한국정보전자통신기술학회논문지(jkiiect)'20-10, Vol.13 No.5
논문 20-13-05-313 http://dx.doi.org/10.17661/jkiiect.2020.13.5.313

데이터 증강을 이용한 혀 영역 분할 성능 개선

진 홍*, 정성태**

Enhancement of Tongue Segmentation by Using Data Augmentation

Hong Chen*, Sung-Tae Jung**

요 약 많은 양의 데이터는 딥 러닝 모델의 견고성을 향상시키고 과적합 문제를 방지할 수 있게 해준다. 자동 혀 분할에서, 혀 영상 데이터 세트를 실제로 수집하고 라벨링하는 데에는 많은 어려움이 수반되므로 많은 양의 혀 영상 데이터를 사용하기 쉽지 않다. 데이터 증강은 새로운 데이터를 수집하지 않고 레이블 보존 변환을 사용하여 학습 데이터 세트를 확장하고 학습 데이터의 다양성을 증가시킬 수 있다. 이 논문에서는 이미지 자르기, 회전, 뒤집기, 색상 변환과 같은 7 가지 데이터 증강 방법을 사용하여 확장된 혀 영상 학습 데이터 세트를 생성하였다. 데이터 증강 방법의 성능을 확인하기 위하여 InceptionV3, EfficientNet, ResNet, DenseNet 등과 같은 전이 학습 모델을 사용하였다. 실험 결과 데이터 증강 방법을 적용함으로써 혀 분할의 정확도를 5~20% 향상시켰으며 기하학적 변환이 색상 변환보다 더 많은 성능 향상을 가져올 수 있음을 보여주었다. 또한 기하학적 변환 및 색상 변환을 임의로 선형 조합한 방법이 다른 데이터 증강 방법보다 우수한 분할 성능을 제공하여 InveptionV3 모델을 사용한 경우에 94.98 %의 정확도를 보였다.

Abstract A large volume of data will improve the robustness of deep learning models and avoid overfitting problems. In automatic tongue segmentation, the availability of annotated tongue images is often limited because of the difficulty of collecting and labeling the tongue image datasets in reality. Data augmentation can expand the training dataset and increase the diversity of training data by using label-preserving transformations without collecting new data. In this paper, augmented tongue image datasets were developed using seven augmentation techniques such as image cropping, rotation, flipping , color transformations. Performance of the data augmentation techniques were studied using state-of-the-art transfer learning models, for instance, InceptionV3, EfficientNet, ResNet, DenseNet and etc. Our results show that geometric transformations can lead to more performance gains than color transformations and the segmentation accuracy can be increased by 5% to 20% compared with no augmentation. Furthermore, a random linear combination of geometric and color transformations augmentation dataset gives the superior segmentation performance than all other datasets and results in a better accuracy of 94.98% with InceptionV3 models.

Key Words : Data augmentation, Deep Learning, Tongue segmentation, Transfer learning

Ⅰ. Introduction

The digitization and standardization of tongue diagnosis are the main development directions of modern tongue diagnosis[1]. When collecting tongue images, images usually contain tongue and background areas

This research was supported by Wonkwang University in 2019.
*College of Information and Computer Engineering, Pingxiang University
Department of Computer Engineering, Wonkwang University
**Corresponding Author : Department of Computer Engineering, Wonkwang University(stjung@wku.ac.kr)
Received August 27, 2020 Revised September 09, 2020 Accepted September 14, 2020

such as face, lip and teeth, and these areas have nothing to do with tongue diagnosis. In order to guarantee accuracy of analysis results and eliminate interference of irrelevant factors, the tongue body needs to be segmented from the complex background information. It ensures that tongue images analyzed by the automated analysis system have a pure color background (usually black or white) except for the part of tongue body. Accurate tongue segmentation helps to obtain correct automatic tongue diagnosis results[2].

In recent years, deep learning models show remarkable performance in the field of computer vision, such as object detection and semantic segmentation[3]. However, the limited amount of training data can inhibit the performance of deep learning model which often need very large quantities of data on which to train to avoid overfitting. In automatic tongue segmentation, the availability of annotated tongue images is often limited because of the difficulty of collecting and labeling the tongue image datasets in reality. In this case, the trained model may suffer from problems such as degraded accuracy and poor generalization ability, which affects the actual use of the model.

