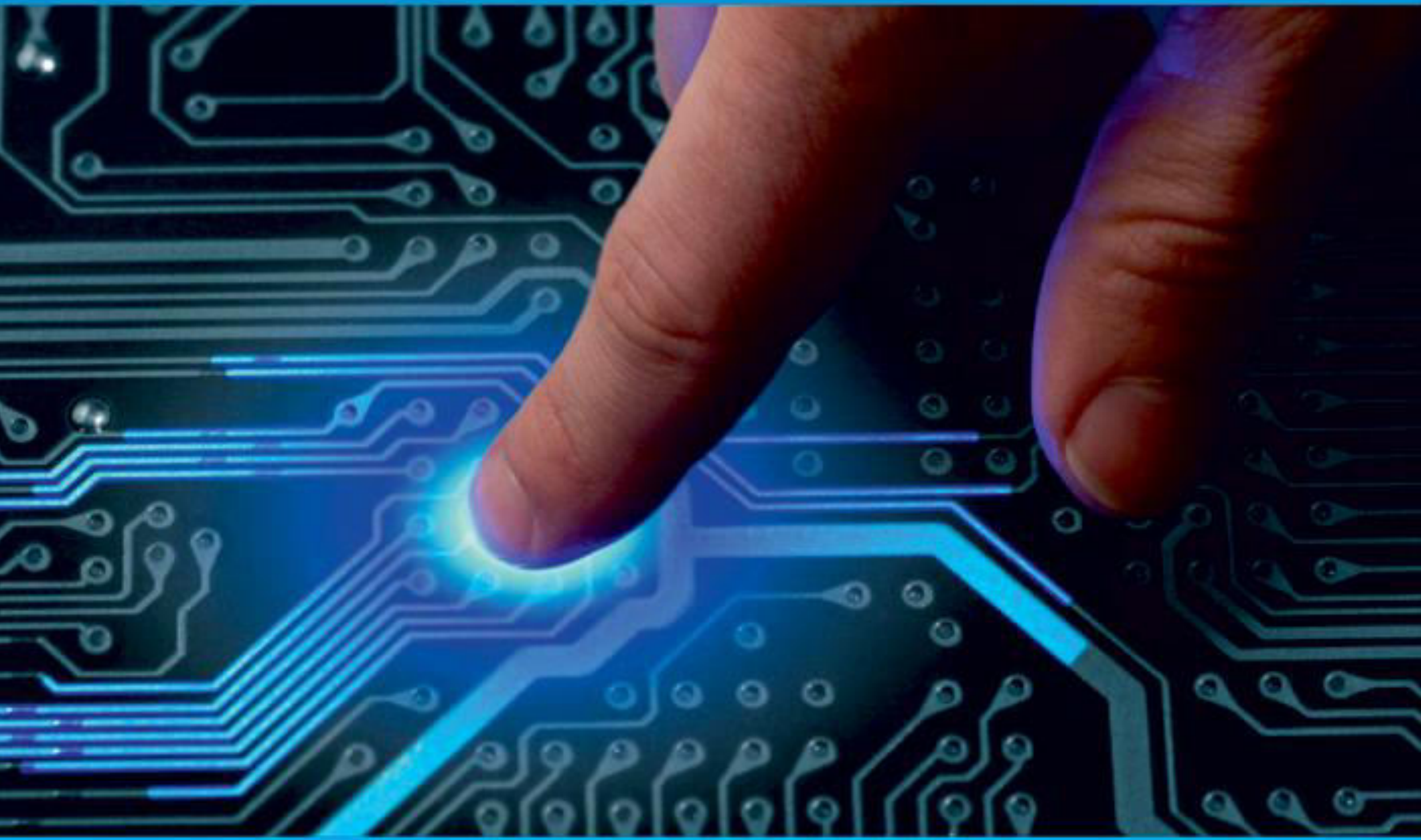
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 10, October 2024

ISSN
INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.625

Data-Augmentation Techniques for Vision Transformer-Based Medical Image Segmentation

Manjunatha A, Neelappa

Research Scholar VTU & Assistant Professor, Department of E & C, Government Engineering College, Hassan, Karnataka, India

Associate Professor, Department of E & C, Government Engineering College, Hassan, Karnataka, India

