HEURISTIC DROPOUT: AN EFFICIENT REGULARIZATION METHOD FOR MEDICAL IMAGE SEGMENTATION MODELS

Dachuan Shi[1] *Ruiyang Liu*[2] *Linmi Tao*[2⋆] *Chun Yuan*[1⋆]

[1]Shenzhen International Graduate School, Tsinghua University, China
[2]Department of Computer Science and Technology, BNRist, Tsinghua University, China
[2]Key Laboratory of Pervasive Computing, Ministry of Education, China

ABSTRACT

For medical image segmentation in a real scenario, the amount of accurate annotation data at the pixel level is typically small, which tends to cause an overfitting problem. This manuscript goes deep into the research of the Dropout algorithm, which is commonly used in neural networks to alleviate the overfitting problem. From the perspective of solving the co-adaptation problem, this manuscript explains the basic principles of the Dropout algorithm and discusses the existing limitations of its derivative methods. Furthermore, we propose a novel Heuristic Dropout algorithm to address these limitations. The proposed algorithm takes information entropy and variance as heuristic rules. It guides our algorithm to drop features suffering from co-adaptation problem more efficiently and thus can better alleviate the overfitting problem of small-scale medical image segmentation datasets. Experiments on medical image segmentation datasets and models show that the proposed algorithm significantly improves the performance of these models.

***Index Terms*—** Medical image segmentation, Overfitting problem, Dropout algorithm, Information entropy

1. INTRODUCTION

As a fundamental part of current computer-aided medical diagnosis (CAD) systems, the accuracy of medical image segmentation greatly affects the performance of CAD systems. In recent years, CAD systems have become more and more involved in real medical diagnosis tasks. Therefore, it is of great importance and application value to improve the accuracy and reliability of medical image segmentation models.

In the field of medical image segmentation, deep learning models such as U-Net [1], nnU-Net [2], TransUNet [3] have already demonstrated their better performance in various tasks than traditional methods [4, 5, 6, 7]. Compared to the natural image segmentation, the data annotation of the medical image segmentation is highly dependent on expert knowledge and requires pixel-level accurate annotation. Therefore, the amount of accurate annotation data at the pixel level with expert guidance is typically small. The small-scale datasets tend to cause the overfitting problem, especially when the amount of segmentation model parameters are large.

There are various methods [8, 9, 10] to tackle the overfitting problem, and Dropout algorithm [11] is one of these methods that is simple but effective. It randomly drops neurons in the model with a certain probability during the training process, which mitigates the co-adaptation problem [11] and thus alleviates the overfitting problem of deep learning models. Co-adaptation [11] refers to the phenomenon that features learned by each neuron often have to be combined with contexts, i.e., other specific neurons, to provide helpful information during the training process. However, such empirical dependencies learned from small-scale medical image segmentation datasets are fragile and may not be trustworthy when faced with the distribution of the test set. Therefore too many dependencies among neurons tend to trigger the overfitting problem. Drop operation in the Dropout algorithm reduces the dependencies among neurons in deep learning models and prevents some neurons from being overly dependent on other neurons, thus avoiding the overfitting problem to some extent.

According to whether the drop process is completely random or not, the derivative methods of the Dropout algorithm can be divided into two categories. The first category is completely random methods, such as Spatial Dropout [12] which randomly drops the unit in the channel dimension, DropBlock [13] which considers 2d blocks as units and randomly drop them, and Stochastic Depth [14] which randomly drops residual connections. The second category is rule-based methods, such as Weighted Channel Dropout [15] which takes the activation value of channels as the guiding rule, Focused Dropout [16] which takes the activation value of 2d blocks as the guiding rule. However both categories of existing methods are not without limitations. The first category, completely random methods, lacks guiding rules and thus could be inefficient that dropped features are not necessarily the ones suffering

⋆ Corresponding authors. This work is supported by the Science and Technology Innovation 2030 - "New Generation Artificial Intelligence" Project of the Ministry of Science and Technology of China. Project No. 2021ZD0113800.

from the co-adaptation problem. Among the second category, rule-based methods, the existing guiding rules are not accurate enough, and the efficiency of dropping exact features suffering from the co-adaptation problem still has room for improvement. Therefore, this manuscript proposes a novel guide rule combining information entropy and variance. Guided by this rule, the efficiency of the proposed algorithm to drop features suffering from the co-adaptation problem is further improved. Experiments on several medical image segmentation datasets and models show that the proposed algorithm improves the model accuracy significantly.