In image semantic segmentation tasks, small datasets usually use deep transfer learning and data augmentation to improve the accuracy and generalization ability of the model[4,5]. The transfer learning strategy allows the network to simply extract the primary features of network weights from model trained on a large source dataset(e.g., ImageNet ILSVRC) and apply them to image segmentation tasks. Meanwhile, transfer learning can avoid overfitting and speed up network training. Data augmentation has been widely used to deal with insufficient training data by artificially expanding the training dataset with label preserving transformation[6,7,8]. The goal of data augmentation is to increase the diversity of training dataset and improve the robustness of the model. The most common data augmentation techniques include flipping, rotation, random cropping, color transformation and noise injection techniques. Other complex data augmentation techniques synthesize a new image from Neural Style Transfer(NST) or from Generative Adversarial Nets(GANs) [3,9,10].

This paper studies data augmentation techniques for deep transfer learning-based tongue segmentation to find a better data augmentation method which can improve segmentation accuracy. The remaining part of this paper is organized as follows. Section 2 introduces current researches and related applications of data augmentation techniques in deep learning models. Section 3 describes the details of the proposed method. Section 4 reports the experiments and results. Finally, Section 5 presents the conclusion and future works.

II. Related Work

Data augmentation has been widely used in CNN architecture. At present, common convolutional networks include AlexNet, VGGNet, GoogleNet, ResNet , Inception-V3 and so on. As early as 2012, the AlexNet CNN architecture developed by Krizhevsky et al. adopted data augmentation to effectively improve the training network performance. In their experiments , they use data augmentation to increase the size of the dataset by a magnitude of 2048. This is done by randomly cropping 224 × 224 patches from the original images, flipping them horizontally, and changing the intensity of the RGB channels using PCA color augmentation[11]. This data augmentation was helpful for reducing overfitting when training a deep neural network and degraded the error rate of the model by over 1%, which also inspired researchers to study and apply data augmentation methods.

L. Huang et al.[12] evaluate three different data augmentation methods, i.e. rotation, flipping, and Gaussian noise via a state-of-the-art deep learning-based modulation classifier. Their results show that all three augmentation methods can improve the classification accuracy and lead to a simplified deep learning model and a shorter classification response time. M. Frid-Adar et al.[13] present a data augmentation method that first use classical data augmentation such as translation, rotation, scaling, flipping and shearing to enlarge the training set and then further enlarge the data size and its diversity by using Generative Adversarial Networks (GANs) to generate synthetic medical images. They demonstrate that the classical data augmentation techniques yielded 78.6% sensitivity and 88.4% specificity, while the results significantly increased to 85.7% sensitivity and 92.4% specificity by applying GAN techniques for synthetic data augmentation. Both are more excellent than the original dataset.

These state-of-the-art studies have shown that data augmentation plays a crucial role in the final recognition performance and generalization ability of deep convolutional neural networks. Expanding the training dataset is the fastest way to improve accuracy. However, the existing researches on tongue segmentation are almost all focused on new and improved models [1, 3], and there are few related work on data augmentation for tongue segmentation in the literature. The most relevant work is the data augmentation method proposed in [14]. The author adopts the method of rotation and horizontal flip to generate new training samples after rotating each image in the training dataset by an angle of 10°, a rotation angle of 20°, a rotation angle of 30°, and a horizontal flip sequentially. Although the rotation and flip augmentation methods improve the accuracy of segmentation, these methods do not take into count that the accuracy of tongue segmentation is affected by various factors such as tongue shape, color and the illumination changes. Therefore, an efficient augmentation method for insufficient tongue image dataset is still absent.

In this work, we further investigate the application of data augmentation for tongue segmentation by comparing 7 different data augmentation schemes, including a novel random linear combination augmentation, and the impact of the same data augmentation on different backbone networks. Unlike the existing methods mentioned above, we simulate changes in color and light due to camera settings and tongue characteristics by modifying the saturation, contrast, and brightness and utilize affine transformation to reproduce the distortions of the camera and create new tongue shape. The tongue image dataset is expanded through random linear combination. Experimental results show that the proposed data augmentation method is not only feasible, but also effective.