ABSTRACT: Medical image segmentation often encounters challenges due to the limited availability of labeled data, which can adversely affect model generalization and robustness. Vision Transformers (ViTs), known for its ability to capture global dependencies within images, offer a promising alternative to traditional convolutional neural networks (CNNs) in this field. ViTs have proven effective in modeling complex image structures by capturing long-range dependencies, an advantage particularly valuable in medical image analysis. However, these models are susceptible to overfitting, especially in scenarios with limited data, making data augmentation a crucial strategy. While traditional augmentation methods have been widely used in CNNs, they may not directly transfer effectively to transformers due to structural differences between these architectures. ViTs benefit more from augmentation techniques that introduce additional complexity without disrupting spatial coherence. This study is designed to systematically evaluate a variety of augmentation techniques specifically suitable for ViT based segmentation models, with the goal of identifying methods that most effectively enhance model generalization and robustness in medical imaging tasks.

KEYWORDS: Augmentation, Cutout, CutMix, MixUp, ViT

I. INTRODUCTION

Simple image data augmentation techniques such as flipping, random cropping and random rotation are commonly used to train large models and these methods generally perform well across many datasets and problem types. However, in real-world scenarios, substantial shifts in data distribution can occur. Are our models truly robust to such data shifts and corruption?

Currently, models do not generalize well to shifts in data distribution. If models could accurately identify when they are likely to make mistakes, or if they could reliably estimate uncertainty in their predictions, it could help mitigate the effects of this fragility. In the field of deep learning, particularly in computer vision, models often learn to focus on the most discriminative features of an image, sometimes at the cost of overlooking less distinctive but essential features. To address this, regional dropout strategies have emerged as a class of data augmentation techniques designed to improve model generalization, robustness and attention to detail. Regional dropout strategies involve removing or altering specific regions of an image during training, forcing the model to learn from a broader set of features instead of relying solely on dominant characteristics. This concept extends the traditional dropout technique, which randomly removes a subset of neuron activations within the network, to the input image space. By selectively obscuring parts of the image, these strategies encourage the model to make predictions based on a variety of visual cues, resulting in a more robust and adaptable feature extraction process.

While regional dropout strategies have shown some improvements in classification and localization performance, they typically involve zeroing out or filling deleted regions with random noise, which significantly reduces the amount of informative data available in each training image. This poses a challenge, as CNNs generally require large amounts of data to perform well. So, how can we make better use of these removed regions while still benefiting from the generalization and localization advantages that regional dropout offers?

To address this question, augmentation techniques like CutMix and Cutout were introduced. In CutMix, instead of simply removing pixels, the masked region is replaced with a patch from a different image. The ground truth labels are also adjusted proportionally based on the area of the combined images. This allows the model to learn from two sources of information in a single sample, maximizing the use of available pixels while encouraging it to focus on diverse visual features. CutMix shares some similarities with MixUp, which also blends two samples by interpolating both the images and their labels. However, while MixUp enhances classification performance, the resulting mixed images can sometimes appear unnatural, potentially confusing the model in certain cases. By contrast, CutMix maintains more realistic image compositions, which can make it more effective for tasks requiring strong localization and generalization.

II. RELATED WORK

Regional Dropout: Methods that involve removing random regions in images [3,4] have been proposed to enhance the generalization capabilities of CNNs. Certain object localization techniques [5,2] also utilize regional dropout strategies to improve the localization abilities of CNN models. CutMix[21] is similar to these approaches but introduces a key difference: instead of leaving removed regions blank, CutMix fills them with patches from other training images. Another approach, DropBlock [6], extends the regional dropout concept into the feature space, showing improved generalization as well.

Synthesizing Training Data: Some studies have explored the synthesis of training data to improve model generalizability. For instance, Stylizing ImageNet [7,8] focuses the model more on shape rather than texture, resulting in better performance in both classification and object detection tasks. Similarly, CutMix generates new training samples by cutting and pasting patches within mini-batches, which enhances performance across various computer vision tasks [8]. In object detection, object insertion techniques [10,9] have been developed to synthesize objects in backgrounds, aiming to create well-represented samples of individual objects, while CutMix generates combined samples that may contain multiple objects.