2. METHODOLOGY

We propose a novel Heuristic Dropout algorithm that uses information entropy and variance as guiding rules to execute the drop operation. The proposed algorithm can efficiently drop features suffering from co-adaptation and thus can mitigate the overfitting problem of medical image segmentation tasks to a great extent.

2.1. Heuristic Metric

In order to efficiently drop features suffering from more severe co-adaptation, we take information entropy as a heuristic rule. Information entropy can measure the uncertainty of a distribution.

$$H(x) = -\sum_{x \in X} p(x) log p(x)$$

where X is a random variable, $p(x)$ is the probability density function, and $H(x)$ is the information entropy about the random variable X.

As illustrated in Figure 7 of Dropout manuscript [11], features suffering from severe co-adaptation have uncertain visual meaning and thus have a high value of information entropy, while features suffering from slight co-adaptation have certain visual meaning, e.g., looking like points, edges or geometric contours of objects, and these features have a low value of information entropy. Taking information entropy as a guiding rule, we drop features suffering from severe co-adaptation problems with a higher probability. Furthermore, we also need variance as another heuristic rule. Considering an extreme case, when the distribution is close to a constant distribution, it is known that the information entropy will close to a minimum. However, features with a constant distribution provide little helpful information for the training. Therefore, take variance as another heuristic rule, we drop features that are closer to a constant distribution with a greater probability.

2.2. Heuristic Dropout Algorithm

Combining the two heuristic rules of information entropy and variance, Algorithm 1 is obtained.

Algorithm 1 Heuristic Dropout

Input: Output activations of a previous layer that contains c channels $A = [a_1, a_2, ..., a_c]$, drop rate p, k
Output: Activations after dropout $A^\star$

if *mode = Inference,* **then**
  Return $A^\star = A$
end
foreach a_i *in A* **do**
  Calculate information entropy of a_i by Algorithm 2 as e_i;
  Calculate variance of a_i as v_i;
  Calculate heuristic metric $m_i \leftarrow e_i + \frac{k}{v_i + \epsilon}$, where k is a hyperparameter, ϵ is a smoothing number;
end
Get rearranged A' by descending sort according to m_i;
for $j = 1; j \leq p \times c$ **do**
  $a_j^\star \leftarrow a_j' \otimes mask$
end
Return $A^\star$

We calculate the information entropy e_i and the variance v_i for each channel of input feature maps. We use $e_i + \frac{k}{v_i + \epsilon}$ as the guiding rule. Because the values of feature maps are continuous distributions, we have first to quantize the values and then calculate the information entropy based on the histogram as shown in the Algorithm 2. It is also found that using a 3×3 Laplace filter instead of an all-zero filter as the drop mask will bring a little boost to the model performance.

Algorithm 2 Information Entropy for Quantized features

Input: A certain channel a_i of A, number of bins b
Output: Information Entropy e_i

Standardization by $tanh$, $a_i \leftarrow tanh(a_i)$;
Quantize by $round$, $a_i \leftarrow round(a_i \times b)$;
Calculate the corresponding histogram h_i according to the quantized a_i and bin number b;
Calculate information entropy e_i according to the definition, $e_i \leftarrow -\sum_{p(x) \in h_i} p(x) log p(x)$
Return e_i

Our algorithm can be seamlessly inserted into various models. Taking U-Net architecture as an example, we insert the proposed algorithm between the two successive convolutional layers in each stage of U-Net's encoder and decoder, i.e., after the activation function of the previous convolutional layer and exactly before the next convolutional layer.