III. Proposed Method

Fig. 1. shows the proposed transfer learning model training method based on data augmentation. We use the U-net model for transfer learning in the deep learning framework of TensorFlow and Keras and choose different popular CNN models trained on ILSVRC ImageNet 2012 datasets as pre-trained models. After the original tongue image dataset is expanded with different data augmentation schemes, we can get various augmented datasets that can be used for further transfer learning training to obtain a new network model. Finally, we conduct the comparative experiments on the same test dataset with these new models and examine the segmentation results to find the best data augmentation scheme.

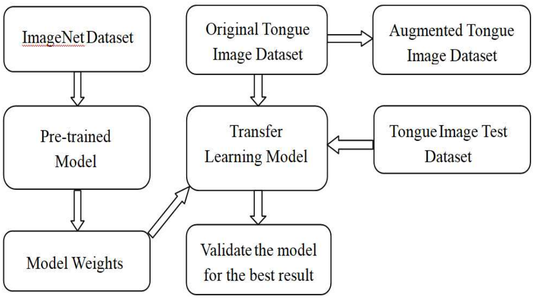

ImageNet Dataset
Original Tongue Image Dataset
Augmented Tongue Image Dataset
Pre-trained Model
Transfer Learning Model
Tongue Image Test Dataset
Model Weights
Validate the model for the best result

Fig. 1. Transfer Learning Model Training Based on Data Augmentation

3.1 Dataset and Data Augmentation

3.1.1 Dataset and Preprocessing

The tongue image dataset contains 180 tongue images for training and 60 tongue images for validation downloaded from the Internet. The validation dataset has the same situation with the training dataset. Fig. 2. shows some sample images from the original tongue image dataset. All images are in BMP format and have a size of 576×768 pixels.

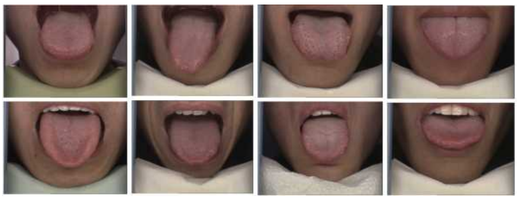


Fig. 2. Sample Data from Original Tongue Image Dataset

We applied these different augmentations described in Section 3.1.2 to the 180 training images. After every augmentation multiple images were generated and saved in a separate folder with the name of the corresponding augmentation technique. The size of each augmented dataset was 2040 images. Finally, we utilized these augmented datasets to fine-tune the pre-trained model. The full dataset was divided into a training dataset containing 1980 tongue images and a validation dataset containing 60 images.

Furthermore, we built the tongue image test dataset in an open acquisition environment. The test dataset was composed of 200 tongue images in five different situations, some of which were captured by using different image acquisition devices such as mobile phones and cameras in different environments and some of which were downloaded from the Internet. Therefore, the tongue images in test dataset have complex and variable lighting environments, different sizes and shapes of tongues, and varying positions compared with the original training dataset. Some sample images of the test dataset are shown in Fig. 3.

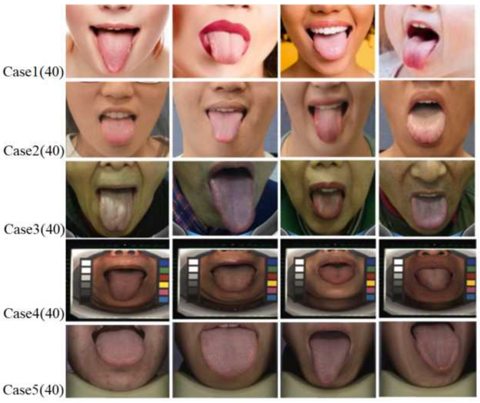


Fig. 3. Sample Data from Tongue Image Test Dataset

3.1.2 Data Augmentation Techniques

We used seven different data augmentation schemes to generate seven new training datasets. Each new training dataset consists of the original training images in addition to the training images augmented by one of the techniques shown in Table 1. Fig. 4. shows sample augmented images of all schemes.

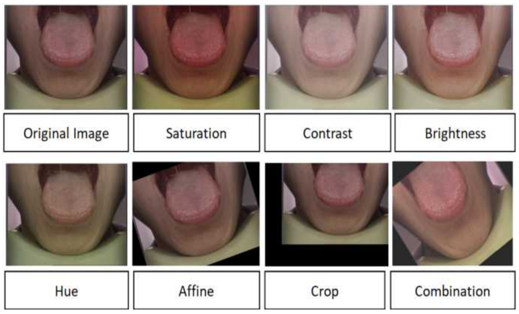


Fig. 4. Examples of Augmented Images

3.2 Transfer Learning Model

The emergence of U-net has greatly promoted the research of medical image segmentation[15]. It won the Data Science Bowl Cells' Nuclei Segmenting Challenge 2018 by a large margin. We used the U-net model based on different pre-trained weights backbone on the tongue segmentation task. All backbones have weights trained on 2012 ILSVRC ImageNet dataset. We chose these networks as they represent the state of the art in image segmentation.