Mixup: CutMix shares similarities with Mixup [11], as both techniques combine two samples by blending both the images and labels, with the new sample label being a linear interpolation of the original labels. However, Mixup can produce locally ambiguous and unnatural samples, which may confuse the model, especially for localization tasks. Recent Mixup variations [12, 13, 14] have attempted feature-level interpolation and other transformations. However, these studies generally lack in-depth analysis, particularly concerning localization and transfer-learning performance. We observed that CutMix offers benefits across a wide range of tasks, including image classification, localization and transfer learning.

Training Techniques for Deep Networks: Efficient training of deep networks is a fundamental challenge in computer vision, as these models require significant computational power and data. Techniques such as weight decay, dropout [16] and batch normalization [17] are widely used to enhance the efficiency of deep network training. Recently, methods involving the addition of noise to the internal features of CNNs [18, 19] or incorporating additional pathways in the architecture [20, 21] have been introduced to boost image classification performance. Unlike these internal methods, CutMix operates directly on the data, modifying the training images themselves without changing the network's internal structure or architecture [21].

III. DATA-AUGMENTATION METHODS SPECIFIC TO VISION TRANSFORMERS

In this section, we explore three advanced data augmentation techniques like Cutout, CutMix and Mixup which are designed to enhance deep learning model generalization in medical image segmentation.

A) Cutout Augmentation Technique:
Cutout is a simple yet effective data augmentation technique aimed at improving the generalization of convolutional neural networks (CNNs). This technique involves masking out random, square regions of the input image and creating "occluded" training samples that encourage the model to rely on more contextual information rather than specific, highly discriminative features. Cutout is presented as an extension of dropout but applied to the input space rather than hidden layers, promoting a more comprehensive understanding of the entire image.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

One of the most common uses of noise to improve model accuracy is Dropout, which stochastically(randomly) drops neuron activations during training and discourages the co-adaptation of feature detectors. Dropout tends to work well for fully connected layers but lacks that regularization power for convolutional layers, because of two reasons:

- Convolutional layers require less regularization since they have much fewer parameters than fully-connected layers.
- The second factor is that neighboring pixels in images share much of the same information.

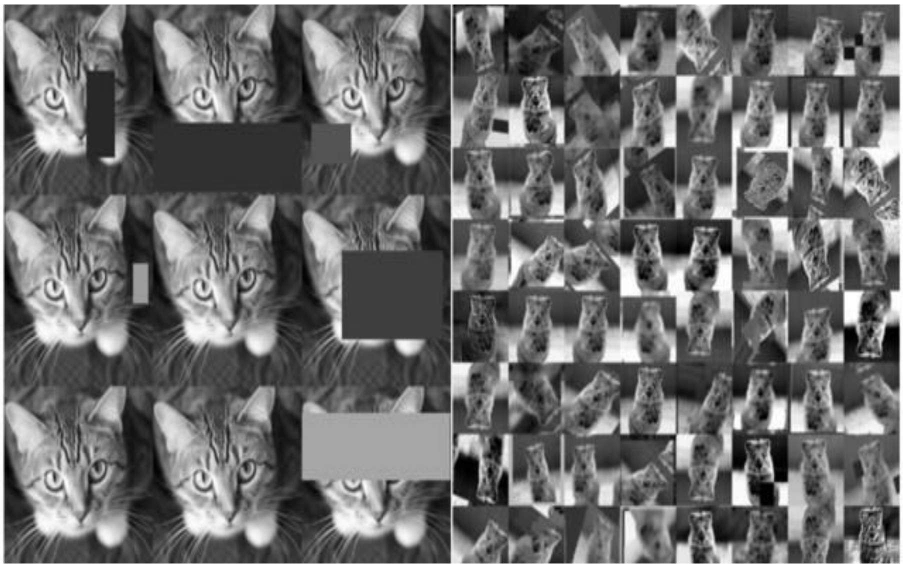


Fig. 1. An example of applying the Cutout augmentation to an image.

The Cutout augmentation technique, as shown fig.1involves masking or "cutting out" random square sections from an image, essentially replacing these regions with a solid color. This technique aims to make the model less sensitive to specific features or regions, thereby encouraging it to focus on the broader patterns and contextual information within the image. Unlike traditional dropout, which targets neuron activations in hidden layers, Cutout applies to the input layer of CNNs. By zeroing out contiguous regions of the input, Cutout forces the model to learn from the remaining visible regions, enhancing its robustness to partial occlusions or missing information in real-world scenarios.