3. RESULTS

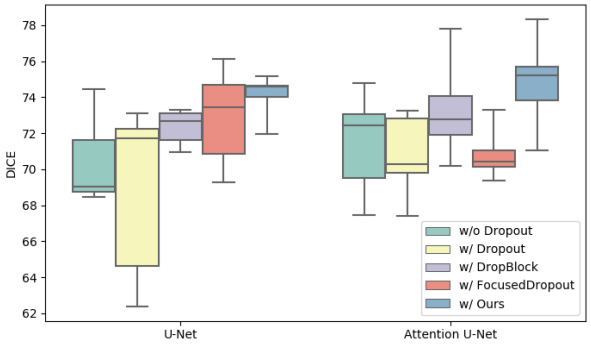
3.1. Datasets

We conduct experiments on the Pancreas-CT Dataset [19] and BAGLS Dataset [20]. Considering that in a real application environment, training samples labeled by experienced experts

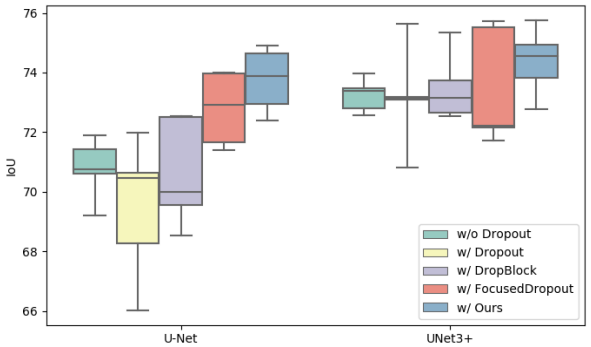
Method	Pancreas-CT Dataset				BAGLS Dataset			
	U-Net [1]		Attention U-Net [17]		U-Net [1]		UNet3+ [18]	
	Dice (↑)	Δ (↑)	Dice (↑)	Δ (↑)	IoU (↑)	Δ (↑)	IoU (↑)	Δ (↑)
Baseline [1, 17, 18]	70.40 ± 2.30	0	71.46 ± 2.62	0	70.78 ± 0.91	0	73.25 ± 0.50	0
+ Dropout [11]	68.82 ± 4.42	-1.58	70.7 ± 2.14	-0.76	69.48 ± 2.10	-1.30	73.17 ± 1.53	-0.08
+ DropBlock [13]	72.33 ± 0.91	+1.93	73.36 ± 2.56	+1.90	70.63 ± 1.10	-0.15	73.47 ± 1.78	+0.22
+ FocusedDropout [16]	72.88 ± 2.50	+2.48	70.85 ± 1.35	-0.61	72.79 ± 1.62	+ 2.01	73.49 ± 1.02	+0.24
+ Ours	**74.07 ± 1.11**	**+3.67**	**74.83 ± 2.39**	**+3.37**	**73.75 ± 0.96**	**+2.97**	**74.37 ± 1.01**	**+1.12**

Table 1. Quantitative and overall comparison results of comparing the proposed algorithm and other Dropout derivative methods on the Pancreas-CT dataset and BAGLS Dataset.

are very few in general, we randomly extracted a subset of these datasets respectively to carry out our experiments. For the Pancreas-CT Dataset, we randomly select 12 scans, which are then converted into 2545 512×512 2D slices for the convenience of training models. Among these 12 3D CT scans, we randomly select 8 scans as the training set, 2 scans as the validation set, and 2 scans as the test set. For the BAGLS Dataset, we randomly select 3000 slices as the training set, while the size of the validation and test sets remain the same as the original setting.



(a) On the Pancreas-CT dataset.



(b) On the BAGLS Dataset.

Fig. 1. Box plots of the comparison on the Pancreas-CT Dataset and BAGLS Dataset.

3.2. Evaluation Metrics

For quantitative analysis of experimental results, we adopt DICE value and IoU value which are widespread used in the field of medical image segmentation as our evaluation metric:

$$DICE = 2\frac{|X \cap Y|}{|X| + |Y|}, \quad IoU = \frac{|X \cap Y|}{|X \cup Y|}$$

where X denotes the mask outputted from models, and Y denotes the ground truth corresponding to the input image.