Table 1. Augmentation Scheme

SCHEME	DISCRIPTION
Saturation	Modify the saturation of images by sampling random values from the discrete uniform range [-50, 50], and adding them to the saturation, i.e. to the S channel in HSV colorspace.
Contrast	Adjust image contrast by scaling pixel values to 255*((P/255)**gamma).where P is a pixel value and gamma is sampled uniformly from the interval [0.5, 2.0] (once per image)
Brightness	Modify the brightness of images. Convert each image to a colorspace with a brightness-related channel, extract that channel, add between -50 and 50 and convert back to the original colorspace.
Hue	Add random values to the hue of images. Sample random values from the discrete uniform range [-16,16], convert them to angular representation and add them to the hue, i.e. to the H channel in HSV colorspace.
Affine	Rotate the image by -30 to 30 degrees, shear with the value ranges between -20 and 20, scale to 80-120% of their size.
Crop And Pad	Randomly crop the original image. The crop has -0.25 to 0.25 of the original size (negative values result in cropping, positive in padding)
Linear Combination	Linearly combine the above 6 augmentations together and randomly flip the image horizontally.

The U-Net architecture[16] based on VGG16 backone is shown in Fig. 5. It consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. The contracting path has a FCN-like architecture that extracts features with two 3 × 3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. The expanding path uses a 2x2 convolution ("up-convolution") for upsampling the feature map to reduce the number of feature maps while increasing their dimensions. Feature maps from the contracting path are copied to the expanding path to avoid losing border pixels. At the final layer a 1x1 convolution is used for processing the feature maps to generate a segmentation map that categorizes each pixel of the input image.

IV. Results and Discussion

4.1 Evaluation Metrics

In order to observe the impact of data augmentation on the tongue segmentation model, it is necessary to evaluate all the models trained by using augmented datasets. We used the common evaluation metric for semantic image segmentation, Mean Intersection over Union (MIoU) to measure the quality of the training model. The higher the MIoU, the closer the predicted tongue area to the real tongue area, that is, the data augmentation scheme used can better improve the segmentation accuracy of the model. MIoU is calculated as follows:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}$$

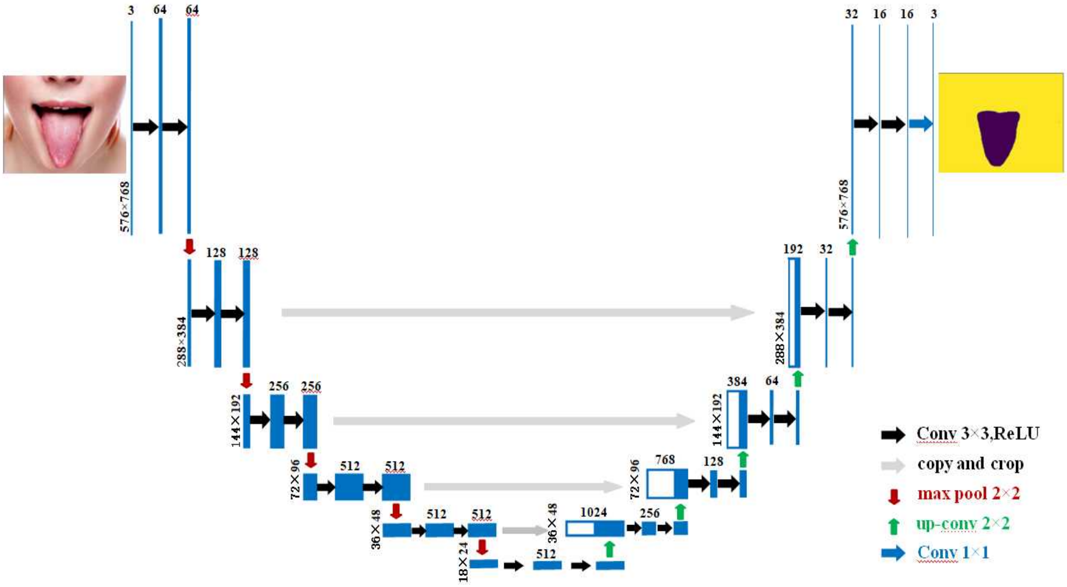


Fig. 5. The U-net Model Architecture Based on VGG16 Backbone

In above equation, i is the ground truth and j is the predicted value while p_{ij} represents the total number that i is predicted as j.