Methodology: To implement Cutout, a fixed-size square mask is randomly applied to each image during training. The process is computationally inexpensive and can run parallel to other augmentation steps.

B) Mix-Up Augmentation Technique

Mix-up is a data augmentation method that generates synthetic training examples by combining two images from the dataset. This technique works by linearly blending the pixel values of two images along with their labels, creating a new, hybrid image. The mathematical representation for Mix-up is:

$\tilde{x}= \lambda x_i + (1- \lambda)\ x_j$

$\tilde{y}= \lambda y_i + (1- \lambda)\ y_{j;}$ where $\tilde{x}$ & $\tilde{y}$ represent the synthetic image and label, x_i and x_j are the original images, y_i and y_j are their labels and λ is a random value between 0 and 1 that controls the mix ratio.

Mix-up has been shown to enhance the generalization and robustness of image classification models by encouraging them to learn smoother decision boundaries. However, it can sometimes produce unrealistic image outputs with ambiguous labels, limiting its effectiveness in tasks requiring precise localization, such as object detection and segmentation.

Other image mixing techniques include CutMix, which creates new training samples by cutting out patches from one image and pasting them onto another. This approach maintains more of the visual structure of the original images, often making it better suited for localization-based tasks where spatial information is crucial.

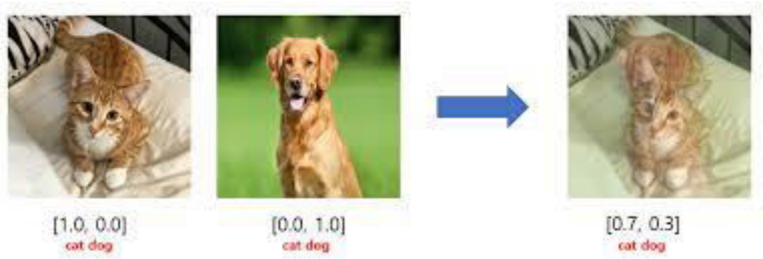


Fig.2. Mixup example.

C) CutMix Augmentation Technique

CutMix creates new training samples by cutting a rectangular patch from one image and pasting it onto another. The labels are also mixed according to the area of the patch, promoting the model to learn from partial views and less discriminative object parts. This strategy aims to overcome the limitations of previous techniques like Cutout and Mixup. While Cutout loses information by masking parts of the image and Mixup introduces unrealistic pixel interpolations, CutMix preserves more natural features by replacing parts of one image with another.

CutMix Methodology:

The method involves generating a binary mask to define a patch area on one image, removing that section and filling it with a patch from another image. Mathematically, the mixed sample ($\tilde{x}$, $\tilde{y}$) is computed as:

$\tilde{x} = M\odot x_A + (1\text{-}M)\ x_B$

$\tilde{y} = \lambda yA + (1\text{-}\lambda)\ y_B$; where M is a binary mask, λ controls the mix ratio, and the two images (x_A, x_B) and labels (y_A, y_B) are combined accordingly.

While Mixup blends the entire images and labels linearly, it often produces unnatural image samples whereas CutMix achieves better localization and maintains spatial coherence by replacing regions instead.