3.3. Experimental Settings

We train all models using an Adam optimizer with learning rate 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The batch size is set as maximum values that can be executed on a GeForce RTX 2080 Ti with mixed precision. We use CrossEntropy and train 100 epochs on the Pancreas-CT Dataset. We use a hybrid loss function combining CrossEntropy, DiceLoss, and SSIMLoss [21] and train 30 epochs on the BAGLS Dataset. We use standard data augmentation on the training datasets. No post-processing on the output results of models is conducted. We independently repeat all comparison experiments 5 times and report the average results.

3.4. Comparison with Dropout Derivative Methods

To verify the effectiveness of the proposed algorithm, we conduct experiments on the Pancreas-CT dataset and BAGLS Dataset. We add our algorithm and other Dropout derivative methods [11, 13, 16] into several models [1, 17, 18]. Figure 1 demonstrate the box plots of the experimental results and Table 1 shows the quantitative and overall comparison.

The experimental results demonstrate that our algorithm outperforms other Dropout derivative methods on both datasets. On the Pancreas-CT Dataset, our algorithm gains 3.67 and 3.37 improvement of DICE value for U-Net and Attention U-Net, respectively. On the BAGLS Dataset, our algorithm gains 2.97 and 1.12 improvement of IoU value for U-Net and UNet3+, respectively. It is convinced that our algorithm can improve the performance of medical image segmentation models more effectively.

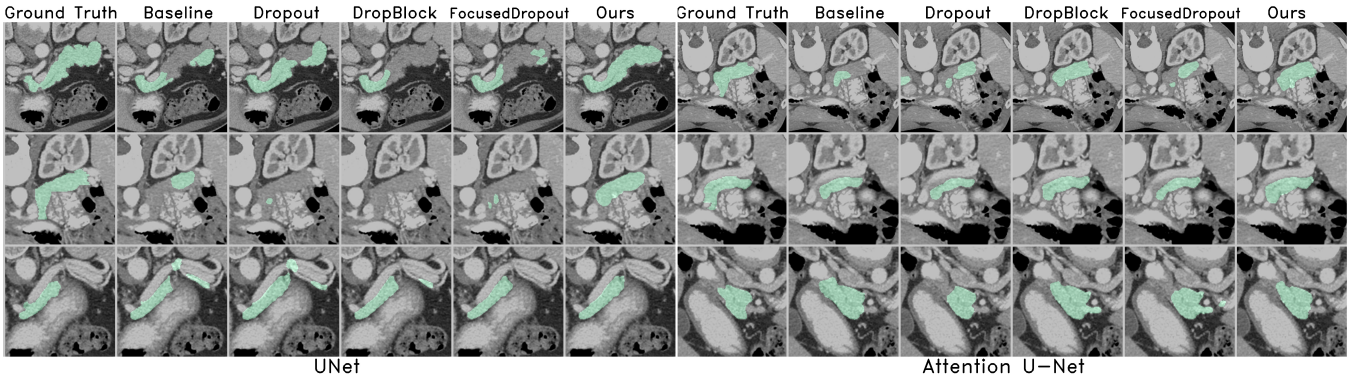


Fig. 2. Visual comparison results while using our algorithm and other Dropout derivative methods.

3.5. Comparison Study on Hyperparameter k

Based on U-Net and Pancreas-CT Dataset, We investigate the effect of hyperparameter k. With the k increasing, the performance tends to increase and then decrease, and the best performance is achieved when k is 3. Besides, from the variance of the box is clear that when k is 2, the model performance is more stable and predictable than that when k is 3.

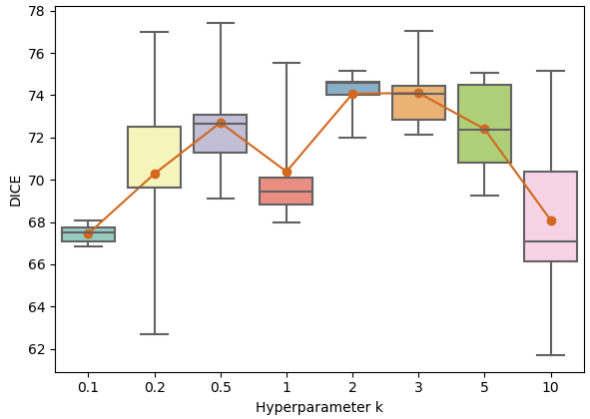


Fig. 3. Performance given by different hyperparameter k.