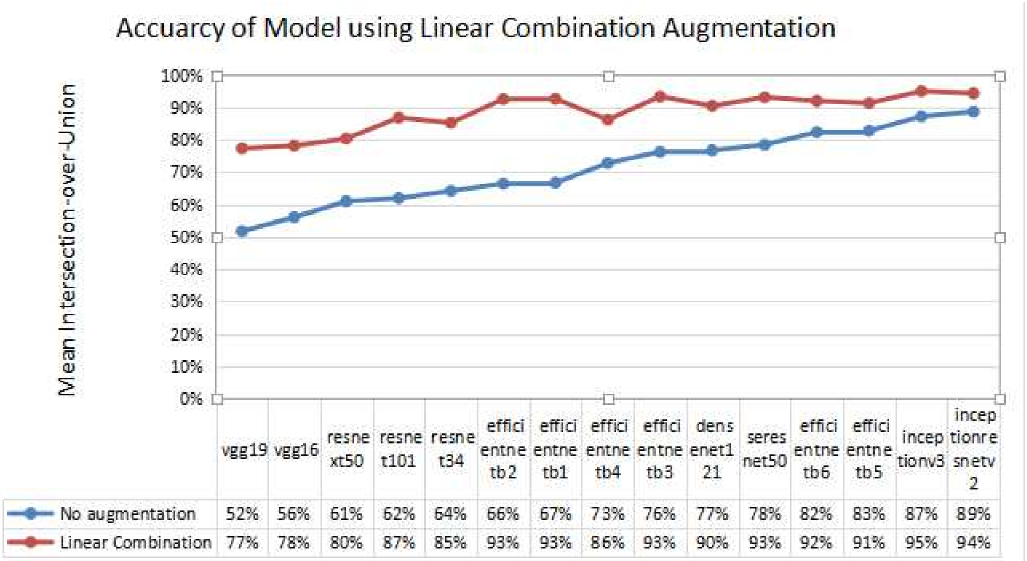
4.2 Experiment and Analysis

We implemented the proposed method in the Ubuntu 18.04 operating system using NVIDIA GeForce RTX 2080Ti with 11GB of memory. Keras and TensorFlow were used as deep learning framework in python to build the network model, and GPU technology was utilized to accelerate the experimental process. In this work, we performed data augmentation schemes to training dataset and apply transfer learning to fine-tune the pre-trained models. To validate the performance of the proposed method, a set of experiments were conducted using an original dataset combined with generated dataset by various augmentations described in Section 3.1.2. We used 15 different pre-trained network models(such as VGG, ResNext, ResNet, EfficientNet, DenseNet, SE-ResNet, inceptionV3, inceptionResNetV2 and etc.) on the test dataset described in section 3.1.1 for comparative experiments to find a network with the highest MIoU under the same conditions. Table 2 summarizes the experimental results of model training based on various augmentation schemes.

The experimental result illustrates that the segmentation accuracy of the pre-trained model using the dataset generated by the Linear Combination scheme is much higher than other datasets. The average accuracy of traditional methods(Saturation, Contrast, Brightness, Hue, Affine, Crop and Pad) was between 70% and 79% while the average segmentation accuracy can be increased by 10% to 19% with the linear combination augmentation scheme. The Linear Combination scheme greatly improves the segmentation

Table 2. Experimental Results of Model Training Based on Various Augmentation Schemes

