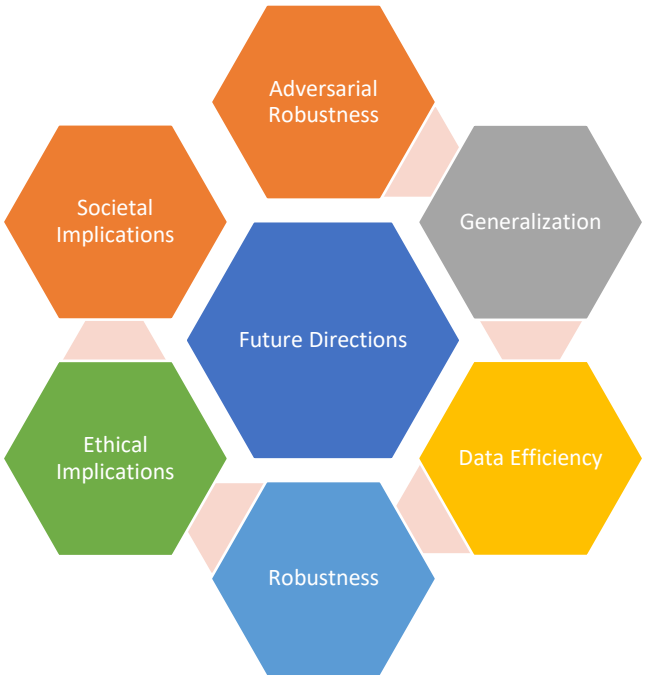

Figure 5 Future Directions for enhancing Robustness and Generalization.

6.2 Generalization Across Domains: Generalizing deep learning models across different domains remains a challenging problem, particularly when there is a significant distribution shift between the training and deployment environments. [19] Future research should focus on developing more effective domain adaptation techniques that can learn domain-invariant representations from limited labeled data. This may involve exploring unsupervised, semi-supervised, and transfer learning approaches that leverage domain-specific knowledge or priors.

6.3 Data Efficiency and Transfer Learning: Deep learning models often require large amounts of labeled data for training, which may be costly or impractical to obtain in certain domains. Future research should focus on improving data efficiency and transfer learning techniques that can leverage knowledge from pre-trained models or auxiliary tasks to improve performance on target tasks with limited labeled data. This may involve exploring new pre-training strategies, multi-task learning approaches, and domain adaptation techniques.

6.4 Robustness to Distribution Shifts: Deep learning models are often sensitive to distribution shifts in the input data, leading to performance degradation in real-world scenarios. Future research

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

should focus on developing more robust models that can adapt to changes in the input data distribution over time. This may involve exploring techniques such as continual learning, domain generalization, and meta-learning approaches that can improve the model's ability to generalize to unseen distributions.

6.5 Interpretability and Explainability: As deep learning models are increasingly deployed in safety-critical applications, there is a growing need for interpretable and explainable models that can provide insights into their decision-making process. [20] Future research should focus on developing interpretable deep learning models and explainability techniques that can provide human-understandable explanations for model predictions. This may involve exploring techniques such as attention mechanisms, feature visualization, and model distillation approaches.

6.6 Ethical and Societal Implications: As deep learning models become more ubiquitous in society, there are growing concerns about their ethical and societal implications. Future research should focus on addressing issues related to fairness, bias, privacy, and accountability in deep learning systems. This may involve developing new evaluation metrics, fairness-aware algorithms, and regulatory frameworks to ensure that deep learning systems are deployed responsibly and ethically.

7. Conclusion: - In conclusion, the pursuit of enhancing robustness and generalization in deep learning models for image processing represents a critical endeavor with far-reaching implications for various real-world applications. Throughout this paper, we have explored a myriad of approaches and techniques aimed at bolstering the reliability, adaptability, and performance of deep learning systems in the realm of image processing. From adversarial training to domain adaptation, regularization techniques to data augmentation strategies, the landscape of methodologies for enhancing robustness and generalization is vast and multifaceted. Adversarial training equips models with resilience against adversarial attacks, while domain adaptation techniques facilitate better generalization across diverse data distributions. Regularization methods mitigate overfitting and improve model generalization, while data augmentation enriches the training data with diverse variations, enabling models to better handle unforeseen conditions. However, as we look toward the future, several challenges and opportunities lie ahead. Adversarial robustness, domain adaptation across disparate environments, data efficiency, and ethical considerations are just a few areas ripe for exploration and innovation. Addressing these challenges will require interdisciplinary collaboration, novel algorithmic developments, and a commitment to responsible and ethical AI deployment.

In essence, the quest to enhance robustness and generalization in deep learning models for image processing is an ongoing journey marked by continuous innovation and discovery. By leveraging

Power System Technology
ISSN:1000-3673
Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

the insights and methodologies presented in this paper, researchers and practitioners can pave the way for the development of more reliable, adaptable, and trustworthy deep learning systems capable of addressing the diverse and evolving challenges of image processing in the years to come.

References: -

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572.

[2] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv preprint arXiv:1706.06083.

[4] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1), 2096-2030.

[5] Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2017). Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7167-7176).

[6] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

[7] Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. In Advances in Neural Information Processing Systems (pp. 950-957).

[8] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60.

[9] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2223-2232).

[10] Warde-Farley, D., & Goodfellow, I. J. (2016). Adversarial Perturbations of Deep Neural Networks. arXiv preprint arXiv:1412.6572.

[11] Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 770-778).

Power System Technology
ISSN:1000-3673

Received: 04-10-2023 Revised: 12-11-2023 Accepted: 10-12-2023

[13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211-252.

[14] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303-338.

[15] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In European Conference on Computer Vision (ECCV) (pp. 740-755).

[16] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene Parsing through ADE20K Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 633-641).

17] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[18] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 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 (KDD) (pp. 1135-1144).

[19] Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In European Conference on Computer Vision (ECCV) (pp. 3-19).

[20] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2020). Defense against Adversarial Attacks Using Feature Scattering-based Adversarial Training. arXiv preprint arXiv:2003.09311.