3.6. Verify Effectiveness of Alleviating Co-adaptation

To verify that our algorithm alleviates co-adaptation more effectively than traditional Dropout, we randomly mask off a certain percentage of intermediate feature maps before the final output layer of U-Net, on the Pancreas-CT dataset. For models suffering from less co-adaptation, the performance decrease due to masked features should be smaller because there are fewer dependencies among their features. It is shown in Figure 4 that the performance decrease of our algorithm after masking is significantly smaller than the traditional Dropout. The results demonstrate that by using our algorithm, more independent features with fewer dependencies can be learned, and thus our algorithm can alleviate co-adaptation to a greater extent than the traditional Dropout algorithm.

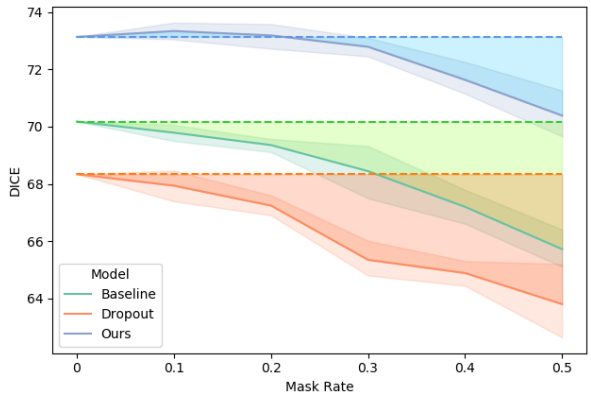


Fig. 4. Performance after masking off features. The colored area between a dashed line and a solid line represents performance decrease as the mask rate increases.

3.7. Visualization of Segmentation Results

The Figure 2 demonstrates the visualization for qualitative analysis. Segmentation results of three slices are displayed from top to bottom. The visualization figure indicates that our algorithm enables models to segment more accurately.

4. CONCLUSION

Our work suggests a novel Heuristic Dropout algorithm to address the overfitting problem for small-scale medical image segmentation datasets. Taking information entropy and variance as heuristic rules, our algorithm alleviates the co-adaptation phenomenon more effectively and thus better mitigates the overfitting problem. Experiments on several datasets and models show the superior performance of our algorithm. Furthermore, we will investigate the compatibility of our algorithm with natural images in future work.

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[4] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[5] Yuri Y Boykov and M-P Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*. IEEE, 2001, vol. 1, pp. 105–112.

[6] Tony F Chan and Luminita A Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[7] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[8] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[9] Song Han, Jeff Pool, John Tran, and William J Dally, "Learning both weights and connections for efficient neural networks," *arXiv preprint arXiv:1506.02626*, 2015.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[11] 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*, vol. 15, no. 56, pp. 1929–1958, 2014.

[12] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christopher Bregler, "Efficient object localization using convolutional networks," 2015.

[13] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le, "Dropblock: A regularization method for convolutional networks," 2018.

[14] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger, "Deep networks with stochastic depth," 2016.

[15] Saihui Hou and Zilei Wang, "Weighted channel dropout for regularization of deep convolutional neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8425–8432.

[16] Tianshu Xie, Minghui Liu, Jiali Deng, Xuan Cheng, Xiaomin Wang, and Ming Liu, "Focuseddropout for convolutional neural network," *arXiv preprint arXiv:2103.15425*, 2021.

[17] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[18] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[19] Holger R Roth, Amal Farag, Evrim B Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers, "Data from pancreas-ct," *The Cancer Imaing Archive*, 2016.

[20] Pablo Gómez, Andreas M Kist, Patrick Schlegel, David A Berry, Dinesh K Chhetri, Stephan Dürr, Matthias Echternach, Aaron M Johnson, Stefan Kniesburges, Melda Kunduk, et al., "Bagls, a multihospital benchmark for automatic glottis segmentation," *Scientific data*, vol. 7, no. 1, pp. 1–12, 2020.

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.