MIoU / Backbone	No Augmentation	Data Augmentation Scheme						
		Saturation	Contrast	Brightness	Hue	Affine	Crop and Pad	Linear Combination
vgg19	0.5176	0.6599	0.6938	0.6377	0.5958	0.6271	0.6391	0.7734
vgg16	0.5606	0.6485	0.6839	0.7045	0.5984	0.6628	0.6428	0.7816
resnext50	0.6100	0.6378	0.6251	0.6102	0.6337	0.6842	0.6592	0.8037
resnet101	0.6194	0.6340	0.6434	0.6335	0.6007	0.6661	0.6115	0.8683
resnet34	0.6418	0.6439	0.6690	0.6245	0.6155	0.6903	0.7588	0.8524
efficientnetb2	0.6644	0.6989	0.8556	0.7755	0.6479	0.8604	0.8501	0.9255
efficientnetb1	0.6673	0.6447	0.8461	0.8658	0.6388	0.8849	0.8900	0.9262
efficientnetb4	0.7281	0.6816	0.8522	0.8454	0.8009	0.8620	0.8066	0.8622
efficientnetb3	0.7628	0.7652	0.8827	0.8445	0.7559	0.8149	0.7883	0.9333
densenet121	0.7676	0.7430	0.6653	0.7130	0.6563	0.7403	0.7895	0.9046
seresnet50	0.7846	0.7470	0.8471	0.8373	0.7222	0.8384	0.8660	0.9312
efficientnetb6	0.8234	0.8532	0.9072	0.8842	0.8592	0.8779	0.8802	0.9197
efficientnetb5	0.8280	0.8579	0.8716	0.8702	0.8544	0.8388	0.8177	0.9132
inceptionv3	0.8720	0.6916	0.8498	0.9095	0.8059	0.8364	0.8615	0.9498
inceptionresnetv2	0.8868	0.7759	0.8673	0.8289	0.7843	0.9066	0.9101	0.9434
Average	0.72	0.71	0.78	0.77	0.70	0.79	0.78	0.89



	vgg19	vgg16	resnext50	resnet101	resnet34	efficientnetb2	efficientnetb1	efficientnetb4	efficientnetb3	densenet121	seresnet50	efficientnetb6	efficientnetb5	inceptionv3	inceptionresnetv2
No augmentation	52%	56%	61%	62%	64%	66%	67%	73%	76%	77%	78%	82%	83%	87%	89%
Linear Combination	77%	78%	80%	87%	85%	93%	93%	86%	93%	90%	93%	92%	91%	95%	94%

Fig. 6. Accuracy of Different Pre-training Mode\ls Based on Linear Combination Scheme

performance of all 15 different pre-trained network models. Fig. 6. visually shows the improvement effect of the Linear Combination scheme on each segmentation model. This scheme yields the best MIoU values for all fifteen networks, for instance, 94.98% for Inception-v3, 94.34% for inceptionresnetv2, 93.33% for efficientnetb3 and also has the highest MIoU for the validation dataset in all fifteen networks.

In general, geometric transformation improves all fifteen networks better than color transformation. Contrast and Brightness schemes resulted in better MIOUs for almost all networks than Saturation and Hue schemes in color transformation. The Saturation scheme shows little improvements for VGG, ResNext, ResNet, EfficientNet networks, but produce worse results than no augmentation with DenseNet, SE-ResNet, inceptionV3 and inceptionResNetV2 networks. Hue scheme performed worse than other augmentations, indicating that the generated images might not able to preserve relevant features from the original tongue image. Geometric transformation-Affine, Crop And Pad had more consistent improvements among almost all networks. This shows that Affine, Crop And Pad can preserve a large quantity of image information in the original training dataset, leading to higher segmentation accuracy. Contrast, Brightness, Affine, Crop And Pad schemes can significantly improve the performance of EfficientNet and SE-ResNet. Compared with no augmentation, the segmentation accuracy can be increased by 5% to 20%.

V. Conclusion

In this paper, we studied tongue image data augmentation techniques for transfer learning based tongue segmentation to provide both a larger dataset for the training of deep neural networks and improve the generalization ability of segmentation models with limited data and enhance the diversity of data. Specifically, seven typical augmentation techniques, i.e., Contrast, Brightness, Saturation, Hue, Affine, Crop And Pad, and Linear Combination were studied based on a well-known U-net model. We first generated seven augmented datasets for training the state-of-the-art transfer learning models to segment the tongue image. Then, various augmentation results were compared and numerically evaluated to show the better augmentation techniques. All numerical results show that geometric transformation achieve higher segmentation accuracy than color transformation. Meanwhile, the random linear combined augmentation strategy of geometric and color transformation can further improve the segmentation accuracy, especially in the case of insufficient training samples. The random linear combination of multiple augmentation techniques based dataset gives a superior result than all other datasets and yields the better accuracy of 94.98%.