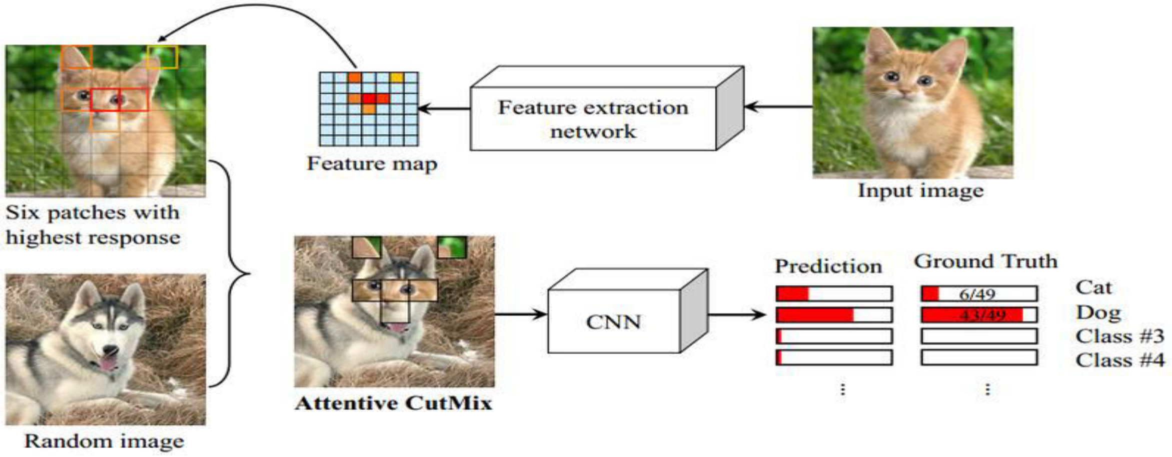


Fig.3. The CutMix image augmentation procedure, attentive regions from one image are overlayed on another image

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Comparison between Cutout, Mixup and CutMix

The table.1. Gives the summarised comparison between cutout, mixup and cutmix augmentation methods.

Table. 1. Comparison between Cutout, Mixup and CutMix

Feature/Aspect	Cutout	Mixup	CutMix
Augmentation Technique	Masks a rectangular region with black or noise	Blends two images by linear interpolation	Replaces a rectangular patch with one from another image
Image Naturalness	Moderate and some information loss	Low and produces blended, unrealistic images	High Naturalness, retains spatial coherence by combining patches
Label Mixing	No label mixing, original label retained	Linear interpolation of labels	Label mixing proportional to patch area
Robustness to Occlusion	High, trained to handle missing parts	Moderate, though not specifically designed for occlusion	Moderate, though not specifically designed for occlusion
Computational Overhead	Low, simple masking	Low, simple blending	Low, patch replacement adds minimal computation
Weaknesses/Drawbacks	Information loss in masked regions	Unnatural samples, less effective for tasks needing spatial precision	Potential label ambiguity in mixed regions

IV. EXPERIMENTAL SETUP AND RESULTS

The experiments were conducted on a brain MRI dataset, focusing on tumor segmentation. This dataset presents unique challenges due to the intricate structures and variability in tumor appearances, making it suitable for evaluating augmentation techniques effectiveness in improving model robustness.

Vision Transformer based Architecture, pre-trained with ImagNet and fine-tuned to the brain MRI data set was used to carryout image segmentation. The performance of Vision Transformer (ViT) models with each augmentation technique applied was summarised and tabulated. We used the Dice coefficient, IoU, and Hausdorff distance as primary evaluation metrics to quantify segmentation performance.

Table 2. Performance comparison

Dataset	Augmentation Technique	Dice Coefficient	IoU	Hausdorff Distance
Brain MRI	Baseline (ViT alone)	0.75	0.68	12.4
	Baseline + Rotation	0.78	0.72	11.5
	Baseline + Cutout	0.80	0.74	10.2
	Baseline + MixUp	0.79	0.73	10.7
	Baseline + CutMix	0.82	0.76	9.8

The results indicate that ViT models benefit significantly from complex augmentations like CutMix, which achieved the highest Dice scores. Cutout also contributed a substantial improvement over the baseline.