REFERENCES

[1] L. Wang, X. He, Y. Tang, P. Chen, and G. Yuan, "Tongue Semantic Segmentation Based on Fully Convolutional Neural Network," *International Conference on Intelligent Computing, Automation and Systems,* pp. 298-301, 2019.

[2] B. Lin, J. Xie, C. Li, and Y. Qu, "Deeptongue: Tongue Segmentation Via Resnet," *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 1035-1039, 2018.

[3] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R.N. Gunn, A. Hammers, D.A. Dickie, M.D. Hernández, K.M. Wardlaw, and D. Rueckert, "GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks," *Available:https://arxiv.org/abs/1810.10863.*

[4] L. Taylor and G. Nitschke, "Improving Deep Learning with Generic Data Augmentation," *IEEE Symposium Series on Computational Intelligence,* pp. 1542-1547, 2018

[5] F. Perez, C.N. Vasconcelos, S. Avila, and E. Valle. "Data Augmentation for Skin Lesion Analysis." *ISIC Skin Image Analysis Workshop and Challenge,* pp. 303-311, 2018.

[6] J. Rama, C. Nalini, and A. Kumaravel, "Image Pre-Processing: Enhance the Performance of Medical Image Classification Using Various Data Augmentation Technique," *ACCENTS Transactions on Image Processing and Computer Vision,* vol. 5, no. 14, pp. 7-14, Feb. 2019.

[7] I. Sirazitdinov, M. Kholiavchenko, R. Kuleev, and B. Ibragimov, "Data Augmentation for Chest Pathologies Classification," *IEEE 16th International Symposium on Biomedical Imaging,* pp. 1216-1219, 2019.

[8] S. Kayal, F. Dubost, H. Tiddens, and M. de Bruijne, "Spectral Data Augmentation Techniques to Quantify Lung Pathology from CT-Images," *IEEE 17th International Symposium on Biomedical Imaging,* pp. 586-590, 2020.

[9] J. Pandian, G. Geetharamani, and B. Annette, "Data Augmentation on Plant Leaf Disease Image Dataset Using Image Manipulation and Deep Learning Techniques," *IEEE 9th International Conference on Advanced Computing,* pp. 199-204, 2019.

[10] P. Dimitrakopoulos, G. Sfikas, and C. Nikou, "ISING-GAN: Annotated Data Augmentation with a Spatially Constrained Generative Adversarial Network," *IEEE 17th International Symposium on Biomedical Imaging,* pp. 1600-1603, 2020.

[11] C. Shorten and T.M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data,* vol. 6, article. 60, pp. 1-48, 2019.

[12] L. Huang, W. Pan, Y. Zhang, L. Qian, N. Gao, and Y. Wu, "Data Augmentation for Deep Learning-Based Radio Modulation Classification," *IEEE Access,* vol. 8, pp. 1498-1506, Dec. 2019.

[13] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," *IEEE 15th International Symposium on Biomedical Imaging,* pp. 289-293, 2018.

[14] T. Yang , Y. Yoshimura, A. Morita, T. Namiki ,and T. Nakaguchi, "Fully Automatic Segmentation of Sublingual Veins from Retrained U-Net Model for Few Near Infrared Images," *The Ninth International Workshop on Image Media Quality and its Applications, Available: https://arxiv.org/abs/1812.09477,* 2018.

[15] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *Available: https://arxiv.org/abs/2001.05566,* 2020.

[16] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention,* pp. 234-241, 2015.

Author Biography

진 홍(Hong Chen) [Member]

- July. 1997 ~ Sept. 2019 :Pingxiang University, China, Professor
- Jan. 2010 : Nanchang University, China, MS
- Mar. 2019 ~ current : Department of Computer Engineering, Wonkwang University, Korea, Ph.D. student

〈Research Interests〉 Image Processing & Machine Learning

정 성 태(Sung-Tae Jung) [Member]

- Feb. 1989 : Computer Enineering of Seoul National University, MS
- Aug. 1994 : Computer Enineering of Seoul National University, Ph.D.
- Mar. 1995 ~ current : Wonkwang Univ., Dept. of Computer and Software Engineering, Professor

〈Research Interests〉 Image Processing & Machine Learning, Computer Graphics