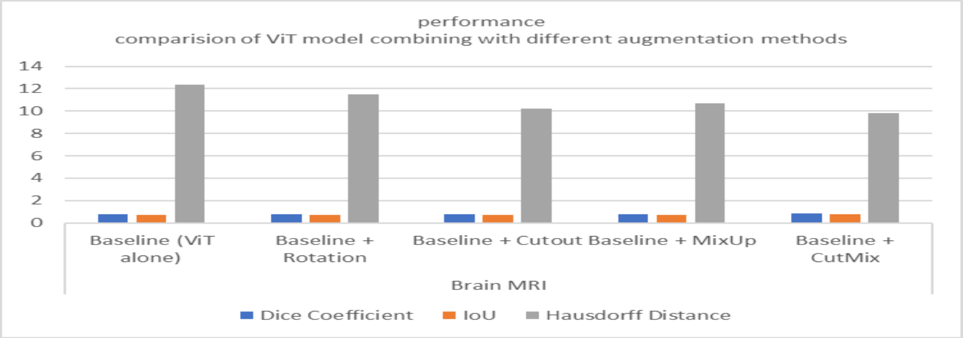


Fig.4. Performance comparison chart

V. CONCLUSION

This study explored the application of advanced data augmentation techniques like Cutout, CutMix and MixUp in enhancing Vision Transformer (ViT) performance on medical image segmentation tasks. Vision Transformers, known for their capacity to capture long-range dependencies, show promising potential in medical image segmentation but are vulnerable to overfitting, especially with small datasets. The experimental results demonstrated that ViT models benefited from augmentation methods. Among the techniques evaluated, CutMix yielded the highest improvements in performance metrics like Dice coefficient and IoU, enhancing segmentation accuracy by promoting better localization and generalization. Cutout also showed significant gains by encouraging the model to focus on diverse visual patterns. MixUp, although effective for enhancing robustness, occasionally introduced label ambiguity due to the blending of images, which makes it less ideal for tasks demanding high spatial accuracy. In conclusion, this study highlights that appropriate augmentation techniques, particularly CutMix, can considerably boost ViT based medical image segmentation.

REFERENCES

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolu- tion, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018

[2] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2219–2228, 2019.

[3] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

[4] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. arXiv preprint arXiv:1708.04896, 2017.

[5] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 3544–3553. IEEE, 2017.

[6] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In Advances in Neural Information Processing Systems, pages 10750–10760, 2018

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, San- jeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231, 2018.

[9] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In Proceedings of the European Conference on Computer Vision (ECCV), pages 364–380, 2018.

[10] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance de- tection. In Proceedings of the IEEE International Confer- ence on Computer Vision, pages 1301–1310, 2017.

[11] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. “mixup: Beyond empirical risk minimization”. arXiv preprint arXiv:1710.09412, 2017.
[12] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. “Manifold mixup: Better representations by interpolating hidden states”. In International Conference on Machine Learning, pages 6438–6447, 2019.
[13] Cecilia Summers and Michael J Dinneen. “Improved mixed- example data ugmentation. In 2019 IEEE Winter Confer- ence on Applications of Computer Vision (WACV)”, pages 1262–1270. IEEE, 2019.
[14] Hongyu Guo, Yongyi Mao, and Richong Zhang. “Mixup as locally linear out-of- manifold regularization”. arXiv preprint arXiv:1809.02499, 2018.
[15] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. “Ricap: Random image cropping and patching data augmentation for deep cnns”. In Asian Conference on Machine Learning, pages 786–798, 2018.
[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. “Dropout: A simple way to prevent neural networks from overfitting”. Journal of Machine Learning Research, 15:1929–1958, 2014.
[17] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In ICML, 2015.
[18] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. “Deep networks with stochastic depth”. In ECCV, 2016.
[19] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. “Gather-excite: Exploiting feature context in convo- lutional neural networks”. In Advances in Neural Information Processing Systems, pages 9423–9433, 2018.
[20] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In arXiv:1709.01507, 2017.
[21] Sangdoo Yun1 et al., “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features”,
[22] Shorten, C., & Khoshgoftaar, T. M. “A Survey on Image Data Augmentation for Deep Learning”. Journal of Big Data, 6(1), 60.
[23] Perez, L., & Wang, J. “The Effectiveness of Data Augmentation in Image Classification using Deep Learning”. arXiv preprint arXiv:1712.04621.
[24] Zhao, T., Liu, Z., Ding, Y., et al. “Data Augmentation for Medical Image Segmentation with One-Class Per Pixel Analysis”. IEEE Access, 7, 156631-156641.
[25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. International Conference on Learning Representations (ICLR).
[26] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. “Medical Transformer: Gated Axial-Attention for Medical Image Segmentation”. Medical Image Analysis, 75, 102306.
[27] Terrance DeVries and Graham W. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout”, arXiv preprint arXiv:1708.04552, 2017
[28] Mikołajczyk, A., & Grochowski, M. “Data Augmentation for Improving Deep Learning in Image Classification Problem”. International Interdisciplinary PhD Workshop (IIPhDW), 117-122.
[29] https://figshare.com/articles/dataset/brain_tumor_dataset/1512427.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details